Ecological and functional capabilities of an uncultured 
*Kordia* sp.

Royo-Llonch M$^{1,4}$, Sánchez P$^1$, González JM$^2$, Pedrós-Alió C$^3$, Acinas SG$^*$

---

$^1$ Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM), CSIC, Barcelona, Spain.

$^2$ Department of Microbiology, University of La Laguna, La Laguna, Spain

$^3$ Systems Biology Program, Centro Nacional de Biotecnología (CNB), CSIC, Madrid, Spain

$^4$ Departament de Genètica i Microbiologia, Facultat de Biociències, Universitat Autònoma de Barcelona, Bellaterra, Spain

*Correspondence:*

Dr. Silvia G. Acinas
sacinas@icm.csic.es;

**Keywords:** SAG, *Kordia*, co-assembly, photoheterotrophy, metagenomics
Abstract

Cultivable bacteria are only a fraction of the diversity in microbial communities. However, the official procedures for classification and characterization of a novel prokaryotic species still rely on isolates. Thanks to Single Cell Genomics, it is possible to retrieve genomes from environmental samples by sequencing them individually, and to assign specific genes to a specific taxon, regardless of their ability to grow in culture. In this study, we performed a complete description of the uncultured *Kordia* sp. TARA_039_SRF, a proposed novel species within the genus *Kordia*, using culture-independent techniques. The type material was a high-quality draft genome (94.97% complete, 4.65% gene redundancy) co-assembled using 10 nearly identical Single Amplified Genomes (SAGs) from surface seawater in the North Indian Ocean from the *Tara* Oceans Expedition. The assembly process was optimized to obtain the best possible assembly metrics and a less fragmented genome. Its closest relative is *Kordia periserrulae*, sharing 97.56% similarity of the 16S rRNA gene, 75% of their orthologs and 89.13% average nucleotide identity. We describe the functional potential of the proposed novel species, that includes proteorhodopsin, the ability to incorporate nitrate, cytochrome oxidases with high affinity for oxygen and CAZymes that are unique features within the genus. Its abundance at different depths and size fractions was also evaluated together with its functional annotation, revealing that its putative ecological niche seems to be particles of phytoplanktonic origin. They can attach to these particles and consume them while sinking to the deeper and oxygen depleted layers of the North Indian Ocean.
Introduction

There is a lack of a consensus on what ecologically coherent units should be used as a proxy for bacterial species in environmental communities [23,52,67,72]. Moreover, for a new bacterial species to be officially recognized, its isolation in pure culture is still a requirement. It becomes necessary to re-define bacterial “species” to fit with the reality of the (still uncultured) majority of bacteria. An update of the validation of novel uncultured high-quality genomes is also needed, as they are given a provisional Candidatus status even after a thorough description [37]. Recent efforts have emerged to facilitate novel uncultured taxa standardization [37], like the Microbial Genome Atlas (MiGA), which infers genome-based taxonomy and quality assessment across genomes from different environments [66].

Nowadays, uncultured genomes can be used to expand knowledge of the genetic and evolutionary differences between representatives of the same species or close relatives, as well as for enriching knowledge of microbial diversity. Uncultured genomes can be retrieved by: i) co-assembling metagenomes (Metagenomic Assembled Genomes or MAGs) or ii) by single cell genomics (Single Amplified Genomes or SAGs). In marine microbiology, assembling MAGs has brought a vast amount of genomes belonging to novel bacterial and archaeal phyla, unveiling key players in the biogeochemistry of the oceans [18,56,57]. Alternatively, Single Cell Genomics allows the assignment of specific functional traits to a specific taxon, an outstanding resolution level for microdiversity studies. Single Cell Genomics has helped link functional roles to relevant taxonomic groups that have changed the current understanding of predominant marine metabolisms (such as chemolithoautotrophy or the role of nitrite-oxidizing bacteria in the deep ocean) [53,77]. While MAGs tend to be longer and more complete than SAGs, MAGs result from a population of genomes and there is still lack of consensus about what taxonomic units are they reflecting. Nevertheless, multiple SAGs can be co-assembled to retrieve more complete genomes, provided they are closely related
phylogenetically. The comparison between MAGs and SAGs assembled from the same Baltic Sea water samples, revealed very high nucleotide identities between the corresponding pairs but a difference in the size and completeness between the genomes [4]. SAGs have also been used in pangenome analyses, which had previously been mostly restricted to cultured microorganisms (especially pathogenic bacteria) [49,51,79]. Comparative genomics and the development of the “pan-genome” concept offered an alternative for understanding the genetic extent and dynamics of bacterial species [49,51,79], dramatically changing the description of a species by studying multiple genomes belonging to the same defined taxa [34,38,65]. In the marine environment, SAG-based pangenome analyses have mostly focused on the highly studied Prochlorococcus (revealing hundreds of co-existing Prochlorococcus populations [31] and the link between their hypervariable genomic islands and the environment [17]) or together with SAR11 (defining endemic gene-level adaptations to specific locations like the Red sea [80]).

In the present study, we carried out the co-assembly of 10 SAGs of a nearly clonal population of a novel species in the genus Kordia to generate a high-quality genome complete enough to: i) allow its putative functional description, ii) determine its preferred niche in the water column from which it was retrieved, and iii) to describe the novel species in comparison with the available sequenced relatives of the genus Kordia. The genus Kordia belongs to the family Flavobacteriaceae and it was first proposed with the isolation in culture of Kordia algicida, which lyses algal cells and feeds on phytoplanktonic bloom exudates [75]. Members of this genus are Gram-negative, strictly aerobic or facultatively anaerobic, non-motile or motile by gliding, rod shaped and showing 34-37% of DNA G+C content [33]. In the last 15 years, new species have been added to the genus, all of them isolated from samples found in the aquatic environment: K. zhangzhouensis thrives in freshwater [41], K. jejudonensis was isolated from the interphase between seawater and freshwater springs [54], K. antarctica and K. aquimarins
were isolated from Antarctic and Taiwanese seawater, respectively [6,26], *K. ulvae*, *K. zosterae* and *Kordia* sp. SMS9 were retrieved from the surface of marine algae [33,59,62] and *K. periserrulae* was found in the gut of a tidal flat polychaete [14]. *Kordia* sp. NORP58 is a MAG assembled from a marine subsurface aquifer [81]. At the time of writing, the genus consists of eight species with validly published names, of which four have had their genomes completely sequenced.

