Seasonal microbial dynamics in the ocean inferred from assembled and unassembled data: a view on the unknown biosphere

Didier Debroas, Corentin Hochart, Pierre Galand

Seasonal microbial dynamics in the ocean inferred from assembled and unassembled data: a view on the unknown biosphere

Didier Debroas[1], Corentin Hochart[2] and Pierre E. Galand[2]

1- Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Genome et Environnement, 63000 Clermont-Ferrand, France

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

Corresponding author: Prof. Didier Debroas didier.debroas@uca.fr

Summary

In environmental metagenomic experiments, a very high proportion of the microbial sequencing data (> 70%) remains largely unexploited because rare and closely related genomes are missed in short-read assemblies. The identity and the potential metabolisms of a large fraction of natural microbial communities thus remain inaccessible to researchers. The purpose of this study was to explore the genomic content of unassembled metagenomic data and test their level of novelty. We used data from a three-year microbial metagenomic time series of the NW Mediterranean Sea, and conducted reference-free and database-guided analysis. The results revealed a significant genomic difference between the assembled and unassembled reads. The unassembled reads had a lower mean identity against public databases, and fewer metabolic pathways could be reconstructed. In addition, the unassembled fraction presented a clear temporal pattern, unlike the assembled ones, and a specific community composition that was similar to the rare communities defined by metabarcoding using the 16S rRNA gene. The rare gene pool was characterised by keystone bacterial taxa, and the presence of viruses, suggesting that viral lysis could maintain some taxa in a state of rarity. Our study demonstrates that unassembled metagenomic data can provide important information on the structure and functioning of microbial communities.

Introduction

Metagenomics studies are based on gene centric approaches often based on assembly followed by contigs binning for building metagenome-assembled genomes (MAGs). However, a relatively low proportion of the reads can be assembled into contigs or/and MAGs. Often the higher proportion of the sequencing data (> 70%) remains largely unexploited in metagenomes because rare and closely related genomes are missed in short-read data assemblies [1]. Indeed, a minimum sequencing depth is often needed for contig assembly. Bacterial species with coverage below 15x in metagenomes typically result in low-quality assemblies [2]. For Luo et al. [3], a species can only be accurately assembled from a complex metagenome when it shows at least 20x coverage. Since rare species within a community typically have low sequencing coverage, they are hardly assembled into long contigs. To reconstruct rare strains from complex assemblages thus requires sometimes an enormous dataset with a very high coverage depth exceeding sometime 1000x [4]. The approach described by Nielsen [5] allows, however, the reconstruction of any species with an adequate sequencing depth (~ 50x according to the simulation) and permits the binning of some rare members with the rarest having 0.02% relative abundance. However, a minimum sequencing depth is often needed, but not always sufficient for accurate contig assembly. Globally, assemblers perform poorly in the presence of multiple similar genomes from closely related species. In that case, unassembled reads can also belong to the flexible or accessory genome of the main components of the community. For instance, members of the wide spread marine Prochlorococcus genus have a huge pangenome, with ~ 1000 common genes (core genomes), and a 'flexible' genome, which is found in only one or a few of the Prochlorococcus genomes [6]. However, by comparing long and short reads, Sharon et al [1] concluded that the majority of unassembled reads in the short-read data were left unassembled because of low coverage and not because of the presence of multiple similar regions.

The rare components of the metagenomics data, bacterial taxa (i.e. rare biosphere) or individual genes (i.e. flexible genome), which may be hard to assemble, could nevertheless play an important role in ecosystem functioning. Regarding genes for instance, genomic and metagenomic data have defined at least 12 major clades among Prochlorococcus and the flexible gene distribution within these clades determines adaptation to the local environment (light, temperature…) [6]. These flexible genes pool, which are not abundant, are still important because they are often associated with specific nutritional requirements (phosphorus, nitrogen or iron, [6]). At the taxa level, rare populations of microorganisms, with their tremendous diversity [7], can also play an important role in ecosystem functioning. The "rare microbial biosphere" [8] was first seen mainly as a seed bank in which some members became dominant at times depending on specific environmental [9]. Some bacteria, for instance, become dominant under anthropogenic pressure [10] or when colonizing a new substrate [11]. Other changes in abundance can occur following climatic fluctuations [12]. These observations illustrate a transient state of rare microorganisms toward the abundant biosphere, or an oscillation within a rare state [13]. Inversely, some rare taxa always remain rare [13]. The fact that some of them exhibit high cell-level metabolic activity [14] could indicate that they are keystone species in ecosystems. Keystone taxa are defined by Banerjee *et al*. [15] as highly connected taxa that exert a considerable influence on microbiome structure and function, irrespective of their abundance across space and time. Thus, some low-abundance taxa that are highly connected in microbial communities can explain compositional turnover better than all the taxa combined [16]. However, the functional role of rare microorganisms remains poorly understood, since they are often phylogenetically distant from referenced cultured or uncultured microbes [14, 17, 18]. Therefore, the microbial rare biosphere may constitute an important genomic reservoir or diversity pool, and a source of genetic novelty with biotechnological potential [19, 20]. Thus, the rare taxa are certainly an important component of the "dark matter" [21], but the metabolic potential of the rare biosphere remains under-explored. A limited number of studies have focused on the genetic content of this biosphere [22, 23].

In this work, we focused on the rare genetic material defined here as the sequencing reads that do not align with assembled contigs. We hypothesize that this genetic material plays an important role in the marine ecosystem functioning. For this purpose, we analyzed a three-year metagenomic time series based on monthly samples from the Bay of Banyuls sur Mer (NW Mediterranean Sea).

Materials and methods

Sampling and sequencing

The sampling strategy was described in Galand *et al.* [24]. Briefly, surface seawater (3 m) was collected monthly from January 2012 to February 2015 (40 samples) by using a 10-L Niskin bottle at the SOLA station (42°31′N, 03°11′E) in the Bay of Banyuls sur Mer (France) in the northwestern Mediterranean. A volume of 5 L was prefiltered through 3-µm pore-size polycarbonate filters (Millipore, Billerica, MA, USA), and the microbial biomass was collected on 0.22-µm pore-size GV Sterivex cartridges (Millipore) and stored at −80 °C until nucleic acid extraction. The physicochemical parameters (Table S1) were provided by the "Service d'Observation en Milieu Littoral" (SOMLIT). After DNA extraction [24] samples were sequenced on eight lanes of a HiSeq 2500 "High-Output" paired-end run (2 × 100 bp). Raw reads were archived in the ENA repository under accession number PRJEB26919.

Assembling

Raw paired-end Illumina reads were preprocessed by removing Nextera adapters with the bbduck program from the BBTools package (12.10.2015 release) (http://jgi.doe.gov/data-and-tools/bbtools/). Reads were then trimmed and filtered using Trimmomatic v. 0.33 [25] based on their quality generating a read length of ca. 85 bp. A total of 34 to 112 million reads per sample remained after filtering (Table S2). For each metagenome, high-quality reads were assembled into contigs with IDBA-UD [26] with the default iterative k-mer assembly and k-mer length increasing from 20 to 100 in steps of 20, the correction option, and with both pair-end reads (-r entry) and single-end reads (--long entry). Two kinds of reads were discriminated by mapping all the reads against the built contigs (Fig. 1). The mapping was conducted with bwa mem algorithm [27] with default parameters, the results by sample are displayed in the Table S2. Thereafter, we term the two fractions as unassembled, as the pool of reads that do not match with contigs formed post-assembly,

and assembled reads. However, algorithms implemented in mappers are different from assemblers and in some cases it can exist some discrepancies between these tools

Community composition, functional abundance table *and OTU abundance table inferred from assembled and unassembled reads*

The composition of the unassembled and assembled read fractions were compared to each other with MetaFast [28], which allows a direct reference-free comparison of shotgun metagenomic data. The Bray-Curtis dissimilarity matrix computed by MetaFast was used to construct a non-metric multidimensional scaling (NMDS) ordination with the vegan package in R [29].

An OTU abundance table based on 16S rRNA gene was built for assembled and non assembled reads separately. The 16S rRNA gene were identified by comparing all preprocessed reads to the SILVA database [30] with BLASTn (identity ≥ 90% and length > 80 bp). An abundance table was built by clustering reads at a 97% similarity against the SILVA sequence collection. In addition, a phylogenetic analysis was conducted based on unique clade-specific marker genes for assembled and unassembled reads with metaphlan2 [31], and the list of taxa and their relative abundance was used with LefSe [32] to identify the taxa that best explained the differences between the fractions. A functional abundance table was built with a reference-guided approach based on the UNIREF (90 and 100) [33] and KEGG databases [34]. Reads were compared against the databases using DIAMOND [35] with the blastx mode and the following parameters: -evalue 1e-5 --sensitive --max-target-seqs 1. Each function in these tables contains reads originating from multiple genomes. The generated abundance tables were characterized by zero-inflation. We removed all genes present as singletons only in the 80 samples (40 assembled and 40 unassembled), or detected in less than 20 samples. Gene loss are presented in the Table S3. Overall, we counted 846 16S rRNA OTUs, 6984 KOs, and 1,210,645 proteins (UNIREF90) in the entire dataset after applying strict filters described in the experimental procedures section (Table S3). The statistical analysis were conducted with the ALDEx2 methods [36] that take into account the compositional nature of the data [37]. Differences

in abundance between the two categories of genes (derived from assembled and unassembled reads) were considered as significant ($P<0.05$) when the Welch and Wilcoxon tests were convergent. The significant results annotated against the KEGG database were used to discriminate metabolic pathways between assembled and unassembled fractions with the "gage" and "pathview" functions implemented in R [38, 39].

Multivariate analyses were conducted with the R MixOmics package [40] by using the "spca" function with centered log ratio transformation (CLR) after replacing zeros with the "cmultRepl" function and the "czm" option included in the zCompositions library [41].

Binning covarying gene groups with assembled and unassembled reads

The most common approach to reconstruct genomes from metagenomes is to build MAGs. MAG construction is based on mapping reads to contigs, but since we cannot obtain contigs from the rare reads, we chose an alternative approach to survey the potential genomic content of the communities. Co-Abundance gene Groups (CAGs) were built separately for the assembled and non assembled datasets, from the table gathering the functional abundance (UNIREF90) and OTU (SILVA) tables, with 3 different approaches: MSPminer [42], canopy [5] and Partial Least Squares regression (PLS) based networks. MSPminer and canopy bin covarying genes by a robust measure of proportionality or correlation between genes, and give a same weight to the proteins and rRNA genes. In our approach, unlike in the original methods cited, we used the abundance of functions rather than a gene catalog. In addition, we introduce a new method to bin genes from abundance tables by associating a Partial Least Squares regression (PLS) and a bipartite network. PLS relates the OTUs (16S rRNA) and the protein tables. The goal was to predict the protein variations from the OTUs dynamics. The regression was computed with the "spls" function associated to the regression method in the MixOmics package in R [40]. In a second step, a bipartite network based on PLS was built linking OTUs and protein genes. The edges with a weight lower than 0.8 and

orphan vertices were deleted by using the igraph package [43]. A CAG was then defined by grouping all the protein genes associated to one OTU.

The quality (completeness and contamination) of the CAGs built by these 3 different approaches were checked with checkM [44] with the option "--genes". In a first step, 149 CAGs were defined and the taxonomy, completeness and contamination was assessed by checkM (Table S4). The temporal dynamics of these different CAGs were assessed from the median of the gene counts at each sampling date, and a network was built based on Spearman correlations. CAGs were considered redundant if their weight (i.e. correlation) in the network was higher than 0.95 to a CAG with the same taxonomy and amino acids identity >95%. This identity was computed with compareM (https://github.com/dparks1134/CompareM). These criteria were based on the histogram of the edge weight (i.e. correlations), manual inspection of the network cluster for the CAG taxonomy and the amino acid identity. The final network, with a correlation coefficient > 0.8 or < -0.8 between edges, included 114 CAGs as well as 3 physicochemical parameters of the water samples. The centrality `indices were computed with the package qgraph [45].

Amplicon sequencing

Amplicon sequencing data were originally published in Lambert et al. [46]. Briefly, specific primer pairs 27F (5'-AGRGTTYGATYMTGGCTCAG) and 519R (5'-GTNTTACNGCGGCKGCTG) were used to target the V1-V3 regions of the bacterial 16S rRNA gene and sequencing was carried out with Illumina MiSeq 2 x 300 bp kits. The analysis of the raw reads was done by constructing amplicon sequence variants (ASVs) following the standard pipeline of the DADA2 package [47]. Abundant ASVs were defined as the ones with a representation > 0.01% within a sample, and rare ASVs as having an abundance < 0.01% within a sample [48]

Results

Temporal dynamics of the assembled and unassembled reads

The reads from the three-year metagenomic time series were classified according to their mapping or not to contigs larger than 1 kb (i.e. assembled and unassembled) (Fig. 1). , A direct comparison of the read composition between time points showed that for the unassembled reads the similarity between samples was highest when samples were taken one year apart (Fig. 2), and similarity was lowest when samples were taken six months apart (Fig. 2a). For the assembled reads, the seasonal pattern of similarity was noisy and the overall pattern was not as clear (Fig. 2b).

The non-metric multidimensional scaling (NMDS) computed from Bray-Curtis index obtained with MetaFast showed that the read composition of the unassembled fraction was different from the read composition of the assembled fraction (Fig. S1). We then identified the reads that were significantly enriched in each fraction (Table 1). From the statistical analysis (ALDEx2 methods) we deduced that a total of 130,450 proteins (10.7% of the total) were significantly enriched in the unassembled fraction and 125,953 (10.4%) in the assembled fraction. Furthermore, 26 16S rRNA (mean reads: 170.5) and 25 KEGG (mean reads: 69.8) annotated genes were only present in the unassembled fraction. Conversely, 2523 UNIREF genes (mean: 209.2) were present only in the assembled fraction (Table 1).

Taxonomic composition

To study the taxonomic composition of the two fractions, we used statistical analysis based on both unique clade-specific marker genes (Fig. 3) and rRNA genes (Fig. S2) found in the reads. In addition, we analyzed the results obtained from high-throughput sequencing of the 16S rRNA gene (Fig. S3). From the shotgun data, both analyses showed that the taxonomic composition of the unassembled fraction was different from that of the assembled fraction. The use of phylogenetic marker genes highlighted differences in prokaryotic and viral compositions (Fig. 3). The analysis showed that the assembled fraction had one characteristic phylum, *Proteobacteria*. At the class

level, *Rhizobiales* and *Betaproteobacteria* with *Burkholderiales* dominated this fraction. The unassembled community had a larger number of signature taxa, including *Verrucomicrobia*, *Actinobacteria*, *Bacteroidetes*, and *Thaumarchaeota*, within *Archaea*. Among this fraction *Proteobacteria*, *Gammaproteobacteria* dominated. Interestingly, this fraction was also characterized by viruses. Since, in this study, the microbial biomass was gathered on 0.2 µm pore-sized filters, viruses were possibly present as prophages or particles in the lytic phase. The ASVs from the amplicon sequencing were separated in two fractions based on an abundance threshold of 0.01% (Fig. S3). The abundant ASVs were dominated by the SAR11 clade whereas the rare ASVs were also more diverse as observed for unassembled metagenomic read fraction. In the rare ASV fraction, the *Gammaproteobacteria, Bacteroidetes Verrucomicrobia* and *Actinobacteria* were more common than in the abundant fraction. Finally, the two fractions based on the assembled/unassembled reads and the reference method for deciphering the rare biosphere based on a threshold (i.e. 0.01%) gave similar results (Fig. 3 and Fig. S3). We can hypothesized that the unassembled reads capture the majority of the rarest fraction of microorganisms.

Identifying metabolic capabilities among the assembled and unassembled fraction

The alignment data showed that for all sampling dates there was a higher proportion of reads that aligned to the UNIREF90 references in the assembled (44.1%) than unassembled fraction (38.5%) (Fig. S4). The overall percentage of aligned reads for both assembled and unassembled reads was low. In addition, a higher proportion of the assembled read alignments had high identity values than those of the unassembled reads (Fig. 4). When comparing both alignment scores and identities for UNIREF90 and UNIREF100, the differences between unassembled and assembled reads were highly significant (ANOVA two ways: assembled/unassembled × sampling dates; Fig. S5). The main factor explaining the variations in identity or scores was “mappability” against contigs and not sampling date.

