Microbiome

RESEARCH

Open Access

Extensive paralogism in the environmental pangenome: a key factor in the ecological success of natural SAR11 populations

Carmen Molina-Pardines[1†], Jose M. Haro-Moreno[1†], Francisco Rodriguez-Valera[1] and Mario López-Pérez[1*]

Abstract

Background The oceanic microbiome is dominated by members of the SAR11 clade. Despite their abundance, challenges in recovering the full genetic diversity of natural populations have hindered our understanding of the eco-evolutionary mechanisms driving intra-species variation. In this study, we employed a combination of single-amplified genomes and long-read metagenomics to recover the genomic diversity of natural populations within the SAR11 genomospecies Ia.3/VII, the dominant group in the Mediterranean Sea.

Results The reconstruction of the first complete genome within this genomospecies revealed that the core genome represents a significant proportion of the genome (~81%), with highly divergent areas that allow for greater strain-dependent metabolic flexibility. The flexible genome was concentrated in small regions, typically containing a single gene, and was located in equivalent regions within the genomospecies. Each variable region was associated with a specific set of genes that, despite exhibiting some divergence, maintained equivalent biological functionality within the population. The environmental pangenome is large and enriched in genes involved in nutrient transport, as well as cell wall synthesis and modification, showing an extremely high degree of functional redundancy in the flexible genome (i.e. paralogisms).

Conclusions This genomic architecture promotes polyclonality, preserving genetic variation within the population. This, in turn, mitigates intraspecific competition and enables the population to thrive under variable environmental conditions and selective pressures. Furthermore, this study demonstrates the power of long-read metagenomics in capturing the full genetic diversity of environmental SAR11 populations, overcoming the limitations of second-generation sequencing technologies in genome assembly.

Keywords SAR11, Pangenome, Mediterranean Sea, Long-read metagenomics, Flexible genome, Paralogs, Single-amplified genomes

Background

The SAR11 clade, also known as the order *Pelagibacterales*, is a group of free-living chemoheterotrophic bacteria that numerically dominate the near-surface epipelagic waters of the ocean, representing 20–40% of all prokaryotic cells [1–5]. Members of this group have been subjected to evolutionary constraints, including streamlined genomes and small cell sizes with high surface-to-volume ratios [6], enabling them to thrive in the oligotrophic conditions that characterise the open ocean.

†Carmen Molina-Pardines and Jose M. Haro-Moreno contributed equally to this work.

*Correspondence:
Mario López-Pérez
mario.lopezp@umh.es
[1] Evolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, Apartado 18, San Juan, Alicante 03550, Spain

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Despite the abundance and cosmopolitan distribution of these microbes, limitations in recovering the full genetic richness of their natural populations have prevented the unravelling of the relationship between evolution and microbial ecology from a genomic point of view. This is due mostly to two factors: first, the inherent difficulty in obtaining pure cultures of these microbes, and second, the enormous microdiversity that paradoxically makes them very resistant to assembly and binning with short-read sequencing [7–9]. Thus, the number of metagenome-assembled genomes (MAGs) is not only scarce and unreliable, but the flexible genome is also under-represented due to its lower representation in metagenomes and high strain-to-strain variability [10, 11]. Consequently, the majority of genomic data is derived from a limited number of successful cultivation efforts and genomes obtained through single-cell genomics [12–17]. The enrichment in single-amplified genomes (SAGs) in databases [17] has, therefore, improved our knowledge about this order. This has enabled the delineation of groups of genomes with a close phylogenomic relationship and similar relative abundances in global metagenomic samples. These groups have been designated as genomospecies [7]. Additionally, evidence has emerged indicating that genomic diversity within these populations is associated with specific ecological contexts. These include adaptations to anoxic conditions in oxygen minimum zones [15] and mesopelagic regions [18], to the availability of certain nutrients [7], geographical distribution [19, 20], evolutionary transitions between fresh and salt water [21], as well as the development of evolutionary models to explain the diversification of distinct lineages within a population [8]. For instance, in the genomospecies Ia.3/VII, the acquisition of a set of genes involved in phosphonate utilisation was found to be correlated with its high abundance in the surface waters of the Mediterranean Sea, which are characterised by limiting phosphate (P) concentrations in comparison to the global ocean [7].

Nevertheless, the full diversity of the flexible genome within a population of SAR11, i.e. cells belonging to the same genomospecies, in their natural environment, as well as the genomic mechanisms and patterns that govern their dynamics, has not yet been addressed with current techniques. Previous studies have investigated the microbial genomic diversity of this group, using highly divergent genomes that are representative of the entire order, due to the lack of reference genomes [19, 22]. Therefore, in this study, we have used a novel approach to recover the full genomic richness of environmental samples by combining third-generation metagenomics applied to four samples from the Western Mediterranean Sea with SAGs [10, 17, 23, 24]. Long-read metagenomics has the potential to provide sequences long enough (typically 5–20 Kb) to carry complete genes, or entire operons, without the need of the assembly, and thereby can contain the flexible genes that make up a genomic island [10]. We focused our work on a single genomospecies, Ia.3/VII (also termed gMED, [7]), as it was the one that recruited the most in the Mediterranean [7]. Reconstruction of the complete first genome within this genomospecies has revealed that the core genome represents a significantly large (~81%) part of their genome. The use of the core genes to delimit the boundaries of flexible genomic regions among gMED SAGs and the reconstructed genome, i.e. genomic islands (GI), showed that the flexible genome was concentrated in small regions containing mostly a single gene. The recovery of genetic diversity from both SAGs and metagenomic reads demonstrated that each GI was associated with a specific gene pool. While some degree of divergence was observed, the biological functionality within the population was found to be equivalent. Therefore, although the environmental pangenome is large and mostly enriched in (a) genes involved in nutrient uptake (transporters) and metabolism and (b) genes related to cell wall synthesis and modification, the degree of functional redundancy (i.e. paralogs) is extremely high. Furthermore, this study demonstrates the suitability of using long-read metagenomics to recover the full genetic diversity within environmental populations of microbes that are difficult to culture in the laboratory, but abundant in nature. This approach overcomes the assembly problems of second-generation sequencing by allowing the direct recovery of large genomic fragments of these microbes and the entire previously unknown flexible gene pool.

Results and discussion

To recover the genomic diversity of the gMED genomospecies, we first collected all available SAR11 genomes from the National Center for Biotechnology Information (NCBI, April 2021). After removing genomes that did not meet the established quality criteria of > 50% completeness and < 5% contamination, a total of 2098 genomes were used to construct a maximum-likelihood phylogenomic tree (Fig. S1). The majority of the genomes were SAGs (~97%), followed by 50 MAGs and 23 pure cultures [7, 17–19, 25]. Sixty-eight SAGs were classified within the gMED genomospecies (Fig. S1, Table S1). Among them, ten SAGs were collected from a single location in the Western Mediterranean Sea [7], and almost 35% of the gMED SAGs (#25) came from a single sample collected at the Bermuda Atlantic Time Series Study (BATS) station, which has similar physicochemical conditions to the Mediterranean. The remaining SAGs (#31) came from different oceanic regions [17] (Table S1). Pairwise

genome comparisons among gMED genomes showed that the minimum average nucleotide identity (ANI) value was 82.1%, with an average of 89.2%. These values are within the normal range for determining genomospecies or "sequence-discrete" populations within the SAR11 clade [7–9].

The absence of pure cultures within gMED represented a significant obstacle to performing a comparative genomics study, as no complete genome could be used as a reference. Fortunately, this increased genomic diversity revealed the presence of five genomes that belonged to the same clonal frame (ANI > 99%). The methodology previously employed in [26–28] was used to subject these genomes to a cross-assembly step, resulting in the generation of a complete and circular reference genome with a size of 1.29 Mb, referred to as SAR11-gMED complete composite genome (SAR11-gMED-CCG) (see Materials and Methods and Table S1).

Determination of the core genome and regions of high intra-population genomic variation

The reconstructed reference genome (SAR11-gMED-CCG) was used to identify the genomic regions exhibiting high genetic variability between strains of the gMED genomospecies (i.e. genomic islands [GIs]). To that end, all predicted genes from the 63 gMED SAGs (excluding the five genomes used to obtain the reference genome) were aligned against SAR11-gMED-CCG (Fig. 1A). To define a gene as part of the core genome, the sequence had to be present in at least 40% of the gMED SAGs with a minimum nucleotide identity of 80% (see Materials and Methods section and Fig. 1A). This threshold was

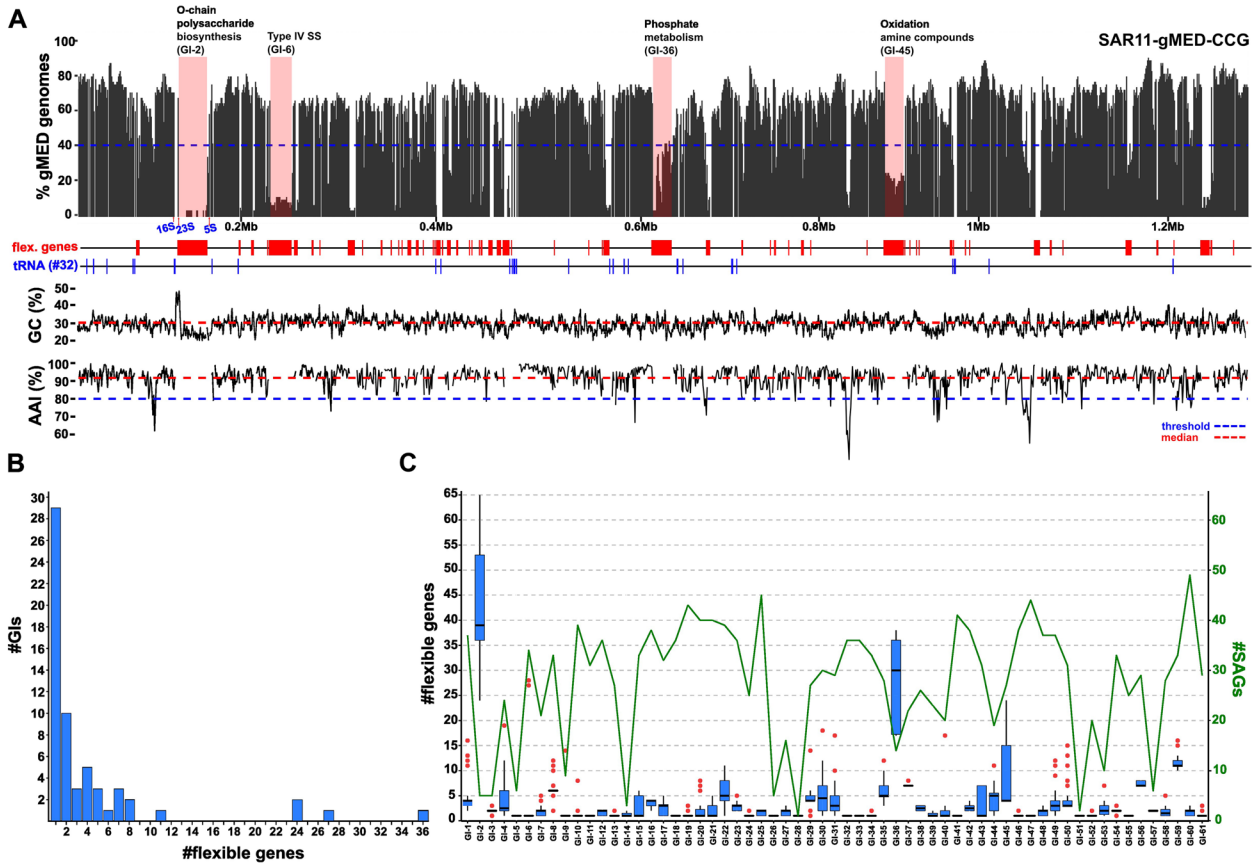

Fig. 1 Strategy to delimit flexible regions within the SAR11-gMED-CCG. **A** Percentage of available gMED SAGs (#63) containing orthologs to the SAR11-gMED-CCG predicted proteins. Flexible regions are reflected by gaps below a threshold of 40% of gMED genomes (blue dashed line), and those larger than 10 Kb are shaded in red. The position in the reference genome of the flexible regions as well as the tRNAs are indicated as red and blue lines respectively. The genome was rearranged with the *dnaA* gene at the beginning. The GC content of the reference genome is plotted, with the median value at 30% indicated by a red dashed line. The AAI of the core genes is also plotted, with a median value of 91.9% (red dashed line), orthologs with median low AAI [median below 80% (blue dashed line)]. **B** Number of flexible genes in GIs present in SAR11-gMED-CCG. **C** Boxplot representing the number of flexible genes of each GI found in gMED SAGs. A secondary axis with a green line shows the number of SAGs where GIs are found. AAI, average of amino acids identity; GIs, genomic islands; SAGs, single-amplified genomes

established due to the incomplete nature of SAGs (average ~ 70%, Table S1).