Despite the fundamental insights in microbial ecology and evolution provided by the analyses of uncultured bacterial and archaeal genomes, there is not yet a proper taxonomy for the uncultured majority. We aim to provide a good example of a complete description of a novel species of the genus *Kordia* analyzed via culture-independent methods to infer its ecological and functional description.
Materials and Methods

Single Amplified Genome generation and phylogeny analysis

A total of 98 Single Amplified Genomes were generated as detailed in [77] from a surface (SRF) seawater sample from the North Indian Ocean (latitude 18.59°N – longitude 66.62°E) during the circumnavigation expedition Tara Oceans [30], station TARA_039 (Sample ID: TARA_G000000266)(SM Table 1).

Multi Locus Sequencing Analysis (MLSA) of the generated SAGs revealed that 84% of them belonged to a novel species of the genus Kordia (referred hereafter as Kordia sp. TARA_039_SRF) and that the amplified genes (16S rRNA, partial 23S rRNA, partial RNA polymerase subunit B and partial proteorhodopsin genes, as well as Internally Transcribed Spacer) were identical in the whole Kordia SAG population. More details on the phylogeny and distribution of these Kordia SAGs can be found in [69].

SAG selection

We chose 10 out of the 98 generated Kordia SAGs for Whole Genome Sequencing. They were selected based on: i) Multiple Displacement Amplification’s (MDA) Cp values, since they were the shortest, ranging from 7:15-8:15, and ii) that amplification was possible for any of the markers tested in previous MLSA.

Sequencing and read treatment

Sequencing of the ten SAGs was carried out in two batches. For the first set of five, the sequence reads were obtained by Illumina MiSeq 2x300 bp technology and by Illumina HiSeq 2x250 bp for the second set. Quality assessment of the raw reads was performed using FastQC and Illumina PhiX174 adapters were removed using Bowtie2 v2.2.9 [43] and Samtools v.1.3.1 [46]. Afterwards, reads were normalized by coverage for each individual SAG using DOE JGI’s BBnorm, setting the maximum coverage at 40X. The normalized paired-end libraries were trimmed and filtered with Trimmomatic [10] and paired-reads were merged with PEAR v0.9.6 [85]. The workable read dataset consisted
of merged reads, pairs of reads that did not merge and also unpaired orphan reads (Figure 1A). The final non-redundant set accounted for 1.75% of the original raw read dataset (3,746,468 reads out of a total of 213,457,114).

Co-assembly optimization and quality control

We chose a strategy that focused on reducing gene redundancy and genome fragmentation to obtain the co-assembly. As a first approach, 36 co-assemblies were generated with assembler Ray v2.2.0 using individual k-mer length values, from 17 to 133, in order to find those that performed better (Figure 1A). The three k-mer values that generated co-assemblies with best metrics were combined using assembler SPAdes v3.10 [7] and options --careful and --sc (recommended for single cell genome assembly). Another co-assembly was gathered with SPAdes default combination of k-mer length values (21, 33, 55) to confirm whether the custom k-mer length combination resulted in a better co-assembly.

Normalizing the pool of reads by coverage and merging them before assembly reduced significantly the memory requirements and duration of assembly. Using default single-cell mode SPAdes assembler with the full read dataset, the co-assembly lasted ~62 hours and consumed 75 Gb of memory. The same process with the workable read dataset (normalized and merged) and optimized k-mer lengths took 11 min and 12 Gb.

Metrics for each co-assembly were generated using a custom Perl script. Genome completeness and contamination (assessed as single copy marker gene redundancy) was calculated with CheckM v1.0.11 [55]. The reference marker gene database chosen was “Family Flavobacteriaceae”, which was the closest to the taxonomical annotation of the co-assembly. Nevertheless, the marker gene PF07659.6 was excluded from the database as it was absent in all complete isolated Kordia genomes. The assembly accuracy was assessed with Assembly Likelihood Estimator ALE [15]. The three k-mer length values that showed better metrics (higher genome completeness, longer co-assembly length, smaller number of contigs and larger N50) and best ALE score were
used as a combination in a final co-assembly (K=55, 61, 117). Overlapping regions among scaffolds were detected after a visual check of gene synteny in their functional annotations. Several refining steps were taken to solve redundancies in the co-assembly (Figure 1B). First, the scaffolds were assembled in Geneious R11, with default settings in the High Sensitivity assembly algorithm. Second, the resulting contigs were split in different datasets by setting a cut-off for a minimum contig length. The final co-assembly was the contig dataset with the highest genome completeness, highest number of copy genes present only once and percent contamination lower than 5% [11,37]. Detailed metrics and evaluation of the different co-assemblies are in SM Table 2.

Alignment of individual SAGs against the co-assembly

Mappings of individual SAGs to the co-assembly were done with the same read dataset as for the co-assembly (clean, trimmed, merged and normalized by coverage reads) using Bowtie2 v2.2.9. The percentage of mapped reads was extracted from Bowtie2’s log file. The vertical coverage of the co-assembly by each SAG individually was calculated using Bedtools v2.27.1 genomecov function. The sum of bases of the co-assembly covered by each SAG was divided by the total length of the co-assembly to obtain the contribution of each SAG to the co-assembly (Table 1).

Ggplot2 [24] was used to visualize the mapping of reads of every individual SAG against the generated co-assembly (Figure SM2).

Phylogeny based on the 16S rRNA gene

The complete 16S rRNA gene of the co-assembly was queried against the Living Tree Project [83] database using SILVA’s SINA ARB v1.2.11 online tool [61]. Its Flavobacteriaceae best hits were exported and aligned in Geneious R11, together with Kordia amplicon OTUs generated from the Malaspina dataset in [50], the 16S rRNA genes of other Kordia spp. deposited in GenBank and two outgroups from a different phylum (SM Table 1). MEGA v7 [40] was used for a Maximum-Likelihood phylogeny
reconstruction using GTR+G model and 500 bootstrap replicates. The tree was later processed in iTol [45].

Phylogeny based on single copy genes

A multi-locus (40 conserved single copy genes) phylogenetic placement of the co-assembly was also carried out using the same taxa as in the 16S rRNA gene phylogeny, also including uncultured genomes like MAGs that belong to family Flavobacteriaceae. Gene prediction was done with Prodigal v2.6.3 [27]. Software FetchMG v1.0 [39] was used with option -v (recommended for reference genomes) to extract the 40 conserved COGs, whose amino acid sequences were concatenated and later aligned with Muscle v3.8.31 [20]. Neighbor-Joining phylogenetic reconstruction was done with MEGA v7.0 and the resulting tree was processed in iTol.

Sequence composition identities against other Flavobacteriaceae

Average Nucleotide Identities (ANI), Average Amino acid Identities (AAI) and tetranucleotide frequency comparisons were done between the co-assembly and the genomes of its closest relatives using fastANI v1.1 [28], compareM v0.0.23 [https://github.com/dparks1134/CompareM] and pyani v0.2.7 [https://github.com/widdowquinn/pyani], respectively.