The sparse principal component analysis (sPCA) based on UNIREF90 and KEGG annotated genes separated the assembled and unassembled fractions (Fig. 5). The multivariate analysis explained 31% (UNIREF90 clusters) and 36% (KEGG clusters) of the variance along axes 1 and 2. By comparing pathways (KO) present in the assembled *vs.* unassembled fractions, we identified two pathways involved in photosynthesis and flagellar assembly, which were enriched in the assembled communities (Fig. S6). The unassembled fraction was not significantly enriched in any of the pathways referenced in the KEGG database. This result is congruent with the previous statistical analysis showing few KOs enriched in this fraction (Table 1).

Covarying gene groups of the assembled and unassembled communities

In total, 114 non-redundant CAGs were identified. The mean completeness was 53.19% (33.47–89.71) for the 56 uCAGs and 47.27% (30.25–80.07) for the 58 aCAGs. The mean contaminations were 4.44% and 4.06% for the uCAGs and aCAGs species, respectively. The uCAGs consisted of 65,787 genes and 59,470 genes for the aCAGs. The UNIREF proteins were linked to KEGG features to identify 3,072 KOs in 78 CAGs. A total of 765 KOs specifically belonged to the uCAGs (37) and 2287 to the aCAGs (41).

Of the 125,257 genes (UNIRE90 + 16S rRNA genes) found to be enriched in the unassembled fraction (Table 1), 16,878 were found in the uCAGs (13.4%). This proportion reached 14.7% for genes enriched in assembled fraction. Three CAGs contained 16S rRNA genes that were found to be significantly enriched in the unassembled fraction (*Gammaproteobacteria*, *Flavobacteriia*, *and Betaproteobacteria*), and one CAG included a 16S rRNA gene present exclusively in the unassembled reads during all sampling dates. This CAG belonged to *Alphaproteobacteria* (*Nisaea* genus).

Key constituents in marine ecosystems deciphered by a network approach

The network built with 49 uCAGs and 46 aCAGs was binned in 18 clusters (Louvain method), of which five had more than three vertices (CAGs or physico-chemical parameters). All of these large clusters included two kinds of CAGs and three were associated with physico-chemical parameters: temperature, oxygen, and nitrite concentration (Fig. 6 and Fig. S7). We identified the main metabolic pathways associated with each cluster by considering the pathways represented by at least 25% of the KEGG orthologs included in the pathway of interest. The major common pathways corresponded mainly to metabolisms involved in amino acid biosynthesis, but photosynthesis pathways also characterized one of these clusters (Fig. S7 - Cluster 17).

When analyzing the temporal dynamics of the CAGs, the spring and summer seasons determined their dynamics (Fig. S7). The network parameters allow us to decipher the main "influencers" or keystone species (Fig. 6), and temperature appears to be the main key parameter. Among the keystone species, uCAGs and aCAGs were present and mainly classified in the *Proteobacteria* phylum (*Alpha* and *Gammaproteobacteria*). Interestingly, *Archaea* classified as *Euryarchaeota* appeared in this top ranking.

Discussion

In this paper we present an overview of the rare genomic content of marine microbial communities based on the reads "mappability" against contigs, and defined for the first time at the taxa or gene level. The congruence between the detection clade-specific marker genes in the assembled and unassembled reads (Fig 3) and metabarcoding results (Fig. S3), separating abundant and rare microbes, indicates that the most part of the unassembled reads belonged to rare marine species. The unassembled reads could also have originated from strain heterogeneity manifested as single nucleotide variations and small insertions or deletions [4]. However, the assembler used in this paper takes into account the coverage ratios between adjacent edges in the assembly graph (*de Bruijn Graph)* to replace it with high-covered alternatives, and acts therefore as a consensus assembly reducing information about individual strains. As only the most abundant microbes are assembled by common bioinformatics tools [49, 50], and because the kind of assembler used performs poorly with strain heterogeneity, the unassembled reads that we focused on most certainly represent members of the rare biosphere.

Community composition of the assembled and unassembled fractions

The comparison of the taxonomy inferred from metabarcoding in the abundant and rare fraction (<0.01%) with those deduced from phylogenetic markers included in assembled and unassembled reads, revealed similar patterns between the two approaches. The unassembled fraction, and the rare 16S rRNA amplicons, were both characterized by a higher community diversity and by a higher abundance of *Gammaproteobacteria*, *Verrucomicrobia*, *Actinobacteria* and *Bacteroidetes*. The similarity between the two data sets is noteworthy since the approaches have different potential biases. Metabarcoding is hampered by well-known PCR bias and the cut-off definition of the rare biosphere is always arbitrary (0.01% here). To date, 16S or 18S rRNA based studies describing the rare biosphere have used a cut off, often ranging between < 1% [51] and < 0.01% [48], which originates from the rank-curve distribution of microbial communities that shows a long 'tail' of low

abundance taxa [13]. In our metagenomic approach, the delineation between rare and abundant pool genes does not depend on an arbitrary cut off, but on sequencing depth and contig length. However, the delineation between rare and abundant may still depend on the sequencing effort. Our approach differs from an earlier metagenomics study that defined rare members as sequence assemblies being in the "tail" of the contig rank abundance curve, or ~0.005% in relative abundance [23]. The two methods that we used, metabarcoding and metagenomic based, allowed to detect the prokaryotes characterizing the abundant fraction, the *Alphaproteobacteria* phylum (SAR11 clade), which dominates marine bacteria [52]. Its ecological importance at our study site was underlined by the network analysis where it appeared among the main keystone. Interestingly, the rare gene pool (unassembled data) was characterized by viruses. These viral genes detected mainly in the rare fraction corresponded likely to the replication of the DNA phage before the cell lysis. The rare community can therefore include some taxa under a strong selection pressure through viral lysis. Earlier experimental work suggested that some rare taxa may indeed have high susceptibility to viral attack [53]. This idea is, however, counter intuitive within the frame of the "kill the winner" hypothesis [54], which suggests that rare microorganisms, because they are not abundant, have a lower probability of encountering virus [55]. The link between predation and rare taxa is then rather seen as an evolutionary advantage for escaping top-down regulation [13]. Our data adds arguments for another hypothesis which suggests that lysis or predation are maintaining some particular taxa in a state of rarity.

Seasonal dynamics and keystone species

Our study showed that the unassembled reads of metagenomes responded strongly to seasonal variations and corresponded certainly to an adaptation of the communities to specific environment conditions (light, temperature, nutrients etc…). This unassembled gene pool, which could correspond mostly to rare taxa as discussed above, displayed a reproducible pattern of temporal dynamics that was stronger than that of the assembled fraction, which in turn could represent the

abundant microorganisms. The rare fraction thus showed a strong seasonal pattern for both similar and dissimilar communities (Fig. 2). Conversely, the rhythm of the abundant fraction (i.e. assembled reads) was noisier, with no patterns for communities sampled during opposite seasons. The abundant gene pool could thus correspond to core marine taxa with few temporal variations or to housekeeping genes. Thus, the overall seasonality of the microbial communities in response to the environment was mainly driven by the rare gene pool. A similar observation was made from coastal sands, where turnover in community composition was no longer observed when 50% of the rare species were removed from the dataset [56], and the Arctic Ocean where the rare biosphere was sensitive to environmental heterogeneity [57]. Rare communities can be classified according to different patterns of seasonal abundance and activity [17]. Within this classification, there is a group defined as rare taxa that never bloom but are active. It has been shown in bacteria, Archaea, and Eukaryotes [14, 17, 51]. These rare but active taxa also have a temporal pattern linked to biotic or abiotic parameters. Even though our metagenomics approach does not allow to infer activity, the reproducible seasonal dynamics of the continually rare community that we observed could suggest that they are active.

Overall, the binning step allowed the reconstruction of the main bacterial and archaeal phyla detected by the metaphlan pipeline (Fig. 3), with the exception of *Thaumarchaeota* (Table S2), and the network provided a good overview of the microbial interactions along the seasonal dynamics. Among the top "influencers" within this network were temperature, abundant microorganisms, and six rare taxa belonging to *Gammaproteobacteria*, *Flavobacteriia*, *Dehalocccoidetes*, and *Euryarchaeota*. The temperature had a significant influence on the microbial components of this network. Such result is not surprising, but it can be viewed as a validation of our approach. This influence is also noticeable at the read scale, since temporal variation was strongly associated with seasonality (Fig. 2). The link between heterotrophic bacterial metabolism and temperature is generally associated with nutrient availability, such as organic matter released from phytoplankton or grazing [58]. *Alphaproteobacteria* (*Rhodobacterales*) appeared twice in the top influencers, but

were also challenged by other taxa, such as *Gammaproteobacteria* and *Bacteroidetes*. Arandia-Gorostidi *et al*. [59] showed that the growth of these taxa was strongly related to temperature changes, whereas *Alphaproteobacteria*, such as SAR11, showed the lowest temperature sensitivity [60]. The *Gammaproteobacteria* class, and more specifically the *Alteromonodales*, dominated the main influencers in this network. After *Alphaproteobacteria*, this class was the most abundant in ICoMM data [60] and *Alteromonodales*, such as *Oceanospirillales* or *Vibrionales,* contains mainly marine species. Therefore, *Alteromonas* could contribute significantly to the flux of dissolved organic carbon and nutrient mineralisation in the upper ocean [61]. Furthermore, *Euryarchaeota* was also found to have a key role. The CAG built in this study does not allow for a precise taxonomy; however, a previous study on the same site highlighted the presence of the MGII clade [17, 63] now defined as an order lineage. The ecological success of the MGII group could be due to the presence of light-harvesting proteins (i.e. proteorhodopsin) [63–65]. Recently, the partially reconstructed MGIIa genome revealed the presence of glycoside hydrolases that are possibly involved in algal substrate breakdown [66, 67].

Rare and abundant gene pools: many unknown functions

This study showed that there was significantly more unknown genes in the rare fraction than in the abundant fraction (Fig 4 and Fig S4). The microbial rare biosphere could thus be seen as a large pool of genes possessing known and unknown functions and considered a reservoir of "genetic novelty" [20, 68]. Since the rare gene pool showed strong temporal dynamics, it indicates that this reservoir of rare functions plays a role in ecosystem functioning. Some of the rare reads could nevertheless be mapped against database references (UNIREF or KEGG). They corresponded to known potential functions, but the identity of these rare genes was significantly lower than that of the abundant ones. This suggests that the rare gene pool harbors different variants of known genes found in abundant microbes. It should be noted that no metabolic pathways could be built from the identified rare KOs. The sequencing depth may have been too shallow to detect all the steps of the

pathways present in the rare microbes, or some of the steps may be conducted by proteins coded by unknown genes.

For the abundant microorganisms, the fraction of the mapped reads against the UNIREF databases (90 or 100) always represented a low proportion of the total clean reads (< 45%). This result at the short-read scale is in agreement with previous studies showing that 40%–60% of the coding genes cannot be assigned to a known function in the marine environment [69, 70]. Even in the human gut microbiome, which has been extensively studied, approximately 40% of the genes have unknown functions, although the "mappability "of the metagenomes against microbial genomes reaches ~ 80% [71]. The unmapped reads can correspond to new functions harbored by known lineages or the dark matter of unknown taxa [69]. Our results showed that little is known about the genes and their coded functions present in marine microbial communities. When analyzing known functions among abundant microbes, some metabolic pathways could be described, but they represented the most common metabolic pathways involved in primary metabolic processes, such as photosynthesis or flagellar assembly (Fig. S6).

Conclusion

In this work, we show that the rare microbial gene pool of the marine environment is made of key species and represents a large number of potentially novel functions. In addition, based on the presence of viruses in the rare fraction, we hypothesized that the state of rarity could be maintained by viral lysis. However, the procedures used in this study were not dedicated to the detection of viruses and thus a large diversity may have escaped detection. A metagenomic based approach helps the challenging characterization of the members of the rare biosphere and promotes the discovery of new putative functions.

406 **References**

407

408 1. Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, et al. Accurate,
409 multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res* 2015; **25**:
410 534–543.
411 2. Bankevich A, Pevzner PA. Joint Analysis of Long and Short Reads Enables Accurate Estimates
412 of Microbiome Complexity. *Cell Systems* 2018; **7**: 192-200.e3.
413 3. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT. Individual genome assembly from
414 complex community short-read metagenomic datasets. *ISME J* 2012; **6**: 898–901.
415 4. Lapidus AL, Korobeynikov AI. Metagenomic Data Assembly – The Way of Decoding
416 Unknown Microorganisms. *Front Microbiol* 2021; **12**.
417 5. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and
418 assembly of genomes and genetic elements in complex metagenomic samples without using
419 reference genomes. *Nat Biotech* 2014; **32**: 822–828.
420 6. Biller SJ, Berube PM, Lindell D, Chisholm SW. Prochlorococcus: the structure and function of
421 collective diversity. *Nat Rev Microbiol* 2015; **13**: 13–27.
422 7. Crespo BG, Wallhead PJ, Logares R, Pedrós-Alió C. Probing the Rare Biosphere of the North-
423 West Mediterranean Sea: An Experiment with High Sequencing Effort. *PLOS ONE* 2016; **11**:
424 e0159195.
425 8. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity
426 in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* 2006; **103**:
427 12115–12120.
428 9. Pedrós-Alió C. Dipping into the Rare Biosphere. *Science* 2007; **315**: 192–193.
429 10. Sauret C, Séverin T, Vétion G, Guigue C, Goutx M, Pujo-Pay M, et al. 'Rare biosphere'
430 bacteria as key phenanthrene degraders in coastal seawaters. *Environmental Pollution* 2014; **194**:
431 246–253.
432 11. Kalenitchenko D, Le Bris N, Peru E, Galand PE. Ultra-rare marine microbes contribute to key
433 sulfur related ecosystem functions. *Mol Ecol* 2018; **27**: 1494–1504.
434 12. Capo E, Debroas D, Arnaud F, Guillemot T, Bichet V, Millet L, et al. Long-term dynamics in
435 microbial eukaryotes communities: a palaeolimnological view based on sedimentary DNA. *Mol*
436 *Ecol* 2016; **25**: 5925–5943.
437 13. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. *Nat Rev Micro* 2015;
438 **13**: 217–229.
439 14. Debroas D, Hugoni M, Domaizon I. Evidence for an active rare biosphere within freshwater
440 protists community. *Mol Ecol* 2015; **24**: 1236–1247.
441 15. Banerjee S, Schlaeppi K, Heijden MGA. Keystone taxa as drivers of microbiome structure and
442 functioning. *Nature Reviews Microbiology* 2018; 1.
443 16. Herren CM, McMahon KD. Keystone taxa predict compositional change in microbial
444 communities. *Environ Microbiol* 2018; **20**: 2207–2217.
445 17. Hugoni M, Taib N, Debroas D, Domaizon I, Dufournel IJ, Bronner G, et al. Structure of the
446 rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *PNAS*
447 2013; **110**: 6004–6009.
448 18. Debroas D, Domaizon I, Humbert J-F, Jardillier L, Lepère C, Oudart A, et al. Overview of
449 freshwater microbial eukaryotes diversity: a first analysis of publicly available metabarcoding data.
450 *FEMS Microbiol Ecol* 2017; **93**.
451 19. Elshahed MS, Youssef NH, Spain AM, Sheik C, Najar FZ, Sukharnikov LO, et al. Novelty and
452 Uniqueness Patterns of Rare Members of the Soil Biosphere. *Appl Environ Microbiol* 2008; **74**:
453 5422–5428.
454 20. Pascoal F, Magalhães C, Costa R. The Link Between the Ecology of the Prokaryotic Rare
455 Biosphere and Its Biotechnological Potential. *Front Microbiol* 2020; **11**.

21. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 2013; **499**: 431–437.
22. Delmont TO, Eren AM, Maccario L, Prestat E, Esen ÖC, Pelletier E, et al. Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. *Frontiers in Microbiology* 2015; **6**.
23. Sachdeva R, Campbell BJ, Heidelberg JF. Rare microbes from diverse Earth biomes dominate community activity. *bioRxiv* 2019; 636373.
24. Galand PE, Pereira O, Hochart C, Auguet JC, Debroas D. A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *The ISME Journal* 2018; **12**: 2470–2478.
25. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**: 2114–2120.
26. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 2012; **28**: 1420–1428.
27. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010; **26**: 589–595.
28. Ulyantsev VI, Kazakov SV, Dubinkina VB, Tyakht AV, Alexeev DG. MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data. *Bioinformatics* 2016; btw312.
29. Dixon P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* 2003; **14**: 927–930.
30. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013; **41**: D590–D596.
31. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods* 2015; **12**: 902–903.
32. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. *Genome Biology* 2011; **12**: R60.
33. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017; **45**: D158–D169.
34. 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–D462.
35. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Meth* 2015; **12**: 59–60.
36. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* 2014; **2**: 15.
37. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* 2017; **8**.
38. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 2009; **10**: 161.
39. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 2013; **29**: 1830–1831.
40. Rohart F, Gautier B, Singh A, Cao K-AL. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology* 2017; **13**: e1005752.
41. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems* 2015; **143**: 85–96.
42. Plaza Oñate F, Le Chatelier E, Almeida M, Cervino ACL, Gauthier F, Magoulès F, et al. MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* 2019; **35**: 1544–1552.
43. Csardi G, Nepusz T. The igraph software package for complex network research. 2006; 1695.

508 44. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the
509 quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*
510 2015; **25**: 1043–1055.
511 45. Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: Network
512 Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software* 2012; **48**: 1–
513 18.
514 46. Lambert S, Tragin M, Lozano J-C, Ghiglione J-F, Vaulot D, Bouget F-Y, et al. Rhythmicity of
515 coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations.
516 *The ISME Journal* 2019; **13**: 388.
517 47. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-
518 resolution sample inference from Illumina amplicon data. *Nat Methods* 2016; **13**: 581–583.
519 48. Galand PE, Casamayor EO, Kirchman DL, Lovejoy C. Ecology of the rare microbial biosphere
520 of the Arctic Ocean. *PNAS* 2009; **106**: 22427–22432.
521 49. Bankevich A, Pevzner PA. Joint Analysis of Long and Short Reads Enables Accurate Estimates
522 of Microbiome Complexity. *Cell Systems* 2018.
523 50. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT. Individual genome assembly from
524 complex community short-read metagenomic datasets. *The ISME Journal* 2011; **6**: 898–901.
525 51. Campbell BJ, Yu L, Heidelberg JF, Kirchman DL. Activity of Abundant and Rare Bacteria in a
526 Coastal Ocean. *Proc Natl Acad Sci USA* 2011; **108**: 12776–12781.
527 52. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al. SAR11 clade
528 dominates ocean surface bacterioplankton communities. *Nature* 2002; **420**: 806–810.
529 53. Bouvier T, del Giorgio PA. Key role of selective viral-induced mortality in determining marine
530 bacterial community composition. *Environ Microbiol* 2007; **9**: 287–297.
531 54. Thingstad TF, Våge S, Storesund JE, Sandaa R-A, Giske J. A theoretical analysis of how strain-
532 specific viruses can control microbial species diversity. *PNAS* 2014; **111**: 7813–7818.
533 55. Pedrós-Alió C. Marine microbial diversity: can it be determined? *Trends in Microbiology* 2006;
534 **14**: 257–263.
535 56. Gobet A, Böer SI, Huse SM, van Beusekom JEE, Quince C, Sogin ML, et al. Diversity and
536 dynamics of rare and of resident bacterial populations in coastal sands. *ISME J* 2012; **6**: 542–553.
537 57 Pascoal F, Costa R, Assmy P, Duarte P, Magalhães C. Exploration of the Types of Rarity in
538 the Arctic Ocean from the Perspective of Multiple Methodologies. *Microb Ecol* 2021.
539 58. Huete-Stauffer TM, Arandia-Gorostidi N, Díaz-Pérez L, Morán XAG. Temperature
540 dependences of growth rates and carrying capacities of marine bacteria depart from metabolic
541 theoretical predictions. *FEMS Microbiol Ecol* 2015; **91**.
542 59. Arandia-Gorostidi N, Huete-Stauffer TM, Alonso-Sáez L, G. Morán XA. Testing the metabolic
543 theory of ecology with marine bacteria: different temperature sensitivity of major phylogenetic
544 groups during the spring phytoplankton bloom. *Environmental Microbiology* 2017; **19**: 4493–4505.
545 60. Giovannoni SJ, Bibbs L, Cho J-C, Stapels MD, Desiderio R, Vergin KL, et al. Proteorhodopsin
546 in the ubiquitous marine bacterium SAR11. *Nature* 2005; **438**: 82–85.
547 61. Yilmaz P, Yarza P, Rapp JZ, Glöckner FO. Expanding the World of Marine Bacterial and
548 Archaeal Clades. *Front Microbiol* 2016; **6**.
549 62. Pedler BE, Aluwihare LI, Azam F. Single bacterial strain capable of significant contribution to
550 carbon cycling in the surface ocean. *Proc Natl Acad Sci USA* 2014; **111**: 7202.
551 63. Pereira O, Hochart C, Boeuf D, Auguet JC, Debroas D, Galand PE. Seasonality of archaeal
552 proteorhodopsin and associated Marine Group IIb ecotypes (Ca . Poseidoniales) in the North
553 Western Mediterranean Sea. *The ISME Journal* 2020; 1–15.
554 64. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. Untangling
555 Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. *Science*
556 2012; **335**: 587–590.
557 65. Pereira O, Hochart C, Auguet JC, Debroas D, Galand PE. Genomic ecology of Marine Group
558 II, the most common marine planktonic Archaea across the surface ocean. *MicrobiologyOpen* 2019;
559 **8**: e00852.

66. Tully BJ. Metabolic diversity within the globally abundant Marine Group II Euryarchaea offers insight into ecological patterns. *Nat Commun* 2019; **10**: 271.
67. Xie W, Luo H, Murugapiran SK, Dodsworth JA, Chen S, Sun Y, et al. Localized high abundance of Marine Group II archaea in the subtropical Pearl River Estuary: implications for their niche adaptation. *Environ Microbiol* 2018; **20**: 734–754.
68. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, et al. Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J* 2017; **11**: 853–862.
69. Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biol Evol* 2018; **10**: 707–715.
70. Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, et al. A global ocean atlas of eukaryotic genes. *Nature Communications* 2018; **9**: 373.
71. Thomas AM, Segata N. Multiple levels of the unknown in microbiome research. *BMC Biology* 2019; **17**: 48.

Aknowledgements

This work was supported by the Agence Nationale de la Recherche (ANR) through the projects EUREKA (ANR-14-CE02-0004-01). We thank the captain and crew of the Nereis II, Eric Maria, and Louise Oriol for assisting with the collection and analysis of samples over the time series. We extend our acknowledgments to all the researchers that were involved in working with the time series over the years. We are grateful to the Mésocentre Clermont Auvergne University (https://mesocentre.uca.fr/) for providing help, computing and storage resources.

Competing interests

The authors declare no competing interests.

Figures

Figure 1: Schematic showing the bioinformatic analysis conducted to separate assembled and un-assembled reads from a 3-year metagenomic time series dataset.

Figure 2. Pairwise comparisons of similarity between communities in relation to the time separating two samples. The similarity was measured by a direct metagenome-to-metagenome comparison of the read content for the unassembled (a) and assembled ones (b).

Figure 3. Cladogram showing the taxonomic position of the unassembled (orange) and assembled (blue) fractions and their relative abundance. Each circle diameter is proportional to the taxon's abundance, and the color represents which branch of the phylogenetic tree is more abundant in each fraction.

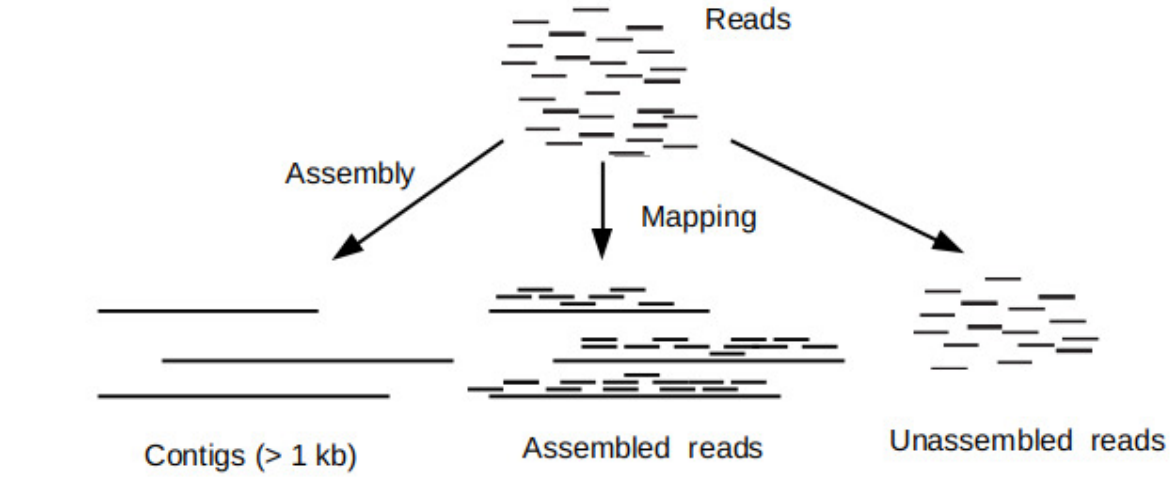
Figure 4. Distribution of the identities between assembled and unassembled reads against the UNIREF90 database.

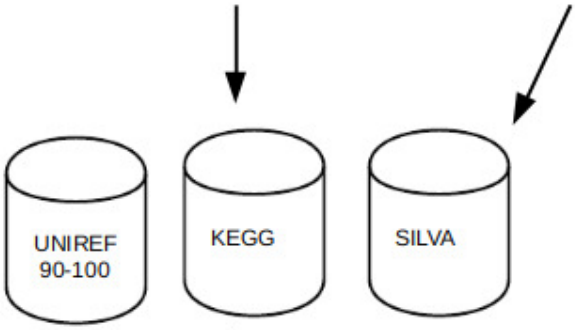
Figure 5. Sparse Principal Component Analysis conducted of the read composition annotated against the UNIREF90 (top) and the KO databases (bottom). The ANOSIM statistics based on the Bray-Curtis similarity were $R=0.63$ ($P<0.01$) for the UNIREF90 dataset and $R=0.90$ ($P<0.01$) for KO results.

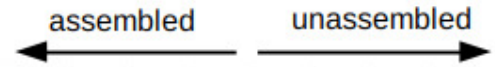
Figure 6. Network representation of the relationship between uCAG (square vertices), aCAG (circle) and physicochemical parameters (rectangle, T: temperature, Ox: oxygen and N: nitrite) and Louvain clusters. Red lines between nodes indicate negative Spearman correlations whereas grey edges correspond to positive correlations. The table below the graphics shows the best keystones in

612 the network inferred from the « ExpectedInfluence » parameter (see Fig. S8). The numbers in the
613 first column correspond to the numbering of the vertex in the network.
614
615 **Tables**
616
617 Table 1. Distribution of the SILVA, UNIREF90 and KEGG clusters among the mapped and
618 unmapped reads. Differences between both categories were considered significant ($P < 0.05$) when
619 the Welch and Wilcoxon tests were convergent; the enrichment were inferred from the log fold
620 computed by the ALDEx2 procedure.
621
622 **Supplementary materials**
623
624 Fig S1. NMDS based on Bray Curtis dissimilarity computed from MetaFast separating assembled
625 and unassembled microbial communities.
626 Fig S2. spca analysis of microbial communities based on 16S rRNA extracted from the assembled
627 and unassembled fractions of the metagenomes.
628 Fig S3. Abundant (top) and rare (bottom) communities deciphered by metabarcoding.
629 Fig S4. Distribution of the assembled and unassembled reads aligned to the UNIREF90 (A) and
630 UNIREF100 (B) databases.
631 Fig S5. Distribution of the scores and identities of the assembled and unassembled reads aligned to
632 the UNIREF90 (A) and UNIREF100 (B) databases.
633 Fig S6. Photosynthesis and flagellar assembly pathways. Green rectangles corresponds to an en-
634 richment in the assembled reads and red in the unassembled reads.
635 Fig S7 Temporal dynamics (z-scores) of the unassembled and assembled CAGs inside the main
636 network clusters assessed by the Louvain methods. The clusters composed of less than 3 vertices
637 are not represented. The grey rectangle represents spring and summer periods. The table displays

638 the mains metabolic pathways in the clusters 16 and 17 (Any pathway with at least 25 % of the KOs
639 were detected in the other clusters),
640 Fig S8. « ExpectedInfluence » parameter computed from the network with the package qgraph un-
641 der R.
642
643 Table S1. Sampling date and environmental parameters
644 Table S2. Sequencing and main statistics.
645 Table S3. Effects of the cleaning procedures on the functional abundance tables
646 Table S4. Completeness, contamination and taxonomy of the CAGs built with the three methods
647 described in the materials and methods (ass : assembled or aCAG - unass : unassembled or uCAG –
648 cano : Canopy method (Nielsen et al. 2014) - miner-msp : MSPminer method (Plaza Oñate et al.
649 2019) - mixo: new approach described in materials and methods section)
650







Functional abundance table:
Features x samples

K00007	148	142	37	74	85	85	260	454
K00008	1606	1357	518	816	819	717	1725	1256
K00009	8	5	1	27	31	36	33	48
K00010	2641	3079	1239	1686	1883	1532	2039	2316
K00011	2	0	0	6	10	6	10	6
K00012	12556	10815	4819	7787	7067	6257	8766	9195
K00013	11475	9557	4588	6933	6455	6223	9303	10041
K00014	1861	1875	821	1293	1387	1344	2206	2358
K00015	3813	2575	1016	1320	1395	1190	1992	1824
K00016	16	14	13	83	88	75	123	48
K00018	240	168	105	195	210	117	209	130
K00019	2292	1807	821	1222	1235	1054	1513	1482