Once the genomic regions containing flexible genes were identified, their variable nature was confirmed by retrieving the full genetic diversity of these regions from the gMED genome SAG collection. To achieve this, the core genes located at the boundaries of the predicted variable regions were used to identify and extract all the flexible gene catalogue from the gMED genome dataset (see Materials and Methods section and Table S2). However, if the same version was identified in all genomes as in the reference genome, then that region was excluded as potentially flexible. Genomic comparison revealed the presence of eight such regions, encompassing a total of 26 genes. Despite meeting the abundance requirements (present in at least 40% of the gMED genomes), these regions had the same version across all SAGs, identical to that observed in the reference genome. However, the identity of these regions fell below the 80% threshold. Although it was considered that the results could be influenced by the number of SAGs used, metagenomic analysis of long reads (as described below) confirmed that these regions were not flexible, as the same version was also found in all reads. To analyse the evolutionary pressure acting on these genes, we examined the average amino acid identity (AAI) over all protein-coding sequences of the core genome using all SAGs of the gMED genomospecies. Figure 1A shows the AAI of each core genome protein of SAR11-gMED-CCG within gMED. The AAI of the core genome for gMED was 91.9%, while the 26 genes that were not included within the flexible genome had a significantly lower AAI compared to the genome-wide AAI [mean AAI 69.17% (standard deviation {SD}, ± 8.42)], suggesting an enrichment in non-synonymous changes. Likewise, the analysis identified 20 additional sequences with a low AAI, all of which were below 80%. The functional classification of these 46 sequences revealed that the most representative categories were those related to the biosynthesis of O-chain polysaccharides and transport systems (15.2% each) (Table S3). Besides, a large number (26.1%) of the sequences could not be functionally annotated.

These results suggest the presence of genomic regions within the core genome that exhibit a high degree of intrapopulation genomic variation. This is in agreement with previous studies that investigated, using a metagenomic screening approach, the impact of the genetic variations on amino acid residues from core genes using a single isolate of the Ia.3/V subclade (HIMB83) as a reference. In these studies, authors attributed these changes to adaptations to variations in abiotic factors, such as large-scale ocean current temperatures [9] or nitrogen concentration [29]. The presence of these allelic variations within the core genome indicates the existence of an additional evolutionary dynamic in microbes with streamlined genomes beyond gene acquisition by horizontal transfer. These highly divergent sequences are maintained in natural populations by high rates of recombination, which homogenises populations and endows the genomospecies with greater metabolic flexibility.

Genetic diversity of the flexible regions in the gMED SAGs population

Once the regions of high divergence were identified as part of the core genome, the total core genome in the reference genome, SAR11-gMED-CCG, was determined to be 1099 genes (1.05 Mb), representing a total of 81% of the genome. These numbers fit very well with previous reports on which the core genome of five SAR11 subclade Ia.1/I isolate genomes had 1060 orthologous clusters [22]. The remaining 258 genes, distributed across 61 GIs, constituted the flexible genome in SAR11-gMED-CCG (Fig. 1A). The two pangenome partitions showed differential genomic characteristics. The percentage of GC content in the core genome was 30, slightly higher compared to the flexible genome (28.5%). In fact, the larger and statistically significant (paired *t*-test, $p < 0.001$) difference was detected in GI-2, with a GC content of ~ 22% (Fig. 1A). This region encodes the O-chain biosynthesis gene cluster (OBC), and the variation in its genomic properties with respect to the core genome has been previously described [23]. The core genome exhibited a minimal intergenic spacing of only two base pairs, resulting in a high coding density of 98%. Conversely, the flexible genome had a lower coding density of 94% and a median intergenic spacer of 13 bp.

The size range observed for the 61 GIs was between 272 bp and 35.5 Kb, with a median of 1.4 Kb. This is consistent with the fact that 47.5% of the GIs contained only one gene. The number of flexible genes for each of the 61 GIs is shown in Fig. 1B and Table S2. According to the classical definition [30, 31], GIs are characterized by fragments longer than 10 Kb. Based on this criterion, only four of the 61 islands met this requirement. These include GI-2, which is related to the OBC (36 genes), GI-6, associated with the type IV secretion system (27 genes), GI-36, which is related to P metabolism (24 genes), and GI-45, which is involved in the oxidation of amine compounds (24 genes) (Fig. 1A). In a previous work, Wilhelm and collaborators [32] described the presence of four large hypervariable regions after metagenomic recruitment of the strain HTCC1062 (genomospecies Ia.1/I) in the Sargasso Sea. No canonical mobile genetic elements were found in any of the GIs, which is a common feature of microbes with streamlined genomes. Only one integrase was found to be associated with GI-35. The analysis of GI

boundaries indicated that seven of them were linked to a tRNA and one to a tmRNA, which are both recognised as integration hotspots (Fig. 1A). In four of them (GI-20, GI-30, GI-50, and GI-58), we found a partial repeat of the same tRNA within the island with more than 90% identity, identifying the insertion of a gene cassette (Fig. S2).

Figure 1C and Table S2 show the genomic features of the 61 GIs recovered from the gMED SAGs. All GIs identified in the reference genome were also present in the gMED genomes, except for GI-28, which was exclusively found in SAR11-gMED-CCG. A total of 60% of the GIs identified in the reference genome were present in at least 40% of the SAGs. Conversely, only 15% of the islands were found in fewer than 10 genomes (Fig. 1C). In some cases, this might be due to the significant length of the islands, such as the OBC genomic island (GI-2; 35.4 Kb), which was entirely recovered from only five SAGs. The variability of flexible genes within each island is shown in Fig. 1C. The analysis revealed that 60.7% of the GIs recovered from the SAG dataset had a median number of genes identical to those in the reference genome (Table S2).

Genomic diversity in natural gMED populations by long-read metagenomics

Given the significant prevalence of SAR11 in environmental samples, we included several PacBio CCS sequenced metagenomes from different ecological niches within the photic zone of an offshore Mediterranean location in our analysis (Table S4). The long-read sequences, capable of spanning flexible genomic regions, provide a powerful approach to enhancing the recovery of gMED genomic diversity.

First, we screened the PacBio CCS metagenomes for all the reads associated with gMED. To consider a good hit, we set a threshold that required that at least 50% of the proteins in a given read to have homology to any of the gMED SAGs with a minimum identity of 80%. To be more precise, reads belonging to other genomospecies were excluded by the search of a database with 2098 well-classified SAR11 genomes (see Materials and Methods section). Thus, the proportion of the resulting reads associated with gMED was 6.1% for reads in the upper photic (UP) sample, 2.4% in the winter sample (MIX), 1.8% in the deep chlorophyll maximum (DCM) sample, and 0.3% in the lower photic (LP) sample (Table S4). This agrees with previous results that indicated a preferential distribution of gMED in the upper epipelagic waters [7].

Following the same methodology as for the SAGs (see above), we identified long metagenomic reads that contained any of the GIs identified in the reference genome (SAR11-gMED-CCG). The majority of the GIs (59 out of 61), except for GI-2 and GI-36, were fully recovered from the reads (Fig. 2A and Table S2). This was expected, as these islands are among the largest and have the highest average number of flexible genes in the SAGs (Table S2). Thus, the metagenomic reads, which have an average size of 7.4 Kb, were not long enough to cover them. Conversely, GI-28, which was not recoverable with SAGs, was here recovered from the pool of environmental reads in the UP sample (Fig. 2A). In this sample, we also observed the highest number of reads associated with GIs, with 816, normalised per 100,000 reads in the metagenomic sample, followed by DCM and MIX (274 and 270 reads, respectively), and lastly the LP sample, with only 40 reads aligning to 54 GIs (Fig. 2A). The range of flexible genes recovered for each GI in the metagenomes varied similarly to that of the SAG collection, and nearly 50% of the GIs in the PacBio CCS data contained a single gene (Fig. S3). This percentage is comparable to that obtained in the reference genome (42.5%) and in the SAGs (47.5%) (Table S2). However, for those GIs with two or more genes, the Wilcoxon test showed a significant difference in the number of flexible genes, being generally a larger number in metagenomic reads than in the SAGs, in nearly one-third of the GIs (Table S2).

In order to evaluate the efficacy of long-read metagenomics (LR) in resolving genetic diversity, we conducted a comparative analysis with the classical short-read assembly (SRa) of the Illumina metagenomic data for the same environmental samples. The results demonstrated that LR consistently outperformed SRa, with an average improvement of over twofold in terms of gMED-associated protein and GI recovery (Fig. 2B). For instance, in the UP sample, where this genomospecies is more abundant, 6985 normalised proteins per 100,000 metagenome proteins were recovered in the LR metagenome, whereas in the SRa metagenome, it was 1328 proteins. Moreover, using SRa, we were unable to recover more than 18 GIs (out of 61 GIs) from SAR11-gMED-CCG in any of the samples, even considering that most of the GIs had only one or two genes (Fig. 2B). However, with LR the minimum recovered in any sample was 54 GIs in the LP sample. This is noteworthy, as these findings serve to reinforce the limited reliability of second-generation sequencing in accurately reconstructing the full genetic diversity, particularly that present in the flexible genome, of microbes with high microdiversity, such as the SAR11 clade. These analyses provide support for the use of LR sequencing as an efficient alternative to recover all the genetic diversity hidden so far in this type of microbes.