Functional annotation of Kordia sp. TARA_039_SRF

Gene prediction and a basic annotation of the co-assembled contigs were carried out in Prokka v1.12 [70]. PFAMs and TIGRFAMs were annotated in the predicted genes with HMMER’s v3.1b2 hmmscan (hmmer.org). KEGG orthologs (KO) were predicted online with BLASTKOALA v2.1 [29] and transporters were predicted using Transporter Classification TC database [21]. Carbohydrate-active enzymes’ annotation was based on dbCAN’s CAZymes database [84] and peptidases were annotated using the Merops database [63]. Polysaccharide Utilization Loci (PULs) were manually located looking for pairs of SusC/SusD genes encoded next to each other. Genomic Islands were predicted
with IslandViewer v4 online tool [9]. Insertion sequences (IS) were predicted and estimated with ISSaga tool [82] and prophage regions were predicted with PHASTER [5]. Secretion systems were detected using MacSyFinder-based TXSScan [1,2]. KEGGmapper Search&Color pathway v3.1 was used to describe the functional potential of the co-assembly, for those gene calls that could be associated to a KO value.

**Phylogeny based on the proteorhodopsin gene**

The complete amino acid sequence of the co-assembly’s proteorhodopsin gene was aligned in Geneious R11 together with its BLAST best hits against NCBI’s Protein database and two outgroup sequences (a proteorhodopsin from a SAR86 group bacterium and one from a “Candidatus Pelagibacter ubique” bacterium). MEGA v7 [40] was used for a Maximum-Likelihood phylogeny reconstruction using LG+G model and 500 bootstrap replicates. The tree was later processed in iTol [45].

**Distribution of Kordia sp. TARA_039_SRF in TARA_039 station with competitive Fragment Recruitment Analysis**

The vertical and size fraction distribution of Kordia sp. TARA_039_SRF was assessed in all available metagenomes for Tara Oceans station TARA_039 as it was the marine sample from which the Kordia SAGs were obtained. Metagenomes from TARA_039 were sampled in three different depths (surface/SRF 5m, Deep Chlorophyll Maximum/DCM 25m and mesopelagic/MES 270m) and size-fractioned for viruses (<0.22 µm), giruses (0.1-0.22 µm), bacteria (0.22-1.6 µm) and protists (0.8-5 µm, 5-20 µm, 20-180 µm and 180-2000 µm) and were generated following different protocols (SM Table 1)[3]. All merged reads (PEAR v0.9.6) from each described metagenome were mapped against all genomes available from Kordia spp. (Kordia sp. TARA_039_SRF, Kordia algicida, Kordia jeudonensis, Kordia zhangzhouensis, Kordia periserrulae, Kordia sp. SMS9 and Kordia sp. NORP58) (SM Table 1) with blastn (BLAST 2.2.8+; options -max_target_seqs 1 –perc_identity 70 –evalue 0.0001). The output was filtered in R: i) coverage between query and subject set at >90%, ii) removal of duplicated reads
and iii) removal of reads mapping to the ribosomal operons [12]. Recruited reads were split based on their nucleotide identities against the reference genome. Those mapping at identities >95% are assumed to belong to the same species as the reference genome and those between 70-95% belonged to the co-occurrent close relatives of the reference genome [12,65]. Normalization of the recruited reads was done by the sequencing depth of each metagenome and considering the complete genomic length.

**Comparative genomics analysis**

Anvi’o v4 pangenomic workflow was used to organize all the genes from the five *Kordia* genomes (*Kordia* sp. TARA_039_SRF, *Kordia algicida*, *Kordia jejudonensis*, *Kordia zhangzhouensis* and *Kordia periserrulae*) into gene clusters of protein sequence similarity. Gene calling was done with Prodigal, amino acid sequence search with NCBI’s blastp v2.7.1, gene clustering with MCL [19] and sequence alignment with Muscle. Using function --anvi-summarize we could export the list of gene clusters and see in which genome they were encoded. These results were plotted using anvi’o and R packages UpSetR v1.1.3 [16] and ggplot2.

Anvi’o phylogenomics workflow was used to produce a phylogenomic tree (Fasttree v2.1.11 [60]) based on the amino acid sequences of the single copy genes shared by all five *Kordia* genomes.

**Accession numbers**

The genome of the novel *Kordia* sp. TARA_039_SRF has been deposited in NCBI’s Bioproject PRJNA524487. The co-assembly’s accession number is SMNH02000000 and Biosample SAMN11028936. Raw reads of the 10 individual SAGs are: AB-193_M23 (SRR8655105), AB-193_M19 (SRR8655106), AB-193_N20 (SRR8655104), AB-193_O13 (SRR8655103), AB-194_L04 (SRR8655108), AB-193_M03 (SRR8655107), AB-193_O22 (SRR8655102), AB-193_I20 (SRR8655109), AB-193_I04 (SRR8655110)
and AB-193_P22 (SRR8655101). A complete list of accession numbers for all the genomes and metagenomes used in this study can be found in SM Table 1.
Results

Co-assembly of the *Kordia* sp. TARA_039_SRF genome

Ten SAGs belonging to the genus *Kordia*, named provisionally *Kordia* sp. TARA_039_SRF, were selected for genome co-assembly and description of gene content. These SAGs were tested for microdiversity through MLSA analyses, resulting in 100% identity at the nucleotide level in each marker [69]. The five markers were the 16S rRNA gene, partial 23S rRNA gene and the Internally Transcribed Spacer, that could be amplified in the 10 SAGs; the partial RNA polymerase subunit B, amplified in 5 SAGs; and partial Proteorhodopsin, amplified in 7 SAGs.

Individual SAG assemblies varied in genome completeness and metrics (Table 1), recovering at most 45% of the estimated complete genome. GC content was consistent in the different SAGs (35.3-36%).

Considering the high genetic similarity among the individual SAGs based on the previous MLSA analyses, their co-assembly into one single genome was the strategy chosen to obtain a more complete genome. All the reads from the 10 SAGs were pooled together, normalized by coverage and merged. They were afterwards assembled using a wide range of k-mer length values, a procedure to find the best metrics and quality values (SM Table 2). The combination of k-mer length values chosen were: 55, 61 and 117 (Figure 1B). After refinement of the optimized co-assembly, the resulting genome was 4,594,716 bp long, fragmented into 27 contigs of N50 429,544 bp. The genome was 94.97% complete using a database of Flavobacteriaceae conserved single copy genes and the contamination based on gene redundancy of these conserved genes was 4.65% (Table 1).

There were 985 ambiguities spread across the final co-assembly (217 Y, 247 W, 92 S, 158 R, 32 N, 140 M and 99 K), representing 0.02% of the total nucleotides.