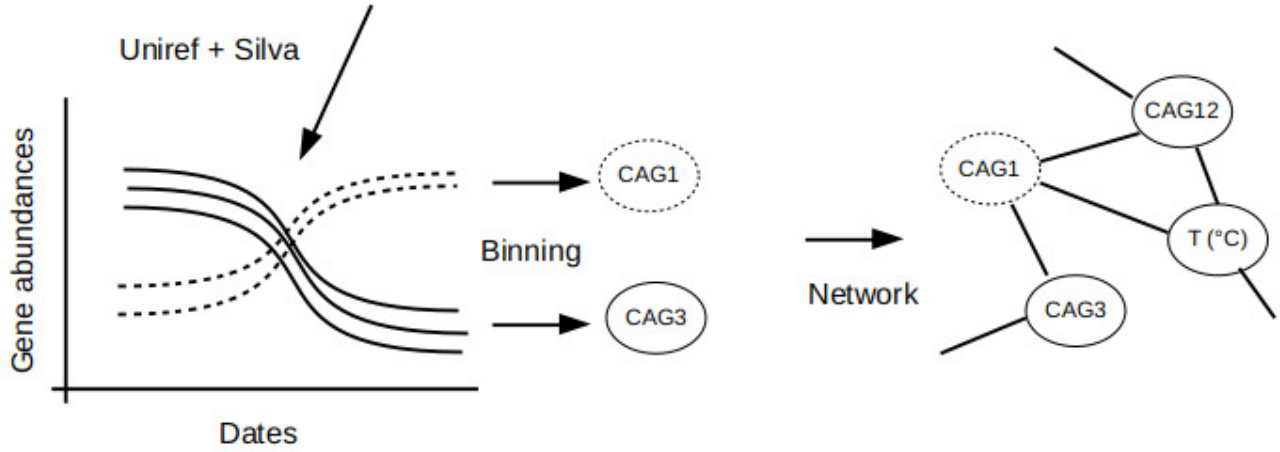


Fig 1

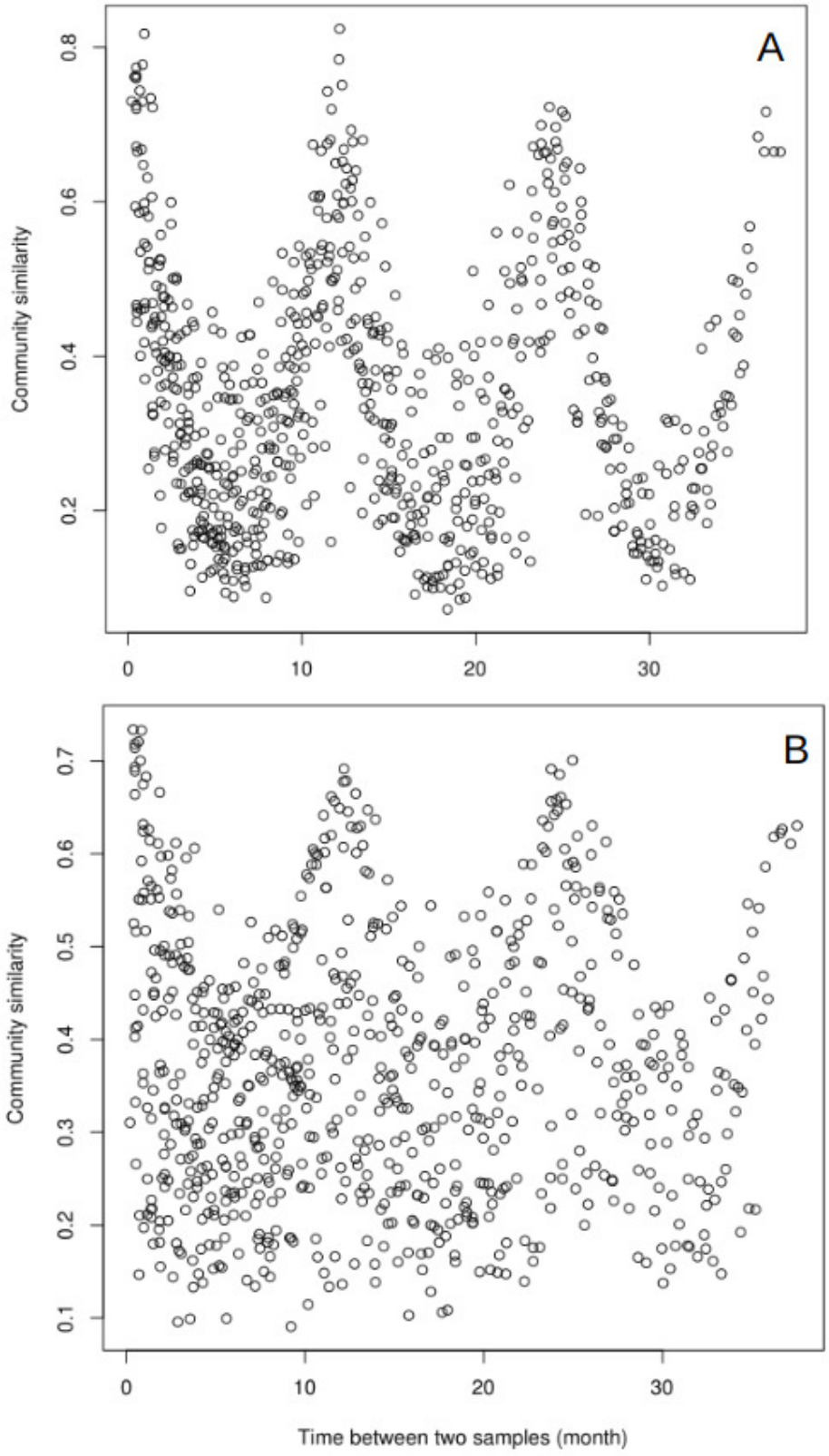

A
B
Community similarity
Community similarity
0.2
0.4
0.6
0.8
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
0
10
20
30
Time between two samples (month)

Fig 2

a: Rhizobiales
b: Alphaproteobacteria_noname
c: Lentisphaerales
d: Flavobacteriales
e: Pseudomonadales
f: Gammaproteobacteria_noname
g: Vibrionales
h: Bdellovibrionales
i: Rickettsiales
j: Oceanospirillales
k: Nitrosopumilales
l: Caudovirales
m: Viruses_noname
n: Xanthomonadales
o: Verrucomicrobia_noname
p: Burkholderiales
q: Methylophilales
r: Rhodobacterales
s: Enterobacteriales
t: Caulobacterales
u: Actinomycetales

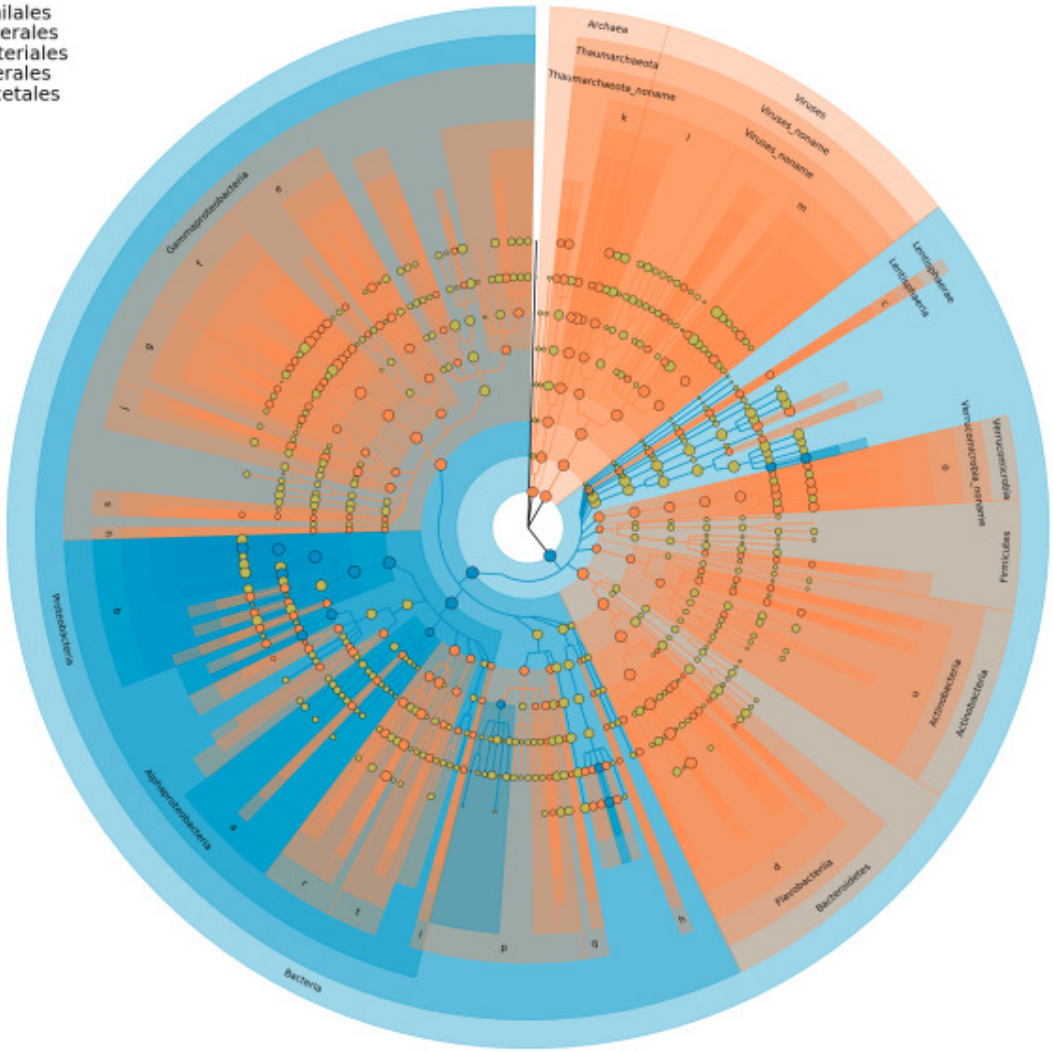


Fig 3:

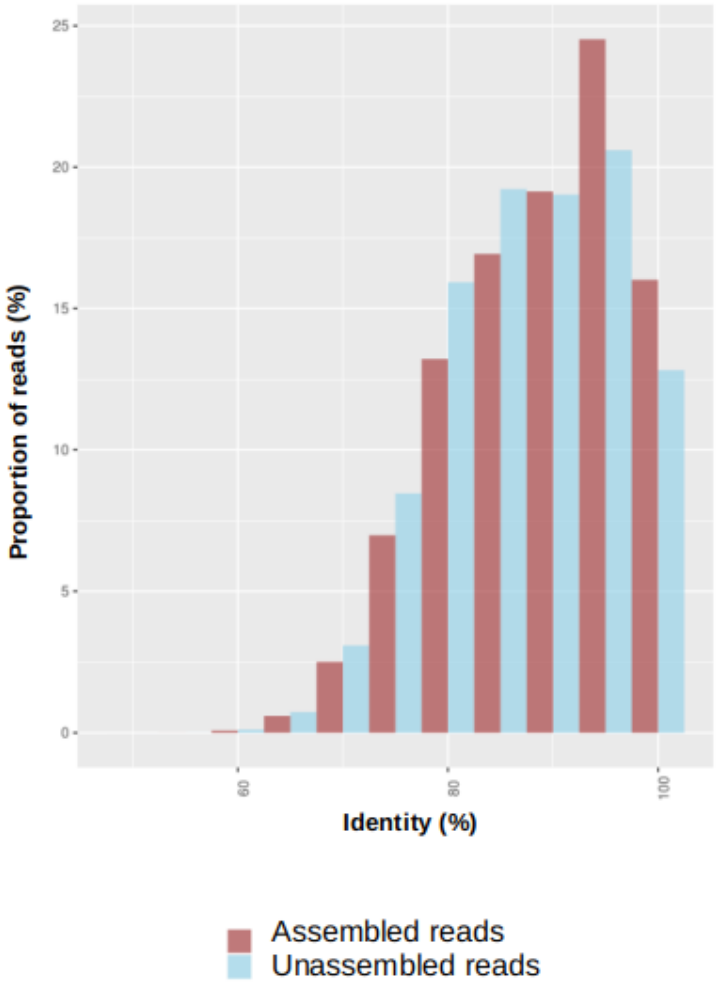

Proportion of reads (%)
25
20
15
10
5
0
60
80
100
Identity (%)
Assembled reads
Unassembled reads

Fig 4

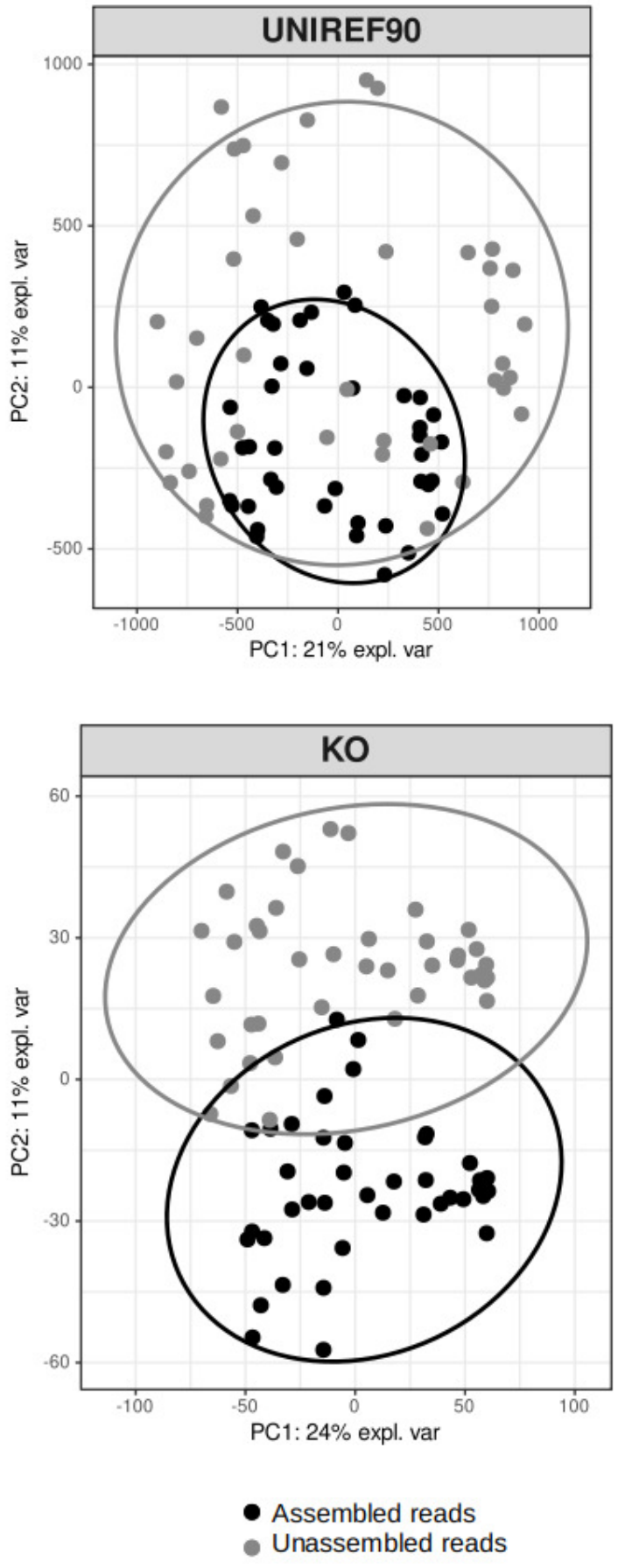
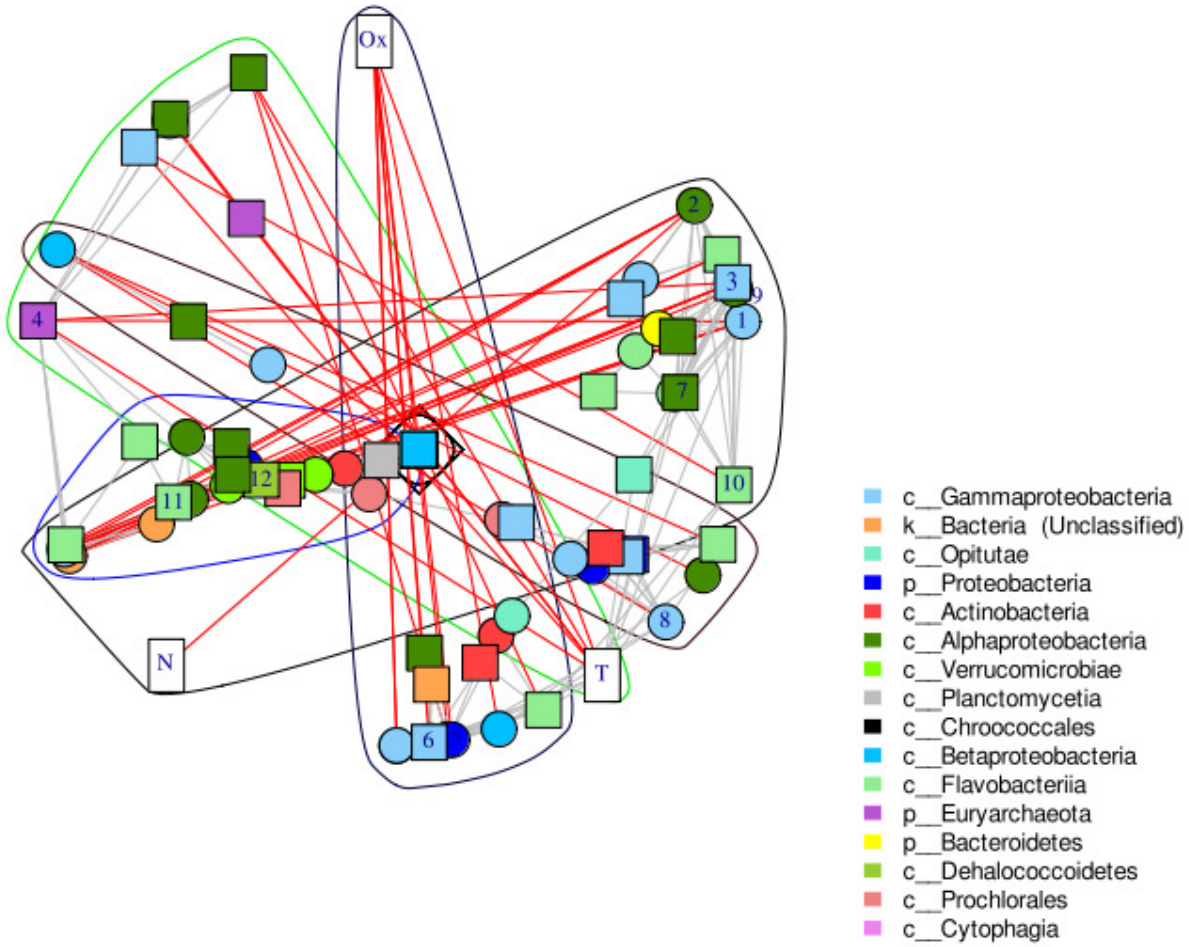

UNIREF90
1000
500
0
-500
PC2: 11% expl. var
-1000
-500
0
500
1000
PC1: 21% expl. var
KO
60
30
0
-30
-60
PC2: 11% expl. var
-100
-50
0
50
100
PC1: 24% expl. var
Assembled reads
Unassembled reads

Fig 5.