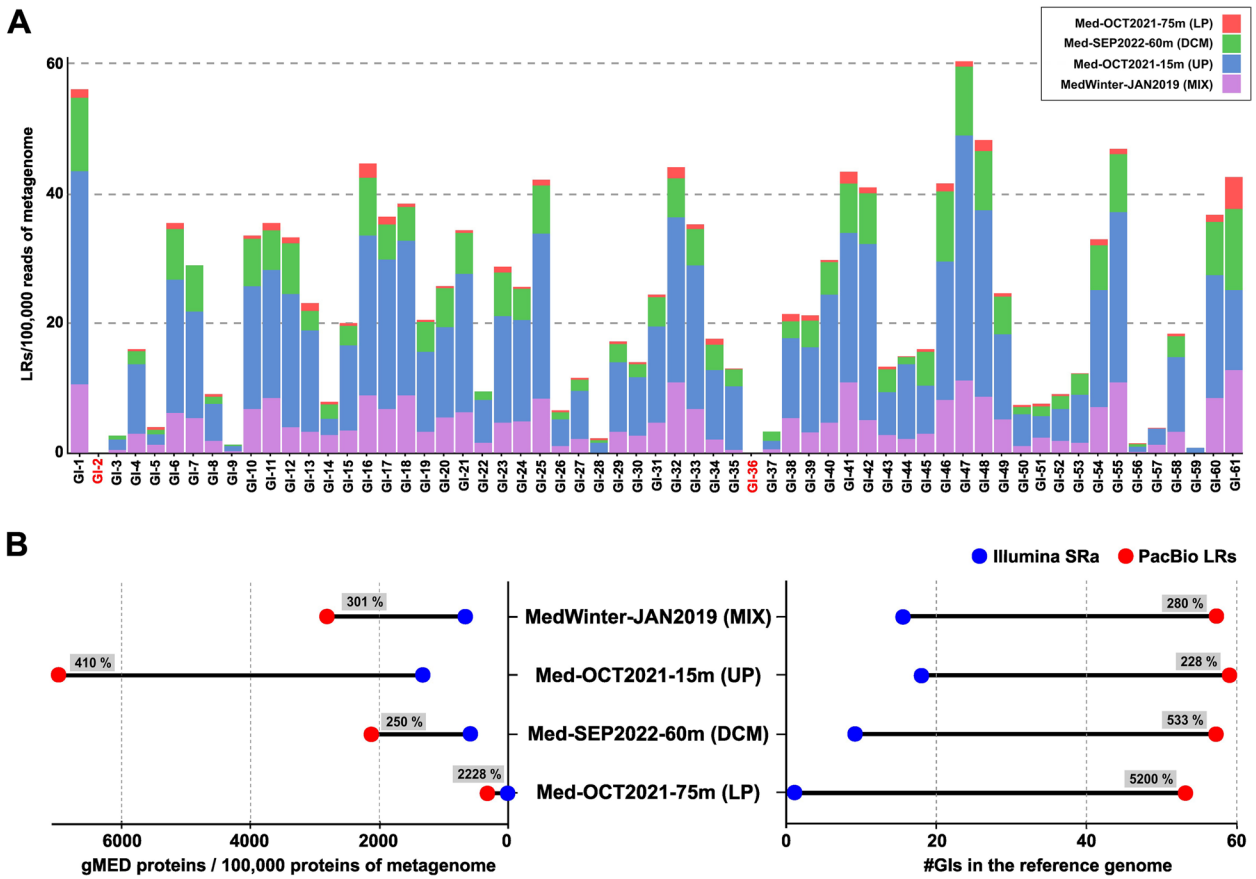

Fig. 2 **A** Number of LRs normalized by 100,000 reads in the metagenome that contained complete GIs. The metagenomes are represented by different colours. The islands that were not recovered in the LRs are marked in red (GI-2 and GI-36). **B** Comparison of the recovery of the number of gMED-associated proteins and GIs defined in SAR11-gMED-CCG, between PacBio (blue) and Illumina (red) metagenomes. The number of gMED proteins recovered was normalised by 100,000 proteins of the metagenome. The percentage increase is indicated for each sample. LRs, long reads; LP, lower photic; DCM, deep chlorophyll maximum; UP, upper photic; SRa, short-read assembly

The flexible gene pool within gMED

Once the flexible gene pool (henceforth the "flexome") associated with the reference genome GIs was recovered from both gMED datasets (SAGs and LRs), we analysed its diversity in terms of protein recovery and function. A total of 14,134 protein sequences were ascribed to the flexome, which were clustered at 80% amino acid identity, leading to a reduced set of 2217 clusters (ca. 60% were singletons at this threshold). We selected this threshold as it is the limit of the genomospecies classification [7], following the classical pangenome analysis in other microbes, that cluster sequences at the species level [28, 33–35]. Thus, considering that the core genome is about 1099 genes, we estimate that the size of the pangenome of the gMED species might be near 3300 genes. However, this estimate likely underrepresents the true size of the pangenome due to some limitations in our approach. First, our analysis used a single genome to identify hypervariable regions, potentially missing additional GIs present in other genomes. Given the small, streamlined and highly syntenic nature of the SAR11 genomes, we hypothesise that the number of additional GIs should be small and limited to one or two genes and that the number of new flexible genes after clustering should be also minimal (see below). Second, GI-2, the one that coded for the OBC, was barely recovered from SAGs, and due to the size of LRs, we could not increase its diversity. This region, of typically 46 Kb long [23], represents the most variable region in the genome, on each SAR11 cell codes for a completely different set of genes (see below). In a previous work, we could extrapolate that nearly 400 different OBCs might be present in a single Mediterranean sample (population), and therefore, the total number of proteins in the flexome should be in the order of tens of thousands [23].

Regardless of the possible limitations mentioned above, clustering at 80% identity, where the number of final clusters represented only 15% of the total flexome (14,134

sequences grouped into 2217 clusters), indicated the existence of high functional redundancy within the flexible genome. To corroborate this, we further clustered the sequences at 30% amino acid identity to group the flexible gene pool into paralogs [36]. In the end, 605 proteins (i.e. clusters) presented at least one paralog, while the remaining 605 clusters were singletons. This is noteworthy, as these 605 clusters containing paralogs represented 95.7% of the flexome (13,529 out of 14,134 proteins) (Fig. 3A). In the core genome, only 42 proteins (4%) were identified with at least one paralog (8.5% of the total proteins being paralogs) (Fig. 3A). The low number of paralogs is a common feature in microbes with streamlined genomes [7, 27, 28]. However, the high percentage of paralogs in the flexible genome of gMED is remarkable. Paralogs imply the same function but with adaptations to different environmental conditions [37]. For example, in the case of transporters paralog clusters might be involved in transport or ligand binding. This consequently results in a notable increase in the number of substrates that can be transported or modified within each cluster. These findings are in accordance with the substantial substrate diversity observed in the open ocean water column [38]. Therefore, these microbes maintain a substantial gene pool with analogous functionality within their natural populations, thereby giving rise to the emergence of numerous subpopulations that coexist in the same environment. This enables the species to respond and adapt expeditiously to a diverse array of environmental conditions, including fluctuations in specific micronutrients. Conversely, the high diversity observed in cell surface components, including genes involved in the synthesis of the outer glycosidic envelope and transporters, likely reflects their critical role as primary targets for phage recognition [39]. Phage-host interactions drive Red Queen dynamics [40], facilitating frequency-dependent negative selection that prevents selective sweeps and helps maintain genetic diversity within the population. These observations highlight the ecological importance of the environmental genomic diversity within SAR11 populations.

In light of the considerable functional redundancy observed within a single species in a natural population, our subsequent objective was to analyse the contribution of each GI to the flexome in terms of both raw (total) and paralogous (clustered with a threshold of 30% identity,

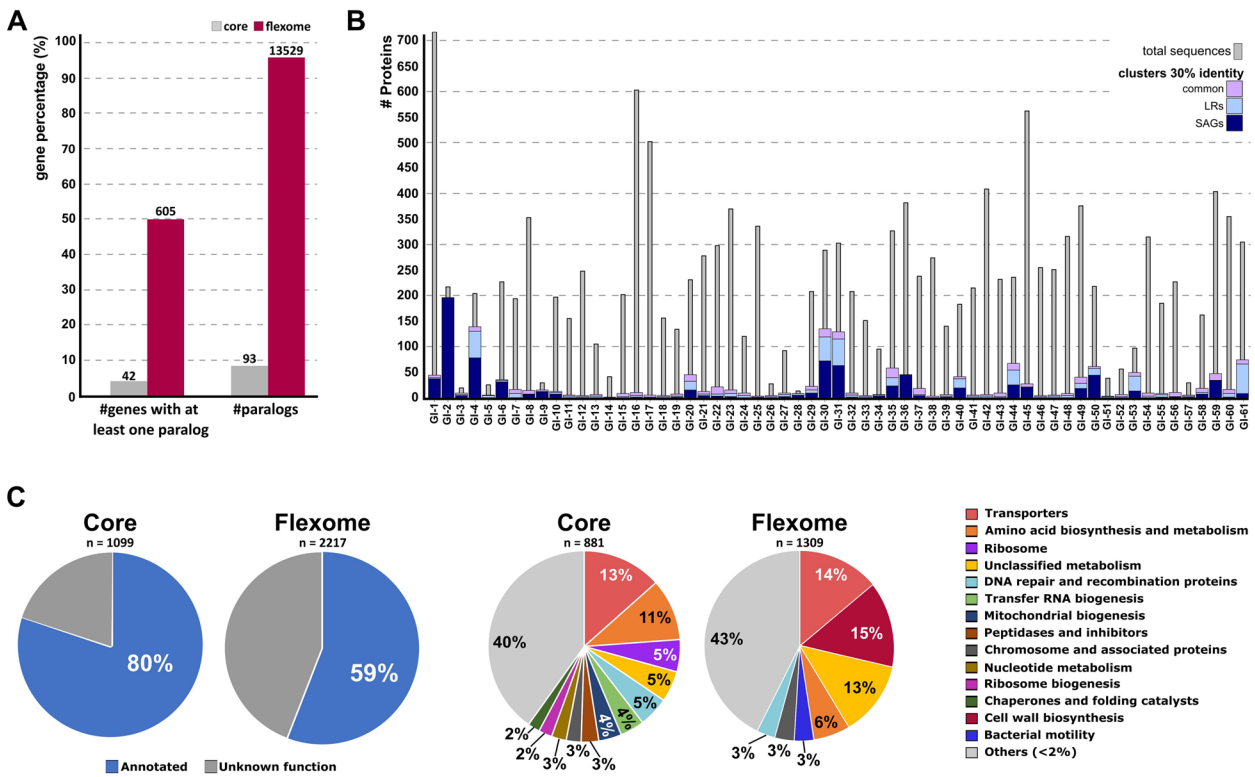

Fig. 3 **A** Percentage of genes that have at least one paralog and the total number of paralogs, both from the flexome and from the core of gMED, obtained from clustering at 30% identity. **B** Number of flexome proteins of each GI in both raw (total) and paralogous (clustered at 30% identity) sequences. Clusters that belong exclusively to SAGs or LRs, or that are present in both datasets, are indicated in different colours. **C** Pie charts with functional annotations for gMED core and flexome from KEGG Orthology database

[36]) sequences. The results demonstrated a distinct disparity in contribution, with islands such as GI-1, GI-16, GI-17, and GI-45 making the most significant contribution to the overall flexome (Fig. 3B and Table S2). However, following the clustering of paralogs, the number of sequences underwent a notable reduction. This phenomenon was observed in the majority of GIs, with the exception of GI-2, followed by GI-4, GI-30, and GI-31. It is noteworthy that most paralogs were derived from either the SAG dataset or the LR sequences, with only a limited number of paralog sequences encompassing proteins from both datasets (Fig. 3B and Table S2). Considering these findings, the incorporation of novel genomic sequences, for example, through LR metagenomics either from the same or from alternative geographical locations, such as the eastern Mediterranean or the Sargasso Sea, where they also exhibit high recruitment values [7], could potentially enhance the discovery of additional paralogs within gMED. The results also demonstrate that most GIs in natural populations possess equivalent biological functions and are located in analogous positions within the species genome, thereby facilitating transfer between subpopulations. Such islands have previously been described as flexible genomic islands (fGIs) [26, 41].

A general functional annotation of both components of the pangenome (core and flexible genome) was conducted using the KEGG Orthology database [42]. The proportion of proteins in the core genome that could be assigned to a functional category was 80%, while for the flexome, it was only 59% (Fig. 3C). This highlights the significant gaps in our knowledge about these microbes. The category of transporters was the most represented within the core genome, accounting for 13% of the total. This was followed by categories related to central metabolic processes, such as amino acid biosynthesis (11%) and translation (ribosomes category, 5%) (Fig. 3C). The flexome, on the other hand, showed an overrepresented in several categories, including glycosyltransferases, O-antigen nucleotide sugar biosynthesis, and lipopolysaccharide biosynthesis proteins. These could be grouped as cell wall biosynthesis and modification, representing 15% of the sequences. A further 14% was related to transporter category, followed by various enzymes with unknown specific targets (e.g. hydrolases, dehydrogenases) (13%) (Fig. 3C). The Pfam domain database [43] was employed to categorise the transporters, with the majority being annotated as ABC-transporter (ATP-binding model), EamA-like transporter, Pst and Phn operons, MlaA and MlaC proteins, Tripartite ATP-independent periplasmic transporter (DctPQM) and Tripartite tricarboxylate transporter (TctCBA), and ammonium transporter family. As a result of the inability to categorise a significant proportion of proteins, an alternative approach was taken, whereby proteins with transmembrane domains and those with signal peptides were also determined. The percentage of proteins with signal peptides, i.e. those that are translocated to the periplasmic or extracellular space, was found to be higher in the flexome than in the core genome (13.1% and 4.6%, respectively). On the other hand, both genome compartments exhibited a comparable proportion of proteins with transmembrane domains (approximately 25%).

