Seasonality of marine prokaryotes using taxonomic and functional diversity approaches

Adrià Auladell Martín

Seasonality of marine prokaryotes using taxonomic and functional diversity approaches

Adrià Auladell Martín

Institut de Ciències del Mar

Tesi doctoral

Programa de doctorat en Microbiologia

Universitat Autònoma de Barcelona
Departament de Genètica i de Microbiologia

Desembre 2021

Seasonality of marine prokaryotes using taxonomic and functional diversity approaches

Estacionalitat de procariotes marins usant mètodes d'anàlisis taxonòmics i funcionals

Estacionalidad de procariotas marinos usando métodos de análisis taxonómicos y funcionales

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

Directors:

Dr. Josep M Gasol
Departament de Biologia Marina i Oceanografia,
Institut de Ciències del Mar (ICM-CSIC)

Dra. Isabel Ferrera Ceada
Centro Oceanográfico de Málaga,
Instituto Español de Oceanografía (IEO-CSIC)

Tutora acadèmica:

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

Barcelona, December 17th 2021

During the realization of this thesis, Adrià Auladell Martín held a PhD Fellowship FPI (BES-2016-077820), funded by the Spanish Ministry of Science and Innovation and was a student of UAB's PhD program in Microbiology. This thesis was also supported by the REMEI (CTM2015-70340-R), ECLIPSE (PID2019-110128RB-I00), and MIAU (RTI2018-101025-B-I00) projects.

Cover image by Sofia Crespo (https://sofiacrespo.com/)
through previous work from Manolo ide (http://manoloide.com/).
Cover design and thesis editing by Gemma Fontanals @gemfontanals.

A la meva mare
Al meu pare
Al meu germà

Us estimo

“There isn’t a way things should be.
There’s just what happens, and what we do.”
— Terry Pratchett, A Hat Full of Sky

Acknowledgments

Bueno, ha arribat el moment d'escriure això. L'agraïmenta és la part que més serà llegida de la tesi i a la que menys temps li dediquem. Com ha de ser, suposo, perquè si t'hi poses de debò et podria sortir una tesi sencera només amb els agraïments. Almenys en aquest cas tot el que faig és gràcies a l'ajuda dels meus. Així doncs, persona desconeguda que està llegint això, o estimat conegut passant el temps, o enemic (conegut o no) volent descobrir qui estimo per fer-me mal, aquí van els meus reconeixements a la gent que em fa gaudir i crear a la vida.

La tradició dicta començar pels que han hagut de suportar-me més: la dolenta escriptura, les idees científiques imberbes i, en general, els que han aconseguit fer de mi un bon científic. A la porta del laboratori de molecular hi ha una foto de la Vanessa amb un cartell on hi diu 'Super jefa!'. Sempre que el veia pensava, amb una mentalitat potser massa competitiva, 'pa super jefa la mia!' i que algun dia també et celebraria igual. Isabel, has sigut tot el que un podria desitjar tenir com a coordinadora. La teva paciència, la teva empatia, el teu respecte, la teva capacitat d'anar al detall, el teu enginy i el teu gust estètic m'han servit de referent tot aquest temps. Amb tu he entès què vol dir ser un bon professional i algú capaç de preocupar-se dels seus. M'he sentit recolzat sempre fent ciència, fins i tot quan discutíem i em posava tossut, cosa que sé que no era sempre fàcil, perquè de tossut ho puc ser bastant. Soc molt feliç d'haver fet el doctorat amb tu i que durant el camí haguem acabat també sent grans amics. I per molts anys més! Pep, no te'm posis gelós, eh? Encara recordo el primer dia que vaig entrar a la teva oficina tot petit i cagat pensant que em diries 'nanai noi, cap a casa'. Recordo també veure el piolet darrere la porta i que d'alguna manera això em relaxés. Moltes gràcies per la tranquil·litat quan la necessitava, per ajudar-me a encarar-me cap a on tocava quan em perdia per matolls científics sense gaire interès i per donar-me mirada elevada quan m'encaparrava en tonteries. Moltes gràcies també per no posar pegues en cap moment a l'hora de formar-me amb cursos, i donar-me llibertat absoluta en les decisions sobre aquests temes. En resum, que he après molt de tu, tant de la ciència com de la vida. Crec que els tres hem fet un equip molt bo i hem pogut fer una molt bona feina. Espero que ho sentiu igual!

També he d'agrair moltíssima gent de l'ICM amb els quals he compartit la feina i la diversió. Olga, a tu t'incloc aquí perquè ets part del centre, vulguis o no! Moltes gràcies per accedir a ser tutora meva, i per ajudar-me amb els dubtes sempre amb diligència i bon rotllo. Moltes gràcies Ramon per les discussions de passadís, de despatx quan venies a buscar l'Aleix i en general per l'escolta atenta quan tenia dubtes. Ramiro, mercès per discutir sobre bioinfo i aportar sempre als diferents estudis. I molts altres PIs del centre: Cèlia, Rafel, Esther, Silvia... gràcies per l'ajuda i bona companyia. Dels postdocs, sempre recordaré l'*horror vacui* que em vau generar un dinar al campito explicant les vostres situacions precàries. Sort que sou tant bona gent que tot això ràpidament s'oblida entre birres i riures. Marta S., moltes gràcies pel bon humor, les discussions científiques i els moments d'Slack. Pablo, gràcies per ensenyar-me i tenir paciència amb les meves típiques rebentades del Marbits amb mmseqs2. Javier, gràcies per escoltar-me i validar l'odi que tinc cap a coses com els culpables del canvi climàtic o simplement els imbècils que hi ha pel món: catàrtic en els temps que corren! Clara R., gràcies pel bon rotllo i accedir a ser suplent, aprenc molt de tu! Massimo, ets la vitalitat en persona i ho encomanes. I mooolta altra gent: Fran, Andrea, Albert... Gràcies per tot! Clara C., aquesta tesi no hauria estat possible sense tu. Moltes gràcies de tot cor pel mostreig constant a Blanes, he gaudit molt anant a mostrejar amb tu i xerrant de tot i de res. Gràcies també Vanessa per les moltes converses durant la tesi. També he d'agrair-te aquest últim any a tu, Maria Yubero. Quan vam fer match de Tinder per part del programa camino del CSIC vaig aconseguir l'excusa perfecte per conèixer-nos i aprendre de tu milers de detalls i intringulis de la vida professional. Moltes gràcies. Algú altre que m'emporto com a referent ets tu Elena Torrecilla! Gràcies per la calma i per tenir clares tantes coses.

Ja passem a la meva tribu, els malamentes, els mala gent, els que cridem als despatxos, que ens celebrem i queixem constantment, la fanfàrria, la palmera, el campito i el cafelico. Som molts, gent. Si no t'hi trobes aquí i creus que hi hauries de ser, vine a queixar-te i t'estimo en persona. Aleix, sento que ni he de dir res, que no cal, així de connectats estem. Ens ho hem passat pipa, i això anirà a més. La depressió post tesi serà no tenir-te porai i poder riure i compartir tantes coses. T'estimo brother. Ets segurament el millor que m'emporto de tot el puto doctorat. Carlota, compi de grup, la quiero un montón. Recordo quan temps enllà l'Andreu em va dir que coneixia algú interessat en l'oferta de feina de l'Isabel, i jo em vaig sentir compromès perquè no volia influir en la decisió. A dia d'avui hauria insistit i molt perquè fossis tu. Queralt, la meva versió femenina, mi vida paral·lela, soc molt feliç que hagis acabat a l'ICM per compartir moments amb tu. Marta R., ets una referent, moltes gràcies per accedir a ser amiga meva. Andrew, hot guy, hot topic, hot Gili, hot hot hot, gràcies per l'humor fi, la poètica i posar-te trist si marxem de la birra. Vals molt. Patri, merci per cuidar-me i ser el centre de la vida malamente, que tira endavant la diversió i que ens ha donat la festa més preciosa de la tesi. Aurelie, thank you for your craziness, I have missed you a lot these two last years. Deju, el meu far de bogeria i diversió, de discussions i de cançons, en breus anirem al bosc com diu aquell. Miguel, aprofito aquest espai per declarar que estic enamorat de tu. Ara ja ho saps. Guillem, a tu només et menjo la boca, dirty sex with you i amb en Miguel em caso. Ari, quan vens a dir bon dia és amb diferència el millor moment del matí. Equipo A per sempre. Janire, gràcies per ser consellera i per les converses profundes inacabables de whatsapp, t'admiro molt. Joan, ànima festiva que corre pel món, convida'm al teu pis malparit, que m'encantes i vull fer més plans amb tu. Sara, des que has arribat a l'ICM t'has convertit en indispensable i molt important per mi. Espero que et quedis molt més temps, la meva petita forta borratxa. Maria, torna, torna, torna a casa. O no, com et vagi bé, però veiem-nos i gaudim. Gràcies pels moments de b12a i els moments fora. I molta altra gent, però ja he de parar: Ari, Elena, Ana Soto, Ana Trini, Anna Arias, Isa, Claudia, Dani, Marina, Manu... Us estimo molt! Veniu i ja us diré coses maques.

A l'altra banda de Barcelona vaig trobar una altra petita tribu d'amazones doctorals de les quals vaig aprendre i gaudir molt. Alícia, que estem acabant, que ja està aquest percal! A prendre pel cul i a gaudir del final. Ja ho celebrarem com cal quan ens veiem. La trobo a faltar molt. Berta, mi amore, mi sindicalista, l'empenta, el somriure sempre. Siusplau, tirem la coope endavant, morim-nos de gana però feliços i contents. Gemma, la calma davant un aparador en flames ple de gent robant. T'aprecio molt. Ja està això noies!

I encara més lluny, a les amèriques del nord, més gent que m'ha ajudat molt. Albert, gràcies per accedir que vingués a fer l'estada, llàstima que la pandèmia dels nassos l'esguerrés pel mig. Mery, recordo molt un dinar al mexicà post escalada, en què vaig adonar-me com de bé em queies. Moltes gràcies per tot, sense tu hauria estat molt miserable aquells mesos. Thanks a ton for everything Gabri and Hannah!

Bueno gent, tots els de la feina fora, ja ho tenim això.

Andreu i Adri, Adri i Adri i Adri, tots som un i un som tots. Us estimo molt, tant como la trucha al trucho, sou el voler ballar el voler cridar i el discutir acaloradament fins esbandir res més. Per molts anys més. Marina, ets la meva calma, ets casa, el teu riure cuinant em treu qualsevol cansament, moltes gràcies per tot. Ainaken, quina sort vam tenir de trobar-te. Portes una alegria molt especial al pis, t'has convertit en indispensable. I els companys del passat. Santa Perpètua! Encara ploro la casa. Sempre la ploraré suposo. Carlos, malparit, especulador, ens vas treure el Shangri-la. Tània, en realitat tu eres el Shangri-la, la casa sense tu no hauria estat mai la festa i la xerinol·la que era. T'estimo, ets la germana que mai he tingut. Eudaldio, tu ets el germà que mai he tingut. Perquè ets com super especial i molt diferent de'n Miqui. Lilas, Guxi, Arnau, Javier, Alba, vam viure molt bons anys junts. L'altra ideòloga d'aquests anys tant bonics vas ser tu Vicky. Grècia, Eivissa, Roma, viatjar fins Galícia, i els milers de moments a Barcelona. Gràcies per tot.

També de vital importància la bioinfointifada! Companys de màster, de jugar al cuquet mentre en Josep ens explicava Perl. Joan i Laura, encara no he superat que marxéssiu lluny. Hauríem pogut fer tant juntets. Però bueno, tenint en compte que sobretot ho vau fer per la merda de sou que ens donen aquí, un es relaxa, entén, i sap que vau fer bé. Moltes gràcies pels un-seminars, els viatges i el gaudi. I molts altres bioinfos importants. Alberto, Sergio, Leo, Pablo, Felipe, trobo a faltar la vostra bogeria.

Suposo que no tenir unes paraules pels Micros seria lleig. Això és llarguíssim, joder. Nanos i nanes, però sobretot nanos, que sou amb qui més he compartit, us aprecio molt!

I la roca també s'ha d'agrair. La roca allibera i treu la pols. I als dos que més he d'agrair la roca sou vosaltres, Roc i Andrés. Gràcies per treure'm a passejar i a gaudir! Mariona, per mi tu també ets roca, roco, i aquestes últimes setmanes has estat casa. Gràcies per cuidar-me.

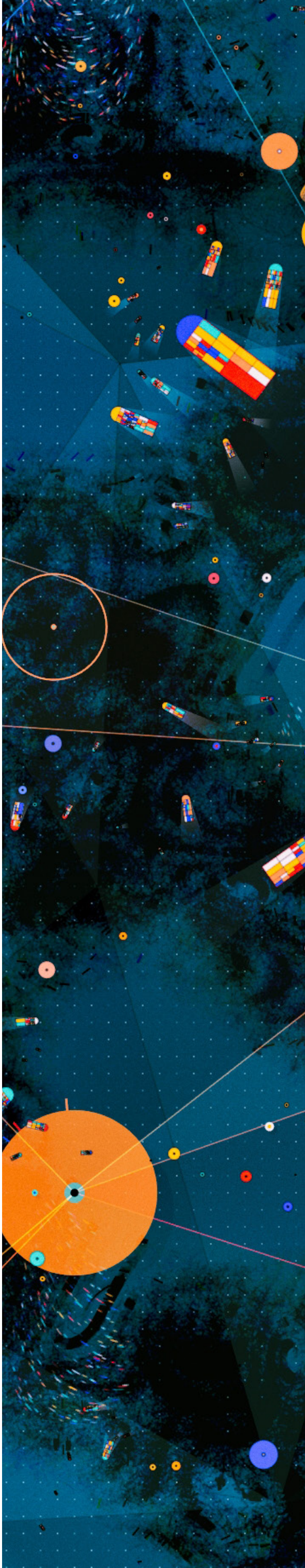
Una altra família que m'ha acompanyat aquests anys són els peixos. Jordi, tu ets peix, ets sanfeta i ets tota la vida. David, cap a Polònia que vinc amb la por que no hi siguis quan arribi. Gemma, tenir-te al costat aquesta última setmana és quelcom que no oblidaré mai. Gràcies infinites per la feinada.

I ja hem arribat a casa. El final d'aquest viatge pesat però maco ple de gent preciosa. Sant Feliu, amb les Darriba, amb les Olivé, amb la Diana, la Anna, la Joana, la Mar. A vosaltres ni us deixo paraules boniques, simplement us anomeno, perquè sou tant de casa que ni us cal. Gràcies per tot.

I Marta. Que no vol agraïmenta, que se li fa rar. Que ella el que vol és l'Adri bé, i el terra ficat, Llançà i la seva costa. Però no, està aquí, llegint, i li fa gràcia aquesta metaescriptura, menys gràcia que la que li fa a l'Adri, però gràcia. Que tot segueixi així, perpètuament, imperfectament, feliçment, tontament. Que el vaivé ens vagi a favor, que trobem pa a Leningrad i que la gana no ens impedeixi gaudir. No em desagrades, com a resum.

Contents

SUMMARY/ RESUM



Summary

The oceans are ecosystems dominated by microbes, in which bacteria and archaea play key roles in biogeochemical cycling. In temperate oceans, seasonal changes in environmental conditions deeply influence the marine microbiome. In this thesis I analyzed the seasonality of the marine microbiome in a coastal ocean site, using the long-term time series of the Blanes Bay Microbial Observatory (BBMO) to understand the seasonal changes through several molecular approaches. Using amplicons of the 16S rRNA gene, I evaluated the dynamics of the main bacterial groups in this coastal oligotrophic station during 11 years and tested how similar the temporal niches of closely related taxa are, and what are the environmental parameters modulating their patterns of seasonality. I further explored how conserved the niche is at higher taxonomic levels. The community presented recurrent patterns of seasonality for 297 out of 6825 amplicon sequence variants (ASVs), which constituted almost half of the total relative abundance (47%). For certain genera, niche similarity decreased as nucleotide divergence in the 16S rRNA gene increased, a pattern compatible with the selection of similar taxa through environmental filtering. Additionally, I observed evidence of seasonal differentiation within various genera as seen by the distinct seasonal patterns of closely related taxa. I then switched the focus to the seasonal patterns of a specific functional group. Using the *pufM* gene as a marker gene for the aerobic anoxygenic phototrophic bacteria (AAPs) –a relevant photoheterotrophic functional group in the marine microbial food web– I evaluated their long-term temporal dynamics through multivariate and co-occurrence analyses. Phylogroup K (Gammaproteobacteria) was the greatest contributor to community structure over all seasons, with phylogroups E and G (Alphaproteobacteria) being prevalent in spring. The diversity indices showed a clear seasonal trend, with maximum values in winter, which was inverse to that of AAP abundance. I later extended these analyses to 21 biogeochemical relevant functions through 7 years of metagenomic data from the BBMO. Most genes presented a seasonal abundance trend: photoheterotrophic processes were enriched during spring, phosphorous-related genes were dominant during summer coinciding with phosphate limitation conditions, and assimilatory nitrate reductases correlated negatively with nitrate availability. Additionally, I identified the main taxa driving each function in each season and showed that, for some groups, the seasonality of bacterial families is different than that of their gene repertoire, so that different taxa within the same group present different functional specialization. Finally, I complemented this descriptive view of the temporal changes with manipulation experiments to test how bottom-up and top-down factors exert selection on specific bacterial genomic species over the seasons. I experimentally modified the presence of predators, viruses, nutrient limitation (by diluting the samples with filtered seawater) and light availability in seawater from the BBMO in different seasons and assessed the growth of different organisms defined by metagenome assembled genomes (MAGs) under the manipulated conditions. Overall, I recovered 262 MAGs mainly from the Rhodobacterales, Flavobacteriales and Alteromonadales classes. Season and treatment greatly influenced community composition, with

26% of the MAGs indicative of the control treatments, 24% of both the control and predator-reduced treatments, 12.8% indicators of both the virus-reduced and the diluted treatments, and 7.3% of the predator-reduced treatment only. *Flavobacteriaceae* MAGs developed mostly in the predator-reduced treatment with distinct species at each season, whereas *Alteromonadaceae* and *Sphingomonadaceae* taxa developed preferably in the virus-reduced and diluted treatments indistinctively of season. Overall, this dissertation provides new insights into the seasonal patterns of key taxonomic and functional groups from the coastal surface ocean through the integration of information obtained using several molecular techniques and approaches applied to a long-term time series.