Nodes		Taxonomy		
T	Temperature			
1	Assembled	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3
2	Assembled	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales
3	Unassembled	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3
4	Unassembled	p__Euryarchaeota		
5	Assembled	p__Proteobacteria	c__Alphaproteobacteria	o__Rhizobiales
6	Unassembled	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3
7	Assembled	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales
8	Assembled	p__Proteobacteria	c__Gammaproteobacteria	
9	Assembled	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales
10	Unassembled	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales
11	Unassembled	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales
12	Unassembled	p__Chloroflexi	c__Dehalococcoidetes	

Fig. 6

Table S1. Sampling date and environmental parameters

	Date	Temperature (°C)	Salinity (PSU)	Oxygen (mL/L)	pH	NH4	NO3 (µmol/L)	NO2 (µmol/L)	PO4 (µmol/L)	SIOH4 (µmol/L)	CHLA (µg/L)	Day Length (h)
Sample 1	03/01/2012	13.8	38.1	13.8	38.1	0.0	0.6	0.2	0.0	0.7	0.7	9.2
Sample 2	31/01/2012	11.8	38.0	11.8	38.0	0.0	1.6	0.4	0.0	0.3	0.3	10.0
Sample 3	21/02/2012	10.5	38.2	10.5	38.2	0.0	2.3	0.2	0.0	0.5	0.5	10.9
Sample 4	07/03/2012	10.9	38.2	10.9	38.2	0.1	1.1	0.1	0.1	0.8	0.8	11.6
Sample 5	13/03/2012	11.2	38.2	11.2	38.2	0.1	0.5	0.2	0.0	0.6	0.6	11.9
Sample 6	04/04/2012	13.8	37.9	13.8	37.9	0.0	0.4	0.1	0.0	0.3	0.3	12.9
Sample 7	23/04/2012	13.3	38.2	13.3	38.2	0.0	1.1	0.1	0.1	0.3	0.3	13.8
Sample 8	09/05/2012	15.5	36.6	15.5	36.6	0.0	0.9	0.1	0.1	1.8	1.8	14.4
Sample 9	07/06/2012	19.5	37.6	19.5	37.6	0.0	0.2	0.0	0.0	0.1	0.1	15.2
Sample 10	12/07/2012	20.1	37.9	20.1	37.9	0.0	0.0	0.0	0.0	0.1	0.1	15.0
Sample 11	06/08/2012	21.7	38.0	21.7	38.0	0.1	0.1	0.0	0.0	0.2	0.2	14.2
Sample 12	20/08/2012	22.9	38.2	22.9	38.2	0.1	0.1	0.0	0.0	0.2	0.2	13.7
Sample 13	22/10/2012	18.2	38.2	18.2	38.2	0.3	0.3	0.1	0.1	0.5	0.5	10.7
Sample 14	05/11/2012	16.6	37.9	16.6	37.9	0.2	0.6	0.1	0.1	0.6	0.6	10.1
Sample 15	19/11/2012	15.4	37.9	15.4	37.9	0.3	0.9	0.2	0.1	0.4	0.4	9.6
Sample 16	12/12/2012	13.1	38.0	13.1	38.0	0.0	1.4	0.2	0.0	0.5	0.5	9.1
Sample 17	15/01/2013	12.7	37.6	12.7	37.6	0.0	0.9	0.3	0.1	1.2	1.2	9.4
Sample 18	04/02/2013	11.1	38.1	11.1	38.1	0.0	1.8	0.3	0.1	1.2	1.2	10.1
Sample 19	11/03/2013	11.5	34.7	11.5	34.7	0.1	5.9	0.2	0.2	2.6	2.6	11.7
Sample 20	22/04/2013	13.1	37.1	13.1	37.1	0.2	1.8	0.2	0.0	0.5	0.5	13.7
Sample 21	06/05/2013	13.9	37.3	13.9	37.3	0.1	1.6	0.2	0.1	0.2	0.2	14.3
Sample 22	03/06/2013	15.0	37.8	15.0	37.8	0.0	0.2	0.0	0.1	1.4	1.4	15.1
Sample 23	01/07/2013	19.1	37.9	19.1	37.9	0.0	0.1	0.0	0.0	0.5	0.5	15.3
Sample 24	26/08/2013	22.6	37.9	22.6	37.9	0.1	0.2	0.0	0.0	0.1	0.1	13.4
Sample 25	28/10/2013	19.0	36.8	19.0	36.8	0.5	1.6	0.1	0.0	1.1	1.1	10.5
Sample 26	13/11/2013	16.5	37.9	16.5	37.9	0.1	0.2	0.0	0.0	1.0	1.0	9.9
Sample 27	12/12/2013	12.7	38.2	12.7	38.2	0.0	2.5	0.2	0.1	0.5	0.5	9.1
Sample 28	24/02/2014	12.7	38.0	5.7	8.1	0.1	1.7	0.3	0.1	2.5	0.7	11.0
Sample 29	07/04/2014	13.5	37.1	6.0	8.0	0.2	1.1	0.1	0.0	0.5	2.5	13.0
Sample 30	22/04/2014	15.2	37.4	5.5	8.2	0.0	0.1	0.1	0.0	0.8	1.2	13.7
Sample 31	10/06/2014	17.7	37.7	5.5	8.1	0.0	0.0	0.0	0.0	0.8	0.3	15.2
Sample 32	21/07/2014	20.5	37.9	5.0	8.1	0.0	0.2	0.0	0.0	2.8	0.3	14.8
Sample 33	04/08/2014	22.0	37.8	4.2	8.1	0.0	0.2	0.0	0.0	0.7	0.3	14.4
Sample 34	01/09/2014	21.7	37.9	5.1	8.6	0.0	0.0	0.1	0.0	0.5	0.2	13.2
Sample 35	12/11/2014	18.2	38.1	5.0	8.1	0.1	0.3	0.1	0.0	1.0	0.3	9.9
Sample 36	24/11/2014	17.3	37.7	5.3	8.1	0.4	0.7	0.1	0.2	1.2	0.6	9.5
Sample 37	08/12/2014	16.2	37.5	5.5	8.1	0.2	1.4	0.3	0.0	3.6	0.3	9.2
Sample 38	08/01/2015	13.3	37.8	5.8	8.1	0.1	0.5	0.3	0.0	1.8	0.7	9.3
Sample 39	22/01/2015	12.7	37.8	5.9	8.1	0.1	0.3	0.2	0.0	1.2	1.1	9.6
Sample 40	02/02/2015	12.6	38.1	5.8	8.2	0.1	1.4	0.2	0.0	3.8	0.6	10.0

Table S2. Sequencing and main statistics.

							UNIREF 90		UNIREF100		KO		SILVA	
Sample	**Raw reads**	**HQ reads**	**Contigs ≥ 1kb**	**Total contig length (nt)**	**unmapped reads**	**mapped reads**	**unmapped reads**	**mapped reads**	**unmapped reads**	**mapped reads**	**unmapped reads**	**mapped reads**	**unmapped reads**	**mapped reads**
Sample 1	59273132	57958251	41244	98262145	44593508	13364743	17744984	6314543	14117390	5010681	6766110	2603287	27983	4990
Sample 2	80826740	79163692	48367	123062579	62841116	16322576	25548060	7813758	21145099	6415748	10024110	3513576	31369	4955
Sample 3	71312636	69857457	41513	111756344	54261976	15595481	22433056	7568176	18701570	6361829	9401751	3708808	29211	4519
Sample 4	80636430	79362166	39939	109512619	63381770	15980396	23922649	6976595	20328038	5995649	11540450	3655534	31165	3547
Sample 5	95291086	93131454	56595	169734749	70714308	22417146	27075890	9352797	22942805	7941408	12795658	4812572	26767	4591
Sample 6	52835444	51942847	35548	98399996	40060850	11881997	14437818	4926161	11961467	4159561	6874015	2527812	20607	3128
Sample 7	51079736	49380567	40508	102758931	38149046	11231521	15608984	5234680	12979380	4352408	6588399	2374903	26193	4676
Sample 8	73445534	71280727	49778	126501323	53386528	17894199	16974682	5948624	13877432	5018846	7253899	2796688	19265	3707
Sample 9	90753470	87683664	48018	126073118	66472599	21211065	23456209	7412645	19771960	6479531	10146005	3388588	27009	7245
Sample 10	104860772	102293843	57819	179149121	77392016	24901827	27983866	9789154	23196981	8293394	12343189	4593072	18863	4016
Sample 11	74392204	71640388	52397	156895851	55492119	16148269	22447996	7455555	18604685	6349977	10385948	3648889	31133	6080
Sample 12	34871574	33772351	31070	76907204	26176962	7595389	10468840	3530427	8736353	3035491	4846031	1726263	14469	2821
Sample 13	72532844	70335589	54369	140748920	54403140	15932449	20873638	7298011	16941868	5980003	8151361	2964819	25543	4931
Sample 14	74872672	72846496	61964	149954009	55958010	16888486	20720685	7510087	16743569	6055242	8069160	3116919	25498	4894
Sample 15	64281530	63027247	53199	120289337	47399605	15627642	18932849	7417287	15201591	5968001	7377584	3093858	23716	4616
Sample 16	106010408	102302513	61926	145187693	78377597	23924916	27240211	10227759	21443828	8207855	10600300	4477385	34273	6052
Sample 17	100810750	98495009	49889	116274315	76504402	21990607	27366704	9556273	22243394	7808796	11608378	4477537	32466	5889
Sample 18	48838402	47515963	37698	89710208	36673413	10842550	16314197	5474541	13579889	4571057	6560683	2516221	21729	3732
Sample 19	46597854	44964848	23994	70893350	35466177	9498671	11144115	3249687	9957275	2894784	6093029	1738209	17482	4302
Sample 20	73729860	71469602	52630	148504862	55238042	16231560	22624288	7468118	18546850	6220009	10637098	3801247	29936	5649
Sample 21	65923978	63996584	44790	145087132	49337751	14658833	19797958	6278121	16556082	5218864	8858459	2838609	26232	7022
Sample 22	58442788	56487206	44179	112636265	44041597	12445609	19554352	5699133	16193963	4777311	8311466	2566026	23809	4256
Sample 23	98347746	95156563	41390	122193270	75858100	19298463	26378686	7150324	21900089	6173804	11792198	3420789	29423	8161
Sample 24	84345622	81841740	50854	161740082	62989261	18852479	23712351	7887602	19228848	6547768	10639223	3693963	30358	7162
Sample 25	60805948	59212336	52256	148645942	44086255	15126081	17889251	6636137	14403477	5384308	7311577	2801849	20559	5512
Sample 26	64913662	63090391	58835	140734121	47145774	15944617	19257860	7558899	15567740	6110545	7886946	3220050	25614	5743
Sample 27	68579126	66588511	44217	105396708	50939677	15648834	19738834	7581930	15372615	6041119	7432843	3202869	25946	5904
Sample 28	79561348	77673692	58673	142607160	59188672	18485020	23569642	9061095	19009395	7386984	9301854	4013124	29089	6043
Sample 29	71354090	69881637	61803	169475995	52749658	17131979	21761777	7349253	18072931	6109384	9626090	3291964	25882	4104
Sample 30	113230060	110798145	51209	156651985	88814754	21983391	28957198	8097672	23718462	6667758	13703444	4004660	33717	6290
Sample 31	114367198	111753381	68044	186932189	86065519	25687862	33574233	11208336	27891691	9520448	15782159	5630437	31003	4906
Sample 32	54582720	52700198	44532	110010378	40640046	12060152	14851597	5027511	12009786	4189769	6521950	2364754	17793	3138
Sample 33	68653980	66202174	57680	155231185	49950779	16251395	21268390	7754553	17368462	6598308	9830173	3873255	30580	4946
Sample 34	70368772	68587662	58720	166717960	50882440	17705222	21513514	8236950	17481664	6889401	9275249	3706257	26127	5610
Sample 35	74795486	72529047	67151	143637806	54130612	18398435	20988244	8631233	16646564	6876798	8052044	3460407	28350	5386
Sample 36	107935088	105270469	88088	220361755	76296829	28973640	27344301	12150638	22016656	9790225	10592706	4766742	34635	8128
Sample 37	74502570	72806977	60035	142623448	54108954	18698023	19647035	8813076	15561830	6547609	7226123	3247426	24875	6020
Sample 38	76479148	74514541	57514	134746061	57119117	17395424	20706268	7716631	16547788	6095887	7811834	3129854	28238	5325
Sample 39	69334770	68072943	49362	130291696	52196745	15876198	21793932	7646863	17591673	6120515	8854864	3325714	27517	4641
Sample 40	50666858	48966052	33115	75897135	37767712	11198340	14552988	5307755	11578548	4217459	5623197	2297260	19865	3619

Table S3. Effects of the cleaning procedures on the functional abundance tables

	Before cleaning		After cleaning	
Databases	Features	Reads	Features	Reads
SILVA	6959	1260545	846	1191644
KEGG	9826	4.97E+08	6984	4.97E+08
uniref90	7725889	1.15E+09	1210645	1.09E+09
Uniref100	8471020	9.40E+08	NA	NA

Table S4. Completeness, contamination and taxonomy of the CAGs built with the three methods described in the materials and methods.
(ass : assembled or aCAG - unass : unassembled or uCAG – cano : Canopy method (Nielsen et al. 2014) - miner-msp : MSPminer method (Plaza Oñate et al. 2019) – mixo: new approach described in materials and methods section)

