Ecogenomics of uncultured marine prokaryotes

Marta Royo Llonch

ECOGENOMICS OF UNCULTURED MARINE PROKARYOTES

PhD THESIS · Marta Royo Llonch

Tesi doctoral

Programa de doctorat en Microbiologia

Universitat Autònoma de Barcelona
Departament de Genètica i Microbiologia

Febrer 2020

Ecogenomics of uncultured marine prokaryotes

Ecogenómica de procariotas marinos no cultivados

Ecogenòmica de procariotes marins no cultivats

Marta Royo Llonch

Departament de Biologia Marina i Oceanografia,
Institut de Ciències del Mar (ICM-CSIC)

Directors:

Dra. Silvia González Acinas
Departament de Biologia Marina i Oceanografia,
Institut de Ciències del Mar (ICM-CSIC)

Dr. Carles Pedrós-Alió
Programa de Biología de Sistemas,
Centro Nacional de Biotecnología (CNB-CSIC)

Tutora acadèmica:

Dra. Olga Sánchez Martínez
Departament de Genètica i Microbiologia,
Universitat Autònoma de Barcelona (UAB)

Barcelona, February 14th 2020

During the realization of this thesis, Marta Royo Llonch held a Ph.D Fellowship FPI (BES-2014-068285) funded by the Spanish Ministry of Science and was a student of UAB's PhD program in Microbiology.

Cover design and thesis editing done by Shook Studio (@shookstudio).

A la meva mare, la meva àvia i la resta de dones rebels.
Gràcies a la seva lluita he pogut fer aquesta tesi,
que anys enrere no hagués tingut dret a fer.
També al meu pare, als meus avis i al Toni.

“It was an unexpected encounter
that slowly altered the course of my life.”
- Patti Smith, M Train

Acknowledgements

Tot i que estigui escribint els agraïments a última hora, porto pensant en aquesta secció des de que vaig saber que podria fer la tesi a l'ICM. Durant aquests anys m'he sentit molt afortunada, no només de poder estar formant-me en un lloc com l'ICM amb un equip científic de referència, sinó també de compartir les vistes al mar amb un munt de gent increïble. I amb el privilegi de poder gaudir-ho ben aprop dels amics i la família.

Primer de tot vull agraïr als meus directors de tesi, el Carles i la Silvia, haver-me donat tantíssimes oportunitats per treure profit als 5 anys que he dedicat a aquest projecte. La FPI, els cursos, els congressos, les reunions, les estades... estic absolutament satisfeta de la feina que he fet i el que he après durant la tesi, espero que vosaltres també. Carles, no se què hagués passat amb la meva vida si no m'haguessis concedit la beca per quedar-me a l'ICM, gràcies a tu he pogut iniciar una carrera científica. Gràcies per la confiança, els consells i per respondre sempre que ho he necessitat. Silvia, gracias por todo! Las reuniones infinitas, los viajes, las comidas, incluirme en proyectos excepcionales y apoyarme personal y profesionalmente durante todo este tiempo. Gràcies als dos per crear aquest espai de confiança en el que he pogut evolucionar amb llibertat.

També vull agraïr a l'Olga, la meva tutora, haver-me ajudat sempre que l'he necessitat, pels consells, els ànims i el bon rotllo.

Quan vaig descobrir que al mar hi havia microorganismes i que tenien un impacte brutal a tot el planeta vaig anar a parlar amb la professora de microbiologia ambiental a veure si em podia recomanar algun llibre de microbiologia marina. Sense saber-ho, havia coincidit amb la persona clau que va donar inici a tota aquesta historia. Isabel F., gràcies per introduir-me a l'ICM, acollir-me i ensenyar-me tot el que he necessitat per poder treballar de manera rigorosa, per recolzar-me, tenir-me sempre present i la inolvidable campanya al García del Cid.

No m'oblido tampoc dels IPs, post-docs i tècnics del departament. Gràcies per escurçar les distàncies i fer-nos sentir a tots part d'una família, tant als passadissos com al laboratori, als TEMAs o als cafès.

Gracias a Jarone, Jose y Shini por acogerme en vuestros laboratorios durante las estancias que he hecho en estos años. A parte del aprendizaje científico, tener la oportunidad de ir a trabajar temporalmente a sitios distintos me ha enseñado otras maneras de afrontar los retos y gestionar el trabajo. Gracias a vosotros y vuestra gente por haberme hecho sentir como una más del equipo.

Al dream team Acinas, qué suerte formar parte de un grupo como este! Donde no faltan risas, buen humor y freakismo bioinformático! Fran, gracias por los consejos y tu ayuda en mis tiempos de beginner en R. Guillem, gràcies per acollir-me a Zurich i els cafès amb humor i ciència. Isa

S., gràcies per aguantar-me al despatx, deixar-te molestar de tant en tant i haver sigut tan bona companya de viatges. Pablo S., que guay haver coincidit! Gràcies pels breaks, les tonteries, les bromes i tot el que m'has ensenyat. Carla, Marta F. i Andrea, ha sigo un placer haber compartido un trocito de este viaje con vosotras. Y fuera del grupo Acinas, pero como extensión del dream team, gracias a Marta S, Ana Mari, Massimo y Clara R., por el cariño que me habéis demostrado des del principio.

I que no faltin agraïments pels predocs amb qui he compartit aquests anys, els que encara segueixen, els que han acabat i els que van haver de marxar. Al multitudinari grup "malagente", aquest mote no us representa, gràcies pels dinars i cafès, plens de converses interessants i molt de riure. En especial gràcies a l'Adri, l'Ari i al Manu. Gràcies també a les companyes de despatx que he tingut aquests anys Dorleta, Mireia, Anna i Isa S., per mantenir l'ambient zen i compartir descansos, brenars i esmorzars amb bones converses, sempre disposades a motivar-nos i recolzar-nos les unes a les altres. I per acabar, gràcies a la Turón, Silvia F., Raül, Gastón, Soto, Maria, Pablo R., Lau i Pau. Crec que encara no he superat que haguessiu de marxar tots de l'ICM! La vostra amistat i tots els records del que hem fet junts són de les coses més importants que m'emporto d'aquests anys.

Parlant de gent indispensable, Clara i Vane, us toca a vosaltres. Per començar, gràcies per fer la feina i els tràmits tan fàcils, i estar sempre disponibles en moments de dubte o necessitat d'ajuda. Però sobretot, gràcies per la vostra amistat i tots els cafès, viatges, dinars i petits moments que hem compartit. Se que us tindré a la meva vida faci el que faci i allà on vagi (Pau i Lau, això també va per vosaltres!).

Fora de l'ICM també hi ha gent a la qui he de donar les gràcies. Primer a les amigues de tota la vida, per pensar en mi quan senten notícies relacionades amb el mar i per ensenyar-me altres maneres de viure la vida. Anna, Berta, Gloria, Farrés, Fabrés, Aina, Maria i Nuria, ens coneixem des de ben petites i se que part de com sóc ha sigut gràcies a créixer amb vosaltres. A la family biocel, la meva família predoc d'adopció de la UAB: Albert, Tania, Jorge, Alina, Andreu, Blanca, Aina, Celia, Lourdes, Sandra, Inma, Ot, Mireia R, Mireia S, Joan, Teresa, Uri i Marina. Sou un munt de gent molt especial i sobretot bones persones. Gràcies per acollir-me i per la vostra amistat. Les festes, excursions, dragcons, calçotades i brenars han sigut vitals durant la tesi. En especial gràcies a la Tania P. i la Lourdes, heu sigut un gran descobriment. Sou fortes, independents, llestes i sobretot, molt bones persones. Realment em sento molt afortunada de poder-vos tenir al meu costat i haver compartit les calvícies tèsiques amb vosaltres. Finalment, Karen, Tania, Aitor, Nere, Laura, Anna, Jon i Noel. Tot i que no sé què és tenir germans, sé que sou el que més s'hi acosta. Gràcies per estimar-me i fer-m'ho saber sempre de mil maneres.

Tot i que segurament no ho llegiran mai, vull agraïr a RuPaul i Lady Gaga per haver-me acompanyat durant aquesta etapa i haver-me enriquit, no només amb el seu art, sinó també amb la seva passió per la feina ben feta i no dubtar en treballar dur. També crec que he d'agraïr a la comunitat de gent anònima que col·labora en fòrums online, on he resolt mil dubtes de programació. També als desenvolupadors de paquets d'R i programes open-source, sense els quals aquesta tesi no seria la que és.

Finalment a la meva petita família. Chary, Cris, Hugo i Abril, gràcies per cuidar-me i fer-me riure, sobretot en aquesta fase final. Als meus avis, se que amb vosaltres he après a ser pacient, saber escoltar i la importància de treballar dur i ser responsable. Sabieu que me'n sortiria i m'ho feieu saber sempre que podieu, gràcies. Papa i mama, gràcies per fer tot el que heu pogut des de que era petita perquè tingués un futur ple d'oportunitats. Per animar-me, cuidar-me, donar-me consell i creure en mi sempre.

I a tu, Toni. El millor de tot aquest procés ha sigut poguer créixer amb tu. Gràcies per contagiar-me la teva creativitat, curiositat i, sobretot, el teu somriure. Tinc mil ganes de saber quins nous projectes i aventures ens esperen, res ens atura si ho vivim junts.

Contents

SUMMARY/RESUMEN/RESUM

Summary

In the last few decades, novel approaches have been applied to the study of marine microorganism aiming to retrieve taxa that escape isolation in culture. Culture independent methodologies, together with high-throughput sequencing and extensive oceanographic sampling, have provided insight into a previously unknown taxonomic and functional diversity of marine microbes. Marine microbes play a fundamental role in nutrient cycling and climate regulation at a planetary scale. Thus, it is of paramount importance to define their taxonomic classification, distribution patterns, habitat preferences and functional properties in the ocean. Linking taxonomy with function has been a challenge in Microbial Ecology, and in the recent years two alternatives have been developed towards this end. Single Cell Genomics allows the sequencing of individual genomes from environmental samples (Single Amplified Genomes, SAGs) and genome reconstruction from metagenomes allows building genomes from the whole community's DNA content (Metagenomic Assembled Genomes, MAGs). In the present dissertation, I have retrieved SAGs and MAGs from underexplored areas like the North Indian Ocean and the Arctic Ocean.

The North Indian Ocean is subject to seasonal upwelling events that provide surface waters with fresh nutrients, resulting in phytoplankton blooms. Such high primary productivity in the surface waters results in heterotrophic metabolism in the subsurface, by prokaryotes that feed on the products released by primary producers. Such high heterotrophic activity consumes the available oxygen, and together with physical processes than prevent water mixing, generates an oxygen-depleted layer in the water column: the Oxygen Minimum Zone (OMZ). These water layers are predicted to increase due to global warming and have caught the attention of microbial ecologists as they are rich in microbes involved in the cycling of nitrogen and several microaerophilic and anaerobic metabolisms. Even though the North Indian Ocean has one of the most intense and large OMZs, little is known about the prokaryotic diversity of this environment. With Single Cell Genomics I was able to retrieve 98 SAGs of a novel species in the genus Kordia and after genetically screening them for microdiversity patterns, ten were selected for complete sequencing. The ten genomes were co-assembled together and manually curated for the generation of a reference, almost complete, draft genome. I described the novelty of this species based on multiple phylogenies and comparative genomics with the other described species of the genus Kordia. I also defined the functional potential and niche preference of the novel species combining its functional annotation with its distribution in the different metagenomes of the water column of origin, that included multiple depths and size fractions.

The Arctic Ocean has a huge impact in climate regulation of our Planet and is currently being affected severely by global warming. The prokaryotic diversity of its waters has been assessed in sporadic sampling events, mostly focused on a specific season or geographic extension. In the present work I have built 3550 bins from Arctic metagenomes from different regions and seasons that are representative of almost half of the genetic content of the community. Of these, 530 can be classified as MAGs due to their medium and high-quality features and include a majority of novel taxa, especially at the species level but also at higher taxonomic ranks like Class in the case of Bacteria. I have studied their implications for the Arctic's carbon cycle, their distribution patterns and habitat preferences, and have defined habitat generalists and specialists that can serve as future sentinels of climate change in the Arctic.

Overall, this dissertation provides new insights into the taxonomic and functional diversity of uncultured taxa, and proposes new methodologies to improve genome assembly and quality controls in meta-omic mappings.

Resumen

En las últimas décadas se han aplicado nuevas metodologías al estudio de los microorganismos marinos para recuperar taxones que no crecen en cultivo. La combinación de técnicas independientes de cultivo, secuenciación masiva y muestreos oceanográficos a gran escala han permitido explorar la diversidad taxonómica y funcional microbiana a un nivel de resolución previamente desconocido. Los microorganismos marinos juegan un papel fundamental en los ciclos biogeoquímicos y regulación del clima a escala planetaria. Por eso es importante que definamos su taxonomía, distribución, hábitats y propiedades funcionales en el océano. Relacionar la taxonomía con la función a nivel genómico ha sido un reto en la ecología microbiana, pero en los últimos años se han desarrollado dos alternativas con este objetivo. La genómica de células individuales permite secuenciar genomas ambientales y la reconstrucción de genomas a través de metagenomas usa el contenido de ADN de la comunidad como material de partida. En esta tesis, he estudiado genomas usando estas dos estrategias en muestras de agua de zonas relativamente poco exploradas como el Océano Índico Norte y el Océano Ártico.

El Océano Índico Norte está sujeto a afloramientos estacionales de agua profunda rica en nutrientes que favorecen el crecimiento masivo de fitoplancton en superfície. La producción primaria es tal que el metabolismo heterotrófico que se nutre de productos derivados de fitoplancton consume la mayoría la oxigeno disponible, generando zonas mínimas de oxigeno (OMZ). Éstas se mantienen por procesos físicos que impiden su mezcla con otras masas de agua. Se prevé que estas zonas aumentarán debido al calentamiento global. Además han captado el interés de los ecólogos microbianos porque son aguas ricas en microorganismos involucrados en el ciclo del nitrógeno, y en metabolismos micro-aerobios y anaeróbicos. A pesar de que en el Índico Norte exista una de las OMZ más extensas del planeta, su diversidad microbiana ha sido poco estudiada. Mediante genómica de células individuales obtuve 98 genomas ambientales del género Kordia de los cuales, tras su análisis de microdiversidad, se seleccionaron diez para su secuenciación y co-ensamblado. El genoma co-ensamblado fue revisado manualmente para generar un genoma de referencia casi completo. Describí la novedad taxonómica de la especie con filogenias y análisis de genómica comparada con otras especies del mismo género. También definí su potencial metabólico y nicho de preferencia combinando la anotación funcional con su distribución en distintos metagenomas de la columna de agua de origen, de distintas profundidades y tamaños de plancton.

El Océano Árctico tiene un gran impacto en la regulación climática de nuestro planeta, siendo actualmente severamente afectado por el calentamiento global. La diversidad procariótica de sus aguas se ha estudiado en muestreos esporádicos, mayormente centrados en estaciones del año específicas o en ciertas regiones geográficas. En esta tesis he construido 3550 genomas a partir de metagenomes Árticos, de diferentes regiones y durante distintas estaciones del año, los cuales representan casi la mitad del contenido genético de la comunidad. De estos, 530 se pueden clasificar como genomas de calidad media y alta, e incluyen una mayoría de taxones no descritos hasta ahora, cuya novedad taxonómica es incluso a nivel de Clase en el caso de las Bacterias. He estudiado sus implicaciones en el ciclo del carbono en el Ártico, sus patrones de distribución y sus preferencias de hábitat, definiendo generalistas y especialistas que pueden servir como especies centinela en futuros estudios de cambio climático en el Ártico.

En resumen, esta tesis aporta una nueva visión en la diversidad funcional y taxonómica de procariotas marinos no cultivados, y propone nuevas metodologías para mejorar el ensamblaje de genomas y controles de calidad en los mapeos meta-ómicos.

Resum

Durant les últimes dècades s'han aplicat noves estratègies a l'estudi dels microorganismes marins per investigar aquells que no creixen en cultiu. La combinació de metodologies independents de cultiu, la seqüenciació massiva i múltiples mostrejos oceanogràfics a gran escala ens han permès explorar una diversitat taxonòmica i funcional de microbis marins fins ara desconeguda. Els microbis marins juguen un paper fonamental en els cicles dels nutrients i la regulació del clima a escala planetària. Per això, és vital que definim la seva classificació taxonòmica, patrons de distribució, preferències d'hàbitat i propietats funcionals a l'oceà. Relacionar la taxonomia amb la funció sempre ha estat un repte en aquesta disciplina, però en els últims anys s'han desenvolupat dues alternatives que persegueixen aquest objectiu. La genòmica de cèl·lules individuals permet seqüenciar genomes ambientals i la reconstrucció de genomes a partir de metagenomes aprofita el contingut total d'ADN de la comunitat. En aquesta tesi, he generat genomes fent servir aquestes dues estratègies en mostres d'aigua de zones relativament menys explorades com l'Oceà Índic Nord i l'Oceà Àrtic.

L'Oceà Índic Nord és subjecte d'afloraments estacionals d'aigua profunda rica en nutrients, que a la superfície afavoreixen el creixement massiu de fitoplàncton. La producció primària a la superfície és tal que el metabolisme heterotròfic que es nodreix de productes derivats del fitoplàncton consumeix la majoria d'oxigen disponible, generant zones mínimes d'oxigen. Aquestes es mantenen gràcies a processos físics que eviten que es mesclin amb altres masses d'aigües i s'ha predit que augmentaran degut a l'escalfament global. Aquestes àrees han atret l'atenció dels ecòlegs microbians perquè són aigües riques en microorganismes relacionats amb el cicle del nitrogen i diversos metabolismes micro-aerobis i anaerobis. Tot i que l'Índic Nord presenti una de les zones mínimes d'oxigen més extenses del planeta, la diversitat microbiana d'aquest ambient s'ha estudiat poc. Mitjançant genòmica de cèl·lules individuals vaig obtenir 98 genomes ambientals del gènere Kordia, dels que se'n van seleccionar deu per seqüenciar-los, després d'analitzar-ne els patrons de microdiversitat. El seu co-assemblatge es va revisar manualment per generar un genoma de referència gairebé complet. Vaig descriure la nova espècie amb múltiples filogènies i anàlisis de genòmica comparada amb altres espècies del mateix gènere. També vaig definir el seu potencial metabòlic i nínxol de preferència combinant la seva anotació funcional amb la distribució a varis metagenomes de la columna d'aigua d'origen, incloent-hi diferents fondàries i mides de plàncton.

L'Oceà Àrtic té un gran impacte a la regulació climàtica del nostre planeta i està severament afectat per l'escalfament global. La diversitat procariòtica de les seves aigües s'ha estudiat en mostrejos esporàdics, majoritàriament centrats en estacions de l'any concretes o certes regions geogràfiques. En aquesta tesi he construït 3550 genomes a partir de metagenomes de diferents regions àrtiques i durant diverses estacions de l'any, en una circumnavegació de les aigües l'Oceà Àrtic. Representen gairebé la meitat del contingut genètic de les comunitats i d'aquests, 530 es poden classificar com a genomes de qualitat mitjana i alta. Inclouen una elevada novetat taxonòmica, sobretot a nivell d’espècie però fins i tot a nivell de Classe, pel que fa als Bacteris. He estudiat les seves implicacions al cicle del carboni a l‘Àrtic, així com els seus patrons de distribució i les seves preferències d’hàbitat, definint generalistes i especialistes que poden servir com a espès sentinella en futurs estudis de canvi climàtic a l’Àrtic.

En resum, aquesta tesi aporta una nova visió en la diversitat funcional i taxonòmica de procariotes no cultivats i proposa noves metodologies per millor l’assemblatge de genomes i controls de qualitat en els mapejos meta-omics.

INTRODUCTION

Introduction

Marine microbes

Life on Earth may have started in the marine hydrothermal environment and in the form of anaerobic chemolithoautotrophic unicellular entities around 3.7 million years ago (Martin et al., 2016; Weiss et al., 2018). All organisms present on the planet have evolved from these and traditionally, they have been phylogenetically classified according to their conserved ribosomal RNA and core genes into three domains of life: Archaea, Bacteria, and Eukarya (Woese et al., 1990). How are they related in the tree of life, however, is not completely clear and under revision (Weiss et al., 2018; Williams et al., 2013; McInerney et al., 2015). Current marine microbes comprise organisms that belong to all of these domains but the two groups that I have focused on in this thesis are Bacteria and Archaea, also known as prokaryotes due to their lack of nuclear membrane. Since the beginning of life on our planet, these two groups have developed all possible biologically-mediated chemical reactions known to date (Kirchman, 2018), shaping the biological and chemical history of the planet (Falkowski et al., 2008).

In the ocean, one ml of seawater contains approximately 10^6 prokaryotic cells, making 10^{29} cells in the whole ocean (Whitman et al., 1998). In fact, marine microbes are estimated to account for 70-90% of the ocean's biomass (Fuhrman and Azam, 1980; Whitman et al., 1998; Bar-On et al., 2018). Being the most abundant organisms in the marine biosphere, their role in marine nutrient cycling is vital for the functioning of the ecosystem.

Biogeochemical roles of marine prokaryotes

Atmospheric CO_2 fixation in the marine environment is carried out by autotrophs, using the energy of light (photoautotrophs) in a process called primary production. Approximately half of primary production on Earth is carried out in the oceans (Field, 1998), mostly by photosynthetic groups like cyanobacteria (e.g., *Prochlorococcus*, *Synechococcus*) and eukaryotic algae (e.g., diatoms and coccolithophorids among other groups). This light-dependent process occurs in the photic ocean, that ranges from the surface down to about 200 m deep, depending on the chemical composition and amount of particulate matter in the water. Organic carbon from primary producers can enter the trophic food-webs directly by being consumed by heterotrophic microbes and animals such as copepods or salps or indirectly through the microbial-loop in the form of dissolved organic matter (DOM) used by heterotrophic bacteria and archaea (Azam et al., 1983; Fenchel, 2008). Between 1 and 40% of primary production reaches aphotic layers in the form of sinking POM (Ducklow et al., 2001), which can be gradually degraded by heterotrophic microbes (Smith et al., 1992). With the transformation of inorganic carbon into organic matter through primary production and the flux of this organic matter through heterotrophic metabolism, microbes are at the base of the oceanic food web and are key biogeochemical regulators. Even though CO_2 is released back to the atmosphere as a result of heterotrophic respiration, a small fraction of the carbon biomass produced by primary production will be removed from the carbon cycle for very long periods of time. This is due to the biological carbon pump, which is based on the sedimentation of particulate organic matter, exported from the photic layers, in various forms like detrital matter, fecal pellets or dead cells (Bopp et al., 2015).

Light-independent carbon fixation processes (chemoautotrophy) occur in coastal, shelf and open-ocean waters, in the whole water column, including the sinking particulate organic matter (POM), and sediments. This so-called dark carbon fixation is estimated to total less than 2% of the amount of carbon fixed by photosynthesis (Middelburg, 2011). Nevertheless, the impact of chemoautotrophy based on the oxidation of reduced sulfur, carbon monoxide or methane, which have been reported to be active in the dark ocean (Swan et al., 2011), is yet unknown. The main energy sources for the dark carbon fixation have been postulated to be nitrification through ammonia (Könneke et al., 2005; Wuchter et al., 2006; Alonso-Sáez et al., 2012) and nitrite oxidation (Pachiadaki et al., 2017).

The marine cycling of carbon is tightly coupled with key transformations in the cycles of nitrogen and phosphorus. Together with hydrogen, sulfur and oxygen they are the elements that constitute the building blocks for all biochemical macromolecules (Schlesinger, 1991). Thus, they are limiting for microbial growth, affecting all trophic stages in the ecosystem and the stable chemical composition of the oceans and the atmosphere. The ocean harbors a plethora of niches with different concentrations of O_2, from the highly oxygenated surface waters of colder latitudes, to the oxygen depleted layers, or the microaerobic niches in organic matter particles along the water column and anaerobic sediments. In the water column, oxygen depends on the equilibrium between oxygenic photosynthesis and heterotrophic respiration and its availability defines the biological redox of the environment and its occurring metabolic reactions.

The marine environment

The oceans comprise the largest continuous ecosystem on Earth, but the availability of nutrients and oxygen is not homogeneous. Higher temperature and lower salinity of the surface layer of the ocean, compared to the waters underneath, lead to a vertical stratification of the water column. This prevents mixing of the (eventually) oligotrophic top-layer with the deeper, colder waters, that are richer in inorganic nutrients. Stratification can vary seasonally. In polar waters, and less strongly in temperate waters, stratification of the water column during spring and summer is followed by a mixing during autumn and winter. In tropical and sub-tropical waters, seasonal variation is not as pronounced, leading to areas where stratification becomes permanent and results in a profound nutrient depletion of surface waters. Such oligotrophy can be found in over 40% of the planet's surface, mostly in the five subtropical oceanic gyres (Tomczak and Godfrey, 1994; Kirchman, 2018).

Stratification can be altered by wind-driven movement of water masses like upwellings. These replace warm surface waters with colder and nutrient richer deep waters that stimulate primary production. Upwellings are common in equatorial waters, in the Southern Ocean and western coasts globally (Anderson and Lucas, 2008). Often there are defined oxygen-depleted water layers called Oxygen Minimum Zones (OMZ), in association with upwelling regions, These OMZs occur at mid-water depths (between 200-1000 meters) and below the pycnocline, and their decrease in oxygen is due to a combination of a high organic matter respiration, poor ventilation and sluggish circulation (Wyrtki, 1962; Ulloa et al., 2012). Sinking organic matter generated by photosynthesis at the surface is consumed by heterotrophic bacteria in these depths (Diaz et al., 2013), leading to a decrease in O_2 concentrations. When anaerobic microbial process occur, significant amounts of nitrite and nitrate are converted to N_2 (Lam and Kuypers, 2011), converting OMZ into sinks of "fixed nitrogen" that are estimated to represent 30-50% of all the nitrogen lost to N_2 in the oceans (Codispoti et al., 2001).

Wind-generated currents have no effect below 200 m depth, where a certain homogeneity in temperature, nutrient concentration and salinity is found. The dark ocean also differs from the epipelagic in higher pressure, lower temperature, higher concentrations of inorganic nutrients and darkness. Most of the carbon demand of the deep ocean is satisfied by POM sinking from the surface (Arístegui et al., 2009), together with autotrophic production via chemolithoautotrophy (Herndl and Reinthaler, 2013). Carbon transfer efficiency from the surface to the deep ocean is high in high latitudes, intermediate in the tropics and low in subtropical gyres (Weber et al., 2016). Most of this carbon is respired in the mesopelagic zone by prokaryotes and returned to the atmosphere, while the rest can be respired into CO_2 at the bathypelagic zone, where it remains sequestered until the current circulation transports it to upper layers in exchange with the atmosphere. One third of biological CO_2 production in the oceans is estimated to occur in the deep ocean (del Giorgio and Duarte, 2002; Arístegui et al., 2009; Reinthaler et al., 2010).

Marine microbial ecology

Prokaryotic richness on planet Earth is estimated to comprise between 10^{11}-10^{12} species (Locey and Lennon, 2016; Eguíluz et al., 2019). However, the distribution and abundance of these species is not homogeneous whatsoever. The Baas-Becking hypothesis for microbial organisms states that "everything is everywhere: but the environment selects" (Baas Becking, 1934) and species sorting has been shown to be an important determinant for the assembly of microbial communities (Van der Gucht et al., 2007; Logue and Lindström, 2008). Changes in the environmental conditions generate a response in the community, which is suggested to be mediated by a combination of adjustment to changes by genome plasticity and horizontal gene transfer, replacement (e.g., high dispersal rates) and species interaction mechanisms (Allison and Martiny, 2008; Comte and del Giorgio, 2011).

In the marine environment, there are several patterns found in microbial community structure and composition, that can be classified into the following three types: the abundant vs. rare biosphere rank-abundance curve, patterns in the spatial dimension and seasonal patterns. First, when looking at the rank-abundance distribution of a community's taxa, we find that a few species are very abundant, while some are moderately present and a large number are represented by a small amount of individuals (Pedrós-Alió, 2012). The position of species in this curve is dynamic, as changes in environmental conditions may favour rare taxa into becoming abundant while processes affecting population size (i.e viral lysis and grazing) target abundant taxa, making them move to the "rare" tail of the curve (Pedrós-Alió, 2012). Second, variation in community composition in the spatial dimension includes: i) that samples from different layers in the water column are more diverging than samples from the same layer but different oceanic region (DeLong et al., 2006; Sunagawa et al., 2015a), clearly separating the photic communities (i.e surface and DCM) from the mesopelagic ones at a global scale; ii) that the latitudinal gradient of diversity observed in macroorganisms, stating that species richness increases in latitudes closer to the equator due to smaller latitudinal range of organisms at lower latitudes, is also applied to microbes (Amend et al., 2013; Ibarbalz et al., 2019; Sul et al., 2013) and iii) that polar zones, even though they are very distant from each other, are more similar in community composition than compared to communities of more temperate or tropical latitudes (sharing nevertheless less than a third of their species diversity) (Ghiglione et al., 2012; Sul et al., 2013). Third, bacterial communities change over time, defining robust seasonal patterns (Sintes et al., 2013; Díez-Vives et al., 2014; Fuhrman et al., 2015) in which the microbial community composition is resilient, but certain populations increase or decrease in abundance (Caporaso et al., 2012).

Microbial communities are biological networks in which direct and indirect links between their members make the success of a species dependable on the performance of others or their biological interactions (Doney et al., 2012; Lima-Mendez et al., 2015). Exploring the whole diversity of microorganisms, their abundance and interactions is limited by technological advances. In addition, quantifying species richness in marine samples depends on the decision of what makes a microbe different enough from another to call it a distinct species, which is still under discussion.

The technological factor in the exploration of marine microbes

Traditional isolation methods limit the knowledge of the diversity of marine prokaryotic communities

Traditionally, microbiology relied on isolation in culture of microbes from environmental samples as the first step in biodiversity studies, which limited the known microbial diversity of natural communities to what was able to grow in a chosen medium under the given incubation conditions. In fact, only a small portion of the bacterial community in an environmental sample can from colonies on agar media, compared to the number of cells examined by microscopy. This is the so-called "great plate count anomaly" and was coined in 1985 by Staley and Konopka (Staley and Konopka, 1985). When compared to the diversity obtained with conventional molecular methods, isolation in culture yields both abundant and very rare bacteria from the ecosystem. The ability to form colonies, or "culturability", of an environmental microbe, depends on many factors. First of all, the composition of the growth medium and incubation conditions, that should resemble the physico-chemical conditions of the environment (i.e. nutrient composition, temperature, light, pressure). Second, the physiological status of the cell, for example, dormancy can determine the duration of a lag phase for the cell to be active and grow again, if it were ever to happen (Buerger et al., 2012). Third, the enzymatic array of a cell confers certain metabolic strategies for thriving under culture-like conditions. It has been the case for bacterioplankton predominating in high-nutrient summer seawaters, compared to those living the oligotrophic waters of a subtropical gyre. Lastly, co-evolved microbial interactions, as have been seen on organic matter aggregates, can also challenge the success of culturing (Azam, 1998).

Direct staining techniques, and later 16S rRNA sequencing, of environmental samples revealed the underestimation of bacterial abundances and diversity in culture-dependent studies. The proportion of bacterial cells to produce colonies in standard plating techniques has been estimated to range between 0.01-0.1% (Kogure et al., 1979) when plating surface seawater, increasing to 3% in deep ocean seawaters (Sanz-Sáez et al. 2020, in prep.) and having an overall median of 0.5% when considering multiple terrestrial, marine and host environments (Lloyd et al., 2018). However, the common paradigm of the proportion of environmental microbial culturability is of 1%, usually attributed to Torsvik and Øvreås (2002) or Amann et al. (1995). The "1% culturability paradigm" has been recently discussed and proposed to be revised by Martiny (2019) and Steen et al. (2019), as its precise meaning is often unclear and the underlying data are hard to find.

Despite the isolation efforts that have enriched our census of marine cultured bacteria (Ziegler et al., 1990; De Bruyn et al., 1990; Button et al., 1993; Eilers et al., 2001; Zengler et al., 2002; Connon and Givannoni, 2002; Kaeberlein et al., 2002; Ferrari et al., 2005; Giovannoni and Stingl, 2007; Buerger et al., 2012; Das et al., 2015; Crespo et al., 2016; Sanz-Sáez et al., 2019), many prokaryotic phyla remain without isolated representatives (Dick and Baker, 2013; Hug et al., 2016).

Culture-independent microbiology sheds light into the "black box" of marine microbial taxonomic and functional diversity

Culture-independent techniques are responsible for the explosion of the marine microbial diversity knowledge along the last 30 years. For bacterioplankton, these have been based on the 16S rDNA gene, a marker gene reflective of the evolutionary history of the respective organisms (Woese and Fox, 1977). The first approach to genetic community profiling was the sequencing of cloned 16S rRNA genes amplified from marine seawater samples using universal primers and polymerase chain reaction (PCR) (Pace et al., 1986). This revealed the immense species richness in marine samples compared to what had been found with cultures, for bacteria (Giovannoni et al., 1990; Ward et al., 1990; Schmidt et al., 1991; Mullins et al., 1995; Acinas et al., 1999) and archaea (DeLong, 1992; Fuhrman et al., 1992; Massana et al., 2000). The approach, however, had biases caused at each step of the methodology, including sample collection, cell lysis, nucleic acid extraction, PCR amplification and cloning (Wintzingerode et al., 1997). Even the most widely used "universal" primer pairs have been biased towards cultured taxa. In addition, the varying copy numbers of this gene in different taxa and microdiversity within the operons of an organism, can lead to distorted estimates of diversity (Acinas et al., 2004a,b). The definition of the most abundant groups of marine archaea and bacterioplankton (i.e the cyanobacteria *Prochlorococcus* and *Synechococcus*, and proteobacteria like SAR11, SAR83 or *Alteromonas*) was based on this approach (Salazar and Sunagawa, 2017). It also served for the first studies on biogeography of prokaryotic taxa and the phylogenetic structure of natural communities (Hagström et al., 2002; Acinas et al., 2004b; Fuhrman et al., 2006; Hughes et al., 2016; Pommier et al., 2007). Nevertheless, researchers were aware of the limitations of this technique: it captured only the most abundant microbes, leaving the majority of rare taxa unknown (Pedrós-Alió, 2006) and it could not provide any functional information.

Since 2007, the emergence of high-throughput sequencing (HTS) methods like 454 and Illumina enabled community nucleic acid sequencing at an unprecedented scale. Metagenomes were generated from sequencing the community's DNA content and metatranscriptomes resulted from sequencing the community's retro-transcribed RNA. Massive amounts of short sequences were generated from a sample, without the need for targeting a specific gene or a PCR amplification step. Compared to the traditional Sanger sequencing, high-throughput sequencing delivers 5 to 7 orders of magnitude more sequences per run (Glenn, 2011; Loman et al., 2012; Goodwin et al., 2016), although reads are shorter. In the last decade, new sequencing approaches have emerged in the field (i.e PacBio, Nanopore, MinION) that rely on sequencing of individual DNA molecules and producing longer reads (Goodwin et al., 2016).

With HTS, the access to a larger fraction of the diversity and functional potential of the community was made possible (Sogin et al., 2006). The advent of HTS together with global marine diversity surveys, has provided millions of publicly available sequences and publications on the major questions in marine microbial ecology.

Global surveys give the first overview of marine planktonic taxonomic and functional diversity, community dynamics and set a precedent in the release of open data and standardized methodologies

The first approximation to the planktonic diversity of the global ocean started in 2003 with the Global Ocean Survey (GOS) expeditions that lasted until 2010. These revealed some 1,300 different 16S rRNA gene sequences in surface seawater samples from the Sargasso Sea (Venter et al., 2004) and millions of gene families in samples in the North Atlantic through the Panama Canal, ending in the South Pacific (Rusch et al., 2007). Contemporary to the GOS, in 2005, the International Census of Marine Microbes (ICoMM) meant to use standardized sampling and data-analysis procedure to study the microbial diversity in a multitude of marine habitats, using 454 sequencing of the 16S rRNA gene profiles (Amaral-Zettler et al., 2010). With this, Sogin and collaborators (Sogin et al., 2006) realized that the largest fraction of diversity in all communities showed very low abundances, which they described as "the rare biosphere".

Between 2009 and 2013, the Tara Oceans Expedition sampled the global ocean from surface waters to the mesopelagic layer of temperate and polar latitudes. It used standardized sampling procedures to obtain seven size fractions of planktonic diversity, from viruses to small metazoans (Pesant et al., 2015). Focusing on genetic diversity, the Ocean Microbial Reference Gene Catalog was built with a total of > 40 million-non-redundant genes, of which over 80% of the sequences were found to be novel (Sunagawa et al., 2015a). This extensive effort has resulted in many significant results: like the genetic repertoire of the sampled global ocean for prokaryotes (Sunagawa et al., 2015a) and eukaryotes (Carradec et al., 2018), the planktonic interactions occurring in the photic ocean (Lima-Mendez et al., 2015), the diversity of eukaryotic plankton (de Vargas et al., 2015) and viruses (Gregory et al., 2019; Brum et al., 2015) or the impact of ocean warming on community composition and gene expression (Salazar et al., 2019) and on the current state of spatial patterns of plankton diversity (Ibarbalz et al., 2019). Between 2010 and 2011, the Malaspina Expedition targeted the microbial diversity of deep waters, sampling vertical profiles that reached the bathypelagic layer (1,000-4,000 meters), and elucidating: the diversity of deep-sea pelagic prokaryotes (Salazar et al., 2016), the diversity of heterotrophic protists in the deep ocean (Pernice et al., 2014), the vertical connectivity in the ocean microbiome due to sinking particles (Mestre et al., 2018), the phylogenetic distribution of particle-association lifestyle of bathypelagic prokaryotes (Salazar et al., 2015) or the metabolic architecture of the deep ocean microbiome (Acinas et al., 2019).

Culture-independent microbiology gravitates towards genome-centric studies

Metagenomics of a marine sample provides an overview of the functional potential of the community. In large-scale surveys, the multitude of samples can provide patterns of functional diversity: in the different oceanic regions (horizontal axis), along the different layers of the water column (vertical axis) and in the free-living or particle-attached fractions (Sunagawa et al., 2015a; Acinas et al., 2019). Functional profiling of marine communities through metagenomics has revealed the rather low level of functional redundancy of marine microbial systems (Galand et al., 2018; Salazar et al., 2019), postulated to be the opposite based on isolated marine taxa (Louca et al., 2016). Important insights have been revealed complementing metagenomic functional profiling with gene expression data from metatranscriptomics. For example, the ecological importance of *Prochlorococcus* in photosynthesis, carbon fixation and ammonium uptake in the tropical and subtropical ocean, where their most highly expressed genes encode these functions; the unexpected high contribution of picocyanobacteria in the community in terms of transcripts, considering their lower presence in metagenomes; the opposite case with heterotrophic bacteria like the highly abundant SAR11 (Frias-Lopez et al., 2008; Shi et al., 2011; Dupont et al., 2015) or the recent finding that nitrogen fixation is detected in the mesopelagic Arctic waters, in an thorough biogeographic analysis of the abundance and expression of the nitrogenase gene nifH from pole to pole (Salazar et al., 2019).

Even so, assigning function to taxonomy, which has been an essential goal in the Microbial Ecology of uncultured prokaryotes, requires the genetic information to be considered in a genomic context.

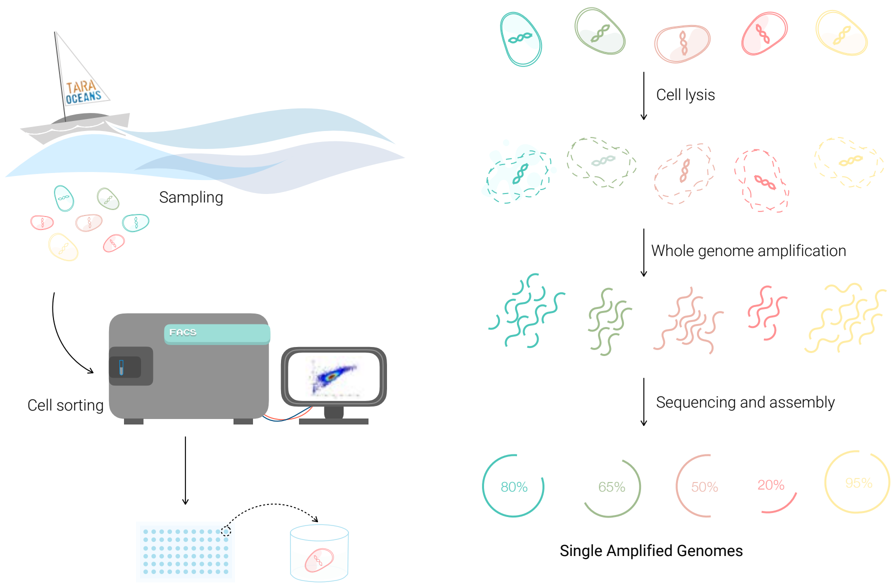


FIGURE 1. Simplified workflow for SAG generation from seawater samples.

Whereas direct analysis of metagenomes can only provide a community overview, other strategies have been developed to either access individual environmental genomes without the need for cultivation (Single Cell Genomics) or grouping the community's metagenomic information into meaningful genomic units, reflective of a population of very close taxa.

Single Amplified Genomes (SAGs). SAGs are generated from the direct amplification of DNA from previously sorted individual cells, their sequencing and assembly (Figure 1). SAGs are environmental genomes sequenced from the most fundamental units of life (Stepanauskas and Sieracki, 2007; Woyke et al., 2009; Stepanauskas, 2012; Blainey, 2013). Individual cells are selected from a sample by fluorescent-activated cell sorting (FACS) or microfluidics. The next step involves cellular lysis and whole genome amplification by several methodologies: degenerate oligonucleotide-primed PCR, multiple displacement amplification (MDA) or X-WGA (Stepanauskas et al., 2017). This is followed by sequencing and genome assembly. Single-cell genomics (SCG) retrieves all the DNA molecules of a cell, unveiling microbial intimate interactions in their natural environment otherwise overlooked, like infections, symbioses and predation (Yoon et al., 2011; Martínez-García et al., 2014; Roux et al., 2014; Labonté et al., 2015; Cornejo-Castillo et al., 2019). As SCG circumvents the taxonomic binning used in metagenomic assembly, it improves understanding of microevolutionary processes in the environment (Kashtan et al., 2014). Valuable lessons learnt through SCG studies include the chemolithoautotrophic potential of the deep ocean (Swan et al., 2011), confirmation that genome streamlining is also happening in uncultured marine free-living bacteria (Swan et al., 2013) and multiple studies on functional diversity of bacterioplankton and marine archaea.

Metagenomic Assembled Genomes (MAGs). Metagenomic reads can be assembled into contigs and later binned into genomes, the so-called Metagenomic Assembled Genomes (Figure 2).

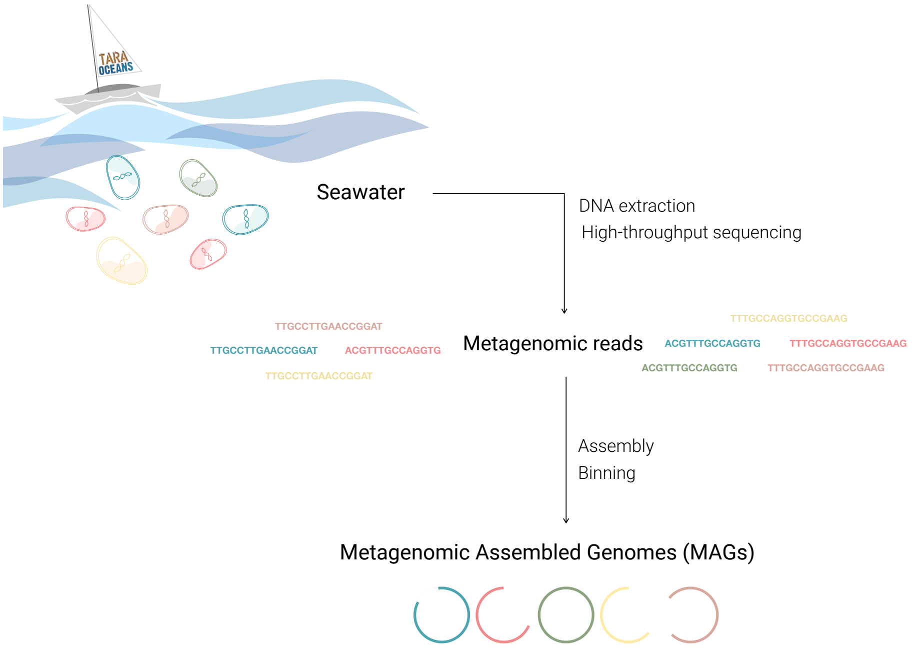


FIGURE 2. Simplified workflow for Metagenomic Assembled Genome generation from seawater samples.

MAGs are composite genomes of populations from natural communities. The first attempts at reconstructing genomes from environmental communities started in the early 2000s (Tyson et al., 2004; Martín et al., 2006), but more reliable methodologies and bigger-scale results emerged in the last decades (Wrighton et al., 2012; Albertsen et al., 2013; Sharon and Banfield, 2013; Alneberg et al., 2014; Parks et al., 2017). Nowadays, thousands of genomes have been recovered from marine metagenomes, both from isolated sampling events and oceanographic cruises (Acinas et al., 2019; Tully et al., 2017b, 2018a; Delmont and Eren, 2018; Delmont et al., 2018) following different strategies in each step of the process.

In the last five years, there has been an increase in the development of tools for estimating genome completion and contamination (Eren et al., 2015; Parks et al., 2015) and guidelines on genome quality standards and complementary analyses for the correct deposition in public databases (Bowers et al., 2017; Konstantinidis et al., 2017). Both MAGs and SAGs have been successfully combined to derive conclusions from organisms and the ecosystem and to improve the binning quality of single-cell assemblies or metagenome binning performance. The performances of the two strategies have been compared in a study aiming to investigate the functional potential of dominant bacterial populations of the Baltic Sea (Alneberg et al., 2018).

The prokaryotic species issue

Since the late 80s, we have a coherent framework on which a true phylogenetic classification of prokaryotic species can be built, based on the comparison of 16S rRNA gene sequences (Oren and

Garrity, 2014). Despite having a universal system for prokaryote classification, microbiologists do not have a single concept describing what prokaryotic species are, beyond the pragmatic proposals of the last years.

The prokaryotic species concept

Even though microbial ecology studies are based on the classification of the individuals in a community into species, a definition of prokaryotic species comparable to that of a biological species, proposed by Ernst Mayr in 1942 is lacking (Mayr, 1942). That is, because prokaryotes reproduce asexually and their genetic repertoire is subject to homologous recombination events (Fraser et al., 2007; Papke et al., 2007) and lateral gene transfer between similar or distant relatives (Doolittle and Papke, 2006).

Several concepts have been proposed (Rosselló-Mora and Amann, 2001; De Queiroz, 2005; Nesbø et al., 2006; Staley, 2006), but none has been officially accepted. In fact, these have become more of a pragmatic and operational (rather than conceptual) description based mainly on finding the boundaries between close monophyletic groups in terms of genetic coherence (Rosselló-Mora and Amann, 2001). A certain degree of this genetic coherence had been found when looking at the average identity between nucleotides between pairs of genomes of close pathogenic strains (Konstantinidis and Tiedje, 2005a), which had a minimum of 95%. However, it was thought that the frequency of appearance of new branches of genetic diversity might be more continuous in the environment than in culture or host-related habitats (Dethlefsen et al., 2007; Achtman and Wagner, 2008; Konstantinidis et al., 2006). Thus, finding a defined boundary might be more difficult than with pure cultures. Years later, with the analysis of the GOS dataset of marine metagenomes, this coherence was also found between environmental marine metagenomic reads and their reference genomes (Caro-Quintero and Konstantinidis, 2012).

Even within delineated species, there is no homogeneity in the genetic content or in their total nucleotide composition. This is called microdiversity (Acinas et al., 2004a; Fuhrman and Campbell, 1998) and it can be seen as "bushy tips" of distinct sequence clusters in phylogenetic reconstructions (Cohan, 2001; Giovannoni, 2004). These differences may persist thanks to forces like periodic selection (Cohan, 2001) and homologous recombination (Whitaker et al., 2005) leading to separate ecotypes. The ecotype concept describes a collection of strains that show some ecological distinctiveness within its species, by accumulated neutral mutations (Cohan, 2001) or recombination processes (Fraser et al., 2007; Konstantinidis and DeLong, 2008; Shapiro et al., 2012). Ecotypes preserve nearly the full phenotypic and ecological potential of the species with slight changes in their genetic collection that enable them to exploit a slightly different ecological niche.

A widely accepted view of microbial species is the pan-genome concept, developed after a comparative genomics study of the pathogen Streptococcus agalactiae (Tettelin et al., 2005). The pan-genome classifies de genetic repertoire of a species into the core genome, that includes all genes shared between all individuals categorised as the same species, and the flexible genome, which includes the gene pool that is partially shared or strain-specific (Tettelin et al., 2005; Mira et al., 2010). Thus, the genomes of multiple representatives of the species are needed to accurately define the genetic potential and size of the pan-genome. Even though the core genome contains the essence of the species and is indispensable, it is the flexible genome that can confer selective advantages like niche adaptation, new host colonization or antibiotic resistance and contributes to the species diversity (Tettelin et al., 2005, 2008). The size of the pan-genome is determined by the size of the genomes, and it can be open or closed depending on the multitude of niches in which the species is able to live (Medini et al., 2005).

The prokaryotic species definition

Through the years, the definition of microbial species has become a polyphasic concept. Before culture-independent technologies emerged, species definition was achieved with the combination of the phylogenetic relationship inferred by 16S rRNA gene sequence similarity between pairs of taxa (Stackebrandt et al., 1985; Ludwig and Schleifer, 1994), by the percentage of hybridization between sequenced genomes (DDH, DNA-DNA hybridization) (De Ley and De Smedt, 1975) and phenotypical traits (Johnson, 1973; Vandamme et al., 1996). These delineated the first arbitrary universal thresholds for bacterial species definition. Microdiversity was assessed by alignment of multiple single copy genes (MLSA, multi-locus sequence analysis) (Maiden et al., 1998; Feil, 2004) and by sequence similarity in the internal transcribed spaces (ITS) within the ribosomal operon (Brown and Fuhrman, 2005). In shotgun-sequencing studies, 16S rRNA gene reads are clustered based on their sequence similarity into "operation taxonomic units" (OTUs). As more studies focused on the microdiversity within closely related clades, the clustering methodology has evolved, since 97% of similarity was not enough (Acinas et al., 2004a; Nguyen et al., 2016). Approaches that rely on comparisons of genomic composition like Average Nucleotide Identity (ANI), Average Aminoacid Identity (AAI) and the tetranucleotide frequency are recent improvements (Richter and Rosselló-Móra, 2009; Jain et al., 2018). They depend on sequenced genomes, which was an issue before the advent of single-cell genomics and genome recovery from metagenomes, as the majority of bacteria have not yet been cultured, but currently these are the most used approaches.

Towards a unified and official description of the uncultured

Despite the enormous advances in our knowledge of the ecology of marine microbes that culture-independent methodologies have enabled in the last 40 years, the current classification of novel taxa in the official microbial taxonomy system is limited to pure cultures (Konstantinidis et al., 2017; Parker et al., 2019). Their deposition in two culture collections is indispensable to be recognized. The term Candidatus was proposed in 1994 by Murray and Schleifer (Murray and Schleifer, 1994) meaning that the proposed novel species could not be described well enough to establish a novel taxon and it has been used ever since, but it has never been accepted by the International Code of

Nomenclature for Prokaryotes (ICNP). This has resulted in a lack of review of the nomenclatures given to new taxa, with consequent errors in over 30% of the names (Oren, 2017). The use of alphanumerical codes instead of Linnean binomial names has also brought many confusions. The increasing number of genomic sequences retrieved by single-cell genomics or reconstructed from metagenomes has highlighted the need of a standardized process in the classification of uncultured microorganisms, which has been proposed to include the combination of several analyses: phylogenetic affiliations based on the 16S rRNA and conserved marker genes, pairwise comparisons of nucleotide and aminoacid composition (ANI and AAI), phenotypic characterization extracted from bioinformatic gene predictions and, if possible, distribution analyses using metagenomics and microscopy pictures (Rosselló-Móra and Amann, 2015; Sutcliffe, 2015; Konstantinidis et al., 2017).

AIMS AND OBJECTIVES

Aims, outline and objectives of the thesis

The general aim of this thesis is to explore the possibilities that genome-centric, culture-independent techniques like Single Cell Genomics and reconstruction of genomes from metagenomes can offer to further our knowledge on the ecology of uncultured marine prokaryotes.

This work can be divided into **two strong objectives**: on the one hand, there is the development of curated methodological workflows that go beyond standard procedures and that focus on the biology and ecology behind the sequences. On the other hand, there is the contribution to marine microbial ecology with genome-resolved ecological studies and characterization of marine prokaryotic populations and communities, sampled from underexplored marine areas like the North Indian Ocean and the Arctic Ocean.

The thesis is arranged in **three chapters**. In the first chapter, I analyzed the microdiversity within a population of 98 co-occurrent single amplified genomes assigned as an unknown species of the genus *Kordia* using the MLSA approach, and defined their putative distribution patterns in multiple metagenomes of different size fractions, using sequenced SAGs that shared the same 16S rRNA gene sequence. The apparent clonality within the SAG population led to the work presented in the second chapter, in which ten of these SAGs were sequenced aiming to produce a high-quality complete co-assembly used as type material for the description of the novel species "*Candidatus* Kordia photophila". This characterization does not only include a complete functional description of the metabolic potential of the novel species, but it also features a comparative genomics overview with the available genomes of the Kordia genus and a thorough analysis of the distribution and abundance of the species in the water column from where it was sampled. Following the recommended procedure for the description of novel uncultured species, it is meant to serve as an example on the curation of genome assemblies to serve as type material and their complete characterization. In the third chapter, I shift to defining a co-assembly strategy of multiple metagenomes from the Arctic waters, as a means to reconstruct the genomes with a greater ecological impact in this vital ecosystem, as complete as possible in order to spot key players and define their biogeographical patterns, niche breadth and metabolic potential and activity. The methodological challenges of this chapter have been to perform a quality control of the assemblies and their read recruitments from a pangenomic point of view.

An overview of the objectives and how they have been approached in the different chapters is presented here:

Objective 1: to curate the methodological procedures regarding genome assembly, read recruitment and their quality controls, to serve as type material and generate reliable ecological conclusions.

Single Cell Genomics applied to environmental samples started in the last decade, and today, genome assemblers perform fairly reliable assemblies with algorithms that contemplate the differential genome coverage produced during genome amplification, that is afterwards sequenced. SAGs are usually incomplete, due to the nature of the amplification process, and strategies like the co-assembly of multiple SAGs can resolve the generation of a complete genome, composite of a population of close genomes. Nevertheless, a pool of multiple SAGs adds microdiversity, a notable sequence redundancy and an even more differential genome coverage, increasing computing time and memory needs, and producing shorter contigs due to variability hotspots.

In Chapter 2, an extensive study of assembly approaches is performed with the following aims:

- To generate a co-assembled genome as complete as possible, for the proper characterization of the novel species *Candidatus* Kordia photophila.
- To reduce the effects of microdiversity in the co-assembly process, that is, to obtain longer and less redundant contigs.
- To improve computation efficiency during co-assembly.

Genome reconstruction from metagenomes is a strategy for the recovery of uncultured genomes from environmental samples. In the marine environment, it has been applied to the Tara Oceans Expedition that sampled the temperate ocean from surface waters down to the mesopelagic layer, in the form of assemblies from individual samples (Tully et al., 2018a) and also co-assembling multiple samples based on their geographical distribution (Delmont et al., 2018). Many ecological analyses rely on the abundance of species, that in the metagenomics era has been extrapolated to the amount of reads that are recruited by the reconstructed genomes. Quality control of mappings has been set at the read alignment level, but quality thresholds at the whole genome level are scarce or inexistent.

In Chapter 3, I have defined a methodological workflow aiming:

- To define ecologically meaningful groups of samples to be co-assembled, in order to retrieve a less redundant dataset of genomes and with higher completeness.
- To find a quality control of read mappings in order to consider a genome present in a certain sample, for an increased reliability of the derived ecological conclusions.
- To understand the taxonomic units that the built MAGs are representing.

Objective 2: to gain insight into the taxonomic and functional diversity of uncultured prokaryotes from the North Indian Ocean and the Arctic Ocean.

The Arabian Sea, in the North Indian Ocean is subject to seasonal diatom blooms following coastal upwellings and contains the second-most intense oxygen minimum zone (OMZ) in the tropical oceans in the world. Prokaryotic taxonomic diversity in the water column of the Arabian Sea has been scarcely assessed by metagenomics, and their functional repertoire remains unexplored.

In Chapter 1, I examine a population of 98 Single Amplified Genomes from the same genus Kordia, retrieved from surface seawater in the North Indian Ocean during the Tara Oceans expedition. The aims are:

- To phylogenetically classify the putative novel species.
- To examine their microdiversity patterns by sequence comparison of several marker genes, in what is known as Multi Locus Sequence Amplification.
- To infer their putative niche using the sequenced genome of a SAG identical at the 16S rRNA gene level.

In Chapter 2, I co-assembled ten of the SAGs studied in chapter 1 with the following objectives:

- To confirm the taxonomic novelty of the species.
- To characterize the species preferred niche in the water column, combining a thorough description of its functional potential and its distribution in the different water layers and size fractions.

The Arctic Ocean is key in climate regulation of the planet and very sensitive to global warming, but our census of the prokaryotic diversity in its seawater is very poor. Knowing its key prokaryotic players is crucial for a better understanding of nutrient cycling in the arctic, which is vital for proper modelling of climate change projections. In 2013, the Tara Ocean Polar Circle Expedition circumnavigated the Arctic Ocean, providing a dataset of multiple metagenomic samples around the different marine protected areas in the Arctic Ocean and from different depths and seasons.

In Chapter 3, the reconstruction of MAGs from the Arctic Ocean metagenomic dataset from the Tara Oceans Polar Circle aims:

- To generate the first dataset of MAGs genomes representative of key uncultured arctic prokaryotic genomes.
- To explore their biogeographical and activity patterns.
- To define unique polar taxa and describe generalists and specialist polar prokaryotes.
- To establish their niche breadth, so that their resilience in a situation of climate change may be estimated.
- To explore their metabolic potential in terms of carbon fixation, and metabolism of sulfur, nitrogen and light.