Resum

Els oceans són ecosistemes dominats per microbis, i els bacteris i els arqueus hi juguen papers clau en els cicles biogeoquímics. En oceans temperats, els canvis estacionals determinen la composició del microbioma a través de les adaptacions de nínxol de les diferents espècies. En aquesta tesi he analitzat l'estacionalitat del microbioma marí usant una sèrie temporal de llarga durada obtinguda a l'Observatori Microbià de la Badia de Blanes per entendre els canvis estacionals mitjançant diverses aproximacions moleculars. A partir de seqüències d'amplicons del gen de l'RNA ribosòmic (16S) he avaluat la dinàmica estacional dels principals grups bacterians durant onze anys, examinant com són de similars els nínxols temporals de taxons relacionats estretament, i quins són els paràmetres que modulen els seus patrons d'estacionalitat. També he explorat com de conservat és aquest nínxol en els nivells taxonòmics més alts. La comunitat presenta patrons estacionals de recurrència en 297 de les 6725 variants d'amplicons que apareixen, la qual cosa suposa gairebé la meitat de l'abundància relativa total (47%) de seqüències. Per a determinats gèneres, la similitud de nínxol disminueix amb l'increment de divergència en nucleòtids del gen del 16S rRNA, un patró compatible amb selecció de taxons similars per mitjà del filtratge ambiental. També he observat diferents patrons estacionals entre taxons del mateix gènere. A continuació vaig centrar l'anàlisi en els patrons estacionals d'un grup funcional concret. Utilitzant el gen *pufM* com a marcador dels bacteris aeròbics anoxigènics fotoheterotròfics –un grup funcional rellevant a la xarxa tròfica marina– he avaluat les seves dinàmiques temporals a través d'anàlisis multivariants i de co-ocurrència. El filogrup K (Gammaproteobacteria) és el grup dominant a l'estructura de la comunitat durant totes les estacions de l'any, amb els filogrups E i G (Alphaproteobacteria) dominants durant la primavera. Els índexs de diversitat presenten un patró estacional clar, amb els valors màxims durant l'hivern i presentant una relació inversa amb l'abundància. Després vam ampliar aquest anàlisi a 21 funcions biogeoquímiques fent ús de set anys de dades metagenòmiques de l'observatori de Blanes. La majoria dels gens presenten un patró estacional d'abundància: els processos fotoheterotròfics enriquits durant la primavera, els gens relacionat amb l'adquisició de fòsfor dominant durant l'estiu coincidint amb una major limitació de fòsfor, i els enzims de reducció assimilatòria de nitrat correlacionant negativament amb la disponibilitat de nitrat. També he identificat els taxons principals que contenen cada gen funcional, i he demostrat que, per alguns grups, l'estacionalitat a nivell de família és diferent de la del seu repertori gènic, indicant que els taxons dins del mateix grup presenten especialització funcional. Finalment, he complementat la visió descriptiva dels canvis temporals amb experiments de manipulació per avaluar com els processos *bottom-up* i *top-down* influencien la selecció d'organismes durant les diferents estacions. He modificat experimentalment la presència de depredadors, de virus, la limitació per nutrients (diluint les mostres amb aigua sense microorganismes) i la llum en mostres de la Badia de Blanes en diferents estacions i he avaluat el creixement de diferents organismes definits a partir de genomes construïts a partir de metagenomes (MAGs, de les sigles en anglès). Vaig recuperar 262 MAGs, principalment de les classes Rhodobacterales,

Flavobacteriales i Alteromonadales. L'estació de l'any i el tractament influeixen la composició de la comunitat, amb el 26% dels MAGs identificats com a indicadors dels tractaments control, el 24% indicant tant el tractament control com el de reducció de depredadors, el 12.8% indicant tant el tractament de reducció de virus com el tractament diluït, i el 7.3% indicant el tractament de reducció de depredadors. Els MAGs afiliats a *Flavobacteriaceae* creixien majoritàriament al tractament amb reducció de depredadors, amb diferents espècies a cada estació, mentre que les especies afiliades a *Alteromonadaceae* i *Sphingomonadaceae* creixien preferentment als tractaments de reducció vírica i diluït indistintament de l'estació. En termes generals, aquesta tesi presenta nous resultats sobre els patrons estacionals de grups taxonòmics i funcionals rellevants a l'oceà costaner superficial per mitjà de la integració d'informació obtinguda usant diverses tècniques moleculars i diverses aproximacions experimentals aplicades a sèries temporals de llarga durada.

INTRODUCTION

Introduction

Marine microbes

The marine environment is the largest ecosystem on Earth. Perhaps ironically, it is dominated by the tiniest forms of life, hidden to humans for thousands of years. Microbes contribute about two thirds to the total biomass of marine organisms (Bar-On and Milo, 2019). Bacteria and archaea specifically represent around 10^{29} cells, making up to 27% of the total marine biomass (Whitman *et al.*, 1998; Bar-On and Milo, 2019). Half of the total planetary primary production occurs in the ocean (Field *et al.*, 1998), 90% of which is performed by microorganisms (Duarte and Cebrián, 1996). These small producers, representing 0.66 gigatons of carbon, sustain most of the marine trophic system through a fast turnover time, on a timescale of days (Kirchman, 2016; Bar-On and Milo, 2019). Microbes are also responsible for most of the respiration occurring in the seas (del Giorgio and Duarte, 2002). Most of the organic matter produced by phytoplankton is consumed by bacterioplankton, channeling these compounds up the food chain in a process known as "microbial loop" (Azam *et al.*, 1983). The many discoveries that microbial ecologists and oceanographers have come across during the last decades have shown that bacteria and archaea present the most diverse metabolic repertoire of the ocean, driving the biogeochemical cycles and channeling matter and energy on a planetary scale (Falkowski *et al.*, 2008; Kirchman, 2008).

The sunlit ocean as a microbial ecosystem

The ocean is vast, spanning 3.6 x 10^{8} km^{2} and containing 1.4 x 10^{21} liters of water, 97% of the total water on Earth (Eakins and Sharman, 2010). Its vastness is not homogeneous but rather contains a myriad of different ecosystems in perpetual change. From all this environmental heterogeneity, the sunlit marine waters are one of the most studied ecosystems, both for logistic reasons and importance to humans. This region comprises the first 200 meters of the water column (the so-called epipelagic), in which light allows the growth of phytoplankton and other primary producers, particularly in the upper zone. In these waters, each microbial type is challenged by physical, chemical and biological conditions that altogether determine the global community structure. Coastal waters in particular are subjected to the variability in continental, atmospheric and oceanic forcing, since these areas are the boundary between these three environments. Eddies, water mass stratification and tides have an impact on microbe dispersion on a regional scale (Hanson *et al.*, 2012). The abiotic environment is moreover determined by latitudinal and seasonal changes in most of their relevant features. Temperature, oxygen, nutrient availability and salinity, among others, play important roles in microbial community assembly (Hewson *et al.*, 2006). The increase in solar irradiance with decreasing latitude and seasonal cycles rises water temperature and promotes stratification, which in turn impacts community assembly (Sunagawa *et al.*, 2015; Logares *et al.*, 2020). Inorganic nutrient availability is dictated mainly by physical forcing such as the mixing of deeper waters (Margalef, 1978). At the coast, however, episodic disturbances such as storms and precipitations are also similarly important (Duarte *et al.*, 1999). Rainfall and wind storms cause an input of allochthonous material, and heavy rainfall events result sometimes in land runoff, introducing both inorganic and organic particulate and dissolved nutrients (Guadayol *et al.*, 2009).

At a microscopic scale, biotic interactions between microbes gain importance and modulate bacterial physiology and growth (Lima-Mendez *et al.*, 2015; Worden *et al.*, 2015). Through the direct release of dissolved organic matter through exudates, phytoplankton interact with heterotrophic bacteria that remineralize it (Seymour *et al.*, 2017). Inversely, the micronutrient production by heterotrophic bacteria facilitates the growth of several phytoplanktonic groups (Johnson *et al.*, 2020). Moreover, the negative biotic interactions range from the grazing by protists, viral infections, competition for nutrients, to even antibiotic warfare between bacteria (Sherr and Sherr, 2002; Kirchman, 2008; Sánchez *et al.*, 2020; Niehus *et al.*, 2021). Altogether, these factors determine how communities assemble in the marine microbial ecosystem.

The field of microbial ecology: the last 50 years

Since the discovery that bacteria could play an important role in marine nutrient cycling (Pomeroy, 1974), the microbial ecology field has been pushed by technologies that have facilitated the exploration of these ‘invisible’ living forms (Figure 1). The initial approaches to study marine microbial diversity relied on culture isolation from environmental samples. These efforts found however that only a small portion of the bacterial and archaeal community was able to grow forming colonies on agar media compared to the total cell numbers seen by microscopy. This discrepancy was referred to as the “Great Plate Count Anomaly” theory (Staley and Konopka, 1985). The ability to grow forming colonies depends on multiple factors, and the traditional plate count approach generally mimics the eutrophic conditions rarely found in the marine ecosystem and beneficial to some specific microbial groups only. The most common paradigm is that only ~1% of the organisms in nature are culturable (Amann *et al.*, 1995; Torsvik and Øvreås, 2002) and the rest forms the so-called ‘uncultured majority’. The adoption of technologies from biomedicine, such as flow cytometry, opened new endeavors. For example, it allowed the discovery of *Prochlorococcus*, the most abundant primary producer in the ocean (Chisholm *et al.*, 1988), and fueled studies combining multiple stains which probed cellular activity and growth (del Giorgio and Gasol, 2008).

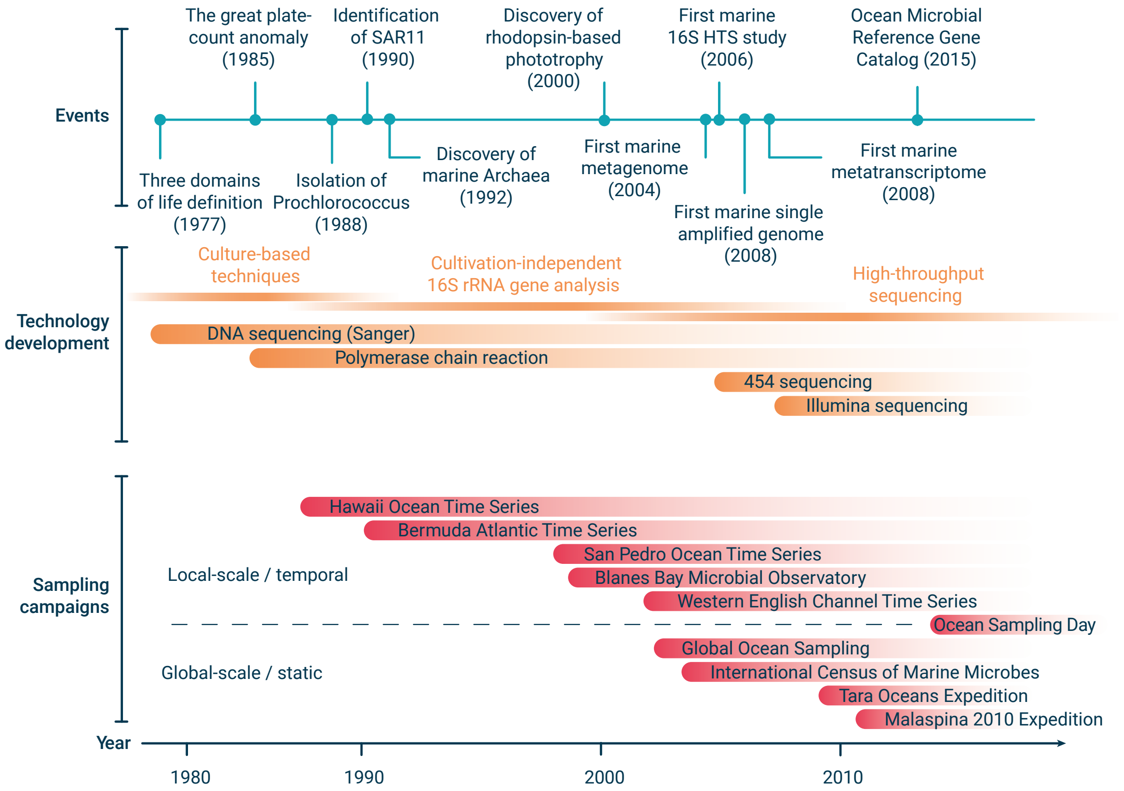

Events
The great plate-count anomaly (1985)
Identification of SAR11 (1990)
Discovery of rhodopsin-based phototrophy (2000)
First marine 16S HTS study (2006)
Ocean Microbial Reference Gene Catalog (2015)
Three domains of life definition (1977)
Isolation of Prochlorococcus (1988)
Discovery of marine Archaea (1992)
First marine metagenome (2004)
First marine single amplified genome (2008)
First marine metatranscriptome (2008)
Technology development
Culture-based techniques
Cultivation-independent 16S rRNA gene analysis
High-throughput sequencing
DNA sequencing (Sanger)
Polymerase chain reaction
454 sequencing
Illumina sequencing
Sampling campaigns
Local-scale / temporal
Hawaii Ocean Time Series
Bermuda Atlantic Time Series
San Pedro Ocean Time Series
Blanes Bay Microbial Observatory
Western English Channel Time Series
Ocean Sampling Day
Global-scale / static
Global Ocean Sampling
International Census of Marine Microbes
Tara Oceans Expedition
Malaspina 2010 Expedition
Year
1980
1990
2000
2010

Figure 1: A selection of the main scientific events in the history of marine microbial ecology are represented through time in the upper panel. The major technological developments related to DNA sequencing are represented in the middle panel. Large-scale sampling campaigns, both in time and space, are indicated in the lower panel. Modified from Salazar *et al.* (2015).

Likely the most important development to access the 'uncultured majority' originated from the application of molecular tools, profoundly innovating the field of marine microbial ecology. For bacterioplankton, the key development was the PCR amplification and sequencing of the 16S rRNA gene of marine samples. That gene was found to be a marker gene reflective of the evolutionary history of most organisms (Woese and Fox, 1977). The first approaches relied on cloning the nearly full length 16S rRNA gene and their subsequent sequencing using the Sanger method (Pace *et al.*, 1986). This technique revealed the presence of multiple uncultured species in marine samples for both bacteria and archaea (Giovannoni *et al.*, 1990; Ward *et al.*, 1990; DeLong, 1992; Acinas *et al.*, 1999; Massana *et al.*, 2000). Additionally, the use of fluorescent in situ hybridization (FISH) with oligonucleotide probes targeting rRNA allowed the visualization and enumeration of specific phylogenetic groups in natural samples (DeLong *et al.*, 1989). Nevertheless, researchers were soon aware of the multiple limitations of these techniques. The increase in 16S rRNA gene data showed that some taxonomically broad FISH probes were not able to match all the relevant groups (reviewed in Amann and Fuchs, 2008). The PCR based methods also presented biases, for example the –considered– universal PCR primer pairs tended to amplify more than the abundant cultured taxa (Reysenbach *et al.*, 1992), as these primers were designed with databases including cultured bacteria only. Additionally, many taxa contain several copies of the SSU rRNA gene (Klappenbach *et al.*, 2000), leading to distorted estimates of diversity (Acinas *et al.*, 2004). Another important limitation of that approach is that rRNA marker genes only indicate the presence of a specific species, but they do not provide information on their functional capacity (Rodríguez-Valera, 2004). The obtention of cultures to perform physiological studies and sequence their genome was the best approach to gain information regarding the metabolic capabilities of a specific species. But for the abovementioned 'uncultured majority', the lack of cultures prevented assigning metabolic capabilities and phenotypes to these newly discovered phylotypes, more genomic context was needed. Moreover, the extent and importance of genetic exchange between bacteria could also influence the amount of functional information obtained from the 16S rRNA gene. Using Carl Woese own words: "*In the extreme, interspecies exchanges of genes could be so rampant, so broad spread, that a bacterium would not actually have a history in its own right; it would be an evolutionary chimera, each with its own history*" (Woese, 1987). New approaches were needed to gain access to the genomic properties of the most environmentally relevant microbes to expand ecosystem knowledge.

The first steps towards the current omics methodologies in the marine field relied in deep artisanal molecular work. Stein *et al.* (1996) investigated the properties of the Crenarchaeota marine archaean clade in Hawaiian ocean waters for which there was no cultured representative. After filtering 30 liters of water, they used a fosmid cloning strategy with large DNA fragments (up to 40 Kb), selecting the correct clones to sequence through PCR, retrieving multiple unknown genes from the group (Stein *et al.*, 1996). After the first genomic approaches based on the analysis of cosmids and fosmids (e.g. Béjà *et al.*, 2000), the first *en masse* whole-genome shotgun sequencing was obtained from the Sargasso Sea in the context of the Global Ocean Sampling expedition (Venter *et al.*, 2004). Towards 2007, high-throughput sequencing technologies begun to be fairly common making possible diversity analyses at unprecedented scales, allowing to differentiate thousands of taxa from a single sample using marker genes (reviewed in Goodwin *et al.*, 2016). The new sequencing technologies allowed to uncover the "rare biosphere", a large number of low abundant taxa present in every sample (Sogin *et al.*, 2006). This discovery had important ecological implications linked to the paradigm of "*everything is everywhere, but the environment selects*" (Beijerinck, 1913; Becking, 1934), with these low abundant taxa acting as a reservoir of phylogenetic and functional diversity (reviewed in Pedrós-Alió, 2012). These technologies also meant a substantial improvement in the delineation of biogeographic and temporal patterns for the abundant groups inhabiting our seas (Ghiglione *et al.*, 2012; Gilbert *et al.*, 2012; Salazar *et al.*, 2016). The read outputs from these technologies were from 5 to 7 orders of magnitude higher than the one obtained with Sanger sequencing, albeit presenting shorter reads (Glenn, 2011; Goodwin *et al.*, 2016). These magnitudes allowed to directly sequence DNA fragments directly from the environment, resulting in metagenomic datasets, which enabled the study of the functional potential of whole communities circumventing PCR (reviewed in Grossart *et al.*, 2020). Likewise, it resulted in the obtention of metatranscriptomes, gathering information of the expression of these functions through sequencing the RNA after retro-transcription (Su *et al.*, 2012). Nowadays, we are in the middle of this golden era, able to recover hundreds of genomes from one sample and providing plenty of information characterizing the functionality of the most abundant groups.