CAGs	Completeness	Contamination	Kingdom	Phylum	Class	Order	Family	Genus	Species
ass-cano-CAG0026	61.2	4.13	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae	g__Nisaea	
ass-cano-CAG0035	46.22	0.84	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Rhodobacteraceae		
ass-cano-CAG0040	40.47	5.17	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria				
ass-cano-CAG0045	49.54	1.01	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae		
ass-cano-CAG0048	35.66	0	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria				
ass-cano-CAG0049	49.34	9.44	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3	f__Alteromonadaceae		
ass-cano-CAG0050	34.85	1.72	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3	f__Alteromonadaceae		
ass-cano-CAG0051	52.59	0.92	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Rhodobacteraceae		
ass-cano-CAG0056	35.75	5.17	k__Bacteria	p__Verrucomicrobia	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Verrucomicrobiaceae		
ass-cano-CAG0059	51.77	7.78	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Flavobacteriaceae		
ass-cano-CAG0060	53.72	5.78	k__Bacteria	p__Actinobacteria	c__Actinobacteria				
ass-cano-CAG0061	36.26	5.13	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Sphingomonadales	f__Sphingomonadaceae_3	g__Sphingobium	
ass-cano-CAG0064	67.45	9.1	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria				
ass-cano-CAG0066	52.26	1.63	k__Bacteria	p__Cyanobacteria	c__Chroococcales	o__Chroococcales	f__Cyanobium		
ass-cano-CAG0068	47.53	1.83	k__Bacteria	p__Verrucomicrobia	c__Opitutae	o__Opitutales			
ass-cano-CAG0069	43.1	0	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3	f__Alteromonadaceae		
ass-cano-CAG0070	70.2	1.92	k__Bacteria	p__Cyanobacteria	c__Prochlorales	o__Prochlorales	f__Prochlorococcaceae	g__Prochlorococcus	s__Prochlorococcus_marinus
ass-cano-CAG0072	44.45	9.47	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae		
ass-cano-CAG0073	49.49	4.01	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales			
ass-cano-CAG0075	40.11	3.51	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Flavobacteriaceae		
ass-cano-CAG0077	58.97	0	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales			
ass-cano-CAG0078	61.33	7.91	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria				
ass-cano-CAG0081	39.87	3.79	k__Bacteria	p__Cyanobacteria	c__Prochlorales	o__Prochlorales	f__Prochlorococcaceae	g__Prochlorococcus	s__Prochlorococcus_marinus
ass-cano-CAG0088	37.23	1.72	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Flavobacteriaceae		
ass-cano-CAG0089	47.85	9.63	k__Bacteria	p__Proteobacteria	c__Betaproteobacteria	o__Methylophilales			
ass-cano-CAG0091	31.27	0	k__Bacteria	p__Actinobacteria	c__Actinobacteria				
ass-cano-CAG0093	39.47	0	k__Bacteria	p__Proteobacteria					
ass-cano-CAG0104	43.36	5.18	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria				
ass-cano-CAG0108	31.99	0.54	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales			
ass-cano-CAG0109	46.33	0.66	k__Bacteria	p__Proteobacteria	c__Betaproteobacteria	o__Methylophilales			
ass-cano-CAG0114	35.8	0.74	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales			
ass-miner-msp_023	58.08	3.86	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Rhodobacteraceae		
ass-miner-msp_025	80.07	4.92	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Rhodobacteraceae		
ass-miner-msp_036	50.17	6.31	k__Bacteria	p__Verrucomicrobia	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Verrucomicrobiaceae		
ass-miner-msp_037	47.87	6.49	k__Bacteria	p__Verrucomicrobia	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Verrucomicrobiaceae		
ass-miner-msp_042	39.09	6.53	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3	f__Alteromonadaceae		
ass-miner-msp_049	41.18	1.72	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria				
ass-miner-msp_054	32.47	0	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae		
ass-miner-msp_056	55.25	2.3	k__Bacteria	p__Planctomycetes	c__Planctomycetia	o__Planctomycetales	f__Planctomycetaceae		
ass-miner-msp_058	49.48	1.72	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Rhodobacteraceae		
ass-miner-msp_059	58.37	3.53	k__Bacteria	p__Cyanobacteria	c__Chroococcales	o__Chroococcales	f__Cyanobium		
ass-miner-msp_060	30.83	0	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria				
ass-miner-msp_064	77.2	6.81	k__Bacteria	p__Actinobacteria	c__Actinobacteria	o__Actinomycetales	f__Microbacteriaceae		
ass-miner-msp_068	55.68	9.55	k__Bacteria	p__Proteobacteria	c__Betaproteobacteria	o__Burkholderiales			
ass-miner-msp_073	68.23	1.97	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales			
ass-miner-msp_074	67.28	3.89	k__Bacteria	p__Cyanobacteria	c__Prochlorales	o__Prochlorales	f__Prochlorococcaceae	g__Prochlorococcus	s__Prochlorococcus_marinus
ass-miner-msp_079	36.04	1.73	k__Bacteria	p__Proteobacteria					
ass-miner-msp_082	54.77	2	k__Bacteria	p__Proteobacteria					
ass-miner-msp_083	38.01	0.78	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria				
ass-miner-msp_088	36.36	0	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3	f__Alteromonadaceae		
ass-miner-msp_089	32.4	0.93	k__Archaea	p__Euryarchaeota					
ass-miner-msp_090	50.11	0.54	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Cryomorphaceae		
ass-miner-msp_092	36.21	3.45	k__Bacteria	p__Actinobacteria	c__Actinobacteria				
ass-miner-msp_094	35.42	1.72	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Flavobacteriaceae		
ass-miner-msp_096	40.61	5.31	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae		

ass-miner-msp_099	55.8	3.81	k__Bacteria	p__Bacteroidetes				
ass-miner-msp_100	52.77	0.54	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
ass-miner-msp_103	30.25	0	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
ass-miner-msp_106	33.56	1.58	k__Bacteria	p__Actinobacteria	c__Actinobacteria	o__Actinomycetales		
ass-miner-msp_111	35.93	1.16	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Flavobacteriaceae	
ass-miner-msp_112	48.02	2.94	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
ass-miner-msp_113	38.09	4.39	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhizobiales		
ass-miner-msp_117	48.4	8.97	k__Bacteria	p__Proteobacteria	c__Betaproteobacteria	o__Methylophilales		
ass-miner-msp_118	47.1	3.23	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
ass-miner-msp_119	35.44	4.8	k__Bacteria					
ass-miner-msp_123	54.44	1.17	k__Bacteria	p__Proteobacteria	c__Betaproteobacteria	o__Methylophilales		
ass-miner-msp_129	31.5	0	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae	
ass-miner-msp_134	46.09	5.05	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
ass-miner-msp_139	32.26	1.53	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3	f__Alteromonadaceae	
ass-miner-msp_140	41.45	0.68	k__Bacteria	p__Verrucomicrobia	c__Opitutae	o__Opitutales		
ass-mixo-msp_236	39.77	9.32	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Pseudomonadales	f__Moraxellaceae	
ass-mixo-msp_318	51.63	4.81	k__Bacteria	p__Planctomycetes	c__Planctomycetia	o__Planctomycetales	f__Planctomycetaceae	
ass-mixo-msp_34	45.3	9.79	k__Bacteria					
ass-mixo-msp_349	39.55	8.82	k__Bacteria	p__Verrucomicrobia	c__Opitutae	o__Opitutales		
ass-mixo-msp_352	48.95	7.73	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae	
ass-mixo-msp_444	34.5	3.66	k__Bacteria	p__Proteobacteria				
ass-mixo-msp_502	50.91	4.52	k__Bacteria	p__Proteobacteria				
ass-mixo-msp_83	42.16	1.62	k__Bacteria	p__Actinobacteria	c__Actinobacteria	o__Actinomycetales		
unass-cano-CAG0027	82.18	5.73	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Rhodobacteraceae	
unass-cano-CAG0032	84.12	6.73	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria			
unass-cano-CAG0048	77.79	6.87	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria			
unass-cano-CAG0050	82.27	4.93	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Rhodobacteraceae	
unass-cano-CAG0052	67.08	2.72	k__Bacteria	p__Verrucomicrobia	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Verrucomicrobiaceae	
unass-cano-CAG0056	79.6	6.83	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Rhodobacteraceae	
unass-cano-CAG0062	58.16	7.67	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3	f__Alteromonadaceae	
unass-cano-CAG0066	73.66	7.03	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Cryomorphaceae	
unass-cano-CAG0073	50.36	1.94	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Rhodobacteraceae	
unass-cano-CAG0074	72.06	2.64	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Flavobacteriaceae	
unass-cano-CAG0077	73.92	2.81	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Flavobacteriaceae	
unass-cano-CAG0080	61.47	7.28	k__Archaea	p__Euryarchaeota				
unass-cano-CAG0082	39.14	6.96	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3	f__Alteromonadaceae	
unass-cano-CAG0085	50.57	6.29	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
unass-cano-CAG0088	52.33	5.39	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria			
unass-cano-CAG0089	74.99	4.04	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
unass-cano-CAG0090	52.06	8.09	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
unass-cano-CAG0091	57.75	1.74	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
unass-cano-CAG0095	50.34	0	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae	
unass-cano-CAG0098	78.78	1.37	k__Bacteria	p__Proteobacteria	c__Betaproteobacteria	o__Methylophilales		
unass-cano-CAG0102	41.88	2.15	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
unass-cano-CAG0106	40.24	3.74	k__Archaea	p__Euryarchaeota				
unass-cano-CAG0108	53.06	1.35	k__Bacteria	p__Verrucomicrobia	c__Opitutae	o__Opitutales		
unass-cano-CAG0109	77.05	1.35	k__Bacteria	p__Verrucomicrobia	c__Opitutae	o__Opitutales		
unass-cano-CAG0113	34.5	0.18	k__Bacteria	p__Bacteroidetes	c__Cytophagia	o__Cytophagales	f__Cytophagaceae_2	
unass-cano-CAG0114	43.47	5.17	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria			
unass-cano-CAG0115	37.48	1.61	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
unass-cano-CAG0116	38.97	0.27	k__Bacteria	p__Actinobacteria	c__Actinobacteria	o__Actinomycetales		
unass-cano-CAG0117	46.77	0	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales		
unass-cano-CAG0124	36.64	4.76	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Hyphomonadaceae	g__Oceanicaulis
unass-cano-CAG0133	42.31	0.91	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Thiotrichales		
unass-cano-CAG0136	37.2	3.45	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Legionellales		
unass-cano-CAG0145	36.13	3.45	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae	
unass-miner-msp_030	85.08	6.95	k__Bacteria	p__Proteobacteria	c__Betaproteobacteria	o__Burkholderiales		
unass-miner-msp_033	81.93	8.22	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria			
unass-miner-msp_045	76.68	5.44	k__Bacteria	p__Verrucomicrobia	c__Verrucomicrobiae	o__Verrucomicrobiales	f__Verrucomicrobiaceae	
unass-miner-msp_060	38.56	1.72	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3	f__Alteromonadaceae	

unass-miner-msp_063	35.75	2.6	k__Bacteria	p__Planctomycetes	c__Planctomycetia	o__Planctomycetales	f__Planctomycetaceae		
unass-miner-msp_082	81.53	9.22	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales			
unass-miner-msp_083	56.05	1.96	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae		
unass-miner-msp_086	38.71	5.17	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales			
unass-miner-msp_087	79.74	8.77	k__Bacteria	p__Actinobacteria	c__Actinobacteria	o__Actinomycetales	f__Microbacteriaceae		
unass-miner-msp_088	86.88	3.32	k__Bacteria	p__Proteobacteria	c__Betaproteobacteria	o__Methylophilales			
unass-miner-msp_093	38.9	1.72	k__Bacteria	p__Proteobacteria					
unass-miner-msp_096	43.6	1.08	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales			
unass-miner-msp_099	52.27	2.56	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales_2	f__Hyphomonadaceae		
unass-miner-msp_101	47.73	0.81	k__Bacteria	p__Actinobacteria	c__Actinobacteria	o__Actinomycetales			
unass-miner-msp_102	58.34	9.22	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria				
unass-miner-msp_105	79.63	5.77	k__Bacteria	p__Verrucomicrobia	c__Opitutae	o__Opitutales			
unass-miner-msp_107	50.87	4.5	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodobacterales	f__Hyphomonadaceae	g__Oceanicaulis	
unass-miner-msp_109	58.94	0.75	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales			
unass-miner-msp_110	52.59	3.75	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales			
unass-miner-msp_111	40.83	5.67	k__Bacteria	p__Cyanobacteria	c__Prochlorales	o__Prochlorales	f__Prochlorococcaceae	g__Prochlorococcus	s__Prochlorococcus_marinus
unass-miner-msp_112	33.65	0	k__Bacteria	p__Bacteroidetes	c__Cytophagia	o__Cytophagales	f__Cytophagaceae_2		
unass-miner-msp_119	48.64	0	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales			
unass-miner-msp_130	34.94	4.79	k__Bacteria						
unass-miner-msp_132	41.98	8.2	k__Bacteria	p__Proteobacteria					
unass-miner-msp_146	34.83	3.2	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Legionellales			
unass-miner-msp_155	33.65	1.75	k__Bacteria						
unass-mixo-msp_122	40.22	4.86	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Cryomorphaceae		
unass-mixo-msp_144	45.54	5.62	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales	f__Cryomorphaceae		
unass-mixo-msp_309	40.04	9.44	k__Bacteria	p__Chloroflexi	c__Dehalococcoidetes				
unass-mixo-msp_355	33.47	0.72	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria				
unass-mixo-msp_405	34.04	3.57	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae	g__Nisaea	
unass-mixo-msp_448	36.85	1.72	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria	o__Rhodospirillales	f__Rhodospirillaceae		
unass-mixo-msp_516	89.71	6.67	k__Bacteria	p__Proteobacteria	c__Betaproteobacteria	o__Methylophilales			
unass-mixo-msp_610	43.94	5.54	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia	o__Flavobacteriales			
unass-mixo-msp_654	45.7	6.91	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria	o__Alteromonadales_3	f__Alteromonadaceae		
unass-mixo-msp_677	42.22	7.25	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria				
unass-mixo-msp_679	37.74	5.62	k__Bacteria	p__Proteobacteria	c__Gammaproteobacteria				
unass-mixo-msp_71	36.32	5.13	k__Bacteria	p__Actinobacteria	c__Actinobacteria				

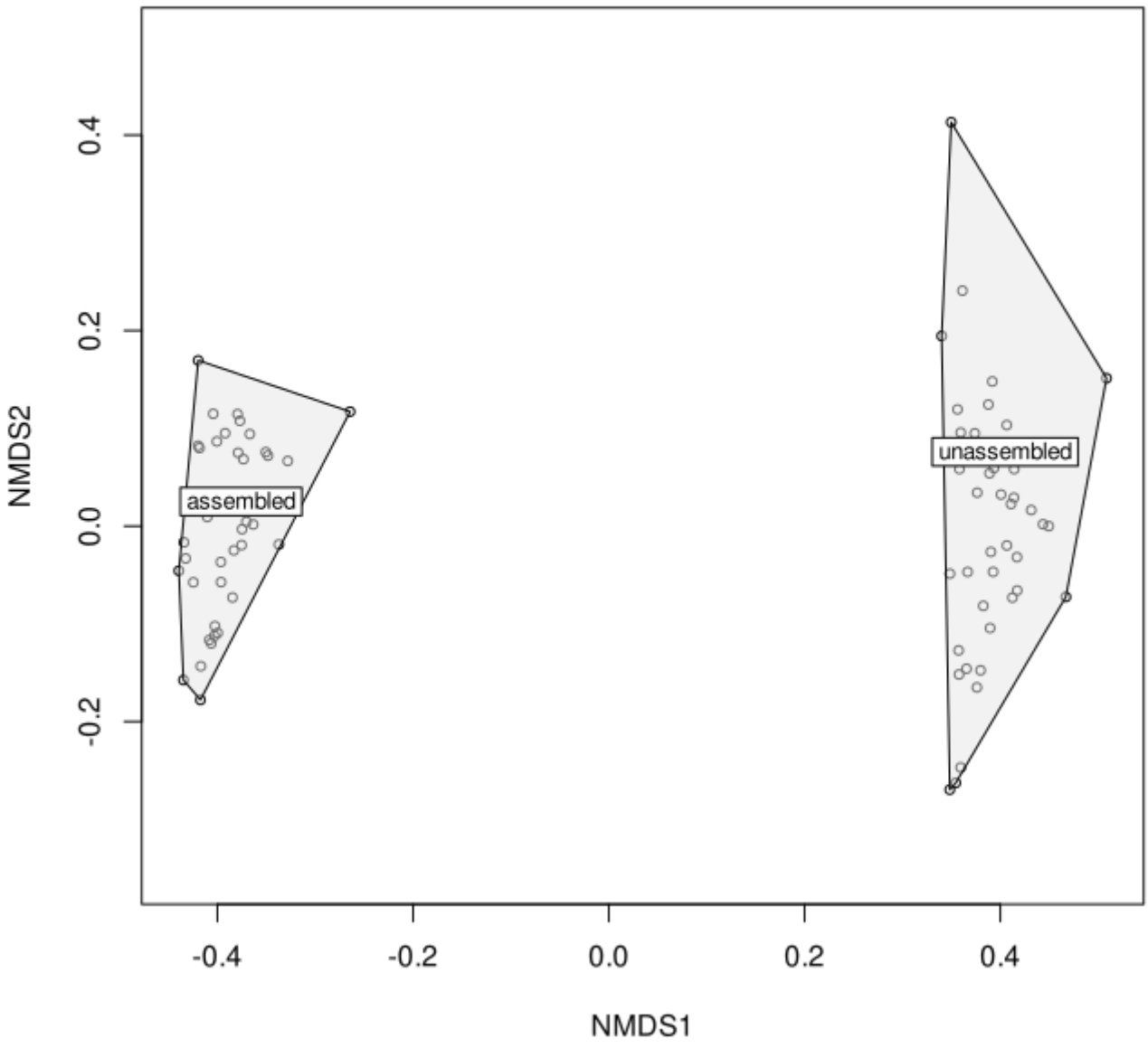


Fig S1. NMDS based on Bray Curtis dissimilarity computed from MetaFast separating assembled and unassembled microbial communities.