To gain a comprehensive understanding of the distribution of this flexible gMED catalogue in other marine microbes that share a similar pelagic marine habitat, we used a large collection of SAGs from global seawater samples [17]. This dataset reliably represents the microbial diversity in the euphotic ocean [11]. At high identities (80–100%), a proportion of 80% of the flexible gMED catalogue was found to be mainly present in SAR11 SAGs. When the identity exceeded 60%, most of the matches (> 90%) were found to correspond to Pelagibacterales. In contrast, at lower identities (20–40%), the proportion was higher (98.6%) and was found in a total of 4751 SAGs (Fig. S4). The most prevalent order was SAR11 (Pelagibacterales) (#395), followed by different orders, such as HIMB59 (10.7%), closely related to SAR11, SAR86 (10.2%), PCC-6307 (Synechococcales) (9.4%), and Flavobacteriales (6.3%) (Fig. S4). The results indicate that a considerable number of flexome gMED sequences possess orthologs in widely distributed and abundant marine microbes.

Presence and expression of the gMED "flexome" in the global ocean

Given the large number of genes in the flexible genome of the gMED genomospecies, we evaluated whether these sequences might have an ecological role in the environment, i.e. they are present and actively transcribed. To achieve this, we recruited the flexible pool of genes against the *Tara* global ocean dataset [44, 45]. We only consider prokaryote-enriched metagenomes and their paired metatranscriptomes from stations that, for a given depth, they collected both kinds of omics samples [45]. Additionally, we filtered out samples where gMED genomes recruited, on average, < 10 reads per kilobase of genome per gigabase of metagenome (RPKG) and their genome coverage was < 70%. In the end, 26 samples were used for analysis (Fig. 4A). Metagenomic recruitment revealed a significant and variable number of flexible genes present per sample (> 10 RPKGs), with an average of 14% of the total flexible gene pool being recruited (Fig. 4A). As expected, the highest values were found in two samples, Tara_018 and Tara_025 (657 [~ 30%] and 506 [~ 23%] of the flexible genes, respectively), both collected in the Mediterranean Sea, where gMED is most

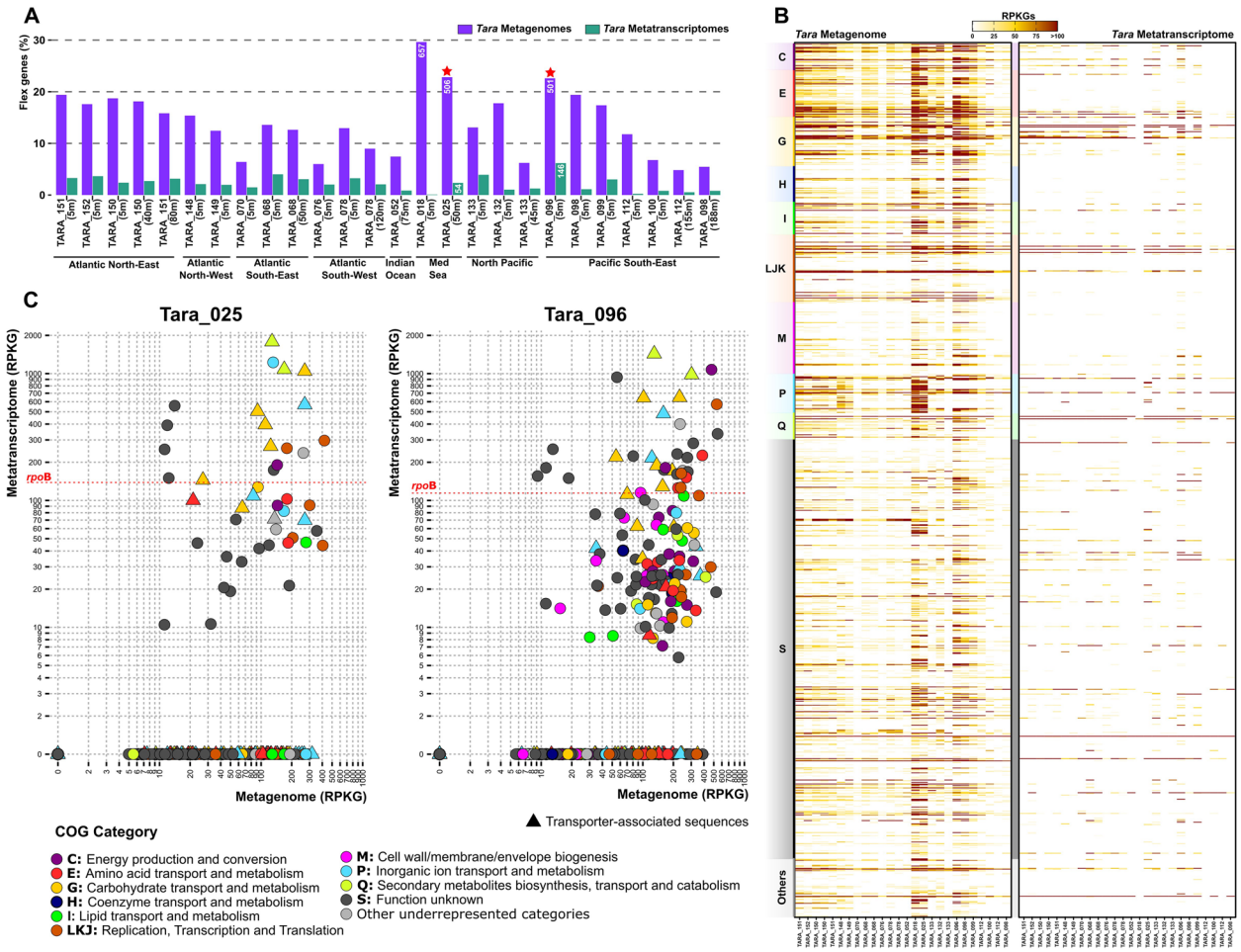

Fig. 4 **A** Percentage of flexome genes present in the metagenome and expressed in the metatranscriptome of different Tara stations. A red star indicates the samples compared in **C**. **B** Recruitment of the flexible gene pool of gMED in the metagenomes and metatranscriptomes of the Tara stations, shown as RPKG values. The categories of the functional annotations are indicated by letters from COG database. **C** Scatter plot showing the relationship between the recruitment in the metagenome and the metatranscriptome of the flexome in Tara_025 (Mediterranean Sea, 50 m depth) and Tara_096 (Pacific South-East, 5 m depth). Expression was compared to the housekeeping gene *rpo*B (red dashed line). COG categories are color-coded, and triangles indicate transporters. RPKG, reads recruited per kilobase of the genome per gigabase of the metagenome

abundant [7]. Tara_096, collected in the South-East Pacific Ocean, was the third metagenome with the greatest number of flexible genes (#501) recruited in a sample. In contrast, metatranscriptomic analysis showed that only a small fraction of the flexible genes (~4% on average) were actively transcribed (Fig. 4A).

Almost half of the dereplicated flexible gene pool (1097 genes) could not be recruited in any of the metagenomic samples (Fig. 4B), indicating that many are absent in these samples. Functional classification according to COG functional categories revealed that the most abundant flexible genes (>50 RPKGs on average) were related to the transport and metabolism of amino acids, carbohydrates, and inorganic ions (categories E, G, and P), as well as cellular processes of replication, transcription, and translation (categories L, K, and J) (Fig. 4B). Finally, an in-depth analysis of the metagenomic and metatranscriptomic recruitment values for two samples, Tara_025 (Mediterranean Sea, 50 m depth) and Tara_096 (Pacific South-East, 5 m depth), selected on the basis of high gMED recruitment values in both types of omics data, showed that, although being present at high values (>50 RPKGs, at least two times more abundant than the core genome) in metagenomic data, most of the flexible genes were less expressed than the housekeeping gene *rpo*B, which encodes the beta subunit of the RNA polymerase (Fig. 4C). Nevertheless, we detected several highly expressed genes (more than *rpo*B) in the above-mentioned metatranscriptomic samples, 18 out of 54 flexible-recruiting genes in the metatranscriptome of Tara_025,

and 33 out of 146 flexible-recruiting genes in Tara_096 metatranscriptome, most of which were classified as transporters (Fig. 4C). Previous studies analysing the metaproteome already revealed that SAR11 cells express abundant transporters to enhance nutrient uptake [46]; thus, it seems logical that these would be the most highly expressed sequences in the flexible gene pool of gMED.

Diversity of transporters in gMED

The high proportion of transmembrane domains and proteins categorised as transporters is not unexpected in an osmotrophic microbe that must compete in an oligotrophic environment such as the ocean. In addition, their high recruitment values in metagenomic and metatranscriptomic studies suggest an important role in the physiology of gMED metabolism. Therefore, we decided to further investigate the diversity of these transporters in the gMED pangenome.

Ammonium transporter

SAR11-gMED-CCG coded, in one of the identified GIs (GI-5; 2299 nucleotides long, Table S2) for a single gene, annotated as an ammonium transporter (Amt-GI5). Amts facilitate the uptake of external ammonium into the cytoplasm, thereby ensuring the fulfilment of the nitrogen requirements of the cell [47, 48]. Thus, it seemed conspicuous that an essential gene was found in a variable region. The search for Amts in the reference genome revealed the presence of two additional sequences in the core genome. Given that Amt-GI5 was only identified in 6 out of 64 gMED SAGs (Table S2) and was detected at scarce values in our long-read metagenomic datasets (Table S2), we improved the numbers by looking for additional Amt protein sequences from the complete SAR11 SAG dataset (see Materials and Methods section). Amino acid alignment and comparison of the sequences, previously clustered at 95% identity (except for those found in gMED), revealed a clear phylogenetic divergence into three major branches (Fig. 5). This distinction was expected, as the proteins exhibited an AAI of ~ 29%. However, while the core Amts had a similar protein size (466 and 433 residues), the Amt-GI5 was significantly larger, with a size of 716 amino acids. Despite the apparent dissimilarity of Amt-GI5, a comparative analysis of its amino acid sequence revealed that the initial 450 residues exhibited a high degree of similarity to those of the other two Amt proteins. These residues were found to adopt a secondary structure comprising eleven transmembrane domains, and the amino acids that are crucial for determining specificity and ammonium binding were also conserved [49] (Fig. S5). Domain annotation using InterPro indicated that Amt-GI5 added a phosphoprotein phosphatase (PPP) (InterPro: IPR001932) at the end of the sequence. Structure prediction with AlphaFold2 [50] of the three Amt paralogs in SAR11-gMED-CCG showed a near-perfect fit over the shared region (template modelling (TM) score of 0.87, root mean square deviation of atomic positions (RMSD) of 1.78), while the PPP region was faced inside the cell, linked with a HAMP-like domain (Fig. 5). The AlphaFold2 predicted local distance difference test (pLDDT) per amino acid of Amt-GI5 showed a very high confidence score in the transmembrane domain (average pLDDT > 90), while the PPP domain was predicted with high confidence (average pLDDT > 70).

The discovery of multiple Amts in the core genome is not a distinctive trait within gMED populations. Indeed, there are several reports reflecting this feature in many organisms [51–53], including marine, such as *Nitrosopumilus* [24, 54, 55], within the SAR86 order of the Gammaproteobacteria [56], and even in SAR11, where a genomic analysis of "*Candidatus* Pelagibacter ubique" strain HTCC1062 revealed up to four distinct paralogs with different patterns of expression under nitrogen-depleted and repleted conditions [57]. The presence of such Amt paralogs has been proposed as a mechanism to cope with different substrate concentrations given a difference in affinity (i.e. high vs low) [47, 56, 58], and the preservation in the core genome in gMED, and even in HTCC1062 (SAR11_0818 and SAR11_1310, [57]) indicate their potential to inhabit such opposed scenarios.