Omics and the importance of the biological unit of study

The basic unit of biological diversity is considered to be the species (Rosselló-Móra and Amann, 2015). The definition of what a species is for bacteria and archaea is difficult because of the ability of these organisms to incorporate foreign DNA into their genome (horizontal gene transference), not always following the mendelian vertical transmission from one generation to the next. Nowadays, a prokaryotic species is considered to be a group of genetically and ecologically similar individuals recognizable as distinct clusters, based on genetic similarity and differences to other species (Rosselló-Móra and Amann, 2015; Shapiro *et al.*, 2016).

Traditionally, marine bacteria have been taxonomically classified following the characteristics identified in cultures (Ammerman *et al.*, 1984). On the other hand, a definition based in genomic similarity stated that a degree of >70% genome to genome cross-hybridization could define the limit between species (Wayne *et al.*, 1987). Although this criterion was useful, many authors tried to find a similar threshold for the 16S rRNA gene. In 1994, Erno Stackerbrandt and Brett Goebel proposed 97% similarity clustering as a species threshold using the full-length 16S rRNA gene sequence (Stackebrandt and Goebel, 1994). With the advent of high throughput sequencing technologies –which generate smaller read lengths than Sanger sequencing– this threshold had to be adapted to shorter fragments, and the amplification focused on the hypervariable regions of the 16S rRNA gene. There are 9 hypervariable regions in this gene, placed among the conserved ones, and using universal primers targeting the flanking conserved regions an amplicon is obtained, that is sequenced afterwards. Even though this amplicon approach was based on a small fraction of the gene, the 97% identity clustering was kept (Schloss and Handelsman, 2005). The biological unit of study was an "operational taxonomic unit" (OTU), a pragmatic definition to discriminate taxa, without a direct relationship with the species concept.

Recent analyses of the power of resolution of the 16S rRNA gene to distinguish species found that 99% similarity would be a better threshold for the full-length gene but that 100% should be used for short amplicons (~250 base pairs; Edgar, 2018; Johnson *et al.*, 2019). Consequently, there have been multiple efforts to avoid merging (clustering) the sequences by identity thresholds and use single variants instead. The most recent algorithms completely avoid this step. Using the quality information of each single base provided by the sequencer, the algorithms predict through statistical models if a read sequence is likely to be a true biological variant or is the result of sequencing errors (Figure 2). Tools such as DADA2 and MED (among others) denoise the reads of what is assumed to be errors from true biological variation (Eren *et al.*, 2013; Callahan *et al.*, 2016). The most used term for the obtained units is *amplicon sequence variants* (ASVs; see the origin of the concept in https://github.com/benjjneb/dada2/issues/62). ASVs are variants with a unique and consistent origin, in contrast to OTUs, which lump sequences together, likely losing true variants in the process, and requiring a choice of the representative sequence for each cluster. Moreover, the process allows to compare sequence variants across different datasets directly (Callahan *et al.*, 2017). Nevertheless, although the use of ASVs has improved the level of resolution at which we can analyze microbial diversity, given that the resolution of each specific hypervariable region is variable for each taxon, it is generally accepted that while the 16S rRNA gene amplicon sequencing can easily be used to classify organisms down to the genus level, it is not well suited to robustly differentiate all the present species (Johnson *et al.*, 2019).

In addition, the amplicon approach can be applied to other marker genes specific for functional groups of interest. Examples include the *amoA* for ammonia oxidizing bacteria (Rotthauwe *et al.*, 1997), *pufM* for bacteriochlorophyll *a*-containing photoheterotrophs (Yutin *et al.*, 2005), and *nifH* for nitrogen fixation (Zehr and McReynolds, 1989), among others. For each of these genes, optimal clustering thresholds have been determined, but in general there is a lack of studies applying the abovementioned threshold-free algorithms to discern their variability. Generally, the use of amplicon sequencing approaches has the advantage that are easily scalable to hundreds of samples for relatively low cost and the low computing analytical cost. Additionally, the use of marker genes for functional groups allows to define biological units beyond the 16S rRNA gene without having to obtain the complete genome.

On the other hand, whole metagenome approaches can be used in different ways that could be classified as gene centric and whole genome analyses (Figure 2). In any of these methods, the thousands to millions of sequenced short (usually 150 base pairs) reads are assembled to reconstruct larger genome fragments (contigs) displaying a wide range of sizes. The gene centric approach uses the gene as a biological unit (Figure 2). It is based in the prediction of all the putative protein-coding genes in each of these contigs, and generally results in millions of genes. To make the procedure computationally tractable, the gene duplicates are pruned through clustering procedures (95% identity generally), generating the so-called reference gene catalog, often containing millions of gene variants. The abundance of each gene is then estimated by assignment of the raw reads to the genes on that gene catalog. This approach has been useful, for example, to obtain new gene variants (Sunagawa *et al.*, 2015), analyze clusters of abundant genes (Minot *et al.*, 2021), or to establish relationships between the presence of a function and its expression (Salazar *et al.*, 2019).

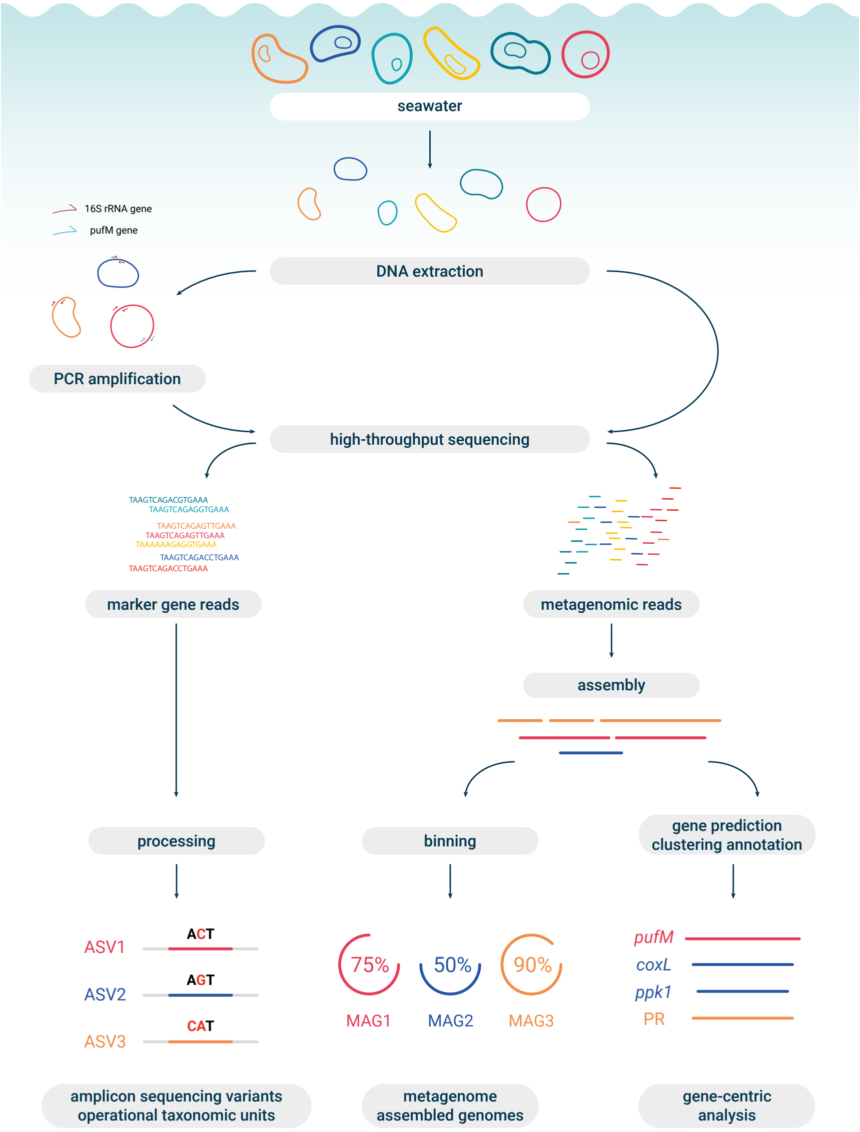


Figure 2: Diagram of the most common approaches to work with amplicon and metagenomic data in marine microbiome analysis.

Another avenue of analysis is based in the obtention of "whole" genomes, in which the assembled contigs are classified (binned) together into Metagenome Assembled Genomes (MAGs, Figure 2). MAGs are composite genomes of populations from natural communities. The binning step classifies the contigs using the tetranucleotide frequency –the nucleotide composition of the sequences, phylogenetically conserved to each organism– and the contig mean sample abundance –the mean coverage, mean read number recruited for each contig– as information. These values can resemble between similar populations within species (strains) and among closely related ones, and thus the binning algorithms can sometimes merge them into the same MAG genome. In this context, the biological unit is the population genome of –ideally– a single species. The first attempts to obtain a genome from metagenomic sequences of environmental communities took place in the early 2000s (Tyson *et al.*, 2004; Martín *et al.*, 2006) but the improvement of the methods was not reached until last decade (Albertsen *et al.*, 2013; Delmont and Eren, 2018; Parks *et al.*, 2018). Recent efforts have reconstructed up to tens of thousands of MAGs from different ecosystems, expanding the known phylogenetic diversity of bacteria and archaea by 44% (Nayfach *et al.*, 2021). The availability of thousands of new genomes coupled with robust bioinformatic platforms have guided the assignment of higher taxa, leading to new phyla, class, order and family designations (Parks *et al.*, 2018, 2020; Rinke *et al.*, 2021). Another popular approach is single-cell genome sequencing (or SAGs, from single amplified genomes), generated after sorting individual cells, direct amplification of the single genomes and its sequencing. The obtention of a single genome (as long as there is a way to physically separate the cells) overcomes the problems of mixing strain genomes in the MAG generation (Macaulay and Voet, 2014). A recent study was able to obtain 12715 SAGs from the surface ocean, allowing to link the genomes of the main taxonomic marine groups with their cell sizes in natural conditions and obtaining key metabolic information (Pachiadaki *et al.*, 2019).

Phylogenetic diversity of bacteria and archaea in surface ocean waters

The characterization of the types of microorganisms present in the sea is a central object of study in the field of marine microbial ecology. As a generality, Alphaproteobacteria, Gammaproteobacteria, Bacteroidetes and Cyanobacteria are the most dominant taxa in surface waters. Within each one, however, there is a wide array of physiologies, metabolisms and ecological strategies. Heterotrophic models such as *Alteromonas* (Gammaproteobacteria), *Flavobacteria* (Bacteroidetes), or *Roseobacter* (Alphaproteobacteria) are copiotrophs able to grow with relatively high concentrations of nutrients, and therefore easily manageable in the laboratory. These groups are usually predominant when there are enough nutrients, either due to the mixing of the water column or a high discharge of organic matter, such as in algal blooms (Buchan *et al.*, 2014). Within the *Rhodobacteraceae* family, the *Roseobacter* is the most well-known clade, with multiple species capable of metabolizing a large number of carbon sources, synthesizing B vitamins that might induce symbiotic interactions with eukaryotes, and able to perform dissimilatory nitrate reduction (Luo and Moran, 2014). Other groups such as the *Flavobacteriaceae* (Bacteroidetes) generally present the metabolic machinery for degrading high molecular weight polysaccharides, such as glycoside hydrolases, polysaccharide lyases and proteases (Buchan *et al.*, 2014; Teeling *et al.*, 2016; Krüger *et al.*, 2019).

The ocean, however, is not eutrophic in general but rather oligotrophic. The most abundant bacterioplankton groups present specializations to develop in conditions of scarcity. Oligotrophs tend to have small cells, allowing a low surface to volume ratio to facilitate the molecule transport from the medium (Giovannoni *et al.*, 2014). They also present highly compacted genomes with low GC content, short intergenic spacers and highly conserved core genomes. The evolutionary origin of these characteristics and genomic features is known as the "streamlining theory" critical to success in nutrient-poor environments, where either gathering a larger share of nutrient resources, or using them more efficiently, can increase success (Giovannoni *et al.*, 2014). This theory has been linked with to the "Black Queen hypothesis", that refers to selection processes favoring minimization of cell size and complexity, usually creating dependency with co-occurring taxa (Morris *et al.*, 2012). These co-dependencies within taxa are one of the features that sometimes difficult cultivation of these types of organisms. In the last two decades, however, there has been successful efforts to obtain information about these oligotrophic clades, and for some of the groups several isolates have been obtained (Rappé *et al.*, 2002; Lee *et al.*, 2019).

Perhaps some of the most studied groups in the surface ocean are *Prochlorococcus* and *Synechococcus*. These cyanobacterial groups contribute around 25% of the primary productivity in the oceans and dominate most of the oceanic oligotrophic surface waters (Flombaum *et al.*, 2013). Considered traditionally photoautotrophs, it has been proven that many of the species present some degree of mixotrophic metabolism (see a review in Muñoz-Marín *et al.*, 2020). *Prochlorococcus* cells are usually smaller than those of *Synechococcus* (~0.6 vs ~0.9 µm), and their fast growth combined with large population sizes have allowed the group to adapt its genome content to the open ocean nutritional conditions (Partensky and Garczarek, 2010; Delmont and Eren, 2018; Ustick *et al.*, 2021). The *Synechococcus* genus is more generalist than *Prochlorococcus*, with some clades able to take advantage of fluctuating environments with higher nutrients (Rocap *et al.*, 2002; Palenik *et al.*, 2003). Another well-studied group presenting a widespread distribution is the SAR11 clade (order Pelagibacterales). The SAR acronym refers to the Sargasso Sea, from which clones were first retrieved (Giovannoni *et al.*, 1990). The first isolate of SAR11 –obtained in 2002 through a dilution culturing method– was proposed as *Candidatus* Pelagibacter ubique (Rappé *et al.*, 2002). The clade represents ~25% of the of the plankton cells in upper regions of the ocean photic zone (Morris *et al.*, 2002; Rusch *et al.*, 2007; Salcher *et al.*, 2011), with streamlined genomes able to oxidize a wide variety of one-carbon compounds (Giovannoni, 2017). The different subclades present multiple specializations and distributions, with subclade Ia.3 adapted to warm surface waters, whereas Ic is adapted to the dark ocean (Giovannoni, 2017). Recent metagenomic analyses defined the biogeography of the group (Delmont *et al.*, 2019), and the capacity to recombine even between distantly related members (López-Pérez *et al.*, 2020). The high recombination rates among this group generate high genomic microdiversity, challenging the operational boundaries to define a microbial species (López-Pérez *et al.*, 2020). Other cosmopolitan clades are the SAR86, SAR116, or the Acidimicrobiales order within the Actinobacteria, from which similar advances in the study of their biogeographic distribution and the recovery of species have recently occurred (Dupont *et al.*, 2012; Mizuno *et al.*, 2015; Roda-Garcia *et al.*, 2021). As an example, SAR86 is one of the most abundant marine clades, belonging to the Gammaproteobacteria class and sharing some traits with SAR11 such as metabolic streamlining, but it also presents a distinct carbon compound specialization that might possibly avoid competition with SAR11 (Dupont *et al.*, 2012). The analyses of the SAR86 pangenome (both the core and flexible genome) indicate that it is composed of different ecotypes with unique geographic distributions (Hoarfrost *et al.*, 2020). The Archaea domain has also presented advances in its knowledge. Since the discovery of archaea thriving in the sunlit ocean (DeLong, 1992), multiple groups have been unveiled, together with its biogeography (reviewed in Santoro *et al.*, 2019). As an example, the Thaumarchaea, initially found in coastal surface waters (DeLong, 1992) generally gain energy from the oxidation of ammonia, and are dominant (sometimes up to 40% of the total cells) in mesopelagic waters, where the presence of nitrogen compounds is higher (Karner *et al.*, 2001). Overall, the characterization of the major clades in the surface ocean has advanced substantially in the recent decades, obtaining an initial picture of the whole community structure. This advance has in turn implicated the discovery of new metabolisms, shaping and changing our view of the biogeochemical cycles in the oceans.

Bacterioplankton photoheterotrophy in the oceans

The classic dichotomy of photoautotrophs as primary producers and heterotrophs as consumers of organic carbon was challenged in the last decades by the realization of the relevance of bacterioplankton photoheterotrophy in the marine system. Early genomic analyses of marine bacterioplankton reported that an uncultured bacterium harbored a gene coding for proteorhodopsin (PR), a light-dependent proton pump able to produce a new type of prototrophy (Béjà *et al.*, 2000). That same year, high signals of bacteriochlorophyll *a* in the surface oligotrophic ocean were detected using infrared fluorometry (Kolber *et al.*, 2000). The latter results suggested that aerobic anoxygenic phototrophic (AAP) bacteria were a substantial component of the marine microbiome. These two reports initiated a change of paradigm in the field of marine microbial ecology, adding the direct effects of light to the well-known deleterious or stimulating effects it had on microbial heterotrophic processes (Ruiz-González *et al.*, 2013). This new knowledge could substantially modify the models of organic carbon fluxes in the ocean. Recent studies have found that proton-pumping proteorhodopsins potentially absorb as much light energy as chlorophyll *a* (Gómez-Consarnau *et al.*, 2019).

The molecular approaches explained above have revealed a large diversity among the PR and AAP bacteria (and archaea for PR), showing that the genes responsible for photoheterotrophy are common among the most abundant microbial taxa of the surface ocean (DeLong and Béjà, 2010; Koblížek, 2015). Proton-pumping rhodopsins are found in marine Proteobacteria (including the Pelagibacterales order), Bacteroidetes, Puniceispirillales and Euryarchaeota (see a review in Pinhassi *et al.*, 2016). Proteorhodopsins consist of only one opsin protein bounded covalently to a pigment (retinal), and this simple structure coupled with a small energetic production cost has favored its lateral gene transfer among distant taxa (Frigaard *et al.*, 2006; Kirchman and Hanson, 2013). Contrarily, the AAP machinery for light-harvesting and energy synthesis consists of several pigments and proteins with a more constrained phylogenetic distribution, being present –in marine samples– mostly in the Alpha- and Gammaproteobacteria. While the broad occurrence of these systems has been well described, less is known about their physiology and ecology. It has been shown that some axenic cultures of AAP and PR bacteria use light to grow more efficiently (Gómez-Consarnau *et al.*, 2007; Hauruseu and Koblízek, 2012; Arandia-Gorostidi *et al.*, 2020), and this has also been confirmed for natural AAP populations (Ferrera *et al.*, 2017). In contrast, other PR isolates do not seem to grow better under light conditions (González *et al.*, 2008), yet proteorhodopsins can facilitate survival during starvation (Gómez-Consarnau *et al.*, 2010). Photoheterotrophs are an illustrative example of how methodological development can improve our comprehension of ecosystem functioning. From its initial finding in 2000, its study and quantification could substantially modify the models of organic carbon fluxes in the ocean. Like photoheterotrophs, the study of other functional groups through marker genes can help understand its role in marine ecosystems.