CHAPTER 1

The content of this chapter was published in Frontiers in Microbiology, July 2017
with the title “Exploring microdiversity in novel *Kordia* sp. (Bacteroidetes) with proteorhodopsin from the tropical Indian Ocean via Single Amplified Genomes”
DOI: 10.3389/fmicb.2017.01317

CHAPTER 1: Exploring Microdiversity in Novel *Kordia* sp. (Bacteroidetes) with Proteorhodopsin from the Tropical Indian Ocean via Single Amplified Genomes

1.1 Abstract

Marine Bacteroidetes constitute a very abundant bacterioplankton group in the oceans that plays a key role in recycling particulate organic matter and includes several photoheterotrophic members containing proteorhodopsin. Relatively few marine Bacteroidetes species have been described and, moreover, they correspond to cultured isolates, which in most cases do not represent the actual abundant or ecologically relevant microorganisms in the natural environment. In this study, we explored the microdiversity of 98 Single Amplified Genomes (SAGs) retrieved from the surface waters of the underexplored North Indian Ocean, whose most closely related isolate is *Kordia algicida* OT-1. Using Multi Locus Sequencing Analysis (MLSA) we found no microdiversity in the tested conserved phylogenetic markers (16S rRNA and 23S rRNA genes), the fast-evolving Internal Transcribed Spacer and the functional markers proteorhodopsin and the beta-subunit of RNA polymerase. Furthermore, we carried out a Fragment Recruitment Analysis (FRA) with marine metagenomes to learn about the distribution and dynamics of this microorganism in different locations, depths and size fractions. This analysis indicated that this taxon belongs to the rare biosphere, showing its highest abundance after upwelling-induced phytoplankton blooms and sinking to the deep ocean with large organic matter particles. This uncultured Kordia lineage likely represents a novel Kordia species (*Kordia* sp. CFSAG39SUR) that contains the proteorhodopsin gene and has a widespread spatial and vertical distribution. The combination of SAGs and MLSA makes a valuable approach to infer putative ecological roles of uncultured abundant microorganisms.

1.2 Introduction

The phylum Bacteroidetes is the third most abundant group of bacteria in the oceans (Kirchman, 2002) but has been poorly studied at the species level compared to the two other main marine microbial phyla, i.e. the Proteobacteria and Cyanobacteria. Bacteroidetes is a cosmopolitan phylum that typically constitutes between 4 and 22% of marine bacterioplankton cells (Glöckner et al., 1999; Cottrell and Kirchman, 2000; Alonso-Sáez et al., 2007; Ruiz-González et al., 2012; Lefort and Gasol, 2013; Acinas et al., 2014). The relative abundance of Bacteroidetes can reach up to 53% and it often correlates with phytoplankton blooms (Abell and Bowman, 2005; van der Meer and Sentchilo, 2003; Fandino et al., 2005). Bacteroidetes are proficient at the degradation of particulate organic matter (POM) (Cottrell and Kirchman, 2000; Gómez-Pereira et al., 2012; Fernández-Gómez et al., 2013; Swan et al., 2013) and present a generally high growth rate when substrate is available (e.g. following phytoplankton blooms) (Kirchman, 2002; Ferrera et al., 2011; Buchan et al., 2014). Some representatives contain the gene coding for the light-dependent proton-pump proteorhodopsin, that has been shown to stimulate bacterial growth or work as a source of additional energy in the presence of light in some strains (Gómez-Consarnau et al., 2007; Stepanauskas and Sieracki, 2007; González et al., 2008; Woyke et al., 2009; Martinez-Garcia et al., 2012). Although Bacteroidetes can be found in the free-living plankton size fraction, they seem to be predominantly particle-attached (DeLong et al., 1993; Schattenhofer et al., 2009; Crespo et al., 2013; Díez-Vives et al., 2014; Salazar et al., 2015).

Population genetics studies of marine Bacteroidetes are scarce, and have been generally conducted using very small populations of isolates or comparing several distant species (Fernández-Gómez et al., 2013; Swan et al., 2013), giving a general overview of common genotypic traits and diferences above the species level. Single cell sequencing enables the retrieval of discrete microbial genomes from natural samples. A direct link between phylogenetic markers and functional information from single cells allows a finer resolution in microdiversity studies (Stepanauskas and Sieracki, 2007). When combined with multi locus sequencing analysis (MLSA), this approach can easily provide a robust phylogenetic reconstruction of the selected population. MLSA has been widely used for population genetics analyses since it is a feasible approach for delineating ecologically relevant units within species (Gevers et al., 2005). It is based on the sequencing of several marker genes and so far, its use in marine bacterial taxa has been limited to a few cultured groups such as *Alteromonas macleodii* (Ivars-Mart\'inez et al., 2008), members of *Vibrio* (Thompson et al., 2005) including *Vibrio cholerae* (Boucher et al., 2011), *Tenacibaculum maritimum* (Habib et al., 2014), or *Prochlorococcus* (Kashtan et al., 2014). Since it is well known that many ecologically relevant bacterial taxa elude isolation in culture (Rappé and Giovannoni, 2003), the combination of single cell genomics of environmental samples and MLSA can help in the exploration of intra-specific genetic variability as well as the different evolutionary processes involved in microbial speciation (such as homologous recombination and divergent selection).

The genus *Kordia* (Bacteroidetes, *Flavobacteriaceae*) contains representatives with isolation sources suggesting different lifestyles and habitats: (i) marine surface seawater for *Kordia antarctica* (Baek et al., 2013), *K. aquimaris* (Hameed et al., 2013), and *K. algicida* (Sohn et al., 2004), (ii) surface freshwater for *K. zhangzhouensis* (Lai et al., 2015), (iii) the interphase between the ocean and a freshwater spring for *K. jejudonensis* (Park et al., 2014), (iv) the surface of the green alga *Ulva* sp. for *K. ulvae* (Qi et al., 2016) or (v) the digestive tract of a marine polychaete for *K. periserrulae* (Choi et al., 2011).

Here we report the first MLSA analysis of the genus *Kordia*, performed with 98 Single Amplified Genomes (SAGs) from the genus *Kordia* that were retrieved from a seawater sample from the North Indian Ocean, a location subjected to seasonal monsoon winds and coastal upwelling events as well as open ocean oligotrophy. As a result, the microbial community structure varies and large phytoplankton blooms are common, especially those formed by diatoms (Landry et al., 1998). One characteristic feature of the North Indian Ocean is the marked Oxygen Minimum Zone (OMZ) layered in the subsurface of the upwelling regions (Roullier et al., 2014). Even though there have been studies of the bacterial community composition of the Indian Ocean seawater (Sunagawa et al., 2015b; Wang et al., 2016) and deep-sea sediments (Houbo Wu, 2011; Khandeparker et al., 2014), the diversity and genomics of the Bacteroidetes populations remain poorly characterized in this area. This is then a good opportunity to analyze whether micro-diversity exists in a particular Bacteroidetes taxon dominating after the phytoplankton bloom event. We used the well-conserved ribosomal RNA genes, the fastevolving internal transcribed spacer (ITS) and the beta subunit DNA-directed RNA polymerase rpoB. We also screened these genomes for a gene that can provide ecological information of the targeted population: proteorhodopsin (PR), which is present in ecologically relevant Bacteroidetes genomes (González et al., 2008; Pinhassi et al., 2016). Fragment Recruitment Analysis (FRA) of metagenomic reads suggested the preferred habitat and an estimated abundance of the novel *Kordia* sp. CFSAG39SUR in different oceans, depths and size fractions.

1.3 Materials and methods

1.3.1 Sampling, generation and selection of Single Amplified Genomes

Surface (SUR) seawater samples (5 m deep, not pre-filtered, Sample ID: TARA_G000000266) were collected for single cell analysis from the North Indian Ocean during the circumnavigation expedition Tara Oceans, station TARA_039 (Figure 3). The environmental setting of the station was inferred from its physical report and can be found at Pangaea website in the following link[1].

Replicated 1 mL aliquots of seawater were cryopreserved with 6% glycine betaine (Sigma) and stored at -80ºC. Samples used in this study were collected from TARA_039 (SUR; 5 m deep) located in the North Indian Ocean (18.59–66.62) on March 18th, 2010 (Table S1). SAGs generation

[1] http://store.pangaea.de/Projects/TARA-OCEANS/Station_Reports/TARA_039_oceanographic_context_report.pdf

and preliminary 16S rRNA screening with primers 27F and 907R (Table S2) were performed at the Bigelow Laboratory Single Cell Genomics. More information can be found in (Swan et al., 2011). Partial 16S rRNA gene sequences (640–850 bp) were compared to previously deposited sequences using the RDP Naive Bayesian rRNA Classifier Version 2.4 tool, and 98 SAGs with >99% similarity to the reference strain *K. algicida* OT-1 (AY195836) were selected for further analyses. These represented 84% of successfully amplified SAGs (total of 117 SAGs). They were named CFSAG39SUR referring to their taxonomic assignment (Cytophaga-Flavobacteria), retrieval method (SAG) and sampling station (TARA_039, SUR). Other SAGs used in this study (AAA285-F05 and AAA242-P21) were collected at 3,000 m at 4,800 m, respectively, in the North Pacific Sub-tropical Gyre. They belong to the Hawaii Ocean Time- Series (HOT, Figure 3A) and were generated following the same procedures.

1.3.2 Amplification of phylogenetic markers 16S rRNA, ITS and 23S rRNA Genes

Amplification of the 16S rRNA gene, ITS and 23S rRNA gene was done by a single Polymerase Chain Reaction (PCR) amplification with the 358F and CF434R primers (Table S2). Amplification was performed in a Biorad Thermocycler by an initial denaturation at 94ºC for 5 min, followed by 40 cycles of 94ºC for 1 min, 55ºC for 1 min and 72ºC for 2 min, and a final extension at 72ºC for 10 min. Each amplification reaction contained: 3 to 20 ng of template DNA, dNTPs (200 µM each), MgCl2 (2 mM), primers (0.5 µM each), Taq DNA polymerase (1.25 U), the PCR buffer supplied by the manufacturer (Invitrogen, Paisley, United Kingdom) and MilliQ water up to the final volume of 25 µl.

1.3.3 Amplification of functional genes: *rpoB* and proteorhodopsin

A set of newly designed primers was used for rpoB amplification (Table S2). Each PCR reaction followed an initial denaturation at 94ºC for 5 min, 40 cycles of 94ºC for 1 min, 40ºC for 45 s and 72ºC for 1 min, and a final extension at 72ºC for 5 min. Each amplification reaction contained: 1.5 to 10 ng of template DNA, dNTPs (0.2 mM each), MgCl2 (1.5 mM), primers (0.25 µM each), KAPA2G Robust HotStart DNA Polymerase (1 U), KAPA2G Buffer A (KAPA BIOSYSTEMS, Wilmington, MA, United States) and MilliQ water up to the final volume of 25 µl. For proteorhodopsin (PR) amplification, the primers used and the PCR protocol were those described by (Yoshizawa et al., 2012) (Table S2) using the KAPA2G Robust HotStart DNA Polymerase and KAPA2G Buffer A from KAPA BIOSYSTEMS. Each PCR reaction followed an initial denaturation at 94ºC for 5 min, 35 cycles of 94ºC for 1 min, 44ºC for 45 s and 72ºC for 1 min, and a final extension at 72ºC for 5 min. Each amplification reaction contained: 3 to 20 ng of template DNA, dNTPs (0.2 mM each), MgCl2 (3 mM), primers (0.5 µM each), Taq DNA polymerase (1.25 U), the PCR buffer supplied by the manufacturer (Invitrogen, Paisley, United Kingdom) and MilliQ water up to 25 µl.

1.3.4 Sequencing and phylogenetic analysis

Polymerase Chain Reaction (PCR) products were purified and sequenced by Genoscreen (Lille, France) with OneShot Sanger sequencing and aligned against NCBI's nBLAST and xBLAST databases for identification. Reference sequences used in this study were retrieved from Integrated Microbial Genomes and Metagenomes (IMG) 4 Data Management or NCBI database. Using Geneious Pro 4.8.5 software all sequences were aligned (using default Geneious progressive pairwise aligner) trimmed and manually checked. The software was also used to obtain the complete 16S rRNA gene assembling the amplicon sequences from MDA screening and the phylogenetic markers amplification. Alignments were processed into phylogenetic trees with Mega 5.2.2 Software choosing Maximum Likelihood statistical methods along with 1000 bootstrap replications. The best-fit substitution models for Maximum Likelihood phylogenies of the studied markers were: GTR with Gamma distribution (16S rRNA gene), Kimura- 2 with Gamma distribution (23S rRNA gene) and WAG with Gamma distribution (PR and RpoB). Representatives from the Proteobacteria and Cyanobacteria phyla were used as outgroups in all phylogenies. Phylogeny reconstruction of functional genes used the closest sequences to the SAGs' genes obtained from the Ocean Microbiome Reference Gene Catalog (Sunagawa et al., 2015a).

1.3.5 Fragment Recruitment Analysis (FRA)

Nucleotide-Nucleotide BLAST 2.2.28+ was used to recruit metagenomic reads from several metagenomic samples similar to all the available sequenced Kordia genomes. Considering the aim of this part ofthe study, which was to determine the presence and distribution of the novel *Kordia* sp. CFSAG39SUR in different oceanic regions, depths and size fractions, a database containing all 4 *Kordia* genomes was generated. This ensured a competitive recruitment between the genomes for each metagenomic sample analyzed. The quality and genome completeness of the four reference genomes in the database was measured with software checkM (Parks et al., 2015) and fetchMG (Sunagawa et al., 2013) (Table S3). The available genomic sequences of *Kordia* AAA285-F05 did not code for any of the marker genes or COGs used by the software, resulting in a genome completeness estimation of 0% despite being 283.58 Kb long. The other three reference genomes (*K. algicida*, *K. jejudonensis*, and *K. zhangzhouensis*) were estimated to be complete and free of contamination. Recruitment regarded only one read per target gene (-max_target_seqs 1), an identity percentage higher than 70% (-perc_identity 70), as well as an e-value lower than 0.000001 (-e-value 0.000001). All other nBLAST settings were set on default. In order to avoid random alignments, a filtering process was applied using R software (**?**) (i) excluding alignments with coverage lower than 90% of the sequence length, (ii) removing duplicated reads within each genome and (iii) masking reads belonging to the ribosomal operon. This operon does not follow the species delimitating recruitment patterns as the rest of the genome does and its presence would overestimate read recruitment (Caro-Quintero and Konstantinidis, 2012). All contigs from each genome were concatenated to a single sequence and the file containing the four sequences (one for each genome) was used to generate the database. This step was done using NCBI's makeblastdb application. As query we selected those metagenomic samples which contained the most metagenomic reads 16S rRNA and metagenomic Illumina tags (miTags) (Logares et al., 2013) annotated as *Kordia* (data not shown) using SILVA database Release 115 (Quast et al., 2013). These were stations TARA_039 (0.2–20 µm) and TARA_085 (0.2–3 µm) from Tara Oceans (Table S1 and Figure 3A) (Sunagawa et al., 2015a). In addition, we used 4 metagenomes from the deep waters of Malaspina 2010 circumnavegation expedition (Acinas et al., in preparation), two of them from the Brazil Basin in the Atlantic Ocean and two of them from Circumpolar Deep Waters (Table S1 and Figure 3A). Due to the flexibility of Bacteroidetes' lifestyle, the metagenomic selection was done from a limited pool of deep ocean metagenomes covering the free-living fraction of prokaryotes (0.2–0.8 µm) and the fraction of prokaryotes attached to particles or aggregates (0.8–20 µm) of a same seawater sample. The coverage of different deep ocean biomes and the highest abundance of Kordia related metagenomic read recruitments was also taken into account for the metagenomes' selection for the study. All queries consisted on merged, paired-end reads with different sequencing depths. Merging was done using Mothur v.1.33.3 (Schloss et al., 2009). Recruited data normalization was done: (i) by reference genome size (resulting in recruited readsgenomic bp), then (ii) upscaling this ratio to 1 Mb (as done in Swan et al., 2011), and finally (iii) by metagenomic sequencing depth. As sequencing depths varied between metagenomes, normalization was performed to the the smallest metagenomic sequencing depth of the study.

1.3.6 Accession numbers

Accession numbers for PCR products of *Kordia* CFSAG39SUR: 16S rRNA gene (MF187452), ITS (MF187454), 23S rRNA gene (MF187453), rpoB (MF187456), and proteorhodopsin (MF187455). Accession numbers for partial 16S rRNA gene sequences are: MF187458 for AAA285-F05 and MF187457 for AAA242-P21. The genomes used as reference for the FRA are the following: Bacteroidetes bacterium SCGC AAA285-F05 (JGI GoldStamp Id Go0060979), *K. algicida* strain OT-1 (NCBI RefSeq NZ_ABIB00000000.1), *K. jejudonensis* strain SSK3-3 (NCBI RefSeq NZ_LBMG00000000.1), *K. zhangzouensis* strain MCCC 1A00726 (NCBI RefSeq NZ_LBMH00000000.1). The Tara Oceans metagenomes used as query are the following (INSDC run accession numbers): TARA_039_DCM_0.22-1.6 (ERR599145), TARA_039_ MES_0.22-1.6 (ERR599037| ERR599172), TARA_085_SRF_0.22-3 (ERR599090| ERR599176), TARA_085_DCM_0.22-3 (ERR599104| ERR599121), TARA_085_MES_0.22-3 (ERR599008| ERR599125). The Expedición Malaspina metagenomes are the following (genome.jgi.doe.gov/): MP1493 (/DeeseametaMP1493_FD), MP1494 (/DeeseametaMP1494_FD), MP0326 (/DeeseametaMP0326_FD), MP0327 (/DeeseametaMP0327_FD) with the previous agreement of the PI of the JGI CSP 602 grant.

1.4 Results

1.4.1 Environmental setting and bacterial diversity of Station TARA_039

On February 2010, one month prior to sampling, a moderate bloom of phytoplankton occurred in the Oman Gulf due to a coastal upwelling under strong wind conditions. It propagated to the central basin of the Arabian Sea reaching only up to station TARA_039 (see Figure 3 in Roullier et al., 2014), which was the last mesotrophic station occupied by the Tara Oceans expedition before entering the warm oligotrophic waters of the central basin. Previous stations occupied during this leg in the North Indian Ocean (TARA_037 and TARA_038) were in the cooler waters of the upwelling, whereas station TARA_039 was at a mesoscale frontal region with the oligotrophic zone (Figure 3B). Particles produced during the phytoplankton bloom decreased in size as the cruise reached the central basin (see Figure 10 in Roullier et al., 2014). Diatom contribution to total phytoplankton decreased from 20 to 5% from stations close to the Gulf of Oman down to station TARA_40, outside the upwelling waters. Sunagawa et al. (2015a) found that the bacterial assemblage at station TARA_039 was dominated by Proteobacteria in both surface and DCM (Deep Chlorophyll Maximum) layers, followed by Cyanobacteria and other, less abundant groups including Euryarchaeota, Actinobacteria, Deferribacteres, and Bacteroidetes (Sunagawa et al., 2015a). In the mesopelagic (MES) zone there was a decrease in the abundance of Proteobacteria, a small increase of Deferribacteres and Actinobacteria and a remarkable increase in Thaumarchaeota and unclassified Bacteria. There was an increase in Bacteroidetes abundance toward the end of the Tara Oceans' sampling leg in the North Indian Ocean (Sunagawa et al., 2015a).

1.4.2 Phylogenetic reconstruction of a natural *Kordia* population

Amplification of the three phylogenetic markers of the ribosomal operon was successful in 78 out of 98 *Kordia* SAGs, and their alignment indicated a 100% identity among them. The resulting sequences contained full 16S rRNA gene (1,495 bp), ITS-1 (536 bp) and partial 23S rRNA gene (415 bp). The alignment of the resulting full 16S rRNA gene sequence against NCBI database assigned the SAGs to the genus Kordia, family *Flavobacteriaceae*. Its best hit with a nucleotide identity of 99% and sequence coverage of 97% was an uncultured Bacteroidetes sampled in the Puerto Rico Trench (clone PRTBB8540, acc. HM798967) at 6,000 m depth. The closest sequenced cultured relatives were *K. algicida* strain OT-1 with a 96.8% of nucleotide identity and 99% coverage (complete sequence, acc. NR027568) and *K. jejudonensis* strain SSK3-3 with the same identity percentage but 96% coverage (partial sequence, NR_126287). A 100% nucleotide identity was found with partial 16S rRNA gene sequences from 17 SAGs collected at 3,000 m (AAA285) and one SAG from 4,800 m depth in the North Pacific Sub-tropical Gyre (AAA242-P21), which belong to the Hawaii Ocean Time-series (HOT) (Stepanauskas, unpublished results). The phylogenetic reconstruction based on the representative sequences of the 16S rRNA genes from the 78 SAGs (Figure 4) strongly indicated that they belonged to the genus *Kordia* (phylum Bacteroidetes, family

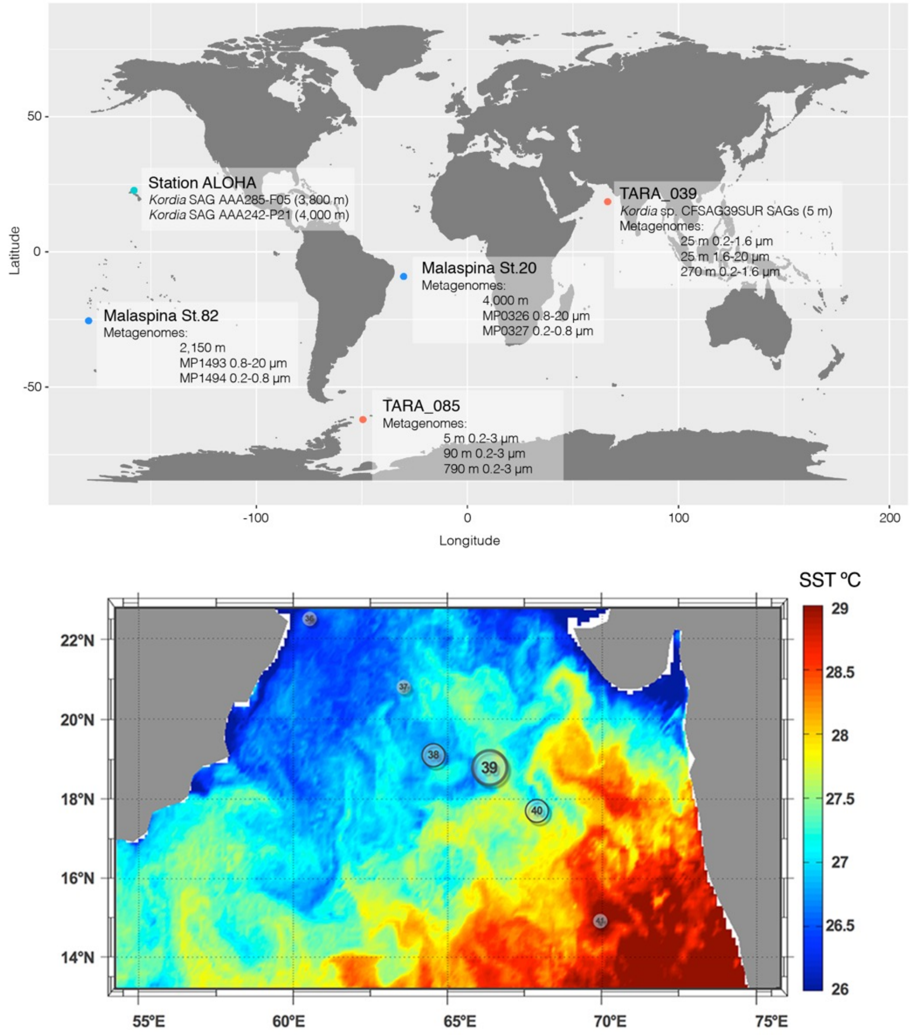


Figure 3. Global map showing the location of all relevant SAGs and metagenomic samples used in this study (A) and Sea Surface Temperature of Tara Oceans stations visited in the north Indian Ocean leg. (B) Station TARA_039 was where the *Kordia* SAGs where retrieved from and is located at a mesoscale front between the warm waters of the Indian Ocean basin and the cooler waters from the coastal upwelling.

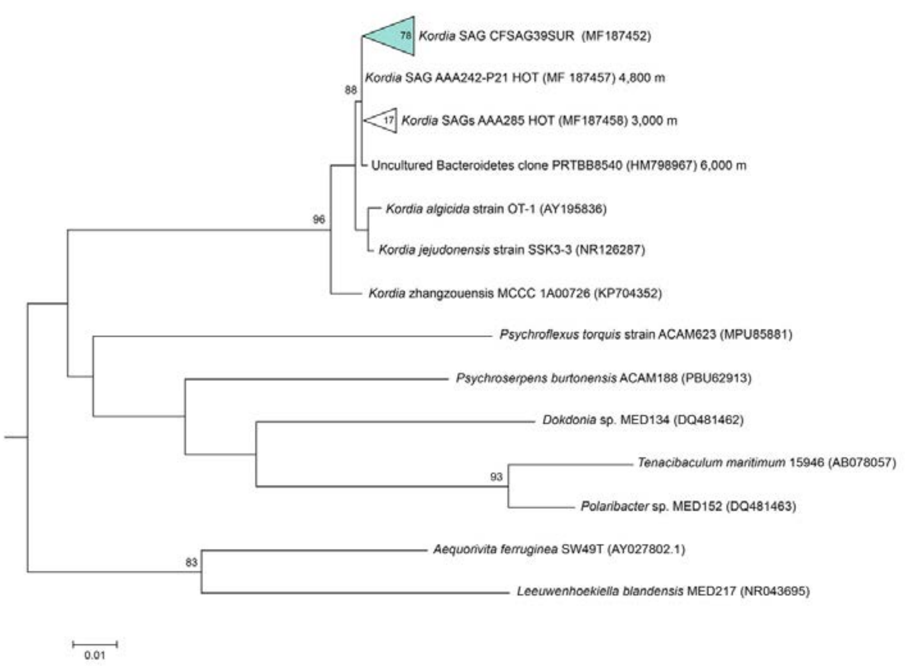

78
Kordia SAG CFSAG39SUR (MF187452)
Kordia SAG AAA242-P21 HOT (MF 187457) 4,800 m
88
17
Kordia SAGs AAA285 HOT (MF187458) 3,000 m
Uncultured Bacteroidetes clone PRTBB8540 (HM798967) 6,000 m
96
Kordia algicida strain OT-1 (AY195836)
Kordia jejudonensis strain SSK3-3 (NR126287)
Kordia zhangzouensis MCCC 1A00726 (KP704352)
Psychroflexus torquis strain ACAM623 (MPU85881)
Psychroserpens burtonensis ACAM188 (PBU62913)
Dokdonia sp. MED134 (DQ481462)
93
Tenacibaculum maritimum 15946 (AB078057)
Polaribacter sp. MED152 (DQ481463)
83
Aequorivita ferruginea SW49T (AY027802.1)
Leeuwenhoekiella blandensis MED217 (NR043695)
0.01

FIGURE 4. Maximum likelihood tree based on partial 16S rRNA gene sequences (805 bp), showing relationships between the representative sequence of the identical 16S rRNA gene from the *Kordia* SAGs (CFSAG39SUR) and members of the Bacteroidetes phylum. Only bootstrap values ≥70% are shown at the nodes. All the sequences have been retrieved from the JGI IMG database with the exception of AAA242-P21 and AAA285 that are unpublished SAG sequences.

Flavobacteriaceae). High bootstrap values supported the cluster formed by the 78 CFSAG39 SAGs with the 17 AAA285-SAGs, the one AAA242-P21 SAG and the uncultured Bacteroidetes clone PRTBB8540. This cluster's sequences were all retrieved from deep waters except for the SAGs reported here. The resulting sequence from the 78 identical partial 23S rRNA genes, when aligned against the NCBI database, showed a lower identity against *K. algicida* OT-1 (90.5%). Nevertheless, it was still its closest culture hit and the phylogenic tree supported the assignment of the SAGs to the genus *Kordia* ITS with high bootstrap values (Figure S1).

The 78 SAGs' ITSs were identical and coded for Ala-tRNA and Ile-tRNA. These tRNAs showed only one polymorphism each when compared to those of *K. algicida*. The nucleotidic change was located at the next-to-last base, not affecting the structure of the functional molecule. The box A conserved region present at the end of the ITS was identical to that of *K. algicida* although the former was located slightly upstream (17 bp) (Figure S2).

Amplification of the beta subunit of RNA polymerase (*rpo*B) was successful in 18 SAGs, obtaining 1̃,400 bp amplicons with 100% of both amino acid and nucleotide identity among them. Their closest hit in NCBI database was the *rpo*B gene of *K. algicida* OT-1 with a 98% amino acid identity. High bootstrap values supported the location of the SAGs' *rpo*B sequence in the phylogenetic tree, within the *Flavobacteriaceae* cluster and closest to *K. algicida* OT-1 (S3). An amplification summary can be found in Table 1.

TABLE 1. Summary on genetic information of Kordia Single Amplified Genomes (SAGs) and closest SAGs, isolates and clones. Identity % against *Kordia* sp. CFSAG39SUR's corresponding genes (16S, ITS and 23S) and proteins (RpoB and PR). For each gene and protein, "# seq" refers to the number of sequences available for comparison

Kordia sequence information	Origin	Depth (m)	16S		ITS		23S		*rpo* B		PR	
			# seq.	id %	#seq	id %	#seq	id %	#seq	id %	#seq	id %
CFSAG39SUR (98 SAGs)	North Indian Ocean	5	78	100	78	100	78	100	18	100	34	100
AAA242-P21 (1 SAG)	NPSG ALOHA	4000	1	100	-	-	-	-	-	-	-	-
AAA285 (17 SAGs)	NPSG ALOHA	3800	1	100	-	-	-	-	-	-	-	-
Kordia algicida OT-1 (1 isolate)	Masan Bay, S.Korea	0	3	96.8	3	76.1	3	90.5	1	89.3	0	-
Kordia sp. PC-4 (1 isolate)	Sagami Bay, Japan	100	1	98.4	-	-	-	-	-	-	1	89.2
Unc. Bact. PRTBB8540 (1 clone)	Puerto Rico Trench	6000	1	99.9	-	-	-	-	-	-	-	-

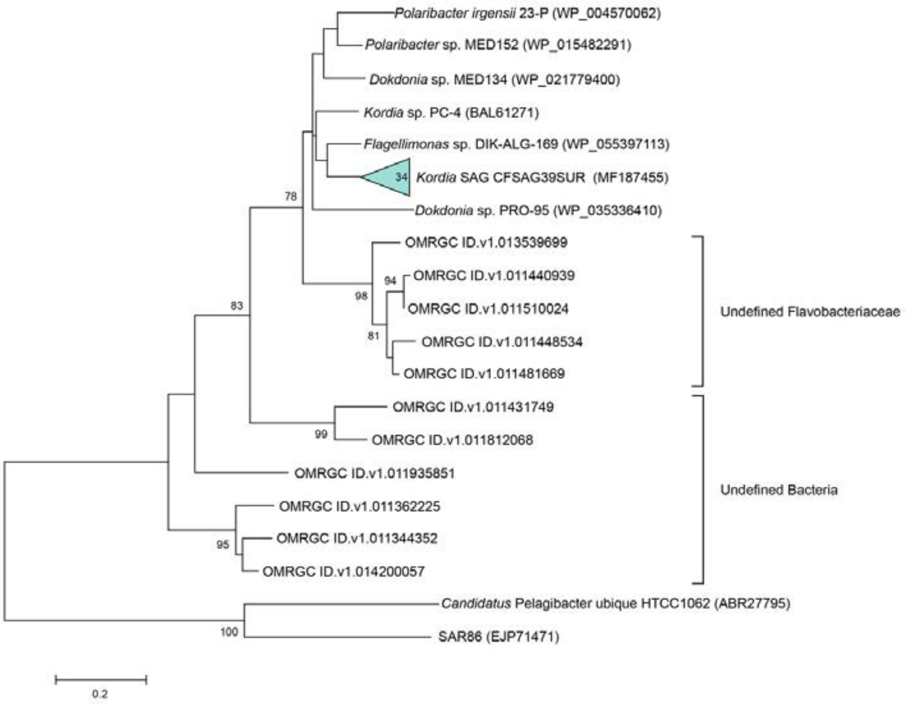

Polaribacter irgensii 23-P (WP_004570062)
Polaribacter sp. MED152 (WP_015482291)
Dokdonia sp. MED134 (WP_021779400)
Kordia sp. PC-4 (BAL61271)
Flagellimonas sp. DIK-ALG-169 (WP_055397113)
34
Kordia SAG CFSAG39SUR (MF187455)
Dokdonia sp. PRO-95 (WP_035336410)
78
OMRGC ID.v1.013539699
94
OMRGC ID.v1.011440939
98
OMRGC ID.v1.011510024
81
OMRGC ID.v1.011448534
OMRGC ID.v1.011481669
83
OMRGC ID.v1.011431749
99
OMRGC ID.v1.011812068
OMRGC ID.v1.011935851
OMRGC ID.v1.011362225
95
OMRGC ID.v1.011344352
OMRGC ID.v1.014200057
Candidatus Pelagibacter ubique HTCC1062 (ABR27795)
100
SAR86 (EJP71471)
Undefined Flavobacteriaceae
Undefined Bacteria
0.2

FIGURE 5. Maximum likelihood tree based on partial proteorhodopsin amino acid sequences (167 aa), showing relationships between the representative sequence of the 34 identical PR genes from the *Kordia* SAGs (CFSAG39SUR), the closest hits according to NCBI and the Ocean Microbiome Reference Gene Catalog. Other sequences were retrieved from the JGI IMG database. Candidatus Pelagibacter ubique and SAR86 act as outgroups. Only bootstrap values ≥70% are shown at the nodes.

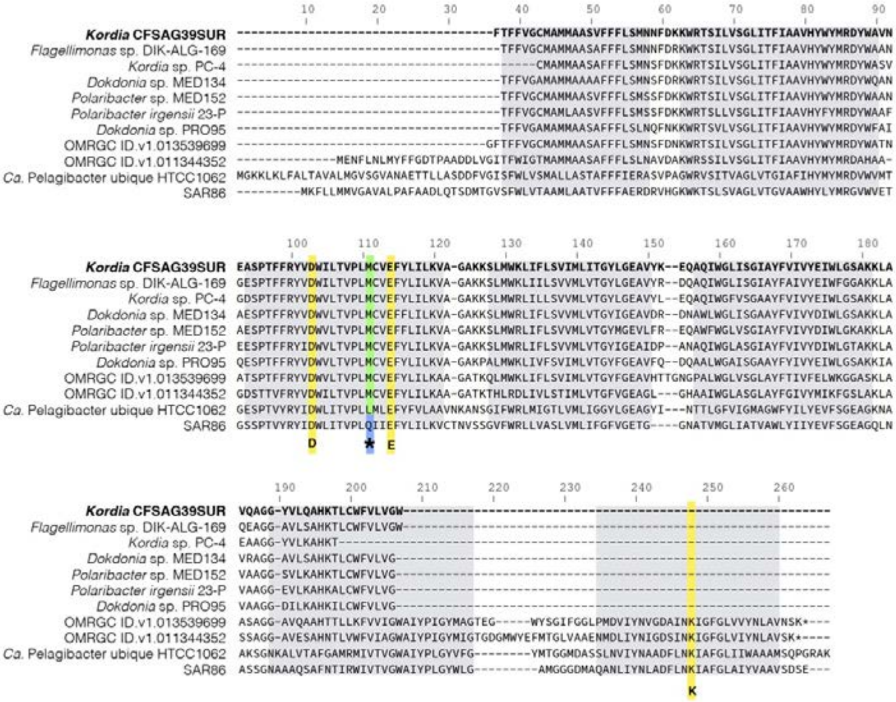


FIGURE 6. Amino acid alignment of full and partial proteorhodopsin representative sequences from Bacteroidetes (*Kordia* SAGs CFSAG39SUR sequence in bold at the top) and some outgroups. Key amino acids for proteorhodopsin functionality are highlighted in yellow: D (Asp) and E (Glu) are the proton acceptor and donor, respectively, conforming the Schiff base. K (Lys) is the amino acid to which retinal binds. The amino acid that plays a role in spectral tuning is marked with *. Methionine (M) and leucine (L) lead to absorption maximum of green light, between 518 and 535 nm whereas glutamine (Q) sets the absorption maximum of blue light, 490 nm. Predicted transmembrane helices are highlighted as gray boxes.

1.4.3 Novel proteorhodopsin gene

The PR gene was amplified in 34 out of the 98 *Kordia* SAGs. The resulting 167 amino acid sequences turned out to be 100% identical to one another. The sequence was closest to the *Flavobacteriaceae Flagellimonas* sp. DIK-ALG-169's proteorhodopsin (89% identity, 100% coverage) (AHN13811) and Kordia sp. PC-4 (Yoshizawa et al., 2012) (88% identity, 91% coverage) in the NCBI database. When comparing sequences in the Ocean Microbiome Reference Gene Catalog (OMRGC), the closest hit shared 84.2% amino acid similarity (unclassified Flavobacteria, NOG136807 ID.v1.022398661). Phylogenetic reconstruction placed the SAGs' proteorhodopsin in a cluster of *Flavobacteriaceae* PR genes (Figure 5). We tried to amplify the proteorhodopsin gene from the 17 SAGs retrieved from 3,000 m at the North Pacific Gyre (AAA285-F05) but the results were negative.

The SAGs' proteorhodopsin alignment with other closely related PR sequences showed that this proteorhodopsin had the same structural features as its relatives. It had the predicted transmembrane helices, key amino acids for functionality and the location of the spectral tuning amino acid, which in this case was methionine (Figure 6), indicating absorbance of light spectrum between 518 and 535 nm (green light).

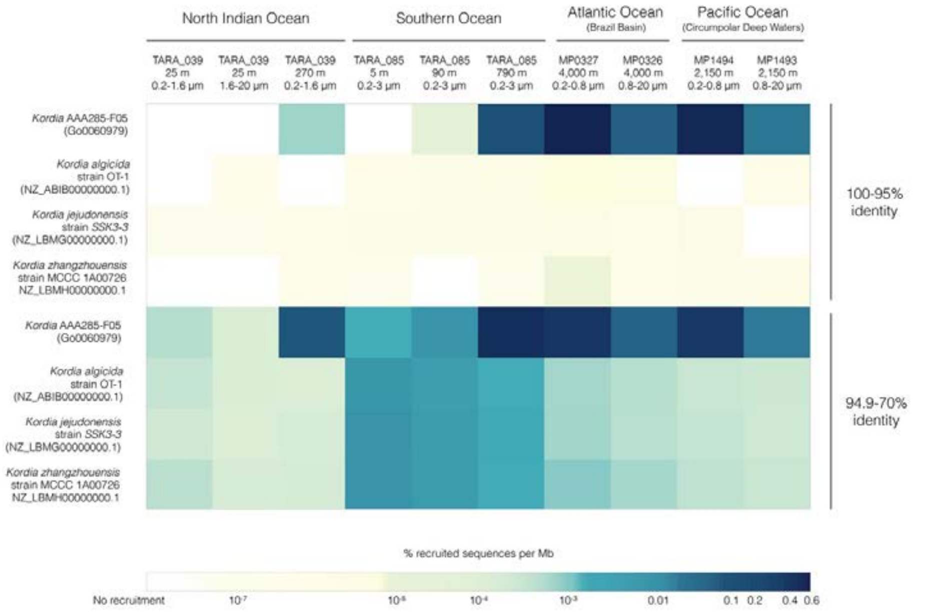


FIGURE 7. FRA results for the four available *Kordia* genomes mapped against 10 metagenomic samples from different locations, depths and size fractions. Heatmap picturing percentage of recruited reads per genomic Mb normalized by metagenomic sequencing depth and reference genome size. The upper four rows depict recruitment at 100-95% identity against reference genome, the bottom four rows depict recruitment identity values between 70 and 94.9%.

1.4.4 Distribution of novel *Kordia* sp. in different oceanic regions

Metagenomic reads from different stations were recruited through a competitive FRA with the available sequenced genomes of the genus *Kordia*: the deep-sea *Kordia* SAG AAA285-F05 (283.58 Kbp), the seawater surface isolated in culture *K. algicida* OT-1 (5.01 Mbp), the freshwater *K. zhangzhouensis* MCCC 1A00726 (4.03 Mbp) and *K. jejudonensis* SSK3-3 (5.3 Mbp), isolated from a region where spring freshwater and seawater meet. We counted reads with 70–100% identity to these reference genomes. For *Kordia* AAA285-F05, the number of recruited reads increased with depth in all stations tested. The percentage of reads recruited per genomic Mbp was several orders of magnitude larger in the free-living (Figure 7; numeric data in Table S4). Highest relative abundances of reads per Mbp with 95–100% identity for *Kordia* AAA285-F05's occurred in the free-living fraction of bathypelagic waters from the Brazil Basin (MP0327, 0.51%) and the Circumpolar Deep Waters of the Pacific Ocean (MP0326, 0.46%). Recruitment percentages for this SAG in these two locations decreased down to 0.04 and 0.01%, respectively. In TARA_039 highest recruitment % per Mbp was 0.0007% at 270 m, while in the Southern Ocean (TARA_085), recruitment increased with depth reaching 0.06% at 790 m. Recruitment was virtually zero for the isolates genomes at this recruitment identity percentage.

Recruitment reads with 70–94.9% identity might be indicative of populations related to, but different from, the reference genome present at the tested metagenomic sample. The recruitment trend observed in the higher identity percentage for *Kordia* SAG AAA285-F05 was consistent at the lower percentage as well. Recruitment % per genomic Mbp increased at all depths of both

TARA_039 and TARA_085, now reaching values of 0.05% at 270 m in TARA_039 and the maximal of 0.38% at 790 m in TARA_085. Similar abundances were obtained for the Brazil Basin (MP0327, 4,001 m) and the Circumpolar Deep Waters (MP1494, 2,150 m) in their free-living fraction (0.3 and 0.27%, respectively). Recruitment decreased in the larger size fraction of the two Malaspina 2010 circumnavigation expedition metagenomes down to 0.04 and 0.01%, respectively. For the other three Kordia genomes competing for recruitment in the same metagenomic samples, recruitment values decreased significantly to values ranging from 0.008% per Mb for *K. zhangzhouensis* in TARA_085 surface waters to values down to 10^{7}% per Mb. The recruitment for these three genomes followed a similar pattern, with highest values in TARA_085 surface decreasing with depth. In their case, bathypelagic samples and North Indian ocean's DCM (25 m) recruited similar values (0.0002%). The lowest values were found at TARA_039 mesopelagic waters (0.00001%).

FRA plots (Figure 8) showed a similar pattern of *Kordia* AAA285-F05 recruitment for the meso- and bathypelagic metagenomes (TARA_085_MES, MP0327, MP1494) of the free-living prokaryotic fraction. There was a very good coverage ofthe whole reference sequence, with the exception of some fragments where there was a visible decrease in the recruitment at all identity percentages. The highest read densities were found above 95% identity. In the TARA_039_OMZ recruitment plot, an overall decrease in read density was observed, identities ranged from 85 to 99%. The two genomic regions with very low recruitment were also apparent in this plot. The three reference *Kordia* genomes (S4) showed read clouds at the lower identity percentage in all cases. There was a good coverage of the length of the genome but not as intense as that observed in Figure 8 for reference genome *Kordia* AAA285-F05.

1.5 Discussion

The aim of this study was to explore the potential microdiversity within a population of 98 SAGs of the genus *Kordia* (*Kordia* sp. CFSAG39SUR), a *Flavobacteriaceae* genus established after describing *K. algicida* strain OT-1 (Sohn et al., 2004). Having these SAGs was a unique opportunity to study the extent of microdiversity within a large population of environmental uncultured organisms, since genomic length and composition of marine uncultured bacterial genomes (SAGs) generally differ from those of similar taxa isolated in pure culture (Swan et al., 2013).

Phylogenetic proximity among *Kordia* sp. CFSAG39SUR SAGs was confirmed by the 100% identity of the full 16S rRNA gene, partial 23S rRNA gene, and the ITS-1 region. The latter is located between the 16S and 23S ribosomal genes and has been widely used when more detailed taxonomical resolution was needed, especially in intra-specific diversity studies (Brown and Fuhrman, 2005). Moreover, its variability in length and composition highlights initial genome diversification and evolutionary speciation (Schloter, 2000). The 100% identity at the nucleotide level of functional genes *rpo*B and PR (exposed to a high degree of horizontal gene transfer, Fuhrman et al., 2008) could suggest that the genetic homogeneity may be present throughout the whole genome. However, only the analysis of the whole genomes would confirm it or reject this hypothesis.

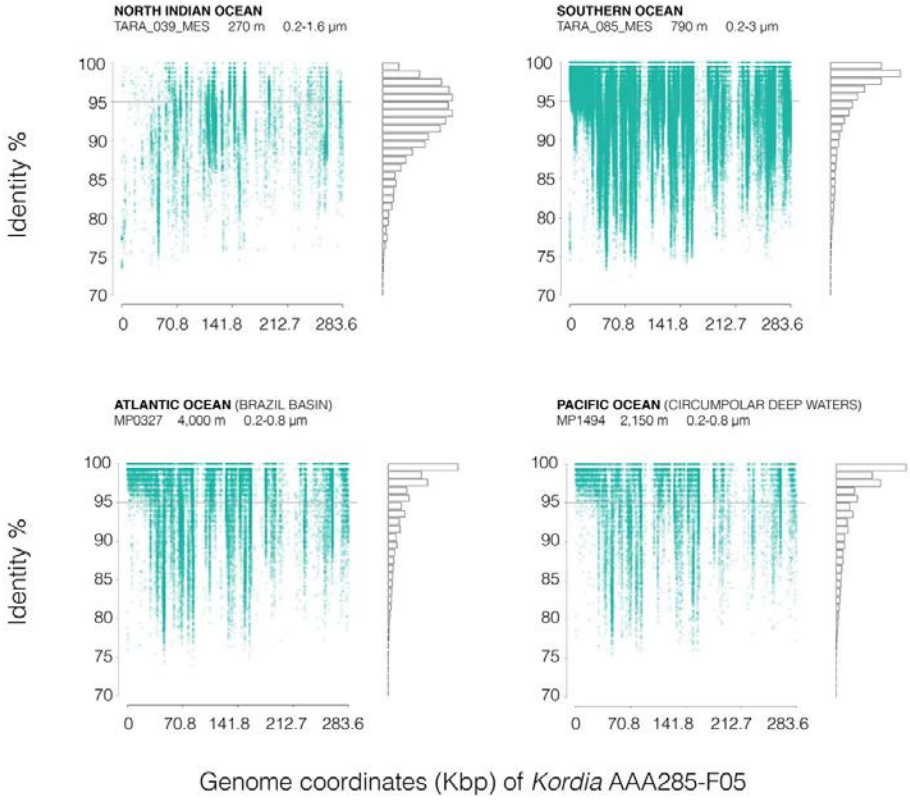


FIGURE 8. FRA plots of *Kordia* SAG AAA285-F05 partial genome (283.58 kb), against (A) two different stations from the Tara Oceans metagenomes (TARA_039_MES and TARA_085_MES) and (B) two metagenomes from the Malaspina 2010 circumnavigation expedition deep metagenomes (MP0326 at 4,000 m deep and MP1494 at 2,150 m deep). FRA has been done competitively between the four available Kordia genomes. Metagenomic reads are displayed according to their position against de reference genome (X-axis, genome coordinates) and their identity percentage (Y-axis). The gray line located at 95% identity in the Y-axis corresponds to 95% average nucleotide identity (ANI), threshold for bacterial species categorization. The bars on the right show the amount of reads mapped at each identity percentage. Duplicates and reads mapped to the ribosomal operon have been removed to avoid bias.

Such genetic similarity among SAGs can be attributed to two main factors: (i) to stable environmental conditions providing fewer mutation-prone events, hence low microdiversity in the population (Cohan, 2006), or (ii) recent population expansion,as nascent populations are usually more genetically homogenous than mature ones, in which products of mutations and recombination may accumulate (Cordero and Polz, 2014). As mentioned above, a month prior to the arrival of the Tara schooner to the North Indian Ocean, a coastal upwelling fertilized the surface waters of the Gulf of Oman with new nutrients, resulting in a phytoplankton bloom that extended through the coast of Iran and south-east to station TARA_039 (Roullier et al., 2014). These blooms have been described to be seasonal, mostly formed by diatoms (Landry et al., 1998). The large particles generated by the bloom were seen to decrease in size through the sampling period (Roullier et al., 2014). Backward Lagrangian particle transport modeling suggested that these particles sunk near station TARA_039 (Roullier et al., 2014), where a frontal system separated the colder upwelled waters from the warmer waters of the oligotrophic basin. Considering the occurrence of such an event before the sampling at TARA_039, it is highly possible that the genetic homogeneity found in the 78 *Kordia* sp. CFSAG39SUR SAGs is due to the second possibility mentioned above, that is, that these bacteria were retrieved as a nascent population derived from the phytoplankton bloom.

There is previous knowledge of the relationship between members ofthe genus Kordia and algae. The first isolate described for the genus, *K. algicida* OT-1, was isolated following a red tide of the diatom *Skeletonema costatum* (Sohn et al., 2004). Additionally, it is the only Bacteroidetes species found to code for R-bodies, a specific system for targeted lysis of eukaryotes (**?**). Likewise, *K. ulvae* was isolated from the surface of the marine alga *Ulva* sp. (Qi et al., 2016). Abundant algal products or exudates could provide a good environment for the rapid proliferation of our *Kordia* sp. CFSAG39SUR SAGs either attached to large particles or free-living in the water column. This would help to explain the apparent lack of microdiversity found at the time of sampling and that would probably be due to nascent populations from a single genotype after the algal bloom. Interestingly, a population of 18 SAGs identical to this *Kordia* sp. CFSAG39SUR population at the available partial 16S rRNA gene sequence were retrieved from the deep waters of station ALOHA in the North Pacific Subtropical Gyre, where upwelling and seasonal diatom blooms are common events (Villareal et al., 2012).

The *Kordia* sp. CFSAG39SUR SAGs made 84% of the sorted heterotrophic genomes retrieved from the sample, suggesting a high abundance of *Kordia* at the moment of sampling. Even though there was no metagenomic sample available from TARA_039 surface waters to support this high abundance and being aware of the environmental context of the sampling, we relied on FRA to test its persistence through the water column. Metagenomic read recruitment of *Kordia* spp. was very low in the TARA_039 DCM metagenome and slightly higher in TARA_039 MES metagenome. These values pointed to *Kordia* sp. CFSAG39SUR SAGs as members of the less abundant taxa known as the rare biosphere (Pedrós-Alió, 2012). Higher abundances (up to 0.51% of recruited reads per genomic Mbp) were found in other metagenomic samples from Tara Oceans and Malaspina 2010 circumnavigation expedition covering different ocean locations, depths (0–4,000 m) and size fractions (0.2–20 μm). In many cases read abundances reached values similar to recruitments in the surface waters of different oceans using as reference isolated Rickettsiales, *Planctomyces* or Rhodobacterales, generally considered to be abundant bacteria. Nevertheless, read abundances never reached those of isolated *Prochlorococcus*, *Synechococcus*, or *Pelagibacter*, the most abundant groups in seawater (see Supplementary Figures S10, S11 by Swan et al., 2011). The scarce recruitment in TARA_039 metagenomes contrasts with the high abundance of SAGs retrieved from the surface waters of the station. We think it is important to highlight the fact that the sampled water processed for single cell genomics was not prefiltered. Together with the slight increase in recruitment at the mesopelagic layer, we cannot reject the possibility that sequences related to *Kordia* sp. CFSAG39SUR SAGs could have been abundant in metagenomes of larger size fractions (20–200 μm) than those available (0.2–20 μm), especially since large particles (2–2.5 mm at 520–560 m, see Figure 10 in Roullier et al., 2014) were reported at Station TARA_039 deeper waters during the Tara expedition. It is possible then, that this new Kordia species becomes abundant only after upwelling-induced phytoplankton bloom events. To further test the preferred niche for *Kordia*, FRA should be performed on metagenomes of higher planktonic size fractions, either from the same stations or from stations where diatom blooms occur after upwelling episodes. Unfortunately, such metagenomes are not currently available.

This suggests a hypothesis where, after colonization of phytoplankton-derived large particles, *Kordia* sp. CFSAG39SUR would sink with the particles to the deep ocean, becoming part of the bacterial seed bank of the rare biosphere. In fact, FRA of reference genome *Kordia* SAG AAA285-F05 displayed higher read recruitments (i.e., relative read abundance per genomic Mbp) in metagenomes from deeper water samples. Intriguingly, our *Kordia* sp. CFSAG39SUR SAGs coded for a light dependent proteorhodopsin, a metabolic advantageous trait in sunlit water layers. The sequence was similar to those of other Flavobacteria which have been shown to absorb green-light, and have fast photocycles characteristic of proton-pumping rhodopsins rather than sensing rhodopsins (Spudich, 2006). This suggests that *Kordia* sp. CFSAG39SUR has an active metabolism in the upper layers of the ocean. Perhaps their activity decreases or becomes dormant when reaching bottom waters. There, they would accumulate attached to the remaining particles or switching to a free-living lifestyle, explaining the highest recruitments at the free-living fraction of meso- and bathypelagic waters. It must be considered, however, that proteorhodopsins have been found in the deep sea (Tang et al., 2015; Yoshizawa et al., 2015) and that both presence and expression ofproteorhodopsin genes were detected in the Arctic Ocean during the Polar night (Nguyen et al., 2015), suggesting that proteorhodopsins may have an unknown function in dark waters.

Comparative genomics has helped us infer a species threshold based on Average Nucleotide Identity (ANI, Konstantinidis and Tiedje, 2005) where 95% identity corresponds to the previous 70% DNA–DNA hybridization threshold, which was the gold standard for taxonomic species identification (Wayne et al., 1987). FRA on metagenomes (Caro-Quintero and Konstantinidis, 2012) using a reference genome makes it possible to infer: (i) similar genetic populations ("species") above >95% ANI and (ii) different sequence-discrete populations within the range of 80–95% ANI. For this study, as the *Kordia* sp. CFSAG39SUR genome sequences are not available, we carried out the analyses with the genome of *Kordia* SAG AAA285-F05 as reference because they shared 100% identity in their 16S rRNA genes. Despite this identity at the 16S rRNA level, there may be significant differences both at the gene content level, as well as in the identity between shared genes. It is common that within a given bacterial species there is a set of genes conserved among all strains for housekeeping functions (core genome), while an array of different functional genes can vary depending on the niche of the specific bacteria (flexible genome) (Medini et al., 2005; Fernández-Gómez et al., 2012). Despite this limitation, the fact that we used *Kordia* AAA285-F05 as our reference genome still may provide an indirect detection of *Kordia* sp. CFSAG39SUR SAGs through the FRA of the fraction of the genome shared between the reference and our SAGs. Since we do not know to what extent *Kordia* sp. CFSAG39SUR resembles *Kordia* SAG AAA285-F05 apart from the 16S rRNA gene in those shared genes, reads recruited between identity values of 95–100% would reveal the presence of *Kordia* belonging to the same species as *Kordia* AAA285-F05. If our *Kordia* sp. CFSAG39SUR on the other hand, was actually a co-occurrent relative of the reference but not the same ecotype (i.e., nucleotidic differences in those shared genes), we would indirectly detect it in those reads recruited below 95% identity (Caro-Quintero and Konstantinidis, 2012).

Being aware of this, we also used all available *Kordia* spp. genomes as reference for the FRAs at the same time. This analysis confirmed that those reads mapping against AAA285- F05's genome at lower identity ranges belonged, indeed, to a new ecotype from the same novel species (likely our *Kordia* sp. CFSAG39SUR SAGs). The non-existent recruitment of reads at 95–100% for those genomes different from *Kordia* AAA285-F05 backs it up strongly.

1.6 Conclusion

This study shows how state-of-the-art single cell genomics can be combined with more traditional techniques such as MLSA to obtain an overview of the genetic composition of the population of study without sequencing whole genomes beforehand. Thus, this can provide both an analysis of population genetics avoiding bias by isolation in culture and an approach to select genomes of interest for further population genomics and biogeography studies. In our case, through MLSA of SAGs we have found that the dominant heterotrophic bacterial taxon in the sample was genetically homogeneous. Moreover, the study of this population in combination with available metagenomic datasets and the oceanographic metadata of the sampling area has helped shedding light on the species distribution, dynamics, and potential ecological niche of a novel *Kordia* species that contains proteorhodopsin.

1.7 Acknowledgements

We thank our fellow scientists and the crew and chief scientists of the different cruise legs involved in Tara Oceans and Malaspina expedition for collecting the samples used in this study. This is Tara Oceans contribution paper number 57. The work described here is not to be associated with any policy of or sanctioned by the United States National Science Foundation.

CHAPTER 2

The content of this chapter was published in Systematic and Applied Microbiology in January 2020 with the title “Ecological and functional capabilities of an uncultured *Kordia* sp.”
DOI: 10.1016/j.syapm.2019.126045

CHAPTER 2: Functional and ecological capabilities of an uncultured *Kordia* sp.

2.1 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.

2.2 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 (Rosselló-Mora and Amann, 2001; Ochman et al., 2005; Fraser et al., 2009; Shapiro and Polz, 2014). 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 (Konstantinidis et al., 2017). Recent efforts have emerged to facilitate novel uncultured taxa standardization (Konstantinidis et al., 2017), like the Microbial Genome Atlas (MiGA), which infers genome-based taxonomy and quality assessment across genomes from different environments (Rodriguez-R et al., 2018).

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 (Parks et al., 2017; Delmont et al., 2018). 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) (Swan et al., 2011; Pachiadaki et al., 2017). 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 (Alneberg et al., 2018). SAGs have also been used in pangenome analyses, which had previously been mostly restricted to cultured microorganisms (especially pathogenic bacteria) (Medini et al., 2005; Tettelin et al., 2008; Mira et al., 2010). Comparative genomics and the development of the "pan-genome" concept offered an alternative for understanding the genetic extent and dynamics of bacterial species (Medini et al., 2005; Tettelin et al., 2008; Mira et al., 2010), dramatically changing the description of a species by studying multiple genomes belonging to the same defined taxa (Konstantinidis and Tiedje, 2005b; Richter and Rosselló-Móra, 2009; Kim et al., 2014). In the marine environment, SAG-based pangenome analyses have mostly focused on the highly studied *Prochlorococcus* (revealing hundreds of co-existing *Prochlorococcus* populations (Kashtan et al., 2014) and the link between their hypervariable genomic islands and the environment (Delmont and Eren, 2018) or together with SAR11 (defining endemic gene-level adaptations to specific locations like the Red sea (Thompson et al., 2019).