Amt-GI5 (as well as SAR11_0049 and SAR11_0050 in HTCC1062) reflects another story. The characterisation of Amts with fused domains involved in cellular signalling, such as histidine kinases or diguanylate cyclases, among others, have been proposed to act as ammonium sensor proteins, rather than transporters [59–61]. We built a hidden Markov model (HMM) alignment of these sequences, with and without the ammonium transporter domain, and performed two additional searches to (a) determine the distribution of similar Amt-GI5 proteins in a reference dataset of nearly 12,000 marine SAGs from the photic layer of the water column [17] and (b) find similar proteins within gMED genomes with a predicted PPP domain that might have a cognate transducer nearby. Results indicated that the fused structure of Amt-GI5 seems to be exclusive for the order SAR11 (at least in our marine dataset). Furthermore, no other PPP domains were found in their genomes, limiting the understanding of the ammonium sensor Amt-GI5 and its effect in the physiology of the cell. Unfortunately, we cannot confidently assess the biological role of Amt-GI5, but the shared structural properties with the well-defined sensors [59–61], and the location in a GI (i.e. it is not essential for the cell) point towards an ammonium sensor.

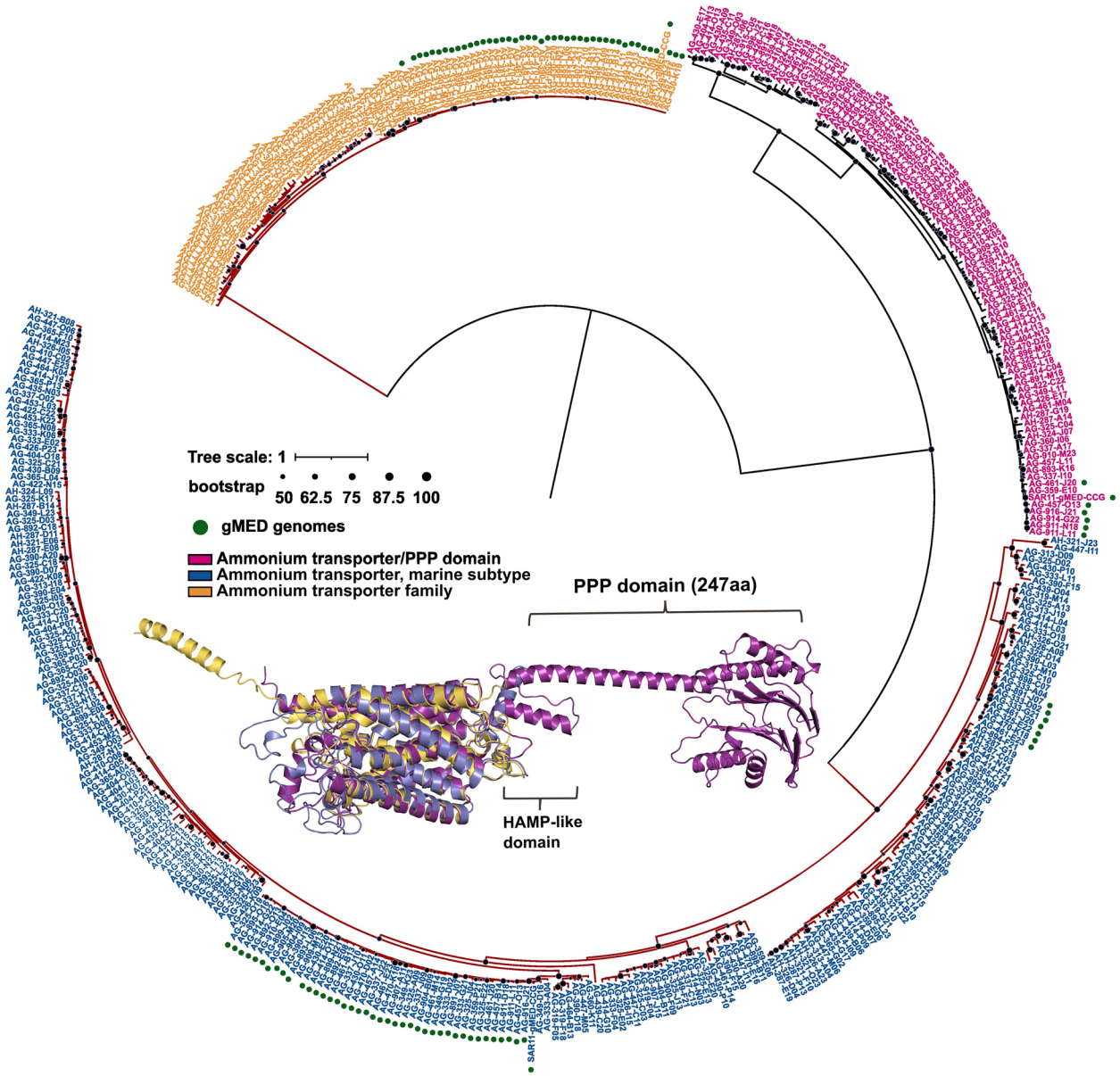

Fig. 5 Maximum likelihood phylogenetic tree of the Amt sequences from SAR11 genomes. The different versions are marked by colours, with the core versions indicated by brown branches. Green dots indicate the Amts from gMED genomes. The structural models defined for the three Amts are shown overlapped and in the same colour as in the tree. PPP domain, phosphoprotein phosphatase domain; HAMP domain, histidine kinases, adenyl cyclases, methyl-accepting proteins, and phosphatases domain; aa, amino acids

TRAP and TTT transporters

The tripartite ATP-independent periplasmic (TRAP) and the tripartite tricarboxylate (TTT) transporters constitute two structurally similar but differentiated (at the level of sequence identity) families of secondary solute transport systems found only in bacteria and archaea [62–65]. Unlike the widely distributed ABC-type transporters, which uses the energy from the hydrolysis of ATP, TRAP and TTT transporters couple the symport of an H^+ or Na^+ (e.g. in marine waters) to the transport of the molecule of interest [66, 67]. Both systems are composed of a soluble periplasmic substrate-binding protein (P and C for the TRAP and TTT systems, respectively) which binds the substrate with high-affinity and specificity, and two transmembrane proteins (subunits Q and M for TRAP, subunits A and B for TTT). Although initially described as C4-dicarboxylate and citrate (tricarboxylate) transporters, their substrate range is much broader,

involving many other sugars and amino acids [62, 68]. Former studies have shown that SAR11 encoded in its streamlined genome for a large proportion of transporters, including TRAP and TTT families, which are heavily expressed in metaproteomic data [46, 69], and thus they make a marked contribution to the assimilation of labile dissolved organic matter [70]. In particular, three paralogs of the DctPQM (TRAP system) were identified in the SAR11-gMED-CCG reference genome, all of which were present in the core genome. A search for this transporter among the SAGs and LRs of gMED yielded 190 and 1,233 additional DctPQMs, respectively. A maximum likelihood phylogenetic tree was constructed to analyse the diversity of this transport system within gMED using a concatenation of DctPQM subunits (Fig. 6A). In addition to the three clusters identified in the reference genome, six further clusters (with an identity of 80%) were identified. In the case of TTT, three different versions of this transporter were identified in the reference genome, one in the core genome and two in the flexible genome.

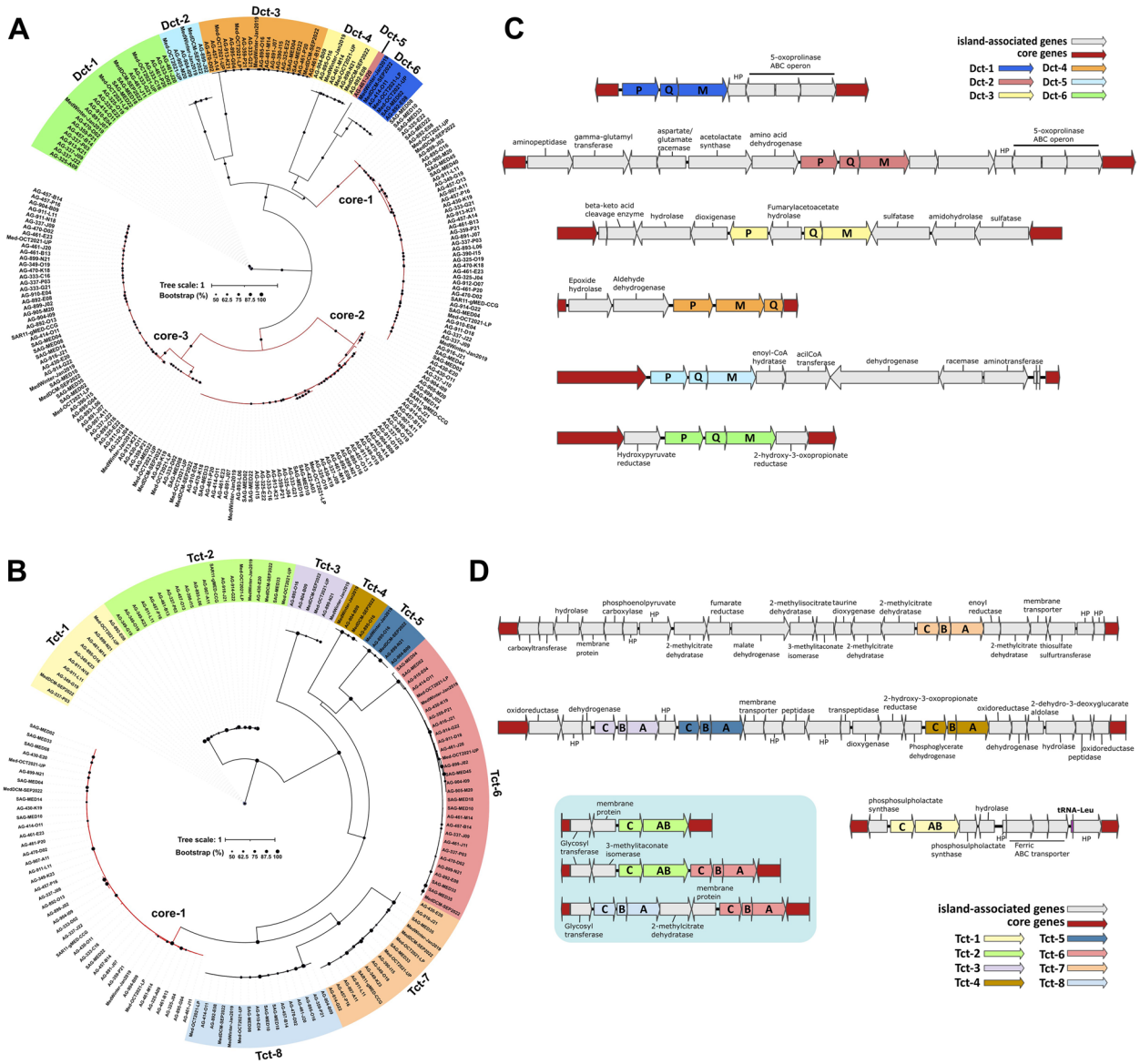

Fig. 6 Maximum likelihood phylogenetic tree of the **A** DctPQM and **B** TctCBA concatenated sequences present in the gMED SAGs and LRs. Different versions of the operons were identified (clustering at 80% of identity), with core versions indicated by brown branches and flexible ones by different colours. For simplicity, one representative sequence from each metagenome is shown in each cluster. Schematic representation of the different versions of the GIs for **C** DctPQM and **D** TctCBA. Three different versions of TctCBA (Tct-2, Tct-6, and Tct-8) are found in the same GI (shaded in light blue). HP, hypothetical protein

A total of 120 and 797 sequences were recovered from the two reference datasets, SAGs and LRs, respectively. The phylogenetic tree constructed by concatenating the TctCBA subunits (see Materials and methods section) revealed nine different versions of the transporter, which clustered with 80% identity (Fig. 6B).