Key microbial marker genes in the biogeochemical marine cycles

Marine biogeochemical cycles are deeply influenced by the genetic repertoire of the marine microbial community, since multiple processes are exclusively driven the different functional groups inhabiting the marine ecosystem. The omics approaches have allowed exploring the functional landscape of the sunlit ocean (reviewed in Ferrera *et al.*, 2015, Figure 3). For example, phytoplankton fix carbon through photosynthesis using the photosystem complex (*psbA*) and the Rubisco enzyme (*rbcL*), incorporate inorganic nutrients and release organic matter (both dissolved and particulate, DOM and POM), dissolved and particulate organic phosphorous (DOP and POP) and dissolved and particulate organic nitrogen (DON and PON). Certain marine phytoplankton groups also produce large quantities of dimethylsulphoniopropionate (DMSP), which accounts in some cases for up to 10% of the carbon fixed (Simó *et al.*, 2002). Bacteria and archaea use this released organic matter and compete with eukaryotic phytoplankton for inorganic nutrients. DMSP specifically provides a substantial fraction of the carbon and sulfur requirements of heterotrophic marine bacteria (Kiene *et al.*, 2000), and acts as a potent chemoattractant towards phytoplankton (Seymour *et al.*, 2017). It can be cleaved by different DMSP lyases (*ddd* genes) to DMS, eventually released to the atmosphere, or demethylated and used as a reduced sulfur source (*dmdA*). DMS release to the atmosphere has been linked to climate regulation through affecting cloud formation (reviewed in Carpenter *et al.*, 2012). Regarding the essential nutrients, phosphorus (P) and nitrogen (N) availability is one of the dominant selective forces driving niche differentiation in bacteria and archaea (Coleman and Chisholm, 2010; Ustick *et al.*, 2021). DOP is mainly composed of phosphoesters (sugar phosphates, vitamins, nucleotides, etc.) and phosphonates (reduced P compounds with a covalent C-P bond), which to be used require the action of alkaline phosphatases (*phoX, phoA, phoD*), and phosphonate genes (*phn* operon), respectively. Multiple marine groups adapt to the deficiency of P through the

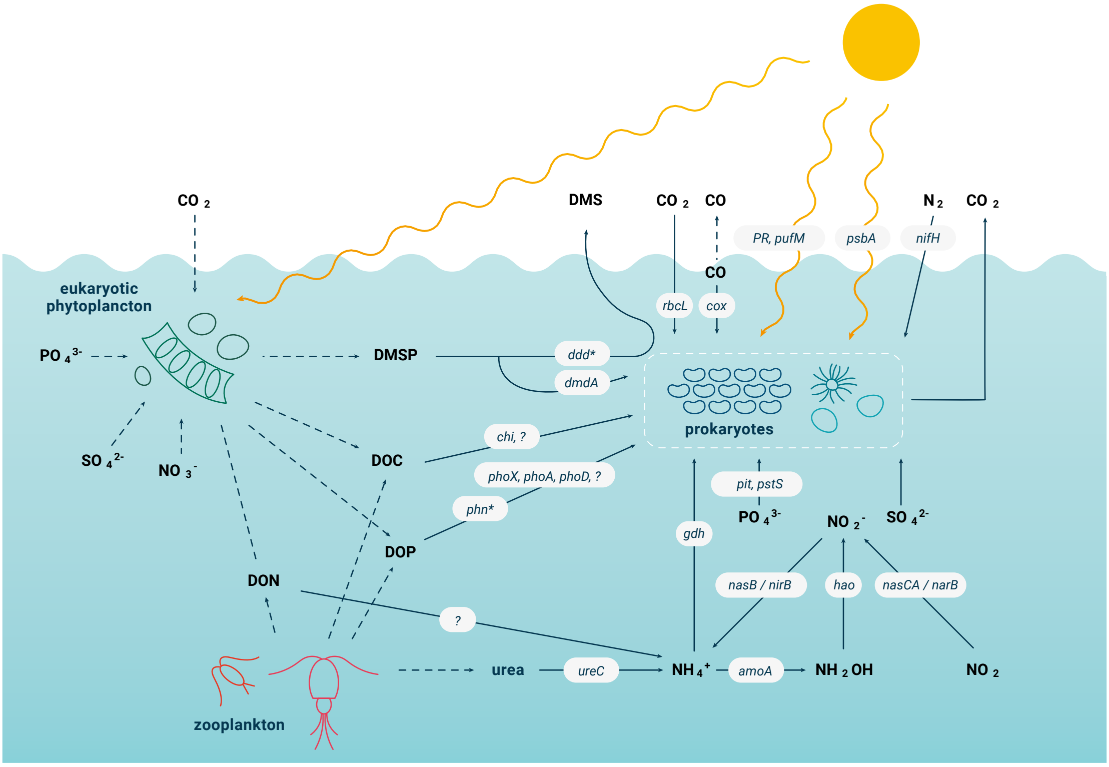


Figure 3: Schematic representation of marine microbial food webs and major biogeochemical cycles in the sunlit ocean. Solid lines indicate prokaryotic mediated processes and the key functional genes involved in the processes are shown in boxes. Asterisks denote groups of genes with a related function. Modified from Ferrera *et al.* (2015).

replacement of membrane phospholipids with alternative non-phosphorous lipids (*plcP*, Sebastián *et al.*, 2016). The N requirements are obtained from direct ammonia (*amt*) or nitrate uptake (*nasA, narB*), urea degradation (*ureC*), and some groups are capable to fix N_2 by means of nitrogenases (encoded in the *nif* operon).

Aside from a nutrient supply, the heterotrophic bacteria also use multiple small molecules as an energy source. The carbon monoxide –formed through the photochemical degradation of organic matter in sunlit waters– can be oxidated to CO_2 by the carbon monoxide dehydrogenase (*cox*), being *Rhodobacteraceae* one of the best known groups presenting this metabolism (King and Weber, 2007; Luo and Moran, 2014). Another molecule used to produce energy is ammonia (through the *amoA* gene). The sequencing of the first marine metagenomes (Venter *et al.*, 2004) retrieved archaeal sequences of this gene, challenging the previous assumption that this function was exclusive of some proteobacterial groups. Nowadays, it is well known that the function in the surface ocean is mainly driven by archaeal groups (reviewed in DeLong, 2021). The discovery and characterization of these and other functional genes has facilitated a more complete understanding of the biogeochemical cycles and the definition of the niche space of the various taxa in the ocean, i.e. what metabolic strategies pelagic bacteria and archaea use to thrive in the oceans. For most of these genes there are however a lot of unknowns yet (Ferrera *et al.*, 2015), among them, their taxonomic distribution and their seasonal variability in the marine ecosystems.

The seasonality of the marine surface microbiome

The sea as an ecosystem presents temporal changes influencing and driving the microbe dynamics. The range of this temporal scale varies from hours to long interannual changes such as seasons. One of the clearest examples of these seasonal changes are phytoplankton blooms. The growth of specific photosynthetic groups due to high temperatures and excess nutrients can cause macroscopic colorful blooms observable sometimes from satellites. These events can be recurrent (Garces, 1999), span distances from meters to kilometers, and sometimes present neurotoxic properties, killing wildlife (Zohdi and Abbaspour, 2019). Humans have even created myths around these events, such as the River of Blood in the bible (Martin and Martin, 1976). The study of these seasonal phytoplankton blooms is a clear example of the importance of studying the temporal scale in marine microbial ecosystems.

Most of the efforts aimed to understand how microbial communities change over time have been concentrated in a few long-term microbial observatories. The establishment of these stations across the globe has allowed to study the seasonality at different latitudes from short- to long-term scales (see reviews by Bunse and Pinhassi, 2017; Buttigieg *et al.*, 2018). Defining seasonality is essential to understand how microbes react to long-term changes in environmental conditions or short-term perturbations. Microbial time series also allow addressing relevant ecological questions, such as the diversity patterns in an ecosystem, the stability and predictability of microbial communities, establishing the interaction among species, and the temporal ecological niche of a given taxon. Disentangling the seasonality of specific taxonomic groups instead of that of the bulk bacterioplankton community could only be investigated with the molecular revolution. The application of fingerprinting methods and clone libraries to samples from the observatories over 1–2 year periods elucidated community shifts over seasons, demonstrating the existence of temporal niches for specific groups (Brown *et al.*, 2005; Alonso-Sáez *et al.*, 2007). Nevertheless, multiple consecutive years were needed to test if these patterns were robust and truly seasonal. Thus, long-term time series were undertaken in oceanic and coastal monitoring stations such as the San Pedro Ocean Time Series (SPOT) and the Hawaii Ocean Time Series (HOT) in the Pacific Ocean, the Bermuda Atlantic Time Series (BATS) in the Atlantic Ocean, the Western Channel Observatory in the English Channel, the Linnaeus Microbial Observatory in the Baltic Sea, or the Service d’Observation du Laboratoire Arago (SOLA Station; Banyuls-sur-Mer, France) and the Blanes Bay Microbial Observatory (BBMO) in the Mediterranean Sea, among others (Buttigieg *et al.*, 2018). Although most of the long-term sites described up to date are located at a similar latitude range, nowadays other stations are generating data enlarging the biogeographic coverage. Examples are the Australian Marine Microbial Biodiversity Initiative (Brown *et al.*, 2018) or the characterization of the seasonal patterns in the Bedford Basin, in the Artic (El-Swais *et al.*, 2015).

In fact, the last decade has become the golden era of time series studies due to the patient year to year sampling and the improvement of high-throughput technologies. The application of molecular fingerprinting methods over monthly samples at SPOT revealed remarkably repeatable and predictable seasonal patterns in the distribution and abundance of microbial taxa (Fuhrman *et al.*, 2006). These patterns were reflected in a dissimilarity analysis, showing that communities are more similar when they are 12 months apart and more dissimilar when they are 6 months apart, a pattern reoccurring during the 10 years of study (Fuhrman *et al.*, 2015). With sequencing technology improvements, other locations confirmed these observations at higher resolution (Eiler *et al.*, 2011; Gilbert *et al.*, 2012; Cram *et al.*, 2015), and unveiled that the rare members of the bacterioplankton community also showed seasonality (Alonso-Sáez *et al.*, 2015). In addition, the long-term time series allowed to distinguish conditionally rare taxa, groups that bloomed when the conditions were favorable (Gilbert *et al.*, 2012; Shade *et al.*, 2014). At the BATS station, the multi-year sampling allowed an improved understanding of the evolutionary diversification of the SAR11 clade (Vergin *et al.*, 2013). The seasonal patterns of other relevant phylogenetic groups such as Flavobacteria (Díez-Vives *et al.*, 2019), Gammaproteobacteria and *Roseobacter* (Teeling *et al.*, 2016) and Archaea (Hugoni *et al.*, 2013) were also determined. Likewise, microbial eukaryotes also were shown to display recurrent seasonal patterns (Lambert *et al.*, 2018; Giner *et al.*, 2019). At a community level, changes in alpha diversity were repetitive, often presenting the highest values in autumn and winter (Gilbert *et al.*, 2012; Giner *et al.*, 2019). The particle attached community also presented seasonal changes, with the communities in the larger size fractions presenting the strongest annual changes (Mestre *et al.*, 2020).

As explained above, current methodologies differentiate the hypervariable regions of the 16S rRNA gene down to single nucleotide differences, allowing to separate closely related taxa that had previously been lumped together (Eren *et al.*, 2013; Callahan *et al.*, 2016). These approaches, coupled to high-frequency sampling over multiple years and network analysis has shown that regardless of the interannual variation in phytoplankton blooms, microbes respond in co-varying modules (Chafee *et al.*, 2018). The pattern of covariation is sometimes linked to specific ecological strategies, with network modules presenting mainly oligotrophic bacteria such as SAR11, and others composed mainly of copiotrophic taxa such as *Tenacibaculum* or *Pseudoalteromonas* species (Lemonnier *et al.*, 2020). In addition, high frequency sampling over a phytoplankton bloom showed that biological interactions among bacteria, archaea, and eukaryotic microorganisms may substantially influence global plankton diversity and dynamics (Needham and Fuhrman, 2016). These results seem to complement the traditional view of the bloom being mainly controlled by physical and chemical processes, and dwell into the importance of biotic interactions such as auxotrophies and grazing by mixotrophic behavior (Johnson *et al.*, 2020). These biotic interactions are not fixed and can change under contrasting environmental conditions (Lambert *et al.*, 2021). Daily sampling during several months has also revealed that coastal microbial plankton can be organized in defined but ephemeral communities whose turnover is rapid, mirroring environmental variability (Martin-Platero *et al.*, 2018). On a functional basis, however, studies in time-series are scarce. Initial analyses with a small sample number (eight samples) hinted that the metagenomic patterns were linked to seasonality whereas the transcriptomic patterns were better explained by diel patterns and shifts in specific functional genes (Gilbert *et al.*, 2010). These day-night activity dynamic shifts for heterotrophic bacteria have been linked to both direct solar radiation and the products or photosynthesis by phytoplankton (Gifford *et al.*, 2014). Analyzing 3 years of data, Galand *et al.* (2018) showed the seasonal pattern for some specific functions such as the fixation of carbon or the flagellar assembly, demonstrating that on a seasonal scale the functional redundancy in marine waters is rather low. Recently, a metatranscriptomic study using two years of data has shown that the expression of key marker genes change through seasons (Alonso-Sáez *et al.*, 2020). Altogether, long-term series have provided evidence for seasonal and interannual recurrence of some microbial taxa and highly resolved time series have shown community fluctuation on a daily and monthly scale alongside changes in environmental conditions.

Unraveling the factors regulating the microbiome using experimental manipulations

Although long-term observational studies are key to understand nature, teasing apart the factors that mechanistically shape community structure is challenging only using descriptive data. Experimental approaches on the other hand are well suited to solve these difficulties allowing conceptual and causality-driven hypotheses. In recent years, the introduction of new techniques to microbial ecology such as the omics analyses have facilitated the descriptive approaches sometimes at the expenses of experimentation, and some authors request a renewed focus in hypothesis-driven approaches (Prosser, 2020).

Marine microbial ecology has a large experimental tradition focused on disentangling the ecological processes governing the ecosystem. Among the large range of conditions, the effect of top-down (mortality, including predation or viral lysis) and bottom-up (resource availability, either nutrients, carbon or energy resources) factors have been thoroughly studied, since they drive the community assembly through selection (Vellend, 2010). The upper limit of the bacterial population size is defined by the bottom-up factors. The concentration of dissolved and particulate organic carbon, nitrogen or phosphorous altogether with other trace nutrients influence the maximum growth of the organisms, and the importance of each element varies between ecosystems. As an example, in the Mediterranean Sea, Pinhassi *et al.* (2006) proved experimentally that phosphorous was the most limiting nutrient, being this limitation more pronounced during spring and summer. Complementing the selection by nutrient concentration, the actual (or realized population size) population is controlled by top-down mortality factors (McQueen *et al.*, 1986; Ducklow and Carlson, 1992). Experiments with natural heterotrophic flagellate assemblages have shown that predators have preferences for certain preys (e.g. Massana *et al.*, 2009). Other experiments have differentiated through FISH which specific groups are favored by the removal of predation; Alteromonadales, Bacteroidetes and Rhodobacterales have considerably higher growth rates after predator removal (Ferrera *et al.*, 2011; Sánchez *et al.*, 2017). Similarly, functional groups such as the aerobic anoxygenic phototrophic bacteria are also tightly regulated by predation (Ferrera *et al.*, 2011, 2017). Viruses can also regulate community structure directly through lysis and indirectly by providing DOM and nutrients through the lysis. Initial experiments found that virus are capable of causing up to 50% of the bacterial mortality in several aquatic environments (Fuhrman and Noble, 1995; Guixa-Boixareu *et al.*, 1996), although their presence and impact is variable (Noble and Fuhrman, 2000). A more recent experiment linked the ecological strategies of the viruses to the productive state of the ecosystem, identifying lysogenic strategies when bacterial productivity is low and switching to the lytic strategy when bacterial production increases (Brum *et al.*, 2016). Most of these experiments however were focused on bulk community estimations determining the community composition at the order level. Nowadays, studies have tried to improve and work at a higher taxonomic resolution, such as OTUs (Teira *et al.*, 2019) or ASVs (Fecskeová *et al.*, 2021). The next logical step to understand the links between microbial composition and biogeochemical processes would be to test hypotheses formulated on a metagenomic basis (Grossart *et al.*, 2020). Some recent experiments have pointed towards this direction; Beier *et al.* (2017) assessed the functional redundancy of communities along environmental gradients through metatranscriptomics and mesocosms, Bertrand *et al.* (2015) tested the effect of nutrient limitation in phytoplankton-bacterial interactions using on-ship bottles, and Haro-Moreno *et al.* (2019) used metagenomics to evaluate the microbial succession dynamics of communities collected at various depths. These studies show the usefulness of experiments with replicates using metagenomics to provide insights into the natural dynamics, complexity and function of marine microbial communities.

References

Acinas, S.G., Antón, J., and Rodríguez-Valera, F. (1999) Diversity of free-living and attached bacteria in offshore Western Mediterranean waters as depicted by analysis of genes encoding 16S rRNA. *Appl Environ Microbiol* 65: 514–522.

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

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

Alonso-Sáez, L., Balagué, V., Sà, E.L., Sánchez, O., González, J.M., Pinhassi, J., *et al.* (2007) Seasonality in bacterial diversity in north-west Mediterranean coastal waters: Assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiol Ecol* 60: 98–112.

Alonso-Sáez, L., Díaz-Pérez, L., and Morán, X.A.G. (2015) The hidden seasonality of the rare biosphere in coastal marine bacterioplankton. *Environ Microbiol* 17: 3766–3780.

Alonso-Sáez, L., Morán, X.A.G., and González, J.M. (2020) Transcriptional Patterns of Biogeochemically Relevant Marker Genes by Temperate Marine Bacteria. *Front Microbiol* 11: 465.

Amann, R. and Fuchs, B.M. (2008) Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nat Rev Microbiol* 6: 339–348.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59: 143–169.

Ammerman, J., Fuhrman, J., Hagström, Å., and Azam, F. (1984) Bacterioplankton growth in seawater: I. Growth kinetics and cellular characteristics in seawater cultures. *Mar Ecol Prog Ser* 18: 31–39.