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 (Sohn et al., 2004). 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 (Kim et al., 2017). 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 (Lai et al., 2015), *K. jejudonensis* was isolated from the interphase between seawater and freshwater springs (Park et al., 2014), *K. antarctica* and *K. aquimaris* were isolated from Antarctic and Taiwanese seawater, respectively(**?**Hameed et al., 2013), *K. ulvae*, *K.zosterae* and *Kordia* sp. SMS9 were retrieved from the surface of marine algae (Qi et al., 2016; Kim et al., 2017; Pinder et al., 2019) and *K. periserrulae* was found in the gut of a tidal flat polychaete (Choi et al., 2011). *Kordia* sp. NORP58 is a MAG assembled from a marine subsurface aquifer (Tully et al., 2018b). 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.

2.3 Materials and methods

2.3.1 Single Amplified Genome generation and phylogeny analysis

A total of 98 Single Amplified Genomes were generated as detailed in (Swan et al., 2011) from a surface (SRF) seawater sample from the North Indian Ocean (latitude 18.59ºN – longitude 66.62ºE) during the circumnavigation expedition Tara Oceans (Karsenti et al., 2011), station TARA_039 (Sample ID: TARA_G000000266)(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 (Royo-Llonch et al., 2017).

2.3.2 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.

2.3.3 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 (Langmead and Salzberg, 2012) and Samtools v.1.3.1 (Li et al., 2009). 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 (Bolger et al., 2014) and paired-reads were merged with PEAR v0.9.6 (Zhang et al., 2014b). The workable read dataset consisted of merged reads, pairs of reads that did not merge and also unpaired orphan reads (Figure 9A). 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).

2.3.4 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 9A). The three k-mer values that generated co-assemblies with best metrics were combined using assembler SPAdes v3.10 (Bankevich et al., 2012) 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 (**?**). 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 (Clark et al., 2013). 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 9B). 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% (Bowers et al., 2017; Konstantinidis et al., 2017). Detailed metrics and evaluation of the different co-assemblies are in Table S6.

2.3.5 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 2).

Ggplot2 (Gómez-Rubio, 2017) was used to visualize the mapping of reads of every individual SAG against the generated co-assembly (Figure S6).

2.3.6 Phylogeny based on the 16S rRNA gene

The complete 16S rRNA gene of the co-assembly was queried against the Living Tree Project (Yarza et al., 2020) database using SILVA's SINA ARB v1.2.11 online tool (Pruesse et al., 2012). Its *Flavobacteriaceae* best hits were exported and aligned in Geneious R11, together with *Kordia* amplicon OTUs generated from the Malaspina dataset in (Mestre et al., 2018), the 16S rRNA genes of other Kordia spp. deposited in GenBank and two outgroups from a different phylum (Table S5). MEGA v7 (Kumar et al., 2016) was used for a Maximum-Likelihood phylogeny reconstruction

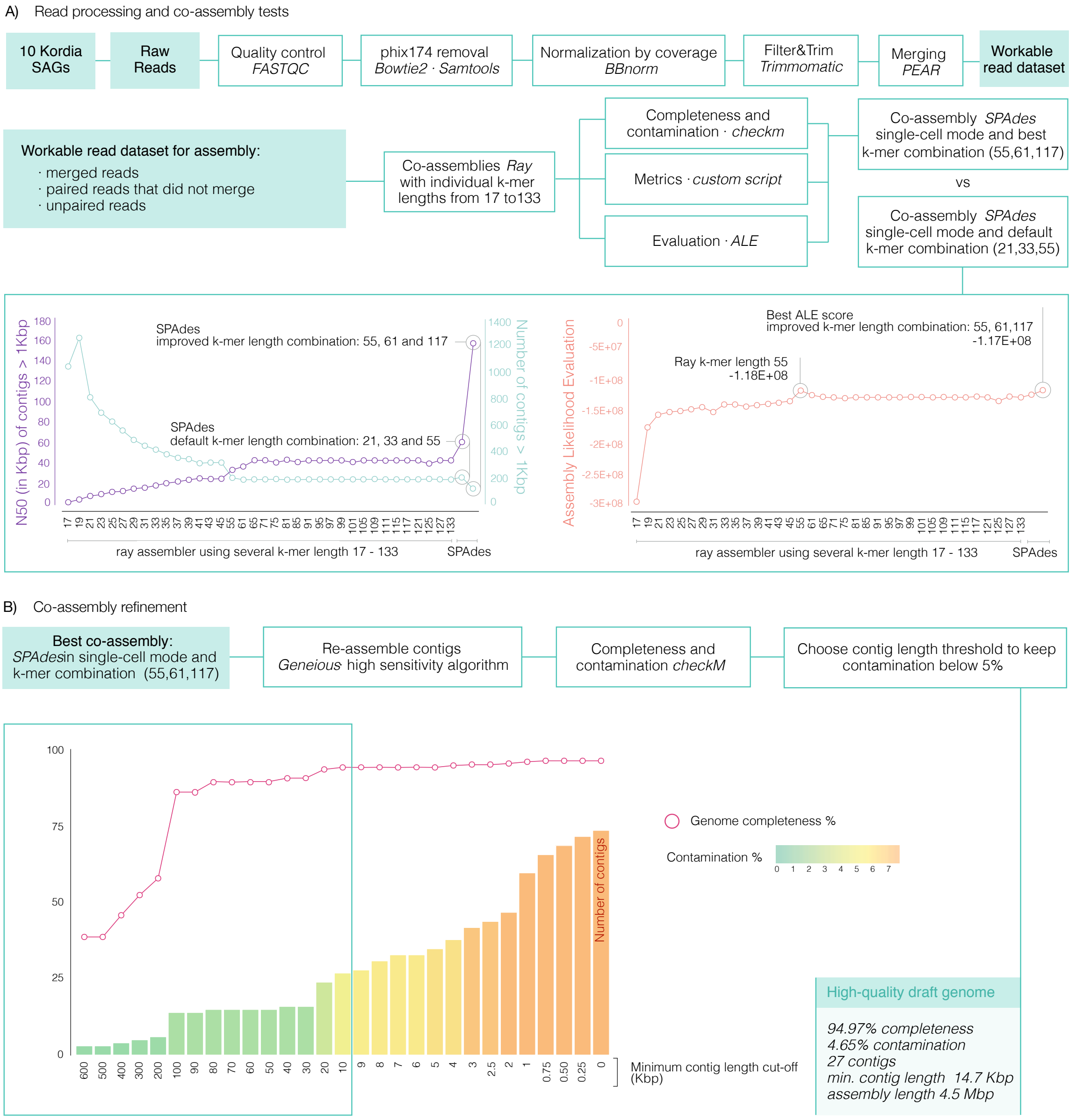

A) Read processing and co-assembly tests
10 Kordia SAGs
Raw Reads
Quality control *FASTQC*
phix174 removal *Bowtie2 · Samtools*
Normalization by coverage *BBnorm*
Filter&Trim *Trimmomatic*
Merging *PEAR*
Workable read dataset
Workable read dataset for assembly:
· merged reads
· paired reads that did not merge
· unpaired reads
Co-assemblies *Ray* with individual k-mer lengths from 17 to133
Completeness and contamination · *checkm*
Metrics · *custom script*
Evaluation · *ALE*
Co-assembly *SPAdes* single-cell mode and best k-mer combination (55,61,117)
vs
Co-assembly *SPAdes* single-cell mode and default k-mer combination (21,33,55)
N50 (in Kbp) of contigs > 1Kbp
Number of contigs > 1Kbp
SPAdes improved k-mer length combination: 55, 61 and 117
SPAdes default k-mer length combination: 21, 33 and 55
ray assembler using several k-mer length 17 - 133
SPAdes
Assembly Likelihood Evaluation
Best ALE score improved k-mer length combination: 55, 61,117 -1.17E+08
Ray k-mer length 55 -1.18E+08
B) Co-assembly refinement
Best co-assembly: *SPAdes*in single-cell mode and k-mer combination (55,61,117)
Re-assemble contigs *Geneious* high sensitivity algorithm
Completeness and contamination *checkM*
Choose contig length threshold to keep contamination below 5%
Genome completeness %
Contamination %
0 1 2 3 4 5 6 7
Number of contigs
Minimum contig length cut-off (Kbp)
High-quality draft genome
94.97% completeness
4.65% contamination
27 contigs
min. contig length 14.7 Kbp
assembly length 4.5 Mbp

FIGURE 9. 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-assemblying 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%.

using GTR+G model and 500 bootstrap replicates. The tree was later processed in iTol (Letunic and Bork, 2019).

2.3.7 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 (Hyatt et al., 2010). Software FetchMG v1.0 (Kultima et al., 2012) 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 (Edgar, 2004). Neighbor-Joining phylogenetic reconstruction was done with MEGA v7.0 and the resulting tree was processed in iTol.

2.3.8 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 (Jain et al., 2018), compareM v0.0.23[2] and pyani v0.2.7[3], respectively.

2.3.9 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 (Seemann, 2014). 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 (Kanehisa et al., 2016) and transporters were predicted using Transporter Classification TC database (Elbourne et al., 2017). Carbohydrate-active enzymes' annotation was based on dbCAN's CAZymes database (Yin et al., 2012) and peptidases were annotated using the Merops database (Rawlings et al., 2014). 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 (Bertelli et al., 2017). Insertion sequences (IS) were predicted and estimated with ISsaga tool (Varani et al., 2011) and prophage regions were predicted with PHASTER (Arndt et al., 2016). Secretion systems were detected using MacSyFinder-based TXSScan (Abby et al., 2014; Abby and Rocha, 2017). 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.

[2] https://github.com/dparks1134/CompareM

[3] https://github.com/widdowquinn/pyani

2.3.10 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 (Kumar et al., 2016) was used for a Maximum-Likelihood phylogeny reconstruction using LG+G model and 500 bootstrap replicates. The tree was later processed in iTol (Letunic and Bork, 2019).

2.3.11 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 (Table S5) (**?**). 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 jejudonensis*, *Kordia zhangzhouensis*, *Kordia periserrulae*, *Kordia* sp. SMS9 and *Kordia* sp. NORP58) (Table S5) 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 (Caro-Quintero and Konstantinidis, 2012). 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(Richter and Rosselló-Móra, 2009; Caro-Quintero and Konstantinidis, 2012). Normalization of the recruited reads was done by the sequencing depth of each metagenome and considering the complete genomic length.

2.3.12 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 (van Dongen and Abreu-Goodger, 2012) 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 (Conway et al., 2017) and ggplot2. Anvi'o phylogenomics workflow was used to produce a phylogenomic tree (Fasttree v2.1.11 (Price et al., 2010)) based on the amino acid sequences of the single copy genes shared by all five *Kordia* genomes.

2.3.13 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 (common code AB-193) are: M23 (SRR8655105), M19 (SRR8655106), N20 (SRR8655104), O13 (SRR8655103), L04 (SRR8655108), A(SRR8655107), O22 (SRR8655102), I20 (SRR8655109), I04 (SRR8655110) and P22 (SRR8655101). A complete list of accession numbers for all the genomes and metagenomes used in this study can be found in Table S5.

2.4 Results

2.4.1 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 (Royo-Llonch et al., 2017). 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 2), 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 (Table S6). The combination of k-mer length values chosen were: 55, 61 and 117 (Figure 9B). 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 2).

TABLE 2. Metrics of the individual assemblies and the co-assembly of the ten Kordia SAGs

Kordia sp. TARA_039_SRF genomes	Individual SAGs AB-193										refined co-assembly
	I04	I20	L04	M03	M19	M23	N20	O13	O22	P22	
Assembly metrics											
Total N of contigs	179	111	367	124	197	244	256	179	221	220	27
Contig N50 (Kbp)	20.9	30.0	33.0	54.2	40.2	38.4	27.5	42.7	49.8	25.5	42.9
Mean contig length (Kbp)	5.9	8.6	5.6	10.2	10.4	9.3	5.9	9.5	9.1	6.6	170.1
Min. contig length (bp)	118	118	118	118	118	118	118	118	118	118	14743
Max. contig length (Kbp)	59.5	80.7	113.4	135	137	139.7	107.7	93.4	106.1	113	682.4
Total assembly size (Mbp)	1.05	0.96	2.05	1.27	2.06	2.27	1.51	1.7	2.01	1.47	4.59
Quality parameters											
Genome completeness	24.36	11.15	39.94	26.14	44.11	43.32	33.02	36.96	44.78	32.31	94.97
Contamination	0.65	0.00	1.14	0.49	1.14	2.16	1.84	0.65	0.43	0.69	4.65
Strain heterogeneity	100.00	0.00	0.00	50.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Read mappings against co-assembly											
Contribution to co-assembly (%)	24.56	21.70	45.96	28.75	45.30	50.04	34.56	37.82	44.99	32.84	-
Mapped clean reads (%)	97.84	97.64	97.76	98.21	98.38	98.40	98.50	98.11	98.48	98.15	-
Genomic features											
Predicted ORFs	1012	944	1855	1142	1891	2118	1357	1548	1864	1334	4192
GC (%)	35.4	35.5	35.3	35.4	35.6	35.5	35.4	35.5	35.8	36.0	36.33
ANIm against co-assembly (%)	99.98	99.98	99.97	99.99	99.98	99.98	99.97	99.98	99.96	99.98	100
Identity of tetrant. freq. With co-assembly (%)	99.53	99.42	99.83	99.64	99.87	99.84	99.76	99.83	99.85	99.56	100

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 2). The average nucleotide identities of the individual SAGs against the co-assembly were higher than 99.9% in all cases (Table 2). 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 (Figure S5; Table S7). 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 (Stepanauskas, 2012), 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.

2.4.2 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 10), as well as through whole genome sequence composition comparisons of ANI, AAI and tetranucleotide signature (Table S8). 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).

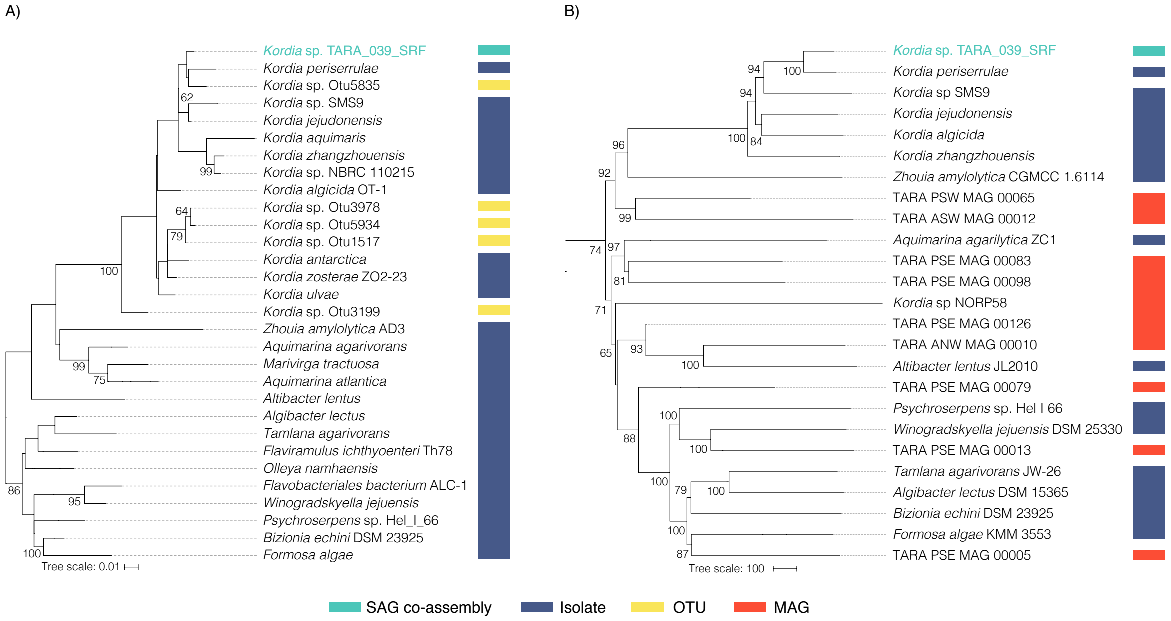


FIGURE 10. Phylogenetic reconstruction based of *Kordia* sp. TARA_039_SRF and its closest relatives based on the 16S rRNA gene (A) and 40 conserved single copy genes (B).

Otu5835, amplified from Malaspina expedition seawater samples, also falls within the *Kordia* sp. TARA_039_SRF and *K. periserrulae* clade.

2.4.3 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 11A; Table S9). 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 CO_2 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 CO_2, and H_2O and HCO_{3-}, 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 CO_2 and HCO_{3-}. *Kordia* sp. TARA_039_SRF encodes a third anaplerotic strategy driven by the malic enzyme, which converts pyruvate into L-malate with NADPH and CO_2. One of the copies is encoded in the core genome of Kordia whereas another copy is located in a genomic island.

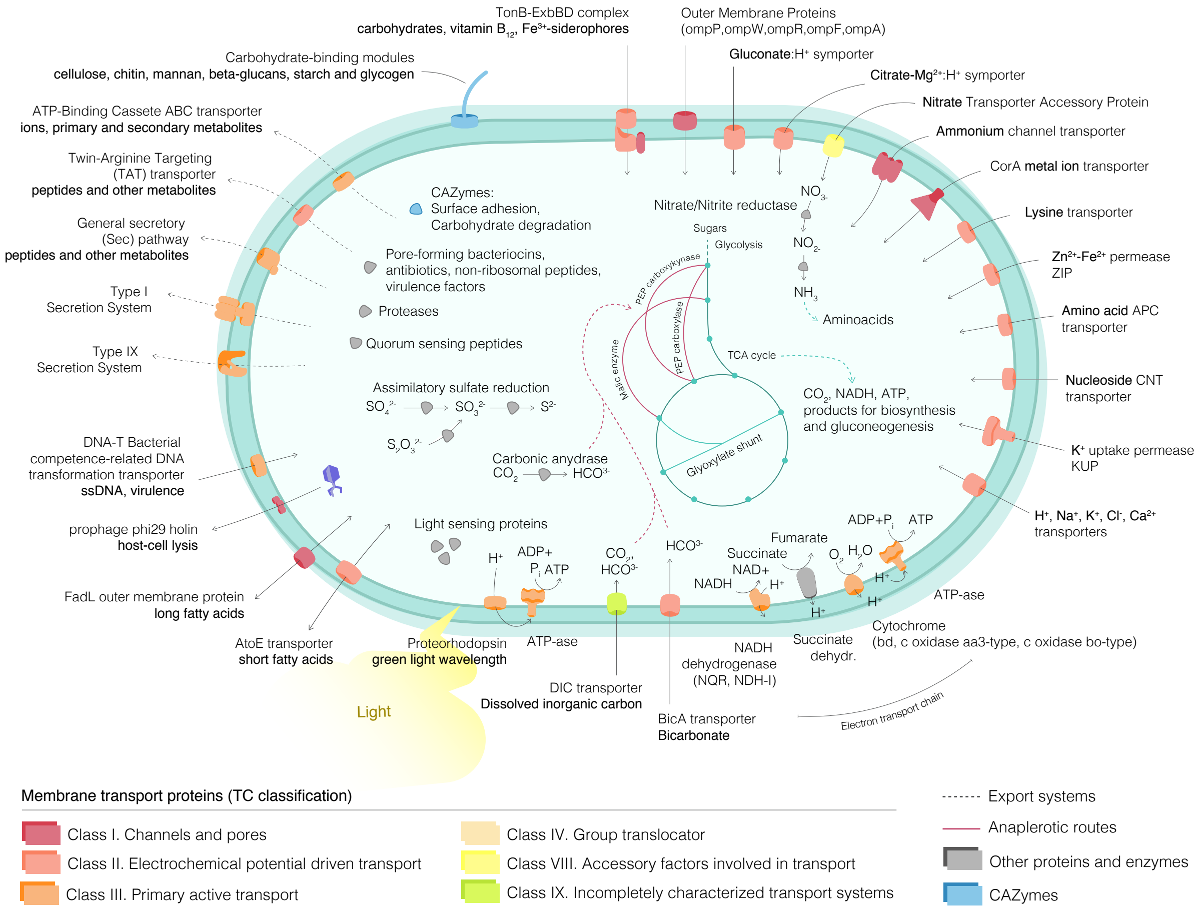

TonB-ExbBD complex
carbohydrates, vitamin B_{12}, Fe^{3+}-siderophores
Outer Membrane Proteins
(ompP,ompW,ompR,ompF,ompA)
Gluconate:H^+ symporter
Citrate-Mg^{2+}:H^+ symporter
Nitrate Transporter Accessory Protein
Ammonium channel transporter
CorA metal ion transporter
Lysine transporter
Zn^{2+}-Fe^{2+} permease
ZIP
Amino acid APC
transporter
Nucleoside CNT
transporter
K^+ uptake permease
KUP
H^+, Na^+, K^+, Cl^-, Ca^{2+}
transporters
Carbohydrate-binding modules
cellulose, chitin, mannan, beta-glucans, starch and glycogen
ATP-Binding Cassete ABC transporter
ions, primary and secondary metabolites
Twin-Arginine Targeting
(TAT) transporter
peptides and other metabolites
General secretory
(Sec) pathway
peptides and other metabolites
Type I
Secretion System
Type IX
Secretion System
DNA-T Bacterial
competence-related DNA
transformation transporter
ssDNA, virulence
prophage phi29 holin
host-cell lysis
FadL outer membrane protein
long fatty acids
AtoE transporter
short fatty acids
CAZymes:
Surface adhesion,
Carbohydrate degradation
Pore-forming bacteriocins,
antibiotics, non-ribosomal peptides,
virulence factors
Proteases
Quorum sensing peptides
Assimilatory sulfate reduction
SO_4^{2-} → SO_3^{2-} → S^{2-}
$S_2O_3^{2-}$
Carbonic anydrase
CO_2 → HCO^{3-}
Nitrate/Nitrite reductase
NO_3^-
NO_2^-
NH_3
Aminoacids
Sugars
Glycolysis
PEP carboxykynase
PEP carboxylase
Malic enzyme
TCA cycle
Glyoxylate shunt
CO_2, NADH, ATP,
products for biosynthesis
and gluconeogenesis
Light sensing proteins
H^+
ADP+
P_i ATP
CO_2,
HCO^{3-}
HCO^{3-}
Succinate
Fumarate
NAD+
NADH
H^+
O_2 H_2O
ADP+P_i ATP
ATP-ase
Cytochrome
(bd, c oxidase aa3-type, c oxidase bo-type)
Proteorhodopsin
green light wavelength
ATP-ase
Light
DIC transporter
Dissolved inorganic carbon
NADH
dehydrogenase
(NQR, NDH-I)
Succinate
dehydr.
BicA transporter
Bicarbonate
Electron transport chain
Membrane transport proteins (TC classification)
Class I. Channels and pores
Class II. Electrochemical potential driven transport
Class III. Primary active transport
Class IV. Group translocator
Class VIII. Accessory factors involved in transport
Class IX. Incompletely characterized transport systems
Export systems
Anaplerotic routes
Other proteins and enzymes
CAZymes

FIGURE 11. 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.

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 (Begley et al., 2001).

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 (Table S10). 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 (Table S11) were found, that are also present in other Flavobacteria (González et al., 2008). The novel *Kordia* sp. TARA_039_SRF does not encode a complete gliding motility complex.

The co-assembly encodes four Polysaccharide Utilization Loci (PULs) (Table S9), 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 (Figure S7), 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 11; Table S12).

Kordia sp. TARA_039_SRF codes for two complete secretory systems: T1SS and the Bacteroidetes specific T9SS (Table S13). There are 85 peptides encoded in *Kordia* sp. TARA_039_SRF that contain the Por secretion system domain (Table S14), 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 (Table S15). 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 (Table S16). 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 (Table S17) and three predicted incomplete prophage regions, with a total size of 26,876 bp (0.057% of the total genome length) (Table S18).

2.4.4 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 12A, Table S19). 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 12B) and against both the core genome and the accessory genome (Table S20). 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 (Figure S8, S9 and S10).

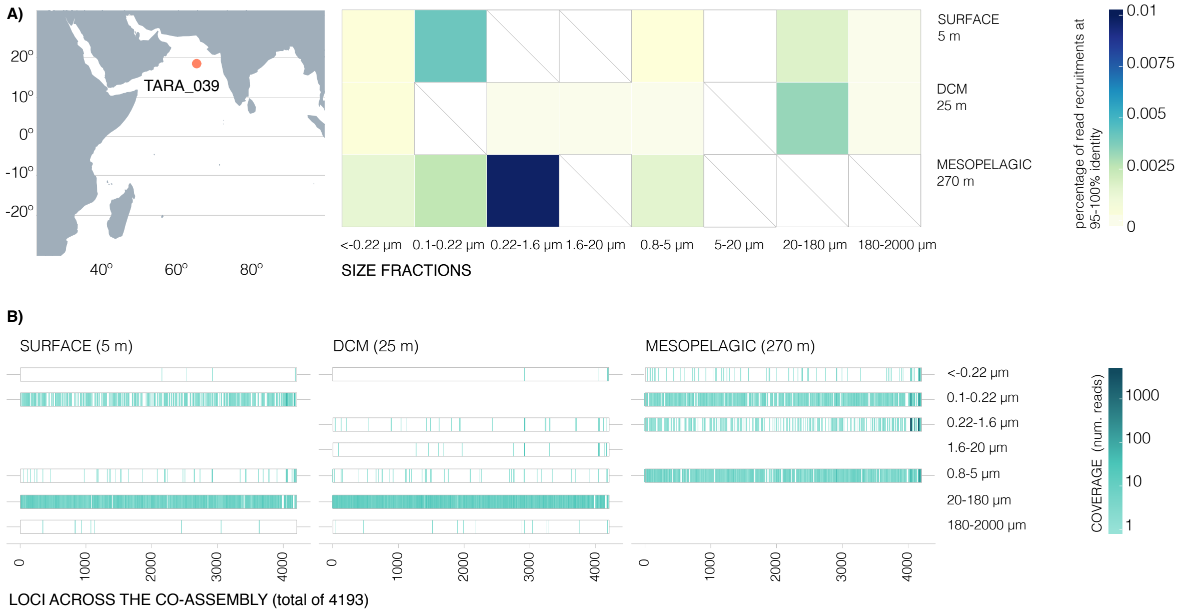


FIGURE 12. Relative abundances of recruited reads in all available metagenomes from station TARA_039 in the north Indian ocean (A). Unavailable metagenomic samples are depicted as slashed in the heatmap, samples where there was no read recruitment whatsoever appear in white. Horizontal genomic coverage 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.

2.4.5 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 3). 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 13, Table S21) 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 13).

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

TABLE 3. Description of the five Kordia genomes used in the comparative genomics analysis.

	Kordia sp. TARA_039_SRF	*K. algicida*	*K. jejudonensis*	*K. zhanghzouensis*	*K. periserrulae*
Genomic features					
Length (Mbp)	4.59	5.01	5.35	4.03	4.72
GC (%)	35.8	34.2	33.9	33.8	36.2
Coding density (%)	88.45	87.58	88.09	90.98	88.93
Total number of ORFs	4192	4492	4782	3585	4105
Predicted protein coding genes (CDS)	4124	4411	4714	3527	4051
Completeness (%)	94.83	100	100	100	100
Pangenome					
# clusters of core genes (% of total)	2286 (57.4%)	2286 (55.4%)	2286 (50.3%)	2286 (65.8%)	2286 (58.1%)
# clusters of genes shared with Kordia spp. (% of total) Accessory shared genome	971 (24.4%)	1038 (25.1%)	969 (21.3%)	772 (22.2%)	1050 (26.6%)
# clusters of unique flexible genes (% of total) Accessory strain specific genome	723 (18.1%)	801 (19.4%)	1287 (28.3%)	414 (11.9%)	599 (15.2%)
Total # of gene clusters	3980	4125	4542	3472	3935
ANIm (%)					
Kordia sp. TARA_039_SRF	100	-	-	-	-
Kordia algicida	81.13	100	-	-	-
Kordia jejudonensis	79.83	81.47	100	-	-
Kordia zhangzhouensis	79.62	79.56	79.23	100	-
Kordia periserrulae	89.13	80.97	79.83	79.63	100
AAI (%)					
Kordia sp. TARA_039_SRF	100	-	-	-	-
Kordia algicida	81.05	100	-	-	-
Kordia jejudonensis	77.96	80.14	100	-	-
Kordia zhangzhouensis	79.93	80.62	78.06	100	-
Kordia periserrulae	92.12	81.03	77.97	80.06	100
Ribosomal RNA					
16S rRNA id. % with *Kordia* sp. TARA_039_SRF	100	-	-	-	-
16S rRNA id. % with K. algicida (coverage)	97.03 (100%)	100	-	-	-
16S rRNA id. % with K. jejudonensis (coverage)	98.00 (95%)	96.69 (95%)	100	-	-
16S rRNA id. % with K. zhangzhouensis (coverage)	96.85 (98%)	95.85 (98%)	97.24 (00%)	100	-
16S rRNA id, % with K. periserrulae (coverage)	97.56 (97%)	96.74 (97%)	97.24 (98%)	96.34 (99%)	100
# of tRNA types	21	21	20	19	20
# of rRNA operons	1	3	1	1	1

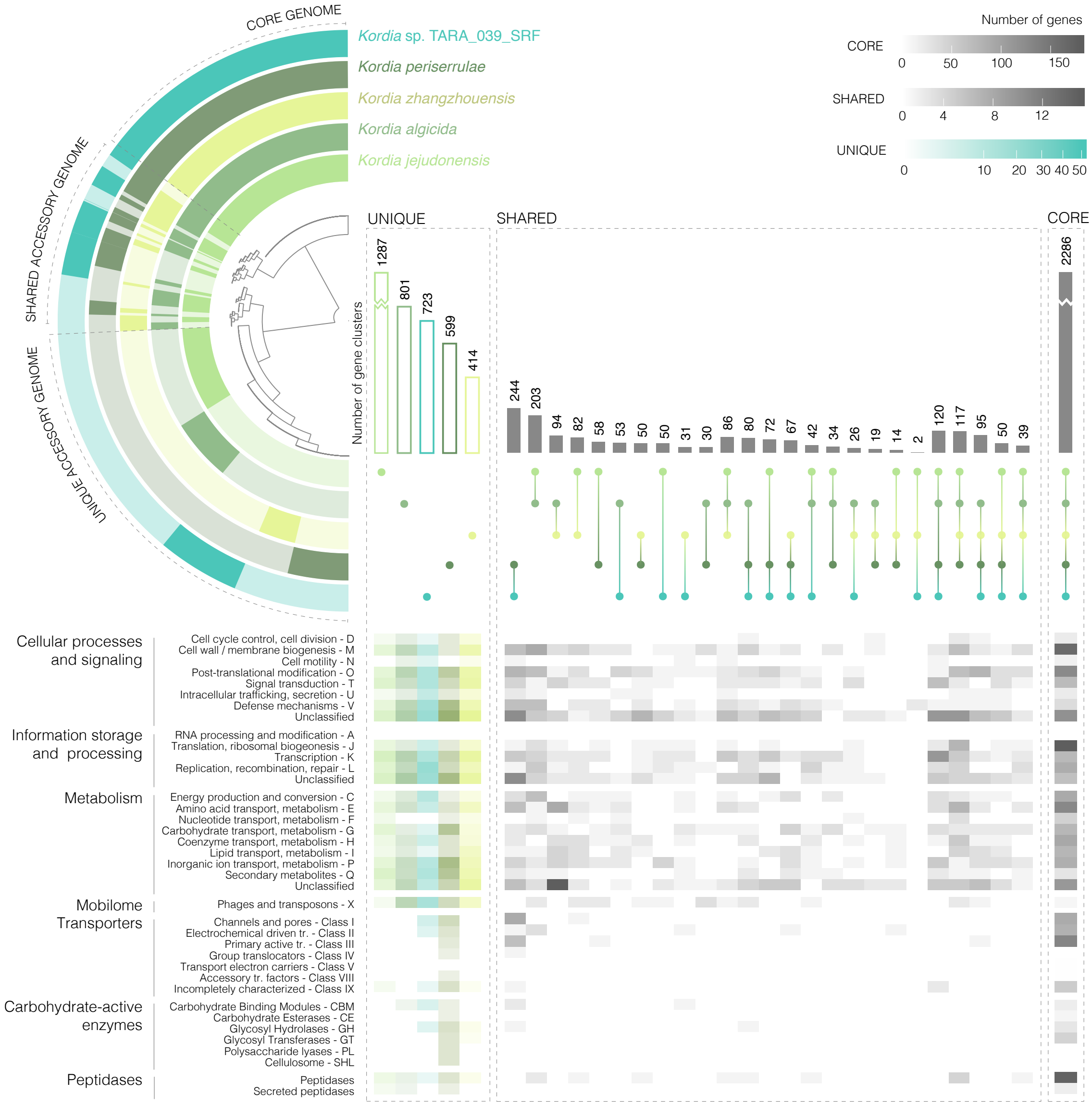


FIGURE 13. Comparative genomics of *Kordia* sp. TARA_039_SRF, *K. algicida*, *K. jejudonensis* and *K. zhangzhouensis*. Top left dendogram represents clusters of genes ordered by the number of genomes encoding them, regardless of copy numbers. The phylogenomics tree connecting *Kordia* spp. labels was 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

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 3, Figure 13, 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.

2.5 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 (Royo-Llonch et al., 2017) into a high-quality draft genome, which is 94.83% complete and meets the set standards regarding contamination (<5%) (Bowers et al., 2017; Konstantinidis et al., 2017). 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 (Mangot et al., 2017; Kogawa et al., 2018). 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 (Sharon et al., 2015). The combination of different k-mer lengths during co-assembly generated remarkably longer contigs than default assembly parameters (Chikhi and Medvedev, 2014). 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-assembly's taxonomic novelty was confirmed through phylogenies based on the 16S rRNA gene and highly conserved single copy genes (Kultima et al., 2012) 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 (Shoseyov et al., 2006). It also encodes Polysaccharide Utilization Loci, surrounded by CAZymes as previously described in Bacteroidetes (carbohydrate-binding modules, glycosyl hydrolases, glycosyl transferases, carbohydrate esterases) (Lapébie et al.). 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 (Kisker et al., 2013), 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 (Landry et al., 1998) and has one of the most intense and large Oxygen Minimum Zones (OMZ) (Paulmier and Ruiz-Pino, 2009). A diatom bloom was registered a month prior to the sampling period and Roullier et al. (2014) 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) (Roullier et al., 2014). 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 (Royo-Llonch et al., 2017). 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 (Ren and Pauisers, 2005), 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 (González et al., 2008).

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 (Fernández-Gómez et al., 2013)), ii) genome streamlining with less non-coding regions and iii) lower GC content (Swan et al., 2013). 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 (Segerman, 2012), 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 (López-Pérez and Rodriguez-Valera, 2016) and Prochlorococcus (Kettler et al., 2007).

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.

2.6 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 photoheterotrophically 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 forO_2than 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 rhanmnogalacturonyl 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.

2.7 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 69.

CHAPTER 3

CHAPTER 3: Ecogenomics of key prokatyotes in the Arctic Ocean

3.1 Abstract

The remote Arctic Ocean is the fastest changing habitat on Earth and the least studied. Predicting the future of this ecosystem remains a challenge and is particularly relevant for the microorganisms that form the base of the marine food web. To enhance our understanding of the prokaryotic diversity in Arctic seawaters, as well as their biogeographic patterns and metabolic potential, we reconstructed a total of 3550 metagenomic assembled genomes (MAGs) of the Tara Oceans Polar Circle expedition (TOPC) covering different Arctic Marine Areas (AMA), depths, and seasons. We described prokaryotic panarctic/ bipolar generalist and specialist taxa with different habitat and functional preferences and defined their expression patterns, providing new insights into the taxonomic composition and metabolic processes of key arctic microbial species. This catalogue of Arctic prokaryotic genomes (TOPC MAGs) is proposed to serve as a baseline in prospective studies of the state and future of the Arctic Ocean.

3.2 Introduction

The Arctic is under increasing pressure from climate change, industrialization, urbanization and tourism. We need to understand how Arctic microorganisms adapt and thrive in the marine ecosystem to forecast their fate in a future ocean impacted by anthropogenic change, as they are the foundation of the marine food web. In addition, the possible invasion of the Arctic Ocean by southern species is an unknown factor with the potential to alter the dynamics of marine ecosystems, affecting from microbes to large animals (Grebmeier, 2012). Therefore, it is fundamental to build a catalogue of truly panarctic/polar prokaryotes, that not only identifies their taxonomy but also their functional capabilities, as a baseline for future climate change scenarios.

The Arctic Ocean's marine ecosystem is subject to the drastic seasonal variation in solar irradiance, ice cover, temperature and extraordinary riverine inflow, together with the influence of inflowing waters from the Pacific and the Atlantic Oceans (Meltofte, 2013). Thus, organisms inhabiting the upper water column have to adapt to a highly dynamic environment (Thomas, 2016). Primary production in this extreme environment happens during spring and summer seasons, when light availability and increased temperatures enhance phytoplankton growth, that bloom under the ice cover and surrounding the melting ice (Wassmann and Reigstad, 2011). Such blooms are followed by a succession of bacterial populations, mostly heterotrophs from the phyla Bacteroidetes and Proteobacteria (Bunse and Pinhassi, 2017). During winter, the lack of light makes productivity almost negligible, resulting in very low vertical carbon export from the surface layers (Olli et al., 2002; Forest et al., 2008). As photosynthesis is restricted, the growth of heterotrophic bacteria and protists is enhanced (Riedel et al., 2007, 2008). In the dark season, mixotrophy (Moorthi et al., 2009; Alonso-Sáez et al., 2010) and chemolithoautotrophy (Alonso-Sáez et al., 2008, 2012; Boetius et al., 2015) also increase in importance in archaea and bacteria.

The Arctic Ocean can be split into eight regional divisions based on the different features of ecological significance, called the Arctic Marine Areas (AMAs) (AMAP/CAFF/SDWG, 2013). Several planktonic biodiversity assessments have sampled a selection of AMAs, like the Arctic Ocean Survey (AOS) during 2005-2010 (Thaler and Lovejoy, 2015), the International Census of Marine Microbes (ICoMM) (Amaral-Zettler et al., 2010), sampling events performed at local scales and formal long term monitoring exercises performed by many countries in direct contact with the Arctic waters (State of the Arctic Marine Biodiversity Report, 2017). However, in contrast with the existing information about diversity and distribution patterns of arctic microbial eukaryotes (Lovejoy et al., 2006), that includes the definition of endemic Arctic pycoplanktonic species (Lovejoy et al., 2007; Terrado et al., 2013), the record of the prokaryotic diversity of the Arctic Ocean is scarce and generally reduced to local surveys, greatly dependent on molecular approaches (Huse et al., 2008; Galand et al., 2009a; Kirchman et al., 2010; Pedrós-Alió et al., 2015). The functional potential of arctic prokaryotes has been approached with radioactive incubation methods, MAR-FISH assays and metagenomic functional annotation studies, with great value for understanding of the prokaryote's role in the ecosystem's functioning (Alonso-Sáez et al., 2014, 2012, 2010). Nevertheless, going beyond the community level with genome-centric approaches that are independent of isolation in culture, remains essential to identify key players in the Arctic's prokaryotic diversity, and their impact on the ecosystem. To fill the gaps in planktonic diversity and function in the Arctic Ocean, the Tara Oceans Polar Circle expedition sampled five AMAs (Atlantic Arctic, Kara-Laptev Sea, Pacific Arctic, Arctic Archipelago and Davis-Baffin) through a circumnavigation around the polar circle, starting in May 2013 (spring) and finishing in October (autumn) the same year. The sampling covered six size fractions, from viruses to protists combining multiple state-of-the-art techniques for a complete description of the sampled biodiversity (Karsenti et al., 2011; Pesant et al., 2015).

In this study, we have built 3550 metagenomic assembled genomes (MAGs) from the 41 metagenomes of the Tara Oceans Polar Circle cruise, following a co-assembly strategy of metagenomes pooled together by their resemblance in community composition, rather than geographical distribution (Delmont et al., 2018) or individually (Tully et al., 2017a). Their distribution in metagenomic samples from the different marine areas, vertical layers and seasons has been used to define their biogeographical distribution. Their metatranscriptomic recruitments have been used to determine their expression patterns and to explore the presence of key marker genes related to different pathways of the dark carbon fixation and to the metabolisms of sulfur, nitrogen and methane, in the different seasons and areas. We have defined two subsets of genomes based on their niche breadth: the generalists, evenly distributed in the majority of samples; and the specialists; with an uneven distribution, usually peaking in abundance in a few samples (Levins, 1968; Colwell and Futuyma, 1971). The variety of niches a taxon can occupy (or niche breadth) seems to be an indicator of their susceptibility to changes in environmental factors, hence their correlations with environmental factors have also been determined with both metagenomic and metatranscriptomic recruitments. Since the community turnover in response to ocean warming may be strongest in polar regions (Salazar et al., 2019), we have focused on those MAGs that are only present in the Arctic or that show a bipolar distribution. Likely being sentinel species of climate change in the Arctic, we have analyzed their metatranscriptomic recruitments aiming to spot the most expressed in every sample together with their metabolic potential in the Arctic ocean.

We provide a pan-Arctic ecogenomics approach to the study of key arctic prokaryotic species. The unveiling of their relevance in the different carbon fixation processes occurring in the seasons and marine areas studied is aimed to serve as a starting point for future experimental approaches dedicated to the processes occurring in the Arctic's marine ecosystem, and in prospective studies of the state and future of the Arctic Ocean.

3.3 Materials and methods

3.3.1 Sample and environmental data collection and processing

As described previously (Salazar et al., 2019) genetic and environmental data were collected during the Tara Oceans expedition (2009-2013), which includes the Tara Oceans Polar Expedition (TOPC, 2013). Polar stations had absolute latitudes above 60º. Sampling was conducted within the epipelagic (surface / SRF, 5-10 m and deep chlorophyll maximum / DCM, 20-200 m) and mesopelagic layer (MES, 20-200 m). The sampling strategy and methodology have been described elsewhere (Pesant et al., 2015). Environmental data measured or inferred at the depth of sampling are published at the PANGAEA database[4].

3.3.2 Extraction and sequencing of DNA and cDNA

Metagenomic DNA and RNA were extracted from prokaryote-enriched size fraction filters as previously described (Alberti et al., 2017). As detailed in (Salazar et al., 2019): For the DNA libraries, extracted DNA was sonicated to a size range of 100-800 bp. The DNA fragments were subsequently end-repaired and 30-adenylated before Illumina adapters were added using the NEBNext Sample Reagent Set (New England Biolabs). The ligation products were then purified by Ampure XP (Beckmann Coulter), and the DNA fragments (> 200 bp) were PCR- amplified with Illumina adaptor-specific primers and Platinum Pfx DNA polymerase (Invitrogen). The amplified fragments were then size selected (300 bp) on a 3% agarose gel. For the metatranscriptomic libraries, 'low-input' cDNA synthesis methods adapted to prokaryotic prokaryotic mRNA were used (Alberti et al., 2014). Briefly, total RNA was depleted of rRNA using the Ribo-Zero Magnetic Kit for Bacteria (Epicentre) and then concentrated to 10 mL total volume with the RNA Clean and Concentrator-5 kit (ZymoResearch). The amount of depleted RNA was measured by Qubit RNA HS Assay quantification, and 40 ng or less was used to synthesize cDNA with the SMARTer Stranded RNA-Seq Kit (Clontech). Additional details are described elsewhere (Alberti et al., 2017). All libraries (DNA and RNA) were subjected to profile analysis using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and qPCR (MxPro, Agilent Technologies, USA), and then sequenced with 101 base-length read chemistry in a paired-end flow cell on Illumina HiSeq2000 sequencing machines (Illumina, USA).

3.3.3 Generation of Metagenomic Assembled Genomes

Co-assembly: In order to reconstruct as much genomes from the dataset, we chose an approach that involved the co-assembly of several samples together, hence increasing the sequencing space for each co-assembly while keeping the computational needs attainable. The pools of samples to

[4] https://doi.org/10.1594/PANGAEA.875582

be co-assembled were chosen based on their taxonomic composition (Salazar et al., 2019), so as samples that clustered together in a NMDS based on 16S miTag taxonomy were assembled jointly with megahit (v1.1.2, –presets meta-large –min-contig-len 2000; Table S22; SM Figure 1) (Li et al., 2015). All assembled contigs were pooled together and de-replicated with cd-hit-est (v4.6.8-2017-0621, compiled from source with MAX_SEQ=10000000, options -c 0.99 -T 64 -M 290000 -n 10) (Li and Godzik, 2006), reducing the dataset from 3.95 M to 1.91 M contigs.

Binning and curation: The reads of the input metagenomes reads were back-mapped to the remaining contigs with bowtie2 (v2.3.2) (Langmead and Salzberg, 2012) with default options, keeping only mapping hits with quality larger than 10 (samtools v1.5; options -q 10 -F 4) (**?**Kang et al., 2015). Mapping hits were processed with jgi_summarize_bam_contig_depths from metabat2 (v2.12.1) (Kang et al., 2015) with options –minContigLength 2000 –minContigDepth 1 and then binned with metabat2 with default options.

The completeness and contamination of each bin, as well as a first estimation of their taxonomic classification, based on single-copy marker genes was assessed with checkM (v1.0.11) (Parks et al., 2015) with the lineage_wf workflow.

Contigs of 96 bins with estimated completeness larger than 95% and contamination lower than 5% were reassembled in Geneious (v10.2.4) with minimum overlap identity 95%, maximum mismatches per read 5, no minimum overlap and don't allow gaps options to find overlaps that allowed to reduce the genome fragmentation, and the results were curated manually. These were considered to be high quality (HQ) MAGs. Additionally, contigs of 434 bins with estimated genome completeness larger than 50% and contamination lower than 10% were also re-assembled with cap3 (v021015) (Huang and Madan, 1999) with overlap length and percent identity cut-offs of 25 bp and 95% respectively. These were considered to be medium quality (MQ) MAGs.

Annotation of bins: All 3550 bins were annotated, including gene prediction, tRNA, rRNA and CRISPR detection with prokka (v1.13) (Seemann, 2014) with default options, using the estimated Domain classification from checkM as argument of the –kingdom option. Additionally, HQ MAGs' predicted coding sequences were annotated against the KEGG orthology database (KEGG release 2019-02-11) (Kanehisa et al., 2019) with diamond (v0.9.22) (Buchfink et al., 2015) with options blastp -e 0.1 –sensitive, and against the PFAM database (release 31.0) with hmmer (v3.1b2) (Wheeler and Eddy, 2013) with options –domtblout -E 0.1.

Taxonomic annotation and genome quality: All 3550 bins were classified taxonomically with GTDBTk (v0.3.2) (Chaumeil et al., 2019) with the classify_wf workflow. Genome estimated completeness and contamination of HQ MAGs was reassessed with checkM as above (Table S23). For those MAGs encoding the 16S rRNA gene, taxonomic annotation was done using SILVA's database and SINA aligner tool v1.2.11 with minimum of 50% of identity (higher thresholds could not classify the ribosomal genes) and Last Common Ancestor algorithm (Table S24).

ANI and AAI calculations: Average nucleotide identity (ANI) was calculated with fastANI v1.2

and default options (Jain et al., 2018) was estimated for each possible pair of all MAGs with more than 50% of genome completeness and less than 10% of genome contamination to check whether the reconstructed genomes could belong to the same species (defined at >95% ANI). As alignment fraction between genomes lower than 20% may provide espurious large ANIs, the average amino acid identity (AAI), that takes into account only the fraction of orthologous genes, was also estimated (compareM[5] v0.0.23 with default options).

3.3.4 Read recruitments

Selection and subsampling of samples: The samples chosen for read recruitment include all the metagenomes from TOPC Stations (Figure 14A and 14B), excepting those containing mixed water layers (TARA_XXX_IZZ or TARA_XXX_ZZZ), the five Tara Oceans metagenomes sampled in the Southern Ocean and a selection of 28 Tara Oceans expedition metagenomes from temperate latitudes (Table S25, SM Figure 2). These were selected based on their sequencing depth (that had to be at least as large as the smallest TOPC metagenome), geographic location (covering the different oceans and seas sampled by the Tara Oceans expedition) and the coverage of different water layers. For the metagenomic samples selected, recruitments were also done with their available metatranscriptomes. Paired-end libraries were used individually for Fragment Recruitment Analysis after cleaning and a step of random subsampling. The latter was done with DOE JGI's BBTools' reformat.sh script (v38.08; https://sourceforge.net/projects/bbmap/), selecting as subsampling value the smallest sequencing depth of the TOPC meta-omic dataset (i.e 140,658,260 and 45,212,614 fragments for metagenomic and metatranscriptomic libraries respectively).

Competitive fragment recruitment analysis: Nucleotide-Nucleotide BLAST v2.7.1+ was used to recruit metagenomic and metatranscriptomic reads similar to any of the 3550 TOPC MAGs. Blast is slower than other high-throughput (HT) aligners, but allows for finer-tuned alignment parameters, plus it is the gold standard against which all HT aligners are compared. Recruitment was competitive, meaning that individual samples were aligned against the pooled contigs of all 3550 bins. Blast alignment parameters were the following: -perc_identity 70, -evalue 0.0001. Only those reads with more than 90% coverage and mapping at identities equal to or higher than 95% were considered to be representative of the MAG. In case of hits with the same e-value, larger bitscore or larger alignment length were used sequentially to choose the best hit. If ties persisted, the best hit was selected at random from the candidate reads. Best hits that corresponded to rRNAs (according to the prokka annotation) were also discarded.

Detection and filtering of false positive recruitments: Putative false positive recruitments were detected and excluded considering their horizontal genomic coverage which was calculated using R package GenomicRanges (Lawrence et al., 2013). A minimum horizontal genomic coverage threshold was set testing the effect of different thresholds on the final number of MAGs recruiting

[5] https://github.com/dparks1134/CompareM

(richness) and the number of samples in which they recruited (occurrence). The variation of species richness in each metagenome was tested for a range of increasing minimum horizontal genomic coverage thresholds (0, 0.1, 0.5, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 98, 100). Recruitments in which the horizontal coverage was equal to or higher than the thresholds, were considered true and those covering a smaller percentage of their genome than the cut-off value were discarded.

The number of species present in each metagenome decreased with increasing minimum horizontal coverage, reaching a clear saturation in richness when the minimum horizontal coverage was 20% for metagenomes from temperate latitudes (SM Figure 3).

Setting a horizontal genomic coverage threshold has an effect on the occurrence of each MAG in the metagenomic samples. In all metagenomic datasets (arctic, southern ocean and temperate), the distribution of occurrence vs mean abundance of MAGs stabilizes when the minimum horizontal coverage is 10% or higher. Lower thresholds show different patterns of distribution, increasing the number of higher occurrences at very low mean abundances (SM Figure 4). To date, there is no consensus about the minimum horizontal coverage thresholds to discard false mappings, some previous studies used 30% (Neyfach et al., under rev.), 50% (Delmont and Eren, 2018) or 60% (Varghese et al., 2015). Based on our analyses, we chose 20% as the minimum horizontal genomic coverage to consider a recruitment valid.

3.3.5 Abundance and distribution of MAGs

Abundance and occurrence calculation: Only those recruited reads aligning at an identity equal to or larger than 95% where considered to be representative of the MAGs. Recruitments passing the minimum horizontal genomic coverage threshold of 20% where considered to represent a true presence of the MAG in the sample, while those with a horizontal genomic coverage lower than 20% were considered not representative of the MAG, thus absent. These read recruitments were normalised by assembly size and sequencing depth with their transformation to RPKGs (recruited reads per genome kilobase and sample gigabase). RPKG filtered at 20% minimum horizontal coverage of metagenomic reads are in Table S26. Metatranscriptomic RPKG recruitments as RPKG are in Table S27.

Sample distribution based on MQ and HQ MAGs composition: The ordination of samples based on their MQ and HQ MAG composition, with RPKG as abundance estimate, was done with a Non-metric Multidimensional Scaling (NMDS) approach using function metaMDS from the vegan package in R.

3.3.6 Niche breadth and classification of specialists/generalists

Habitat specialist-generalist patterns in the Arctic Ocean: Specialist-generalist classification of MAGs was based on Levin's Index (B). In order to avoid sampling biases, function spec.gen from R package EcolUtils[6] was used to calculate B for a 1000 random permutations of the metagenomic RPKG table and categorise MAGs into generalists if the original B index was larger than its confidence interval (CI 95) or specialists if the original B index was smaller than its confidence interval (CI 95) (SM Figure 5). As the sampling occurred in a spatial and temporal gradient, each individual sample was considered as a habitat. Niche breadth classifications are in Table S23.

3.3.7 Functional analysis

Functional annotation: MAGs were processed with Prokka (1.13) to obtain predicted genes, rRNAs, tRNAs and CRISPRs and their functional annotation (COGs and ECs), using default options and passing the domain inferred by checkM (1.0.11) to option –kingdom. Additionally, MAGs were further functionally annotated for PFAMs (HMMER 31b2, PFAM release 31.0) and KEGG orthologs (KO; Diamond 0.9.22, KEGG release 2019-02-11).

Selection of key genes and metabolic pathways: To explore the ubiquity of representative biogeochemical cycling metabolisms related to carbon, sulfur, nitrogen and methane, a selection of 101 marker genes (Table S28) were searched in the TOPC MAGs dataset.

3.4 Results and Discussion

3.4.1 Generating the TOPC MAG dataset

The co-assembly of 41 Tara Oceans Polar Circle metagenomes, representative of different depths in 5 Arctic Marine Areas or CAFF regions (Figure 14A), resulted in the reconstruction of 3550 bins. There are alternative assembly strategies used in the recovery of environmental genomes from metagenomes, like individual assemblies (Tully et al., 2018a) or co-assembly of geographically bounded samples (Delmont et al., 2018). We co-assembled pools of samples that were most similar in their 16S miTAG composition as a way to obtain a less redundant set of bins with higher genome completeness (SM Figure 1), even though due to strain heterogeneity in the co-assembly procedure, contigs may be more prone to fragmentation than in the assembly of individual samples (Sharon et al., 2015; Roux et al., 2017; Sczyrba et al., 2017). Strain heterogeneity could also affect the binning process (Alneberg et al., 2014).

Based on genome completeness and quality values, the 3550 bins have been classified into four quality groups, following existing thresholds (Bowers et al., 2017) (Figure 14B): 96 high quality

[6] https://github.com/GuillemSalazar/EcolUtils

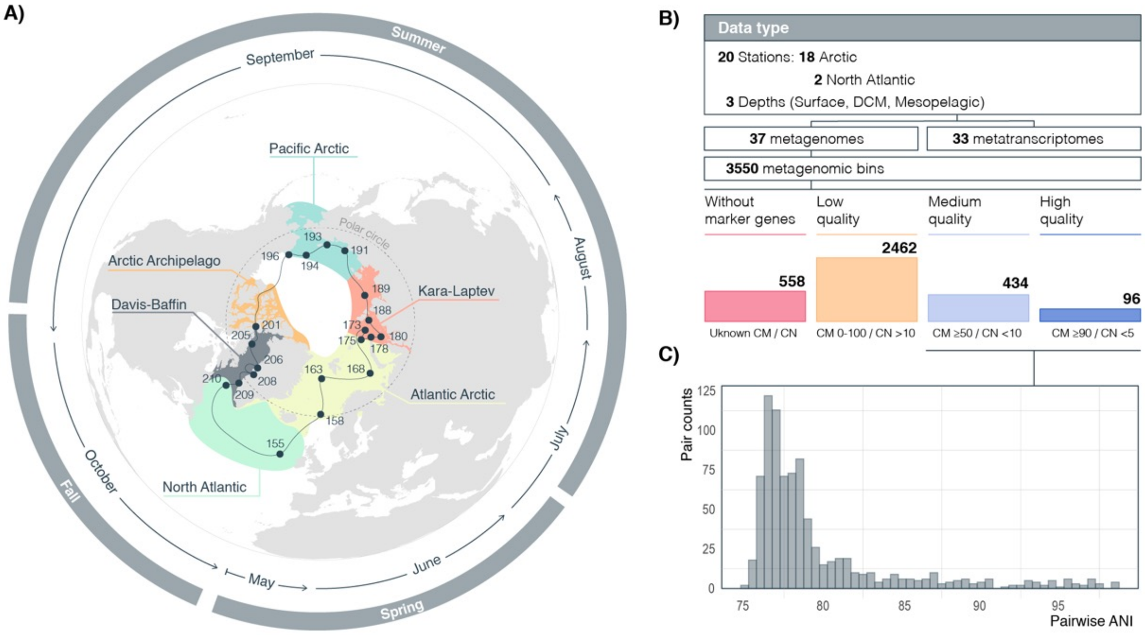


FIGURE 14. Tara Oceans Polar Circle MAGs dataset. (A) Tara Oceans Polar Circle expedition trajectory and the stations from which meta-omics samples are available. Colored regions belong to the six sampled Arctic Marine Areas or CAFF regions. The Polar Circle is highlighted with a dashed line. Outer circles present the month and season of sampling during the circumnavigation.(B) Outline of the polar meta-omics dataset, the number of bins assembled from metagenomic samples and their quality-based classification, that is measured combining genome completeness (CM) and contamination (CN). (C) Histogram of pairwise ANI comparisons of 530 MQ and HQ MAGs.

MAGs (HQ), 434 medium quality MAGs (MQ), 2642 low quality bins and 558 bins whose quality cannot be estimated due to a lack of phylogenetic marker genes. Only those assemblies with sufficient quality ratings were considered MAGs and were used in further ecological analyses, that is MQ MAGs with completeness values ≥50% and contamination <10% and HQ MAGs with completeness ≥90% and contamination <5%. HQ MAGs were manually curated and re-assembled in order to resolve contig fragmentation.

Only 8 combinations (0.006%) out of 140185 genome pairs within the MQ and HQ MAGs dataset (530 MAGs) could be considered to be closely related species, with ANIs larger than 96% (Figure 14C), ten times less than massive single-cell genomics analysis of the ocean's microbiome by (Pachiadaki et al., 2019). Six of these combinations were also confirmed by AAI larger than 96%, and only 3 with orthologous fractions over 80% (SM Figure 6). Nevertheless, considering previous polyphasic approaches to the classification of genomes into species based on whole-genome comparisons, only one pair of HQ MAGs were likely redundant (ANI > 95% and tetranucleotide frequency > 99%) (Richter and Rosselló-Móra, 2009) (SM Figure 7). Based on AAI results, our MQ and HQ MAGs dataset represent the consensus genome of different species and shows a great diversity at the class level (AAI between 40-50%), and to a lesser extent at the order level (AAI between 50-60%) (SM Figure 6).