Given the presence of an additional set of genes within the GIs, we attempted to infer the putative substrate of the TRAP and TTT transporters, following the premise that these genes might be involved in its metabolism. This approach has previously been employed for the identification of substrates of TRAP transporters [68]. To gain insight into the function of the neighbouring genes, they were annotated against several databases (see Materials and Methods section). Furthermore, the structure of these genes was predicted with AlphaFold2 [50]. In some cases, the substrate transporter could be partially inferred. For instance, Dct-1 was observed to transport and degrade 5-oxo-L-proline, whereas Dct-3 was inferred to transport and metabolise 4-hydroxyphenylpyruvate, which is subsequently catabolised to fumarate and acetoacetate by the action of the dioxygenase, hydrolase, the beta-keto acid cleavage enzyme, and finally by the fumarylacetoacetate hydrolase (Fig. 6C). However, we did not consider the two sulfatases and the amidohydrolase near the DctPQM operon, but they may potentially play a role in substrate modification. Dct-6, appears to facilitate the transport of two structurally analogous compounds, hydroxypyruvate and 2-hydroxy-3-oxopropanoate. These compounds are situated near the DctPQM operon, where two enzymes act to reduce them to glycerate, which can subsequently be incorporated into glycolysis (Fig. 6C). We also identified several arrangements of the operon. In Dct-3, the gene coding for the DctP subunit was found to be separate from the other two subunits and in the opposite direction. In Dct-4, the gene coding for DctQ was located at the end of the operon, rather than in the middle (Fig. 6C).

The same approach was applied to the TTT transporters (Fig. 6D). For instance, the GI containing the TTT paralog Tct-7 seemed to facilitate the transport of a substrate associated with propionate metabolism via the methylcitrate cycle, potentially methylcitrate. This was evidenced by the detection of multiple methylcitrate dehydratases (prpD) and a methylitanonate isomerase (prpF), in addition to enzymes involved in the tricarboxylic acid cycle (malate dehydrogenase, fumarate reductase) (Fig. 6D). Similarly, Tct-8 also appeared to transport citrate or a methylcitrate derivative, as prpD and prpF genes were also detected. It is noteworthy that within this same genomic region, multiple versions of the GI were identified, each containing a distinct genetic cassette, which in turn gave rise to up to three TTT paralogs (blue inset, Fig. 6D). This indicates that the GI may be capable of utilising a range of related substrates. Tct-1, on the other hand, coded for two phosphosulfolactate synthases at the ends of the Tct paralog. Consequently, the proposed substrate is believed to be phosphosulfolactate, which is transformed into phosphoenolpyruvate, a precursor of pyruvate in the final step of glycolysis.

EamA-like transporter

A total of ten EamA-like transporters were identified in the core genome of SAR11-gMED-CCG, while two were associated with its flexible genome. EamA represents a large family of transporters that are part of the superfamily of membrane drug/metabolite transporters (DMTs). These transporters are widely distributed in both prokaryotes and eukaryotes and harbour a myriad of protein sequences and structures, which has hindered the comprehensive classification of their functions. However, they have been postulated to function as nucleotide-sugar transporters [71]. The average number of EamA-like copies per megabase of genome within gMED SAGs was determined to be 11, with some genomes exhibited up to 17. In view of the considerable number of such transporters present in a streamlined microbe, an exhaustive search for all EamA-like transporters was conducted in both SAGs and metagenome-associated reads. A maximum likelihood phylogenetic tree of the EamA-like revealed a total of twenty-two different clusters (identity > 80%), of which ten corresponded to the core genome of the reference genome (Fig. S6).

In conclusion, while some of these transporter types are well conserved in the core genome of the genomospecies (e.g. gMED), the results demonstrate that most of the observed variability is concentrated within the GIs. This variability not only increases the number of transporter paralogs, but also leads to an increase in the number of accessory genes located in the GIs accompanying these transporters. These genes modify the assimilable substrates, consequently increasing the number of putative substrates that can be assimilated. These findings have significant implications for the physiology of SAR11, as different cells encode distinct transporters and enzymes that facilitate access to a diverse range of dissolved nutrients in the marine environment. This offers an ecological rationale for the notable intraspecies diversity observed in this clade and its potential role in their evolutionary success. Additionally, this structural variation and the maintenance of transporter diversity could be driven by selective pressure, such as the need to evade phages. These processes (nutritional adaptation and phage-driven selection) are not mutually exclusive and likely act interdependently, collectively shaping the genomic and ecological dynamics of SAR11 populations.

Whilst genomic data provide valuable hypotheses regarding the eco-evolutionary dynamics of SAR11 clade, experimental validation is crucial for a comprehensive understanding of these mechanisms. A significant limitation of this study is the difficulty in obtaining pure cultures of these microorganisms using conventional methods, compounded by their extremely slow growth rates, which further hinder experimental investigations. Nevertheless, the data generated here provide a strong foundation for future research. It is therefore essential that targeted efforts are made towards the objective of obtaining experimental evidence to support these findings, in order to unravel the complexities behind the evolutionary success of this clade.

Conclusions

The oligotrophic nature of the ocean surface, characterised by extremely low concentrations of organic and inorganic nutrients, exerts significant environmental pressure on microbial communities. In response, individual cells of SAR11 have undergone adaptations in line with the "streamlining theory", including a reduced surface-to-volume ratio, resulting in very small cell sizes, as well as compact, highly efficient genomes. However, at the population level, our findings suggest a more intricate evolutionary framework, where population dynamics play a central role. The conserved positioning of variable regions within the genome facilitates the exchange of small DNA fragments, typically comprising one or two genes, via homologous recombination. Each variable region was linked to a specific set of genes that, while displaying some divergence, retained equivalent biological functionality across the population. This process not only ensures the preservation of essential genes during selective sweeps but also maintains functional redundancy within GIs, safeguarding a broad environmental gene reservoir. This reservoir provides the population with the capacity for rapid and adaptive responses to environmental fluctuations.

The enrichment of cell surface proteins encoded within GIs, particularly those involved in cell wall biosynthesis and modification, reflects an evolutionary strategy to evade phage recognition through "constant diversity" dynamics [39]. Additionally, the diversification of transporters enhances the efficient utilisation of limited resources within natural populations, while GI-encoded accessory genes expand the range of assimilable organic and inorganic nutrients by modifying substrates. It is probable that these processes occur concurrently and are interdependent, thus elucidating their extensive functional redundancy within the flexible genome. This intricate interplay promotes polyclonality, maintaining a stable balance of genetic variation within the population. By reducing both intraspecific and interspecific competition, these mechanisms collectively enable the population to thrive under variable environmental conditions and selective pressures. Altogether, these findings underscore the interplay between individual and population-level strategies that drive the ecological and evolutionary success of SAR11 in nutrient-limited marine environments.

Methods

Phylogenomic classification of SAR11 order

A total of 2098 available SAR11 genomes with a completeness > 50% and contamination < 5% estimates with CheckM2 [72] were downloaded from NCBI (April 2021) and used to perform a phylogenomic analysis using Phylophlan 3.0 [73]. The resulting tree was analysed using iTOL [74] and following the well-established SAR11 nomenclature described in [7] to classify them. A subset of 68 SAGs of gMED [26–28] genomospecies was used to conduct the following analysis. The pyani [75] package was utilised to calculate the average nucleotide identity (ANI) values among genomes, with the ANIblastall subcommand argument.

Genome reconstruction

A gMED genome cluster with > 99% ANI enabled us to construct a complete composite genome (CCG) by co-assembling contigs, following the methodology previously described [26–28]. This was done using Flye v2.9 [76] and reassembling all contigs from SAGs AG-359-I17, AG-899-K23, AG-911-L13, AG-914-O09, and AG-916-P21, with the "subassemblies" option and an expected genome size of 1.4 Mb. The resulting genome, named SAR11-gMED-CCG (BioSample SAMN40626952), was rearranged with the *dnaA* gene at the beginning. The genome exhibits 99.1% completeness and 0.1% contamination, as estimated by CheckM2. Besides, it was determined that the genome is complete, as the number of transfer RNA (tRNA) genes (#32), ribosomal protein genes (51 vs 52), and tRNA synthetase genes (19 vs 20) are similar to those observed in the strain HTCC7211 [14].

Genomic islands from reference genome

To identify hotspots in the gMED reference genome, a gene alignment between SAR11-gMED-CCG and the rest of gMED SAGs was performed using BLASTn v2.12.0 [77], establishing a threshold of 80% identity between genes, which is near the genomospecies identity boundaries, and a requirement that they have to be present in at least in 40% of gMED genomes to be considered a core gene, taking into account that the average of completeness of the 63 gMED genomes used was 70.5% and the estimated core genome for the whole order was 53% [22].

The genes that remain between the core genes, as defined by the aforementioned thresholds, were considered flexible genes.

Core genome divergence

Core genome analysis was analysed at the protein level due to the high genomic level of heterogeneity within this genomospecies (ANI 80%). The SAR11-gMED-CCG genome was used as a reference and an all-against-all BLASTp v2.12.0 [77] was performed using all proteins from the gMED genomes. The AAI for each core reference protein was calculated, and the threshold to classify a core protein as conserved was set at a minimum of 80% identity. Proteins with AAI below this threshold were considered part of the hypervariable region within the core genome.

Genomic islands recovery from gMED genomes, PacBio CCS15 reads, and Illumina assemblies

The hotspots previously defined in the reference genome were identified in the remaining gMED genomes to maximise the diversity of the gMED flexible genome. This was achieved by blasting the core genes flanking each SAR11-gMED-CCG GI with a minimum of 80% identity and extracting all the flexible genes located between the core genes. LRs metagenomes and SRa were used to increase the diversity of the flexible gMED genome. Three samples were collected from a single off-shore location (200 m bottom depth) in the western Mediterranean (37.35361°N, 0.286194°W), comprising the three layers that characterise the photic zone [78]. There were the upper photic (UP) zone (Med-OCT2021-15 m), the deep chlorophyll maximum (DCM) zone (Med-SEP2022-60 m), and the lower photic (LP) zone (Med-OCT2021-75 m). The samples were collected during the annual stratification period, as well as an additional sample during the winter, when the water column was mixed (MIX; MedWinter-JAN2019) (Table S4). Details regarding sampling procedures and sequencing of these samples have been already published [10, 23, 24]. To identify GIs in the datasets of both technologies (LR and SRa), we first searched for reads specific to gMED. To do this, reads belonging to other SAR11 genomes were excluded using a database of 2098 well-sorted genomes for the search. Sequences from LR and SRa were compared to these genomes using a BLASTn, and we set a threshold that required at least 50% of the sequence proteins to have homology to one of the gMED SAGs with a minimum identity of 80%. Once the subset of gMED reads was obtained, a search for the core genes defined by SAR11-gMED-CCG was performed on the LR and SRa metagenomes with a minimum threshold of 90% identity. Given the small size of the LR, we used this value to guarantee a good hit to the gMED genome. Statistical analyses between the recovery from SAGs and LRs were performed in Python using the SciPy package [79], specifically, the Wilcoxon signed-rank test with the function stats.wilcoxon with default options.

The flexible catalogue was obtained by clustering all proteins recovered from SAGs and LRs, all together using CD-HIT v4.8.1 [80] at 80% identity. Finally, those proteins were functionally annotated against KEGG (KEGG Mapper, Reconstruct Brite, KEGG Orthology, [42]) through the tool BlastKOALA V.2.2 [81] to obtain a general functional classification. Categories, such as transporters, were further refined using Pfam database v35.0 [43]. Moreover, paralogs were defined using CD-HIT iterating from 90 to 30%, in steps of 20% amino acid identity.