Arandia-Gorostidi, N., González, J.M., Huete-Stauffer, T.M., Ansari, M.I., Morán, X.A.G., and Alonso-Sáez, L. (2020) Light supports cell-integrity and growth rates of taxonomically diverse coastal photoheterotrophs. *Environ Microbiol* 22: 3823–3837.

Azam, F., Fenchel, T., Field, J., Gray, J., Meyer-Reil, L., and Thingstad, F. (1983) The Ecological Role of Water-Column Microbes in the Sea. *Mar Ecol Prog Ser* 10: 257–263.

Bar-On, Y.M. and Milo, R. (2019) The Biomass Composition of the Oceans: A Blueprint of Our Blue Planet. *Cell* 179: 1451–1454.

Becking, L.G.M.B. (1934) Geobiologie of inleiding tot de milieukunde, W.P. Van Stockum & Zoon.

Beier, S., Shen, D., Schott, T., and Jürgens, K. (2017) Metatranscriptomic data reveal the effect of different community properties on multifunctional redundancy. *Mol Ecol* 26: 6813–6826.

Beijerinck, M.W. (1913) De infusies en de ontdekking der bakterien, Johannes Müller.

Béjà, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P., *et al.* (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* 289: 1902–1906.

Bertrand, E.M., McCrow, J.P., Moustafa, A., Zheng, H., McQuaid, J.B., Delmont, T.O., *et al.* (2015) Phytoplankton–bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge. *PNAS* 112: 9938–9943.

Boras, J.A., Sala, M.M., Vázquez-Domínguez, E., Weinbauer, M.G., and Vaqué, D. (2009) Annual changes of bacterial mortality due to viruses and protists in an oligotrophic coastal environment (NW Mediterranean). *Environ Microbiol* 11: 1181–1193.

Brown, M.V., van de Kamp, J., Ostrowski, M., Seymour, J.R., Ingleton, T., Messer, L.F., *et al.* (2018) Systematic, continental scale temporal monitoring of marine pelagic microbiota by the Australian Marine Microbial Biodiversity Initiative. *Sci Data* 5: 180130.

Brown, M.V., Schwalbach, M.S., Hewson, I., and Fuhrman, J.A. (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. *Environ Microbiol* 7: 1466–1479.

Brum, J.R., Hurwitz, B.L., Schofield, O., Ducklow, H.W., and Sullivan, M.B. (2016) Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. *ISME J* 10: 437–449.

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

Bunse, C. and Pinhassi, J. (2017) Marine Bacterioplankton Seasonal Succession Dynamics. *Trends in Microbiol* 25: 1–12.

Buttigieg, P.L., Fadeev, E., Bienhold, C., Hehemann, L., Offre, P., and Boetius, A. (2018) Marine microbes in 4D-using time series observation to assess the dynamics of the ocean microbiome and its links to ocean health. *Curr Opin Microbiol* 43: 169–185.

Callahan, B.J., McMurdie, P.J., and Holmes, S.P. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 11: 2639–2643.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 13: 581.

Chafee, M., Fernàndez-Guerra, A., Buttigieg, P.L., Gerdts, G., Eren, A.M., Teeling, H., and Amann, R.I. (2018) Recurrent patterns of microdiversity in a temperate coastal marine environment. *ISME J* 12: 237–252.

Chisholm, S.W., Olson, R.J., Zettler, E.R., Goericke, R., Waterbury, J.B., and Welschmeyer, N.A. (1988) A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* 334: 340–343.

Coleman, M.L. and Chisholm, S.W. (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. *PNAS* 107: 18634–18639.

Cram, J.A., Chow, C.-E.T., Sachdeva, R., Needham, D.M., Parada, A.E., Steele, J.A., and Fuhrman, J.A. (2015) Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *ISME J* 9: 563–580.

Delmont, T.O. and Eren, A.M. (2018) Linking Pangenomes and Metagenomes: The Prochlorococcus Metapangenome. *PeerJ* 6: e4320.

Delmont, T.O., Kiefl, E., Kilinc, O., Esen, O.C., Uysal, I., Rappé, M.S., *et al.* (2019) Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *eLife* 8: e46497.

DeLong, E.F. (1992) Archaea in coastal marine environments. *PNAS* 89: 5685–5689.

DeLong, E.F. (2021) Exploring Marine Planktonic Archaea: Then and Now. *Front Microbiol* 11: 616086.

DeLong, E.F. and Béjà, O. (2010) The Light-Driven Proton Pump Proteorhodopsin Enhances Bacterial Survival during Tough Times. *PLoS Biol* 8: e1000359.

DeLong, E.F., Wickham, G.S., and Pace, N.R. (1989) Phylogenetic Stains: Ribosomal RNA-Based Probes for the Identification of Single Cells. *Science* 243: 1360–1363.

Díez-Vives, C., Nielsen, S., Sánchez, P., Palenzuela, O., Ferrera, I., Sebastián, M., *et al.* (2019) Delineation of ecologically distinct units of marine Bacteroidetes in the Northwestern Mediterranean Sea. *Mol Ecol* 28: 2846– 2859.

Duarte, C.M., Agustí, S., Kennedy, H., and Vaqué, D. (1999) The Mediterranean climate as a template for Mediterranean marine ecosystems: the example of the northeast Spanish littoral. *Prog Oceanogr* 44: 245–270.

Duarte, C.M. and Cebrián, J. (1996) The fate of marine autotrophic production. *Limnol Oceanogr* 41: 1758–1766.

Ducklow, H.W. and Carlson, C.A. (1992) Oceanic Bacterial Production. In Adv Microb Ecol. Marshall, K.C. (ed). Boston, MA: Springer US, pp. 113–181.

Dupont, C.L., Rusch, D.B., Yooseph, S., Lombardo, M.-J., Alexander Richter, R., Valas, R., *et al.* (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* 6: 1186–1199.

Eakins, B.W. and Sharman, G.F. (2010) Volumes of the World's Oceans.

Edgar, R.C. (2018) Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34: 2371–2375.

Eiler, A., Hayakawa, D.H., and Rappé, M.S. (2011) Non-Random Assembly of Bacterioplankton Communities in the Subtropical North Pacific Ocean. *Front Microbiol* 2: 140.

El-Swais, H., Dunn, K.A., Bielawski, J.P., Li, W.K.W., and Walsh, D.A. (2015) Seasonal assemblages and short-lived blooms in coastal north-west Atlantic Ocean bacterioplankton: Bacterial diversity in Bedford Basin. *Environ Microbiol* 17: 3642–3661.

Eren, A.M., Maignien, L., Sul, W.J., Murphy, L.G., Grim, S.L., Morrison, H.G., and Sogin, M.L. (2013) Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol* 4: 1111–1119.

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

Fecskeová, L.K., Piwosz, K., Šantić, D., Šestanović, S., Tomaš, A.V., Hanusová, M., *et al.* (2021) Lineage-Specific Growth Curves Document Large Differences in Response of Individual Groups of Marine Bacteria to the Top-Down and Bottom-Up Controls. *mSystems* 6: e00934-21.

Ferrera, I., Borrego, C.M., Salazar, G., and Gasol, J.M. (2014) Marked seasonality of aerobic anoxygenic phototrophic bacteria in the coastal NW Mediterranean Sea as revealed by cell abundance, pigment concentration and pyrosequencing of *pufM* gene. *Environ Microbiol* 16: 2953–2965.

Ferrera, I., Gasol, J.M., Sebastián, M., Hojerová, E., and Kobížek, M. (2011) Comparison of growth rates of aerobic anoxygenic phototrophic bacteria and other bacterioplankton groups in coastal mediterranean waters. *Appl Envir Microbiol* 77: 7451–7458.

Ferrera, I., Sánchez, O., Kolářová, E., Koblížek, M., and Gasol, J.M. (2017) Light enhances the growth rates of natural populations of aerobic anoxygenic phototrophic bacteria. ISME J 11: 2391–2393.

Ferrera, I., Sebastián, M., Acinas, S.G., and Gasol, J.M. (2015) Prokaryotic functional gene diversity in the sunlit ocean: Stumbling in the dark. *Curr Opin Microbiol* 25: 33–39.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281: 237–240.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincon, J., Zabala, L.L., Jiao, N., *et al.* (2013) Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *PNAS* 110: 9824–9829.

Frigaard, N.U., Martinez, A., Mincer, T.J., and DeLong, E.F. (2006) Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* 439: 847–850.

Fuhrman, J.A., Cram, J.A., and Needham, D.M. (2015) Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol* 13: 133–146.

Fuhrman, J.A., Hewson, I., Schwalbach, M.S., Steele, J.A., Brown, M.V., and Naeem, S. (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. *PNAS* 103: 13104–13109.

Fuhrman, J.A. and Noble, R.T. (1995) Viruses and protists cause similar bacterial mortality in coastal seawater. *Limnol Oceanogr* 40: 1236–1242.

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

Garcés, E. (1999) A recurrent and localized dinoflagellate bloom in a Mediterranean beach. *J Plankton Res* 21: 2373–2391.

Ghiglione, J.-F., Galand, P.E., Pommier, T., Pedrós-Alió, C., Maas, E.W., Bakker, K., *et al.* (2012) Pole-to-pole biogeography of surface and deep marine bacterial communities. *PNAS* 109: 17633–17638.

Gifford, S.M., Sharma, S., and Moran, M.A. (2014) Linking activity and function to ecosystem dynamics in a coastal bacterioplankton community. *Front Microbiol* 5: 185.

Gilbert, J.A., Field, D., Swift, P., Thomas, S., Cummings, D., Temperton, B., *et al.* (2010) The Taxonomic and Functional Diversity of Microbes at a Temperate Coastal Site: A 'Multi-Omic' Study of Seasonal and Diel Temporal Variation. *PLoS ONE* 5: e15545.

Gilbert, J.A., Steele, J.A., Caporaso, J.G., Steinbrück, L., Reeder, J., Temperton, B., *et al.* (2012) Defining seasonal marine microbial community dynamics. *ISME J* 6: 298–308.

Giner, C.R., Balagué, V., Krabberød, A.K., Ferrera, I., Reñé, A., Garcés, E., *et al.* (2019) Quantifying long-term recurrence in planktonic microbial eukaryotes. *Mol Ecol* 28: 923–935.

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

del Giorgio, P.A. and Gasol, J.M. (2008) Physiological Structure and Single-Cell Activity in Marine Bacterioplankton. In *Microbial Ecology of the Oceans.* John Wiley & Sons, Ltd, pp. 243–298.

Giovannoni, S.J. (2017) SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Ann Rev Mar Sci* 9: 231–255.

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

Giovannoni, S.J., Cameron Thrash, J., and Temperton, B. (2014) Implications of streamlining theory for microbial ecology. *ISME J* 8: 1553–1565.

Glenn, T.C. (2011) Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 11: 759–769.

Gómez-Consarnau, L., Akram, N., Lindell, K., Pedersen, A., Neutze, R., Milton, D.L., *et al.* (2010) Proteorhodopsin Phototrophy Promotes Survival of Marine Bacteria during Starvation. *PLoS Biol* 8: e1000358.

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

Gómez-Consarnau, L., Raven, J.A., Levine, N.M., Cutter, L.S., Wang, D., Seegers, B., *et al.* (2019) Microbial rhodopsins are major contributors to the solar energy captured in the sea. *Sci Adv* 5: eaaw8855.

González, J.M., Fernández-Gómez, B., Fernàndez-Guerra, A., Gómez-Consarnau, L., Sánchez, O., Coll-Lladó, M., *et al.* (2008) Genome analysis of the proteorhodopsin-containing marine bacterium Polaribacter sp. MED152 (Flavobacteria). *PNAS* 105: 8724–8729.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17: 333–351.

Grossart, H., Massana, R., McMahon, K.D., and Walsh, D.A. (2020) Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnol Oceanogr* 65: S2–S20.

Guadayol, Ò., Peters, F., Marrasé, C., Gasol, J., Roldán, C., Berdalet, E., *et al.* (2009) Episodic meteorological and nutrient-load events as drivers of coastal planktonic ecosystem dynamics: a time-series analysis. *Mar Ecol Prog Ser* 381: 139–155.

Guixa-Boixareu, N., Calderón-Paz, J., Heldal, M., Bratbak, G., and Pedrós-Alió, C. (1996) Viral lysis and bacterivory as prokaryotic loss factors along a salinity gradient. *Aquat Microb Ecol* 11: 215–227.

Hanson, C.A., Fuhrman, J.A., Horner-Devine, M.C., and Martiny, J.B.H. (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol* 10: 497–506.

Haro-Moreno, J.M., Rodriguez-Valera, F., and López-Pérez, M. (2019) Prokaryotic Population Dynamics and Viral Predation in a Marine Succession Experiment Using Metagenomics. *Front Microbiol* 10: 2926.

Hauruseu, D. and Koblížek, M. (2012) Influence of light on carbon utilization in aerobic anoxygenic phototrophs. *App Environ Microbiol* 78: 7414–7419.

Hewson, I., Steele, J., Capone, D., and Fuhrman, J. (2006) Temporal and spatial scales of variation in bacterioplankton assemblages of oligotrophic surface waters. *Mar Ecol Prog Ser* 311: 67–77.

Hoarfrost, A., Nayfach, S., Ladau, J., Yooseph, S., Arnosti, C., Dupont, C.L., and Pollard, K.S. (2020) Global ecotypes in the ubiquitous marine clade SAR86. *ISME J* 14: 178–188.

Hugoni, M., Taib, N., Debroas, D., Domaizon, I., Jouan Dufournel, I., Bronner, G., *et al.* (2013) Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *PNAS* 110: 6004–9.

J. Carpenter, L., D. Archer, S., and Beale, R. (2012) Ocean-atmosphere trace gas exchange. *Chem Soc Rev* 41: 6473–6506.

Johnson, J.S., Spakowicz, D.J., Hong, B.-Y., Petersen, L.M., Demkowicz, P., Chen, L., *et al.* (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 10: 5029.

Johnson, W.M., Alexander, H., Bier, R.L., Miller, D.R., Muscarella, M.E., Pitz, K.J., and Smith, H. (2020) Auxotrophic interactions: a stabilizing attribute of aquatic microbial communities? *FEMS Microbiol Ecol* 96: fiaa115.

Karner, M.B., DeLong, E.F., and Karl, D.M. (2001) Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* 409: 507–510.

Kiene, R.P., Linn, L.J., and Bruton, J.A. (2000) New and important roles for DMSP in marine microbial communities. *J Sea Res* 43: 209–224.

King, G.M. and Weber, C.F. (2007) Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nat Rev Microbiol* 5: 107–118.

Kirchman, D.L. (2016) Growth Rates of Microbes in the Oceans. *Ann Rev Mar Sci* 8: 285–309.

Kirchman, D.L. ed. (2008) Microbial ecology of the oceans, 2nd ed., rev. Hoboken, N.J: Wiley-Blackwell.

Kirchman, D.L. and Hanson, T.E. (2013) Bioenergetics of photoheterotrophic bacteria in the oceans: Phototrophy bioenergetic. *Env Microbiol Rep* 5: 188–199.

Klappenbach, J.A., Dunbar, J.M., and Schmidt, T.M. (2000) rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. *Appl Environ Microbiol* 66: 1328–1333.

Koblížek, M. (2015) Ecology of aerobic anoxygenic phototrophs in aquatic environments. *FEMS Microbiol Rev* 39: 854–870.

Kolber, Z.S., Van Dover, C.L., Niederman, R. a, and Falkowski, P.G. (2000) Bacterial photosynthesis in surface waters of the open ocean. *Nature* 407: 177–179.

Krüger, K., Chafee, M., Ben Francis, T., Glavina del Rio, T., Becher, D., Schweder, T., *et al.* (2019) In marine Bacteroidetes the bulk of glycan degradation during algae blooms is mediated by few clades using a restricted set of genes. *ISME J* 13: 2800–2816.

Lambert, S., Lozano, J.-C., Bouget, F.-Y., and Galand, P.E. (2021) Seasonal marine microorganisms change neighbours under contrasting environmental conditions. *Environ Microbiol* 23: 2592-2604.

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., and Galand, P.E. (2018) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *ISME J* 13: 388–401.

Lee, J., Kwon, K.K., Lim, S.-I., Song, J., Choi, A.R., Yang, S.-H., *et al.* (2019) Isolation, cultivation, and genome analysis of proteorhodopsin-containing SAR116-clade strain Candidatus Puniceispirillum marinum IMCC1322. *J Microbiol* 57: 676–687.

Lemonnier, C., Perennou, M., Eveillard, D., Fernandez-Guerra, A., Leynaert, A., Marié, L., *et al.* (2020) Linking Spatial and Temporal Dynamic of Bacterioplankton Communities With Ecological Strategies Across a Coastal Frontal Area. *Front Mar Sci* 7: 376.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., *et al.* (2015) Determinants of community structure in the global plankton interactome. *Science* 348: 1262073.

Logares, R., Deutschmann, I.M., Junger, P.C., Giner, C.R., Krabberød, A.K., Schmidt, T.S.B., *et al.* (2020) Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome* 8: 55.

López-Pérez, M., Haro-Moreno, J.M., Coutinho, F.H., Martinez-Garcia, M., and Rodriguez-Valera, F. (2020) The Evolutionary Success of the Marine Bacterium SAR11 Analyzed through a Metagenomic Perspective. *mSystems* 5: e00605-20.

Luo, H. and Moran, M.A. (2014) Evolutionary Ecology of the Marine Roseobacter Clade. *Microbiol Mol Biol Rev* 78: 573–587.

Macaulay, I.C. and Voet, T. (2014) Single Cell Genomics: Advances and Future Perspectives. *PLoS Genetics* 10: e1004126.

Margalef, R. (1978) Life-forms of phytoplankton as survival alternatives in an unstable environment. *Oceanol Acta* 1: 493–509.

Martin, D.F. and Martin, B.B. (1976) Red tide, red terror. Effects of red tide and related toxins. *J Chem Educ* 53: 614.

Martín, H.G., Ivanova, N., Kunin, V., Warnecke, F., Barry, K.W., McHardy, A.C., *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* 24: 1263–1269.

Martin-Platero, A.M., Cleary, B., Kauffman, K., Preheim, S.P., McGillicuddy, D.J., Alm, E.J., and Polz, M.F. (2018) High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nat Commun* 9: 266.

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

Massana, R., Unrein, F., Rodríguez-Martínez, R., Forn, I., Lefort, T., Pinhassi, J., and Not, F. (2009) Grazing rates and functional diversity of uncultured heterotrophic flagellates. *ISME J* 3: 588–596.