3.4.2 Abundance and expression of TOPC bins

The 3550 TOPC bins represent a unique dataset but to what extent do these bins represent abundant uncultured genomes? What fraction of the community are we retrieving? One of the weak points of reconstructing genomes from metagenomes of marine seawater origin is the fact that, when mapping the metagenomic reads back to the bins, they usually recruit a rather low fraction (about 10%) compared to other culture-independent techniques like assemblies (Salazar et al., 2019) or single cell genomics (Pachiadaki et al., 2019). In our case, the full set of 3550 arctic bins recovers almost half of the Tara Oceans Polar Circle subsampled metagenomic reads (43.3% of the Arctic metagenomes and 35.1% of North Atlantic metagenomes) (Figure 15A), which could be due both to the benefits of co-assemblying samples similar in taxonomic composition or to the fact that polar prokaryotic communities tend to be less rich in diversity than the temperate ocean (Ibarbalz et al., 2019). These numbers are similar to those of the 4,741 marine SAGs from the GORG-tropics database (Pachiadaki et al., 2019), and between two and almost ten times greater than previous big-scale studies of metagenomic assembled bins (Parks et al., 2017; Delmont et al., 2018; Klemetsen et al., 2018; Tully et al., 2018a) and single amplified genomes (Swan et al., 2013). Nevertheless, in order to test if the co-assembly strategy proposed in this study actually enhances the diversity coverage of the metagenomic sample, it should be tested with the rest of big-scale metagenomic datasets.

In addition to the 37 Arctic metagenomes, we also used 27 metagenomes from temperate seawater and four from the Southern Ocean, collected during the Tara Oceans expedition (TO), to define

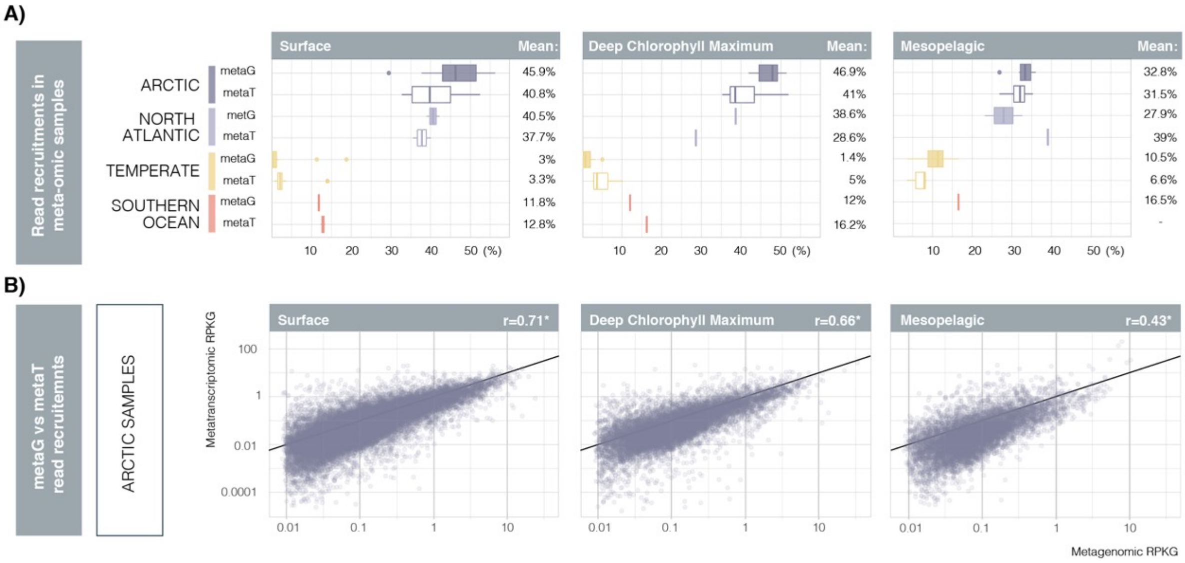


FIGURE 15. Trends in meta-omic read recovery by all 3550 bins. (A) Distribution of meta-omic reads' recovery by all 3550 reconstructed genomes per sample. Samples are divided into their layer (columns) and latitudinal range (purple boxes for Tara Oceans Polar Circle, yellow box for Tara Oceans Expedition and red box for Southern Ocean samples from the Tara Oceans Expedition). Metagenomic samples are represented by filled boxplots, metatranscriptomic samples are represented by empty boxplots. Mean read recovery per group of samples is indicated at the right side of each plot. (B) Pearson correlations between individual read recruitments in metagenomic and metatranscriptomic samples. It includes all read recruitments in arctic samples by all 3550 bins by water column layer. Asterisks represent significant correlations (p-value <0.05). Normalized recruitment units are RPKG (reads per genome kilobase and sample gigabase).

polar pan-arctic and global patterns. Compared to these, TOPC MAGs recover 13% of Tara Oceans' Southern Ocean metagenomic reads and 4.8% of the non-polar.

Mesopelagic layers show a more homogeneous amount of diversity in the latitudinal gradient than photic layers, reaching values similar to those of temperate waters (Ibarbalz et al., 2019), which could explain the decrease in the mean recruited reads of samples in the mesopelagic layers compared to the photic of both Arctic and North Atlantic samples (Figure 15A). This occurs in metagenomic and metatranscriptomic recruitments, with the exception of the mesopelagic North Atlantic, that recruits slightly more reads than surface seawaters. On the other hand, mean metagenomic read recovery increases with depth in temperate and Southern Ocean metagenomes suggesting that some Arctic MAGs may have a bi-polar distribution in the mesopelagic layers. This could be due to less variable conditions of the deep waters compared to the surface and the poles' connectivity through ocean circulation (Aagaard et al., 1985; Ghiglione et al., 2012). Nevertheless, the same increase with depth is not followed by metatranscriptomic read recovery in temperate waters. Southern Ocean photic metatranscriptomes recovered four-fold those of temperate latitudes.

Metagenomic read recruitments by individual MAGs were very similar at different latitudes, with a broader recruitment range in photic Arctic samples reaching up to 8% of total community reads (SM Figure 8) by LQ bins like Crenarchaeota MAGs ICM_TOPC-bin-748 and ICM_TOPC-bin-491. Highest percentages in MQ and HQ MAGs reach ~1.5% by Gammaproteobacteria ICM_TOPC-bin-1423 (Thioglobus sp.) or ICM_TOPC-bin-755 (*Oleispira* sp.). Indeed, the TOPC MAGs dataset includes abundant representatives of the community but also members of the rare biosphere. Individual metatranscriptomic recruitments by our TOPC MAGs dataset tend to be lower in temperate latitudes in all layers (SM Figure 8), which could suggest that even though the deep currents could connect polar prokaryotes, most might remain in resting stages during transit through non-polar latitudes until reaching favorable latitudes like the Southern Ocean (Jones and Lennon, 2010).

As a general trend, there was a significant and good correlation between whole-genome metagenomic and metatranscriptomic read recruitment by individual MAGs in Arctic samples (Figure 15B). Similar correlations have been found at the gene level in the marine microbiome (Salazar et al., 2019) or in the beef rumen and human gut, suggesting that, as may be expected, expression profiles are dependent on gene abundance (Franzosa et al., 2014; Li et al., 2019). Nevertheless, the strength of the correlation decreased with depth and was weaker in temperate latitudes (Figure 14B, SM Figure 9), showing a tendency of MAGs recruiting less metatranscriptomic reads than metagenomic reads. This could be explained by genomes that have been vertically exported from their putative preferred photic niches to the mesopelagic and those that are being transported by the deep currents to more temperate latitudes. These results also reinforce the polar habitat preference of a significant fraction of our TOPC MAGs.

This significant positive correlation in metatranscriptomic vs metagenomic recruitment is strong in 37% of phyla (Nitrospinota, Dadabacteria, Verrucomicrobiota, SAR324, Planctomycetota, Proteobacteria and Crenarchaeota), moderate in 47% (Bacteroidota, Marinisomatota, Acidobacteriota, Latescibacterota, Asgardarchaeota, Actinobacterota, Gemmatimonadota and Chloroflexota) and weak in 16% (Cyanobacteria, Myxococcota and Thermoplasmatota)(SM Figure 10).

The 3550 TOPC bins are a good representation of Arctic marine taxa, as they: i) recover half of the sequenced metagenomic dataset from the Arctic metagenomes, including some of the communities' abundant members but also a majority of low-abundance and rare prokaryotes; and ii) they are actively expressed members in the ecosystem.

3.4.3 Insights into the ecology and metabolic processes of the TOPC HQ and MQ MAGs

Taxonomic classification and novelty

Since only 27 of the MAGs (5%) encoded full or partial ribosomal RNA genes (SM Figure 11), the taxonomic annotation and novelty of the MQ and HQ TOPC MAGs was assessed with a phylogenomics approach against a database that includes both cultured and uncultured taxa (Parks et al., 2018a). The taxonomic nomenclature proposed in Parks et al. (2018a) is the one followed throughout this work.

The MQ and HQ MAG dataset is composed of 473 Bacteria and 58 Archaea, all of them assigned to a known phylum (Figure 16A). Proteobacteria and Bacteroidota were the predominant phyla in the Bacteria domain (182 and 111 MAGs respectively, making 34.3% and 20.9%), as previously reported for Arctic waters (Pedrós-Alió et al., 2015). Thermoplasmatota was the dominant phylum in Archaea (50 MAGs, 9.4%), which had been previously assigned to the Euryarchaeota phylum in ribosomal phylogenies but constituting a new phylum according to phylogenomics (Parks et al., 2018a). Most represented classes in these three dominant phyla were Alphaproteobacteria (61, 11.5%), Gammaproteobacteria (121, 22.8%), Bacteroidia (111, 20.9%) and Poseidoniia (50, 9.4%). Poseidoniia Archaea (or MGII) show a global distribution (Zhang et al., 2015) and they have been previously found in the Arctic waters (Bano et al., 2004), especially in the Beaufourt sea, where waters were richer in labile organic matter of surrounding land (Galand et al., 2006). Alpha- and Gammaproteobacteria representation in the Arctic waters tends to vary through the year, as it has been postulated that Alphaproteobacteria is favored by longer open water periods, implying that high proportions of Gammaproteobacteria may be linked to seasonal ice-formation (Pedrós-Alió et al., 2015). Together with Bacteroidetes, they are commonly the most abundant phyla in Arctic seawater studies (Pedrós-Alió et al., 2015). Betaproteobacteria are typical freshwater bacteria, common in the Arctic due to the high river influence of its waters (Carmack, 2007; Galand et al., 2008) and have also been found as ephemeral blooms (Alonso-Sáez et al., 2014). Due to a recent standardized bacterial taxonomy (Parks et al., 2018a), Betaproteobacteria MAGs in the dataset are included in the Gammaproteobacteria class. Other Bacteria phyla represented by more than 18 MAGs (3.3%) in the dataset are Chloroflexota, Verrucomicrobiota, Actinobacterota, Planctomycetota and Marinisomatota. Archaea MAGs include Halobacterota (previous Euryarchaeota) and Crenarchaeota phyla (previous Thaumarchaeota). Members of the phylum Chloroflexota had been reported in many cryosphere habitats like permafrost, sea ice, arctic soil and seawater (Bano and Hollibaugh, 2002; Boetius et al., 2015; McCann et al., 2016). There is genomic evidence for their ability to degrade terrestrial organic matter (Colatriano et al., 2018), which makes them key species in current increase of terrestrial dissolved organic matter (DOM) fluxes and loadings into the Arctic Ocean (Frey and McClelland, 2009; Vonk et al., 2012; Bintanja and Andry, 2017; Colatriano et al., 2018). Verrucomicrobiota are found in a diversity of habitats globally in low abundances, while in the Arctic's Svalbard Smeerenburgfjord they were found especially dominant in the deep waters, but also abundant in the surface waters and the sediment's surface (Cardman et al., 2014). The gram-positive Actinobacteria have been extensively explored via culture-dependent and culture-independent methods in polar sediments (Bienhold et al., 2012; Jorgensen et al., 2012; Zhang et al., 2014a; González-Rocha et al., 2017; Millán-Aguiñaga et al., 2017), as well as Antarctic lakes, mats and sea ice (Wilkins et al., 2013) and it has also been found as a major clade in Arctic seawater (Bowman et al., 2012) and ice (Collins et al., 2010). Planctomycetota had previously been found in the Arctic Mid Ocean Ridge (Storesund and Øvreås, 2013), in Antarctic continental shelf sediment (Bowman and McCuaig, 2003), also in river biofilms (Brümmer et al., 2004). Their abundance in Arctic sediments was found greater than in Arctic seawater (Teske et al., 2011). Marinisomatota, previously known as Marinimicrobia or SAR406 clades, has been observed in winter polar surface waters (Ghiglione et al., 2012), under the sea ice in Beaufort sea (Collins et al., 2010) and more abundant in early spring and autumn than during

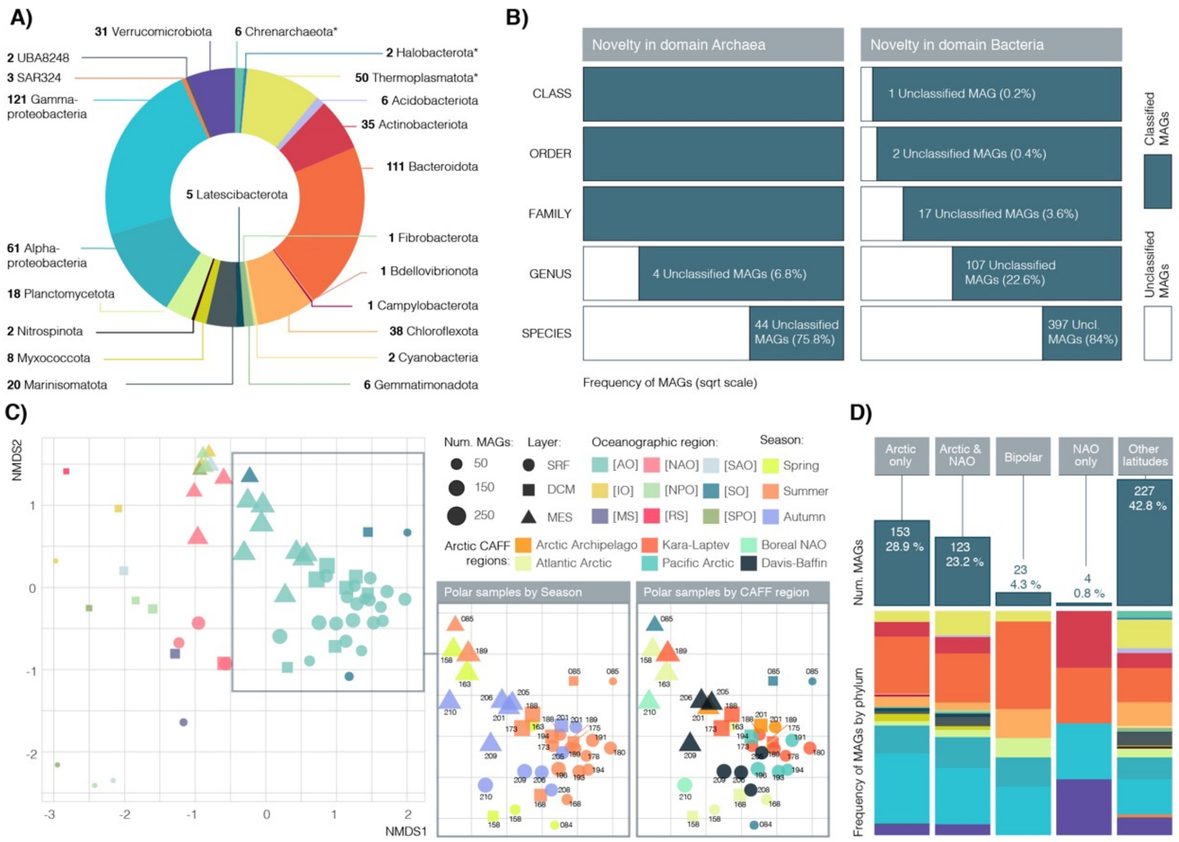


FIGURE 16. Medium and high-quality MAG dataset. (A) Phylogenomics-based taxonomic classification of the 530 medium and high-quality MAG dataset at the phylum level, excepting Proteobacteria that have been split at the class level. Archaea phyla are highlighted with an asterisk, annotations without asterisk belong to the Bacteria domain. (B) Taxonomic novelty quantification of the medium and high-quality MAG dataset in stacked-bars with square-root scale in the X axis. (C) Ordination of metagenomic samples based on their composition of medium and high-quality MAGs using the Non-metric Multidimensional Scaling (NMDS) approach. Shape defines the sample's layer in the water column and the dot size represents the MAG richness of the sample. The plot on the left is colored based on the oceanographic region of the samples. Polar samples are further coloured by season (in the middle) and CAFF region (on the right) and labeled with their Station number. (D) Biogeographic categorization of medium and high-quality MAGs. Bar plots represent the number of MAGs into every category and the percentage within the medium and high-quality dataset. The lower stacked bar plot shows their taxonomic annotations following the color coding of panel (A).

summer in Arctic seawaters (Sipler et al., 2017) and the free-living size fraction in Antarctic waters (Luria et al., 2016). Crenarchaeota (Thaumarchaeota) had been found to be dominant in polar winters (Alonso-Sáez et al., 2012), while Halobacterota (Euryarchaeota) were found unexpectedly dominating the archaeal assemblage of the deep Atlantic water masses from the central Arctic (Galand et al., 2009b) but very low throughout the year in surface waters of the Western Arctic (Kirchman et al., 2007; Alonso-Sáez et al., 2008).

The degree of taxonomical novelty within this group of Medium Quality MAGs is notorious (Figure 16B). In the Archaea domain 4 MAGs cannot be classified further than family rank, and 44 (75% of the Archaea MAG subset) cannot be classified as any known species. In the Bacteria domain, one MAG cannot be classified further than phylum Latescibacterota, another cannot resolve further than class Lentisphaeria (Verrucomicrobiota). Novelty increases at the family rank, 22% of

Bacteria MAGs cannot be assigned a genus and 16% belong to a known species. For 16S rRNA taxonomic annotations, despite 58% of 16S rRNA encoding MAGs could not be classified into a domain (at 95% identity), we also find an increasing novelty with more resolutive taxonomic ranks.

Towards a biogeographical mapping of TOPC MAGs

We have delineated the geographical distribution of the HQ and MQ MAGs mapping them against metagenomic reads from 68 metagenomic samples, covering various regions of the global ocean at different water layers. These include the 37 TOPC metagenomes, including those metagenomes from true Arctic latitudes and North Atlantic metagenomes, and 31 Tara Oceans metagenomes from temperate latitudes (27 samples) and the Southern Ocean (4 samples).

Narrowing the metagenomic dataset to only true polar latitudes, there are 153 MAGs found only in Arctic samples (28.9%) and 23 (4%) show a bipolar distribution (Figure 16D). As the co-assembly included metagenomes from the Tara Oceans Polar Circle expedition that belong to the North Atlantic Ocean, they were separated from the temperate latitudes but also from the Arctic metagenomic dataset. Almost 25% of MAGs were present only in the Tara Oceans Polar Circle Arctic and North Atlantic samples, while 4 MAGs were exclusive of the North Atlantic metagenomes. These North Atlantic specific MAGs could be future signatures of the "atlantification" degree of the Arctic Ocean due to global warming, that has been already detected in the Barents Sea (Barton et al., 2018).

Drivers of MAG community composition

Depth stratification of the marine microbiome had previously been described for the global temperate ocean, including samples from the Southern Ocean (Sunagawa et al., 2015a). Sample ordination based on their TOPC MAG composition sets polar samples and temperate samples apart, and also shows that depth stratification, rather than season or geographic location, drives the composition of MAGs in the Arctic Ocean (Figure 16C).

Niche breadth of MQ and HQ MAGs in the Arctic

In order to explore which MAGs were representative of ecologically relevant Artic prokaryotic taxa, we calculated their niche breadth and delineated habitat generalists and specialists across the Arctic metagenomic dataset. Habitat generalists can use a broader range of habitat types than specialists. The availability of habitats is subject to environmental changes, which is proposed to have a stronger effect on specialists than on generalists (Pandit et al., 2009; Logares et al., 2013; Liao et al., 2016). Specialists might have strict environmental requirements, reducing their ability to survive under unfavorable conditions during dispersal, even undergo extinction after drastic environmental disturbances. Generalists, on the other hand, are less dependent on environmental conditions, have a wide habitat tolerance, and high functional plasticity (Székely et al., 2013). In the current state of climate change, it is imperative to identify the Arctic's taxa that are either more susceptible or resilient to environmental change.

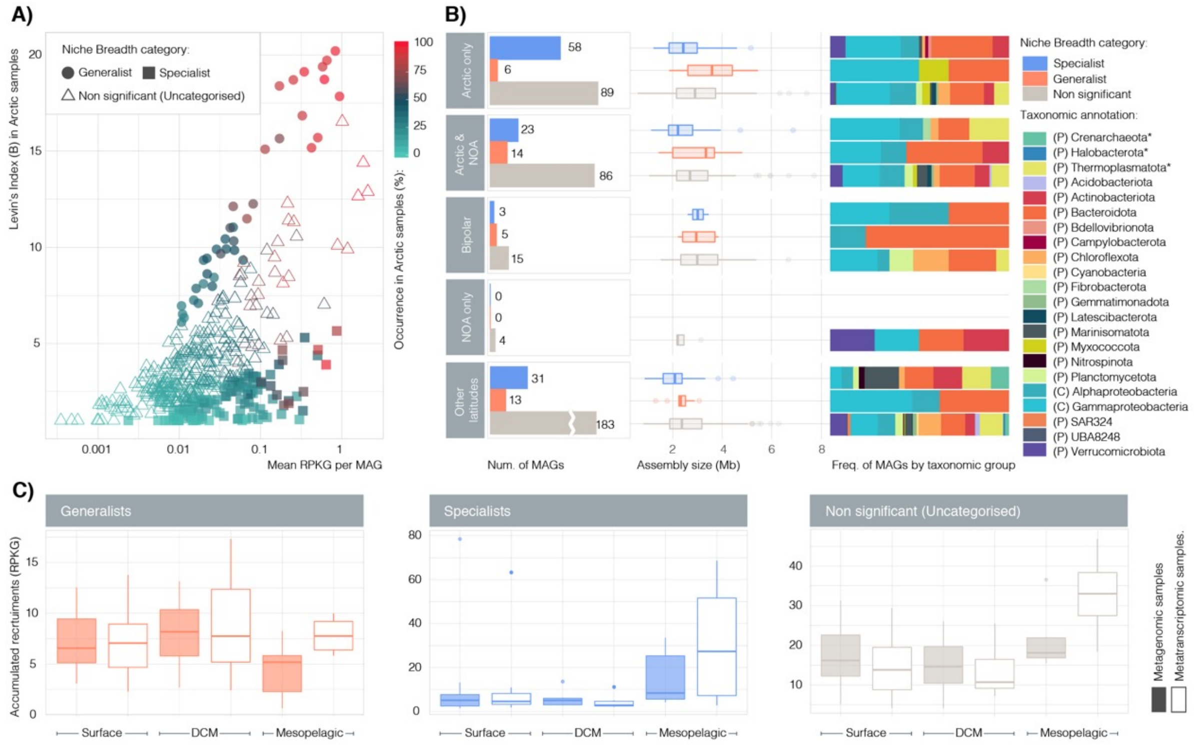


Figure 17. Niche breadth of medium and high-quality MAGs. (A) Distribution of habitat generalists or specialists based on their mean read recruitments in metagenomic samples (RPKG, X axis) and their Levin's Index (Y axis). Color gradient depicts the occurrence in the Arctic metagenomic dataset and shape defines their niche breadth category. (B) Quantification of habitat generalists (orange), specialists (blue) and uncategorised MAGs (grey) by biogeographic category in bar plots. Adjacent boxplots show the distribution of assembly sizes within each subcategory (upscaled to 100% of genome completeness). Stacked barplots are representative of their taxonomic annotation. Asteristsk in the taxonomic annotation legend indicate phyla from domain Archaea, a lack of asterisk indicates domain Bacteria. (C) Abundances of generalist (orange), specialists (blue) and uncategorised (grey) MAGs in metagenomic (filled boxplots) and metatranscriptomic (empty boxplots) samples. There are no significant differences between the groups. Note that Y axis scales are different between categories.

In our dataset, each Arctic sample was considered an individual habitat, as their geographical location, depth in the water column, and season of sampling differed. In line with previous studies (Liao et al., 2016; Lindh et al., 2016), the majority of MAGs (71%) could not be categorized into generalists or specialists, while 21% (115) were habitat specialists and 7% (38) were generalists (Figure 17A). Higher numbers of habitat specialists are also in agreement with previous observations in species-rich communities (Romanuk and Kolasa, 2005; van der Gast et al., 2011; Logares et al., 2013; Liao et al., 2016). Both generalist and specialist MAGs show a similar range in their mean abundances, with the least widespread MAGs in the Arctic samples being specialists. These results are consistent with other studies (Pandit et al., 2009; Logares et al., 2013; Liao et al., 2016), but also contradict those of Székely and Langenheder (2014). Generalist MAGs show a narrower standard deviation in their mean recruitments than specialists, suggesting a greater evenness in their distribution across arctic samples (SM Figure 12).

As habitat generalists may likely adapt to a wider range of habitats due to their functional plasticity (Székely et al., 2013), their genome size would be expected to be larger than habitat specialists.

This difference is only apparent in the median genome size of MAGs with an Arctic and North Atlantic distribution (Figure 17B). MAGs with panarctic/ bipolar distribution make up 53% of total specialists and 30% of all generalists. All of them belong to the Bacteria domain, while we find eight Thermoplasmatota MAGs (previously classified as Euryarchaeota) in the pool of uncategorised polar MAGs.

While generalists are assigned to Bacteria phyla Actinobacteria, Proteobacteria, Bacteroidetes and Myxococcota (Figure 17B), specialists displayed larger taxonomic diversity, being additionally classified into nine other phyla including Archaea's Thermoplasmatota and Crenarchaeota. It is suggested that one can predict the phylogenetic depth at which a trait is conserved (Martiny et al., 2013) and it had been proposed that habitat specialization is a complex trait conserved at the phyla or class level (Lennon et al., 2012; Székely and Langenheder, 2014). We find that habitat specialization is conserved at the phyla or class level in our set of HQ and MQ MAGs with some exceptions such as in phyla Actinobacteria (it is conserved at the family level), Bacteroidota (conservation is reached at genus level for Cytophagales but no conservation is found for Flavobacteriales) and Proteobacteria (where no conservation is found for Rhizobiales, Pseudomonadales and SAR86).

Interestingly, generalist MAGs are recruit more meta-omics reads in the Arctic photic layers than the mesopelagic, while specialists MAGs recruit more reads in mesopelagic samples (Figure 17C). Despite these differences, the distribution of recruited reads per sample by both specialists and generalists in the photic layers are very similar. Photic arctic samples show wider gradients of temperature and salinity than mesopelagic waters, which could be a reason why habitat generalists, that have the ability to be resilient and thrive in various environmental conditions, are enriched in the photic depths. Habitat specialists are also present in photic waters, but in lower magnitudes compared to the mesopelagic. Specialization in the deeper layers could probably be linked to a narrower range of potential niches, such as different composition of organic matter and/or available nutrients.

Bacteria generalists and specialists recruit more metagenomic reads in the summer samples (Kara-Laptev and Pacific Arctic regions) when looking into the photic layer (SM Figure 13). Generalists are more abundant in the Kara-Laptev waters, depleted in nitrate and richer in bicarbonate and iron than the Pacific Arctic, where specialists are more present and there are higher concentrations in nitrite, ammonium and phosphate. In the mesopelagic, high recruitments of generalists happen during autumn samples (Davis-Baffin, Atlantic Arctic), while specialists dominate the Atlantic Arctic and summer Kara-Laptev regions. Overall, it seems like habitat generalists are more present in areas where phytoplankton blooms (spring or fall) have occurred, independently of the layer.

In the case of specialist Archaea (SM Figure 14), in surface waters there is more metagenomic and metatranscriptomic recruitment in the Davis-Baffin region, sampled during autumn. In the DCM, highest presence and activity is found in the Atlantic Arctic, which also shows the highest amount of reads in the mesopelagic layer, together with the Kara-Laptev and Davis-Baffin region.

We find that the environmental features that correlate stronger with meta-omics distribution include (SM Figure 15): i) carbonate and bicarbonate concentrations for generalist SAR86 Gammaproteobacteria and specialists Chitinophagales (Flavobacteria), Fibrobacterales and Enterobacteria; ii) nitrate for generalists Woeseiales (Gammapr.), Pelagibacterales (Alphapr.) and Cytophagales and specialists Marinisomatota and Plantomycetes; and iii) ammonium concentration for generalists Rhizobiales, UBA6615 and SAR86 and specialists Poseidoniales (Thermoplasmatota) and Chitinophagales (Flavobacteria).

For the set of MAGs that could not be classified into generalists or specialists, there is a common positive correlation with nitrate concentrations in both meta-omics abundances, which is especially strong in Archaea, Myxococcota, Planctomycetes, Alphaproteobacteria, SAR324 and Verrucomicrobia (SM Figure 16). Salinity is also positively correlated, albeit weaker, in the majority of taxonomic groups.

Photoheterotrophy in the Polar Arctic TOPC MAGs

Photoheterotrophic light harvesting via rhodopsin and bacteriochlorophyll is widely utilized by aquatic microorganisms as a supplementary source of energy (Beja, 2000; Pinhassi et al., 2016). In the recent sequencing effort of more than 12,000 SAGs (GORG-Tropics database; (Pachiadaki et al., 2019)) genes related to photoheterotrophy were found in 58% of the genomes, while in our MQ and HQ TOPC MAGs dataset, proteorhodopsin was found in 24% of MAGs and pufM in 2% of MAGs.

Generally, green-light proteorhodopsins show fast photo-cycles and are associated with proton pumping, whereas the slower photocycles of blue-light proteorhodopsins are associated with environmental sensing (Fuhrman et al., 2008). MAGs encoding green- or blue-light types of proteorhodopsin were expressing these genes in all the CAFF regions in all samples, with the exception of the mesopelagic waters from the Pacific Arctic and Arctic Archipelago (Figure 18). They were found in nearly all samples in Actinobacteria, Proteobacteria and Bacteroidota, while Verrucomicrobia encoding this light-dependent proteins expressed the genes in North Atlantic samples and in late-summer/early autumn stations (SM Figure 17B). Interestingly, we find blue-light sensory proteorhodopsins in MAGs that are only expressed in the deep waters, Chloroflexota and Marinisomatota (previously proposed as Marinimicrobia). To our knowledge, the only case of a proteorhodopsin encoding Chloroflexi was found in the hypolimnion of Lake Michigan (Denef et al., 2016). The bin showed a ubiquitous distribution in the water column during spring, preferring the deep during summer stratification. A previous record of a proteorhodopsin encoding Marinimicrobia bin was most expressed in surface waters (Hawley et al., 2017). On the contrary to the Arctic Chloroflexi and Marinisomatota, these bins show an increased proteorhodopsin expression in the surface waters. Nevertheless blue-light penetrates further in the water column, which could explain why the sensory proteorhodopsin would be an ecological advantage in their preferred habitat, the mesopelagic, and the high expression of Chloroflexi in surface waters during darker seasons like autumn (Figure 18B). Interestingly, none of the Thermoplasmatota Archaea

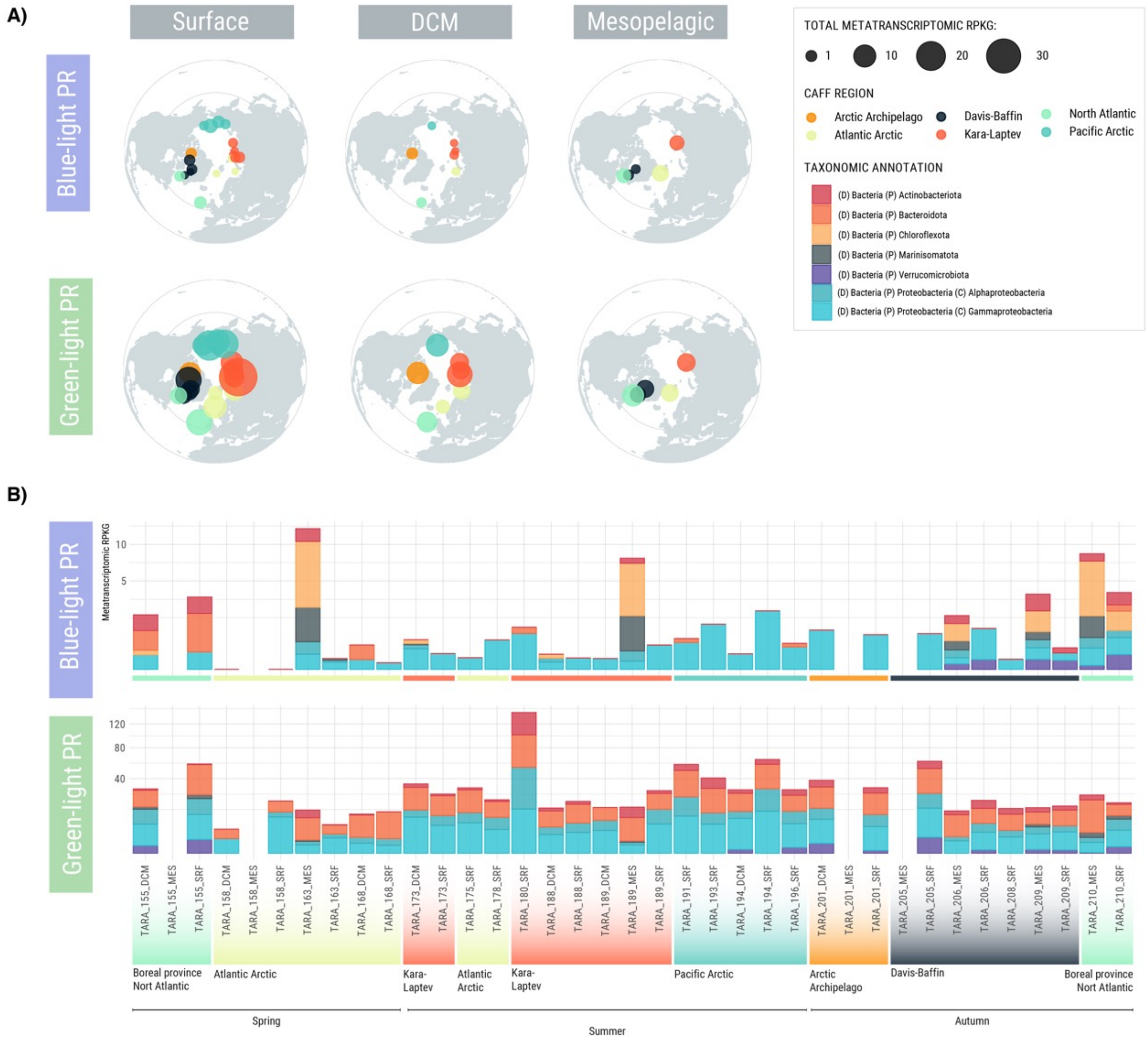


FIGURE 18. Accumulated expression of HQ and MQ MAGS encoding key PFAM PF01036 for protorhodopsin, classified by light wavelength tuning (blue or green). Polar maps with the accumulated metatranscriptomic RPKGs of all MAGs encoding proteorhosospin. Absent bars mean that no recruitment was found for that specific metabolism, domain and/or layer.

MAGs from the class Poseiidonia encode proteorhodopsin. Photoheterotrophy in this group was extensively studied in (Rinke et al., 2019) in temperate and Antarctic waters distributed globally and is described as a characteristic feature of the clade, excepting some genera that lost its PR gene together with other light-related genes to adapt to deeper-layers (Rinke et al., 2019) and are classified as obligate mesopelagic genomes (Tully, 2019).

Aerobic anoxygenic photoheterotrophy (AAP) is relatively common in the global euphotic ocean (Schwalbach and Fuhrman, 2005) and we find that MAGs encoding the diagnostic gene for the anoxygenic photosystem II, pufM, were expressed in the majority of Arctic samples, with a preference for the surface and especially the Atlantic-influenced regions (Figure 19A) and were members of the Alpha and Gammaproteobacteria classes (Figure 19B). Abundance of AAP or PR containing bacteria had been reported to decrease from summer to winter in the Arctic Ocean

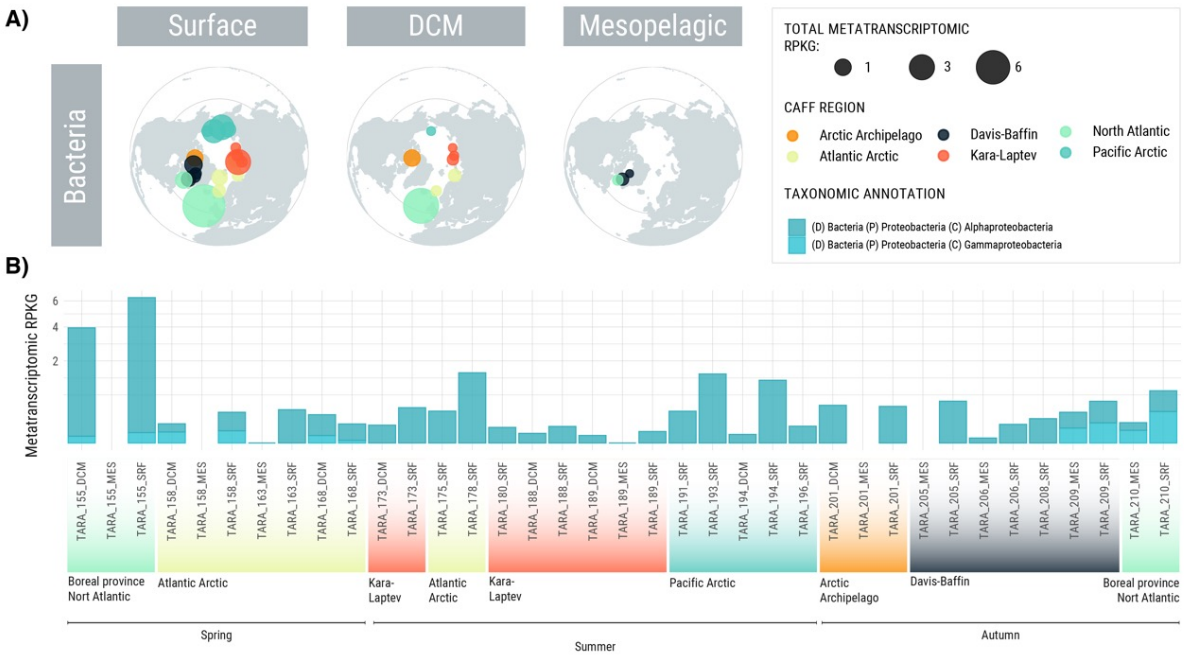


FIGURE 19. Accumulated expression of HQ and MQ MAGS encoding key marker gene pufM K08929 from the anoxygenic photosystem II. Polar maps with the accumulated metatranscriptomic RPKGs of all Aerobic Anoxygenic Phototrophs (AAPs). Absent bars mean that no recruitment was found for that specific metabolism/domain/layer.

(Cottrell and Kirchman, 2009), a trend that fits expression patterns of AAPs (Figure 19B) but not PR-encoding genomes. Nguyen et al., (2015) had also reported expression patterns increasing especially during the onset of summer (July). The amount of recruited reads by green-PR coding MAGs seems slightly higher during that period of time, but are not as different compared to late spring and autumn recruitments as previously reported.

Carbon metabolism by Polar Arctic TOPC MAGs

Inorganic carbon fixation has been previously detected in the Arctic deep oceans and surface waters during the winter period in the Canadian Archipelago, Western Arctic, in which Archaea of the phylum Thaumarchaeota played a main role performing nitrification (Alonso-Sáez et al., 2012). However, the relevance and ubiquity of different inorganic carbon fixation pathways across different CAFF regions, depths and seasons is unknown, as well as who the potential key players are. Calvin cycle was the main strategy for inorganic carbon fixation in our 530 HQ and MQ MAGs. MAGs encoding rubisco or rubisco and phosphoribulokinase genes are expressed in the 6 different CAFF regions at all depths (Figure 20A). Photosynthetic Synechococcus MAGs are only present in North Atlantic samples, thus potential chemolithoautotrophy by RuBisCo encoding MAGs appear to be a ubiquitous process occurring from spring to autumn by mostly Gammaproteobacteria and Actinobacteria (Figure 20B). Interestingly, we reconstructed MAGs encoding RuBisCo that belong to the phylum Crenarchaeota. RuBisCo had been previously reported for Archaea genomes (Kono et al., 2017). Nevertheless, our MAGs lack a homologue of the phosphoribulokinase gene that is essential in carbon fixation in this Domain, which has only been found in methanogenic Archaea

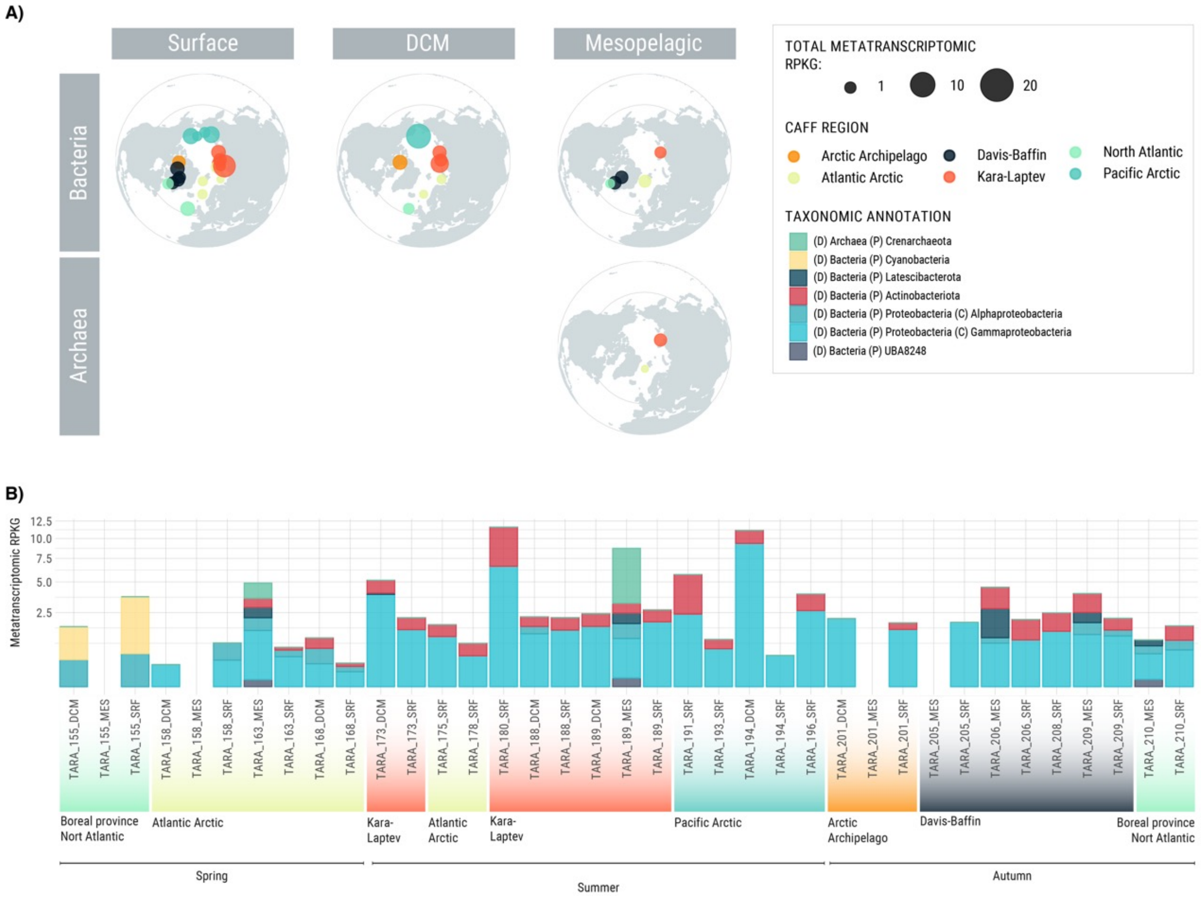


FIGURE 20. Accumulated metatranscriptomic RPKGs of HQ and MQ MAGS encoding RuBisCo in the Arctic Ocean. (A) Polar maps with the accumulated metatranscriptomic RPKGs of all MAGs encoding at least one subunit of RuBisCo for the Calvin cycle pathway (K01601, K01602). Absent plots mean that no recruitment was foudn for that specific metabolism/domain/layer. (B) Stacked bars with the accumulated metatranscriptomic RPKGs of all MAGs encoding at least one subunit of RuBisCo for the Calvin cycle pathway (K01601, K01602). Absent bars mean that no recruitment was found for that specific metabolism/domain/layer. Bars are colored based on taxonomic annotation at the phylum level.

so far (Kono et al., 2017). RuBisCo coding Bacteria MAGs that were actively expressing the genes included the phyla Latescibacterota and UBA8248 previously unrecognized even in the recent GORG-Tropics SAGs database (Pachiadaki et al., 2019). In the mesopelagic waters, RuBisCo encoding MAGs belonged to the mentioned taxa and to Actinobacteria, and Proteobacteria, and were all active in the summer Atlantic Arctic and Kara-Laptev Sea, while only Bacteria were actively fixing CO_2 in the autumn waters of the Davis-Baffin and North Atlantic stations.

There were also MAGs that encoded certain genes related with the 3-Hydroxypropionate bi-cycle and the Hydroxypropionate/4-Hydroxybutyrate Cycle. The marker genes malyl-CoA/(S)-citramalys-CoA lyase and propionyl-coA carboxylase from the 3-Hydroxypropopionate bicycle were present in 40 Bacteria MAGs from phyla Actinobacteriota, Chloroflexota, Myxococcota and Alphaproteobacteria, and were expressed widely in the photic layers and restricted to the autumn samples and Kara-Laptev region in the mesopelagic (SM Figure 19). As they lacked complementary genes for carbon fixation like the acetyl-CoA carboxylase, their autotrophic potential remains putative. A similar case was that of MAGs with key marker genes in the 3-Hydroxypropionate/4-Hydroxybutyrate cycle (SM Figure 20), like the 4-hydroxybutyryl dehydratase and the 3-hydroxypropionyl-coenzyme A. They were active in the same regions as the 3-Hydroxypropionate containing MAGs with a mesopelagic preference, but lacked the three extra enzymes for the pathway to be functional (3-hydroxypropionate dehydrogenase (NADP+), 3-hydroxypropionyl-coenzyme A synthetase, acryloyl-coenzyme A reductase).

It is well know that heterotrophic microorganisms can incorporate CO_2 in the dark through different metabolic pathways without any net carbon assimilation and linked to fatty acid biosynthesis, or anaplerotic reactions. Mixotrophy has been proposed to be relevant for specific Arctic heterotrophs (Alonso-Sáez et al., 2010) and therefore our TOPC MAGs may be not all performing chemolithoautotrophy but rather mixotrophy. Future experimental validation would be necessary to confirm such processes.

Additionally, CO-oxidation was explored as a potential alternative energy source for heterotrophy microorganisms in which mixotrophy is possible (Cordero et al., 2019). The oxidation of carbon monoxide is catalyzed by carbon monoxide dehydrogenase (CODH; cox genes) (Ragsdale, 2004; King and Weber, 2007) and it is carried out by Actinobacteria, Proteobacteria, and taxa from Bacteroidetes and Chloroflexi (Cordero et al., 2019; Martin-Cuadrado et al., 2009). MAGs encoding the genes for the oxidation of carbon monoxide (CO-oxidation) were widely distributed in all samples (Figure 21A) and belonged to twelve different phyla (Figure 21B). This metabolism is putatively present in all niche breadth categories (generalist, specialists and uncategorized), associated to Alphaproteobacteria Rhodobacteraceae and Gammaproteobacteria Burkholderiaceae, Spongiibacteraceae and HTCC2089 (Figure 22).

3.4.4 Genome resolved Ecology of the Arctic Ocean prokaryotic sentinels

In order to define key prokaryotic genomes specific of the poles we chose those MAGs displaying a polar-only metagenomic distribution and that showed the highest expressed in every sample, within their niche breadth category. Here we present a summary of the most ecologically relevant polar specific taxa, that should be present in future monitoring events or seasonal samplings with the aim to inspect the health of the Arctic Ocean and its evolution in the climate change scenario.

Generalist MAGs showed a higher number of genomes with bipolar distribution, especially (photo-)heterotrophic *Polaribacter* spp. with the most intense expression in the vast majority of photic samples (Figure 22). Polaribacter is one of the most common genera in polar waters and its bipolar distribution had been defined previously (Staley and Gosink, 1999). Another heterotrophic Flavobacteria (UA16 family) dominated gene expression in the mesopelagic, together with a MAG from the Myxococcota family UBA4427 that is a potential chemolithoautotroph (Figure 22). Metatranscriptomic recruitment from generalist Alphaproteobacteria is highest in specific samples from the photic Pacific Arctic and Arctic Archipelago, and remarkable in the Davis-Baffin's

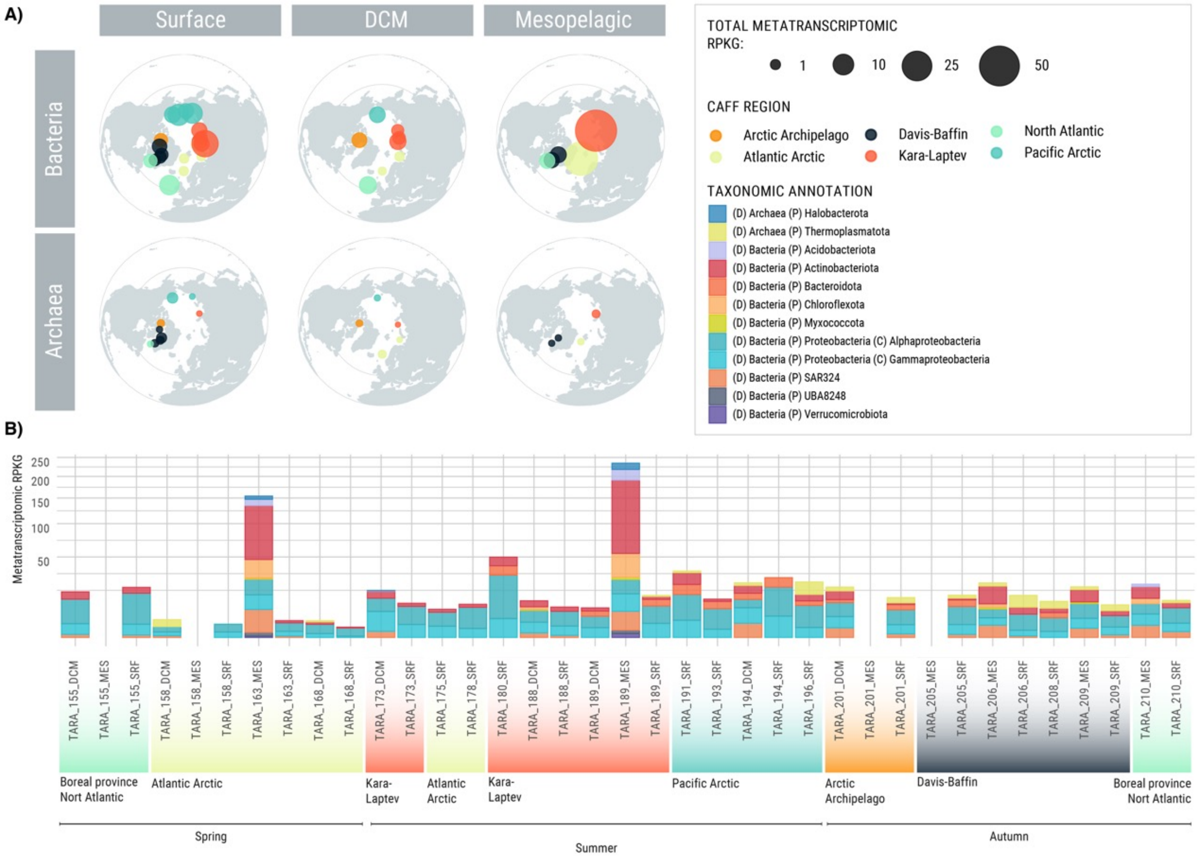


FIGURE 21. Accumulated expression of HQ and MQ MAGS encoding key marker gene coxL K03520 from the aerobic carbon-monoxide dehydrogenase. Polar maps show the accumulated metatranscriptomic RPKGs of all coxL coding MAGS. Absent bars mean that no recruitment was found for that specific metabolism/domain/layer.

mesopelagic, being all of them heterotrophs, some with genes related to the oxidation of CO and thiosulfate.

The maximum metatranscriptomic recruitments of specialist MAGs was not dominated by any particular family (Figure 22). The majority of highest recruitments took place by Gammaproteobacteria and Alphaproteobacteria in the summer photic samples, whereas spring and autumn showed highest expression of *Polaribacter* and other Flavobacteria and Verrucomicrobia. The majority of these Gammaproteobacteria were heterotrophic with nitryfing and denitrifying potential and we also found the ability to fix CO_2 via the Calvin cycle in a MAG from the family *Methylophilaceaea*. Alphaproteobacteria also had CO_2 fixing genes via the 3-Hydroxypropionate pathway, one of these, from the genus Sulfitobacter (*Rhodobacteraceae*) also had the pufM gene from the anoxygenic photosystem II. In the mesopelagic, the most expressed specialists were MAGs from the phylum Verrucomicrobiota in spring, Chloroflexota in summer and Marinisomatota in the autumn. Whereas Verrucomicrobiota encode genes related to the metabolism of nitrogen and are putative heterotrophs, as previously described (Cardman et al., 2014), Chloroflexota does not show any markers for chemoautotrophy or other nitrogen or sulfur redox abilities, a part from CO oxidation. From those MAGs that did not fall into a niche breadth category, we found that those showing more expression in surface waters through spring and summer were heterotrophic and potentially chemolitoautotrophic. Alphaproteobacteria, heterotrophic Bacteroidetes, some with denitrifying genes and especially an Archeaea MAG from the family *Poseidoniaceae* MGIIa during the autumn in the Davis-Baffin region (Figure 22). In the DCM spring sample, the same MGIIa MAG was the most expressed. In DCM summer samples, it was mostly photoheterotrophic *Flavobacteriaceae* that showed highest metatranscriptomic recruitments, while putative chemolithoautotrophic Gammaproteobacteria showed more expression in the late summer and autumn samples. Highest expression in the mesopelagic was dominated by MAGs that were either absent or lowly expressed in the photic layers, belonging to the *Thalassocoarchaeaceae* MGIIb of phylum Thermoplastota, two Chloroflexi MAGs and an unclassified MAG within the Myxococcota phylum. These had genes related to CO_2 fixation and several, but not diagnostic, genes belonging to the 3-Hydropropyonate bicycle. The Planctomycetes MAG, expressed preferably in deeper layers, does not encode any unique marker for chemolithoautotrophy or annamox metabolism, despite being previously described in the phylum (Strous et al., 1999; Storesund and Øvreås, 2013).

Overall, for the metabolic pathways that were explored, specialist MAGs displayed a wider metabolic variety than generalists, making them candidates to colonize a wider variety of niches with different environmental settings. Even though photic generalists were mostly heterotrophic, we also found that Myxococcota was a generalist in the mesopelagic with a putative autotrophic metabolism, both being active during the seasons and the geographic locations.

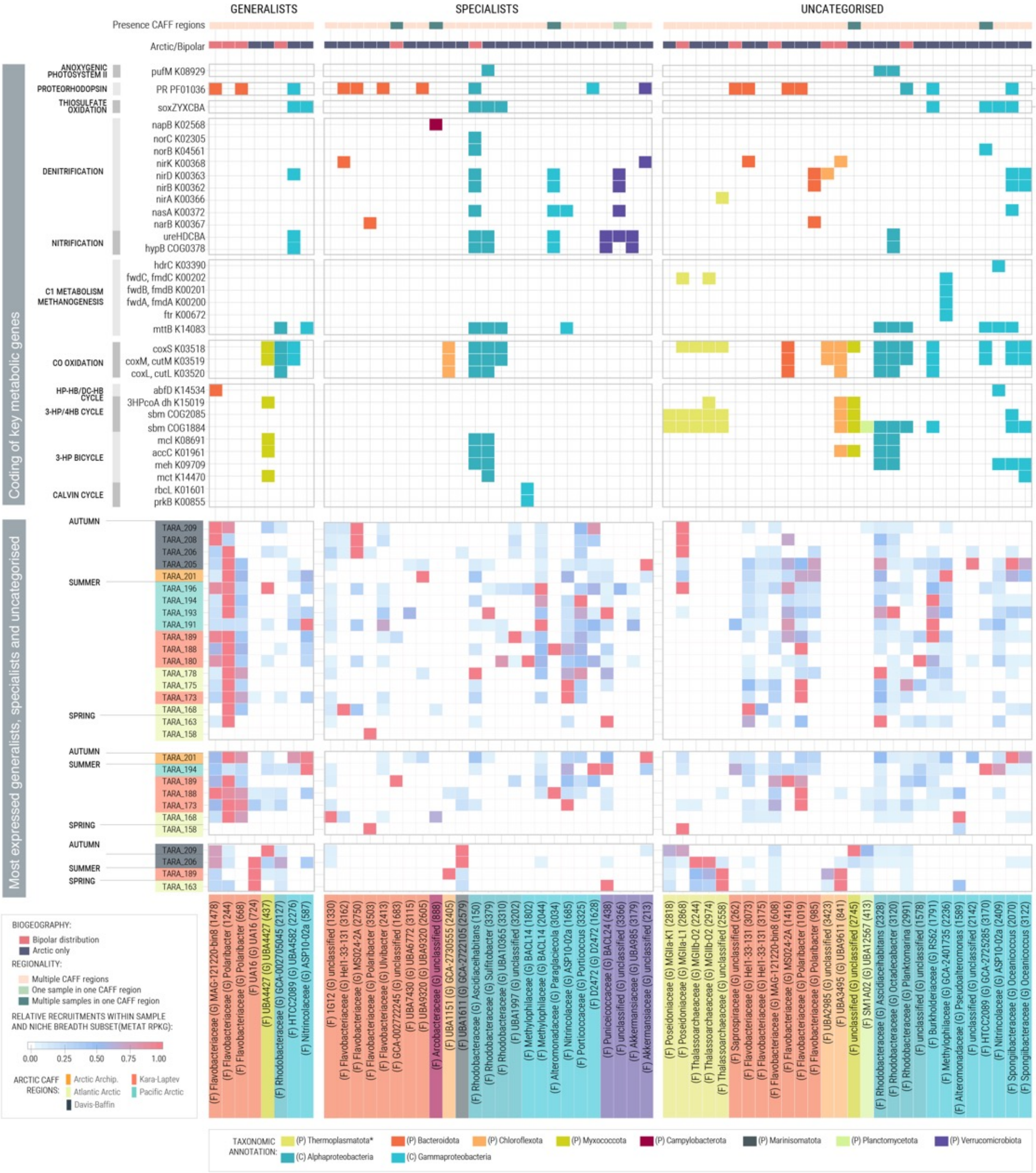


FIGURE 22. Generalists, specialists and uncategorised polar mags showing highest gene expression per sample. Top heatmap represents which of the selected marker genes are encoded per each MAG. Bottom heatmap represents the relative expression of each MAG (X axis) within each sample (Y axis). Normalizations were done for every niche breadth category. Samples in Y axis are coloured based on their CAFF region and the sampling season has been indicated. MAGs in X axis are coloured based on phyla and number in brackets corresponds to the MAGs identification number.