Metagenomic and metatranscriptomic comparisons of the flexible gene pool

Metagenomic and metatranscriptomic samples collected from the Tara Oceans expedition [44, 45] were downloaded from the European Nucleotide Archive under project accessions PRJEB1787 and PRJEB6608, respectively. Ilumina raw reads were trimmed with Trimmomatic v0.39 [82], and sequences > 50 bp were kept. Trimmed reads were then used to recruit the abundance (metagenomes) and expression (metatranscriptomes) of the gMED flexible gene pool using BLASTn with an identity threshold of 90% and an alignment size of 90 bp. In addition, genomes from the gMED genomospecies were also recruited in the metagenomes using BLASTn (> 90% identity), and only those samples on which they recruited > 10 reads per kilobase of the genome per gigabase of the metagenome (RPKGs) with a genome coverage of at least 50% were considered. Only samples containing both a metagenome and a metatranscriptome (e. g., TARA_025) were further analysed.

Ammonium transporter (Amt) diversity analysis

Ammonium transporter versions were recovered by using the database the ammonium transporter family (PF00909) and the probable ammonium transporter, marine subtype (TIGR03644.1) for core ones. A custom database of ammonium transporters with phosphatase domain (for the flexible transporter versions) was constructed using sequences found in gMED genomes and metagenomes, as well as a set of sequences from NCBI with 99% of coverage and a range until 80% of identity, aligning them by Muscle v5 [83] and building Hidden Markov models (HMMs) using hmmbuild v3.3 [84]. The screening of putative ones in the bulk of SAR11 SAGs was performed using the tool HMMscan v3.3 [84]. Sequences were dereplicated at 95% identity, except for

gMED sequences, to simplify the construction of a maximum likelihood phylogenetic tree by iqtree v1.6.12 [85] with 5000 ultrafast bootstraps and the LG + G4 model. Structure models of the three Amts were obtained using AlphaFold2 [50] in a local machine using the reduced database and the maximum template date fixed to December 2022.

TRAP and TTT transporter diversity analysis

The retrieval of the Dct and Tct operons from gMED genomes and LRs was performed using a custom database containing DctP, DctQ, and DctM, as well as TctC, TctB, and TctA protein sequences. This was accomplished through the application of HMMs, following the same approach. To analyse the phylogenetic diversity, a maximum likelihood phylogenetic tree was generated using a concatenation of DctPQM subunits and TctCBA subunits, with one representative sequence of each metagenome for simplicity. Clustering at 80% identity and manual curation were used to define the different transporter versions. To infer the putative substrate for the TRAP and TTT families, the genes surrounding the transporter operon were annotated using the InterPro database [86], as well as their structure elucidated with AlphaFold2 [50]. These protein structures were then compared to a database containing both modelled and experimentally characterised structures using FoldSeek [87]. Only models with a template modelling score (TM-score) > 0.85 and a root mean square deviation of atomic positions (RMSD) < 2 were considered bona fide hits. These parameters measure how well the model of the protein of interest matches the one in the database.

EamA diversity analysis

EamA sequences were identified in the gMED genomes and LRs using HMMscan with the Pfam model PF00892. A maximum likelihood phylogenetic tree was constructed following the same method as described above. To simplify the representation, only one sequence per metagenome was included. Subsequently, clustering analysis was performed using CD-HIT at 80% identity, and the resulting clusters were manually curated to define the different transporter versions.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-025-02037-6.

Additional file 1: Fig. S1. Maximum likelihood phylogenomic tree of 2098 SAR11 genomes, classified into the different subclades. The gMED genomes are highlighted in blue, with yellow dots for those that come from the Mediterranean Sea and purple dots for those from BATS. Fig. S2. Schematic representation of the different GIs associated with tRNA duplications. The indicated percentage shows the identity for each duplication. Fig. S3. Boxplot representing the number of flexible genes of each GI found in gMED LRs. The islands that were not recovered in the LRs are marked in red (GI-2 and GI-36). Fig. S4. Donut chart representing the taxonomic orders in which the flexible gene pool is present at different identity ranges in a set of SAGs from the marine environment. The percentage of the flexible catalogue present in these genomes is shown in the grey bar plot. Fig. S5. Alignment of the protein sequence of the different versions of Amt. Fig. S6. Maximum likelihood phylogenetic tree of the EamA sequences present in the gMED SAGs and LRs. Different versions were identified (clustering at 80% of identity, grey shaded), with core versions found in the reference genome indicated by brown branches and flexible ones by different colours. For simplicity, one representative sequence from each metagenome is shown in each cluster. Table S1. Genomic features of gMED genomes. Table S2. Genomic features of GIs identified in SAR11-gMED-CCG. Table S3. Functional classification of sequences with high divergence in the core genome. Table S4. Summary statistics of PacBio metagenomes CCS15 reads.

Acknowledgements

Not applicable.

Authors' contributions

MLP conceived the study. CMP and JHM analyzed the data. MLP, CMP, and JHM contributed to write the manuscript. All authors revised the manuscript and approved the final version.

Funding

This work was supported by grant "FLEX3GEN" (PID2020-118052 GB-I00) to MLP and FRV as well as "MICRO3GEN" (PID2023-150293NB-I00) to MLP, both from the Spanish Ministerio de Economía, Industria y Competitividad (co-financed with FEDER funds). CMP was supported by a PhD fellowship from the Spanish Ministerio de Economía y Competitividad (PRE2021-098122). JMH-M was supported with a PhD fellowship from the Margarita Salas program, cofounded by the Spanish Ministerio de Universidades and the European Union—Next Generation EU (2021/PER/00020).

Data availability

The genome of SAR11-gMED-CCG has been submitted to the NCBI and is accessible via the accession number JBJJMP000000000.1 in the GenBank database, associated with the BioProject PRJNA1092564 and the BioSample number SAMN40626952.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 11 October 2024 Accepted: 10 January 2025
Published online: 04 February 2025

References

1. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al. SAR11 clade dominates ocean surface bacterioplankton communities. Nature. 2003 420:6917. 2002;420:806–10. Available from: https://www.nature.com/articles/nature01240.
2. Malmstrom RR, Cottrell MT, Elifantz H, Kirchman DL. Biomass production and assimilation of dissolved organic matter by SAR11 bacteria in the Northwest Atlantic Ocean. Appl Environ Microbiol. 2005;71:2979–86. Available from: https://journals.asm.org/journal/aem.

3. Malmstrom RR, Straza TRA, Cottrell MT, Kirchman DL. Diversity, abundance, and biomass production of bacterial groups in the western Arctic Ocean. Aquat Microb Ecol. 2007;47:45–55. Available from: https://www.int-res.com/abstracts/ame/v47/n1/p45-55/.
4. Lefort T, Gasol JM. Global-scale distributions of marine surface bacterioplankton groups along gradients of salinity, temperature, and chlorophyll: a meta-analysis of fluorescence in situ hybridization studies. Aquat Microb Ecol. 2013;70:111–30.
5. Giovannoni SJ. SAR11 bacteria: the most abundant plankton in the oceans. Ann Rev Mar Sci. 2017;9:231–55. Available from: https://www.annualreviews.org/content/journals/10.1146/annurev-marine-010814-015934.
6. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014;8:1553–65. Available from: https://www.nature.com/articles/ismej201460.
7. Haro-Moreno JM, Rodriguez-Valera F, Rosselli R, Martinez-Hernandez F, Roda-Garcia JJ, Gomez ML, et al. Ecogenomics of the SAR11 clade. Environ Microbiol. 2020;22:1748–63. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/1462-2920.14896.
8. López-Pérez M, Haro-Moreno JM, Coutinho FH, Martinez-Garcia M, Rodriguez-Valera F. The evolutionary success of the marine bacterium SAR11 analyzed through a metagenomic perspective. mSystems. 2020;5. Available from: https://journals.asm.org/doi/10.1128/msystems.00605-20.
9. Delmont TO, Kiefl E, Kilinc O, Esen OC, Uysal I, Rappé MS, et al. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. Elife. 2019;8. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6721796/.
10. Haro-Moreno JM, López-Pérez M, Rodriguez-Valera F. Enhanced recovery of microbial genes and genomes from a marine water column using long-read metagenomics. Front Microbiol. 2021;12:708782. Available from: www.frontiersin.org.
11. Chang T, Gavelis GS, Brown JM, Stepanauskas R. Genomic representativeness and chimerism in large collections of SAGs and MAGs of marine prokaryoplankton. Microbiome. 2024;12:1–14. Available from: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-024-01848-3.
12. Rappé MS, Connon SA, Vergin KL, Giovannoni SJ. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature. 2002;418:630–3. Available from: https://www.nature.com/articles/nature00917.
13. Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, Vergin KL, et al. Proteorhodopsin in the ubiquitous marine bacterium SAR11. Nature. 2005;438:82–5. Available from: https://www.nature.com/articles/nature04032.
14. Stingl U, Tripp HJ, Giovannoni SJ. Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic Time Series study site. ISME J. 2007;1:361–71. Available from: https://www.nature.com/articles/ismej200749.
15. Tsementzi D, Wu J, Deutsch S, Nath S, Rodriguez-R LM, Burns AS, et al. SAR11 bacteria linked to ocean anoxia and nitrogen loss. Nature. 2016;536:179–83. Available from: https://www.nature.com/articles/nature19068.
16. Jimenez-Infante F, Ngugi DK, Vinu M, Blom J, Alam I, Bajic VB, et al. Genomic characterization of two novel SAR11 isolates from the Red Sea, including the first strain of the SAR11 Ib clade. FEMS Microbiol Ecol. 2017;93:83. https://dx.doi.org/10.1093/femsec/fix083.
17. Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, et al. Charting the complexity of the marine microbiome through single-cell genomics. Cell. 2019;179:1623-1635.e11.
18. Cameron TJ, Temperton B, Swan BK, Landry ZC, Woyke T, Delong EF, et al. Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. ISME J. 2014;8(7):1440–51. Available from: https://www.nature.com/articles/ismej2013243.
19. Thompson LR, Haroon MF, Shibl AA, Cahill MJ, Ngugi DK, Williams GJ, et al. Red Sea SAR11 and Prochlorococcus single-cell genomes reflect globally distributed pangenomes. Appl Environ Microbiol. 2019;85:369–88. Available from: https://journals.asm.org/doi/10.1128/AEM.00369-19.
20. Larkin AA, Hagstrom GI, Brock ML, Garcia NS, Martiny AC. Basin-scale biogeography of Prochlorococcus and SAR11 ecotype replication. ISME J. 2022;17(2):185–94. Available from: https://www.nature.com/articles/s41396-022-01332-6.
21. Henson MW, Lanclos VC, Faircloth BC, Thrash JC. Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. ISME J. 2018;12(7):1846–60. Available from: https://www.nature.com/articles/s41396-018-0092-2.
22. Grote J, Cameron Thrash J, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, et al. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. mBio. 2012;3. Available from: https://journals.asm.org/doi/10.1128/mbio.00252-12.
23. Haro-Moreno JM, López-Pérez M, Molina-Pardines C, Rodriguez-Valera F. Large diversity in the O-chain biosynthetic cluster within populations of Pelagibacterales. bioRxiv. 2024;2024.03.20.585866. Available from: https://www.biorxiv.org/content/10.1101/2024.03.20.585866v2.
24. Suárez-Moo P, Haro-Moreno JM, Rodriguez-Valera F. Microdiversity in marine pelagic ammonia-oxidizing archaeal populations in a Mediterranean long-read metagenome. Environ Microbiol. 2024;26:e16684. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/1462-2920.16684.
25. Berube PM, Biller SJ, Hackl T, Hogle SL, Satinsky BM, Becker JW, et al. Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. Sci Data. 2018;5:1–11. Available from: https://www.nature.com/articles/sdata2018154.
26. Molina-Pardines C, Haro-Moreno JM, López-Pérez M. Phosphate-related genomic islands as drivers of environmental adaptation in the streamlined marine alphaproteobacterial HIMB59. mSystems. 2023;8. Available from: https://journals.asm.org/doi/10.1128/msystems.00898-23.
27. Roda-Garcia JJ, Haro-Moreno JM, Rodriguez-Valera F, Almagro-Moreno S, López-Pérez M. Single-amplified genomes reveal most streamlined free-living marine bacteria. Environ Microbiol. 2023;25:1136–54. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/1462-2920.16348.
28. López-Pérez M, Haro-Moreno JM, Iranzo J, Rodriguez-Valera F. Genomes of the "Candidatus actinomarinales" order: highly streamlined marine epipelagic actinobacteria . mSystems. 2020;5. Available from: https://journals.asm.org/doi/10.1128/msystems.01041-20.
29. Kiefl E, Esen OC, Miller SE, Kroll KL, Willis AD, Rappé MS, et al. Structure-informed microbial population genetics elucidate selective pressures that shape protein evolution. Sci Adv. 2023;9. Available from: https://www.science.org/doi/10.1126/sciadv.abq4632.
30. Juhas M, Van Der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev. 2009;33:376–93. https://doi.org/10.1111/j.1574-6976.2008.00136.x.
31. Hacker J, Kaper JB, editors. Pathogenicity islands and the evolution of pathogenic microbes. 2002;264/1. Available from: http://link.springer.com/10.1007/978-3-662-09217-0.
32. Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ. Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. Biol Direct. 2007;2:1–19. Available from: https://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-2-27.
33. López-Pérez M, Rodriguez-Valera F. Pangenome evolution in the marine bacterium alteromonas. Genome Biol Evol. 2016;8:1556. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC4898812/.
34. Manzano-Morales S, Liu Y, González-Bodí S, Huerta-Cepas J, Iranzo J. Comparison of gene clustering criteria reveals intrinsic uncertainty in pangenome analyses. Genome Biol. 2023;24:1–27. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-03089-3.
35. Iranzo J, Wolf YI, Koonin EV, Sela I. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. Nat Commun. 2019;10:1–10. Available from: https://www.nature.com/articles/s41467-019-13429-2.
36. Pushker R, Mira A, Rodríguez-Valera F. Comparative genomics of gene-family size in closely related bacteria. Genome Biol. 2004;5:R27. Available from: https://europepmc.org/article/pmc/395786.
37. Bratlie MS, Johansen J, Sherman BT, Huang DW, Lempicki RA, Drabløs F. Gene duplications in prokaryotes can be associated with environmental adaptation. BMC Genomics. 2010;11:1–17. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-588.
38. Riedel T, Dittmar T. A method detection limit for the analysis of natural organic matter via Fourier transform ion cyclotron resonance mass