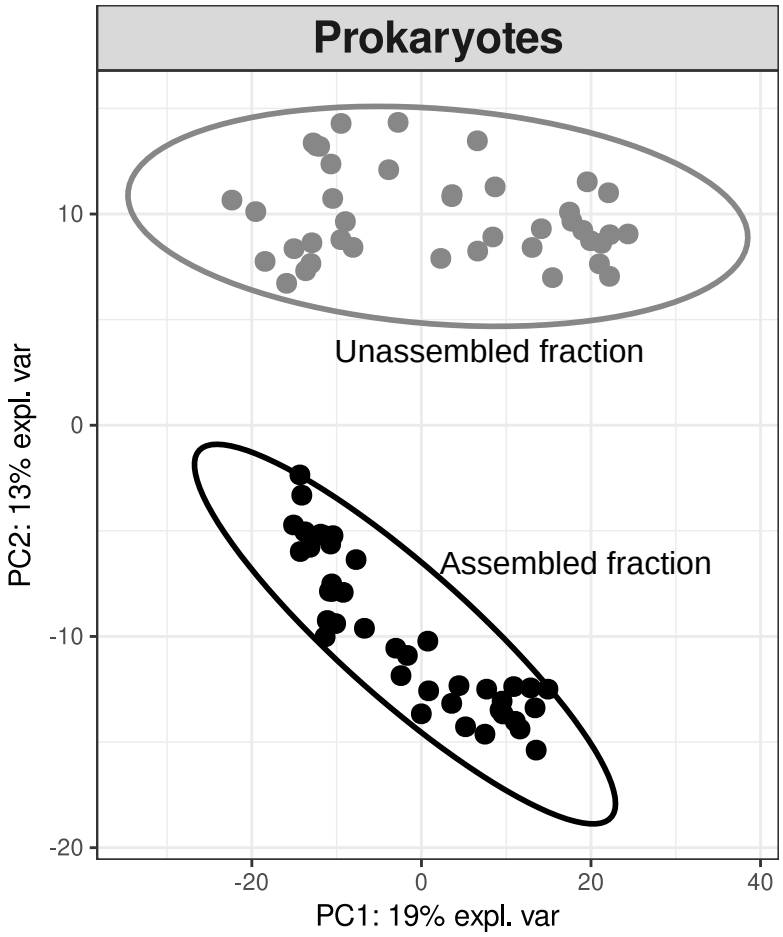


Fig S2. spca analysis of microbial communities based on 16S rRNA extracted from the assembled and unassembled fractions of the metagenomes.

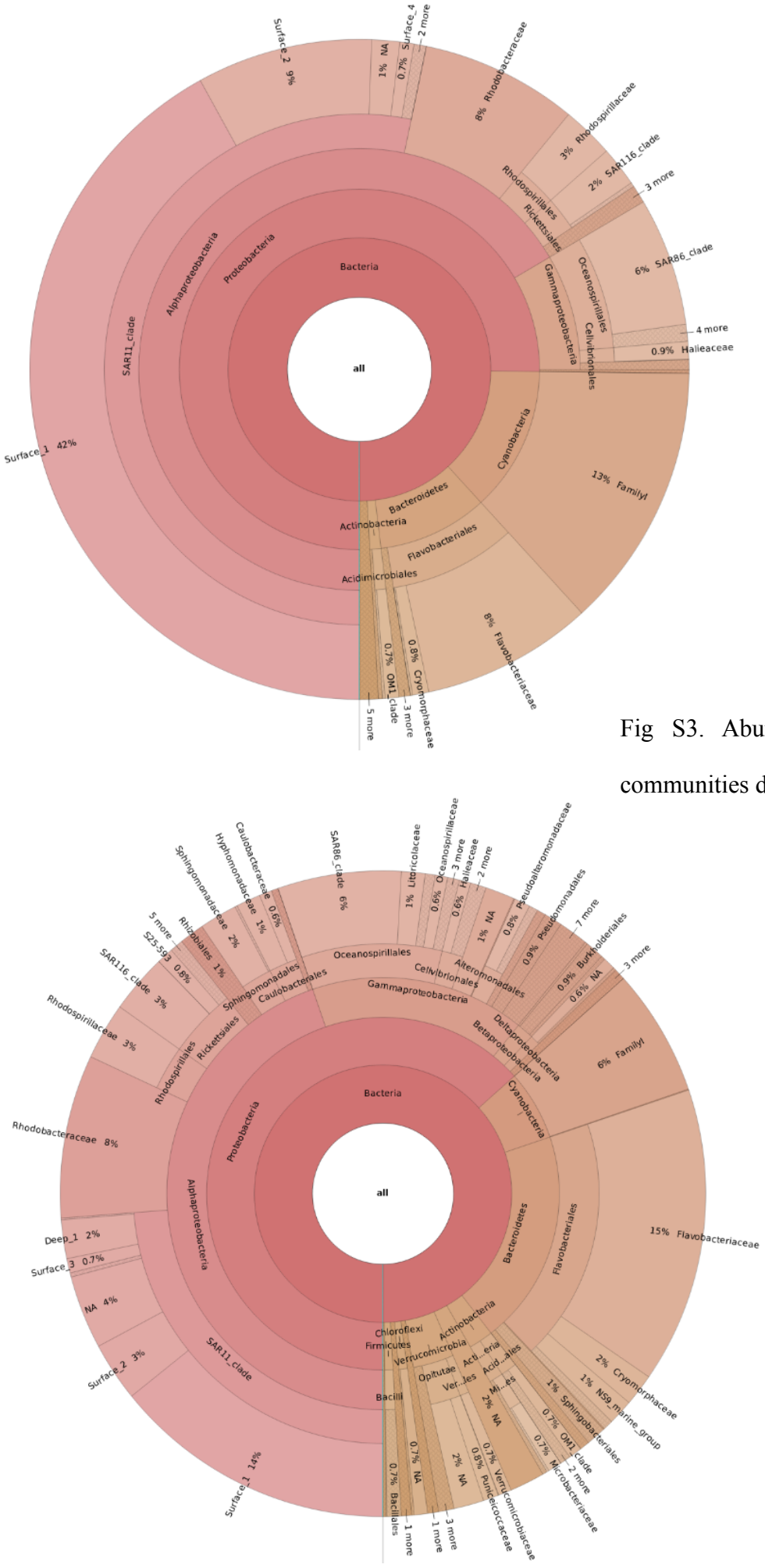


Fig S3. Abundant (top) and rare (bottom) communities deciphered by metabarcoding.

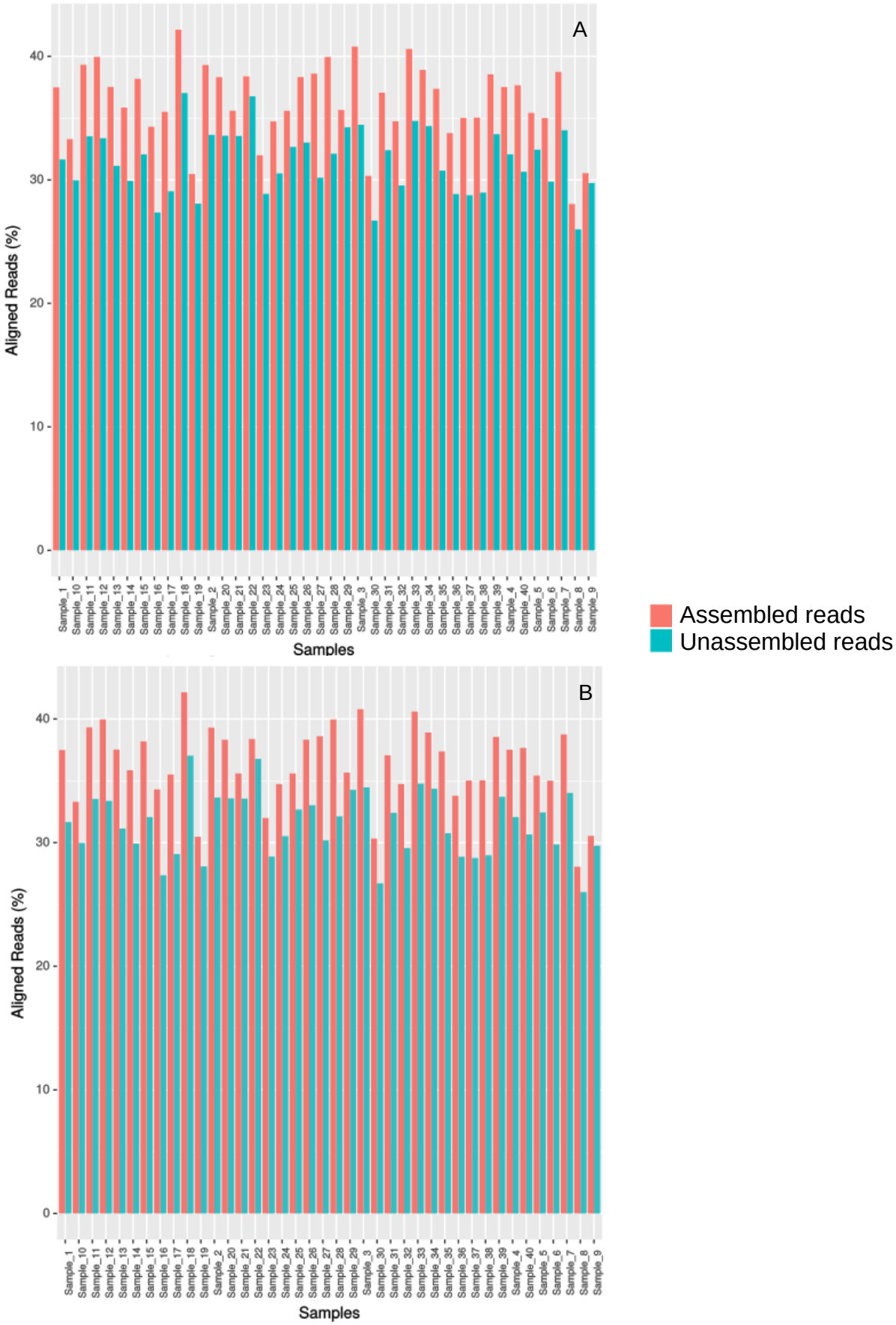
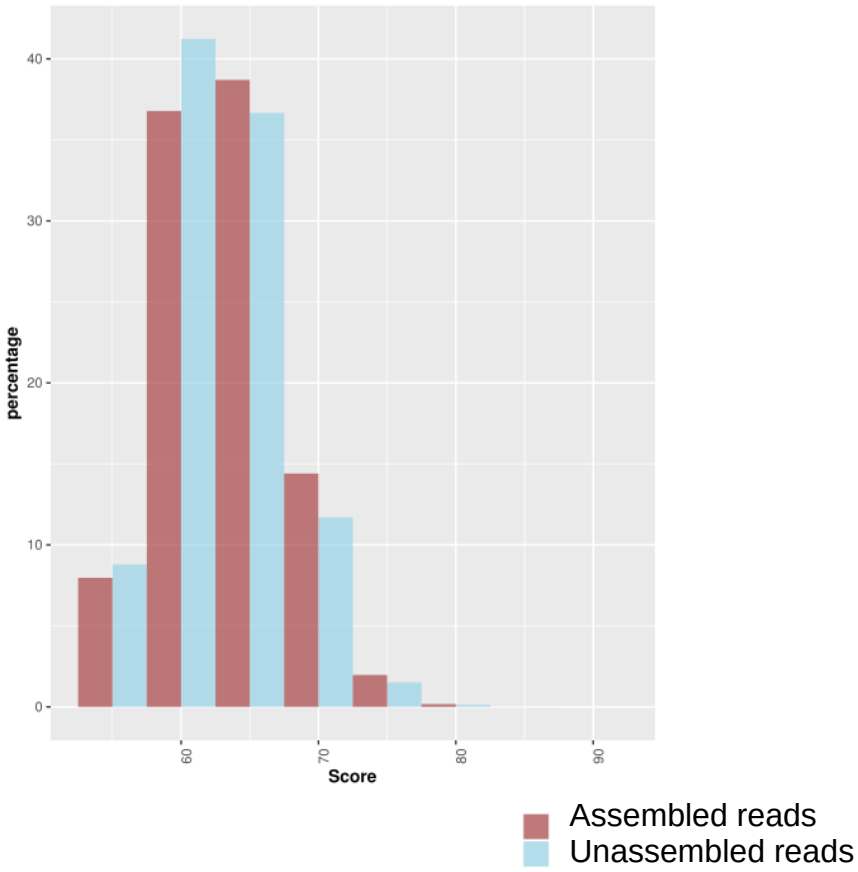


Fig S4. Distribution of the assembled and unassembled reads aligned to the UNIREF90 (A) and UNIREF100 (B) databases.