The proportion of reads that mapped back to the refined co-assembly varied between 97.6 and 98.1% depending on the SAG. Contribution of each SAG to the co-assembly
ranged between 21 and 50% (Table 1). The average nucleotide identities of the individual SAGs against the co-assembly were higher than 99.9% in all cases (Table 1). Tetranucleotide frequency was almost identical between each SAG and the co-assembly, with similarities ranging between 99.6 and 99.8%.

Comparing the completeness and gene redundancy of all the possible SAG combinations to co-assemble (1023 for 10 SAGs), the highest completeness value (97.48%) is reached with the co-assembly of all 10 SAGs, with 8.48% contamination (SM Figure 1; SM Table 3). Nevertheless, there is saturation in completeness values ~94-95% starting at combinations of seven and eight SAGs. Their contamination values were also higher than the standards for high-quality draft genomes (>5%) with the exception of a combination of eight SAGs, that produced a co-assembly 94.07% complete with 4.92% of contamination. Considering that genome amplification in SAGs is random [76], it is impossible to know which part of the genome was amplified prior to sequencing. Despite the fact that we sequenced 10 SAGs, the co-assembly of just eight of them would have been enough to reach completeness values similar to the refined co-assembly.

Novelty of Kordia sp. TARA_039_SRF

The novelty of Kordia sp. TARA_039_SRF was confirmed with phylogenies of the 16S rRNA gene and 40 conserved single copy genes (Figure 2), as well as through whole genome sequence composition comparisons of ANI, AAI and tetranucleotide signature (SM Table 4).

Both phylogenies placed Kordia sp. TARA_039_SRF within family Flavobacteriaceae, more accurately within genus Kordia, in 100% of the bootstrap replicates. Its closest described relative is Kordia periserrulae, with 97.56% of identity in the 16S rRNA gene (97% coverage) and 89.13% ANI (75% coverage). Otu5835, amplified from Malaspina expedition seawater samples, also falls within the Kordia sp. TARA_039_SRF and K. periserrulae clade.
Functional description of *Kordia* sp. TARA_039_SRF

Having an almost complete genome (94.97%) allowed to investigate the potential functional capabilities and energy metabolism of this novel species (*Figure 3; SM Table 5*). Nevertheless, 2357 of 4125 predicted protein coding genes (57.1%) were annotated as hypothetical proteins of unknown function.

Central metabolism

The co-assembled *Kordia* sp. TARA_039_SRF has the ability to shut down the loss of CO₂ of the regular TCA cycle by the glyoxylate shunt, like other *Kordia* spp., and it is also able to replenish the TCA cycle of certain intermediates through anaplerotic routes: it encodes both PEP carboxykynase and PEP carboxylase, which convert PEP to oxaloacetate using ATP and CO₂, and H₂O and HCO₃⁻, respectively. This mechanism is also found in other *Kordia* spp. Bicarbonate uptake can be achieved through specific bicarbonate membrane transporters (BicA), of which four copies were found. One of these copies is next to a carbonic anhydrase gene, whose product interconverts CO₂ and HCO₃⁻. *Kordia* sp. TARA_039_SRF encodes a third anaplerotic strategy driven by the malic enzyme, which converts pyruvate into L-malate with NADPH and CO₂. One of the copies is encoded in the core genome of *Kordia* whereas another copy is located in a genomic island.

Unlike other *Kordia* spp., *Kordia* sp. TARA_039_SRF encodes several enzymes that allow a complete biosynthesis of CoA from pyruvate. This cofactor plays a key role in the biosynthesis and breakdown of fatty acids, as well as in the biosynthesis of polyketides and non-ribosomal peptides [8].

*Kordia* sp. TARA_039_SRF shows several mechanisms for oxidative phosphorylation, of which some are also common in other *Kordia* spp.: i) succinate dehydrogenase / fumarate reductase, ii) cytochrome c oxidase, cbb3-type, iii) NADH dehydrogenase NQR, iv) NADH dehydrogenase NDH-I, v) cytochrome bd subunits CydA and CydB,; and
some are unique: i) cytochrome c oxidase aa3-type genes CoxABC, QoxA, COX10 and COX11, and ii) cytochrome c oxidase bo-type genes CyoDW.

**Carbohydrate-active enzymes, PULs, surface adhesion and motility**

*Kordia* sp. TARA_039_SRF encodes a CAZymes array most similar to *K. periserrulae* and larger than the rest of *Kordia* spp. included in the study (SM Table 6). It shows the second highest density of carbohydrate-active enzymes (41 enzymes per genomic Mbp) after that of *K. zhangzhouensis*. It encodes 34 non-catalytic carbohydrate-binding modules (CBM), mostly specific for complex carbohydrates like cellulose, chitin, mannan, beta-glucans, starch, and glycogen. The novel species also encodes 1.1 glycosyltransferases per genomic Mb (a total of 52). These outer membrane proteins are involved in the synthesis of surface adhesion polysaccharides. A total of 51 predicted adhesion-related genes, such as fasciclin, fibronectin, or lectins (SM Table 7) were found, that are also present in other Flavobacteria [25]. The novel *Kordia* sp. TARA_039_SRF does not encode a complete gliding motility complex.

The co-assembly encodes four Polysaccharide Utilization Loci (PULs) (SM Table 5), three of them being surrounded by 6 families of glycosyl hydrolases, one family of glycosyl transferases, one family of carbohydrate esterases, two families of carbohydrate-binding modules and peptidases like serine aminopeptidase and prolyl oligopeptides.

**Sulfur and nitrogen metabolism**

Another unique feature of *Kordia* sp. TARA_039_SRF within its genus is that it codes for all the genes involved in the assimilatory sulfate reduction pathway that converts sulfate to sulfide, and it is also able to transform thiosulfate to sulfite.

The genome of *Kordia* sp. TARA_039_SRF is the only *Kordia* genome sequenced so far that encodes both NasA, a nitrate/nitrite transporter, and NasF, the substrate binding protein in the nitrate/nitrite transport system. The bacterium is predicted to assimilate
nitrate for further incorporations into amino acids as it encodes the nitrate reductase/nitrite oxidoreductase. Moreover, like other *Kordia* spp., it has the potential to use environmental ammonium as a nitrogen source importing it through an ammonium channel transporter.

**Light sensing and environmental information processing**

The co-assembled *Kordia* encodes a copy of a green-light absorbing proteorhodopsin, like *Kordia periserrulae* and *Kordia* sp. SMS9. Their phylogenetic reconstruction clusters the co-assembly’s and *K.periserrulae*’s proteins in the same clade with a 100% bootstrap value (*SM Figure 3*), and their pairwise alignment shows that they are identical in 95.5% of the sites (232 of 243 bp).