3.5 Conclusion

In this study we have presented a unique catalogue of 3550 bins generated from metagenomes collected during the Tara Oceans Polar Circle Expedition, that circumnavigated the Arctic Ocean from May to November 2013. Samples have been pooled for co-assembly based on their composition in 16S miTAGs, a strategy proposed for the first time in this work, and the resulting bins are able to recover almost half of the metagenomic dataset, which had only been previously achieved by a massive generation of SAGs from the global ocean. Metagenomic distribution of the 530 MAGs meeting medium- and high-quality standards has uncovered a greater number of habitat specialists than generalists in the Arctic waters, with generalists being more abundant in photic samples and specialists preferring the mesopelagic. This could be explained by nutrient availability and niche compartmentalization in the deeper waters versus the wider gradients in nitrate, temperature and salinity of the upper Arctic Ocean and is coherent with the findings of higher metabolic versatility in the specialist pool of polar-specific MAGs. Metatranscriptomic recruitments together with functional annotation has provided the base to propose chemoautotrophic carbon fixation via the Calvin Cycle to be a widely distributed metabolic strategy in Arctic prokaryotes, along the seasons and throughout the water column of the sampled Arctic Ocean. Photoheterotrophy is also ubiquitously expressed from spring to autumn with novel cases of sensory rhodopsins being highly expressed by mesopelagic and autumn related MAGs like Chloroflexota and Marinisomatota. By selecting those MAGs with a polar-specific distribution we have found a larger bipolar distribution within the generalists, and due to even distribution of generalist polar MAGs in the majority of Arctic samples, we have selected those showing higher expression in each sample as putative sentinel species of climate change in the arctic. Additionally, we have also found signature taxa from the North Atlantic Ocean, which would serve as markers for future Atlantification of the Arctic Ocean.

GENERAL DISCUSSION

General discussion

This thesis aimed at gaining insight into the genomics and ecological significance of uncultured marine prokaryotes with different approaches and at different scales, using the powerful Tara Oceans expedition (2009-2013) resources. This dataset provides meta-omics data covering multiple size fractions of planktonic diversity, together with Single Cell Genomics from each sampling event and environmental metadata, opening up new venues to explore the uncultured prokaryotes at fine scale (Karsenti et al., 2011; Pesant et al., 2015). In the present thesis, I have used state-of-art, culture-independent technologies such as Single Cell Genomics and genome reconstruction from metagenomes (MAGs) to put a fraction of the uncultured prokaryotic diversity under the spotlight. Applied in two different environmental settings, the North Indian Ocean and the Arctic Ocean, I have been able to: i) detect an apparently clonal population of novel Flavobacteria, demonstrate its taxonomic novelty and describe its putative relevance in the succession of phytoplankton blooms, with the ability to live in sinking organic-matter particles under low-oxygen conditions and ii) reconstruct the key genomes of multiple microorganisms around the Arctic Ocean, detecting species with a pan-arctic or bipolar distribution and inferring their distribution and expression patterns. At the same time, the co-assembly strategy common in both cases reveals its potential to retrieve high-quality type material for the description of novel taxa and the generation of a rich catalogue of MAGs representative of nearly half of the metagenomic dataset.

The importance of reference genomes in the culture-independent era

Obtaining reference genomes is crucial for the advance of all branches of microbiology and techniques like Single Cell Genomics (SCG) or metagenomic genome reconstruction can fill the gap left by the limitations of isolation in culture. Single Cell Genomics can sort cells at random or it can provide an oriented selection of the microbial population of interest, as cell individualization is based on the combination of microfluidics and cell sorting (Stepanauskas and Sieracki, 2007). Just recently, a randomized Single Cell Genomics approach was applied at large scale with more than 12,000 SAGs offering also quantitative analyses of the distribution of relevant prokaryotic lineages (Pachiadaki et al., 2019). In Chapter 1, we conducted a targeted SAG generation that focused on the heterotrophic prokaryotic fraction of a surface seawater sample from the North Indian Ocean. The pool of heterotrophic SAGs was dominated by a population of unknown *Kordia* spp., representing 98 out of the 117 amplified genomes. The first approach to this dataset relied on analyses alternative to whole-genome sequencing. Firstly, microdiversity was assessed on 78 SAGs with Multi-locus Sequencing Analyses of both core and laterally transferred genes, a technique extensively used in intra-species diversity analysis (Maiden et al., 1998; Feil, 2004). The study provided evidence for an apparent genome clonality within the population. After its phylogenetic placement, based on the 16S rRNA gene sequence, we used the genome of its closest relative to infer their habitat preference. Combining this reference genome and different metagenomes from various locations, depths and fractions, we hypothesized that this putatively novel taxon episodically thrives in the surface ocean after phytoplankton blooms, following a particle-attached lifestyle and sinking to the depths while colonizing the particles. This hypothesis was backed up by the results of Chapter 2, after the sequencing and co-assembly of ten of these North Indian Ocean Kordia SAGs and an extensive functional description and metagenomic mapping. Nevertheless, despite the fact that the North Indian Ocean *Kordia* SAGs and the reference genome AAA285-F05 shared an identical 16S rRNA gene sequence, the two genomic sequences shared less than 70% of their orthologs, so neither their ANI nor AAI identities were reliable for their classification into the same species. This could be possibly due to the small size of the AAA285-F05 SAG fragment (285Kb) and the fact that it did not encode single-copy core genes, suggesting that the genomic fragment may belong to the flexible genome of the species. The massive sequencing of uncultured genomes that has occurred in the last ten years, reaching 11,723 MAG projects and 2,168 projects in the JGI GOLD database (Goh et al., 2019), has been vital for the enrichment of databases and the advancement of the field. Nevertheless high-quality genomic sequences cannot always be achieved. Additionally, the use of reference genomes close to the taxon of interest does not circumvent the fact that the flexible genome of a species' pan-genome can be quite variable, thus interpretation of results should be cautious. In line with these concerns, the use of phylogenomics in genome classification has become more widely used, with the development of curated databases like MiGA (Rodriguez-R et al., 2018) and GTDB (Parks et al., 2018b).

The generation of high quality genomic references should be a priority but single cell genomics and MAG reconstruction can be limiting in terms of genome completeness. In the last years, there has been an investment in the improvement of SAG amplification (Stepanauskas et al., 2017) and assembly (Bankevich et al., 2012), exploration of metagenomic assembly strategies (Tully et al., 2017a; Delmont and Eren, 2018; Delmont et al., 2018; Tully et al., 2017b; Tully, 2019) and MAG binning (Albertsen et al., 2013; Alneberg et al., 2014). The co-assembly of multiple genomes has also been successful in improving genome completeness of SAGs from both protist (Mangot et al., 2017) and bacteria (Rinke et al., 2013; Clingenpeel et al., 2014). However, it is seldom used since even the most similar genomes can have strain heterogeneity, resulting in contig breakage (Sharon et al., 2015; Roux et al., 2017; Sczyrba et al., 2017). To avoid this, in Chapter 2, we performed an analysis on the effect of k-mer length in the co-assembly process with the hypothesis that with such low microdiversity, longer k-mer lengths might help reduce unresolvable branches found in variability hotspots, like in mobile elements and genomic islands (Ricker et al., 2012). After combining the best-performing k-mer lengths in the assembly of normalized and merged pool of reads from the 10 Kordia SAGs, we were able to generate an assembly substantially less fragmented than with default assembly parameters, improving as well computation time and requirements. Manual curation resulted in a high-quality draft genome representative of a novel species that consists of 27 contigs and genomic completeness and contamination estimations of 94.83% and 4.65%, respectively. Tools have been developed to ease manual curation of assemblies (Eren et al., 2015), find the most suitable k-mer for assembly (Chikhi and Medvedev, 2014) and remove chimeras in co-assembled SAGs (Kogawa et al., 2018). Co-assembly has also been used in the generation of marine MAGs. Delmont et al. (2018) pooled Tara Oceans samples based on their geographical location. Nevertheless, considering that the prokaryotic communities sampled in Tara Oceans are structured by depth rather than region of origin (Sunagawa et al., 2015a), we did not follow Delmont's co-assembly strategy. The co-assembly groups in Chapter 3 were defined by the similarities in composition of the metagenomic communities, with the aim to produce a dataset of composite genomes with greater completeness. With the set of high-quality MAGs we found that the generated genomes belonged to different species. Considering that genomes from the same species but different geographic origins were mixed into single MAGs, we cannot discard the fact that MAGs might be a composite of ecotypes. Some species redundancy was found in the genomes of lower quality, which could be explained by differential abundance within the different metagenomes, resulting in them not binning together. Compared to previous big scale MAG generation projects, the Arctic MAGs recruited between two and four times more reads than those of other studies (Parks et al., 2017; Delmont et al., 2018; Tully et al., 2017a). In fact, only the read recovery of the GORG-Tropics dataset of SAGs (Pachiadaki et al., 2019) was similar to our MAGs dataset. One could think that the co-assembly strategy based on community composition is more resolutive than individual assembly or co-assembly of geographically bounded metagenomes, but the difference could be also due to the amount of diversity found in the Arctic, that is lower than the temperate latitudes from which the lower-recruiting MAGs have been generated (Ibarbalz et al., 2019). Assembly, binning and curation are highly variable in the different studies, so a methodological comparison between the different approaches would be very valuable for future studies.

Effect of horizontal genome coverage and rare taxa on recruitment of metagenomic reads

Metagenomic read mapping, also referred to as fragment recruitment, is a widely used approach to estimate a genome's presence and abundance in different samples, especially useful when working with uncultured taxa. Those reads aligning with a sufficient coverage and identity against the reference are considered to be representative of that species in the specific sample (Rusch et al., 2007; Caro-Quintero and Konstantinidis, 2012). Setting a detection limit of target genomes is key for the interpretation of recruitment results, as we have seen in Chapter 3. Recruitment of highly conserved genomic areas is not diagnostic of the reference's presence in the metagenome (Castro et al., 2018). For example, Delmont and Eren (2018) found Prochlorococcus populations recruiting 0.01% of metagenomic reads form the Southern Ocean, where they are virtually absent (Flombaum et al., 2013), and so they assumed that a genome was detected in a given metagenome when at least 50% of their nucleotide positions had 1X read coverage. In Chapter 2 we considered as reliable those mappings spread homogeneously along the genome and in Chapter 3 we analyzed the effect of an array of minimal horizontal coverages in: i) the richness estimation of the 68 metagenomes used in the study and ii) the distribution of MAGs based on their occurrence and mean abundance in the arctic metagenomic dataset. We decided on the threshold from which variation in richness and abundance/occurrence decreased, that is 20%. Even though we followed restrictive, commonly used, parameters for the individual read mapping, considering a minimal horizontal genomic coverage definitely had an impact on the results of our mapping analyses and the rest of ecological analyses that relied on them.

Each step of the metagenomic sequencing workflow may preferentially measure some taxa over others, resulting in systematically distorted estimations of relative abundance (Brooks et al., 2015). Rare genomes might be actually rare, appear rare due to insufficient sampling or sequencing, or made rare because steps in the experimental procedure (i.e. preservation, lysis, nucleic acid extraction) has negatively selected them (Brooks et al., 2015; Costea et al., 2017; Hugerth and Andersson, 2017). In species abundance-based ecological analyses like calculations of niche breadth in Chapter 3, these biases directly affect the output and conclusions. To avoid doubtful conclusions, previous niche breadth studies discarded low recruiting OTUs (Logares et al., 2013; Liao et al., 2016), but we circumvented this issue by statistically classifying actual niche breadth values after 1000 randomized permutations.

Insights into the ecogenomics of uncultured taxa

In Chapter 2 we follow the recommendations of Konstantinidis et al. (2017) for the description of the uncultured novel species *Candidatus* Kordia photophila, which include a thorough demonstration of taxonomic novelty by 16S rRNA gene phylogeny, phylogenomics, and genome composition analyses such as ANI and AAI; defining their coding potential by an accurate description of their functional annotation; and niche preference or distribution by the recruitment analysis of multiple metagenomes its station of origin. Combining its distribution and genetic content, we propose this as a novel taxon able to grow on phytoplankton derived particulate organic matter by an extensive array of carbohydrate-active enzymes, encoding proteorhodopsin for extra energy acquisition in the sunlit ocean, but also able to continue colonizing and consuming particles as they sink into the characteristic OMZ layer of the North Indian Ocean, as it encodes machinery for micro-aerobic metabolism and virulence factors for out-competing other members of the community. It also encodes sensors and genetic potential for the assimilation of nitrate, a nutrient that was in highest abundance in the mesopelagic layer, that was the same depth as the OMZ and in which the taxon seems to switch from a particle-attached lifestyle to a free-living stage.

The Circumpolar Biodiversity Monitoring Program (CBMP) from the Conservation of Arctic Flora and Fauna (CAFF) association reported in 2017 that monitoring activities of plankton are poor in the Atlantic Arctic, inexistent in the Kara-Laptev seas and sporadic in the rest of arctic regions (Conservation of Arctic Flora and Fauna, 2017). Complementing our current knowledge on Arctic prokaryotic diversity, that has been mostly retrieved by expeditions of limited geographical range and during specific seasons, in Chapter 3 we provide a pan-Arctic view of prokaryotic diversity, distribution and expression patterns. Arctic MAGs are represented by common groups previously found in Arctic seawaters but also a large amount of taxonomic novelty, reaching novelty at the class level for Bacteria MAGs and increasing novel taxa in decreasing taxonomic ranks. Moreover, we have been able to define Arctic habitat specialists and generalists, that could be used as sentinel species for climate change and "atlantification" of the Arctic ocean in future monitoring events and found that mesopelagic waters are the preferred niche for specialist MAGs, that show higher metabolic diversity than generalists.

In Chapter 3, we have also assessed the potential for chemoautotrophy via the Calvin Cycle in the waters of the Arctic Ocean from spring until autumn. Prokaryotic heterotrophy has been found to dominate the Arctic summer waters, due to large organic matter availability from late spring phytoplankton blooms (Boetius et al., 2015) and increased nutrient and organic matter provisions by coastal run-off and riverine influence (Wheeler et al., 1996, 1997; Anderson, 2002). Chemolithoautotrophic processes, on the other hand, have been found to increase in importance during winter (Boetius et al., 2015), with ammonia-oxidizing Thaumarchaeaota constituting up to 16% of cells in the winter surface waters of Western Arctic (Alonso-Sáez et al., 2008). By selecting those MAGs encoding RuBisCo and looking at their expression patterns in metatranscriptomes we have found that prokaryotic inorganic carbon fixation via the Calvin Cycle is expressed in all the Arctic regions covered by the TOPC expedition from spring to autumn. The dominant groups encoding RuBisCo are Proteobacteria and Actinobacteria, while Cyanobacteria are only present and expressing their genes in the North Atlantic photic samples. RuBisCo encoding phyla Latescibacterota, UBA8248 and Crenarchaeota (previous Thaumarchaeota) are only showing genetic expression in the mesopelagic samples. Photo-heterotrophs are also widely represented in metatranscriptomes in the temporal, horizontal and vertical axes, further analysis of the metatranscriptomic reads recruited specifically in the PR and pufM genes should be done to confirm previous trends of increased expression during summer (Cottrell and Kirchman, 2009; Nguyen et al., 2015). We have not found any complete metabolic pathway related to other autotrophic strategies, but there could be a widespread use of the 3-hydroxypropionate bicycle by bacteria in the different regions, layers and seasons, especially by Proteobacteria, Chloroflexota, Actinobacteria and Myxococcota. The 3-hydroxypropionate/4-hydroxybutyrate cycle, described previously in ammonia oxidizing Thaumarchaeota during winter (Alonso-Sáez et al., 2012) could also be actively occurring in the mesopelagic samples of late spring, early summer and autumn. The oxidation of CO has also been found to be encoded within 70% of Arctic Bacteria MAGs and 5% of Archaea, expressed in all the sampled regions in the photic ocean and the spring, early summer and autumn mesopelagic samples. CO oxidation has been found widespread in temperate marine genomes, enhancing long-term survival in oligotrophic conditions of aerobic heterotrophic bacteria (Cordero et al., 2019). The Arctic Ocean presently acts as a sink for atmospheric CO_2, playing a disproportionate role in the global oceanic uptake relative to is surface area (Bates and Mathis, 2009). In spite of its global importance and the likelihood of significant change due to global warming, the Arctic Carbon cycle remains poorly quantified (MacGilchrist et al., 2014). Eventhough more accurate expression analyses at the gene level should be done to confirm their expression, the findings of Chapter 3 related to the widespread mechanisms for autotrophic inorganic carbon fixation, light-enhanced heterotrophy and oxidation of atmospheric CO provides valuable new information that could significantly alter the current views of the carbon cycle in the Arctic.

Despite the importance of the genome-based functional description of uncultured taxa like those of Chapter 2 and 3, including the estimations of active genomes by meta-transcriptomic read mapping done in Chapter 3, experimental validation should follow these studies to identify which metabolic processes are actually occurring in the environment and their magnitude. Visualization is often key in understanding the ecology of microbes (Sebastián and Gasol, 2019) and many methodologies can be applied to environmental samples in order to detect active members of the community. Some approaches like MAR-FISH (microautoradiography combined with fluorescent in situ hybridization) (Sintes and Herndl, 2006) or nanoSIMS (nanoscale secondary ion mass spectrometry) (Gao et al., 2016) can be insightful in metabolic processes visualized at the single cell level. Nevertheless, their application is destructive, meaning that no further analyses can be performed with the active cells detected. Other alternatives like SIP (stable isotope probing) (Dumont and Murrell, 2005; Gao et al., 2016) or BONCAT (biorthogonal non-canonical amino-acid tagging) (Hatzenpichler et al., 2014; Smriga et al., 2014; Leizeaga et al., 2017) can detect active cells and be combined with Raman microscopy (Lorenz et al., 2017), FISH or FACS (fluorescence-activated cell sorting) for their targeted sorting and further molecular exploration. Microscopic observations were key in the validation of network-generated hypothesis of plankton interaction in the global ocean (Lima-Mendez et al., 2015), BONCAT-FISH and BONCAT-FACS were essential in the identification of key methane anaerobic oxidizers in deep methane seep sediments (Hatzenpichler et al., 2016), FISH combined to SCG was key in the characterization of the endosymbiotic relationship between the nitrogen-fixing UCYN-A and its host (Zehr and Kudela, 2011) and nanoSIMS was decisive in determining the magnitude of UCYN-A's nitrogen fixation abilities (Martínez-Pérez et al., 2016). In winter Arctic waters, MAR-FISH was essential in determining that Thaumarchaeota were active inorganic CO_2 fixers using ammonia oxidation (Alonso-Sáez et al., 2012) as electron donor. CARD-FISH of radiolabeled bicarbonate incubated cultures and MAR-FISH showed heterotrophic bacteria actively incorporating CO_2 in the dark (Alonso-Sáez et al., 2010).

The descriptive work based on meta-omics and single-cell genomics of marine samples in this thesis provides a significant advance in our knowledge of the ecology (biogeography, habitat preferences) and functional capabilities of uncultured key bacteria and archaea, many of them representing novel families, classes, genera and species. At the same time, it serves as a starting point in the design of experimental procedures that can identify the active players in the environment at the sampled locations and seasons exposed in this thesis, and measure the magnitude of their metabolic activity.

CONCLUSIONS

Conclusions

1. Targeted single-cell genomics (SCG) was successful in the retrieval of heterotrophic prokaryotes genomes that were first approached with sequencing independent analyses, determining a lack of microdiversity and allowing their selection for whole-genome sequencing and co-assembly.

2. Normalization by coverage and read merging substantially reduces computation time and resources in co-assembly of multiple genomes, as well as finding the k-mer combinations producing the best metrics and highest assembly quality scores. Longer k-mers during assembly have provided a less fragmented genome, which was later improved by re-assembly and manual curation.

3. Complementary sampling of size fractions bigger and smaller of the typical bacterial size fraction was essential in determining the preferred niche of *Kordia* SAGs, as well as a chemical and oceanographical characterization of the station of origin of the SAGs and the surrounding region.

4. The novel photo-heterotrophic *Candidatus* Kordia photophila has the ability to bind to specific phytoplankton-derived particulate organic matter and degrade them. Micro-aerobic cytochrome oxidases enable its metabolic activity the characteristic oxygen-minimum zone (OMZ) of the water column it inhabits.

5. We have enriched our knowledge on genome-wise prokaryotic arctic seawater diversity, by the generation of 3550 bins of which 530 are of medium and high-quality metagenomic assembled genomes (MAGs) that recruit almost half of the metagenomic dataset, the highest so far in big-scale MAG studies.

6. Setting a minimum horizontal genome coverage to validate read recruitments has been essential in discarding false mappings that would affect further ecological conclusions.

7. Taxonomic novelty is found in the majority of MAGs at the species level, but some Archaea MAGs cannot be classified further than family, and Bacteria MAGs show novelty at the class level. At the phyla level, MAGs are dominated by previously described abundant taxa in the Arctic Ocean: Proteobacteria, Bacteroidetes, Chloflexota, Halobacterota, Verrucomicrobiota, Actinobacteriota, Marinisomatota and Planctomycetota.

8. A third of MAGs are distributed only in polar regions. Of these 11 genomes are habitat generalists, which are evenly distributed along the majority of samples. These could serve as sentinels of climate change in the Arctic. Four MAGs are unique of the North Atlantic latitudes closer to the polar circle, their presence in higher latitudes in future studies would be indicative of the climate-change induced atlantification of the Arctic Ocean.

9. Polar specialists show greater metabolic diversity than the mostly heterotrophic polar generalists and include a majority of arctic-specific MAGs. Polar generalists on the other hand include more bipolar taxa.

10. Habitat generalist MAGs, show greater abundances in the Arctic photic waters while mesopelagic waters show greater amounts of specialists. It could be due to wider environmental gradients in surface waters, and specific range of nutrient types and availability in the deeper layers.

11. Carbon metabolism from spring to autumn in the arctic MAGs is diverse. Inorganic carbon fixation via Calvin cycle may be widespread spatially and from spring to autumn, photoheterotrophy is ubiquitously expressed in the photic Arctic and CO oxidation is a widespread energy providing mechanism in the Arctic MAGs, likely for survival in oligotrophic waters. Further expression analyses at the gene level should be made to corroborate the expression of these mechanisms, as well as experimental validation.

BIBLIOGRAPHY

Bibliography

Aagaard, K., J. H. Swift, and E. C. Carmack
1985. Thermohaline circulation in the Arctic Mediterranean Seas. *Journal of Geophysical Research: Oceans*, 90(C3):4833–4846.

Abby, S. S., B. Néron, H. Ménager, M. Touchon, and E. P. Rocha
2014. MacSyFinder: A program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS ONE*, 9(10).

Abby, S. S. and E. P. Rocha
2017. Identification of protein secretion systems in bacterial genomes using MacSyFinder. *Methods in Molecular Biology*, 1615(February):1–21.

Abell, G. C. and J. P. Bowman
2005. Ecological and biogeographic relationships of class Flavobacteria in the Southern Ocean. *FEMS Microbiology Ecology*, 51(2):265–277.

Achtman, M. and M. Wagner
2008. Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, 6(6):431–440.

Acinas, S. G., J. Antón, and F. Rodríguez-Valera
1999. Diversity of free-living and attached bacteria in offshore Western Mediterranean waters as depicted by analysis of genes encoding 16S rRNA. *Applied and environmental microbiology*, 65(2):514–22.

Acinas, S. G., I. Ferrera, H. Sarmento, C. Díez-Vives, I. Forn, C. Ruiz-González, F. M. Cornejo-Castillo, G. Salazar, and J. M. Gasol
2014. Validation of a new catalysed reporter deposition-fluorescence in situ hybridization probe for the accurate quantification of marine Bacteroidetes populations. *Environmental microbiology*.

Acinas, S. G., V. Klepac-Ceraj, D. E. Hunt, C. Pharino, I. Ceraj, D. L. Distel, and M. F. Polz
2004a. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, 430(6999):551–4.

Acinas, S. G., L. A. Marcelino, V. Klepac-ceraj, and M. F. Polz
2004b. Divergence and Redundancy of 16S rRNA Sequences in Genomes with Multiple rrn Operons. *Journal of bacteriology*, 186(9):2629–2635.

Acinas, S. G., P. Sánchez, G. Salazar, F. M. Cornejo-Castillo, M. Sebastián, R. Logares, S. Sunagawa, P. Hingamp, H. Ogata, G. Lima-Mendez, S. Roux, J. M. González, J. M. Arrieta, I. S. Alam, A. Kamau, C. Bowler, J. Raes, S. Pesant, P. Bork, S. Agustí, T. Gojobori, V. Bajic, D. Vaqué, M. B. Sullivan, C. Pedrós-Alió, R. Massana, C. M. Duarte, and J. M. Gasol
2019. Metabolic Architecture of the Deep Ocean Microbiome. *bioRxiv*, P. 635680.

Alberti, A., C. Belser, S. Engelen, L. Bertrand, C. Orvain, L. Brinas, C. Cruaud, L. Giraut, C. Da Silva, C. Firmo, J.-M. Aury, and P. Wincker
2014. Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics*, 15(1):912.

Alberti, A., J. Poulain, S. Engelen, K. Labadie, S. Romac, I. Ferrera, G. Albini, J.-M. Aury, C. Belser, A. Bertrand, C. Cruaud, C. Da Silva, C. Dossat, F. Gavory, S. Gas, J. Guy, M. Haquelle, E. Jacoby, O. Jaillon, A. Lemainque, E. Pelletier, G. Samson, M. Wessner, P. Bazire, O. Beluche, L. Bertrand, M. Besnard-Gonnet, I. Bordelais, M. Boutard, M. Dubois, C. Dumont, E. Ettedgui, P. Fernandez, E. Garcia, N. G. Aiach, T. Guerin, C. Hamon, E. Brun, S. Lebled, P. Lenoble, C. Louesse, E. Mahieu, B. Mairey, N. Martins, C. Megret, C. Milani, J. Muanga, C. Orvain, E. Payen, P. Perroud, E. Petit, D. Robert, M. Ronsin, B. Vacherie, S. G. Acinas, M. Royo-Llonch, F. M. Cornejo-Castillo, R. Logares, B. Fernández-Gómez, C. Bowler, G. Cochrane, C. Amid, P. T. Hoopen, C. De Vargas, N. Grimsley, E. Desgranges, S. Kandels-Lewis, H. Ogata, N. Poulton, M. E. Sieracki, R. Stepanauskas, M. B. Sullivan, J. R. Brum, M. B. Duhaime, B. T. Poulos, B. L. Hurwitz, S. G. Acinas, P. Bork, E. Boss, C. Bowler, C. De Vargas, M. Follows, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, J. Raes, C. Sardet, M. E. Sieracki, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, S. Pesant, E. Karsenti, P. Wincker, G. T. Team, S. G. Acinas, M. Royo-Llonch, F. M. Cornejo-Castillo, R. Logares, B. Fernández-Gómez, C. Bowler, G. Cochrane, C. Amid, P. T. Hoopen, C. De Vargas, N. Grimsley, E. Desgranges, S. Kandels-Lewis, H. Ogata, N. Poulton, M. E. Sieracki, R. Stepanauskas, M. B. Sullivan, J. R. Brum, M. B. Duhaime, B. T. Poulos, B. L. Hurwitz, T. O. C. Coordinators, S. Pesant, E. Karsenti, and P. Wincker
2017. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific Data*, 4:170093.

Albertsen, M., P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen
2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, 31(6):533–538.

Allison, S. D. and J. B. H. Martiny
2008. Resistance, resilience, and redundancy in microbial communities. *Proceedings of the National Academy of Sciences*, 105(Supplement 1):11512 LP – 11519.

Alneberg, J., B. S. Bjarnason, I. De Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince
2014. Binning metagenomic contigs by coverage and composition. *Nat. Methods*, 11(11):1144–1146.

Alneberg, J., C. M. G. Karlsson, A.-M. Divne, C. Bergin, F. Homa, M. V. Lindh, L. W. Hugerth, T. J. G. Ettema, S. Bertilsson, A. F. Andersson, and J. Pinhassi
2018. Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome 2018 6:1*, 6(1):173.

Alonso-Sáez, L., V. Balagué, E. L. Sà, O. Sánchez, J. M. González, J. Pinhassi, R. Massana, J. Pernthaler, C. Pedrós-Alió, and J. M. Gasol
2007. Seasonality in bacterial diversity in north-west Mediterranean coastal waters: assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiology Ecology*, 60(1):98–112.

Alonso-Sáez, L., P. E. Galand, E. O. Casamayor, C. Pedrós-Alió, and S. Bertilsson
2010. High bicarbonate assimilation in the dark by Arctic bacteria. *ISME Journal*, 4(12):1581–1590.

Alonso-Sáez, L., O. Sánchez, J. M. Gasol, V. Balagué, and C. Pedrós-Alio
2008. Winter-to-summer changes in the composition and single-cell activity of near-surface Arctic prokaryotes. *Environmental Microbiology*, 10(9):2444–2454.

Alonso-Sáez, L., A. S. Waller, D. R. Mende, K. Bakker, H. Farnelid, P. L. Yager, C. Lovejoy, J.-É. J.-E. Tremblay, M. Potvin, F. Heinrich, M. Estrada, L. Riemann, P. Bork, C. Pedrós-Alió, and S. Bertilsson
2012. Role for urea in nitrification by polar marine Archaea. *Proceedings of the National Academy of Sciences*, 109(44):17989–17994.

Alonso-Sáez, L., M. Zeder, T. Harding, J. Pernthaler, C. Lovejoy, S. Bertilsson, and C. Pedrós-Alió
2014. Winter bloom of a rare betaproteobacterium in the Arctic Ocean. *Frontiers in Microbiology*, 5(September 2015):425.

Amann, R. I., W. Ludwig, and K. H. Schleifer
1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1):143–69.

AMAP/CAFF/SDWG
2013. Identification of Arctic Marine Areas of Heightened Ecological and Cultural Significance: Arctic Marine Shipping Assessment (AMSA) IIc. P. 114 pp.

Amaral-Zettler, L., L. F. Artigas, J. Baross, L. Bharathi P.A., A. Boetius, D. Chandramohan, G. Herndl, K. Kogure, P. Neal, C. Pedrós-Alió, A. Ramette, S. Schouten, L. Stal, A. Thessen, J. de Leeuw, and M. Sogin
2010. A Global Census of Marine Microbes.

Amend, A. S., T. A. Oliver, L. A. Amaral-Zettler, A. Boetius, J. A. Fuhrman, M. C. Horner-Devine, S. M. Huse, D. B. M. Welch, A. C. Martiny, A. Ramette, L. Zinger, M. L. Sogin, and J. B. H. Martiny
2013. Macroecological patterns of marine bacteria on a global scale. *Journal of Biogeography*, 40(4):800–811.

Anderson, L. G.
2002. DOC in the Arctic Ocean. In *Biogeochemistry of Marine Dissolved Organic Matter*, D. A. Hansell and C. A. B. T. B. o. M. D. O. M. Carlson, eds., Pp. 665–683. San Diego: Elsevier.

Anderson, T. R. and M. I. Lucas
2008. Upwelling Ecosystems. In *Encyclopedia of Ecology*, S. E. Jørgensen and B. D. Fath, eds., Pp. 3651–3661. Oxford: Academic Press.

Arístegui, J., J. M. Gasol, C. M. Duarte, and G. J. Herndld
2009. Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography*, 54(5):1501–1529.

Arndt, D., J. R. Grant, A. Marcu, T. Sajed, A. Pon, Y. Liang, and D. S. Wishart
2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research*, 44(W1):W16–W21.

Azam, F.
1998. OCEANOGRAPHY: Microbial Control of Oceanic Carbon Flux: The Plot Thickens. *Science*, 280(5364):694–696.

Azam, F., T. Fenchel, J. Field, J. Gray, L. Meyer-Reil, and F. Thingstad
1983. The Ecological Role of Water-Column Microbes in the Sea. *Marine Ecology Progress Series*, 10:257–263.

Baas Becking, L. G. M.
1934. *Geobiologie of inleiding tot de milieukunde*. Den Haag: W.P. Van Stockum & Zoon.

Baek, K., A. Choi, I. Kang, K. Lee, and J. C. Cho
2013. Kordia antarctica sp. nov., isolated from Antarctic seawater. *International Journal of Systematic and Evolutionary Microbiology*, 63(PART10):3617–3622.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner
2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477.

Bano, N. and J. T. Hollibaugh
2002. Phylogenetic composition of bacterioplankton assemblages from the Arctic Ocean. *Applied and environmental microbiology*, 68(2):505–518.

Bano, N., S. Ruffin, B. Ransom, and J. T. Hollibaugh
2004. Phylogenetic Composition of Arctic Ocean Archaeal Assemblages and Comparison with Antarctic Assemblages. *Applied and Environmental Microbiology*, 70(2):781 LP – 789.

Bar-On, Y. M., R. Phillips, and R. Milo
2018. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences of the United States of America*, 115(25):6506–6511.

Barton, B. I., Y.-D. Lenn, and C. Lique
2018. Observed Atlantification of the Barents Sea Causes the Polar Front to Limit the Expansion of Winter Sea Ice. *Journal of Physical Oceanography*, 48(8):1849–1866.

Bates, N. R. and J. T. Mathis
2009. The arctic ocean marine carbon cycle: evaluation of air-sea co_2 exchanges, ocean acidification impacts and potential feedbacks. *Biogeosciences*, 6(11):2433–2459.

Begley, T. P., C. Kinsland, and E. Strauss
2001. The biosynthesis of coenzyme A in bacteria. *Vitamins and hormones*, 61:157–171.

Beja, O.
2000. Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. *Science*, 289(5486):1902–1906.

Bertelli, C., M. R. Laird, K. P. Williams, B. Y. Lau, G. Hoad, G. L. Winsor, and F. S. Brinkman
2017. IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research*, 45(W1):W30–W35.

Bienhold, C., A. Boetius, and A. Ramette
2012. The energy–diversity relationship of complex bacterial communities in Arctic deep-sea sediments. *The ISME Journal*, 6(4):724–732.

Bintanja, R. and O. Andry
2017. Towards a rain-dominated Arctic. *Nature Climate Change*, 7(4):263–267.

Blainey, P. C.
2013. The future is now: single-cell genomics of bacteria and archaea. *FEMS microbiology reviews*, 37(3):407–27.

Boetius, A., A. M. Anesio, J. W. Deming, J. A. Mikucki, and J. Z. Rapp
2015. Microbial ecology of the cryosphere: sea ice and glacial habitats.

Bolger, A. M., M. Lohse, and B. Usadel
2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Bopp, L., C. Bowler, L. Guidi, E. Karsenti, and C. de Vargas
2015. The Ocean: a Carbon Pump. *Ocean Climate*, Pp. 12–17.

Boucher, Y., O. X. Cordero, and A. Takemura
2011. Local mobile gene pools rapidly cross species boundaries to create endemicity within global Vibrio cholerae populations. *mBio*, 2(2):1–8.

Bowers, R. M., N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A.

Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, N. C. Kyrpides, L. Schriml, G. M. Garrity, P. Hugenholtz, G. Sutton, P. Yilmaz, F. Meyer, F. O. Glöckner, J. A. Gilbert, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, T. Woyke, G. S. Consortium, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, and T. Woyke
2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*, 35(8):725–731.

Bowman, J. P. and R. D. McCuaig
2003. Biodiversity, Community Structural Shifts, and Biogeography of Prokaryotes within Antarctic Continental Shelf Sediment. *Applied and Environmental Microbiology*, 69(5):2463 LP – 2483.

Bowman, J. S., S. Rasmussen, N. Blom, J. W. Deming, S. Rysgaard, and T. Sicheritz-Ponten
2012. Microbial community structure of Arctic multiyear sea ice and surface seawater by 454 sequencing of the 16S RNA gene. *The ISME Journal*, 6(1):11–20.

Brooks, J. P., D. J. Edwards, M. D. Harwich, M. C. Rivera, J. M. Fettweis, M. G. Serrano, R. A. Reris, N. U. Sheth, B. Huang, P. Girerd, J. F. Strauss, K. K. Jefferson, G. A. Buck, and V. M. C. a. Members)
2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology*, 15(1):66.

Brown, M. and J. Fuhrman
2005. Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. *Aquatic Microbial Ecology*, 41(1):15–23.

Brum, J. R., J. C. Ignacio-Espinoza, S. Roux, G. Doulcier, S. G. Acinas, A. Alberti, S. Chaffron, C. Cruaud, C. de Vargas, J. M. Gasol, G. Gorsky, A. C. Gregory, L. Guidi, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, B. T. Poulos, S. M. Schwenck, S. Speich, C. Dimier, S. Kandels-Lewis, M. Picheral, S. Searson, T. O. Coordinators, P. Bork, C. Bowler, S. Sunagawa, P. Wincker, E. Karsenti, and M. B. Sullivan
2015. Patterns and ecological drivers of ocean viral communities. *Science*, 348(6237).

Brümmer, I. H. M., A. D. M. Felske, and I. Wagner-Döbler
2004. Diversity and seasonal changes of uncultured Planctomycetales in river biofilms. *Applied and environmental microbiology*, 70(9):5094–5101.

Buchan, A., G. R. LeCleir, C. A. Gulvik, and J. M. González
2014. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat Rev Micro*, 12(10):686–698.

Buchfink, B., C. Xie, and D. H. Huson
2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60.

Buerger, S., A. Spoering, E. Gavrish, C. Leslin, L. Ling, and S. S. Epstein
2012. Microbial Scout Hypothesis, Stochastic Exit from Dormancy, and the Nature of Slow Growers. *Applied and Environmental Microbiology*, 78(9):3221–3228.

Bunse, C. and J. Pinhassi
2017. Marine Bacterioplankton Seasonal Succession Dynamics. *Trends in Microbiology*, 25(6):494–505.

Button, D. K., F. Schut, P. Quang, R. Martin, and B. R. Robertson
1993. Viability and isolation of marine bacteria by dilution culture: theory, procedures, and initial results. *Applied and environmental microbiology*, 59(3):881–91.

Caporaso, J. G., K. Paszkiewicz, D. Field, R. Knight, and J. A. Gilbert
2012. The Western English Channel contains a persistent microbial seed bank. *The ISME Journal*, 6(6):1089–1093.

Cardman, Z., C. Arnosti, A. Durbin, K. Ziervogel, C. Cox, A. D. Steen, and A. Teske
2014. Verrucomicrobia Are Candidates for Polysaccharide-Degrading Bacterioplankton in an Arctic Fjord of Svalbard. *Applied and Environmental Microbiology*, 80(12):3749–3756.

Carmack, E. C.
2007. The alpha/beta ocean distinction: A perspective on freshwater fluxes, convection, nutrients and productivity in high-latitude seas.

Caro-Quintero, A. and K. T. Konstantinidis
2012. Bacterial species may exist, metagenomics reveal. *Environmental Microbiology*, 14(2):347–355.

Carradec, Q., E. Pelletier, C. Da Silva, A. Alberti, Y. Seeleuthner, R. Blanc-Mathieu, G. Lima-Mendez, F. Rocha, L. Tirichine, K. Labadie, A. Kirilovsky, A. Bertrand, S. Engelen, M.-A. Madoui, R. Méheust, J. Poulain, S. Romac, D. J. Richter, G. Yoshikawa, C. Dimier, S. Kandels-Lewis, M. Picheral, S. Searson, S. G. Acinas, E. Boss, M. Follows, G. Gorsky, N. Grimsley, L. Karp-Boss, U. Krzic, S. Pesant, E. G. Reynaud, C. Sardet, M. Sieracki, S. Speich, L. Stemmann, D. Velayoudon, J. Weissenbach, O. Jaillon, J.-M. Aury, E. Karsenti, M. B. Sullivan, S. Sunagawa, P. Bork, F. Not, P. Hingamp, J. Raes, L. Guidi, H. Ogata, C. de Vargas, D. Iudicone, C. Bowler, P. Wincker, and T. O. Coordinators
2018. A global ocean atlas of eukaryotic genes. *Nature Communications*, 9(1):373.

Castro, J. C., L. M. Rodriguez-R, W. T. Harvey, M. R. Weigand, J. K. Hatt, M. Q. Carter, and K. T. Konstantinidis
2018. imGLAD: accurate detection and quantification of target organisms in metagenomes. *PeerJ*, 6(11):e5882.

Chaumeil, P.-A., A. J. Mussig, P. Hugenholtz, and D. H. Parks
2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*.

Chikhi, R. and P. Medvedev
2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1):31–37.

Choi, A., H.-M. Oh, S.-J. Yang, and J.-C. Cho
2011. Kordia periserrulae sp. nov., isolated from a marine polychaete Periserrula leucophryna, and emended description of the genus Kordia. *International journal of systematic and evolutionary microbiology*, 61(Pt 4):864–9.

Clark, S. C., R. Egan, P. I. Frazier, and Z. Wang
2013. ALE: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, 29(4):435–443.

Clingenpeel, S., A. Clum, P. Schwientek, C. Rinke, and T. Woyke
2014. Reconstructing each cell's genome within complex microbial communities - dream or reality? *Frontiers in Microbiology*, 5(DEC):1–6.

Codispoti, L. A., J. A. Brandes, J. P. Christensen, A. H. Devol, S. W. Naqvi, H. W. Paerl, and T. Yoshinari
2001. The oceanic fixed nitrogen and nitrous oxide budgets: Moving targets as we enter the anthropocene? *Scientia Marina*, 65(SUPPLEMENT 2):85–105.

Cohan, F. M.
2001. Bacterial Species and Speciation. *Systematic Biology*, 50(4):513–524.

Cohan, F. M.
2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 361(1475):1985–96.

Colatriano, D., P. Q. Tran, C. Guéguen, W. J. Williams, C. Lovejoy, and D. A. Walsh
2018. Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean Chloroflexi bacteria. *Communications Biology*, 1(1):90.

Collins, R. E., G. Rocap, and J. W. Deming
2010. Persistence of bacterial and archaeal communities in sea ice through an Arctic winter. *Environmental microbiology*, 12(7):1828–41.

Colwell, R. K. and D. J. Futuyma
1971. On the Measurement of Niche Breadth and Overlap. *Ecology*, 52(4):567–576.

Comte, J. and P. A. del Giorgio
2011. Composition Influences the Pathway but not the Outcome of the Metabolic Response of Bacterioplankton to Resource Shifts. *PLOS ONE*, 6(9):e25266.

Connon, S. A. and S. J. Givannoni
2002. High-throughput methods for culturing microorganisms in very-low-nutrient media. *Appl Environ Microbiol*, 68(8):3878–3885.

Conway, J. R., A. Lex, and N. Gehlenborg
2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940.

Cordero, O. X. and M. F. Polz
2014. Explaining microbial genomic diversity in light of evolutionary ecology. *Nature reviews. Microbiology*, 12(4):263–73.

Cordero, P. R. F., K. Bayly, P. Man Leung, C. Huang, Z. F. Islam, R. B. Schittenhelm, G. M. King, and C. Greening
2019. Atmospheric carbon monoxide oxidation is a widespread mechanism supporting microbial survival. *The ISME Journal*, Pp. 2868–2881.

Cornejo-Castillo, F. M., M. d. C. Muñoz-Marín, K. A. Turk-Kubo, M. Royo-Llonch, H. Farnelid, S. G. Acinas, and J. P. Zehr
2019. UCYN-A3, a newly characterized open ocean sublineage of the symbiotic N 2 -fixing cyanobacterium Candidatus Atelocyanobacterium thalassa. *Environmental Microbiology*, 21(1):111–124.

Costea, P. I., G. Zeller, S. Sunagawa, E. Pelletier, A. Alberti, F. Levenez, M. Tramontano, M. Driessen, R. Hercog, F.-E. Jung, J. R. Kultima, M. R. Hayward, L. P. Coelho, E. Allen-Vercoe, L. Bertrand, M. Blaut, J. R. M. Brown, T. Carton, S. Cools-Portier, M. Daigneault, M. Derrien, A. Druesne, W. M. de Vos, B. B. Finlay, H. J. Flint, F. Guarner, M. Hattori, H. Heilig, R. A. Luna, J. van Hylckama Vlieg, J. Junick, I. Klymiuk, P. Langella, E. Le Chatelier, V. Mai, C. Manichanh, J. C. Martin, C. Mery, H. Morita, P. W. O'Toole, C. Orvain, K. R. Patil, J. Penders, S. Persson, N. Pons, M. Popova, A. Salonen, D. Saulnier, K. P. Scott, B. Singh, K. Slezak, P. Veiga, J. Versalovic, L. Zhao, E. G. Zoetendal, S. D. Ehrlich, J. Dore, and P. Bork
2017. Towards standards for human fecal sample processing in metagenomic studies. *Nature Biotechnology*, 35(11):1069–1076.

Cottrell, M. T. and D. L. Kirchman
2000. Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low- and high-molecular-weight dissolved organic matter. *Applied and environmental microbiology*, 66(4):1692–7.

Cottrell, M. T. and D. L. Kirchman
2009. Photoheterotrophic microbes in the arctic ocean in summer and winter. *Applied and Environmental Microbiology*, 75(15):4958–4966.

Crespo, B. G., T. Pommier, B. Fernández-Gómez, and C. Pedrós-Alió
2013. Taxonomic composition of the particle-attached and free-living bacterial assemblages in

the Northwest Mediterranean Sea analyzed by pyrosequencing of the 16S rRNA. *Microbiology-Open*, 2(4):541–552.

Crespo, B. G., P. J. Wallhead, R. Logares, and C. Pedrós-Alió
2016. Probing the Rare Biosphere of the North-West Mediterranean Sea: An Experiment with High Sequencing Effort. *PLOS ONE*, 11(7):e0159195.

Das, N., N. Tripathi, S. Basu, C. Bose, S. Maitra, and S. Khurana
2015. Progress in the development of gelling agents for improved culturability of microorganisms. *Frontiers in Microbiology*, 6(JUN):1–7.

De Bruyn, J. C., F. C. Boogerd, P. Bos, and J. G. Kluenen
1990. Floating filters, a novel technique for isolation and enumeration of fastidious, acidophilic, iron-oxidizing, autotrophic bacteria. *Applied and Environmental Microbiology*, 56(9):2891–2894.

De Queiroz, K.
2005. Ernst Mayr and the modern concept of species. *Systematics and the Origin of Species: On Ernst Mayr's 100th Anniversary*, Pp. 243–263.

de Vargas, C., S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horák, O. Jaillon, G. Lima-Mendez, J. Lukeš, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, Tara Oceans Coordinators, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, J. Raes, M. E. Sieracki, S. Speich, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, and E. Karsenti
2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science (New York, N.Y.)*, 348(6237):1261605.

del Giorgio, P. A. and C. M. Duarte
2002. Respiration in the open ocean. *Nature*, 420(6914):379–384.

Delmont, T. O. and A. M. Eren
2018. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ*, 6:e4320.

Delmont, T. O., C. Quince, A. Shaiber, O. C. Esen, S. T. Lee, S. Lücker, and A. Murat Eren
2018. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in the surface ocean. *Nature Microbiology*, 3(July).

DeLong, E. F.
1992. Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences of the United States of America*, 89(12):5685–5689.

DeLong, E. F., D. G. Franks, and A. L. Alldredge
1993. Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnology and Oceanography*, 38(5):924–934.

DeLong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N.-U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl
2006. Community Genomics Among Stratified Microbial Assemblages in the Ocean&#039;s Interior. *Science*, 311(5760):496 LP – 503.

Denef, V. J., R. S. Mueller, E. Chiang, J. R. Liebig, and H. A. Vanderploeg
2016. Chloroflexi CL500-11 Populations That Predominate Deep-Lake. *Applied and environmental microbiology*, 82(5):1423–1432.

Dethlefsen, L., M. McFall-Ngai, and D. A. Relman
2007. An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature*, 449(7164):811–818.

Diaz, R. J., H. Eriksson-Hägg, and R. Rosenberg
2013. Chapter 4 - Hypoxia. In *Managing Ocean Environments in a Changing Climate*, K. J. Noone, U. R. Sumaila, and R. J. Diaz, eds., Pp. 67–96. Boston: Elsevier.

Dick, G. J. and B. J. Baker
2013. Omic Approaches in Microbial Ecology: Charting the Unknown. *Microbe Magazine*, 8(9):353–360.

Díez-Vives, C., J. M. Gasol, and S. G. Acinas
2014. Spatial and temporal variability among marine Bacteroidetes populations in the NW Mediterranean Sea. *Systematic and Applied Microbiology*, 37(1):68–78.

Doney, S. C., M. Ruckelshaus, J. Emmett Duffy, J. P. Barry, F. Chan, C. A. English, H. M. Galindo, J. M. Grebmeier, A. B. Hollowed, N. Knowlton, J. Polovina, N. N. Rabalais, W. J. Sydeman, and L. D. Talley
2012. Climate Change Impacts on Marine Ecosystems. *Annual Review of Marine Science*, 4(1):11–37.

Doolittle, W. F. and R. T. Papke
2006. Genomics and the bacterial species problem. *Genome biology*, 7(9):116.

Ducklow, H., D. Steinberg, and K. Buesseler
2001. Upper Ocean Carbon Export and the Biological Pump. *Oceanography*, 14(4):50–58.

Dumont, M. G. and J. C. Murrell
2005. Stable isotope probing — linking microbial identity to function. *Nature Reviews Microbiology*, 3(6):499–504.

Dupont, C. L., J. P. McCrow, R. Valas, A. Moustafa, N. Walworth, U. Goodenough, R. Roth, S. L. Hogle, J. Bai, Z. I. Johnson, E. Mann, B. Palenik, K. A. Barbeau, J. Craig Venter, and A. E. Allen
2015. Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *The ISME Journal*, 9(5):1076–1092.

Edgar, R. C.
2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Eguíluz, V. M., G. Salazar, J. Fernández-Gracia, J. K. Pearman, J. M. Gasol, S. G. Acinas, S. Sunagawa, X. Irigoien, and C. M. Duarte
2019. Scaling of species distribution explains the vast potential marine prokaryote diversity. *Scientific Reports*, 9(1):18710.

Eilers, H., J. Pernthaler, J. Peplies, F. O. Glöckner, G. Gerdts, and R. Amann
2001. Isolation of Novel Pelagic Bacteria from the German Bight and Their Seasonal Contributions to Surface Picoplankton. *Applied and Environmental Microbiology*, 67(3-12):5134–5142.

Elbourne, L. D. H., S. G. Tetu, K. A. Hassan, and I. T. Paulsen
2017. TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Research*, 45(D1):D320–D324.

Eren, A. M., Ö. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, and T. O. Delmont
2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319.

Falkowski, P. G., T. Fenchel, and E. F. Delong
2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320(5879):1034–1039.

Fandino, L. B., L. Riemann, G. F. Steward, and F. Azam
2005. Population dynamics of Cytophaga-Flavobacteria during marine phytoplankton blooms analyzed by real-time quantitative PCR . *Aquatic Microbial Ecology*, 40(3):251–257.

Feil, E. J.
2004. Small change: Keeping pace with microevolution. *Nature Reviews Microbiology*, 2(6):483–495.

Fenchel, T.
2008. The microbial loop – 25 years later. *Journal of Experimental Marine Biology and Ecology*, 366(1-2):99–103.

Fernández-Gómez, B., A. Fernàndez-Guerra, E. O. Casamayor, J. M. González, C. Pedrós-Alió, and S. G. Acinas
2012. Patterns and architecture of genomic islands in marine bacteria. *BMC Genomics*, 13(1):1–19.

Fernández-Gómez, B., M. Richter, M. Schüler, J. Pinhassi, S. G. Acinas, J. M. González, and C. Pedrós-Alió
2013. Ecology of marine Bacteroidetes: a comparative genomics approach. *The ISME journal*, 7(5):1026–37.

Ferrari, B. C., S. J. Binnerup, and M. Gillings
2005. Microcolony cultivation on a soil substrate membrane system selects for previously uncultured soil bacteria. *Applied and Environmental Microbiology*, 71(12):8714–8720.

Ferrera, I., J. M. Gasol, M. Sebastian, E. Hojerova, and M. Koblizek
2011. Comparison of Growth Rates of Aerobic Anoxygenic Phototrophic Bacteria and Other Bacterioplankton Groups in Coastal Mediterranean Waters. *Applied and Environmental Microbiology*, 77(21):7451–7458.

Flombaum, P., J. L. Gallegos, R. A. Gordillo, J. Rincon, L. L. Zabala, N. Jiao, D. M. Karl, W. K. W. Li, M. W. Lomas, D. Veneziano, C. S. Vera, J. A. Vrugt, and A. C. Martiny
2013. Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proceedings of the National Academy of Sciences*, 110(24):9824–9829.

Forest, A., M. Sampei, R. Makabe, H. Sasaki, D. G. Barber, Y. Gratton, P. Wassmann, and L. Fortier
2008. The annual cycle of particulate organic carbon export in Franklin Bay (Canadian Arctic): Environmental control and food web implications. *Journal of Geophysical Research*, 113(C3):1–14.

Franzosa, E. A., X. C. Morgan, N. Segata, L. Waldron, J. Reyes, A. M. Earl, G. Giannoukos, M. R. Boylan, D. Ciulla, D. Gevers, J. Izard, W. S. Garrett, A. T. Chan, and C. Huttenhower
2014. Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences*, 111(22):E2329 LP – E2338.

Fraser, C., E. J. Alm, M. F. Polz, B. G. Spratt, and W. P. Hanage
2009. The Bacterial Species Challenge : Making Sense of Genetic and Ecological Diversity. *Science*, 323(February):741–746.

Fraser, C., W. P. Hanage, and B. G. Spratt
2007. Recombination and the nature of bacterial speciation. *Science*, 315.

Frey, K. E. and J. W. McClelland
2009. Impacts of permafrost degradation on arctic river biogeochemistry. *Hydrological Processes*, 23(1):169–182.

Frias-Lopez, J., Y. Shi, G. W. Tyson, M. L. Coleman, S. C. Schuster, S. W. Chisholm, and E. F. DeLong
2008. Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences*, 105(10):3805 LP – 3810.

Fuhrman, J. A. and F. Azam
1980. Bacterioplankton secondary production estimates for coastal waters of british columbia, antarctica, and california. *Applied and environmental microbiology*, 39(6):1085–95.

Fuhrman, J. a. and L. Campbell
1998. Microbial microdiversity. *Nature*, 393(6684):410–411.

Fuhrman, J. A., J. A. Cram, and D. M. Needham
2015. Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology*, 13(3):133–146.

Fuhrman, J. A., I. Hewson, M. S. Schwalbach, J. A. Steele, M. V. Brown, and S. Naeem
2006. Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences of the United States of America*, 103(35):13104–13109.

Fuhrman, J. A., K. McCallum, and A. A. Davis
1992. Novel major archaebacterial group from marine plankton. *Nature*, 356(6365):148–149.

Fuhrman, J. A., M. S. Schwalbach, and U. Stingl
2008. Proteorhodopsins: an array of physiological roles? *Nature*, 6.

Galand, P. E., E. O. Casamayor, D. L. Kirchman, and C. Lovejoy
2009a. Ecology of the rare microbial biosphere of the Arctic Ocean. 106(52).

Galand, P. E., E. O. Casamayor, D. L. Kirchman, M. Potvin, and C. Lovejoy
2009b. Unique archaeal assemblages in the Arctic Ocean unveiled by massively parallel tag sequencing. *The ISME journal*, 3(7):860–869.

Galand, P. E., C. Lovejoy, J. Pouliot, M.-È. Garneau, and W. F. Vincent
2008. Microbial community diversity and heterotrophic production in a coastal Arctic ecosystem: A stamukhi lake and its source waters.

Galand, P. E., C. Lovejoy, and W. F. Vincent
2006. Remarkably diverse and contrasting archaeal communities in a large arctic river and the coastal Arctic Ocean. *Aquatic Microbial Ecology*, 44(2):115–126.

Galand, P. E., O. Pereira, C. Hochart, J. C. Auguet, and D. Debroas
2018. A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *The ISME Journal*, 12(10):2470–2478.

Gao, D., X. Huang, and Y. Tao
2016. A critical review of NanoSIMS in analysis of microbial metabolic activities at single-cell level. *Critical Reviews in Biotechnology*, 36(5):884–890.

Gevers, D., F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F. L. Thompson, and J. Swings
2005. Opinion: Re-evaluating prokaryotic species. *Nature reviews. Microbiology*, 3(9):733–9.

Ghiglione, J.-F., P. E. Galand, T. Pommier, C. Pedrós-Alió, E. W. Maas, K. Bakker, S. Bertilson, D. L. Kirchmanj, C. Lovejoy, P. L. Yager, A. E. Murray, D. L. Kirchman, C. Lovejoy, P. L. Yager, and A. E.

Murray
2012. Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 109(43):17633–8.

Giovannoni, S.
2004. Oceans of bacteria. *Nature*, 430(6999):515–516.

Giovannoni, S. and U. Stingl
2007. The importance of culturing bacterioplankton in the 'omics' age. *Nature Reviews Microbiology*, 5(10):820–826.

Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field
1990. Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, 345(6270):60–63.

Glenn, T. C.
2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, 11(5):759–769.

Glöckner, F. O., B. M. Fuchs, and R. Amann
1999. Bacterioplankton compositions of lakes and oceans: a first comparison based on fluorescence in situ hybridization. *Applied and environmental microbiology*, 65(8):3721–6.

Goh, K. M., S. Shahar, K.-G. Chan, C. S. Chong, S. I. Amran, M. H. Sani, I. I. Zakaria, and U. M. Kahar
2019. Current Status and Potential Applications of Underexplored Prokaryotes. *Microorganisms*, 7(10):468.

Gómez-Consarnau, L., J. M. González, M. Coll-Lladó, P. Gourdon, T. Pascher, R. Neutze, C. Pedrós-Alió, and J. Pinhassi
2007. Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature*, 445(7124):210–213.

Gómez-Pereira, P. R., M. Schüler, B. M. Fuchs, C. Bennke, H. Teeling, J. Waldmann, M. Richter, V. Barbe, E. Bataille, F. O. Glöckner, and R. Amann
2012. Genomic content of uncultured Bacteroidetes from contrasting oceanic provinces in the North Atlantic Ocean. *Environmental microbiology*, 14(1):52–66.

Gómez-Rubio, V.
2017. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). *Journal of Statistical Software*, 77(Book Review 2):3–5.