McQueen, D.J., Post, J.R., and Mills, E.L. (1986) Trophic Relationships in Freshwater Pelagic Ecosystems. *Can J Fish Aquat Sci* 43: 1571–1581.

Mestre, M., Höfer, J., Sala, M.M., and Gasol, J.M. (2020) Seasonal Variation of Bacterial Diversity Along the Marine Particulate Matter Continuum. *Front Microbiol* 11: 1590.

Minot, S.S., Barry, K.C., Kasman, C., Golob, J.L., and Willis, A.D. (2021) geneshot: gene-level metagenomics identifies genome islands associated with immunotherapy response. *Genome Biol* 22: 135.

Mizuno, C.M., Rodriguez-Valera, F., and Ghai, R. (2015) Genomes of Planktonic Acidimicrobiales: Widening Horizons for Marine Actinobacteria by Metagenomics. *mBio* 6: e02083-14.

Morris, J.J., Lenski, R.E., and Zinser, E.R. (2012) The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio* 3: e00036-12.

Morris, R.M., Rappé, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A., and Giovannoni, S.J. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* 420: 806–810.

Muñoz-Marín, M.C., Gómez-Baena, G., López-Lozano, A., Moreno-Cabezuelo, J.A., Díez, J., and García-Fernández, J.M. (2020) Mixotrophy in marine picocyanobacteria: use of organic compounds by Prochlorococcus and Synechococcus. *ISME J* 14: 1065–1073.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., *et al.* (2021) A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 39: 499–509.

Needham, D.M. and Fuhrman, J.A. (2016) Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol* 1: 16005.

Niehus, R., Oliveira, N.M., Li, A., Fletcher, A.G., and Foster, K.R. (2021) The evolution of strategy in bacterial warfare via the regulation of bacteriocins and antibiotics. *eLife* 10: e69756.

Noble, R.T. and Fuhrman, J.A. (2000) Rapid virus production and removal as measured with fluorescently labeled viruses as tracers. *Appl Environ Microbiol* 66: 3790–3797.

Pace, N.R., Stahl, D.A., Lane, D.J., and Olsen, G.J. (1986) The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. In *Adv Microb Ecol.* Marshall, K.C. (ed). Boston, MA: Springer US, pp. 1–55.

Pachiadaki, M.G., Brown, J.M., Brown, J., Bezuidt, O., Berube, P.M., Biller, S.J., *et al.* (2019) Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* 179: 1623–1635.

Palenik, B., Brahamsha, B., Larimer, F.W., Land, M., Hauser, L., Chain, P., *et al.* (2003) The genome of a motile marine Synechococcus. *Nature* 424: 1037–1042.

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., and Hugenholtz, P. (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* 38: 1079–1086.

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

Parsons, R.J., Breitbart, M., Lomas, M.W., and Carlson, C.A. (2012) Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *ISME J* 6: 273–284.

Partensky, F. and Garczarek, L. (2010) Prochlorococcus: Advantages and Limits of Minimalism. *Ann Rev Mar Sci* 2: 305–331.

Pedrós-Alió, C. (2012) The rare bacterial biosphere. *Ann Rev Mar Sci* 4: 449–466.

Pinhassi, J., DeLong, E.F., Béjà, O., González, J.M., and Pedrós-Alió, C. (2016) Marine Bacterial and Archaeal Ion-Pumping Rhodopsins: Genetic Diversity, Physiology, and Ecology. *Microbiol Mol Biol Rev* 80: 929–954.

Pinhassi, J., Gómez-Consarnau, L., Alonso-Sáez, L., Sala, M., Vidal, M., Pedrós-Alió, C., and Gasol, J. (2006) Seasonal changes in bacterioplankton nutrient limitation and their effects on bacterial community composition in the NW Mediterranean Sea. *Aquat Microb Ecol* 44: 241–252.

Pomeroy, L.R. (1974) The Ocean's Food Web, A Changing Paradigm. *BioScience* 24: 499–504.

Prosser, J.I. (2020) Putting science back into microbial ecology: a question of approach. Phil Trans R Soc B 375: 20190240.

Rappé, M.S., Connon, S.A., Vergin, K.L., and Giovannoni, S.J. (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* 418: 630–633.

Reysenbach, A.L., Giver, L.J., Wickham, G.S., and Pace, N.R. (1992) Differential amplification of rRNA genes by polymerase chain reaction. *Appl Environ Microbiol* 58: 3417–3418.

Rinke, C., Chuvochina, M., Mussig, A.J., Chaumeil, P.-A., Davín, A.A., Waite, D.W., *et al.* (2021) A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* 6: 946–959.

Rocap, G., Distel, D.L., Waterbury, J.B., and Chisholm, S.W. (2002) Resolution of Prochlorococcus and Synechococcus Ecotypes by Using 16S-23S Ribosomal DNA Internal Transcribed Spacer Sequences. *Appl Environ Microbiol* 68: 1180–1191.

Roda-Garcia, J.J., Haro-Moreno, J.M., Huschet, L.A., Rodriguez-Valera, F., and López-Pérez, M. (2021) Phylogenomics of SAR116 clade reveals two subclades with different evolutionary trajectories and important role in the ocean sulfur cycle, Microbiology.

Rodríguez-Valera, F. (2004) Environmental genomics, the big picture? *FEMS Microbiol Lett* 231: 153–158.

Rosselló-Móra, R. and Amann, R. (2015) Past and future species definitions for Bacteria and Archaea. *Syst Appl Microbiol* 38: 209–216.

Rotthauwe, J.H., Witzel, K.P., and Liesack, W. (1997) The ammonia monooxygenase structural gene *amoA* as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations. *Appl Environ Microbiol* 63: 4704–4712.

Ruiz-González, C., Simó, R., Sommaruga, R., and Gasol, J.M. (2013) Away from darkness: a review on the effects of solar radiation on heterotrophic bacterioplankton activity. *Front Microbiol* 4: 131.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* 5: e77.

Salazar, G., Cornejo-Castillo, F.M., Benítez-Barrios, V., Fraile-Nuez, E., Álvarez-Salgado, X.A., Duarte, C.M., *et al.* (2016) Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J* 10: 596–608.

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., *et al.* (2019) Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* 179: 1068-1083.e21.

Salcher, M.M., Pernthaler, J., and Posch, T. (2011) Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'that rule the waves' (LD12). *ISME J* 5: 1242–1252.

Sánchez, O., Ferrera, I., Mabrito, I., Gazulla, C.R., Sebastián, M., Auladell, A., *et al.* (2020) Seasonal impact of grazing, viral mortality, resource availability and light on the group-specific growth rates of coastal Mediterranean bacterioplankton. *Sci Rep* 10: 19773.

Sánchez, O., Koblížek, M., Gasol, J.M., and Ferrera, I. (2017) Effects of grazing, phosphorus and light on the growth rates of major bacterioplankton taxa in the coastal NW Mediterranean: Growth rates of bacterioplankton. *Environ Microbiol Rep* 9: 300–309.

Santoro, A.E., Richter, R.A., and Dupont, C.L. (2019) Planktonic Marine Archaea. *Ann Rev Mar Sci* 11: 131–158.

Schloss, P.D. and Handelsman, J. (2005) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Appl Environ Microbiol* 71: 1501–1506.

Sebastián, M., Smith, A.F., González, J.M., Fredricks, H.F., Van Mooy, B., Koblížek, M., *et al.* (2016) Lipid remodelling is a widespread strategy in marine heterotrophic bacteria upon phosphorus deficiency. *ISME J* 10: 968–978.

Seymour, J.R., Amin, S.A., Raina, J.-B., and Stocker, R. (2017) Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nat Microbiol* 2: 17065.

Shade, A., Jones, S.E., Caporaso, J.G., Handelsman, J., Knight, R., Fierer, N., and Gilbert, A. (2014) Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio* 5: 1–9.

Shapiro, B.J., Leducq, J.-B., and Mallet, J. (2016) What Is Speciation? *PLoS Genet* 12: e1005860.

Sherr, E.B. and Sherr, B.F. (2002) Significance of predation by protists in aquatic microbial food webs. *ALJMAO* 81: 293–308.

Simó, R., Archer, S.D., Pedrós-Alió, C., Gilpin, L., & Stelfox-Widdicombe, C.E. (2002). Coupled dynamics of dimethylsulfoniopropionate and dimethylsulfide cycling and the microbial food web in surface waters of the North Atlantic. *Limnol Oceanogr*, 47: 53-61.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere." *PNAS* 103: 12115–12120.

Stackebrandt, E. and Goebel, B.M. (1994) Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int J Syst Evol,* 44: 846–849.

Staley, J.T. and Konopka, A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* 39: 321–346.

Stein, J.L., Marsh, T.L., Wu, K.Y., Shizuya, H., and DeLong, E.F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* 178: 591–599.

Su, C., Lei, L., Duan, Y., Zhang, K.Q., and Yang, J. (2012) Culture-independent methods for studying environmental microorganisms: Methods, application, and perspective. *Appl Microbiol Biotech* 93: 993–1003.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., *et al.* (2015) Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348: 1261359.

Teeling, H., Fuchs, B.M., Bennke, C.M., Krüger, K., Chafee, M., Kappelmann, L., *et al.* (2016) Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. *eLife* 5: e11888.

Teira, E., Logares, R., Gutiérrez-Barral, A., Ferrera, I., Varela, M.M., Morán, X.A.G., and Gasol, J.M. (2019) Impact of grazing, resource availability and light on prokaryotic growth and diversity in the oligotrophic surface global ocean. *Environ Microbiol* 21: 1482–1496.

Torsvik, V. and Øvreås, L. (2002) Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol* 5: 240–245.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428: 37–43.

Unrein, F., Massana, R., Alonso-Sáez, L., and Gasol, J.M. (2007) Significant year-round effect of small mixotrophic flagellates on bacterioplankton in an oligotrophic coastal system. *Limnol Oceanogr* 52: 456–469.

Ustick, L.J., Larkin, A.A., Garcia, C.A., Garcia, N.S., Brock, M.L., Lee, J.A., *et al.* (2021) Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science* 372: 287.

Vellend, M. (2010) Conceptual Synthesis in Community Ecology. *Q Rev Biol* 85: 183–206.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66–74.

Vergin, K.L., Beszteri, B., Monier, A., Cameron Thrash, J., Temperton, B., Treusch, A.H., *et al.* (2013) High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J* 7: 1322–1332.

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

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., *et al.* (1987) Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *Int J Syst Evol* 37: 463–464.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: The unseen majority. *PNAS* 95: 6578–6583.

Woese, C.R. (1987) Bacterial evolution. *Microbiol Rev* 51: 221–271.

Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *PNAS* 74: 5088–5090.

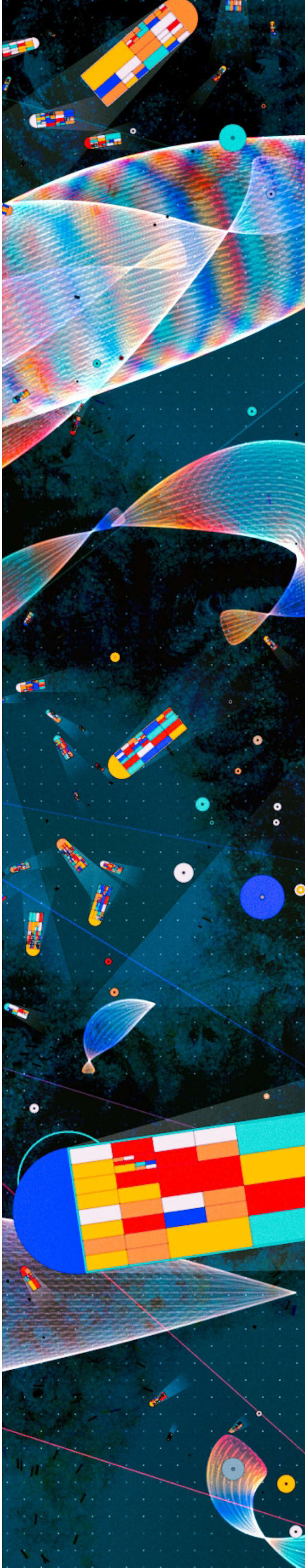
Worden, A.Z., Follows, M.J., Giovannoni, S.J., Wilken, S., Zimmerman, A.E., and Keeling, P.J. (2015) Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science* 347: 1257594.

Yutin, N., Suzuki, M.T., and Béjà, O. (2005) Novel primers reveal wider diversity among marine aerobic anoxygenic phototrophs. *Appl Environ Microbiol* 71: 8958–8962.

Zehr, J.P. and McReynolds, L.A. (1989) Use of degenerate oligonucleotides for amplification of the nifH gene from the marine cyanobacterium Trichodesmium thiebautii. *Appl Environ Microbiol* 55: 2522–2526.

Zohdi, E. and Abbaspour, M. (2019) Harmful algal blooms (red tide): a review of causes, impacts and approaches to monitoring and prediction. *Int J Environ Sci Technol* 16: 1789–1806.

AIMS AND OBJECTIVES



Aims of the thesis

The general aim of this thesis is to understand how marine bacteria and archaea community structure is assembled seasonally through multiple omics approaches applied to a long-term time series and manipulation experiments.

The thesis is arranged in four chapters. In the first chapter (*Seasonal niche differentiation among closely related marine bacteria*, ISME J. 2021), the temporal patterns of the whole community at the Blanes Bay Microbial Observatory in the North Western Mediterranean Sea using the analysis of amplicons of the 16S rRNA gene are presented. Through 11 years of data, we studied the temporal niche distribution among closely related taxa using amplicon sequence variants. The second chapter (*Long-term seasonal and interannual variability of marine aerobic anoxygenic photoheterotrophic bacteria*, ISME J. 2019) focuses on the seasonality of a particular functional group that thrives in the sunlit ocean, the aerobic anoxygenic photoheterotrophic bacteria (AAPs). Although the temporal patterns of AAPs had been studied for a whole year (Ferrera *et al.*, 2014), the interannual changes were unknown. The third chapter (*Seasonality of biogeochemically relevant microbial genes in a coastal ocean microbiome,* unpublished) delineates the seasonal patterns at the whole gene and at the single-variant level of multiple key biogeochemical marker genes using metagenomics. Finally, to bridge the gap between in situ descriptive observations and the top-down and bottom-up factors modulating microbial community structure, the fourth chapter (*Seasonal influence of predation, viral mortality and nutrient limitation in a marine microbiome assessed through metagenome assembled genomes*, unpublished) presents the results of experimental manipulations combined with metagenomics performed at different seasons. By these means, the seasonal influence of predation, viruses, light and nutrient limitation was studied, focusing on the response of metagenome assembled genomes and the importance of the genomic repertoire in this selection.

Objectives of the thesis

Objective 1. Characterize the seasonality of the bacterial community structure differentiating closely related taxa.

- Quantify how many ASVs present seasonality and what is the temporal distribution of specific taxonomic groups.
- Evaluate how similar the temporal niche between closely related taxa within the same genus is.
- Extend this comparison to broader taxonomic levels, testing how conserved seasonality is, as we move from genus to higher taxonomic levels (i.e., family, order, and class).

Objective 2. Explore the temporal patterns of aerobic anoxygenic photoheterotrophic bacteria (AAPs).

- Explore long-term seasonality among the different AAPs phylogroups and identify the main environmental drivers of this functional group.
- Explore whether the phylotypes of this functional group are ecologically cohesive, or contrarily, there is temporal niche differentiation within each phylogroup.
- Compare the results from amplicon sequences to metagenomic approaches to understand the biases of the PCR approach.

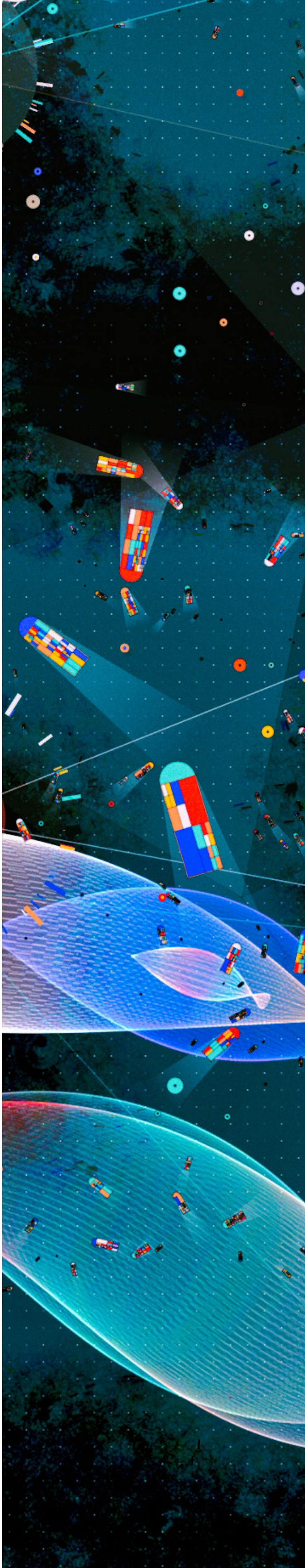
Objective 3. Determine the temporal patterns of key biogeochemical functional genes.

- Find whether a suite of biogeochemically relevant functional genes are seasonal, and if so, determine in which season each function is likely to prevail in the system.
- Obtain a detailed picture of the main taxonomic groups involved in each function at each season.
- Explore the seasonality of each gene at the single-variant level, inspecting whether the distribution of some functions change among genera within the same taxonomic family.

Objective 4. Assess how individual species respond seasonally to top-down and bottom-up factors.

- Test whether particular species are enriched under particular conditions: predation removal, viral mortality reduction, manipulation of nutrient availability, and light availability.
- Establish links between the genetic repertoire of the selected organisms and the treatments in which they dominate.
- Explore the presence and distribution of strains through population genomics to evaluate how deterministic growth within each species is.