It also codes for other putative light sensors usually found in proteorhodopsin containing marine bacteria: one (6-4) photolyase, seven copies of phytochrome-like proteins, one cryptochrome DASH domain and one cryptochrome-like protein.

The co-assembly codes for DNA repair systems like the entire nucleotide excision repair (NER) complex UvrABC, which recognizes and removes light induced DNA lesions.

The two-component systems encoded in the co-assembly have been described to detect the following environmental signals: phosphate, ferric ion, oxygen, nitrate/nitrite and chemotaxis related attractant/repellent substances. It also codes for sigma 54 factor and putative quorum-quenching lactonases, which have been related in some cases to virulence and biofilm formation.

**Transporters**

The co-assembly encodes 95 different transporter families and ~88 transporters per Mbp. We find 4 different types of pore-forming toxins, phage-related transporters like the FadL outer membrane protein and a holin belonging to the *Bacillus subtilis* phage φ29 transporter family, two different light-absorption driven transporters and uptake systems specific for amino acids, potassium, fatty acids, nitrate, dissolved inorganic carbon, iron and magnesium (*Figure 3; SM Table 8*).
Kordia sp. TARA_039_SRF codes for two complete secretory systems: T1SS and the Bacteroidetes specific T9SS (SM Table 9). There are 85 peptides encoded in Kordia sp. TARA_039_SRF that contain the Por secretion system domain (SM Table 10), meaning that they are secreted through the T9SS to the extracellular environment or the cell surface. Their most abundant functions relate to surface adhesion, protein degradation, and carbohydrate hydrolysis. Out of the 184 peptidases found in the co-assembly, 8 of them are putatively secreted (SM Table 11). Other secreted peptides show domains that have a role in carbohydrate binding, environmental sensing through two-component system, and quorum sensing sensors. Additionally, we also found proteins with domains associated with prophage endonucleases, prokaryotic endonucleases, bacterial toxins and proteinase inhibitors.

Pigment and vitamin synthesis
Kordia sp. TARA_039_SRF is the only described Kordia spp. that can potentially synthesize beta-carotene without relying on any exogenous intermediates as it codes for the complete terpenoid backbone synthesis. Another exclusive feature is the putative ability to synthesize biotin from Malonyl-ACP and Pimeloyl-CoA. As other Kordia spp., Kordia sp. TARA_039_SRF can synthesize Riboflavin (vitamin B2) and it also has genes related to the synthesis of folate (B9), pantothenate (B5) and pyridoxine (B7). It can putatively import vitamins from the environment with specific binding proteins and an outer membrane channel coupled to a TonB transporter.

Genomic islands, mobile elements and prophages
Kordia sp. TARA_039_SRF’s genome contains 12 predicted genomic islands with a total length of 252,407 bp (5.5% of the total genome and 28% of the unique accessory genome). They code for virulence genes such as non-ribosomal peptides (NRPs), porins, toxins and invasion proteins and other genes such as those involved in oxidative phosphorylation, nitrate/nitrite transport, and sulfur carrier proteins (SM Table 12). There are 16 different kinds of transporters encoded in ten genomic islands, some of them
represented by several copies, related to oxidative phosphorylation pathways, secretory pathways, Ompf-OmpA environmental oxygen sensing and bacteriocin releasing systems. There are also 15 complete insertion sequences (IS) belonging to 8 different transposase families (SM Table 13) and three predicted incomplete prophage regions, with a total size of 26,876 bp (0.057% of the total genome length) (SM Table 14).

**Abundance of Kordia sp. TARA_039_SRF in the North Indian Ocean**

*Kordia* sp. TARA_039_SRF’s read recruitments at the species level (alignment identities >95%) did not exceed 0.01% in station TARA_039 (Figure 4A, SM Table 15). Nevertheless, surface and DCM metagenomic reads of the 20-180 µm fraction mapped against 40 and 53% of the co-assembly’s predicted genes, respectively (8.5 and 13% of horizontal genomic coverage). Relative abundances in these samples are low (0.0018 and 0.0033%) but reads map homogeneously across the genome (Figure 4B) and against both the core genome and the accessory genome (SM Table 16). Read mappings also occur homogeneously in the 0.8-5 µm and 0.1-0.22 µm fractions of the mesopelagic layer, with abundances of 0.0016 and 0.0026% each but the matched number of genes and horizontal coverage is smaller in both cases (between 24-28% of genes and ~5% of genomic coverage). The highest abundance of recruited reads was found in the mesopelagic 0.22-1.6 µm size fraction (0.009%) but it is spread through 8% of the co-assembly’s ORFs and specially in house-keeping genes like transcriptional regulators, RNA polymerase subunits and a serine aminopeptidase.

None of the other *Kordia* spp. used in the competitive Fragment Recruitment Analysis recruited enough reads at the species level to consider them present at the moment of sampling. Nevertheless, recruitments at identities between 80-90% increasing with depth by all *Kordia* genomes might suggest that a co-occurrent close relative to the co-assembly is also part of the community (SM Figure 4, 5 and 6).

**Comparative genomics of the genus Kordia**
The five *Kordia* genomes included in this comparative study belong to five different species: four *Kordia* genomes available in public databases retrieved from isolated cultures and our novel *Kordia* co-assembled genome (**Table 2**). The ANIm values between genome pairs were ~79-89% (with alignment coverages between 44-75% of the genes). Similarities at the 16S rRNA gene level ranged from 96.34 to 98.00% (with alignment coverage values between 95-100%). In both cases, the closest relative to the co-assembled *Kordia* sp. was *Kordia periserrulae*. Genomic length ranged between 4.59 and 5.35 Mb for the seawater genomes, while *K. zhangzhouensis*, retrieved from the interphase between seawater and freshwater, was the shortest (4.03 Mbp). *Kordia* sp. TARA_039_SRF had the second highest GC content (35.8%), following *K. periserrulae* (36.2). This value ranged between 33.8 and 34.2 for the other *Kordia* spp. The mean coding density of the studied genomes was 88.8%.

The core genome of the analyzed *Kordia* spp. consisted of a mean of 57% of the total number of gene clusters of every other *Kordia* genome (**Figure 5, SM Table 17**) and ~53% of them could be classified into a Cluster of Orthologous Genes (COGs) category. A phylogenomics analysis of these gene clusters defined two clades, one grouping *K. zhangzhouensis*, *K.algicida* and *K.jejdonensis* and another with the co-assembled *Kordia* sp. and *K.periserrulae* (**Figure 5**).

For the rest of gene clusters in these *Kordia* spp., a mean of 24% are shared between two, three or four genomes. These are defined as the shared accessory genome. The highest number of gene clusters in this category is shared between *K. periserrulae* and the co-assembled *Kordia* sp. TARA_039_SRF (204).