González, J. M., B. Fernández-Gomez, A. Fernández-Guerra, L. Gomez-Consarnau, O. Sánchez, M. Coll-Lladó, J. del Campo, L. Escudero, R. Rodríguez-Martínez, L. Alonso-Sáez, M. Latasa, I. Paulsen, O. Nedashkovskaya, I. Lekunberri, J. Pinhassi, and C. Pedrós-Alió
2008. Genome analysis of the proteorhodopsin-containing marine bacterium Polaribacter sp. MED152 (Flavobacteria). *Proceedings of the National Academy of Sciences*, 105(25):8724–8729.

González-Rocha, G., G. Muñoz-Cartes, C. B. Canales-Aguirre, C. A. Lima, M. Domínguez-Yévenes, H. Bello-Toledo, and C. E. Hernández
2017. Diversity structure of culturable bacteria isolated from the Fildes Peninsula (King George Island, Antarctica): A phylogenetic analysis perspective. *PLOS ONE*, 12(6):e0179390.

Goodwin, S., J. D. McPherson, and W. R. McCombie
2016. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351.

Grebmeier, J. M.
2012. Shifting patterns of life in the Pacific Arctic and sub-Arctic seas. *Ann. Rev. Mar. Sci.*, 4:63–78.

Gregory, A., A. Zayed, N. Conceiçao-Neto, B. Temperton, B. Bolduc, A. Alberti, M. Ardyna, K. Arkhipova, M. Carmicheal, C. Cruaud, C. Dimier, G. Dominguez-Huerta, J. Ferland, S. Kandels-Lewis, Y. Liu, C. Marec, S. Stéphane, M. Picheral, S. Pisarev, J. Poulain, J.-É. Tremblay, D. Vik, T. O. Coordinators, M. Babin, C. Bowler, A. Culley, C. de Vargas, B. Dutilh, D. Iudicone, L. Karp-Boss, S. Roux, S. Sunagawa, P. Wincker, and M. Sullivan
2019. Marine DNA Viral Macro-and Micro-Diversity From Pole to Pole. *Cell*, Pp. 1–15.

Habib, C., A. Houel, A. Lunazzi, J. F. Bernardet, A. B. Olsen, H. Nilsen, A. E. Toranzo, N. Castro, P. Nicolas, and E. Duchaud
2014. Multilocus sequence analysis of the marine bacterial genus Tenacibaculum suggests parallel evolution of fish pathogenicity and endemic colonization of aquaculture systems. *Applied and Environmental Microbiology*, 80(17):5503–5514.

Hagström, Å., T. Pommier, F. Rohwer, K. Simu, W. Stolte, D. Svensson, and U. L. Zweifel
2002. Use of 16S Ribosomal DNA for Delineation of Marine Bacterioplankton Species. *Applied and Environmental Microbiology*, 68(7):3628–3633.

Hameed, A., M. Shahina, S. Y. Lin, J. C. Cho, W. A. Lai, and C. C. Young
2013. Kordia aquimaris sp. nov., a zeaxanthin-producing member of the family Flavobacteriaceae isolated from surface seawater, and emended description of the genus Kordia. *International Journal of Systematic and Evolutionary Microbiology*, 63(PART 12):4790–4796.

Hatzenpichler, R., S. A. Connon, D. Goudeau, R. R. Malmstrom, T. Woyke, and V. J. Orphan
2016. Visualizing in situ translational activity for identifying and sorting slow-growing archaeal–bacterial consortia. *Proceedings of the National Academy of Sciences*, 113(28):E4069 LP – E4078.

Hatzenpichler, R., S. Scheller, P. L. Tavormina, B. M. Babin, D. A. Tirrell, and V. J. Orphan
2014. In situ visualization of newly synthesized proteins in environmental microbes using amino acid tagging and click chemistry. *Environmental Microbiology*, 16(8):2568–2590.

Hawley, A. K., M. K. Nobu, J. J. Wright, W. E. Durno, C. Morgan-Lang, B. Sage, P. Schwientek, B. K. Swan, C. Rinke, M. Torres-Beltrán, K. Mewis, W. T. Liu, R. Stepanauskas, T. Woyke, and S. J.

Hallam
2017. Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients. *Nature Communications*, 8(1):1–9.

Herndl, G. J. and T. Reinthaler
2013. Microbial control of the dark end of the biological pump. *Nature Geoscience*, 6(9):718–724.

Houbo Wu
2011. Composition of bacterial communities in deep-sea sediments from the South China Sea, the Andaman Sea and the Indian Ocean. *African Journal of Microbiology Research*, 5(29):5273–5283.

Huang, X. and A. Madan
1999. CAP3: A DNA sequence assembly program. *Genome Res.*, 9(9):868–877.

Hug, L. A., B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield
2016. A new view of the tree of life. *Nature Microbiology*, 1(5):16048.

Hugerth, L. W. and A. F. Andersson
2017. Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. *Frontiers in Microbiology*, 8:1561.

Hughes, J. B., J. J. Hellmann, T. H. Ricketts, B. J. M. Bohannan, L. Sinclair, O. A. Osman, S. Bertilsson, A. Eiler, V. Sala, E. De Faveri, C. Li, K. M. K. Lim, K. R. Chng, N. Nagarajan, S. Nishida, Y. Ono, K. Sekimizu, I. Hanning, S. Diaz-Sanchez, L. B. Smith, S. Kasai, J. G. Scott, Z.-J. Chu, Y.-J. Wang, S.-H. Ying, X.-W. Wang, M.-G. Feng, G. Benelli, A. Lo Iacono, A. Canale, H. Mehlhorn, A. Aldersley, A. Champneys, M. Homer, D. Robert, L. V. Ferguson, D. E. Heinrichs, B. J. Sinclair, B. M. Ott, M. Cruciger, A. M. Dacks, R. V. M. Rio, S. S. Andreadis, and A. Michaelakis
2016. Counting the Uncountable : Statistical Approaches to Estimating Microbial Diversity MINIREVIEW Counting the Uncountable : Statistical Approaches to Estimating Microbial Diversity. *Applied and environmental microbiology*, 10(1):4399–4406.

Huse, S. M., L. Dethlefsen, J. A. Huber, D. M. Welch, D. A. Relman, and M. L. Sogin
2008. Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing. *PLOS Genetics*, 4(11):e1000255.

Hyatt, D., G.-l. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser
2010. Prodigal : prokaryotic gene recognition and translation initiation site identification.

Ibarbalz, F. M., N. Henry, F. Lombard, C. Bowler, L. Zinger, G. Busseni, and H. Byrne
2019. Global Trends in Marine Plankton Diversity across Kingdoms of Life AR OCEANS EXPEDITION Article Global Trends in Marine Plankton Diversity across Kingdoms of Life. Pp. 1084–1097.

Ivars-Mart\'inez, E., A. . B. Martin-Cuadrado, G. D'Auria, A. Mira, S. Ferriera, J. Johnson, R. Friedman, and F. Rodr\'iguez-Valera
2008. Comparative genomics of two ecotypes of the marine planktonic copiotroph Alteromonas macleodii suggests alternative lifestyles associated with different kinds of particulate organic matter. *ISMEJ*, 2.

Jain, C., L. M. Rodriguez-R, A. M. Phillipy, K. T. Konstantinidis, and S. Aluru
2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(5114):1–8.

Johnson, J. L.
1973. Use of nucleic acid homologies in the taxonomy of anaerobic bacteria. *International Journal of Systematic Bacteriology*, 23(4):308–315.

Jones, S. E. and J. T. Lennon
2010. Dormancy contributes to the maintenance of microbial diversity. *Proceedings of the National Academy of Sciences*, 107(13):5881–5886.

Jorgensen, S. L., B. Hannisdal, A. Lanzén, T. Baumberger, K. Flesland, R. Fonseca, L. Øvreås, I. H. Steen, I. H. Thorseth, R. B. Pedersen, and C. Schleper
2012. Correlating microbial community profiles with geochemical data in highly stratified sediments from the Arctic Mid-Ocean Ridge. *Proceedings of the National Academy of Sciences*, 109(42):E2846 LP – E2855.

Kaeberlein, T., K. Lewis, and S. S. Epstein
2002. Isolating "uncultivabte" microorganisms in pure culture in a simulated natural environment. *Science*, 296(5570):1127–1129.

Kanehisa, M., Y. Sato, M. Furumichi, K. Morishima, and M. Tanabe
2019. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*, 47(D1):D590–D595.

Kanehisa, M., Y. Sato, and K. Morishima
2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology*, 428(4):726–731.

Kang, D. D., J. Froula, R. Egan, and Z. Wang
2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165.

Karsenti, E., S. G. Acinas, P. Bork, C. Bowler, C. De Vargas, J. Raes, M. Sullivan, D. Arendt, F. Benzoni, J.-M. Claverie, M. Follows, G. Gorsky, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, U. Krzic, F. Not, H. Ogata, S. Pesant, E. G. Reynaud, C. Sardet, M. E. Sieracki, S. Speich, D. Velayoudon, J. Weissenbach, and P. Wincker
2011. A holistic approach to marine eco-systems biology. *PLoS biology*, 9(10):e1001177.

Kashtan, N., S. E. Roggensack, S. Rodrigue, J. W. Thompson, S. J. Biller, A. Coe, H. Ding, P. Marttinen, R. R. Malmstrom, R. Stocker, M. J. Follows, R. Stepanauskas, and S. W. Chisholm
2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science (New York, N.Y.)*, 344(6182):416–20.

Kettler, G. C., A. C. Martiny, K. Huang, J. Zucker, M. L. Coleman, S. Rodrigue, F. Chen, A. Lapidus, S. Ferriera, J. Johnson, C. Steglich, G. M. Church, P. Richardson, and S. W. Chisholm
2007. Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genetics*, 3(12):2515–2528.

Khandeparker, R., R. M. Meena, and D. Deobagkar
2014. Bacterial Diversity in Deep-Sea Sediments from Afanasy Nikitin Seamount, Equatorial Indian Ocean. *Geomicrobiology Journal*, 31(10):942–949.

Kim, D. I., J. H. Lee, M. S. Kim, and C. N. Seong
2017. Kordia zosterae sp. nov., isolated from the seaweed, Zostera marina. *International Journal of Systematic and Evolutionary Microbiology*, 67(11):4790–4795.

Kim, M., H. S. Oh, S. C. Park, and J. Chun
2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, 64(PART 2):346–351.

King, G. M. and C. F. Weber
2007. Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nature Reviews Microbiology*, 5(2):107–118.

Kirchman, D. L.
2002. The ecology of Cytophaga-Flavobacteria in aquatic environments. *FEMS Microbiology Ecology*, 39(2):91–100.

Kirchman, D. L.
2018. *Microbial Ecology of the Oceans: Third Edition*.

Kirchman, D. L., M. T. Cottrell, and C. Lovejoy
2010. The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environmental Microbiology*, 12(5):1132–1143.

Kirchman, D. L., H. Elifantz, A. I. Dittel, R. R. Malmstrom, and M. T. Cottrell
2007. Standing stocks and activity of Archaea and Bacteria in the western Arctic Ocean. *Limnology and Oceanography*, 52(2):495–507.

Kisker, C., J. Kuper, and B. Van Houten
2013. Prokaryotic Nucleotide Excision Repair. *Cold Spring Harbor Perspectives in Biology*, 5(3):a012591–a012591.

Klemetsen, T., I. A. Raknes, J. Fu, A. Agafonov, S. V. Balasundaram, G. Tartari, E. Robertsen, and N. P. Willassen
2018. The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res.*, 46(D1):D692–D699.

Kogawa, M., M. Hosokawa, Y. Nishikawa, K. Mori, and H. Takeyama
2018. Obtaining high-quality draft genomes from uncultured microbes by cleaning and co-assembly of single-cell amplified genomes. *Scientific Reports*, 8(1):1–11.

Kogure, K., U. Simidu, and N. Taga
1979. A tentative direct microscopic method for counting living marine bacteria. *Canadian Journal of Microbiology*, 25(3):415–420.

Könneke, M., A. E. Bernhard, J. R. de la Torre, C. B. Walker, J. B. Waterbury, and D. A. Stahl
2005. Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, 437(7058):543–546.

Kono, T., S. Mehrotra, C. Endo, N. Kizu, M. Matusda, H. Kimura, E. Mizohata, T. Inoue, T. Hasunuma, A. Yokota, H. Matsumura, and H. Ashida
2017. A RuBisCO-mediated carbon metabolic pathway in methanogenic archaea. *Nature Communications*, 8(1):14007.

Konstantinidis, K. T. and E. F. DeLong
2008. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *The ISME journal*, 2(10):1052–65.

Konstantinidis, K. T., A. Ramette, and J. M. Tiedje
2006. The bacterial species definition in the genomic era. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 361(1475):1929–40.

Konstantinidis, K. T., R. Rosselló-Móra, and R. Amann
2017. Uncultivated microbes in need of their own taxonomy. *The ISME Journal*, Pp. 1–8.

Konstantinidis, K. T. and J. M. Tiedje
2005a. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2567–72.

Konstantinidis, K. T. and J. M. Tiedje
2005b. Towards a genome-based taxonomy for prokaryotes. *Journal of bacteriology*, 187(18):6258–6264.

Kultima, J. R., S. Sunagawa, J. Li, W. Chen, H. Chen, D. R. Mende, M. Arumugam, Q. Pan, B. Liu, J. Qin, J. Wang, and P. Bork
2012. MOCAT : A Metagenomics Assembly and Gene Prediction Toolkit. 7(10):1–6.

Kumar, S., G. Stecher, and K. Tamura
2016. MEGA7 : Molecular Evolutionary Genetics Analysis Version 7 . 0 for Bigger Datasets Brief communication. 33(7):1870–1874.

Labonté, J. M., B. K. Swan, B. Poulos, H. Luo, S. Koren, S. J. Hallam, M. B. Sullivan, T. Woyke, K. Eric Wommack, and R. Stepanauskas
2015. Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME Journal*, 9(11):2386–2399.

Lai, Q., Z. Shao, Y. Liu, Y. Xie, J. Du, and C. Dong
2015. Kordia zhangzhouensis sp. nov., isolated from surface freshwater. *International Journal of Systematic and Evolutionary Microbiology*, 65(10):3379–3383.

Lam, P. and M. M. Kuypers
2011. Microbial Nitrogen Cycling Processes in Oxygen Minimum Zones. *Annual Review of Marine Science*, 3(1):317–345.

Landry, M. R., S. L. Brown, L. Campbell, J. Constantinou, and H. Liu
1998. Spatial patterns in phytoplankton growth and microzooplankton grazing in the Arabian Sea during monsoon forcing. *Deep-Sea Research Part II: Topical Studies in Oceanography*, 45(10-11):2353–2368.

Langmead, B. and S. L. Salzberg
2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359.

Lefort, T. and J. Gasol
2013. Global-scale distributions of marine surface bacterioplankton groups along gradients of salinity, temperature, and chlorophyll: a meta-analysis of fluorescence in situ hybridization studies. *Aquatic Microbial Ecology*, 70(2):111–130.

Leizeaga, A., M. Estrany, I. Forn, and M. Sebastián
2017. Using Click-Chemistry for Visualizing in Situ Changes of Translational Activity in Planktonic Marine Bacteria. *Frontiers in Microbiology*, 8:2360.

Lennon, J. T., Z. T. Aanderud, B. K. Lehmkuhl, and D. R. Schoolmaster Jr.
2012. Mapping the niche space of soil microorganisms using taxonomy and traits. *Ecology*, 93(8):1867–1879.

Letunic, I. and P. Bork
2019. Interactive Tree Of Life (iTOL) v4 : recent updates and. 47(April):256–259.

Levins, R.
1968. *Evolution in changing environments : some theoretical explorations*. Princeton, N.J.: Princeton University Press.

Li, D., C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam
2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676.

Li, F., T. C. A. Hitch, Y. Chen, C. J. Creevey, and L. L. Guan
2019. Comparative metagenomic and metatranscriptomic analyses reveal the breed effect on the rumen microbiome and its associations with feed efficiency in beef cattle. *Microbiome*, 7(1):6.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup
2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, W. and A. Godzik
2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

Liao, J., X. Cao, L. Zhao, J. Wang, Z. Gao, M. C. Wang, and Y. Huang
2016. The importance of neutral and niche processes for bacterial community assembly differs between habitat generalists and specialists. *FEMS Microbiology Ecology*, 92(11):1–10.

Lima-Mendez, G., K. Faust, N. Henry, J. Decelle, S. S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-Espinosa, S. Roux, F. Vincent, L. Bittner, Y. Darzi, J. Wang, S. S. Audic, L. L. Berline, G. Bontempi, A. M. Cabello, L. Coppola, F. M. Cornejo-Castillo, F. D'Ovidio, L. De Meester, I. Ferrera, M.-J. M.-J. Garet-Delmas, L. Guidi, E. Lara, S. S. Pesant, M. Royo-Llonch, G. Salazar, P. Sanchez, M. Sebastian, C. Souffreau, C. C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, G. Gorsky, F. Not, H. Ogata, S. Speich, L. Stemmann, J. Weissenbach, P. Wincker, S. G. Acinas, S. Sunagawa, P. Bork, M. B. Sullivan, E. Karsenti, C. Bowler, C. de Vargas, J. Raes, T. O. Coordinators, P. Sánchez, M. Sebastian, C. Souffreau, C. C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Coordinators, G. Gorsky, F. Not, H. Ogata, S. Speich, L. Stemmann, J. Weissenbach, P. Wincker, S. G. Acinas, S. Sunagawa, P. Bork, M. B. Sullivan, E. Karsenti, C. Bowler, C. de Vargas, and J. Raes
2015. Determinants of community structure in the global plankton interactome. *Science*, 348(6237):1–10.

Lindh, M. V., J. Sjöstedt, M. Casini, A. Andersson, C. Legrand, and J. Pinhassi
2016. Local Environmental Conditions Shape Generalist But Not Specialist Components of Microbial Metacommunities in the Baltic Sea. *Frontiers in Microbiology*, 07(DEC):1–10.

Lloyd, K. G., A. D. Steen, J. Ladau, J. Yin, and L. Crosby
2018. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems*, 3(5):1–12.

Locey, K. J. and J. T. Lennon
2016. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970 LP – 5975.

Logares, R., E. S. Lindström, S. Langenheder, J. B. Logue, H. Paterson, J. Laybourn-Parry, K. Rengefors, L. Tranvik, and S. Bertilsson
2013. Biogeography of bacterial communities exposed to progressive long-term environmental change. *The ISME Journal*, 7(5):937–948.

Logue, J. B. and E. S. Lindström
2008. Biogeography of Bacterioplankton in Inland Waters. *Freshwater Reviews*, 1(1):99–114.

Loman, N. J., C. Constantinidou, J. Z. Chan, M. Halachev, M. Sergeant, C. W. Penn, E. R. Robinson, and M. J. Pallen
2012. High-throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology*, 10(9):599–606.

López-Pérez, M. and F. Rodriguez-Valera
2016. Pangenome evolution in the marine bacterium <i>Alteromonas</i>. *Genome Biology and Evolution*, 8(5):evw098.

Lorenz, B., C. Wichmann, S. Stöckel, P. Rösch, and J. Popp
2017. Cultivation-Free Raman Spectroscopic Investigations of Bacteria. *Trends in Microbiology*, 25(5):413–424.

Louca, S., L. W. Parfrey, and M. Doebeli
2016. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272 LP – 1277.

Lovejoy, C., R. Massana, and C. Pedrós-Alió
2006. Diversity and distribution of marine microbial eukaryotes in the Arctic Ocean and adjacent seas. *Appl. Environ. Microbiol.*, 72(5):3085–3095.

Lovejoy, C., W. F. Vincent, S. Bonilla, S. Roy, M.-J. Martineau, R. Terrado, M. Potvin, R. Massana, and C. Pedrós-Alió
2007. Distribution, phylogeny, and growth of cold-adapted picoprasinophytes in Arctic seas. *Journal of Phycology*, 43(1):78–89.

Ludwig, W. and K. H. Schleifer
1994. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiology Reviews*, 15(2-3):155–173.

Luria, C. M., L. A. Amaral-Zettler, H. W. Ducklow, and J. J. Rich
2016. Seasonal Succession of Free-Living Bacterial Communities in Coastal Waters of the Western Antarctic Peninsula. *Frontiers in Microbiology*, 7:1731.

MacGilchrist, G. A., A. C. Naveira Garabato, T. Tsubouchi, S. Bacon, S. Torres-Valdés, and K. Azetsu-Scott
2014. The Arctic Ocean carbon sink. *Deep Sea Research Part I: Oceanographic Research Papers*, 86:39–55.

Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt
1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6):3140–3145.

Mangot, J.-f., R. Logares, P. Sánchez, F. Latorre, Y. Seeleuthner, S. Mondy, M. E. Sieracki, O. Jaillon, P. Wincker, C. de Vargas, and R. Massana
2017. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Scientific Reports*, 7(January):41498.

Martín, H. G., N. Ivanova, V. Kunin, F. Warnecke, K. W. Barry, A. C. McHardy, C. Yeates, S. He, A. A. Salamov, E. Szeto, E. Dalin, N. H. Putnam, H. J. Shapiro, J. L. Pangilinan, I. Rigoutsos, N. C. Kyrpides, L. L. Blackall, K. D. McMahon, and P. Hugenholtz
2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotechnology*, 24(10):1263–1269.

Martin, W., M. Weiss, S. Neukirchen, S. Nelson-Sathi, and F. Sousa
2016. Physiology, phylogeny, and LUCA. *Microbial Cell*, 3(12):582–587.

Martin-Cuadrado, A.-B., R. Ghai, A. Gonzaga, and F. Rodriguez-Valera
2009. CO dehydrogenase genes found in metagenomic fosmid clones from the deep mediterranean sea. *Applied and environmental microbiology*, 75(23):7436–7444.

Martínez-García, M., F. Santos, M. Moreno-Paz, V. Parro, and J. Antón
2014. Unveiling viral–host interactions within the 'microbial dark matter'. *Nature Communications*, 5:1–8.

Martinez-Garcia, M., B. K. Swan, N. J. Poulton, M. L. Gomez, D. Masland, M. E. Sieracki, and R. Stepanauskas
2012. High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *The ISME Journal*, 6(1):113–123.

Martínez-Pérez, C., W. Mohr, C. R. Löscher, J. Dekaezemacker, S. Littmann, P. Yilmaz, N. Lehnen, B. M. Fuchs, G. Lavik, R. A. Schmitz, J. LaRoche, and M. M. M. Kuypers
2016. The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nature Microbiology*, 1(11):16163.

Martiny, A. C.
2019. High proportions of bacteria are culturable across major biomes. *The ISME Journal*, 13(8):2125–2128.

Martiny, A. C., K. Treseder, and G. Pusch
2013. Phylogenetic conservatism of functional traits in microorganisms. *The ISME Journal*, 7(4):830–838.

Massana, R., E. F. DeLong, and C. Pedrós-Alió
2000. A Few Cosmopolitan Phylotypes Dominate Planktonic Archaeal Assemblages in Widely Different Oceanic Provinces. *Applied and Environmental Microbiology*, 66(5):1777 LP – 1787.

Mayr, E.
1942. *Systematics and the origin of species: An introduction*. Columbia University press, New York.

McCann, C. M., M. J. Wade, N. D. Gray, J. A. Roberts, C. R. J. Hubert, and D. W. Graham
2016. Microbial Communities in a High Arctic Polar Desert Landscape. *Frontiers in Microbiology*, 7(March):1–10.

McInerney, J., D. Pisani, and M. J. O'Connell
2015. The ring of life hypothesis for eukaryote origins is supported by multiple kinds of data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140323.

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

Meltofte, H.
2013. Arctic Biodiversity Assessment Status and Trens in Arctic Biodiversity. Technical report, Conservation of Arctic Flora and Fauna (CAFF).

Mestre, M., C. Ruiz-gonzález, R. Logares, C. M. Duarte, and J. M. Gasol
2018. Sinking particles promote vertical connectivity in the ocean microbiome. 115(29):6799–6807.

Middelburg, J. J.
2011. Chemoautotrophy in the ocean. *Geophysical Research Letters*, 38(24):94–97.

Millán-Aguiñaga, N., K. L. Chavarria, J. A. Ugalde, A.-C. Letzel, G. W. Rouse, and P. R. Jensen
2017. Phylogenomic Insight into Salinispora (Bacteria, Actinobacteria) Species Designations. *Scientific Reports*, 7(1):3564.

Mira, A., A. B. Martín-Cuadrado, G. D'Auria, and F. Rodríguez-Valera
2010. The bacterial pan-genome:a new paradigm in microbiology. *International microbiology : the official journal of the Spanish Society for Microbiology*, 13(2):45–57.

Moorthi, S., D. A. Caron, R. J. Gast, and R. W. Sanders
2009. Mixotrophy: a widespread and important ecological strategy for planktonic and sea-ice nanoflagellates in the Ross Sea, Antarctica.

Mullins, T. D., T. B. Britschgi, R. L. Krest, and S. J. Gioivannoni
1995. Genetic comparisons reveal the same unknown bacterial lineages in Atlantic and Pacific bacterioplankton communities. *Limnology and Oceanography*, 40(1):148–158.

Murray, R. G. and K. H. Schleifer
1994. Taxonomic notes: A proposal for recording the properties of putative taxa of procaryotes. *International Journal of Systematic Bacteriology*, 44(1):174–176.

Nesbø, C. L., M. Dlutek, and W. F. Doolittle
2006. Recombination in thermotoga: Implications for species concepts and biogeography. *Genetics*, 172(2):759–769.

Nguyen, D., R. Maranger, V. Balagué, M. Coll-Lladó, C. Lovejoy, and C. Pedrós-Alió
2015. Winter diversity and expression of proteorhodopsin genes in a polar ocean. *The ISME Journal*, Pp. 1–11.

Ochman, H., E. Lerat, and V. Daubin
2005. Examining bacterial species under the specter of gene transfer and exchange. *Proceedings of the National Academy of Sciences*, 102(Supplement 1):6595–6599.

Olli, K., C. W. Riser, P. Wassmann, T. Ratkova, E. Arashkevich, A. Pasternak, C. Wexels Riser, P. Wassmann, T. Ratkova, E. Arashkevich, and A. Pasternak
2002. Seasonal variation in vertical flux of biogenic matter in the marginal ice zone and the central Barents Sea. *Journal of Marine Systems*, 38(1-2):189–204.

Oren, A.
2017. A plea for linguistic accuracy – also for Candidatus taxa. *International Journal of Systematic and Evolutionary Microbiology*, 67(4):1085–1094.

Oren, A. and G. M. Garrity
2014. Then and now: a systematic review of the systematics of prokaryotes in the last 80 years. *Antonie van Leeuwenhoek*, 106(1):43–56.

Pace, N. R., D. A. Stahl, D. J. Lane, and G. J. Olsen
1986. The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. Pp. 1–55.

Pachiadaki, M. G., J. M. J. J. M. Brown, J. M. J. J. M. Brown, O. Bezuidt, P. M. Berube, S. J. Biller, N. J. Poulton, M. D. Burkart, J. J. La Clair, S. W. Chisholm, and R. Stepanauskas
2019. Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell*, 179(7):1623–1635.e11.

Pachiadaki, M. G., E. Sintes, K. Bergauer, J. M. Brown, N. R. Record, B. K. Swan, M. E. Mathyer, S. J. Hallam, P. Lopez-Garcia, Y. Takaki, T. Nunoura, T. Woyke, G. J. Herndl, and R. Stepanauskas
2017. Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation. *Science*, 358(6366):1046–1051.

Pandit, S. N., J. Kolasa, and K. Cottenie
2009. Contrasts between habitat generalists and specialists: An empirical extension to the basic metacommunity framework. *Ecology*, 90(8):2253–2262.

Papke, R. T., O. Zhaxybayeva, E. J. Feil, K. Sommerfeld, D. Muise, and W. F. Doolittle
2007. Searching for species in haloarchaea. *Proceedings of the National Academy of Sciences*, 104(35):14092 LP – 14097.

Park, S., Y.-T. Jung, and J.-H. Yoon
2014. Kordia jejudonensis sp. nov., isolated from the junction between the ocean and a freshwater spring, and emended description of the genus Kordia. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 64(Pt 2):657–662.

Parker, C. T., B. J. Tindall, and G. M. Garrity
2019. International code of nomenclature of Prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, 69(1):S1.

Parks, D. H., M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, and P. Hugenholtz
2018a. A proposal for a standardized bacterial taxonomy based on genome phylogeny. *bioRxiv*, P. 256800.

Parks, D. H., M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. A. Chaumeil, and P. Hugenholtz
2018b. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, 36(10):996–1004.

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

Parks, D. H., C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, and G. W. Tyson
2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2(11):1533–1542.

Paulmier, A. and D. Ruiz-Pino
2009. Oxygen minimum zones (OMZs) in the modern ocean. *Progress in Oceanography*, 80(3-4):113–128.

Pedrós-Alió, C.
2006. Marine microbial diversity: can it be determined? *Trends in microbiology*, 14(6):257–63.

Pedrós-Alió, C.
2012. The rare bacterial biosphere. *Annual review of marine science*, 4:449–66.

Pedrós-Alió, C., M. Potvin, and C. Lovejoy
2015. Diversity of planktonic microorganisms in the Arctic Ocean. *Progress in Oceanography*, 139:233–243.

Pernice, M. C., I. Forn, A. Gomes, E. Lara, L. Alonso-Sáez, J. M. Arrieta, F. del Carmen Garcia, V. Hernando-Morales, R. MacKenzie, M. Mestre, E. Sintes, E. Teira, J. Valencia, M. M. Varela, D. Vaqué, C. M. Duarte, J. M. Gasol, and R. Massana
2014. Global abundance of planktonic heterotrophic protists in the deep ocean. *The ISME Journal*.

Pesant, S., F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot, G. Gorsky, D. Iudicone, E. Karsenti, S. Speich, R. Troublé, C. Dimier, S. Searson, S. G. Acinas, P. Bork, E. Boss, C. Bowler, C. De Vargas, M. Follows, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, U. Krzic, F. Not, H. Ogata, S. Pesant, J. Raes, E. G. Reynaud, C. Sardet, M. Sieracki, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, D. Velayoudon, J. Weissenbach, P. Wincker, and T. O. C. Coordinators
2015. Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, 2(Lmd):150023.

Pinder, M. I., O. N. Johansson, A. Almstedt, O. Kourtchenko, A. K. Clarke, A. Godhe, and M. Töpel
2019. Genome Sequence of Kordia sp. Strain SMS9 Identified in a Non-Axenic Culture of the Diatom Skeletonema marinoi . *Journal of Genomics*, 7:46–49.

Pinhassi, J., E. F. Delong, O. Béjà, J. M. González, and C. Pedrós-alió
2016. Marine Bacterial and Archaeal Ion-Pumping Rhodopsins : Genetic Diversity , Physiology , and Ecology. 80(4):929–954.

Pommier, T., B. Canbäck, L. Riemann, K. H. Boström, K. Simu, P. Lundberg, A. Tunlid, and Å. Hagström
2007. Global patterns of diversity and community structure in marine bacterioplankton. *Molecular Ecology*, 16(4):867–880.

Price, M. N., P. S. Dehal, and A. P. Arkin
2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. 5(3).

Pruesse, E., J. Peplies, F. O. Glöckner, A. Editor, and J. Wren
2012. SINA : Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. 28(14):1823–1829.

Qi, F., Z. Shao, D. Li, Z. Huang, and Q. Lai
2016. Kordia ulvae sp. nov., a bacterium isolated from the surface of green marine algae Ulva sp. *International Journal of Systematic and Evolutionary Microbiology*, 66(7):2623–2628.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner
2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):590–596.

Ragsdale, S. W.
2004. Life with carbon monoxide. *Critical Reviews in Biochemistry and Molecular Biology*, 39(3):165–195.

Rappé, M. S. and S. J. Giovannoni
2003. The Uncultured Microbial Majority. *Annual Review of Microbiology*, 57(1):369–394.

Rawlings, N. D., M. Waller, A. J. Barrett, and A. Bateman
2014. MEROPS : the database of proteolytic enzymes , their substrates and inhibitors. 42(July 2011):503–509.

Reinthaler, T., H. M. van Aken, and G. J. Herndl
2010. Major contribution of autotrophy to microbial carbon cycling in the deep North Atlantic's interior. *Deep-Sea Research Part II: Topical Studies in Oceanography*, 57(16):1572–1580.

Ren, Q. and J. T. Pauisers
2005. Comparative analyses of fundamental differences in membrane transport Capabilities in Prokaryotes and Eukaryotes. *PLoS Computational Biology*, 1(3):0190–0201.

Richter, M. and R. Rosselló-Móra
2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45):19126–31.

Ricker, N., H. Qian, and R. R. Fulthorpe
2012. The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. *Genomics*, 100(3):167–175.

Riedel, A., C. Michel, and M. Gosselin
2007. Grazing of large-sized bacteria by sea-ice heterotrophic protists on the Mackenzie Shelf during the winterspring transition.

Riedel, A., C. Michel, M. Gosselin, and B. LeBlanc
2008. Winter–spring dynamics in sea-ice carbon cycling in the coastal Arctic Ocean.

Rinke, C., F. Rubino, L. F. Messer, N. Youssef, D. H. Parks, M. Chuvochina, M. Brown, T. Jeffries, G. W. Tyson, J. R. Seymour, and P. Hugenholtz
2019. A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.). *ISME Journal*, 13(3):663–675.

Rinke, C., P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J.-F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W.-T. Liu, J. a. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, and T. Woyke
2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–7.

Rodriguez-R, L. M., S. Gunturu, W. T. Harvey, R. Rossell, J. M. Tiedje, J. R. Cole, and K. T. Konstantinidis
2018. The Microbial Genomes Atlas (MiGA) webserver : taxonomic and gene diversity analysis of Archaea and. (June):1–7.

Romanuk, T. N. and J. Kolasa
2005. Resource limitation, biodiversity, and competitive effects interact to determine the invasibility of rock pool microcosms. *Biological Invasions*, 7(4):711–722.

Rosselló-Mora, R. and R. Amann
2001. The species concept for prokaryotes. *FEMS Microbiology Reviews*, 25(1):39–67.

Rosselló-Móra, R. and R. Amann
2015. Past and future species definitions for Bacteria and Archaea. *Systematic and Applied Microbiology*, 38(4):209–216.

Roullier, F., L. Berline, L. Guidi, X. Durrieu De Madron, M. Picheral, A. Sciandra, S. Pesant, and L. Stemmann
2014. Particle size distribution and estimated carbon flux across the Arabian Sea oxygen minimum zone. *Biogeosciences*, 11(16):4541–4557.

Roux, S., J. Emerson, E. Eloe-Fadrosh, and M. Sullivan
2017. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*, 5:e3817.

Roux, S., A. K. Hawley, M. Torres Beltran, M. Scofield, P. Schwientek, R. Stepanauskas, T. Woyke, S. J. Hallam, and M. B. Sullivan
2014. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife*, 3:1–20.

Royo-Llonch, M., I. Ferrera, F. Cornejo-Castillo, P. Sánchez, G. Salazar, R. Stepanauskas, J. González, M. Sieracki, S. Speich, L. Stemmann, C. Pedrós-Alió, and S. Acinas
2017. Exploring microdiversity in novel Kordia sp. (Bacteroidetes) with proteorhodopsin from the tropical Indian Ocean via Single Amplified Genomes. *Frontiers in Microbiology*, 8(JUL).

Ruiz-González, C., T. Lefort, R. Massana, R. Simó, and J. M. Gasol
2012. Diel changes in bulk and single-cell bacterial heterotrophic activity in winter surface waters of the northwestern Mediterranean Sea. *Limnology and Oceanography*, 57(1):29–42.

Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y.-H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter
2007. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology*, 5(3):e77.

Salazar, G., F. M. Cornejo-Castillo, V. Benítez-Barrios, E. Fraile-Nuez, X. A. Álvarez-Salgado, C. M.

Duarte, J. M. Gasol, and S. G. Acinas
2016. Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME Journal*, 10(3):596–608.

Salazar, G., F. M. Cornejo-Castillo, E. Borrull, C. Díez-Vives, E. Lara, D. Vaqué, J. M. Arrieta, C. M. Duarte, J. M. Gasol, and S. G. Acinas
2015. Particle-association lifestyle is a phylogenetically conserved trait in bathypelagic prokaryotes. *Molecular ecology*, 24(22):5692–706.

Salazar, G., L. Paoli, A. Alberti, J. Huerta-Cepas, H.-J. Ruscheweyh, M. Cuenca, C. M. Field, L. P. Coelho, C. Cruaud, S. Engelen, A. C. Gregory, K. Labadie, C. Marec, E. Pelletier, M. Royo-llonch, S. Roux, P. Sanchez, H. Uehara, A. A. Zayed, G. Zeller, M. Carmichael, C. Dimier, J. Ferland, S. Kandels-Lewis, M. Picheral, S. Pisarev, J. Poulain, the Tara Oceans Consortium Coordinators, S. G. Acinas, M. Babin, P. Bork, C. Bowler, C. de Vargas, L. Guidi, P. Hingamp, D. Iudicone, L. Karp-Boss, E. Karsenti, H. Ogata, S. S. Pesant, S. Speich, M. B. Sullivan, P. Wincker, S. Sunagawa, P. Sánchez, H. Uehara, A. A. Zayed, G. Zeller, M. Carmichael, C. Dimier, J. Ferland, S. Kandels, M. Picheral, S. Pisarev, J. Poulain, T. O. Coordinators, S. G. Acinas, M. Babin, P. Bork, C. Bowler, C. de Vargas, L. Guidi, P. Hingamp, D. Iudicone, L. Karp-Boss, E. Karsenti, H. Ogata, S. S. Pesant, S. Speich, M. B. Sullivan, P. Wincker, and S. Sunagawa
2019. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell*, 179(5):1068–1083.

Salazar, G. and S. Sunagawa
2017. Marine microbial diversity. *Current Biology*, 27(11):R489–R494.

Sanz-Sáez, I., G. Salazar, E. Lara, M. Royo-Llonch, D. Vaqué, C. M. Duarte, J. M. Gasol, C. Pedrós-Alió, O. Sánchez, and S. G. Acinas
2019. Diversity patterns of marine heterotrophic culturable bacteria along vertical and latitudinal gradients. *bioRxiv*, P. 774992.

Schattenhofer, M., B. M. Fuchs, R. Amann, M. V. Zubkov, G. A. Tarran, and J. Pernthaler
2009. Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environmental Microbiology*, 11(8):2078–2093.

Schlesinger, W. H.
1991. *Biogeochemistry: an Analysis of Global Change*. Elsevier.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber
2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75(23):7537–7541.

Schloter, M.
2000. Ecology and evolution of bacterial microdiversity. *FEMS Microbiology Reviews*, 24(5):647–660.

Schmidt, T. M., E. F. DeLong, and N. R. Pace
1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology*, 173(14):4371–4378.

Schwalbach, M. S. and J. A. Fuhrman
2005. Wide-ranging abundances of aerobic anoxygenic phototrophic bacteria in the world ocean revealed by epifluorescence microscopy and quantitative PCR. *Limnology and Oceanography*, 50(2):620–628.

Sczyrba, A., P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Dröge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. S. Jørgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, F. Meyer, M. Balvočiūtė, L. H. Hansen, S. J. Sørensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. Don Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.-W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. D. Barton, T. Lingner, H.-H. Lin, Y.-C. Liao, G. G. Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H.-P. Klenk, M. Göker, N. C. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, and A. C. McHardy
2017. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods*, 14(11):1063–1071.

Sebastián, M. and J. M. Gasol
2019. Visualization is crucial for understanding microbial processes in the ocean. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1786):1–7.

Seemann, T.
2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069.

Segerman, B.
2012. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Frontiers in Cellular and Infection Microbiology*, 2(September):1–8.

Shapiro, B. J., J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, G. Szabo, M. F. Polz, E. J. Alm, S. B. J., F. J., C. O. X., P. S. P., T. S. C., G. Szabó, M. F. Polz, E. J. Alm, and J. S. Universit
2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science*, 336(6077):48–51.

Shapiro, B. J. and M. F. Polz
2014. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in microbiology*, 22(5):235–47.

Sharon, I. and J. F. Banfield
2013. Genomes from Metagenomics. *Science*, 342(6162):1057–1058.

Sharon, I., M. Kertesz, L. A. Hug, D. Pushkarev, T. A. Blauwkamp, C. J. Castelle, M. Amirebrahimi, B. C. Thomas, D. Burstein, S. G. Tringe, K. H. Williams, and J. F. Banfield
2015. Accurate , multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Research*, 25(4):534–543.

Shi, Y., G. W. Tyson, J. M. Eppley, and E. F. DeLong
2011. Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *The ISME Journal*, 5(6):999–1013.

Shoseyov, O., Z. Shani, and I. Levy
2006. Carbohydrate Binding Modules: Biochemical Properties and Novel Applications. *Microbiology and Molecular Biology Reviews*, 70(2):283–295.

Sintes, E. and G. J. Herndl
2006. Quantifying Substrate Uptake by Individual Cells of Marine Bacterioplankton by Catalyzed Reporter Deposition Fluorescence In Situ Hybridization Combined with Microautoradiography. *Applied and Environmental Microbiology*, 72(11):7022 LP – 7028.

Sintes, E., H. Witte, K. Stodderegger, P. Steiner, and G. J. Herndl
2013. Temporal dynamics in the free-living bacterial community composition in the coastal North Sea. *FEMS microbiology ecology*, 83(2):413–24.

Sipler, R. E., C. T. E. Kellogg, T. L. Connelly, Q. N. Roberts, P. L. Yager, and D. A. Bronk
2017. Microbial Community Response to Terrestrially Derived Dissolved Organic Matter in the Coastal Arctic. *Frontiers in Microbiology*, 8:1018.

Smith, D. C., M. Simon, A. L. Alldredge, and F. Azam
1992. Intense hydrolytic enzyme activity on marine aggregates and implications for rapid particle dissolution. *Nature*, 359(6391):139–142.

Smriga, S., S. TJ, F. Malfatti, J. Villareal, and F. Azam
2014. Individual cell DNA synthesis within natural marine bacterial assemblages as detected by click chemistr. *Aquatic Microbial Ecology*, 72(3):269–280.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl
2006. Microbial diversity in the deep sea and the underexplored ”rare biosphere”. *Proceedings of the National Academy of Sciences of the United States of America*, 103(32):12115–20.

Sohn, J. H., J. H. Lee, H. Yi, J. Chun, K. S. Bae, T. Y. Ahn, and S. J. Kim
2004. Kordia algicida gen. nov., sp. nov., an algicidal bacterium isolated from red tide. *International Journal of Systematic and Evolutionary Microbiology*, 54(3):675–680.

Spudich, J. L.
2006. The multitalented microbial sensory rhodopsins. *Trends in microbiology*, 14(11):480–7.

Stackebrandt, E., G. E. Fox, and I. Introduction
1985. 16 S Ribosomal. 18.

Staley, J. T.
2006. The bacterial species dilemma and the genomic-phylogenetic species concept. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1475):1899–1909.

Staley, J. T. and J. J. Gosink
1999. Poles Apart: Biodiversity and Biogeography of Sea Ice Bacteria. *Annual Review of Microbiology*, 53(1):189–215.

Staley, J. T. and A. Konopka
1985. Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annual Review of Microbiology*, 39(1):321–346.

Steen, A. D., A. Crits-Christoph, P. Carini, K. M. DeAngelis, N. Fierer, K. G. Lloyd, and J. Cameron Thrash
2019. High proportions of bacteria and archaea across most biomes remain uncultured. *ISME Journal*, 13(12):3126–3130.

Stepanauskas, R.
2012. Single cell genomics: An individual look at microbes. *Current Opinion in Microbiology*, 15(5):613–620.

Stepanauskas, R., E. A. Fergusson, B. Joseph, N. J. Poulton, B. Tupper, J. M. Labonté, E. D. Becraft, J. M. Brown, M. G. Pachiadaki, T. Povilaitis, B. P. Thompson, C. J. Mascena, W. Bellows, and K. A. Lubys
2017. Improved genome recovery and integrated cell-size analyses of individual, uncultured microbial cells and viral particles. *Nature Communications*, Pp. 1–10.

Stepanauskas, R. and M. E. Sieracki
2007. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):9052–7.

Storesund, J. E. and L. Øvreås
2013. Diversity of Planctomycetes in iron-hydroxide deposits from the Arctic Mid Ocean Ridge (AMOR) and description of Bythopirellula goksoyri gen. nov., sp. nov., a novel Planctomycete from deep sea iron-hydroxide deposits. *Antonie van Leeuwenhoek*, 104(4):569–584.

Strous, M., J. A. Fuerst, E. H. M. Kramer, S. Logemann, G. Muyzer, K. T. van de Pas-Schoonen, R. Webb, J. G. Kuenen, and M. S. M. Jetten
1999. Missing lithotroph identified as new planctomycete. *Nature*, 400(6743):446–449.

Sul, W. J., T. A. Oliver, H. W. Ducklow, L. A. Amaral-Zettler, and M. L. Sogin
2013. Marine bacteria exhibit a bipolar distribution. *Proceedings of the National Academy of Sciences*, 110(6):2342–2347.

Sunagawa, S., L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. D'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Coordinators, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, P. Bork, E. Boss, C. Bowler, M. Follows, L. Karp-Boss, U. Krzic, E. G. Reynaud, C. Sardet, M. Sieracki, and D. Velayoudon
2015a. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359–1261359.

Sunagawa, S., L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. D'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Coordinators, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, and P. Bork
2015b. Ocean plankton. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359.

Sunagawa, S., D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. a. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Dore, S. D. Ehrlich, A. Stamatakis, and P. Bork
2013. Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods*, 10(12):1196–1199.

Sutcliffe, I. C.
2015. Challenging the anthropocentric emphasis on phenotypic testing in prokaryotic species descriptions: Rip it up and start again. *Frontiers in Genetics*, 6(JUN):6–9.

Swan, B. K., M. Martinez-Garcia, C. M. Preston, A. Sczyrba, T. Woyke, D. Lamy, T. Reinthaler, N. J. Poulton, E. D. P. Masland, M. L. Gomez, M. E. Sieracki, E. F. DeLong, G. J. Herndl, and R. Stepanauskas
2011. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science (New York, N.Y.)*, 333(6047):1296–300.

Swan, B. K., B. Tupper, A. Sczyrba, F. M. Lauro, M. Martinez-Garcia, J. M. González, H. Luo, J. J. Wright, Z. C. Landry, N. W. Hanson, B. P. Thompson, N. J. Poulton, P. Schwientek, S. G. Acinas,

S. J. Giovannoni, M. A. Moran, S. J. Hallam, R. Cavicchioli, T. Woyke, and R. Stepanauskas
2013. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences of the United States of America*, 110(28):11463–8.

Székely, A. J., M. Berga, and S. Langenheder
2013. Mechanisms determining the fate of dispersed bacterial communities in new environments. *The ISME Journal*, 7(1):61–71.

Székely, A. J. and S. Langenheder
2014. The importance of species sorting differs between habitat generalists and specialists in bacterial communities. *FEMS Microbiology Ecology*, 87(1):102–112.

Tang, K., D. Lin, K. Liu, and N. Jiao
2015. Draft genome sequence of Parvularcula oceani JLT2013T, a rhodopsin-containing bacterium isolated from deep-sea water of the Southeastern Pacific. *Marine Genomics*, 24:211–213.

Terrado, R., K. Scarcella, M. Thaler, W. F. Vincent, and C. Lovejoy
2013. Small phytoplankton in Arctic seas: vulnerability to climate change. *Biodiversity*, 14(1):2–18.

Teske, A., A. Durbin, K. Ziervogel, C. Cox, and C. Arnosti
2011. Microbial Community Composition and Function in Permanently Cold Seawater and Sediments from an Arctic Fjord of Svalbard. *Applied and Environmental Microbiology*, 77(6):2008 LP – 2018.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser
2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–5.

Tettelin, H., D. Riley, C. Cattuto, and D. Medini
2008. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5):472–477.

Thaler, M. and C. Lovejoy
2015. Biogeography of Heterotrophic Flagellate Populations Indicates the Presence of Generalist and Specialist Taxa in the Arctic Ocean. *Applied and Environmental Microbiology*, 81(6):2137 LP – 2148.

Thomas, D. N.
2016. *Sea Ice*. John Wiley & Sons.

Thompson, F. L., D. Gevers, C. C. Thompson, P. Dawyndt, S. Naser, B. Hoste, C. B. Munn, and J. Swings
2005. Phylogeny and Molecular Identification of Vibrios on the Basis of Multilocus Sequence Analysis. *Applied and Environmental Microbiology*, 71(9):5107–5115.

Thompson, L. R., M. F. Haroon, A. A. Shibl, M. J. Cahill, D. K. Ngugi, G. J. Williams, J. T. Morton, R. Knight, K. D. Goodwin, and U. Stingl
2019. Red Sea SAR11 and Prochlorococcus Single-Cell Genomes Reflect Globally Distributed Pangenomes. *Applied and Environmental Microbiology*, 85(13):1–18.

Tomczak, M. and J. S. Godfrey
1994. Regional Oceanography: an Introduction. In *Regional Oceanography*, Pp. 415–422. Amsterdam: Pergamon.

Torsvik, V. and L. Øvreås
2002. Microbial diversity and function in soil: from genes to ecosystems. *Current opinion in microbiology*, 5(3):240–245.

Tully, B., E. Graham, and J. Heidelberg
2017a. The Reconstruction of 2,631 Draft Metagenome-Assembled Genomes from the Global Oceans. *BioRxiv*, (September):1–31.

Tully, B. J.
2019. Metabolic diversity within the globally abundant Marine Group II Euryarchaea offers insight into ecological patterns. *Nature Communications*, 10(1):271.

Tully, B. J., E. D. Graham, and J. F. Heidelberg
2018a. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, 5(1):170203.

Tully, B. J., R. Sachdeva, E. D. Graham, and J. F. Heidelberg
2017b. 290 metagenome-assembled genomes from the Mediterranean Sea: A resource for marine microbiology. *PeerJ*, 2017(7).

Tully, B. J., C. G. Wheat, B. T. Glazer, and J. A. Huber
2018b. A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer. *The ISME Journal*, 12(1):1–16.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield
2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43.

Ulloa, O., D. E. Canfield, E. F. DeLong, R. M. Letelier, and F. J. Stewart
2012. Microbial oceanography of anoxic oxygen minimum zones. *Proceedings of the National Academy of Sciences of the United States of America*, 109(40):15996–6003.

van der Gast, C. J., A. W. Walker, F. A. Stressmann, G. B. Rogers, P. Scott, T. W. Daniels, M. P. Carroll, J. Parkhill, and K. D. Bruce
2011. Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. *The ISME Journal*, 5(5):780–791.

Van der Gucht, K., K. Cottenie, K. Muylaert, N. Vloemans, S. Cousin, S. Declerck, E. Jeppesen, J.-M. Conde-Porcuna, K. Schwenk, G. Zwart, H. Degans, W. Vyverman, and L. De Meester
2007. The power of species sorting: Local factors drive bacterial community composition over a wide range of spatial scales. *Proceedings of the National Academy of Sciences*, 104(51):20404 LP – 20409.

van der Meer, J. R. and V. Sentchilo
2003. Genomic islands and the evolution of catabolic pathways in bacteria. *Curr Opin Biotech*, 14.

van Dongen, S. and C. Abreu-Goodger
2012. *Using MCL to Extract Clusters from Networks*, Pp. 281–295. New York, NY: Springer New York.

Vandamme, P., B. Pot, M. Gillis, P. de Vos, K. Kersters, and J. Swings
1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiological Reviews*, 60(2):407 LP – 438.

Varani, A. M., P. Siguier, E. Gourbeyre, V. Charneau, and M. Chandler
2011. ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology*, 12(3):R30.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. H. Rogers, H. O. Smith, A. Z. Enkavi, B. Weber, I. Zweyer, J. Wagner, C. E. Elger, and E. U. Weber
2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74.

Villareal, T. A., C. G. Brown, M. A. Brzezinski, J. W. Krause, and C. Wilson
2012. Summer diatom blooms in the north Pacific subtropical gyre: 2008-2009. *PLoS ONE*, 7(4):2008–2009.

Vonk, J. E., L. Sánchez-García, B. E. van Dongen, V. Alling, D. Kosmach, A. Charkin, I. P. Semiletov,

O. V. Dudarev, N. Shakhova, P. Roos, T. I. Eglinton, A. Andersson, and Ö. Gustafsson
2012. Activation of old carbon by erosion of coastal and subsea permafrost in Arctic Siberia. *Nature*, 489(7414):137–140.

Wang, J., J. Kan, L. Borecki, X. Zhang, D. Wang, and J. Sun
2016. A snapshot on spatial and vertical distribution of bacterial communities in the eastern Indian Ocean. *Acta Oceanologica Sinica*, 35(6):85–93.

Ward, D. M., R. Weller, and M. M. Bateson
1990. 16S rRNA sequences reveal numerous uncultured inhabitants in a well-studied natural community. *Nature (London)*, 345(6270):63–65.

Wassmann, P. and M. Reigstad
2011. Future Arctic Ocean seasonal ice zones and implications for pelagic-benthic coupling. *Oceanography*, 24(3):220–231.

Wayne, L. G., D. J. Brenner, R. R. Colwell, P. a. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E. Murray, E. Stackebrandt, M. P. Starr, and H. G. Truper
1987. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic Bacteriology*, 37(4):463–464.

Weber, T., J. A. Cram, S. W. Leung, T. DeVries, and C. Deutsch
2016. Deep ocean nutrients imply large latitudinal variation in particle transfer efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, 113(31):8606–8611.

Weiss, M. C., M. Preiner, J. C. Xavier, V. Zimorski, and W. F. Martin
2018. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLOS Genetics*, 14(8):e1007518.

Wheeler, P. A., M. Gosselin, E. Sherr, D. Thibaultc, D. L. Kirchman, R. Benner, and T. E. Whitledge
1996. Active cycling of organic carbon in the central Arctic Ocean. *Nature*, 380(6576):697–699.

Wheeler, P. A., J. M. Watkins, and R. L. Hansing
1997. Nutrients, organic carbon and organic nitrogen in the upper water column of the Arctic Ocean: implications for the sources of dissolved organic carbon. *Deep Sea Research Part II: Topical Studies in Oceanography*, 44(8):1571–1592.

Wheeler, T. J. and S. R. Eddy
2013. nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29(19):2487–2489.

Whitaker, R. J., D. W. Grogan, and J. W. Taylor
2005. Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. *Molecular Biology and Evolution*, 22(12):2354–2361.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe
1998. Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583.

Wilkins, D., S. Yau, T. J. Williams, M. A. Allen, M. V. Brown, M. Z. DeMaere, F. M. Lauro, and R. Cavicchioli
2013. Key microbial drivers in Antarctic aquatic environments. *FEMS Microbiology Reviews*, 37(3):303–335.

Williams, T. A., P. G. Foster, C. J. Cox, and T. M. Embley
2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, 504(7479):231–236.

Wintzingerode, F. V., U. B. Göbel, and E. Stackebrandt
1997. Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews*, 21(3):213–229.

Woese, C. R. and G. E. Fox
1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms (archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny). *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):5088–5090.

Woese, C. R., O. Kandler, and M. L. Wheelis
1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576 LP – 4579.

Woyke, T., G. Xie, A. Copeland, J. M. González, C. Han, H. Kiss, J. H. Saw, P. Senin, C. Yang, S. Chatterji, J.-F. Cheng, J. A. Eisen, M. E. Sieracki, and R. Stepanauskas
2009. Assembling the Marine Metagenome, One Cell at a Time. *PLoS ONE*, 4(4):e5299.

Wrighton, K. C., B. C. Thomas, I. Sharon, C. S. Miller, C. J. Castelle, N. C. VerBerkmoes, M. J. Wilkins, R. L. Hettich, M. S. Lipton, K. H. Williams, P. E. Long, and J. F. Banfield
2012. Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science*, 337(6102):1661 LP – 1665.

Wuchter, C., B. Abbas, M. J. L. Coolen, L. Herfort, J. van Bleijswijk, P. Timmers, M. Strous, E. Teira, G. J. Herndl, J. J. Middelburg, S. Schouten, and J. S. Sinninghe Damsté
2006. Archaeal nitrification in the ocean. *Proceedings of the National Academy of Sciences*, 103(33):12317 LP – 12322.

Wyrtki, K.
1962. The oxygen minima in relation to ocean circulation. *Deep Sea Research and Oceanographic Abstracts*, 9(1-2):11–23.

Yarza, P., M. Richter, J. Euzeby, R. Amann, K.-h. Schleifer, W. Ludwig, F. O. Glo, and R. Rossello
2020. The All-Species Living Tree project : A 16S rRNA-based phylogenetic tree of all sequenced type strains. 31(2008):241–250.

Yin, Y., X. Mao, J. Yang, X. Chen, F. Mao, and Y. Xu
2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, 40(W1):W445–W451.

Yoon, H. S., D. C. Price, R. Stepanauskas, V. D. Rajah, M. E. Sieracki, W. H. Wilson, E. C. Yang, S. Duffy, and D. Bhattacharya
2011. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science*, 332(6030):714–717.

Yoshizawa, S., A. Kawanabe, H. Ito, H. Kandori, and K. Kogure
2012. Diversity and functional analysis of proteorhodopsin in marine Flavobacteria. *Environmental microbiology*, 14(5):1240–8.

Yoshizawa, S., J. Song, K. Kogure, A. Choi, Y. Joung, J.-C. Cho, and M. Im
2015. Aurantivirga profunda gen. nov., sp. nov., isolated from deep-seawater, a novel member of the family Flavobacteriaceae. *International Journal of Systematic and Evolutionary Microbiology*, 65(12):4850–4856.

Zehr, J. P. and R. M. Kudela
2011. Nitrogen Cycle of the Open Ocean: From Genes to Ecosystems. *Annual Review of Marine Science*, 3(1):197–225.

Zengler, K., G. Toledo, M. Rappé, J. Elkins, E. J. Mathur, J. M. Short, and M. Keller
2002. Cultivating the uncultured. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24):15681–15686.

Zhang, G., T. Cao, J. Ying, Y. Yang, and L. Ma
2014a. Diversity and novelty of actinobacteria in Arctic marine sediments. *Antonie van Leeuwenhoek*, 105(4):743–754.

Zhang, J., K. Kobert, T. Flouri, and A. Stamatakis
2014b. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5):614–620.

Ziegler, M., M. Lange, and W. Dott
1990. Isolation and morphological and cytological characterization of filamentous bacteria from bulking sludge. *Water Research*, 24(12):1437–1451.