CHAPTER I



Chapter I

Seasonal niche differentiation among closely related marine bacteria

Adrià Auladell, Albert Barberán, Ramiro Logares, Esther Garcés, Josep M. Gasol, Isabel Ferrera

ISME Journal (2021). DOI: 10.1038/s41396-021-01053-2

Abstract

Bacteria display dynamic abundance fluctuations over time in marine environments, where they play key biogeochemical roles. Here, we characterized the seasonal dynamics of marine bacteria in a coastal oligotrophic time series station, tested how similar the temporal niche of closely related taxa is, and what are the environmental parameters modulating their seasonal abundance patterns. We further explored how conserved the niche is at higher taxonomic levels. The community presented recurrent patterns of seasonality for 297 out of 6825 amplicon sequence variants (ASVs), which constituted almost half of the total relative abundance (47%). For certain genera, niche similarity decreased as nucleotide divergence in the 16S rRNA gene increased, a pattern compatible with the selection of similar taxa through environmental filtering. Additionally, we observed evidence of seasonal differentiation within various genera as seen by the distinct seasonal patterns of closely related taxa. At broader taxonomic levels, coherent seasonal trends did not exist at the class level, while the order and family ranks depended on the patterns that existed at the genus level. This study identifies the coexistence of closely related taxa for some bacterial groups and seasonal differentiation for others in a coastal marine environment subjected to a strong seasonality.

1.1 Introduction

Marine microbial communities display dynamic abundance fluctuations over time, particularly in temperate coastal environments. Community structure changes on a daily, monthly, and annual scale due to the variation of bottom-up factors such as resource availability (including inorganic nutrients and dissolved organic carbon), top-down biotic interactions, and physical properties such as temperature, day length or the presence of eddies and upwelling events (Fuhrman *et al.*, 2015). Given that microbes are key players in the functioning of the biosphere, defining seasonality and understanding how taxa respond to changes in environmental conditions is crucial (Falkowski, 2012).

The establishment of microbial observatories across the globe in combination with the advances in sequencing methodologies has allowed the monitoring of microbial communities over time, from short- to long-term scales (see reviews by Bunse and Pinhassi, 2017; Buttigieg *et al.*, 2018). Various studies have shown remarkably repeatable seasonal patterns in the distribution and abundance of microbial taxa (i.e. Fuhrman *et al.*, 2015; Eiler *et al.*, 2011; Gilbert *et al.*, 2012; Cram *et al.*, 2015; Giner *et al.*, 2019), including those in the rare biosphere (Alonso-Sáez *et al.*, 2015) and despite irregular environmental perturbations (Lambert *et al.*, 2018). Further, investigating the dynamics of individual taxa –or finely resolved taxonomic units– on the short-term scale has revealed sharp turnover of communities mirroring environmental variability (Martin-Platero *et al.*, 2018) and the relevance of interactions among microorganisms, influenced by the dynamics of phytoplankton blooms (Needham and Fuhrman, 2016; Needham *et al.*, 2018). On longer time scales, these high-resolution analyses have shown recurrent co-varying taxa (modules) regardless of interannual variation in phytoplankton blooms (Chafee *et al.*, 2018) or a clear partitioning of modules of oligotrophs and copiotrophs over time (Lemonnier *et al.*, 2020). Nevertheless, these patterns of module covariance can be lost under contrasting environmental conditions, as shown by a recent study (Lambert *et al.*, 2021). In addition, the analysis of closely related populations of photoheterotrophic bacteria has shown that closely related amplicon sequence variants (ASVs) could represent distinct ecotypes occupying temporally different niches (Auladell *et al.*, 2019). What is still missing is an in-depth study exploring the degree of niche similarity among closely related marine bacteria and how conserved the niche is at higher taxonomic levels.

Hutchinson proposed that an 'n-dimensional hypervolume' could define the niche of a species: a set of conditions under which an organism can survive and reproduce (Hutchinson, 1957). Together with abiotic parameters, biotic interactions such as mutualism, cross-feeding, and competition delineate the realized niche of taxa (Cordero and Polz, 2014; Hammarlund *et al.*, 2021). The niche is determined both by homogeneous selection of traits to survive in a specific environment and heterogeneous selection for other traits to reduce competition that would facilitate coexistence (Cordero and Polz, 2014). In bacteria, genomic adaptations can come from horizontal gene transfer, gene polymorphisms, and other mutations mediated by these evolutionary selective processes. The analysis of these processes and how they impact the niche distribution is limited by the taxonomic resolution of the methodology used. Metagenomics has shown how multiple *Prochlorococcus* subpopulations with a distinctive set of flexible genes can temporally coexist (Kashtan *et al.*, 2014), and has also uncovered a large amount of diversity within the SAR11 clade (Haro-Moreno *et al.*, 2020). Although metagenomic data provide highly resolved taxonomic information, the technique is financially and computationally costly, which complicates scaling the analysis of these processes to the full community (Schloss, 2020). On the contrary, 16S rRNA gene amplicon sequencing is a cheap and efficient approach for broad community analyses. The limitation of this technique is however the genetic resolution of the 16S rRNA gene hypervariable regions, in the range of species delineation, yet easily allowing the genus differentiation (Johnson *et al.*, 2019; VanInsberghe *et al.*, 2020). Coupled with time series studies of marine microbial observatories, this approach can thus inform on whether ecological distributions are shared within organisms at the sub-genus level (therein 'closely related taxa', Tromas *et al.*, 2018). Furthermore, it allows to extend this comparison to broader taxonomical groups and obtain insights into the 'phylogenetic scale' at which ecology presents coherence (Philippot *et al.*, 2010; Martiny *et al.*, 2015; Ladau and Eloe-Fadrosh, 2019).

Here we used a monthly sampled time-series spanning 11 years from a coastal marine observatory in the North-Western Mediterranean Sea to explore the long-term seasonal trends in bacterioplankton communities. First, we evaluated how similar the temporal niche is between ASVs within the same genus, and later extended the comparison to broader taxonomic levels in order to answer the following questions: (1) how many ASVs are seasonal and what is the temporal distribution of distinct taxonomic groups, (2) how similar the niche among closely related ASVs within different marine genera is and what are the environmental parameters modulating their distinct ecological responses, and (3) how conserved the realized niche is as we go from genus to higher taxonomic levels (i.e., family, order and class).

1.2 Material and methods

Location and sample collection.

Samples were collected from the Blanes Bay Microbial Observatory (BBMO), a station located in the NW Mediterranean Sea about 1 km offshore over a water column of 20 m depth (41º40'N, 2º48'E) (Gasol *et al.*, 2016). Sampling was conducted monthly over 11 years (January 2003 to December 2013). Water temperature and salinity were measured in situ with a conductivity, temperature and depth probe, and light penetration was estimated using a Secchi disk. Surface seawater was pre-filtered through a 200 µm nylon mesh, transported to the laboratory under dim light in 20 L plastic carboys, and processed within 2 h. Chlorophyll *a* concentration was measured on GF/F filters extracted with acetone and processed by fluorometry (Yentsch and Menzel, 1963). The concentrations of inorganic nutrients (NO_3^-, NO_2^-, NH_4+, PO_4^{3-}, SiO_2) were determined spectrophotometrically using an Alliance Evolution II autoanalyzer (Grasshoff *et al.*, 1983). The abundances of picocyanobacteria, heterotrophic bacteria, and photosynthetic pico- and nanoeukaryotes were determined by flow cytometry as described elsewhere (Gasol and Morán, 2016). Additionally, the abundance of photosynthetic and heterotrophic flagellates of different size ranges was measured by epifluorescence microscopy on 0.6 µm polycarbonate filters stained with 4',6-diamidino-2-phenylindole. Microbial biomass was collected by filtering about 4 L of seawater using a peristaltic pump sequentially through a 20 µm nylon mesh (to remove large eukaryotes), a 3 µm pore-size 47 mm polycarbonate filter, and a 0.2 µm pore-size Sterivex unit (Millipore).

DNA extraction, PCR amplification, and sequencing.

DNA was extracted from the Sterivex unit (0.2 to 3 µm fraction of bacterioplankton) as described in (Massana *et al.*, 1997), purified and concentrated in an Amicon 100 (Millipore) and quantified in a NanoDrop-1000 spectrophotometer (Thermo Scientific). DNA was stored at -80ºC and an aliquot from each sample was used for sequencing using a MiSeq sequencer (2 × 250 bp, Illumina) at the Research and Testing Laboratory (Lubbock, TX, USA; http://rtlgenomics.com/). Primers 341F (5'-CCTACGGGNGGCWGCAG-3') (Herlemann *et al.*, 2011) and 806RB (5'-GGACTACNVGGGTWTC-TAAT-3') (Apprill *et al.*, 2015) were used to amplify the V3-V4 regions of the 16S rRNA gene. A total of 131 samples were successfully sequenced and used in subsequent analyses.

Sequence processing.

DADA2 v1.12 was used to differentiate the partial 16S rRNA gene amplicon sequence variants (ASVs) and to remove chimeras (Callahan *et al.*, 2016). Previously, spurious sequences and primers were trimmed using *cutadapt* v.1.16 (Martin, 2011). Taxonomic assignment of the ASVs was performed with IDTAXA from *DECIPHER* v2.14 package (Wright, 2016) against the Genome Taxonomy Database (GTDB) r89 (Parks *et al.*, 2018). The GTDB has the advantage that incorporates new data from metagenomic assembled genomes (MAGs) and generates phylogenies based on 120 single-copy genes. Additionally, SILVA r138 taxonomy was used for nomenclature correspondence (see ASVs taxonomy in Supplementary Table 1 and the correspondence between databases in Supplementary Table 2) (Quast *et al.*, 2013). Compared to SILVA, GTDB allowed an increase in the assignation at the genus rank (14.6% more sequences) and the differentiation of new groups (e.g. D2472 genus within SAR86). Furthermore, the ASVs assigned to *Synechococcus* were checked against the Cyanorak database v2.1 (Garczarek *et al.*, 2020) through 100% BLAST matches. ASVs classified as Mitochondria or Chloroplast were removed. ASV sequences were also clustered into OTUs (Operational Taxonomic Units) at 99% identity –a typical threshold for the delineation of OTUs in microbiome studies– for comparison purposes. Clustering was performed by calculating the nucleotide sequence distance matrix using the *DECIPHER* package. This matrix was also used to calculate the nucleotide divergence among ASVs.

Community data analyses.

We performed all analyses with the R v3.5 language (R Core Team, 2014). For data processing we used the *phyloseq* v1.26 and *tidyverse* v1.3 packages (McMurdie and Holmes, 2013; Wickham *et al.*, 2019), and *ggplot2* v3.2 for visualization (Wickham, 2016). We defined abundant taxa as those above or equal to 1% relative abundance in at least one sample (Campbell *et al.*, 2011). An ASV always below that cutoff was considered permanently rare. For both abundance groups, we defined three ASV categories based on occurrence: broad (≥75% occurrence), intermediate (>10% and <75% samples), and narrow (≤10% samples) distribution, as termed in Chafee *et al.* (2018). Abundant ASVs were further tested as Conditionally Rare Taxa (CRT), taxa typically in low abundance that occasionally become prevalent (bimodality = 0.9, relative abundance ≥ 1%, Shade *et al.*, 2014).

To estimate alpha diversity and beta diversity we used the *breakaway* v4.6 and *divnet* v0.34 packages, respectively (Willis *et al.*, 2017; Willis and Martin, 2020). These approaches avoid common pitfalls from applying classical ecology indexes (i.e. Chao1, Shannon) to microbiome data, such as the influence of sequencing depth and data compositionality.

Seasonality data analysis.

To test whether each of the ASVs displayed seasonality –that is, recurrent changes over time– we used the Lomb Scargle Periodogram (LSP) as implemented in the *lomb* package v1.2 (Ruf, 1999). The method has previously been used for testing the seasonality of marine microbial communities (Lambert *et al.*, 2018). The LSP determines the spectrum of frequencies composing the dataset. Afterwards, significance is tested through data randomizations (q ≤ 0.05, False Discovery Rate (FDR) correction). For each ASV, we obtain the density distribution for each of the periods and the peak normalized power (PN). The distribution shows which is the most recurrent period and the PN value measures the strength of this period. We considered the results as seasonal only if PN was above 10 and q ≤ 0.05, as in Lambert *et al.* (2018). The non-seasonal fraction is thus comprised of 1) truly non-seasonal ASVs, and 2) seasonal ASVs with no recurrent signal detected likely due to a limitation in our sequencing depth. In addition to the ASV level, we evaluated the seasonality at the class, order, family, and genus ranks. For a specific rank group (e.g. class Alphaproteobacteria), 80% of the ASVs were chosen randomly, aggregated, and the LSP was calculated (using 300 iterations). Out of the 29 classes present in the dataset, only the Alphaproteobacteria, Gammaproteobacteria and Bacteroidia could be evaluated since these were the classes that presented more than one order, family, and genus ranks with at least 10 ASVs.

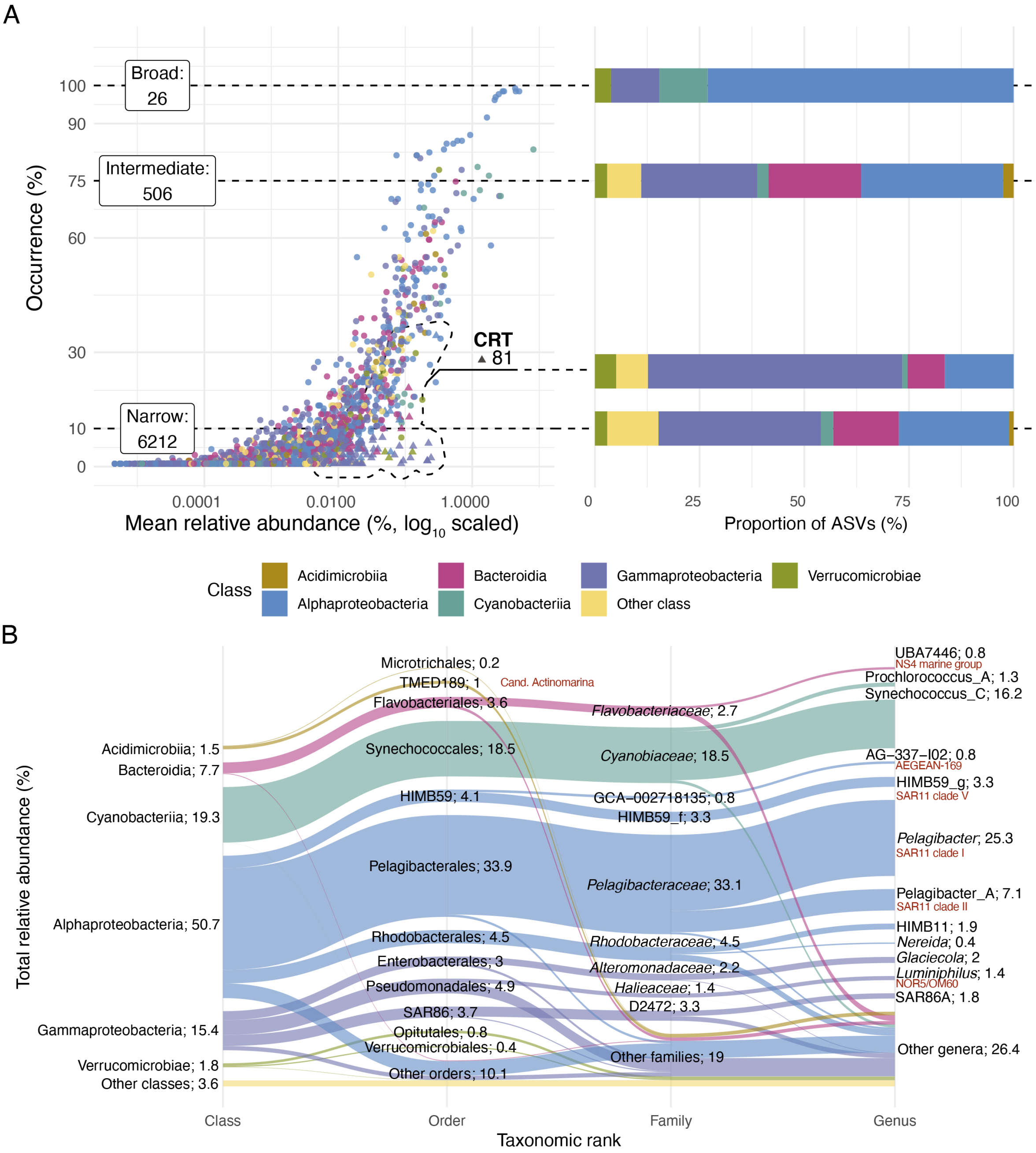


Figure 1: A) Distribution of the different ASV types (broad, narrow or intermediate, and conditionally rare taxa, CRT). The Y axis indicates the occurrence (% of samples) and the X axis corresponds to the mean relative abundance (%) over the time series. Dotted lines delimitate the distributions (the numbers of ASVs of each type are displayed in the label) and connect to a box indicating the number of ASVs for each distribution and a bar plot colored by taxonomy at the class rank. CRT taxa are following a bimodal distribution and present ≥1% relative abundance in at least one sample. B) Alluvial plot showing the total relative abundance distribution of Blanes Bay taxa across different taxonomic ranks (class, order, family and genus). The height of the sections displays the relative abundance (indicated in the text; the total is 100%). The SILVA nomenclature is displayed in red next to the corresponding GTDB database nomenclature.

Further, we tested how the ASVs clustered based on the seasonal abundance patterns. First, we checked the number of possible clusters through the gap statistic from the *cluster* v2.1 package, since the expected number of clusters is unknown beforehand (Tibshirani *et al.*, 2001). Afterwards, we clustered the data through hierarchical clustering. To visually compare the trend of the various seasonal ASVs, each one was fitted through a generalized additive model (GAM) using the *mgcv* v1.8 package (Hastie and Tibshirani, 1986). The Centered Logarithm Ratio values (CLR, adding a pseudocount of 1) were fitted along day of the year, allowing a smoothing parameter with 12 knots (the maximum number of curves, being 12 for the number of months, Pedersen *et al.*, 2019).

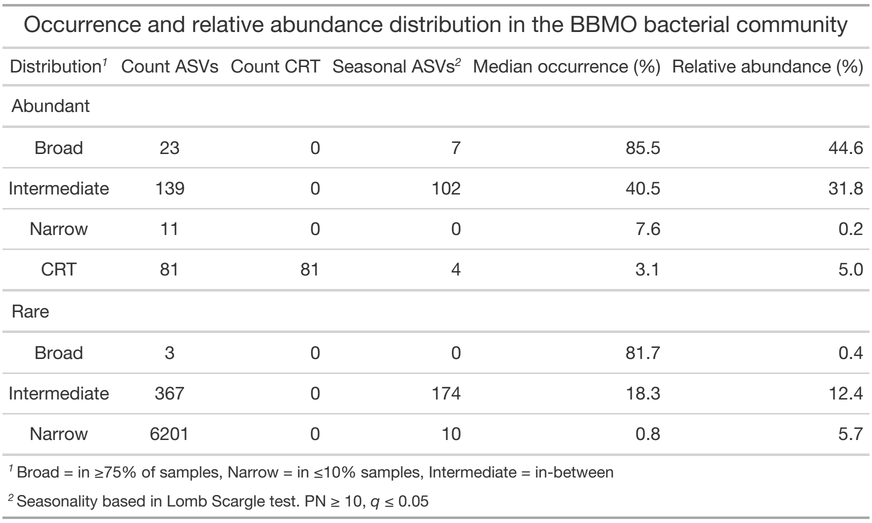
Analyses of niche preference and environmental drivers.