Between 18 and 28% of gene clusters are unique to each species (**Table 2, Figure 5**, constituting the flexible genome. Of those classified into COG categories, between 10-15% are classified into the cellular processes and signaling category, 6-11% are related to information storage and processing, 7-12% belong to metabolic functions and 0.2-2.3% are related to mobilome elements. The co-assembly has the lowest number of
peptidases in its flexible gene dataset (one), 7 transporters (mainly channels and pores and electrochemical potential-driven transporters) and 5 CAZymes (3 different CBM and 2 glycosyl hydrolases).

There is a high number of hypothetical proteins of unknown function in the co-assembly *Kordia* sp. TARA_039_SRF. It reaches a 41% of its core genome, 6% of its shared accessory genome and up to 89% of its strain specific accessory genome.
Discussion

The aim of this study was to characterize the novel uncultured species *Kordia* sp. TARA_039_SRF and shed light on its ecological and functional potential in ocean waters through an optimized co-assembly of ten co-occurring and nearly identical marine *Kordia* SAGs.

We have co-assembled 10 *Kordia* SAGs previously classified as nearly identical genomes [69] into a high-quality draft genome, which is 94.83% complete and meets the set standards regarding contamination (<5%) [11,37]. Recent studies have shown that co-assembling very closely related SAGs helps overcome the main drawback of their genome amplification step: that it rarely exceeds 60% of estimated genome length [36,48]. We have taken the assembly procedure one step further in order to optimize contig length, reduce genome fragmentation, and significantly decrease the needs for computational power and processing time. Read redundancy was expected to be high considering the near genomic clonality of the ten SAGs. After normalization by coverage the workable dataset was reduced to 1.75% of the original, which substantially decreased assembly computational needs and time. The use of merged reads, together with those unmerged, contributed to the reliability of predicted gene synteny in contigs [73]. The combination of different k-mer lengths during co-assembly generated remarkably longer contigs than default assembly parameters [13]. Selecting the best combination of k-mer sizes was approached by comparing different assemblies generated with different individual k-mer length values and finding the best performers. These contigs were assembled again as gene redundancy was observed in the ends of some of them (probably due to variability hotspots in some of the SAGs that forced a split in the assembly process). The resulting reference genome had a small number of very long contigs (27 contigs, N50 0.43 Mb) of high-quality (best Assembly Likelihood Evaluation score) to which the ten individual *Kordia* SAGs shared a pairwise ANIm of 99.9%. We think this co-assembly can serve as a reference genome for comparative
genomics and ecology studies just like a draft genome assembled from bulk DNA of a monoclonal culture.

The co-assemble’s taxonomic novelty was confirmed through phylogenies based on the 16S rRNA gene and highly conserved single copy genes [39] and also using whole genome composition comparisons like ANI and AAI. The closest relative is *Kordia periserrulae*, isolated from the gut of the marine polychaete *Periserrula leucophryna*.

Its genome-based functional analysis suggests that this taxon is a novel photo-heterotrophic, non-motile, bacterium that contains proteorhodopsin and oxidative phosphorylation machinery (together with membrane oxygen sensors) for aerobic and microaerobic environments. It has the putative ability to sense other limiting nutrients and elements in the environment like iron, nitrite, and nitrate. *Kordia* sp. TARA_039_SRF presents exclusive CAZymes in its genus, especially several types of carbohydrate binding modules with an affinity for phytoplankton exopolymers like cellulose or glucomannan, that would promote their hydrolysis before uptake from the extracellular environment [74]. It also encodes Polysaccharide Utilization Loci, surrounded by CAZymes as previously described in Bacteroidetes (carbohydrate-binding modules, glycosyl hydrolases, glycosyl transferases, carbohydrate esterases) [44]. Moreover, the genome encodes almost twice the number of copies of the outer membrane TonB dependent receptor/transport systems than its relatives, hinting at the apparent relevance of importing extracellular compounds for the survival of this species. The presence of virulence factors such as NRPs, antibiotics, proteases, porins, toxins and prophage related enzymes in *Kordia* sp. TARA_039_SRF’s genome (some of them exported through the Bacteroidetes specific secretion system T9SS), suggest that this novel taxon is an active player for niche colonization. It also presents features common to surface seawater bacteria: seven copies of DNA light-induced damage repair system genes [35], the green-light absorbing proteorhodopsin and other light-sensing proteins.
Such functional capabilities match the environmental characteristics of the SAGs’ original habitat, as well as its abundance in specific size fractions and depths of the water column. The SAGs co-assembled as Kordia sp. TARA_039_SRF were retrieved from a surface seawater sample from the Arabian Sea, in the North Indian Ocean (TARA_039), pre-filtered through a 200 µm mesh. This oceanic region is prone to seasonal surface phytoplankton blooms [42] and has one of the most intense and large Oxygen Minimum Zones (OMZ) [58]. A diatom bloom was registered a month prior to the sampling period and Roullier et al. [68] thoroughly described the particle distribution in the water column across the area, as well as seawater nitrate concentrations and size of the OMZ. By the time of sampling in TARA_039, surface waters showed a general decreasing concentration of large particulate matter of 15 particles m$^{-2}$ (LPM; >100 µm) [68]. This is the size range where we find highest read recruitments of the co-assembly in the available sea surface and DCM metagenomes. Even though the abundance of recruited reads in this size fraction is low, the reads spread homogeneously throughout the genome. The co-assembly codes for a large number of enzymes that would allow growth on particles that result from diatom exudates, even on those where microaerobic conditions may arise. Moreover, expression of proteorhodopsin in surface waters would provide extra energy and therefore an ecological advantage for rapid particle colonization. In fact, the very low microdiversity within the SAG population may be due to the sorting of a disaggregated particle, as proposed in [69]. The carbohydrate-binding modules, adhesins and internalization transporters encoded in the co-assembly would provide it with the ability to feed on the phytoplankton particles as they sink, regardless of decreasing environmental light. An intermediate nepheloid layer (INL) formed only by particles < 200 µm was observed in TARA_039, at 250-300 m. This depth was also the oxygen minimum zone and had the highest concentration of nitrate in the sampled water column. Of the available metagenomic samples at this mesopelagic depth, the co-assembly was more abundant in the free-living size fraction (0.22-1.6 µm). This could be related to the novel species’ putative ability to cope with low oxygen conditions and the
potential for nitrate assimilation, and therefore switching life-style from particle-associated (> 1.6 µm size fractions) at the surface to free-living at the OMZ. Metabolic versatility had previously been suggested to be related to a high number of transporters [64], a feature apparently characteristic of the genus *Kordia*, in which the co-assembly shows the highest ratio of encoded transporters per genomic Mb. A lifestyle alternation between attachment to particles and free-living has already been proposed for the marine Flavobacteria *Polaribacter* sp. MED152 [25].