LIST OF FIGURES AND TABLES

List of Figures

List of Tables

SUPPLEMENTARY MATERIAL

Supplementary figures

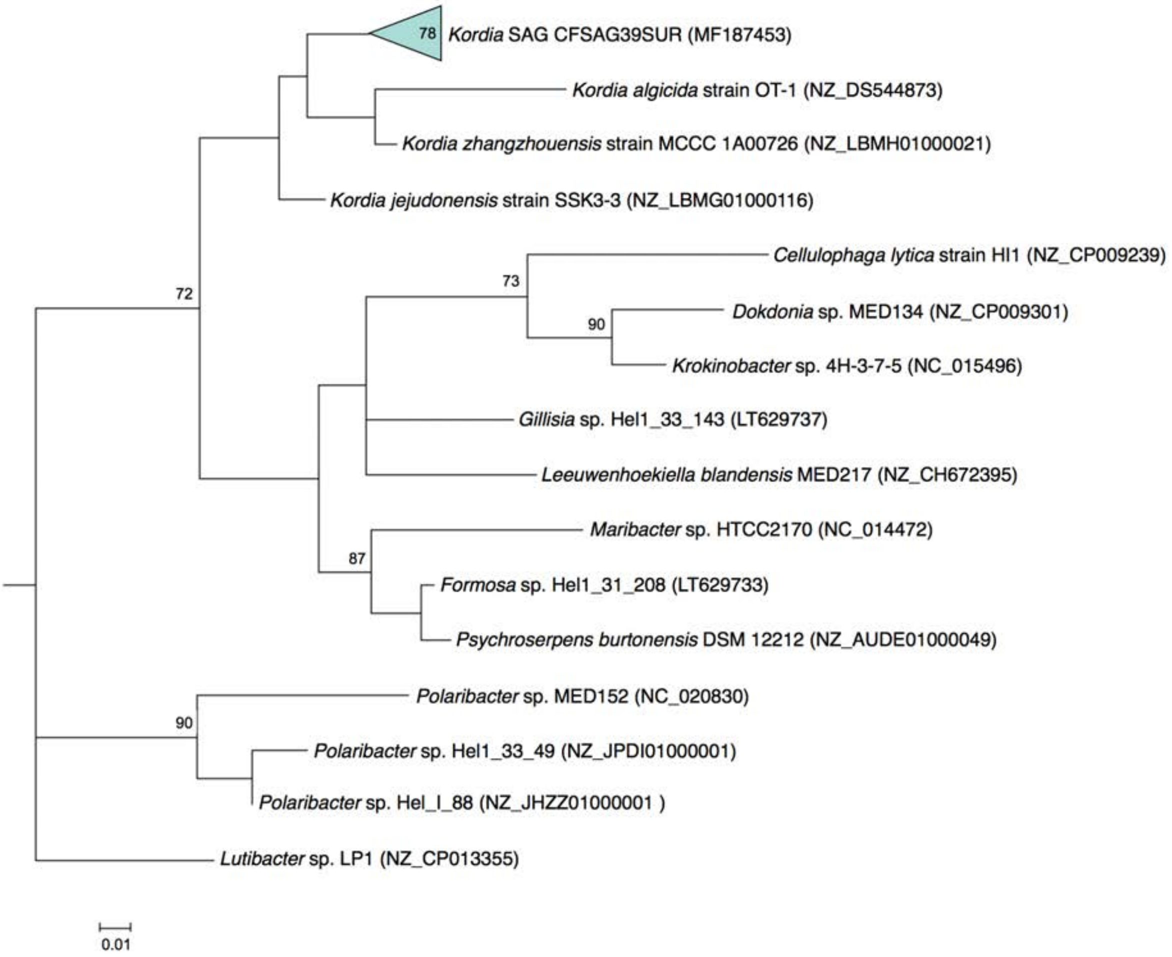
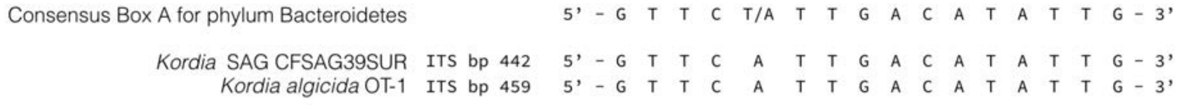

78
Kordia SAG CFSAG39SUR (MF187453)
Kordia algicida strain OT-1 (NZ_DS544873)
Kordia zhangzhouensis strain MCCC 1A00726 (NZ_LBMH01000021)
Kordia jejudonensis strain SSK3-3 (NZ_LBMG01000116)
72
Cellulophaga lytica strain HI1 (NZ_CP009239)
73
Dokdonia sp. MED134 (NZ_CP009301)
90
Krokinobacter sp. 4H-3-7-5 (NC_015496)
Gillisia sp. Hel1_33_143 (LT629737)
Leeuwenhoekiella blandensis MED217 (NZ_CH672395)
Maribacter sp. HTCC2170 (NC_014472)
87
Formosa sp. Hel1_31_208 (LT629733)
Psychroserpens burtonensis DSM 12212 (NZ_AUDE01000049)
Polaribacter sp. MED152 (NC_020830)
90
Polaribacter sp. Hel1_33_49 (NZ_JPDI01000001)
Polaribacter sp. Hel_I_88 (NZ_JHZZ01000001)
Lutibacter sp. LP1 (NZ_CP013355)
0.01

FIGURE S1. Maximum likelihood tree based on partial 23S rRNA gene sequences (415 bp), showing relationships between the representative sequence of the identical 23S rRNA gene from the 78 *Kordia* SAGs (CFSAG39DCM) and members of the Bacteroidetes phylum. Only bootstrap values ≥ 70% are shown at the nodes. All the sequences have been retrieved from the JGI IMG Database.



```
Consensus Box A for phylum Bacteroidetes               5' - G  T  T  C  T/A  T  T  G  A  C  A  T  A  T  T  G - 3'

        Kordia SAG CFSAG39SUR  ITS bp 442          5' - G  T  T  C   A   T  T  G  A  C  A  T  A  T  T  G - 3'
             Kordia algicida OT-1  ITS bp 459          5' - G  T  T  C   A   T  T  G  A  C  A  T  A  T  T  G - 3'
```

FIGURE S2. BoxA sequences of Kordia SAG CFSAG39SUR and Kordia algicida OT-1 aligned to consensus Box A sequence for phylum Bacteroidetes.

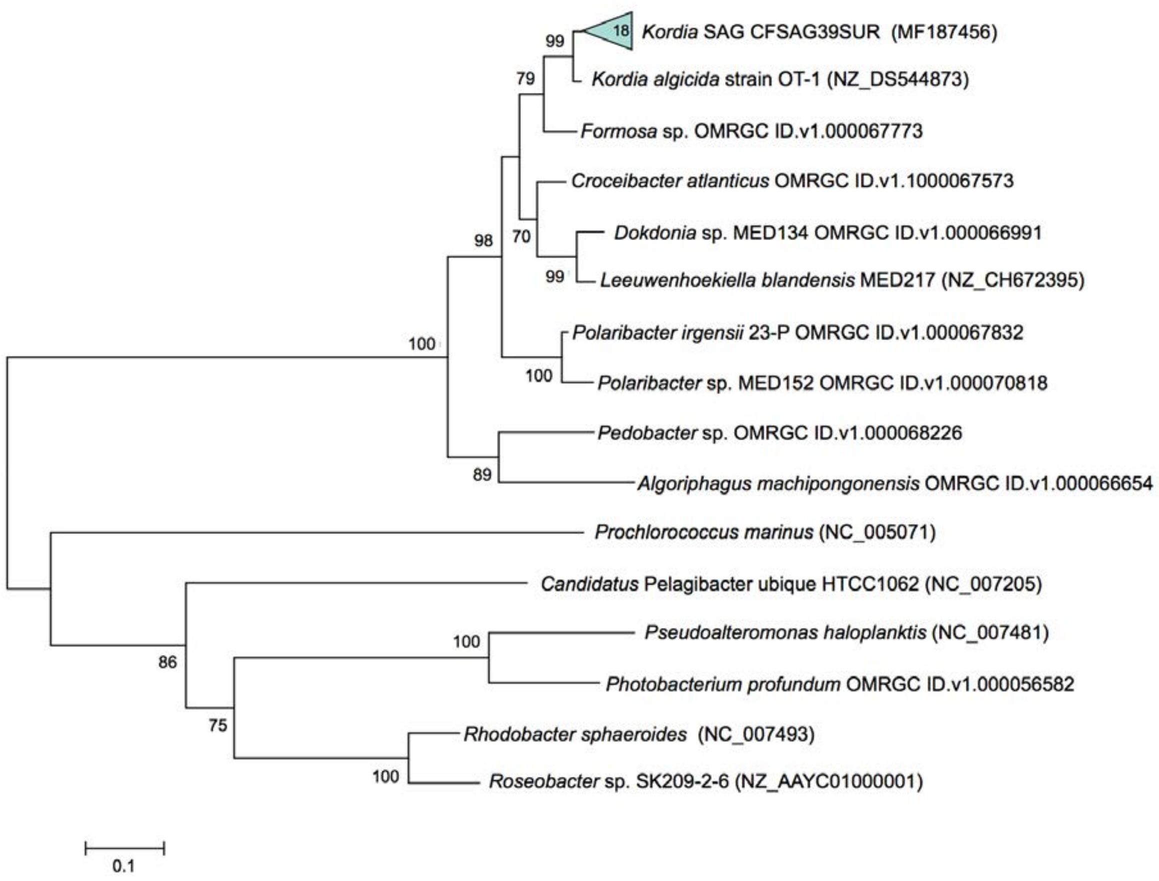

18 Kordia SAG CFSAG39SUR (MF187456)
Kordia algicida strain OT-1 (NZ_DS544873)
Formosa sp. OMRGC ID.v1.000067773
Croceibacter atlanticus OMRGC ID.v1.1000067573
Dokdonia sp. MED134 OMRGC ID.v1.000066991
Leeuwenhoekiella blandensis MED217 (NZ_CH672395)
Polaribacter irgensii 23-P OMRGC ID.v1.000067832
Polaribacter sp. MED152 OMRGC ID.v1.000070818
Pedobacter sp. OMRGC ID.v1.000068226
Algoriphagus machipongonensis OMRGC ID.v1.000066654
Prochlorococcus marinus (NC_005071)
Candidatus Pelagibacter ubique HTCC1062 (NC_007205)
Pseudoalteromonas haloplanktis (NC_007481)
Photobacterium profundum OMRGC ID.v1.000056582
Rhodobacter sphaeroides (NC_007493)
Roseobacter sp. SK209-2-6 (NZ_AAYC01000001)
99
79
98
70
99
100
100
89
86
100
75
100
0.1

Figure S3. Maximum likelihood tree based on same length partial RpoB aminoacid sequences, showing relationships between the consensus sequence of the 18 identical rpoB genes from the Kordia SAGs, the closest hits from the Ocean Microbiome Reference Gene Catalog (OMRGC) (**?**) and other sequences retrieved from the JGI IMG Database. Cyanobacteria and Proteobacteria act as outgroups and only bootstrap values ≥ 70% are shown at the nodes

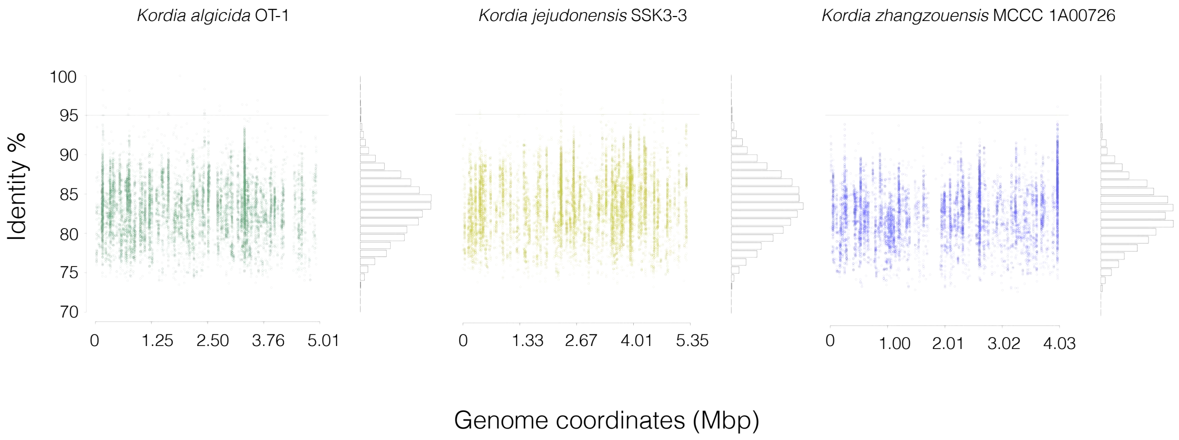


Figure S4. FRA plots for three of the four available *Kordia* genomes mapped against ST_085 mesopelagic sample (270 m, 0.2-3 µm). Metagenomic reads are displayed according to their position against de reference genome (X axis, genome coordinates) and their identity percentage (Y axis). The grey line located at 95% identity in the Y axis corresponds to 95% average nucleotide identity (ANI), threshold for bacterial species categorization. The bars on the right show the amount of reads mapped at each identity percentage (note different scales). These quantitative values are normalized to the sequencing depth and genomic size. Duplicates and reads mapped to the ribosomal operon have been removed to avoid bias. Plots for fourth *Kordia* genome, AAA285-F05 can be found at Figure **??** of this article.

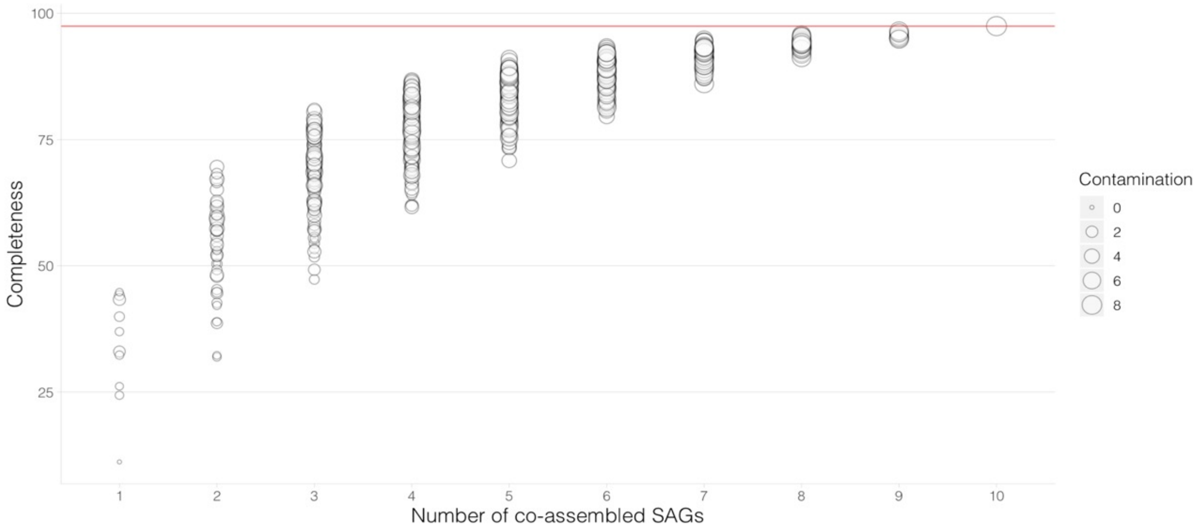


FIGURE S5. Genome completeness curve of all possible co-assemblies from pairs to the combination of 10 SAGs. Dot size varies according to contamination percentage as shown in legend and red horizontal line is located at the maximum genome completeness achieved in the 1023 combinations (97.48%).

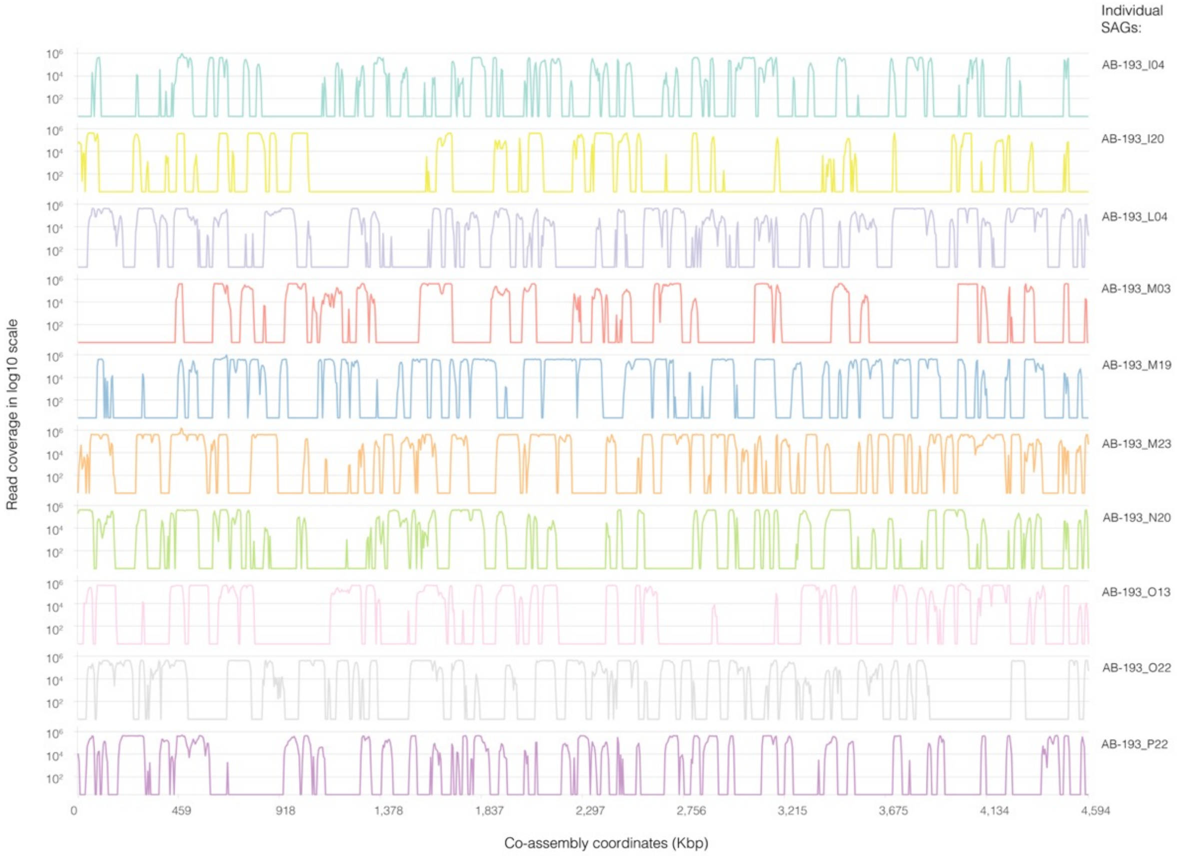


FIGURE S6. Coverage of the co-assembly by every individual MAG. Coverage is calculated as number of reads per nucleotide position and is shown in a log10 scale. Flat lines represent no coverage (0x). For figure execution purposes coverage has been collapsed by bins of 4,594 bp each.

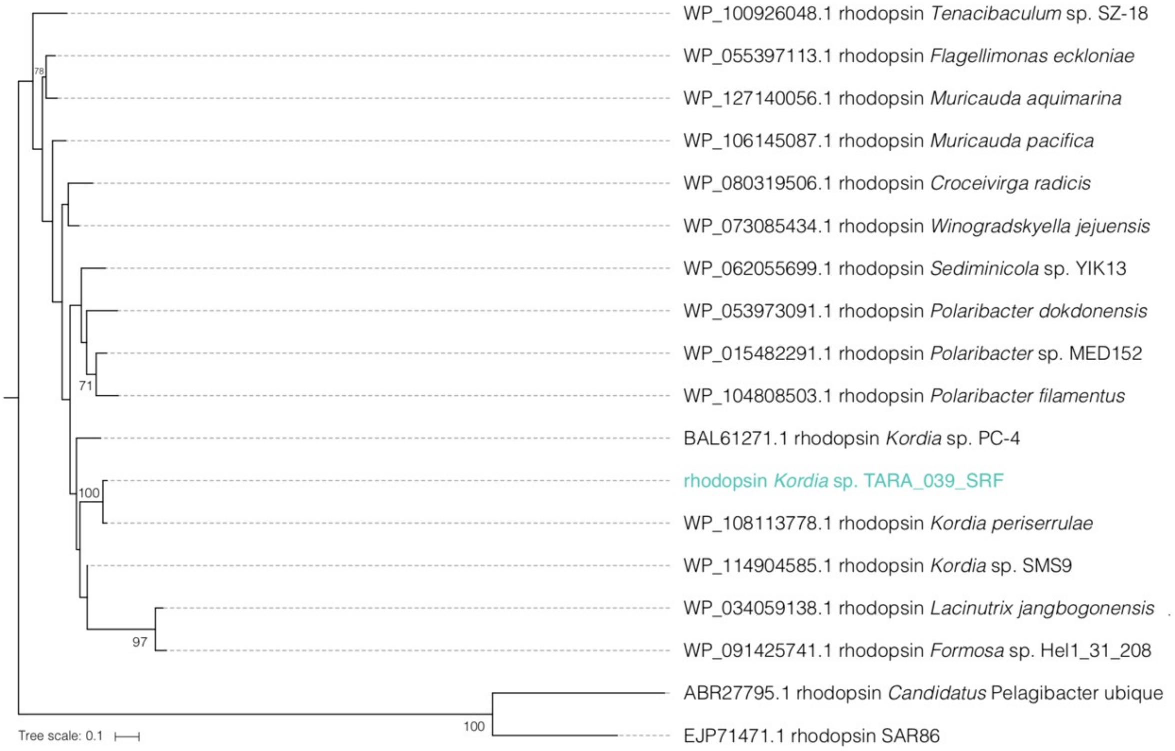

WP_100926048.1 rhodopsin *Tenacibaculum* sp. SZ-18
WP_055397113.1 rhodopsin *Flagellimonas eckloniae*
WP_127140056.1 rhodopsin *Muricauda aquimarina*
WP_106145087.1 rhodopsin *Muricauda pacifica*
WP_080319506.1 rhodopsin *Croceivirga radicis*
WP_073085434.1 rhodopsin *Winogradskyella jejuensis*
WP_062055699.1 rhodopsin *Sediminicola* sp. YIK13
WP_053973091.1 rhodopsin *Polaribacter dokdonensis*
WP_015482291.1 rhodopsin *Polaribacter* sp. MED152
WP_104808503.1 rhodopsin *Polaribacter filamentus*
BAL61271.1 rhodopsin *Kordia* sp. PC-4
rhodopsin *Kordia* sp. TARA_039_SRF
WP_108113778.1 rhodopsin *Kordia periserrulae*
WP_114904585.1 rhodopsin *Kordia* sp. SMS9
WP_034059138.1 rhodopsin *Lacinutrix jangbogonensis*
WP_091425741.1 rhodopsin *Formosa* sp. Hel1_31_208
ABR27795.1 rhodopsin *Candidatus* Pelagibacter ubique
EJP71471.1 rhodopsin SAR86
78
71
100
97
100
Tree scale: 0.1

FIGURE S7. Maximum Likelihood Phylogenetic reconstruction based on the complete proteorhodopsin sequence of *Kordia* sp. TARA_039_SRF and its closest hits

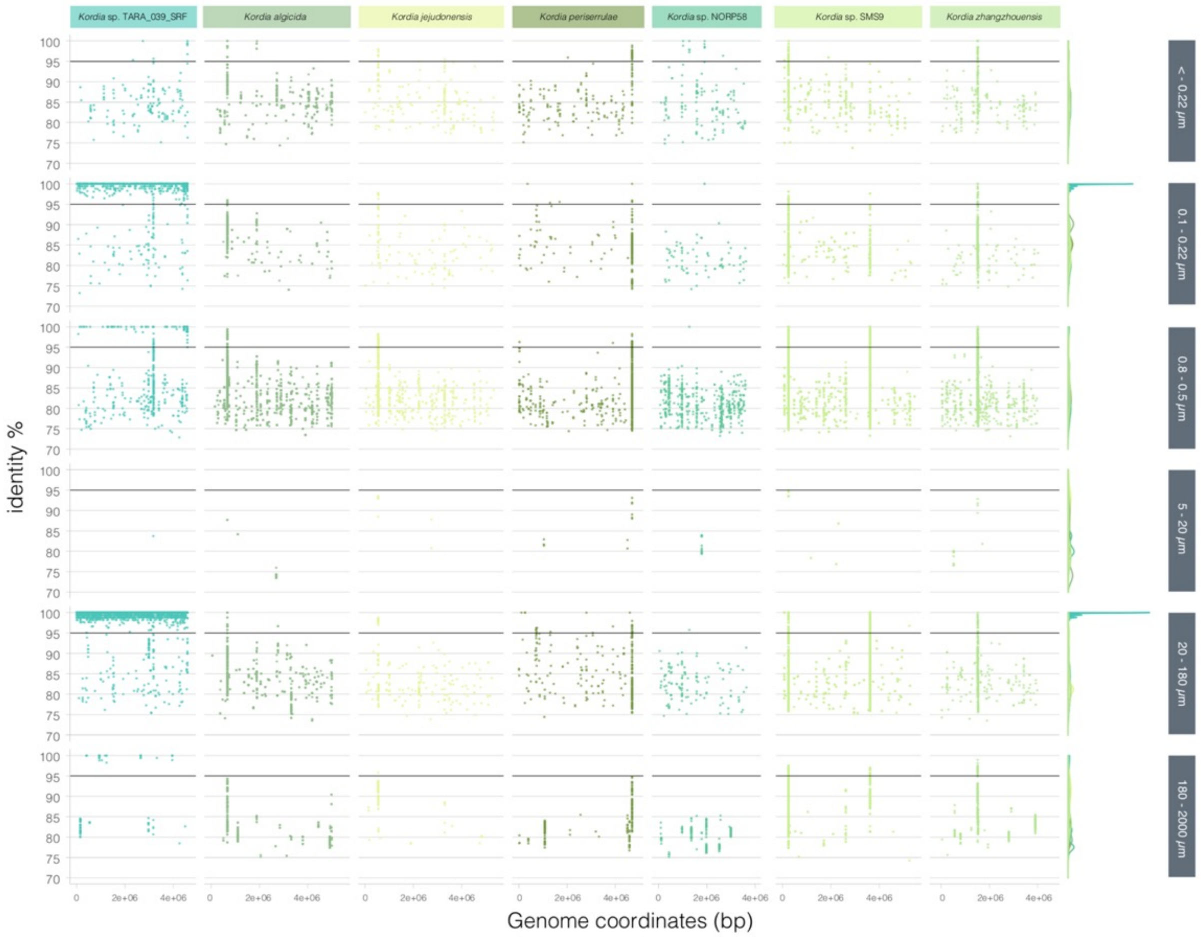


FIGURE S8. Recruitment plots of Kordia genomes against metagenomes of surface seawater (5 m) from station TARA_039. Each dot represents a read that has mapped competitively against a genome at a certain position (X axis) and with a certain identity (Y axis). Right lineplot represents the total of reads recruited by each genome (in its particular color) through the different identity percentages. Scales differ between lineplots in every row. Each row represents a different planktonic size fraction, that can be seen in the dark labelling on the right.

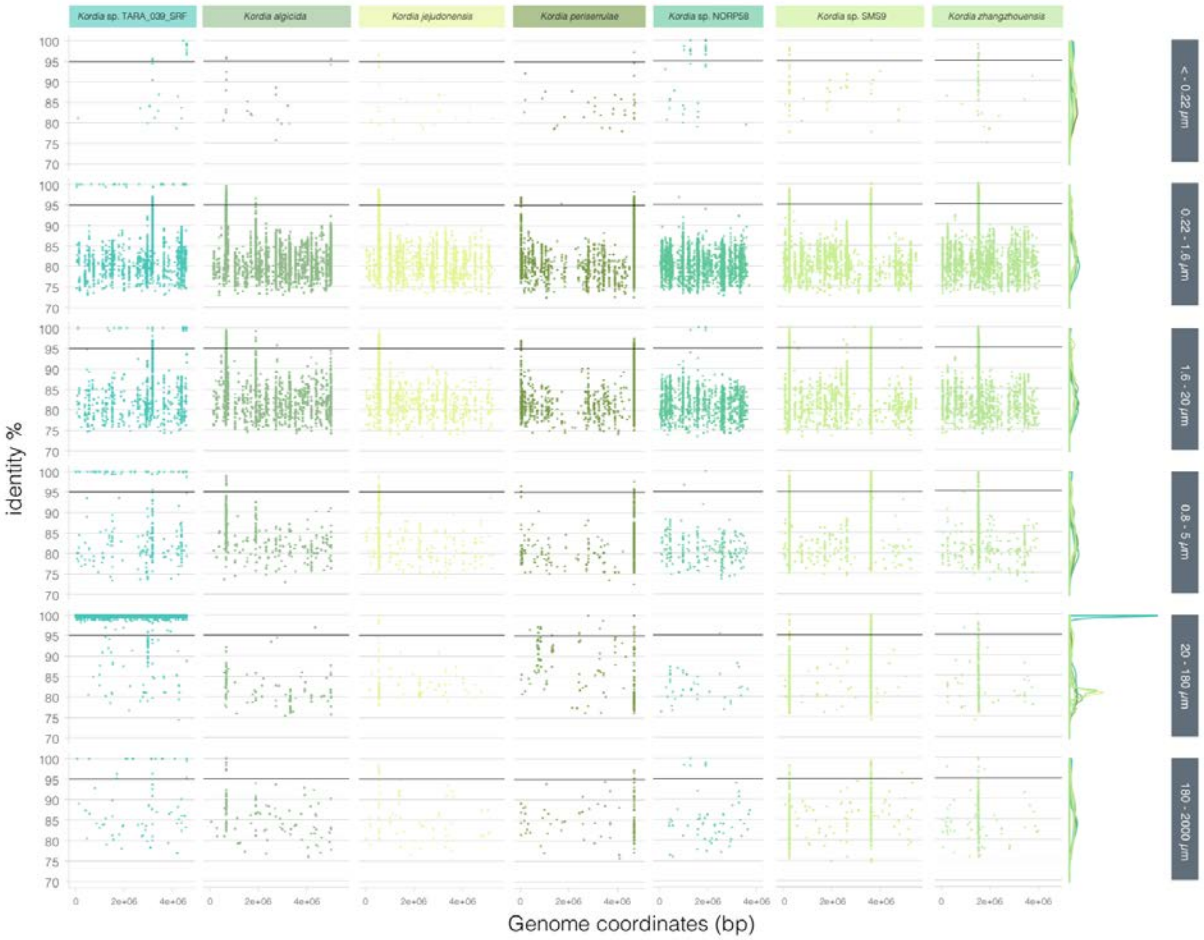


FIGURE S9. Recruitment plot of Kordia genomes against metagenomes of DCM seawater (25 m) from station TARA_039. Each dot represents a read that has mapped competitively against a genome at a certain position (X axis) and with a certain identity (Y axis). Right lineplot represents the total of reads recruited by each genome (in its particular color) through the different identity percentages. Scales differ between lineplots in every row. Each row represents a different planktonic size fraction, that can be seen in the dark labelling on the right.

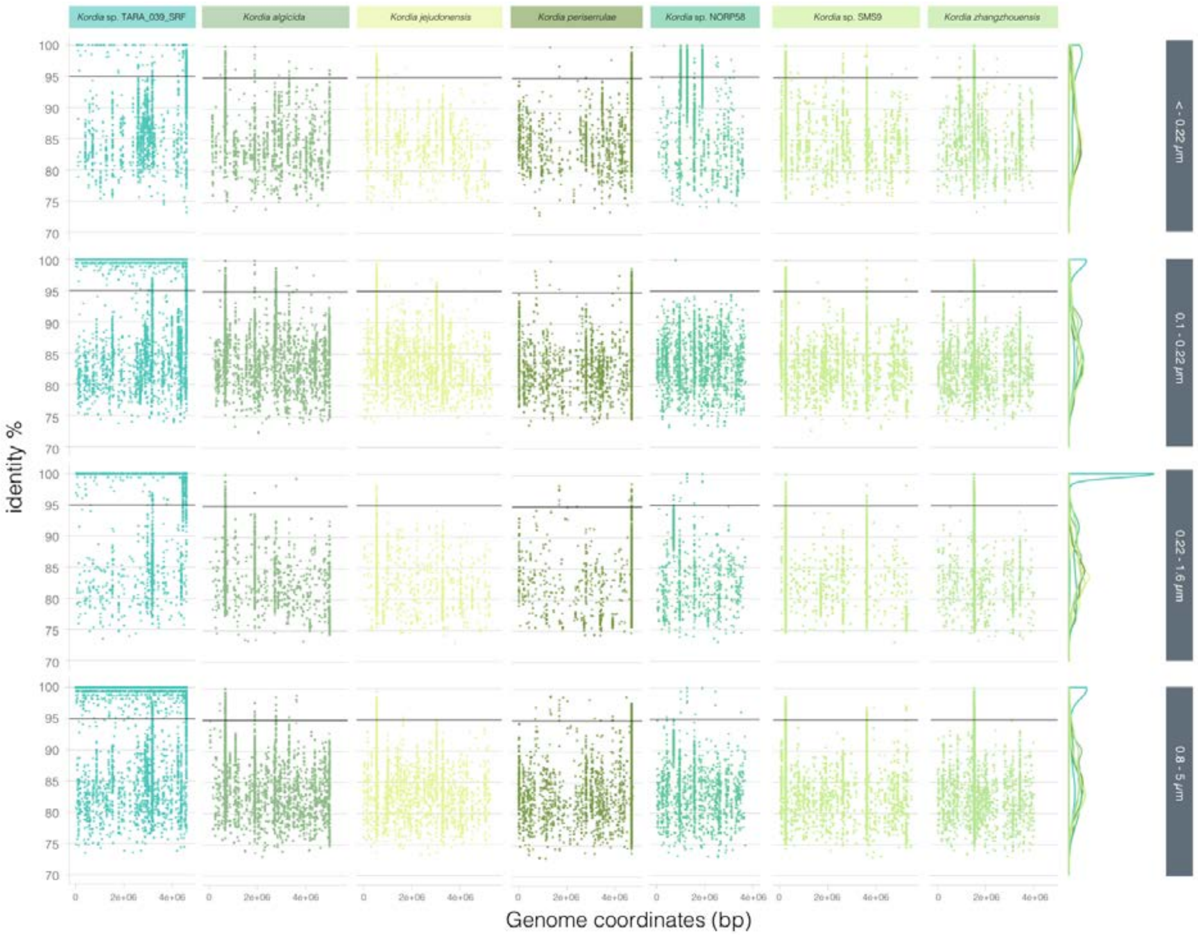


FIGURE S10. Recruitment plot of Kordia genomes against metagenomes of mesopelagic seawater (270 m) from station TARA_039. Each dot represents a read that has mapped competitively against a genome at a certain position (X axis) and with a certain identity (Y axis). Right lineplot represents the total of reads recruited by each genome (in its particular color) through the different identity percentages. Scales differ between lineplots in every row. Each row represents a different planktonic size fraction, that can be seen in the dark labelling on the right.

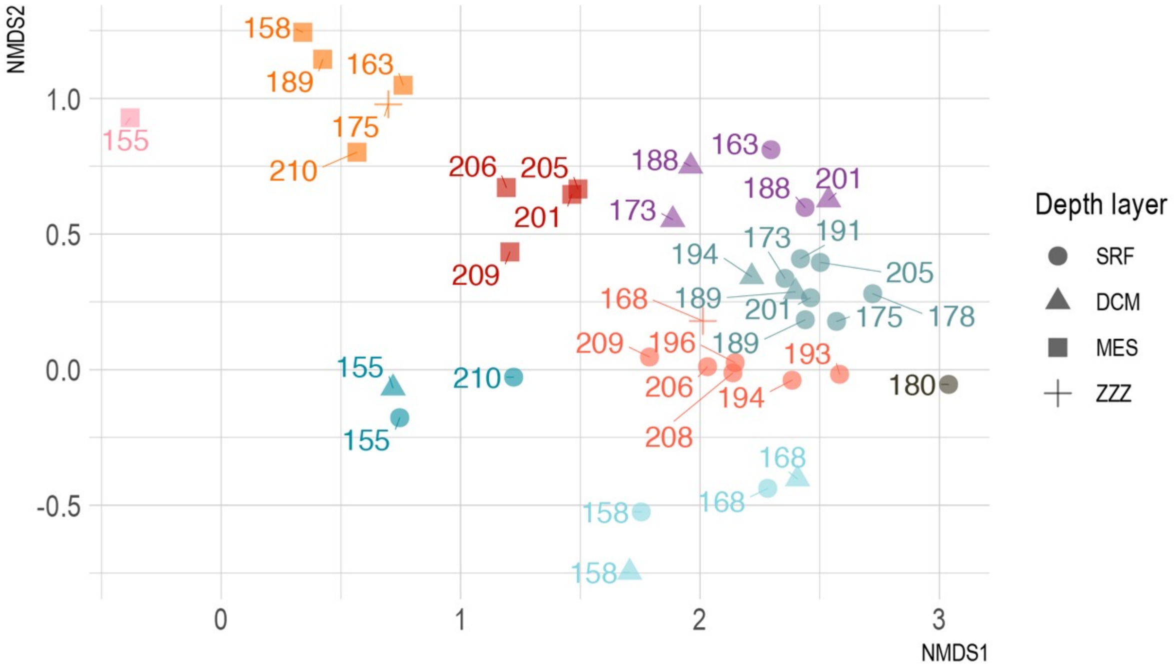


FIGURE S11. NMDS of miTAG community composition of Tara Oceans Polar Circle metagenomes. Colors delimit the grouping of samples for later co-assembly. Shape refers to the layer of origin in the water column of each metagenomic sample.

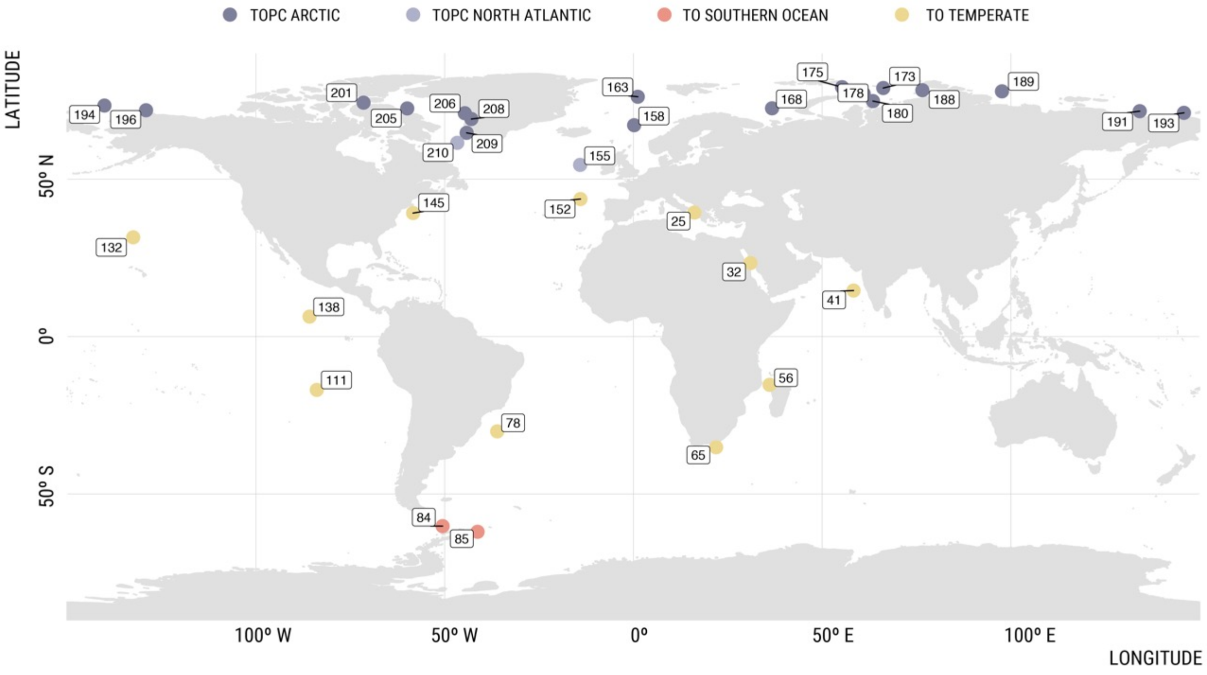


FIGURE S12. Map with all meta-omics samples used in the study. Samples are colored based on the expedition (TOPC - Tara Oceans Polar Circle, TO - Tara Oceans). SM Table 2 contains more details about environmental metadata of these stations.

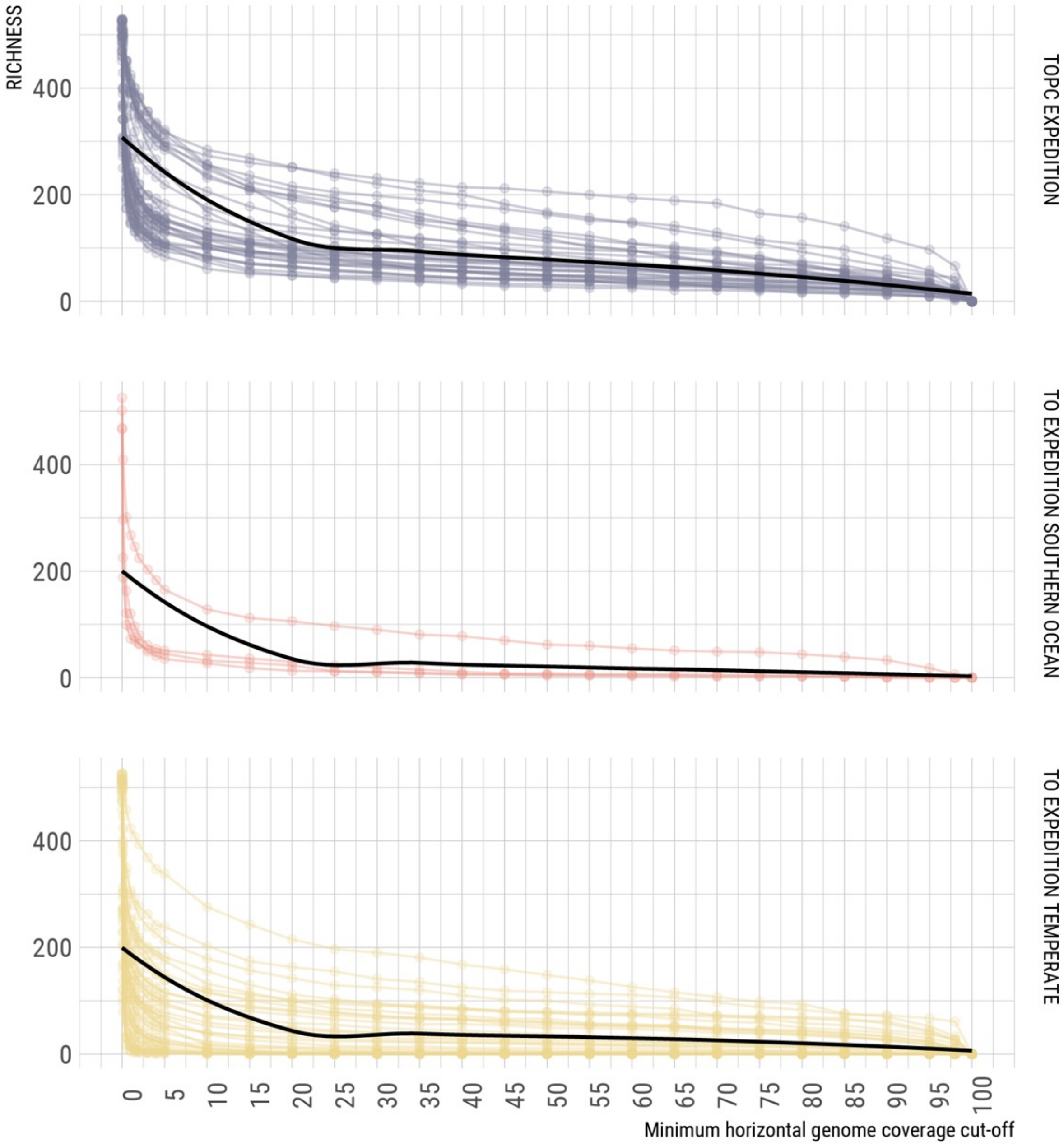


Figure S13. Effect on samples' richness with increasing minimum horizontal coverage thresholds of metagenomic mappings. Purple data represents metagenomic samples from the Tara Oceans Polar Circle dataset (TOPC). Red data represents metagenomic samples from the Tara Oceans Southern Ocean dataset and yellow data represents metagenomic samples from temperate latitudes from the Tara Oceans Expedition.

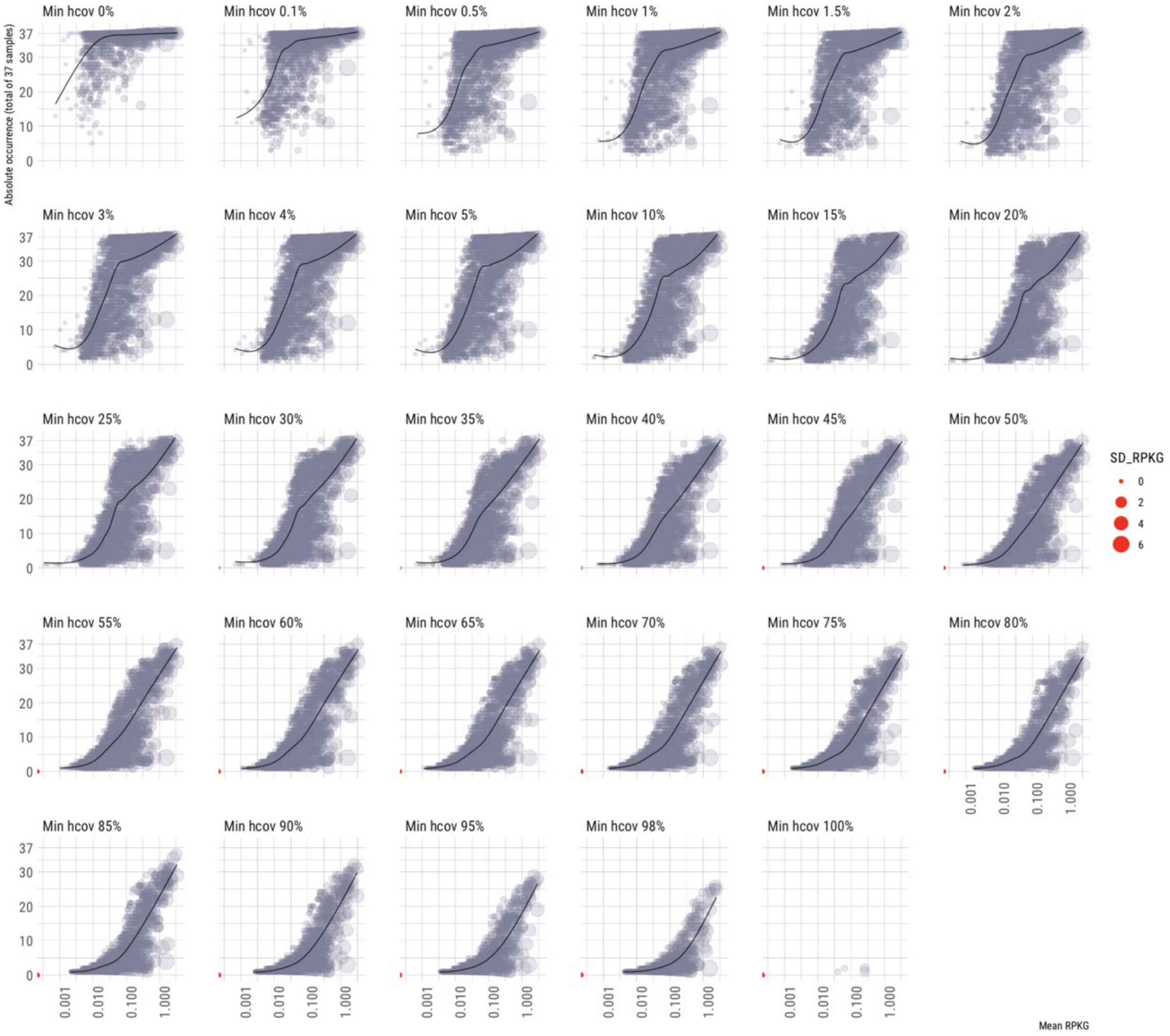


FIGURE S14. Effect of increasingly stringent minimum horizontal coverages in MAG recruitments in TOPC samples. Each dot represents a MAG and its location in the X axis (mean metagenomic RPKG) and Y axis (occurrence in the 37 TOPC samples) varies with every minim horizontal coverage threshold applied to their recruitments (facets). Size of dot represents the standard deviation of the mean RPKG. Red dots indicate that mean abundance and occurrence equal 0.

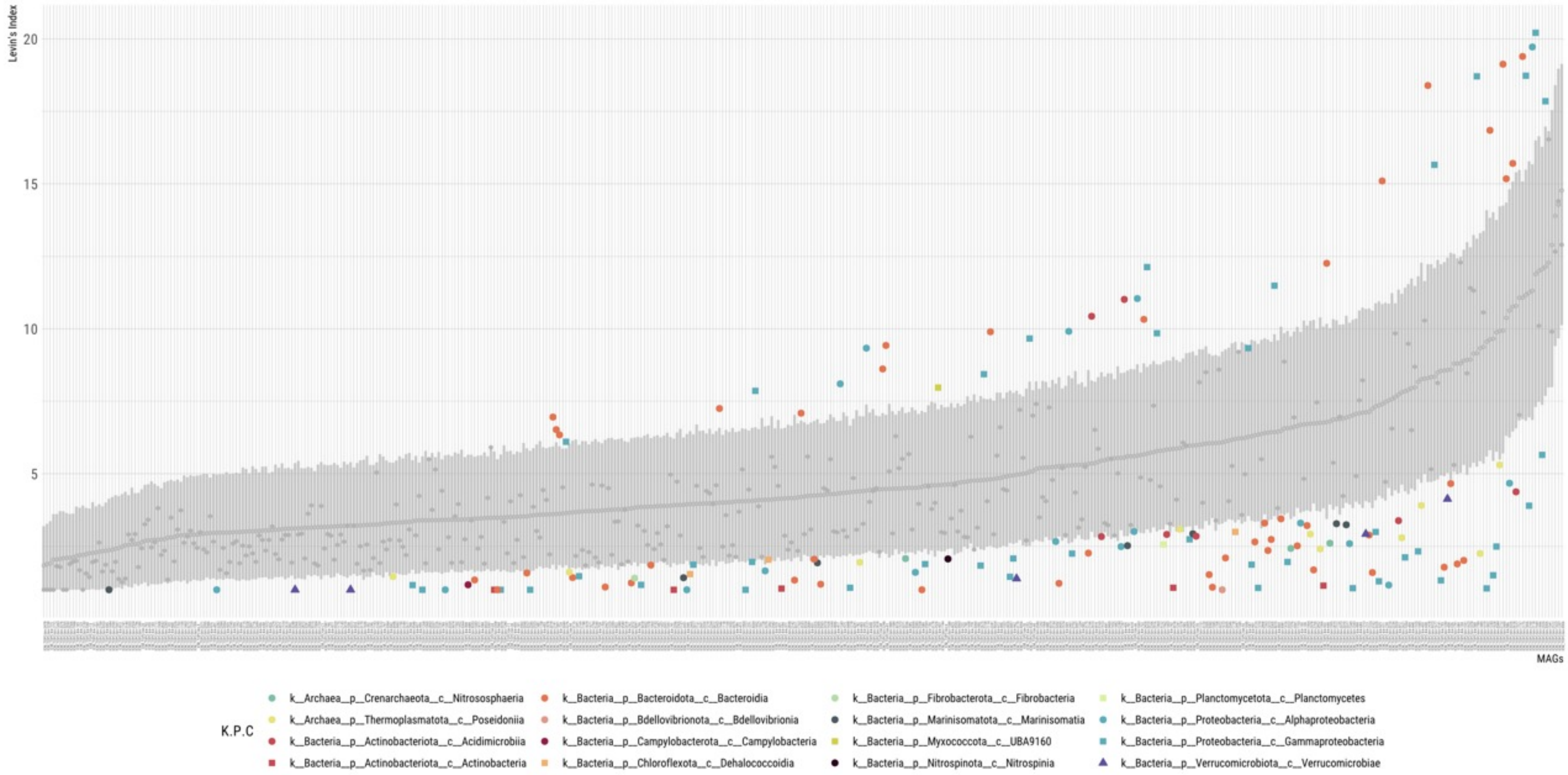


FIGURE S15. Classification of generalists and specialists based on 1000 random permutations of Levin's Index in the 37 Arctic samples. Distribution of Levin's index (B) calculations in the different MQ and HQ MAGs (X axis). MAGs are sorted by the mean B value within the 1000 random permutations (dark grey dot). Grey vertical lines indicate the confidence interval and actual B value of each MAG is depicted in color if located outside the confidence interval or grey if located within. Color of dots is based the phylum annotation of each MAG, and shape corresponds to taxonomic rank class. Dots above the confidence interval represent habitat generalists, while dots below the confidence interval represent habitat specialists.

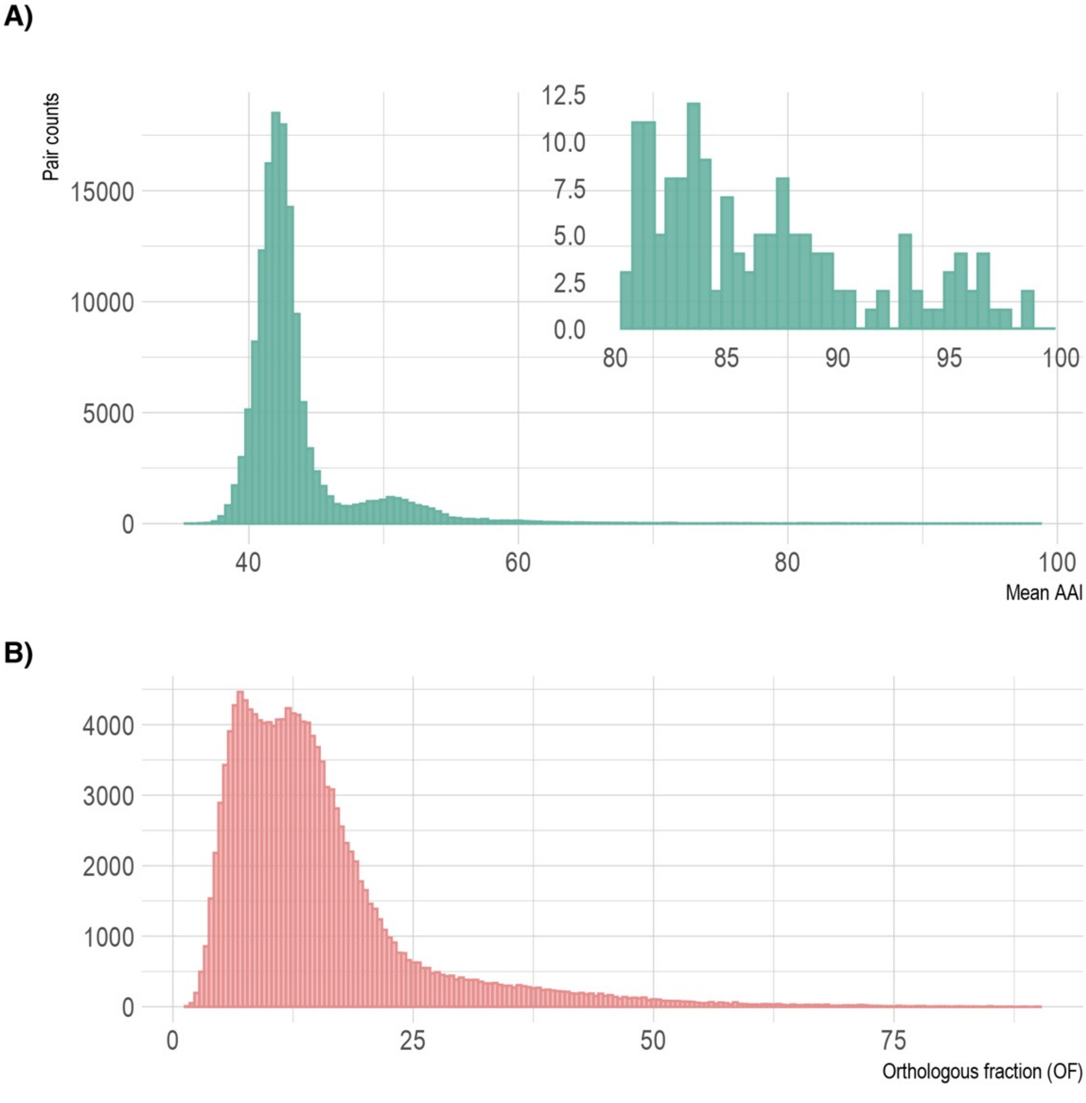


FIGURE S16. Pairwise average aminoacid identities (AAI) of 530 MQ and HQ MAGs. A) Histogram of pairwise AAI between the 530 MQ and HQ MAGs. Top right window shows a zoom of the pair counts with AAI values bigger than 80%. B) Histogram of pairwise shared orthologous fraction between 530 MQ and HQ MAGs.

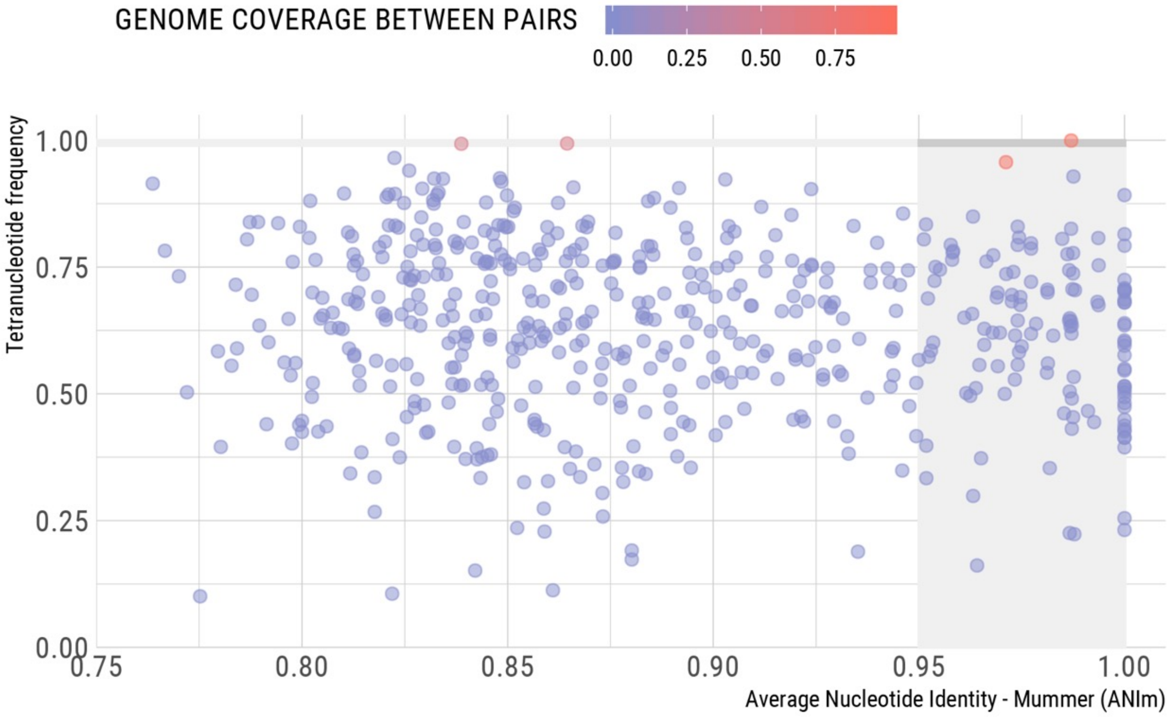


Figure S17. HQ MAGs ANI vs Tetranucleotide frequencies. Dotplot with pairs of HQ MAGs located based on their Average Nucleotide Identity (ANI) and the identity between their tetranucleotide frequencies. Pairs of genomes reported as belonging to the same species show ANI >95% and tetranucleotide frequences >99% (**?**), MAGs of redundant species are located in the area where ANI and tetranucletoide frequency gray boxes converge. Color range of the dots is representative of the percentage of genome that is shared between each pair of MAGs.