A) UNIREF90



B) UNIREF100

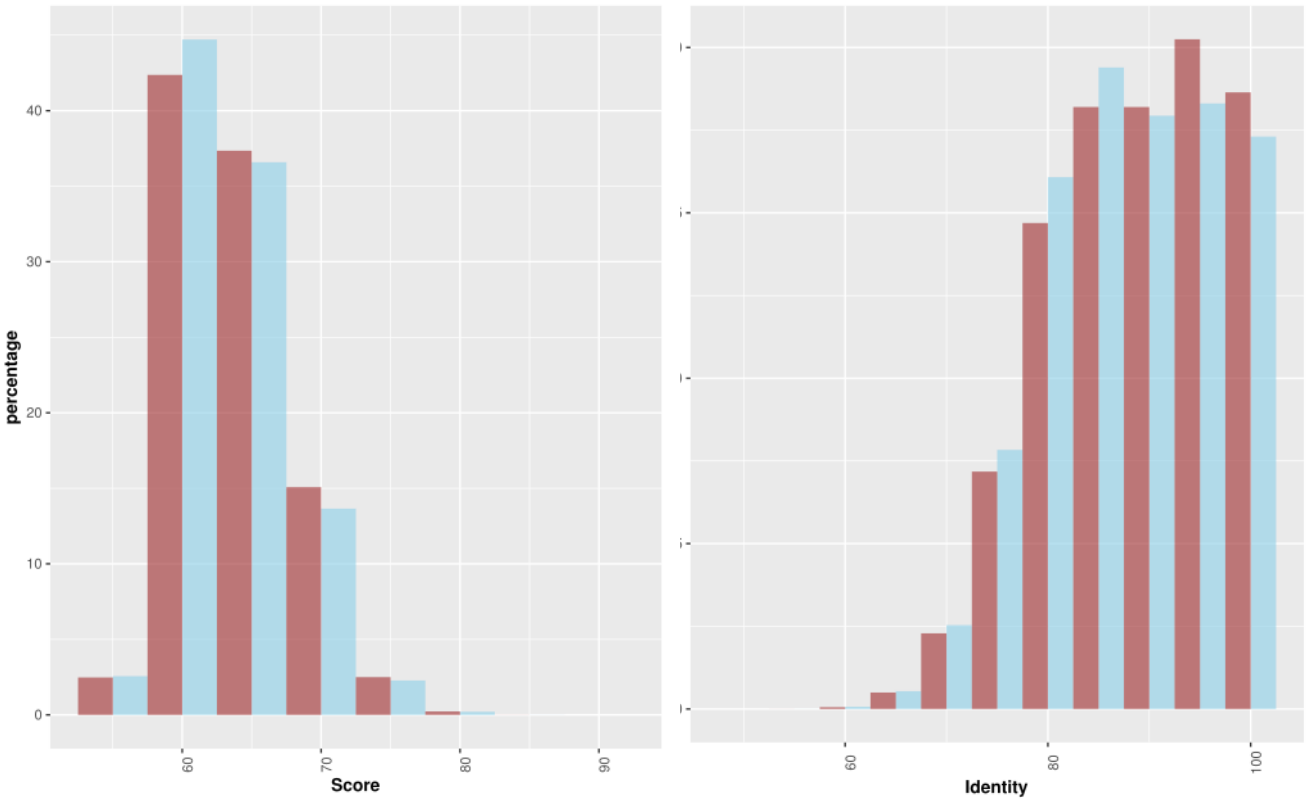
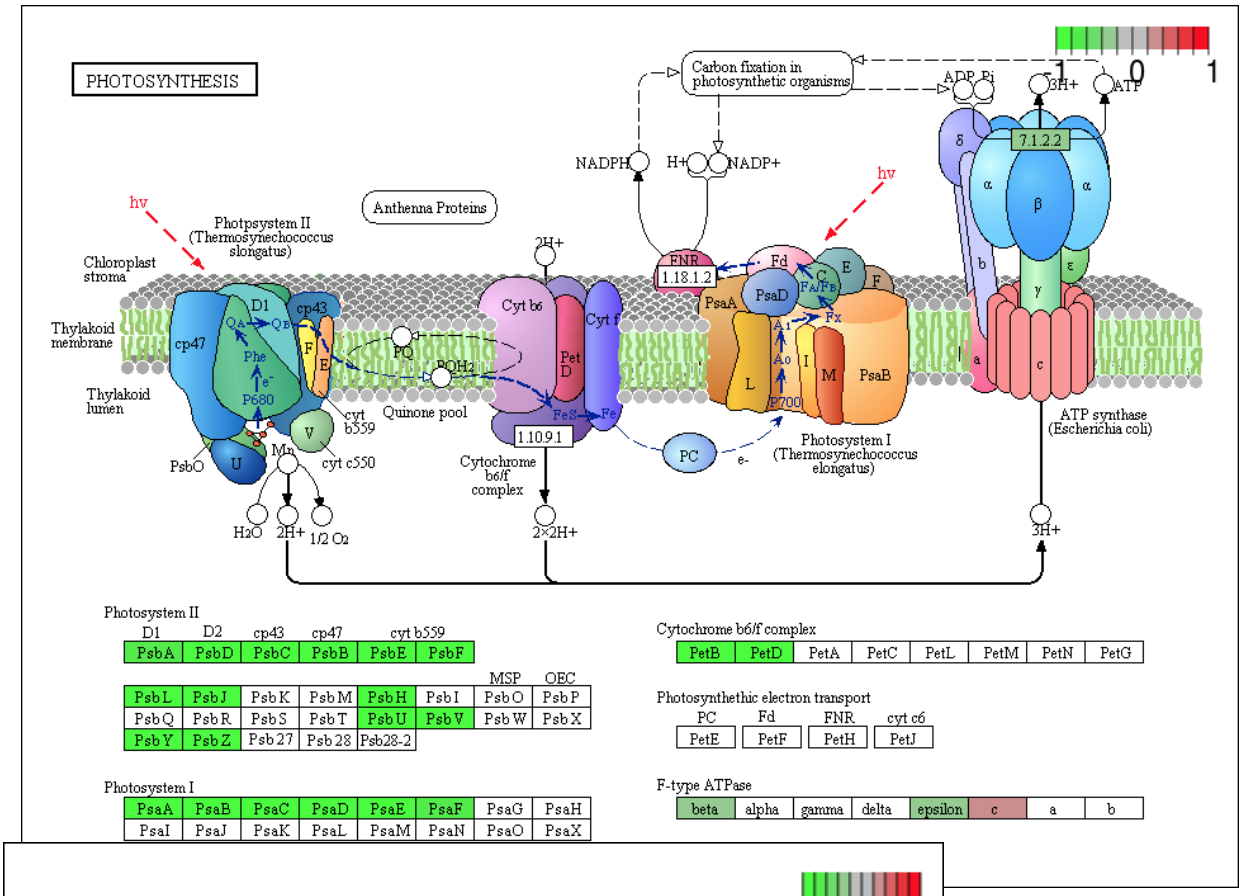


Fig S5. Distribution of the scores and identities of the assembled and unassembled reads aligned to the UNIREF90 (A) and UNIREF100 (B) databases.



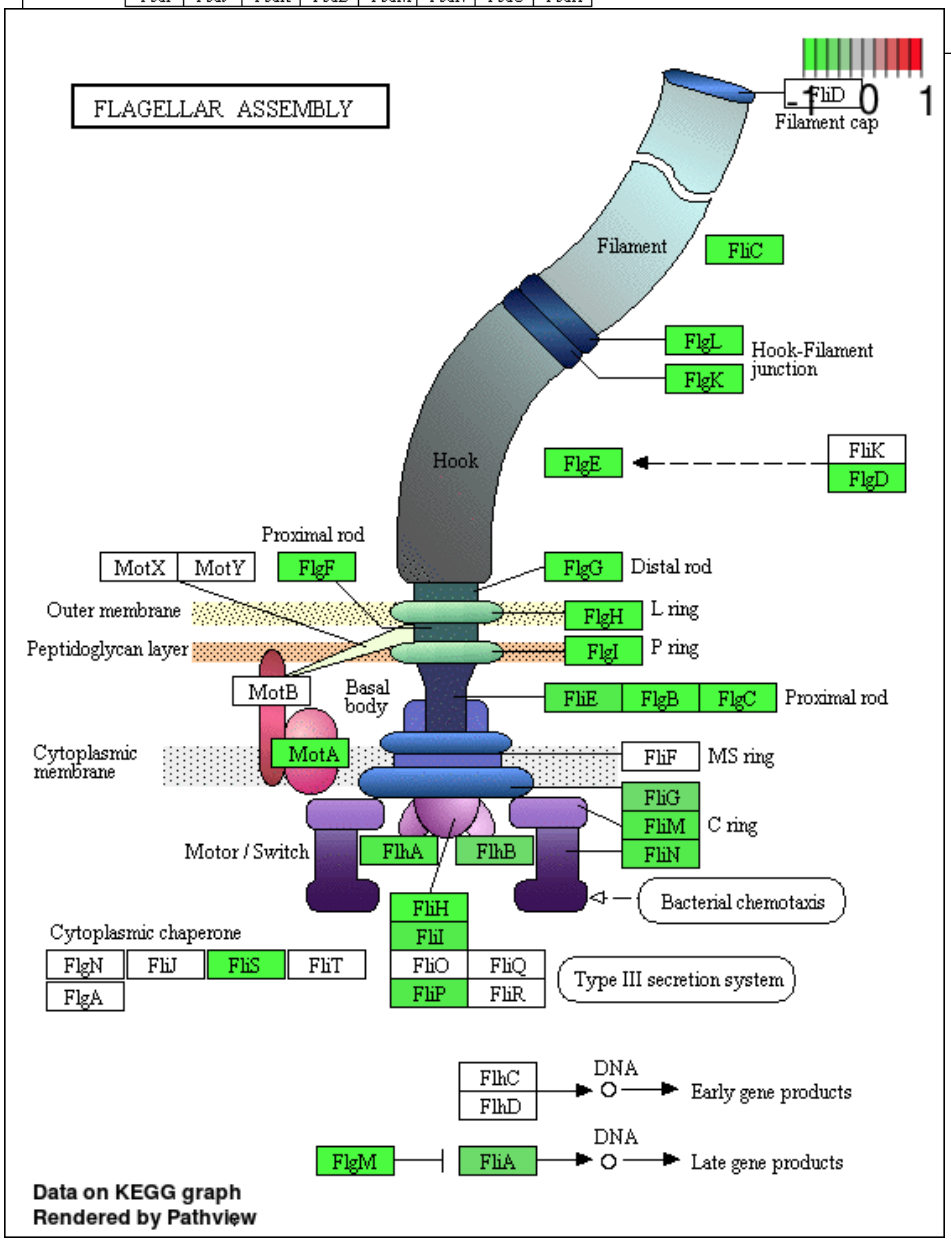
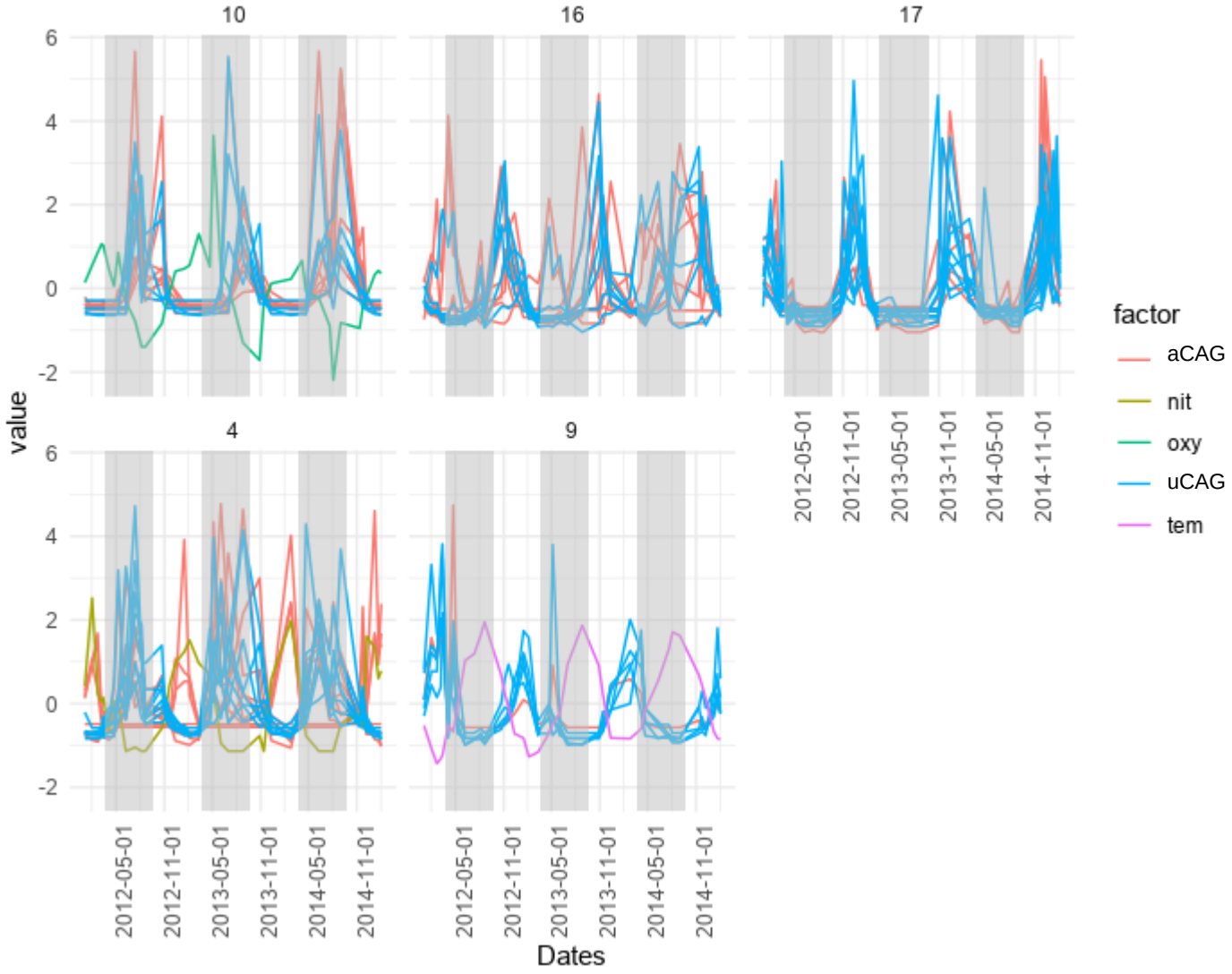
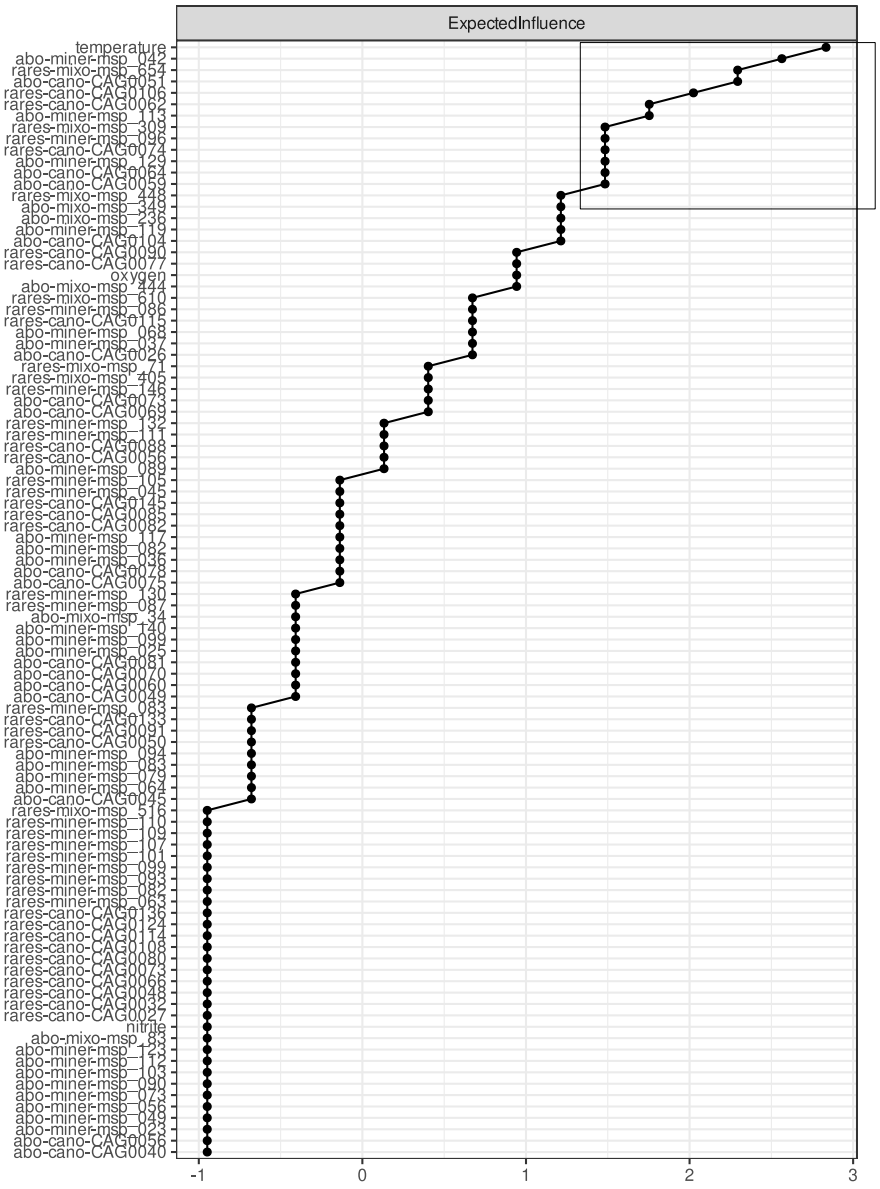


Fig S6. Photosynthesis and flagellar assembly pathways. Green rectangles corresponds to an enrichment in the assembled reads and red in the unassembled reads.



Pathways	KOs in pathways
Cluster 16	
ko00290 Valine, leucine and isoleucine biosynthesis	42%
ko00970 Aminoacyl-tRNA biosynthesis	32%
ko00780 Biotin metabolism	30%
Cluster 17	
ko00290 Valine, leucine and isoleucine biosynthesis	53%
ko00780 Biotin metabolism	39%
ko00970 Aminoacyl-tRNA biosynthesis	35%
ko00473 D-Alanine metabolism	33%
ko01230 Biosynthesis of amino acids	31%
ko00670 One carbon pool by folate	30%
ko00785 Lipoic acid metabolism	27%
ko00770 Pantothenate and CoA biosynthesis	26%
ko00195 Photosynthesis	25%
ko00860 Porphyrin and chlorophyll metabolism	25%

Fig S7 Temporal dynamics (z-scores) of the unassembled and assembled CAGs inside the main network clusters assessed by the Louvain methods. The clusters composed of less than 3 vertices are not represented. The grey rectangle represents spring and summer periods. The table displays the mains metabolic pathways in the clusters 16 and 17 (Any pathway with at least 25 % of the KOs were detected in the other clusters).



Best influencers/
keystone species

Fig S8. « ExpectedInfluence » parameter computed from the network with the package qgraph under R.