The genomic characteristics of the novel *Kordia* sp. TARA_039_SRF contrast with most of the common genomic signatures described in free-living bacterioplankton SAGs by Swan et al. (2013), in accordance with the predicted attachment to particles of *Kordia* sp. TARA_039_SRF as main lifestyle. Free-living bacteria SAGs were characterized by:

i) shorter genomes (also described as characteristic of proteorhodopsin coding Bacteroidetes [22]), ii) genome streamlining with less non-coding regions and iii) lower GC content [78]. In contrast, the co-assembled genome is relatively large (~4.5 Mb), it also shows the second highest GC value of the tested *Kordia* spp. and a high portion of non-coding DNA (12%).

Comparative genomics of *Kordia* sp. TARA_039_SRF, *Kordia algicida*, *Kordia jejudonensis* and *Kordia zhangzhouensis* reveals an inter-species core genome that constitutes ~57% of the genomes included in the analysis. This value is consistent with the findings in [71], where core genome size is analyzed at the intra-species level but also between different species of the same genus. The result is also consistent with the genus-level pangenomes of marine *Alteromonas* [47] and *Prochlorococcus* [32].

The low abundance of read recruitment by the co-assembly in the samples from which the SAGs were retrieved and their low genomic coverage suggests that this novel taxon would have remained unseen using other culture-independent techniques like metagenomic genome assembly, confirming the convenience of single cell genomics in the unveiling of novel microbial taxa.
Description of “Candidatus Kordia photophila” sp. nov.

“Candidatus Kordia photophila” (pho.to’phi.la. Gr. neut. n. phos, photos light; N.L. masc. adj. philus (from Gr. masc. adj. philos) loving; N.L. fem. adj. photophila light-loving).

This bacterium is proposed to grow photolithotrophically attached to phytoplankton derived particles. Its genome-based functional annotation suggests that the taxon is non-motile, it encodes green-light absorbing proteorhodopsin and oxidative phosphorylation machinery (together with membrane oxygen sensors) for aerobic and microaerobic environments. It is the first Kordia species that encodes genes from cytochrome c oxidase types aa3 and bo. The former shows higher affinity for O2 than cbb3-type found in other Kordia spp. and the latter is expressed under iron limitation conditions. It is the first representative to encode both the nitrate/nitrite transporter NasA and the substrate binding protein NasF. Other unique features within the Kordia genus are the putative ability to perform assimilatory sulfate reduction, encoding cellulose and glucomannan specific CBM 16, 22 and 04, N-acetylmuramidase GH108 and the unsaturated rhamnogalacturonyl hydrolase GH types 105 and 88. Regarding membrane transporters absent in other Kordia spp., it encodes: i) 4 types of channels and pores (Phospholemman family, Type B Influenza Virus NB Channel family, General Bacterial Porin family and the channel-forming Colicin family), ii) 2 types of electrochemical potential-driven transporters (Betaine/Carnitine/Choline transporter family and the K+ uptake permease KUP) and iii) the slow voltage-gated K+ channel accessory protein MinK.

The type material of the proposed “Candidatus Kordia photophila” is its genome sequence, co-assembled from 10 SAGs, that can be found under accession number SMNH00000000. The version described in this paper is version SMNH02000000. Its NCBI’s taxonomy id is NCBI:txid2530203.
**Acknowledgements**

High-Performance computing analyses were run at the Marine Bioinformatics Core Facility (MARBITS) of the Institut de Ciències del Mar (ICM-CSIC) in Barcelona. We thank Shook Studio for assistance with figure design and execution. We thank Aharon Oren for helping in the etymology of the candidatus and Dr. Ramon Roselló-Móra and the two other anonymous reviewers for helping us improve our manuscript. We thank our fellow scientists, the crew and chief scientists of the different cruise legs involved in *Tara* Oceans for collecting the samples used in this study. This is a *Tara* Oceans contributed paper number XXX.
Funding

Funding was provided by the Spanish Ministry of Economy and Competitivity grant MAGGY (CTM2017-87736-R) to SGA and CTM2016-80095-C2 to CPA and JMG. PS is funded by MAGGY project and MR-L held a Ph.D. Fellowship FPI (BES-2014-068285) funded by the Spanish Ministry of Economy and Competitivity.
**Author's contributions**

SA designed this research. MR-L performed the laboratory and analyses work and wrote the paper. PS and JG helped with the data analyses. SA funded this research with contributions from CP-A and JG. All authors were involved in critical reading for writing the paper.
Conflicts of interest

The authors declare no conflicts of interest.
Bibliography


Sinking particles promote vertical connectivity in the ocean microbiome 115(29), 6799–807, Doi: 10.1073/pnas.1802470115.


Identified in a Non-Axenic Culture of the Diatom Skeletonema marinoi. J. Genomics 7, 46–9, Doi: 10.7150/jgen.35061.


Figure 1. Workflow used to co-assemble the novel *Kordia* sp. TARA_039_SRF high-quality draft genome from ten *Kordia* SAGs. The first step includes read processing and co-assembly optimization (A). It cleans and merges raw reads to eventually have a workable read dataset consisting of merged reads and clean paired-end reads that did not merge. These are co-assembled multiple times with different individual k-mer sizes, ranging between 17 and 133 bp. N50 length, number of contigs and best assembly evaluation score (ALE score) are the parameters chosen to determine the three best k-mer sizes to combine for the co-assembly that will be used afterwards. The second step aims to reduce the redundancy found in contig ends of this co-assembly and reduce contamination (B). It relies on re-assembling the co-assembly’s contigs, measuring genome completeness and contamination and discarding those shorter contigs that add extra copies of single copy genes into the co-assembly to a final contamination value <5%.
Figure 2. Phylogenetic reconstruction based of Kordia sp. TARA_039_SRF and its closest relatives based on the 16S rRNA gene (A) and 40 conserved single copy genes (B).