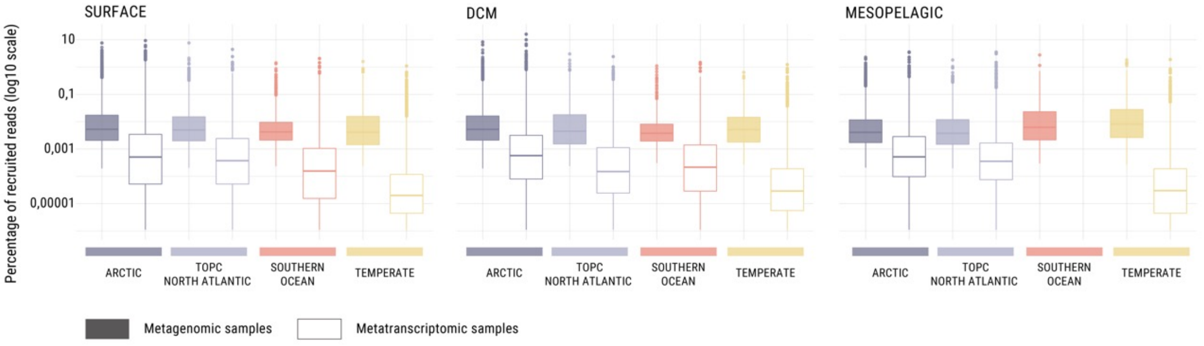


Figure S18. Distribution of individual recruitments by 3550 bins. Filled boxplots represent metagenomic samples, empty boxplots represent metatranscriptomic boxplots. Recruitment are grouped and coloured by region, TOPC North Atlantic stands for the North Atlantic samples collected during the Tara Oceans Polar Circle expedition.

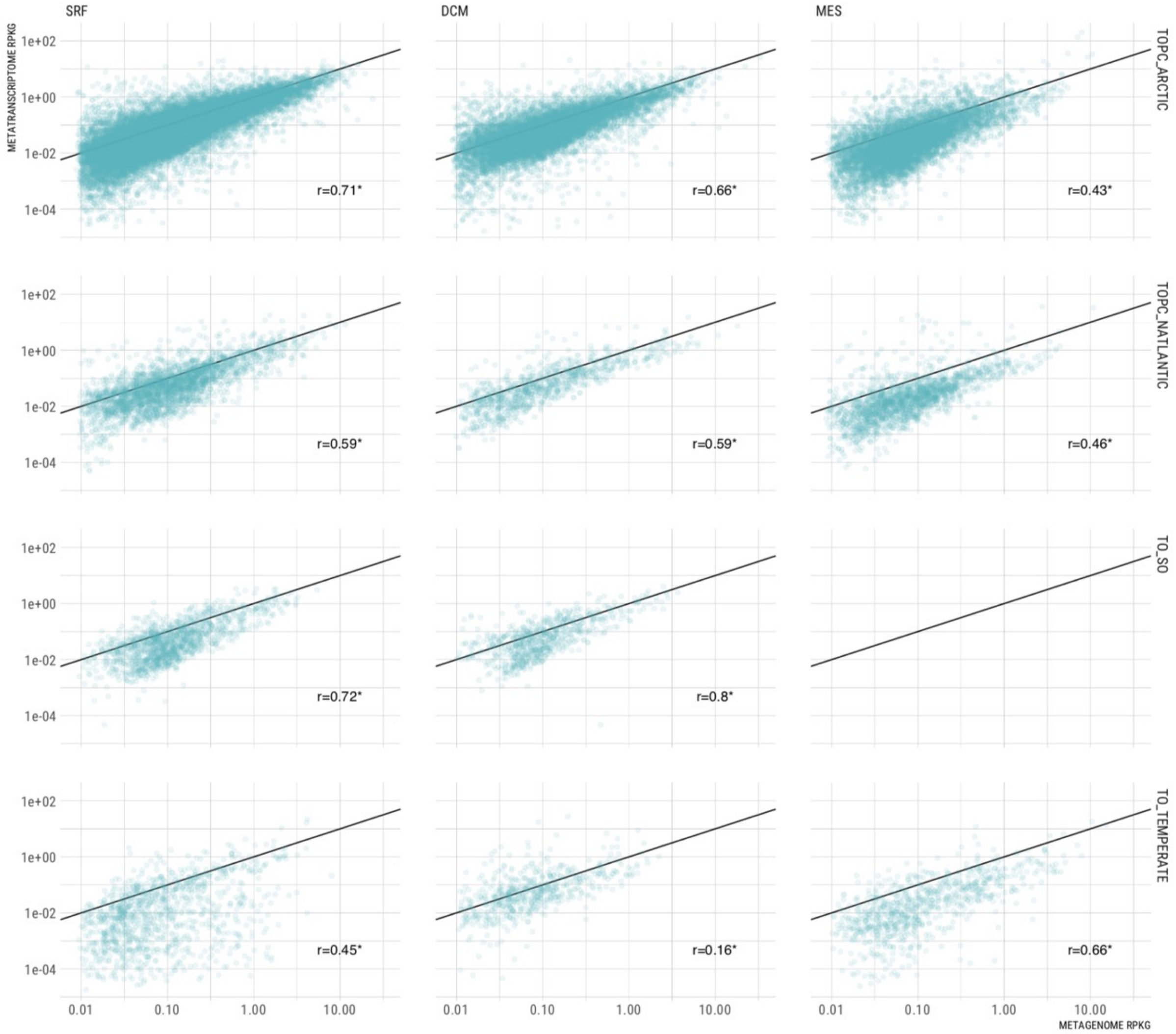


FIGURE S19. Abundance vs activity recruitment plots. Dotplots of metagenomic RPKGs (X axis) vs metatranscriptomic RPKGs (Y axis) of individual bins, in the different layers (columns) and latitudes (rows). Significant Pearson's correlations (p-value <0.05) are shown with an asterisk.

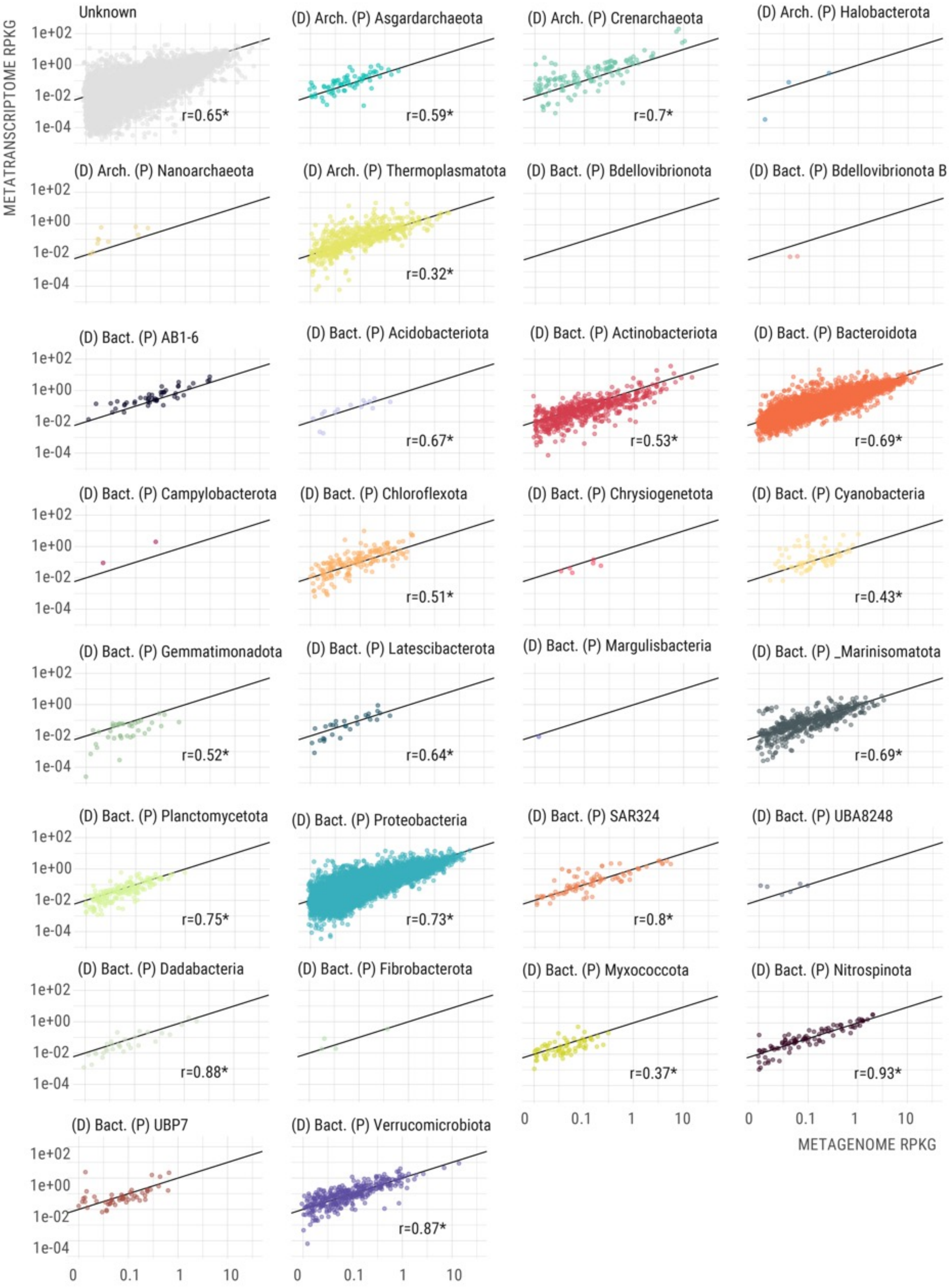


FIGURE S20. Abundance vs activity recruitment plots of different phyla in the Arctic metagenomes. Dotplots of metagenomic RPKGs (X axis) vs metatranscriptomic RPKGs (Y axis) of individual bins separated by phylum. Significant Pearson's correlations (p-value <0.05) are shown with an asterisk.

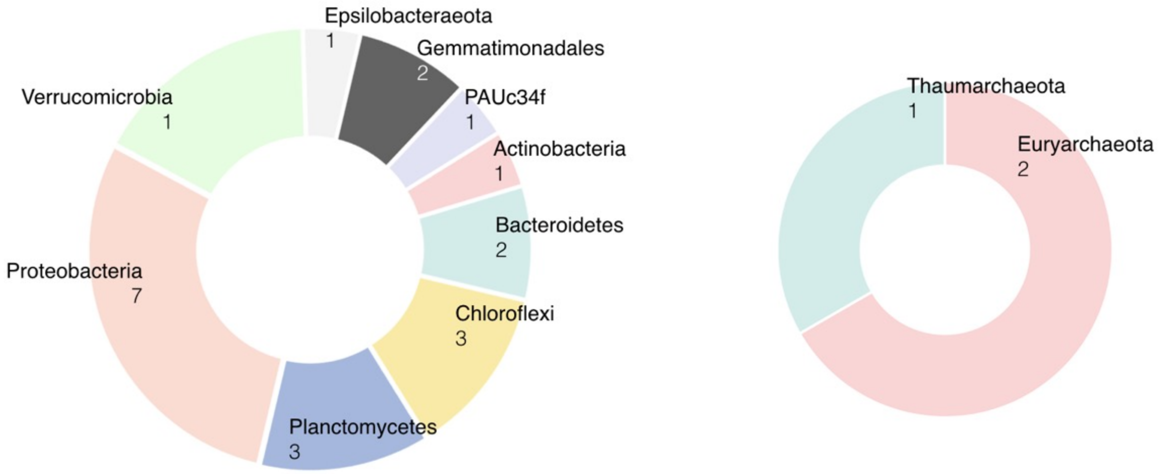


FIGURE S21. Taxonomic classification of the 27 partial ribosomal genes encoded in the 530 MQ and HQ MAGs. A) Bacteria domain. B) Archaea domain.

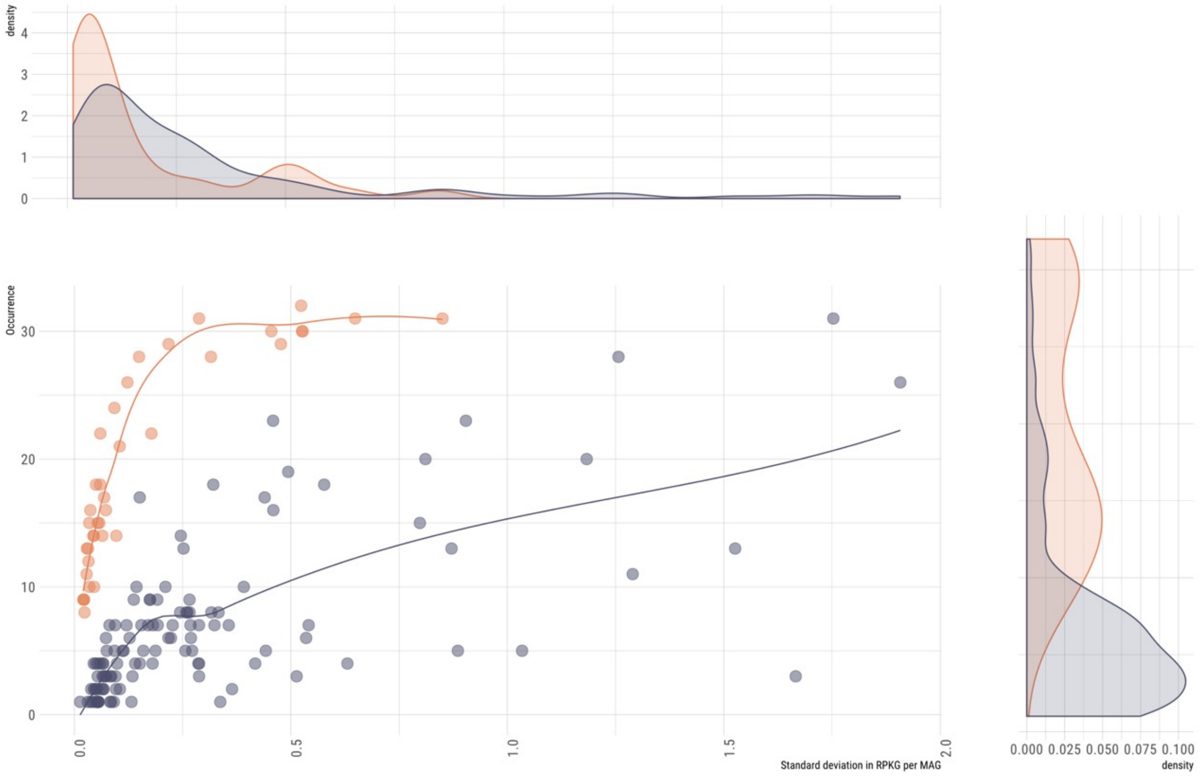


FIGURE S22. Occurrence and standard deviation of generalist and specialist MAGs. Distribution of MAGs according to their occurrence in the arctic metagenomics (Y axis) and the standard deviation of their mean recruitments (X axis, normalized by genome size and sequencing depth as RPKG), which is a proxy for the evenness of their distribution. Generalist MAGs are coloured in orange and specialists in blue.

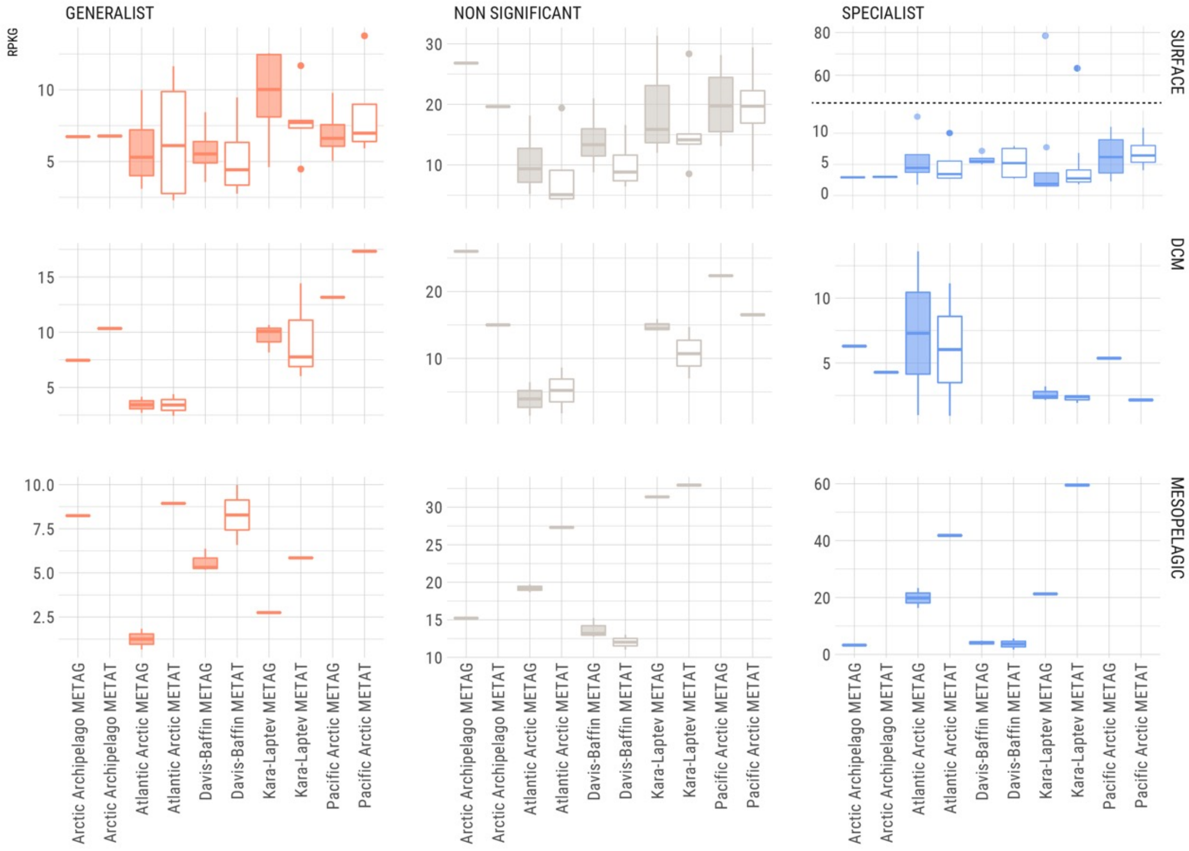


Figure S23. Distribution of samples based on their accumulated RPKGs of Bacteria MQ and HQ MAGs. Boxplots of the distribution of generalists (orange), specialists (blue) and non-significant (grey) groups of domain Bacteria in the different metagenomic (filled boxplots) and metatranscriptomic (empy boxplots) samples. Each row separates samples by their layer in the water column and in the X axis samples are categorised based on the CAFF region they belong and contain the code METAG and METAT that correspond to metagenomic recruitments and metatranscriptomic recruitments, respectively. Dashed line in Y axis of surface specialists indicates that Y scale has been cut.

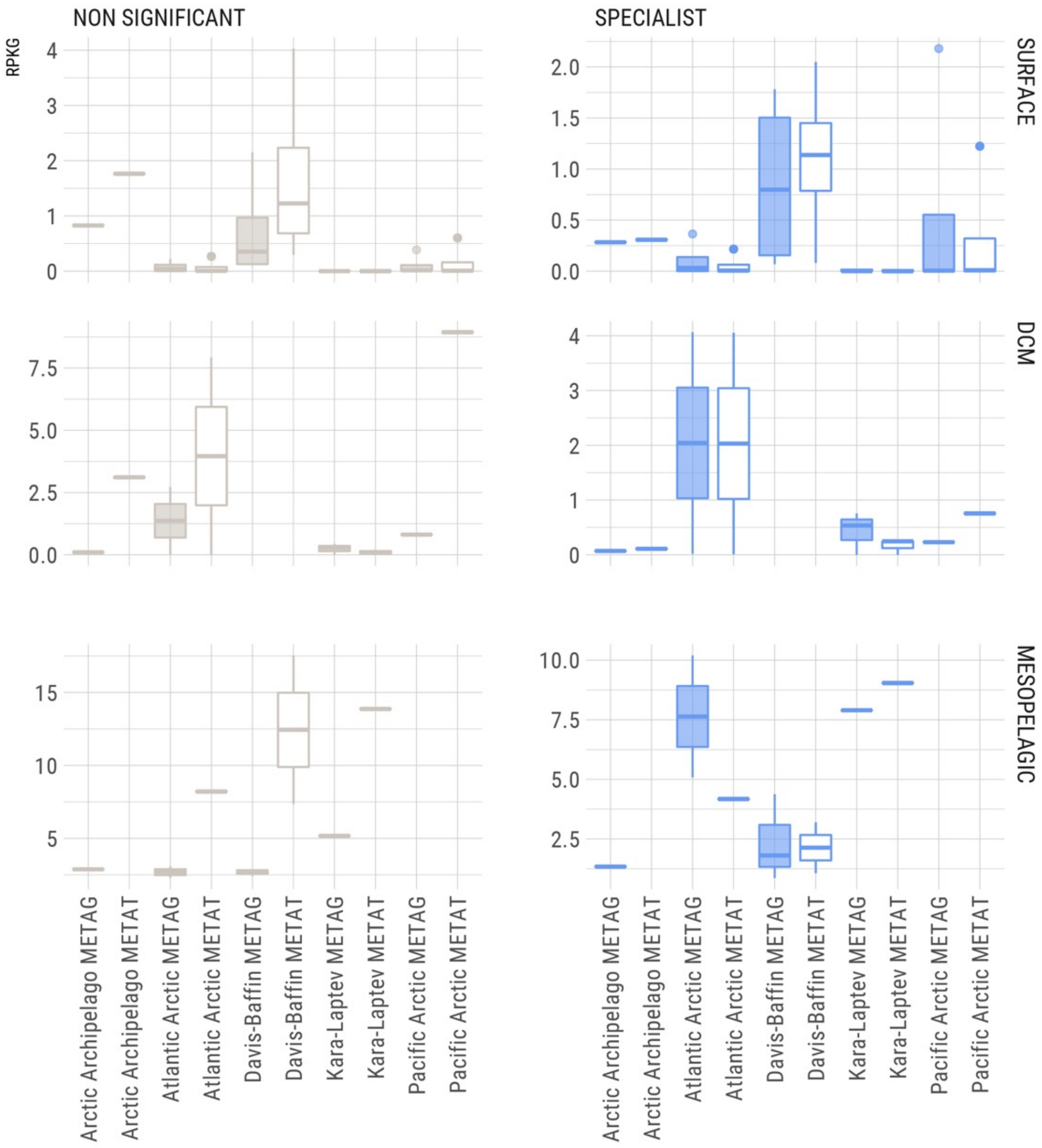


FIGURE S24. Distribution of samples based on their accumulated RPKGs of Archaea MQ and HQ MAGs. Boxplots of the distribution of specialists (blue) and non-significant (grey) groups of domain Archaea in the different metagenomic (filled boxplots) and metatranscriptomic (empy boxplots) samples. Each row separates samples by their layer in the water column and in the X axis samples are categorised based on the CAFF region they belong and contain the code METAG and METAT that correspond to metagenomic recruitments and metatranscriptomic recruitments, respectively.

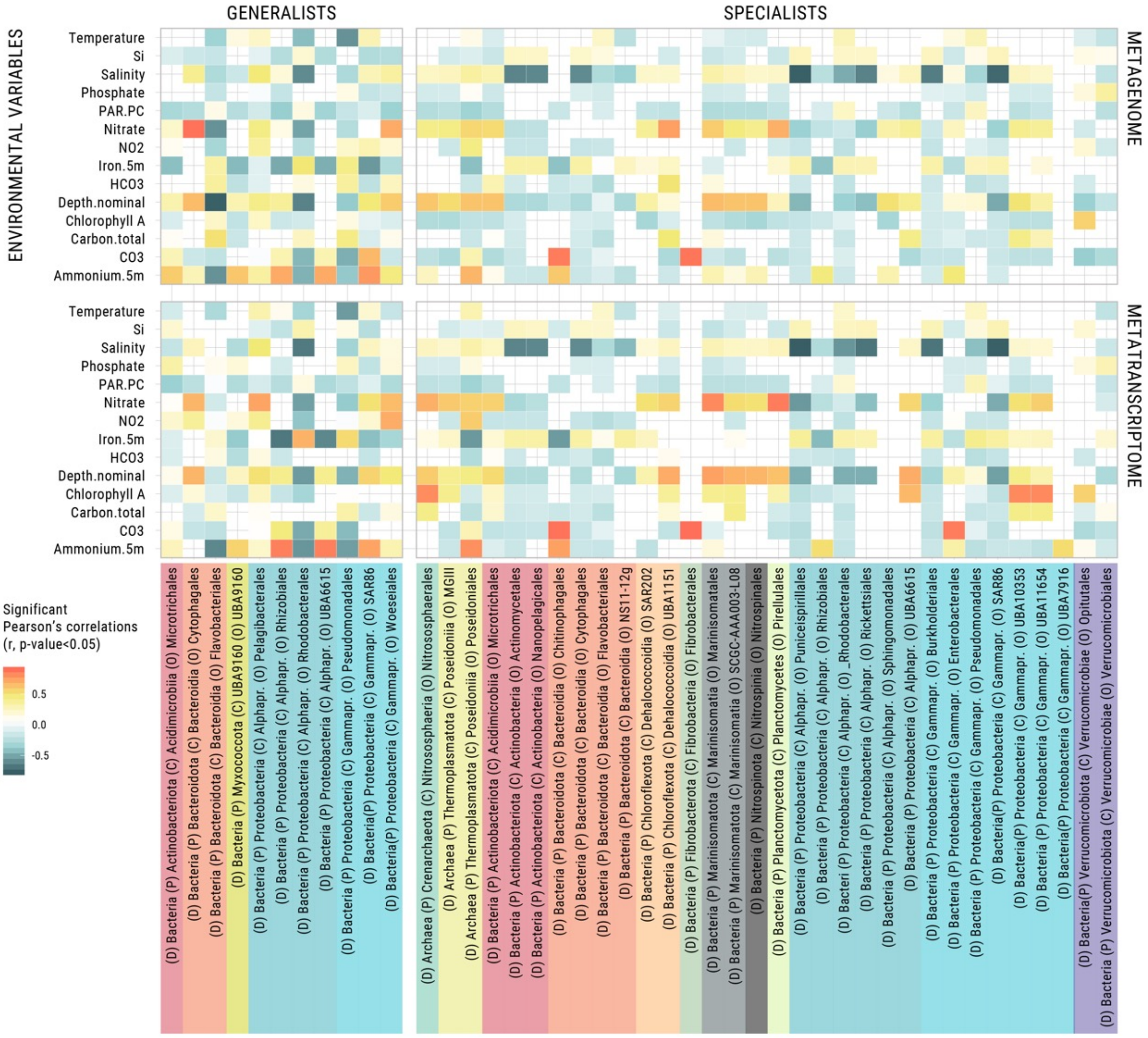


FIGURE S25. Significant correlations between niche breadth of MQ and HQ generalists/specialists and environmental variables. Significant Pearson's correlations (p-value <0.05) between accumulated metagenomic (top row) and metatranscriptomic (bottom row) reads and environmental variables (Y axis). Metagenomic and metatranscriptomic reads are accumulated within taxonomic orders. The left column contains correlations with generalist MAGs and right column contains correlations with specialist MAGs. Non-significant correlations are not shown.

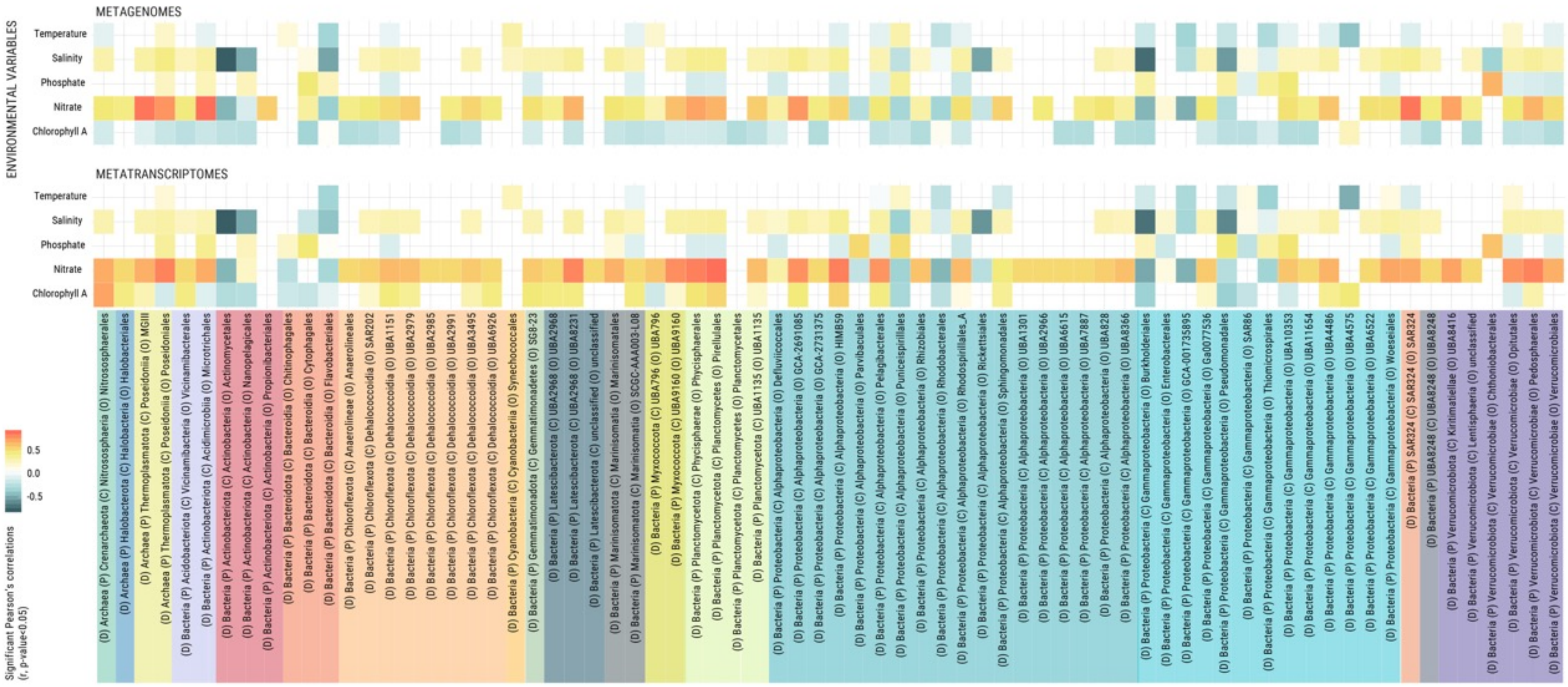


FIGURE S26. Significant correlations between niche breadth of MQ and HQ generalists/specialists and environmental variables. Significant Pearson's correlations (p-value <0.05) between accumulated metagenomic (top row) and metatranscriptomic (bottom row) reads and environmental variables (Y axis). Metagenomic and metatranscriptomic reads are accumulated within taxonomic orders. The left column contains correlations with generalist MAGs and right column contains correlations with specialist MAGs. Non-significant correlations are not shown.

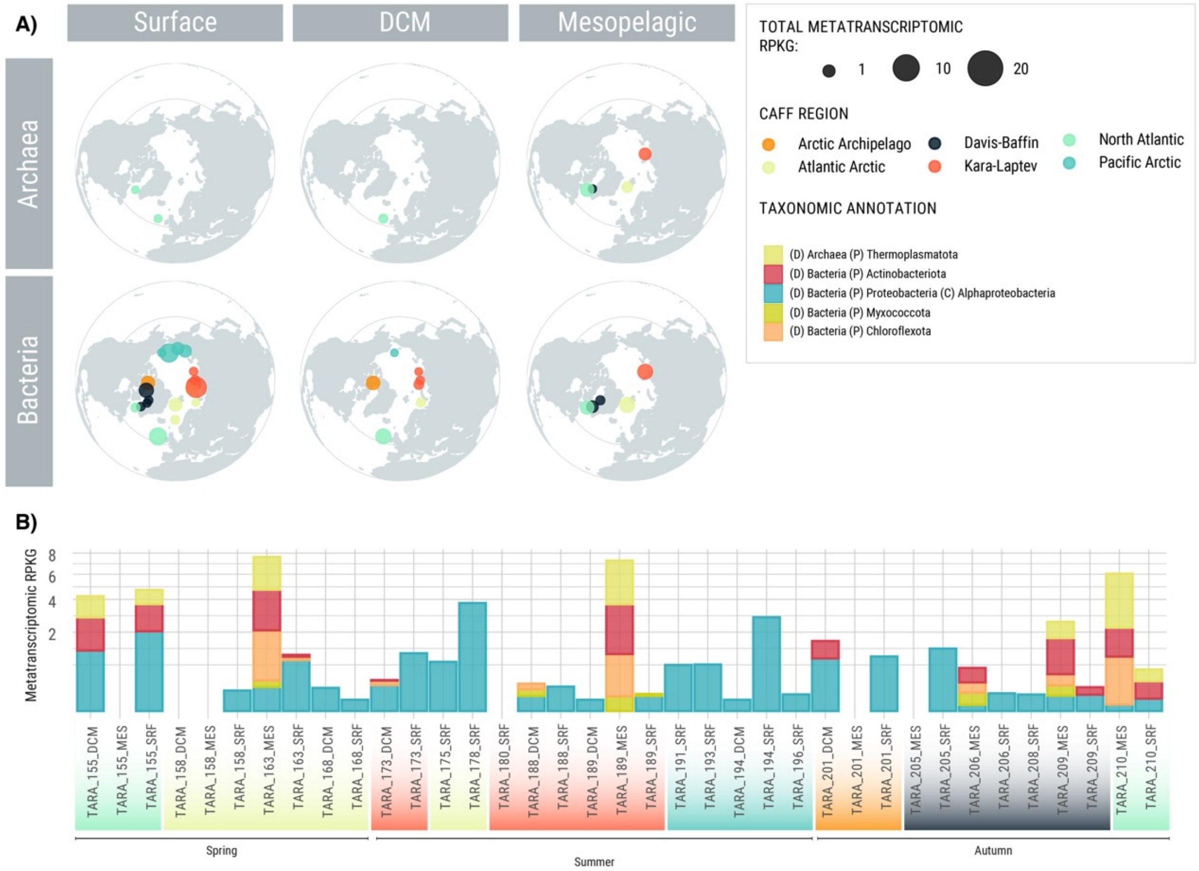


FIGURE S27. Accumulated expression of HQ and MQ MAGS encoding marker genes for the inorganic carbon fixation pathway 3-hydroxypropionate bicycle. Polar maps with the accumulated metatranscriptomic RPKGs of all MAGs encoding at least one of the following key genes: for the 3-Hydroxypropionate bicycle: malyl-CoA/(S)-citramalyl-CoA lyase (K08691) or propionyl-CoA carboxylase (K15052). Absent bars mean that no recruitment was found for that specific metabolism/domain/layer.

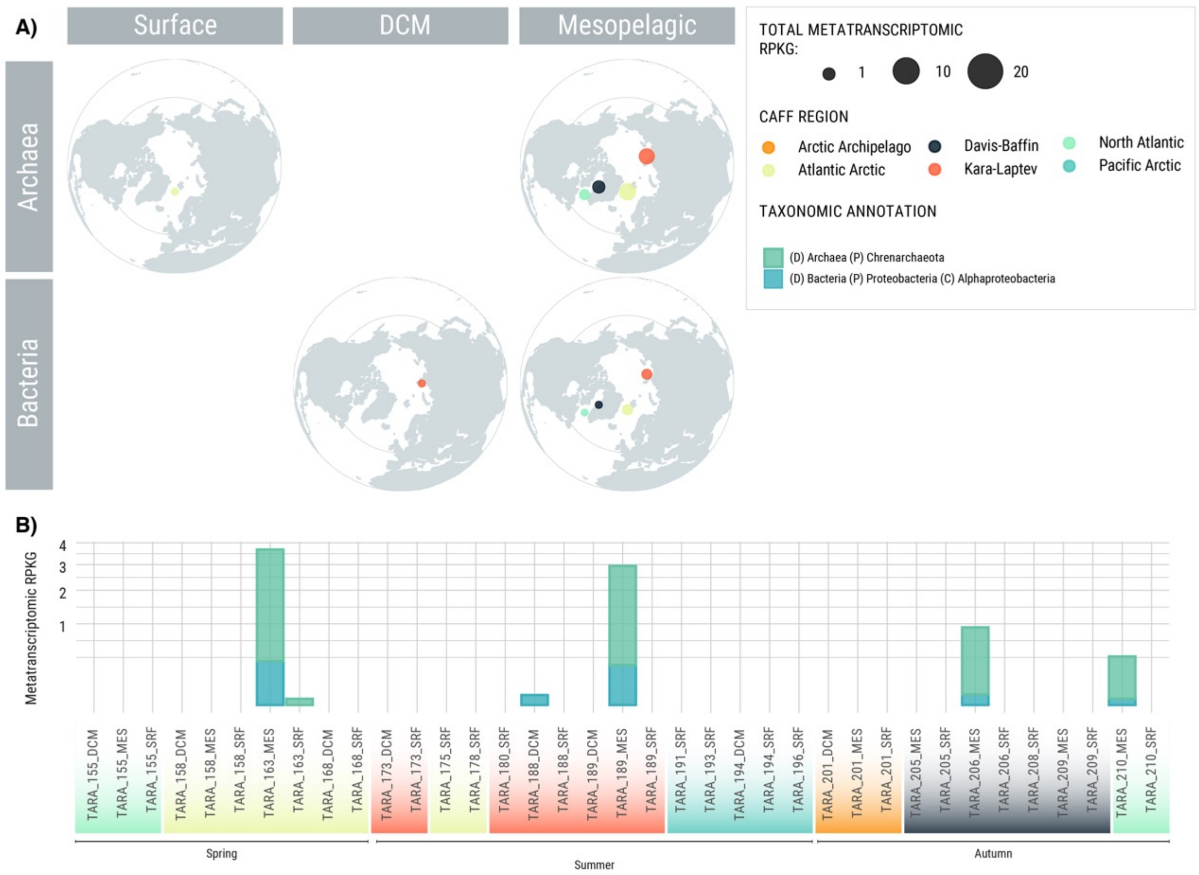


FIGURE S28. Accumulated metatranscriptomic RPKGs of HQ and MQ MAGS encoding marker genes for the inorganic carbon fixation pathway 3-hydroxypropionate/4-hydroxybutyrate cycle. Stacked bars with the accumulated metatranscriptomic RPKGs of all MAGs encoding key enzyme 4-hydroxybutyryl-CoA dehydratase (K14534) together with the complementary gene 3-hydroxypropionyl-coenzyme A dehydratase (K15019). Absent bars/maps mean that no recruitment was found for that specific metabolism/domain/layer. Bars are colored based on taxonomic annotation at the phylum level.

Supplementary tables

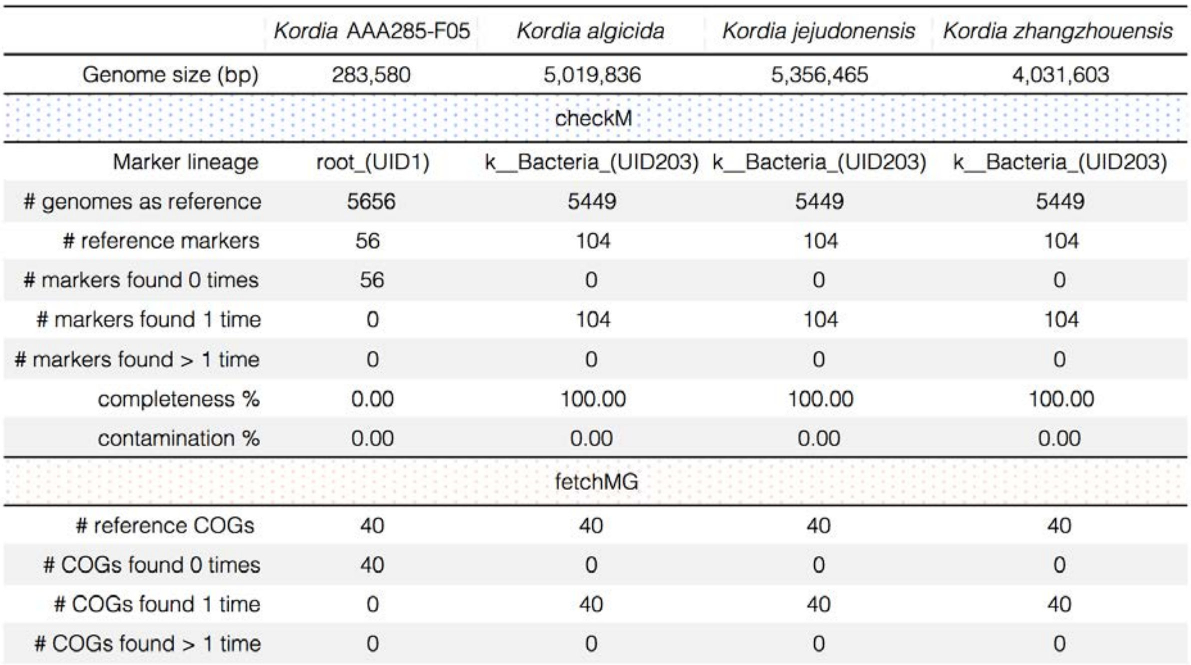
TABLE S1. Samples used in this study.

Expedition	Sample ID	Station	Size Fraction (µm)	Depth (m)	Latitude	Longitude	Ocean	Kind of sample
Tara Oceans	39SUR	TARA_039	All	5	18.59	66.62	Indian	Single Cell Genomics
Tara Oceans	39DCM	TARA_039	0.2-1.6	25	18.57	66.48	Indian	Metagenomics
Tara Oceans	39MES	TARA_039	0.2-1.6	270	18.57	66.48	Indian	Metagenomics
Tara Oceans	39DCM	TARA_039	1.6-20	25	18.57	66.48	Indian	Metagenomics
Tara Oceans	85SUR	TARA_085	0.2-3	5	-62.03	-49.53	Southern	Metagenomics
Tara Oceans	85DCM	TARA_085	0.2-3	90	-62.03	-49.53	Southern	Metagenomics
Tara Oceans	85MES	TARA_085	0.2-3	790	-62.03	-49.53	Southern	Metagenomics
Malaspina	MP0326	20	0.8-20	4,000	-9.12	-30.19	Atlantic	Metagenomics
Malaspina	MP0327	20	0.2-0.8	4,000	-9.12	-30.19	Atlantic	Metagenomics
Malaspina	MP1493	82	0.8-20	2,150	-25.49	-179.52	Pacific	Metagenomics
Malaspina	MP1494	82	0.2-0.8	2,150	-25.49	-179.52	Pacific	Metagenomics

TABLE S2. Primers used in this study.

Primer name	Target gene	Sequence (5’-3’)	Reference
27F	16S rRNA gene	5’-AGRGTTYGATYMTGGCTCAG-3’	Page et al., 2004
907R	16S rRNA gene	5’-CCGTCAATTCMTTTRAGTTT-3’	Lane et al., 1998
358F	16S rRNA gene	5’-CCTACGGGAGGCAGCAG-3’	Muyzer et al., 1993
CF434R	23S rRNA gene	5’-CACTATCGGTCTCTCAGG-3’	Acinas et al. 2014
rpoB-F	Beta Subunit of RNA polymerase (*rpo*B)	5’-ATTCCTTTYAARGGDTCDTGGAT-3’	This study
rpoB-R	Beta Subunit of RNA polymerase (*rpo*B)	5’-CCAATGTTTGGTCCTTCT-3’	This study
PR-Flavo-F	Proteorhodopsin	5’-GAYTAYGTWGSWTTYACDTTYTTTGTRGG-3’	Yoshizawa et al., 2012
PR-Flavo-R	Proteorhodopsin	5’-GCCCAWCCHACWARWACRAACCARCATA-3’	Yoshizawa et al., 2012

TABLE S3. Genome quality estimations of FRA reference genomes using checkM (**?**) and fetchMG (**?**).



	Kordia AAA285-F05	*Kordia algicida*	*Kordia jejudonensis*	*Kordia zhangzhouensis*
Genome size (bp)	283,580	5,019,836	5,356,465	4,031,603
		checkM		
Marker lineage	root_(UID1)	k__Bacteria_(UID203)	k__Bacteria_(UID203)	k__Bacteria_(UID203)
# genomes as reference	5656	5449	5449	5449
# reference markers	56	104	104	104
# markers found 0 times	56	0	0	0
# markers found 1 time	0	104	104	104
# markers found > 1 time	0	0	0	0
completeness %	0.00	100.00	100.00	100.00
contamination %	0.00	0.00	0.00	0.00
		fetchMG		
# reference COGs	40	40	40	40
# COGs found 0 times	40	0	0	0
# COGs found 1 time	0	40	40	40
# COGs found > 1 time	0	0	0	0

Table S4. FRA values of ten different metagenomes from expeditions Tara Oceans and Malaspina 2010 mapped against four reference *Kordia* genomes.

Kordia AAA285-F05

Metagenomic sample	Fraction (*µ*m)	Depth (m)	Sequencing Depth	70-94.9% identity					95-100% identity				
				Number or recruited reads	read/ genome bp	recruitment per Mbp	normalized by subsampled Seq. Depth	% of recruited reads per Mbp	Number or recruited reads	read/ genome bp	recruitment per Mbp	normalized by subsampled Seq. Depth	% of recruited reads per Mbp
TARA_039_DCM	0.2-1.6	25	141787219	197	6,95E-04	6,95E+02	6,26E+01	4,90E-04	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00
TARA_039_DCM	1.6-20	25	176193015	66	2,33E-04	2,33E+02	1,69E+01	1,32E-04	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00
TARA_039_MES	0.2-1.6	270	83476538	13368	4,71E-02	4,71E+04	7,21E+03	5,65E-02	172	6,07E-04	6,07E+02	9,28E+01	7,27E-04
TARA_085_DCM	0.2-3	90	151217436	3327	1,17E-02	1,17E+04	9,91E+02	7,76E-03	12	4,23E-05	4,23E+01	3,57E+00	2,80E-05
TARA_085_MES	0.2-3	790	134267400	145859	5,14E-01	5,14E+05	4,89E+04	3,83E-01	23284	8,21E-02	8,21E+04	7,81E+03	6,12E-02
TARA_085_SUR	0.2-3	5	143758287	600	2,12E-03	2,12E+03	1,88E+02	1,47E-03	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00
MP0326	0.8-20	4000	16469268	1890	6,66E-03	6,66E+03	5,17E+03	4,05E-02	2063	7,27E-03	7,27E+03	5,64E+03	4,42E-02
MP0327	0.2-0.8	4000	23070154	19818	6,99E-02	6,99E+04	3,87E+04	3,03E-01	33385	1,18E-01	1,18E+05	6,52E+04	5,10E-01
MP1493	0.8-20	4000	20947396	1020	3,60E-03	3,60E+03	2,19E+03	1,72E-02	1074	3,79E-03	3,79E+03	2,31E+03	1,81E-02
MP1494	0.2-0.8	4000	12772018	10068	3,55E-02	3,55E+04	3,55E+04	2,78E-01	16785	5,92E-02	5,92E+04	5,92E+04	4,63E-01

Kordia algicida

Metagenomic sample	Fraction (*µ*m)	Depth (m)	Sequencing Depth	70-94.9% identity					95-100% identity				
				Number or recruited reads	read/ genome bp	recruitment per Mbp	normalized by subsampled Seq. Depth	% of recruited reads per Mbp	Number or recruited reads	read/ genome bp	recruitment per Mbp	normalized by subsampled Seq. Depth	% of recruited reads per Mbp
TARA_039_DCM	0.2-1.6	25	141787219	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00
TARA_039_DCM	1.6-20	25	176193015	1179	2,35E-04	2,35E+02	1,70E+01	1,33E-04	6	1,20E-06	1,20E+00	8,66E-02	6,78E-07
TARA_039_MES	0.2-1.6	270	83476538	496	9,88E-05	9,88E+01	1,51E+01	1,18E-04	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00
TARA_085_DCM	0.2-3	90	151217436	35259	7,02E-03	7,02E+03	5,93E+02	4,64E-03	15	2,99E-06	2,99E+00	2,52E-01	1,98E-06
TARA_085_MES	0.2-3	790	134267400	10183	2,03E-03	2,03E+03	1,93E+02	1,51E-03	23	4,58E-06	4,58E+00	4,36E-01	3,41E-06
TARA_085_SUR	0.2-3	5	143758287	54922	1,09E-02	1,09E+04	9,72E+02	7,61E-03	13	2,59E-06	2,59E+00	2,30E-01	1,80E-06
MP0326	0.8-20	4000	16469268	391	7,79E-05	7,79E+01	6,04E+01	4,73E-04	4	7,97E-07	7,97E-01	6,18E-01	4,84E-06
MP0327	0.2-0.8	4000	23070154	775	1,54E-04	1,54E+02	8,55E+01	6,69E-04	7	1,39E-06	1,39E+00	7,72E-01	6,04E-06
MP1493	0.8-20	4000	20947396	290	5,78E-05	5,78E+01	3,52E+01	2,76E-04	1	1,99E-07	1,99E-01	1,21E-01	9,51E-07
MP1494	0.2-0.8	4000	12772018	205	4,08E-05	4,08E+01	4,08E+01	3,20E-04	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00

Kordia jejudonensis

Metagenomic sample	Fraction (*µ*m)	Depth	Sequencing Depth	70-94.9% identity					95-100% identity				
				Number or recruited reads	read/ genome bp	recruitment per Mbp	normalized by subsampled Seq. Depth	% of recruited reads per Mbp	Number or recruited reads	read/ genome bp	recruitment per Mbp	normalized by subsampled Seq. Depth	% of recruited reads per Mbp
TARA_039_DCM	0.2-1.6	25	141787219	1852	3,46E-04	3,46E+02	3,11E+01	2,44E-04	2	3,73E-07	3,73E-01	3,36E-02	2,63E-07
TARA_039_DCM	1.6-20	25	176193015	1072	2,00E-04	2,00E+02	1,45E+01	1,14E-04	3	5,60E-07	5,60E-01	4,06E-02	3,18E-07
TARA_039_MES	0.2-1.6	270	83476538	773	1,44E-04	1,44E+02	2,21E+01	1,73E-04	4	7,47E-07	7,47E-01	1,14E-01	8,95E-07
TARA_085_DCM	0.2-3	90	151217436	40385	7,54E-03	7,54E+03	6,37E+02	4,99E-03	18	3,36E-06	3,36E+00	2,84E-01	2,22E-06
TARA_085_MES	0.2-3	790	134267400	11502	2,15E-03	2,15E+03	2,04E+02	1,60E-03	16	2,99E-06	2,99E+00	2,84E-01	2,22E-06
TARA_085_SUR	0.2-3	5	143758287	66201	1,24E-02	1,24E+04	1,10E+03	8,60E-03	19	3,55E-06	3,55E+00	3,15E-01	2,47E-06
MP0326	0.8-20	4000	16469268	408	7,62E-05	7,62E+01	5,91E+01	4,62E-04	1	1,87E-07	1,87E-01	1,45E-01	1,13E-06
MP0327	0.2-0.8	4000	23070154	841	1,57E-04	1,57E+02	8,69E+01	6,81E-04	5	9,33E-07	9,33E-01	5,17E-01	4,05E-06
MP1493	0.8-20	4000	20947396	297	5,54E-05	5,54E+01	3,38E+01	2,65E-04	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00
MP1494	0.2-0.8	4000	12772018	235	4,39E-05	4,39E+01	4,39E+01	3,44E-04	2	3,73E-07	3,73E-01	3,73E-01	2,92E-06

Kordia zhangzhouensis

Metagenomic sample	Fraction (*µ*m)	Depth	Sequencing Depth	70-94.9% identity					95-100% identity				
				Number or recruited reads	read/ genome bp	recruitment per Mbp	normalized by subsampled Seq. Depth	% of recruited reads per Mbp	Number or recruited reads	read/ genome bp	recruitment per Mbp	normalized by subsampled Seq. Depth	% of recruited reads per Mbp
TARA_039_DCM	0.2-1.6	25	141787219	2516	6,24E-04	6,24E+02	5,62E+01	4,40E-04	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00
TARA_039_DCM	1.6-20	25	176193015	1275	3,16E-04	3,16E+02	2,29E+01	1,79E-04	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00
TARA_039_MES	0.2-1.6	270	83476538	631	1,57E-04	1,57E+02	2,39E+01	1,87E-04	3	7,44E-07	7,44E-01	1,14E-01	8,91E-07
TARA_085_DCM	0.2-3	90	151217436	33545	8,32E-03	8,32E+03	7,03E+02	5,50E-03	0	0,00E+00	0,00E+00	0,00E+00	0,00E+00
TARA_085_MES	0.2-3	790	134267400	8445	2,09E-03	2,09E+03	1,99E+02	1,56E-03	1	2,48E-07	2,48E-01	2,36E-02	1,85E-07
TARA_085_SUR	0.2-3	5	143758287	50153	1,24E-02	1,24E+04	1,11E+03	8,65E-03	1	2,48E-07	2,48E-01	2,20E-02	1,73E-07
MP0326	0.8-20	4000	16469268	431	1,07E-04	1,07E+02	8,29E+01	6,49E-04	1	2,48E-07	2,48E-01	1,92E-01	1,51E-06
MP0327	0.2-0.8	4000	23070154	869	2,16E-04	2,16E+02	1,19E+02	9,34E-04	8	1,98E-06	1,98E+00	1,10E+00	8,60E-06
MP1493	0.8-20	4000	20947396	261	6,47E-05	6,47E+01	3,95E+01	3,09E-04	2	4,96E-07	4,96E-01	3,02E-01	2,37E-06
MP1494	0.2-0.8	4000	12772018	215	5,33E-05	5,33E+01	5,33E+01	4,18E-04	1	2,48E-07	2,48E-01	2,48E-01	1,94E-06

The following tables belong to **Chapter 2** and **Chapter 3** and due to their size they cannot be displayed. You can download them clicking here[1].

TABLE S5. Metadata including accession numbers of all samples used in this study.

TABLE S6. Metrics and evaluation of all co-assemblies using different k-mer length values in the co-assembly optimization step.

TABLE S7. Genome completeness and contamination estimation of all the possible co-assembly combinations, from individual assemblies to a single combination of all 10 SAGs.

TABLE S8. Whole-genome identities between the co-assembly and their closest *Flavobacteriaceae*. NA values correspond to ANIs <75%.

TABLE S9. Functional annotation of *Kordia* sp. TARA_039_SRF co-assembly

TABLE S10. Summary of number and class of CAZymes encoded in *Kordia* sp. TARA_039_SRF and relatives within the genus *Kordia*.

TABLE S11. Presence in *Kordia* sp. TARA_039_SRF of adhesion proteins found in other Bacteroidetes.

TABLE S12. Summary of number and class of transporters encoded in *Kordia* sp. TARA_039_SRF and relatives within the genus *Kordia*.

TABLE S13. Predicted secretory system related genes in *Kordia* sp. TARA_039_SRF.

TABLE S14. Genes whose peptides contain the Por secretion system domain TIGR04183, hence they are secreted to the extracellular environment or cell surface through the secretion system T9SS. PFAM summary description is based on the known functions of PFAMs extracted from PFAM website.

TABLE S15. Summary of number and type of peptidases encoded in *Kordia* sp. TARA_039_SRF and relatives within the genus *Kordia*. Asterisk next to copy number means that some (or all) of these proteins encode the Por domain as well (TIGR04183), meaning that they are secreted to the extracellular space.

[1] http://tarod.cmima.csic.es/tmp/data/marbits/acinasLab/thesis-royo-llonch/

TABLE S16. Predicted genomic islands in the co-assembly with encoded transporters, CAZymes and other interested CDS.

TABLE S17. Insertion Sequences (IS) predicted in the co-assembly.

TABLE S18. Predicted prophages in the co-assembly.

TABLE S19. Fragment Recruitment analysis results.

TABLE S20. Number of genes where the co-assembled *Kordia* sp. TARA_039_SRF shows mapped reads in Fragment Recruitment Analysis.

TABLE S21. Number of genes classified in a COG category, CAZyme type and transporter family for the four *Kordia* genomes used in the study. First considering the whole genome, then the genes belonging to the core genome, those unique strain-specific genes in their flexible genomes and finally the genes that are shared between three or more genomes (shared accessory genome).

TABLE S22. Clusters of samples for co-assembly, based on miTag community composition similarities.

TABLE S23. Phylogenomic taxonomic annotation, assembly metrics and biogeographic information of the complete bin dataset. Niche breadth categories for HQ and MQ MAGs.

TABLE S24. Taxonomic annotation of 16S rRNA coding bins MAGs against SILVA database.

TABLE S25. Environmental metadata of the metagenomes used in this study.

TABLE S26. Metagenomic recruitments normalized as RPKG of all 3550 in the 68 metagenomes of the study.

TABLE S27. Metatranscriptomic recruitments normalized as RPKG of all 3550 in the 68 metagenomes of the study.

TABLE S28. Metabolic marker genes of selected metabolisms of interest.

BARCELONA
FEBRUARY 2020

ICM Institut de Ciències del Mar

UAB Universitat Autònoma de Barcelona