spectrometry. Anal Chem. 2014;86:8376–82. Available from: https://pubs.acs.org/doi/full/10.1021/ac501946m.
39. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–36. Available from: https://www.nature.com/articles/nrmicro2235.
40. Papkou A, Guzella T, Yang W, Koepper S, Pees B, Schalkowski R, et al. The genomic basis of red queen dynamics during rapid reciprocal host–pathogen coevolution. Proc Natl Acad Sci U S A. 2019;116:923–8. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1810402116.
41. Rodriguez-Valera F, Martin-Cuadrado AB, López-Pérez M. Flexible genomic islands as drivers of genome evolution. Curr Opin Microbiol. 2016;31:154–60. Available from: https://pubmed.ncbi.nlm.nih.gov/27085300/.
42. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62. https://dx.doi.org/10.1093/nar/gkv1070.
43. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49:D412–9. https://dx.doi.org/10.1093/nar/gkaa913.
44. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 1979;2015:348. Available from: https://www.science.org/doi/10.1126/science.1261359.
45. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh HJ, Cuenca M, et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell. 2019;179:1068-1083.e21.
46. Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, et al. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. ISME J. 2009;3:93–105. https://dx.doi.org/10.1038/ismej.2008.83.
47. Wacker T, Garcia-Celma JJ, Lewe P, Andrade SLA. Direct observation of electrogenic NH4+ transport in ammonium transport (Amt) proteins. Proc Natl Acad Sci U S A. 2014;111:9995–10000. Available from: https://pubmed.ncbi.nlm.nih.gov/24958855/.
48. Javelle A, Severi E, Thornton J, Merrick M. Ammonium sensing in Escherichia coli. Role of the ammonium transporter AmtB and AmtB-GlnK complex formation. J Biol Chem. 2004;279:8530–8. Available from: https://pubmed.ncbi.nlm.nih.gov/14668330/.
49. Bizior A, Williamson G, Harris T, Hoskisson PA, Javelle A. Prokaryotic ammonium transporters: what has three decades of research revealed? Microbiology (United Kingdom). 2023;169:001360. Available from: https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.001360.
50. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9. Available from: https://www.nature.com/articles/s41586-021-03819-2.
51. Andrade SLA, Dickmanns A, Ficner R, Einsle O. Crystal structure of the archaeal ammonium transporter Amt-1 from Archaeoglobus fulgidus. Proc Natl Acad Sci U S A. 2005;102:14994–9. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.0506254102.
52. Smeulders MJ, Peeters SH, van Alen T, de Bruijckere D, Nuijten GHL, op den Camp HJM, et al. Nutrient limitation causes differential expression of transport- and metabolism genes in the compartmentalized anammox bacterium Kuenenia stuttgartiensis. Front Microbiol. 2020;11:553887. Available from: https://www.frontiersin.org.
53. Williamson G, Harris T, Bizior A, Hoskisson PA, Pritchard L, Javelle A. Biological ammonium transporters: evolution and diversification. FEBS J. 2024;291:3786–810.
54. Nakagawa T, Stahl DA. Transcriptional response of the archaeal ammonia oxidizer Nitrosopumilus maritimus to low and environmentally relevant ammonia concentrations. Appl Environ Microbiol. 2013;79:6911–6. Available from: https://journals.asm.org/doi/10.1128/AEM.02028-13.
55. Santoro AE, Dupont CL, Richter RA, Craig MT, Carini P, McIlvin MR, et al. Genomic and proteomic characterization of "Candidatus Nitrosopelagicus brevis": an ammonia-oxidizing archaeon from the open ocean. Proc Natl Acad Sci U S A. 2015;112:1173–8.
56. Kellom M, Pagliara S, Richards TA, Santoro AE. Exaggerated transmembrane charge of ammonium transporters in nutrient-poor marine environments. Open Biol. 2022;12:220041. Available from: https://royalsocietypublishing.org/doi/10.1098/rsob.220041.
57. Smith DP, Thrash JC, Nicora CD, Lipton MS, Burnum-Johnson KE, Carini P, et al. Proteomic and transcriptomic analyses of "Candidatus Pelagibacter ubique" describe the first PII-independent response to nitrogen limitation in a free-living alphaproteobacterium. mBio. 2013;4. Available from: https://journals.asm.org/doi/10.1128/mbio.00133-12
58. Martens-Habbena W, Berube PM, Urakawa H, De La Torre JR, Stahl DA. Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. Nature. 2009;461:976–9. Available from: https://www.nature.com/articles/nature08465.
59. Tremblay PL, Hallenbeck PC. Of blood, brains and bacteria, the Amt/Rh transporter family: emerging role of Amt as a unique microbial sensor. Mol Microbiol. 2009;71:12–22. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2958.2008.06514.x.
60. Pflüger T, Hernández CF, Lewe P, Frank F, Mertens H, Svergun D, et al. Signaling ammonium across membranes through an ammonium sensor histidine kinase. Nat Commun. 2018;9:1–11. Available from: https://www.nature.com/articles/s41467-017-02637-3.
61. Pflüger T, Gschell M, Zhang L, Shnitsar V, Zabadné AJ, Zierep P, et al. How sensor Amt-like proteins integrate ammonium signals. Sci Adv. 2024;10:9441. Available from: https://www.science.org/doi/10.1126/sciadv.adm9441.
62. Mulligan C, Fischer M, Thomas GH. Tripartite ATP-independent periplasmic (TRAP) transporters in bacteria and archaea. FEMS Microbiol Rev. 2011;35:68–86. Available from: https://pubmed.ncbi.nlm.nih.gov/20584082/.
63. Rosa LT, Bianconi ME, Thomas GH, Kelly DJ. Tripartite ATP-independent periplasmic (TRAP) transporters and tripartite tricarboxylate transporters (TTT): from uptake to pathogenicity. Front Cell Infect Microbiol. 2018;8:324356. Available from: https://www.frontiersin.org.
64. Winnen B, Hvorup RN, Saier MH. The tripartite tricarboxylate transporter (TTT) family. Res Microbiol. 2003;154:457–65. Available from: https://pubmed.ncbi.nlm.nih.gov/14499931/.
65. Rabus R, Jack DL, Kelly DJ, Saier MH. TRAP transporters: an ancient family of extracytoplasmic solute-receptor-dependent secondary active transporters. Microbiology (N Y). 1999;145:3431–45. Available from: https://www.microbiologyresearch.org/content/journal/micro/10.1099/00221287-145-12-3431.
66. Mulligan C, Geertsma ER, Severi E, Kelly DJ, Poolman B, Thomas GH. The substrate-binding protein imposes directionality on an electrochemical sodium gradient-driven TRAP transporter. Proc Natl Acad Sci U S A. 2009;106:1778–83. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.0809979106.
67. Davies JS, Currie MJ, North RA, Scalise M, Wright JD, Copping JM, et al. Structure and mechanism of a tripartite ATP-independent periplasmic TRAP transporter. Nat Commun. 2023;14:1–12. Available from: https://www.nature.com/articles/s41467-023-36590-1.
68. Vetting MW, Al-Obaidi N, Zhao S, San Francisco B, Kim J, Wichelecki DJ, et al. Experimental strategies for functional annotation and metabolism discovery: targeted screening of solute binding proteins and unbiased panning of metabolomes. Biochemistry. 2015;54:909. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC4310620/.
69. Bergauer K, Fernandez-Guerra A, Garcia JAL, Sprenger RR, Stepanauskas R, Pachiadaki MG, et al. Organic matter processing by microbial communities throughout the Atlantic water column as revealed by metaproteomics. Proc Natl Acad Sci U S A. 2018;115:E400–8. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.1708779115.
70. Clifton BE, Alcolombri U, Jackson CJ, Laurino P. Ultrahigh-affinity transport proteins from ubiquitous marine bacteria reveal mechanisms and global patterns of nutrient uptake. bioRxiv. 2023:2023.02.16.528805. Available from: https://www.biorxiv.org/content/10.1101/2023.02.16.528805v1.
71. Västermark Å, Almén MS, Simmen MW, Fredriksson R, Schiöth HB. Functional specialization in nucleotide sugar transporters occurred through differentiation of the gene cluster EamA (DUF6) before the radiation of Viridiplantae. BMC Evol Biol. 2011;11:1–18. Available from: https://bmcecolevol.biomedcentral.com/articles/10.1186/1471-2148-11-123.
72. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods. 2023;20:1203–12. Available from: https://www.nature.com/articles/s41592-023-01940-w.

73. Segata N, Börnigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun. 2013;4:1–11. Available from: https://www.nature.com/articles/ncomms3304.
74. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5. https://dx.doi.org/10.1093/nar/gkw290.
75. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal Methods. 2015;8:12–24. Available from: https://pubs.rsc.org/en/content/articlehtml/2016/ay/c5ay02550h.
76. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods. 2020;17:1103–10. Available from: https://www.nature.com/articles/s41592-020-00971-x.
77. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402. https://dx.doi.org/10.1093/nar/25.17.3389.
78. Haro-Moreno JM, López-Pérez M, de la Torre JR, et al. Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. Microbiome. 2018;6:128. https://doi.org/10.1186/s40168-018-0513-5.
79. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72. Available from: https://www.nature.com/articles/s41592-019-0686-2.
80. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26:680–2. https://dx.doi.org/10.1093/bioinformatics/btq003.
81. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol. 2016;428:726–31.
82. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. https://dx.doi.org/10.1093/bioinformatics/btu170.
83. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7. https://dx.doi.org/10.1093/nar/gkh340.
84. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39:W29-37. https://dx.doi.org/10.1093/nar/gkr367.
85. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74. https://dx.doi.org/10.1093/molbev/msu300.
86. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res. 2023;51:D418–27. https://dx.doi.org/10.1093/nar/gkac993.
87. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2023;42:243–6. Available from: https://www.nature.com/articles/s41587-023-01773-0.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.