Figure 3. Theoretical model with highlights of the functional potential of Kordia sp. TARA_039_SRF. Membrane transporters are colored depending on their TC Family classification and their simplified structure has been inferred from TC database and related bibliography.
Figure 4. Relative abundances of recruited reads in all available metagenomes from station TARA_039 in the north Indian ocean (A). Unavailable metagenomic samples are depicted as slashed in the heatmap, samples where there was no read recruitment whatsoever appear in white. Horizontal genomic coverage of mapped reads at 95-100% in TARA_039 metagenomes (B). The total of 4193 loci encoded in the co-assembly have been collapsed in 420 bins of 10 loci each and the number of reads mapping has been colored. No read mapping is shown in white.
Figure 5. Comparative genomics of *Kordia* sp. TARA_039_SRF, *K. algicida*, *K. jejudonensis* and *K. zhangzhouensis*. Top left dendogram represents clusters of genes ordered by the number of genomes encoding them, regardless of copy numbers. The phylogenomics tree connecting *Kordia* spp. lables was done based on their core genome. Right bar plot quantifies the total number of genes that are unique and shared in all possible combinations between the five *Kordia* genomes. Bottom heatmap quantifies the number of genes classified to a COG category, peptidases, Transport Classification family or CAZymes category for each genome combination and for the unique flexible genome. Each genome has a specific color that is maintained in the different sections of the figure.
### Table 1. Metrics of the individual assemblies and the co-assembly of the 10 *Kordia* SAGs.

<table>
<thead>
<tr>
<th><em>Kordia sp. TARA_039_SRF</em> genomes</th>
<th>Individual SAGs AB-193_</th>
<th>refined co-assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assembly metrics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total N of contigs</td>
<td>179</td>
<td>27</td>
</tr>
<tr>
<td>Contig N50 (Kbp)</td>
<td>20.9</td>
<td>42.9</td>
</tr>
<tr>
<td>Mean contig length (Kbp)</td>
<td>5.9</td>
<td>170.1</td>
</tr>
<tr>
<td>Min. contig length (bp)</td>
<td>118</td>
<td>14743</td>
</tr>
<tr>
<td>Max. contig length (Kbp)</td>
<td>59.5</td>
<td>682.4</td>
</tr>
<tr>
<td>Total assembly size (Mbp)</td>
<td>1.05</td>
<td>4.59</td>
</tr>
<tr>
<td><strong>Quality parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genome completeness</td>
<td>24.36</td>
<td>94.97</td>
</tr>
<tr>
<td>Contamination</td>
<td>0.65</td>
<td>4.65</td>
</tr>
<tr>
<td>Strain heterogeneity</td>
<td>100.00</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Read mappings against co-assembly</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contribution to co-assembly (%)</td>
<td>24.56</td>
<td>-</td>
</tr>
<tr>
<td>Mapped clean reads (%)</td>
<td>97.84</td>
<td>-</td>
</tr>
<tr>
<td><strong>Genomic features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted ORFs</td>
<td>1012</td>
<td>4192</td>
</tr>
<tr>
<td>GC (%)</td>
<td>35.4</td>
<td>36.33</td>
</tr>
<tr>
<td>ANIm against co-assembly (%)</td>
<td>99.98</td>
<td>100</td>
</tr>
<tr>
<td>Identity of tetral. freq. with co-assembly (%)</td>
<td>99.53</td>
<td>100</td>
</tr>
</tbody>
</table>
Table 2. Description of the four Kordia genomes used in the comparative genomics analysis.

<table>
<thead>
<tr>
<th>Genomic features</th>
<th>Kordia sp. TARA_039_SRF</th>
<th>K. algicida</th>
<th>K. jejudonensis</th>
<th>K. zhangzhouensis</th>
<th>K. periserrulae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length (Mbp)</td>
<td>4.59</td>
<td>5.01</td>
<td>5.35</td>
<td>4.03</td>
<td>4.72</td>
</tr>
<tr>
<td>GC (%)</td>
<td>35.8</td>
<td>34.2</td>
<td>33.9</td>
<td>33.8</td>
<td>36.2</td>
</tr>
<tr>
<td>Coding density (%)</td>
<td>88.45</td>
<td>87.58</td>
<td>88.09</td>
<td>90.98</td>
<td>88.93</td>
</tr>
<tr>
<td>Total number of ORFs</td>
<td>4192</td>
<td>4492</td>
<td>4782</td>
<td>3585</td>
<td>4105</td>
</tr>
<tr>
<td>Predicted protein coding genes (CDS)</td>
<td>4124</td>
<td>4411</td>
<td>4714</td>
<td>3527</td>
<td>4051</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>94.83</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pangencode</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># clusters of core genes (% of total)</td>
<td>2286 (57.4%)</td>
<td>2286 (55.4%)</td>
<td>2286 (50.3%)</td>
<td>2286 (65.8%)</td>
<td>2286 (58.1%)</td>
</tr>
<tr>
<td># clusters of genes shared with Kordia spp. (% of total)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accessory shared genome</td>
<td>971 (24.4%)</td>
<td>1038 (25.1%)</td>
<td>969 (21.3%)</td>
<td>772 (22.2%)</td>
<td>1050 (26.6%)</td>
</tr>
<tr>
<td># clusters of unique flexible genes (% of total)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accessory strain specific genome</td>
<td>723 (18.1%)</td>
<td>801 (19.4%)</td>
<td>1287 (28.3%)</td>
<td>414 (11.9%)</td>
<td>599 (15.2%)</td>
</tr>
<tr>
<td>Total # of gene clusters</td>
<td>3980</td>
<td>4125</td>
<td>4542</td>
<td>3472</td>
<td>3935</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANIm (%)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kordia sp. TARA_039_SRF</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kordia algicida</td>
<td>81.13</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kordia jejudonensis</td>
<td>79.83</td>
<td>81.47</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kordia zhangzhouensis</td>
<td>79.62</td>
<td>79.56</td>
<td>79.23</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>Kordia periserrulae</td>
<td>89.13</td>
<td>80.97</td>
<td>79.83</td>
<td>79.63</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AAI (%)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kordia sp. TARA_039_SRF</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kordia algicida</td>
<td>81.05</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kordia jejudonensis</td>
<td>77.96</td>
<td>80.14</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kordia zhangzhouensis</td>
<td>79.93</td>
<td>80.62</td>
<td>78.06</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>Kordia periserrulae</td>
<td>92.12</td>
<td>81.03</td>
<td>77.97</td>
<td>80.06</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ribosomal RNA</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16S rRNA id. % with Kordia sp. TARA_039_SRF</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>16S rRNA id. % with K. algicida (coverage)</td>
<td>97.03 (100%)</td>
<td>100</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>16S rRNA id. % with K. jejudonensis (coverage)</td>
<td>98.00 (95%)</td>
<td>96.69 (95%)</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>16S rRNA id. % with K. zhangzhouensis (coverage)</td>
<td>96.85 (98%)</td>
<td>96.85 (98%)</td>
<td>97.24 (00%)</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>16S rRNA id. % with K. periserrulae (coverage)</td>
<td>97.56 (97%)</td>
<td>96.74 (97%)</td>
<td>97.24 (98%)</td>
<td>96.34 (99%)</td>
<td>100</td>
</tr>
<tr>
<td># of tRNA types</td>
<td>21</td>
<td>21</td>
<td>20</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td># of rRNA operons</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>