To examine if taxa within a given genus covary and, therefore, could share a realized temporal niche, we used the *propr* v4.2 package (Quinn *et al.*, 2017). This package avoids the common pitfalls of compositional data analyzing correlation-like measurements. The raw counts are transformed to ratios, usually between the abundance of the taxon of interest and the geometric mean of all taxa for a specific sample. Then, for all the ratios of taxa A and taxa B, we measure the proportionality of change, Rho, with similar properties to a correlation measurement (see Lovell *et al.* (2015) for a detailed explanation). The results are then filtered with a final estimate of 5% of FDR. Within each genus, we compared the Rho value between pairs of ASVs –acting as a proxy of niche similarity– against the nucleotide divergence among ASVs to see whether there were trends in niche relatedness. A linear model was used to test which genera presented significant relationships ($p \leq 0.05$) between nucleotide divergence and Rho. We analyzed the genera with at least 10 closely related ASVs (at a maximum of 5 nucleotide divergence), which resulted in a total of 8 genera (out of 581). For most of these groups, using the V3 and V4 hypervariable regions of the 16S rRNA gene, 5 nucleotide divergence equals to a median sequence identity of 98.8% between two pairs. This nucleotide distance is the threshold that we used for considering two ASVs as closely related.

Finally, we tested which measured environmental parameters drive the patterns among closely related taxa. From the suite of measured variables, we selected temperature, chlorophyll *a* concentration, inorganic nutrient concentrations, and the abundance of photosynthetic nanoflagellates (PNF) and heterotrophic nanoflagellates (HNF). This selection was based on the expected relevance in modulating the ASV response (bottom up and top-down processes) and also considering the number of missing values in the dataset. Multicollinearity between the parameters was tested using the *HH* v3.1 package, showing no collinearity (Heiberger, 2020). To model each ASV across the different parameters, we used the *corncob* v0.1 package (Martin *et al.*, 2020) (FDR ≤ 5%). Afterwards, a display of the results was created with the GAM approach.

Reproducibility.

All the code including the parameters used for each package is available in the following repository: https://github.com/adriaaulaICM/bbmo_niche_sea. Sequence data have been deposited in the European Nucleotide Archive under project number PRJEB38773.



Occurrence and relative abundance distribution in the BBMO bacterial community

Distribution[1]	Count ASVs	Count CRT	Seasonal ASVs[2]	Median occurrence (%)	Relative abundance (%)
Abundant					
Broad	23	0	7	85.5	44.6
Intermediate	139	0	102	40.5	31.8
Narrow	11	0	0	7.6	0.2
CRT	81	81	4	3.1	5.0
Rare					
Broad	3	0	0	81.7	0.4
Intermediate	367	0	174	18.3	12.4
Narrow	6201	0	10	0.8	5.7

[1] Broad = in ≥75% of samples, Narrow = in ≤10% samples, Intermediate = in-between

[2] Seasonality based in Lomb Scargle test. PN ≥ 10, $q \leq 0.05$

Table 1: Distribution, occurrence and relative abundance of the amplicon sequence variants (ASVs) in the Blanes Bay Microbial Observatory dataset. Distribution indicates the occurrence category: broad (≥75% samples), narrow (≤10% samples) and intermediate. The results are distributed between abundant (≥1% in at least one sample) and rare ASVs. Count ASVs stands for the number of ASVs within each category; Count CRT, the number of Conditionally Rare Taxa; seasonal ASVs, the count of seasonal ASVs (based in the Lomb Scargle test, $q \leq 0.05$, PN ≥ 10); median occurrence, the % of samples in which the ASVs appear; Relative abundance, the total relative abundance of each category.

1.3 Results

Environmental, ecological and taxonomic context.

Surface water temperature at Blanes Bay varied seasonally, with minimal mean values in February (12.6°C) and maximal values in August (24.5°C, Supplementary Figure 1). Inorganic nutrients were higher during autumn and winter while chlorophyll *a* reached the highest values (ca. 1 mg·m-^{3}) during the winter-spring transition. For a detailed description of the seasonality at Blanes Bay, including these and other environmental parameters, see Gasol *et al.* (2016).

We detected a total of 6,825 ASVs in the 11 years of monthly data. The ASV distribution was compared by occurrence (narrow: ≤10% occurrence; intermediate: >10% and <75%; and broad: ≥75%) and abundance (abundant or rare, i.e. <1% in all samples). Most of the ASVs (91%) displayed a narrow distribution (Figure 1A, Table 1). Only 26 ASVs displayed a broad distribution, of which 3 always belonged to the rare fraction. Taxonomically, 19 of the broad ASVs belonged to the Alphaproteobacteria, mostly to the Pelagibacterales (13 ASVs) and HIMB59 (4 ASVs; former SAR11 clade V) orders. The 506 ASVs presenting an intermediate occurrence belonged to 20 different classes. The dominant classes for this category were the Alphaproteobacteria and Gammaproteobacteria (163 and 133 ASVs respectively) followed by the Bacteroidia (106 ASVs), mostly by the Flavobacteriales order (91 ASVs; Figure 1A). We also evaluated if rare ASVs occasionally became abundant (Conditionally Rare Taxa, CRT) and found a total of 81 ASVs. Gammaproteobacteria (48 ASVs) and Alphaproteobacteria (13) were the most common CRTs, while the rest belonged to the Verrucomicrobiae and Bacteroidia classes (Figure 1B).

In terms of alpha diversity, spring and summer displayed lower values than autumn and winter (α richness estimates = 197 vs 334 ASVs respectively, $p \leq 0.01$; Supplementary Figure 2). Using January as intercept, we observed a significant decrease in richness in April (232 ASVs, $p = 0.015$) to regain higher values in October (316 ASVs, $p = 0.87$). Regarding community similarity (i.e. beta diversity), summer and winter displayed the maximum dissimilarity (β Bray Curtis estimate = 0.48, standard error = 0.036), while autumn and spring presented the lowest difference (β estimate = 0.21, standard error = 0.047; Supplementary Figure 3), with similar ranges for all the other comparisons.

ASV seasonality.

A total of 297 ASVs out of 6825 were seasonal (Lomb Scargle Periodogram test $q \leq 0.05$, PN ≥ 10) covering different ranges of occurrence and season maxima. These seasonal ASVs represented on average 47% of the read relative abundance, partitioned in 13% from ASVs exhibiting broad distribution, 34% of intermediate occurrence, and 0.1% of narrow presence. In our study, peak normalized power values –a statistic measuring how strong the recurrence is– ranged between 10 and 43.1. The highest values corresponded to ASVs with distributions that recurrently presented a peak in one particular season, often winter. ASV122, ASV55, and ASV131, belonging to the Acidimicrobiia, Bacteroidia, and Alphaproteobacteria classes respectively, are examples of this pattern (Supplementary Figure 4).

Within the seasonal ASVs, we differentiated 3 significantly different clusters (Supplementary Figure 5). The first group, composed of 23 ASVs, includes most of the broadly distributed ASVs that peaked during summer and autumn. Taxonomically, this cluster was mostly composed of *Cyanobiaceae* and *Flavobacteriaceae* ASVs. The second cluster, of 30 ASVs, includes ASVs that peaked during winter and spring, mainly belonging to *Pelagibacteraceae*. Interestingly, this cluster includes the

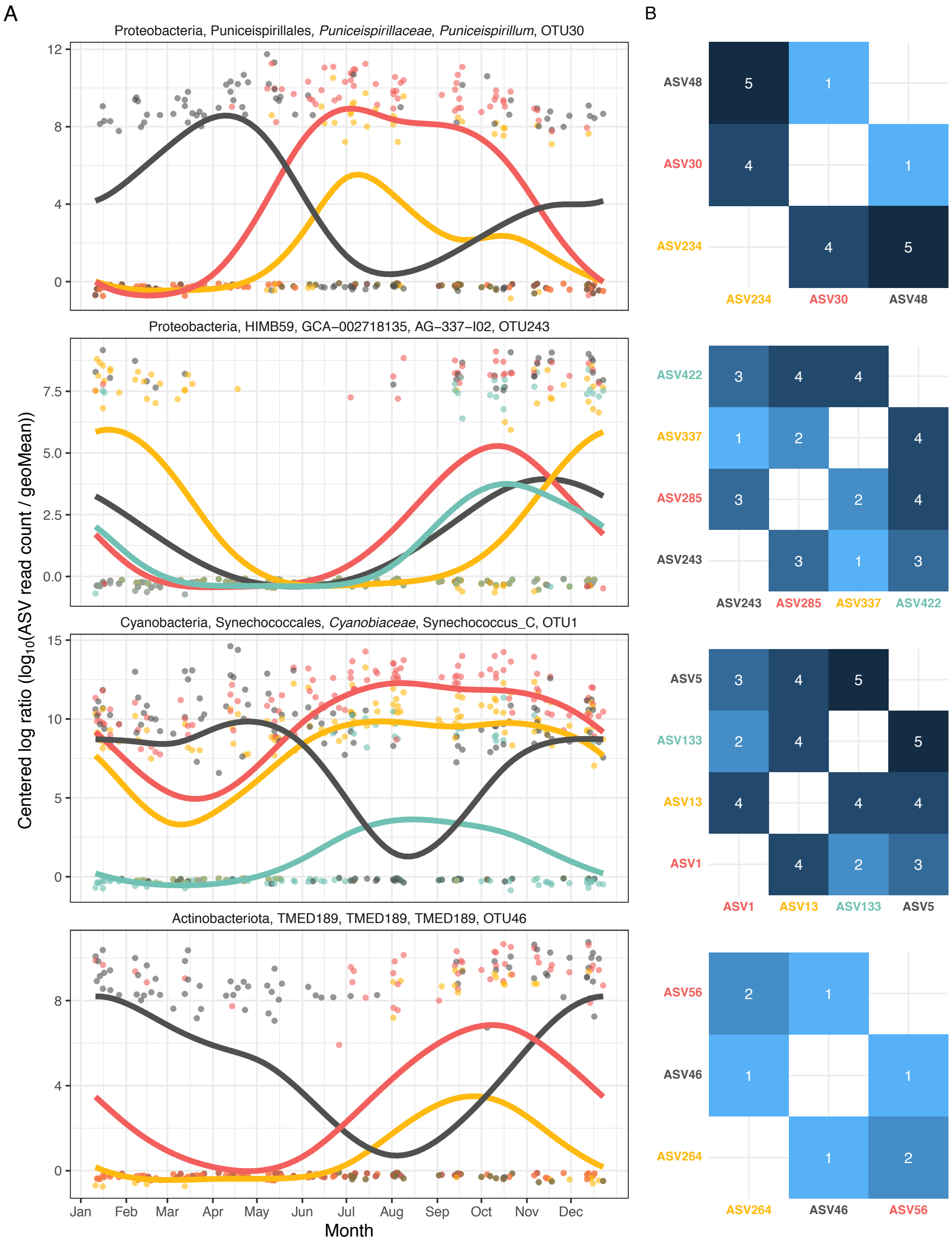


Figure 2: A) Examples of seasonal differentiation among closely related ASVs conforming the same OTU at 99% clustering. The X axis presents the month and the Y axis presents the centered logarithm ratio abundance. A generalized additive model smooth is adjusted to the data points. B) Heatmaps presenting the nucleotide divergence between each ASV pair (number of mismatches after alignment). Five nucleotide divergence equals to a median sequence identity of 98.8%.

understudied group *Marinisoma* that displayed a winter trend in all its seasonal ASVs (5 out of 9 ASVs). Finally, the last cluster was composed of 244 ASVs without a clear seasonal pattern, likely due to their lower occurrence and relative abundance, without the dominance of a particular taxonomic group.

In order to compare the seasonal trend of closely related taxa and investigate how frequent the presence of differentiated seasonal patterns at high sequence similarity is, we checked the ASVs that clustered at 99% similarity. We found 42 OTUs with ASVs presenting multiple ecological patterns. For example, *Pelagibacter* was represented by 20 different OTUs; 3 of them were composed only of seasonal ASVs, 6 OTUs contained both seasonal and non-seasonal ASVs, and 11 OTUs consisted only of non-seasonal ASVs. Similar trends were observed for other genera such as SAR86A and *Luminiphilus*. In general, we found that seasonal differentiation was not common, since only 20% of the OTUs contained ASVs with a clear difference. In total 8 ASVs displayed such behavior, that is, seasonal ASVs within 5 nucleotide mismatches presenting relative abundances with distinct temporal patterns (Figure 2). Most of these patterns could be classified into either an almost complete temporal separation (e.g. ASV48 vs ASV30 within OTU30, affiliated to Puniceispirillales; Figure 2) or a "restriction" of the temporal niche (one of the ASVs is only present in a specific month or season although the other is also present; e.g. ASV285 vs ASV337 within OTU243, affiliated to HIMB59). In fact, seven out of these 8 ASVs displayed the latter pattern of seasonal restriction.

Variability of niche preference within genera.

Here we define the ecological niche of a given taxon as the set of environmental conditions that fluctuate in this marine temperate coastal environment and that allows the growth of the microorganism or its persistence. Cooccurrence and covariance point to a possible niche similarity or mutualism. In our analysis, centered at variability within a genus, our proxy to test for niche overlap among closely related taxa is the Rho measurement (proportional change between two taxa), which can be expressed as a function of the nucleotide divergence between two sequences. A decrease in Rho as nucleotide distance increases denotes that the two taxa decrease their covariance, behaving less similarly as they become more phylogenetically distinct.

Out of the 13 evaluated genera, we found that *Pelagibacter* (Alphaproteobacteria, SAR11 clade I), Pelagibacter_A (Alphaproteobacteria, SAR11 clade II), and less clearly SAR86A (a subclade of SAR86, Gammaproteobacteria) displayed a significant decrease in Rho proportionality when increasing nucleotide divergence (Figure 3; Supplementary Table 3). The distributions within each genus were highly variable. *Pelagibacter* displayed the highest number of ASVs (60) and the variation in the Rho score was likewise the highest, between 0.3 and 0.996. Pelagibacter_A presented fewer ASVs (26) than *Pelagibacter* but a similar Rho distribution. SAR86A had a smaller amount of variation along with the nucleotide change, with a maximum Rho of 0.85. The *Synechococcus* genus (9 ASVs) displayed similarly high proportionality values at low and high nucleotide distances, not showing a decreasing trend. Merging all the non-significant genera, the values did not present a significant tendency (data not shown), suggesting that the decrease is specific to some groups.

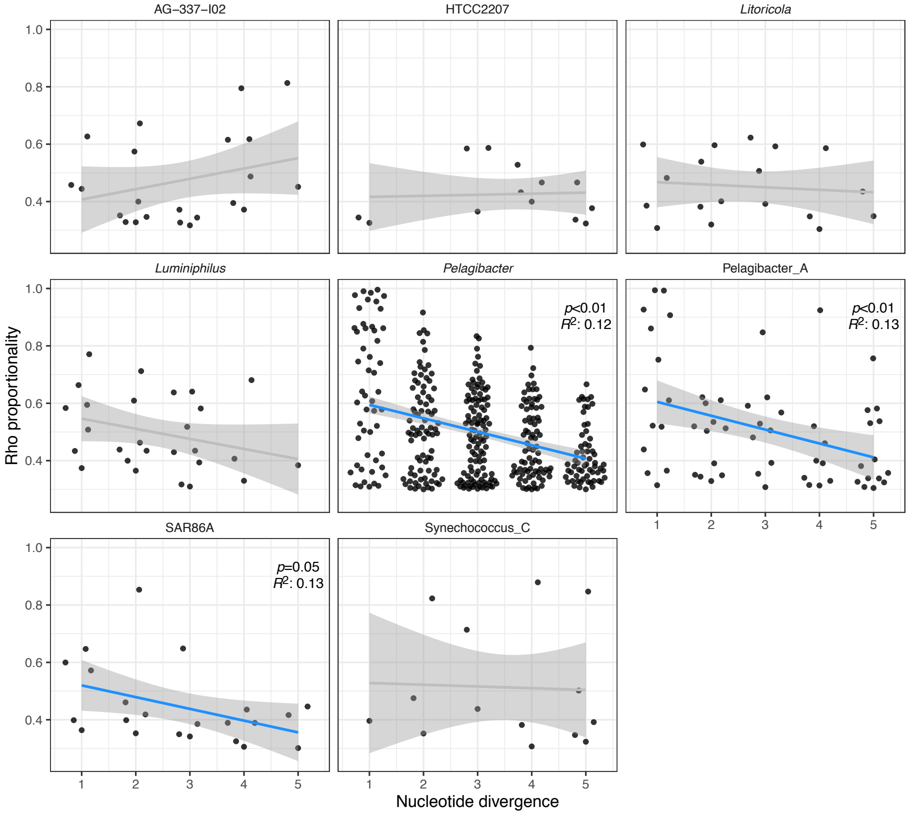


Figure 3: Relationship between the proportionality of change (Rho, Y axis) and the nucleotide divergence (mismatches after alignment, X axis). Only genera with more than 3 ASVs at less than 5 nucleotide divergences were evaluated. Grey and blue lines represent the linear relationship between the two variables (blue indicates statistical significance). The *p* value and the R^2 are displayed for the significant regressions. See Supplementary Table 2 for the correspondence between GTDB and SILVA nomenclature.

Environmental drivers of the observed niche differences within genera.

Given the identified differences in the temporal niche among taxa, we evaluated how different environmental parameters influenced these distributions. For each ASV-parameter pair, we generated a model and the estimated coefficient indicating how the ASV responded (increase or decrease in abundance). A total of 245 out of the 603 response models were significant (FDR ≤ 0.05; Figure 4, Supplementary Figure 6). About two-thirds of the models were polynomial while the rest were linear. Temperature, nitrite, and nitrate concentrations were the parameters appearing most often, followed by the abundance of photosynthetic and heterotrophic nanoflagellates. The different bacterial genera responded divergently to the environmental parameters. *Pelagibacter*,

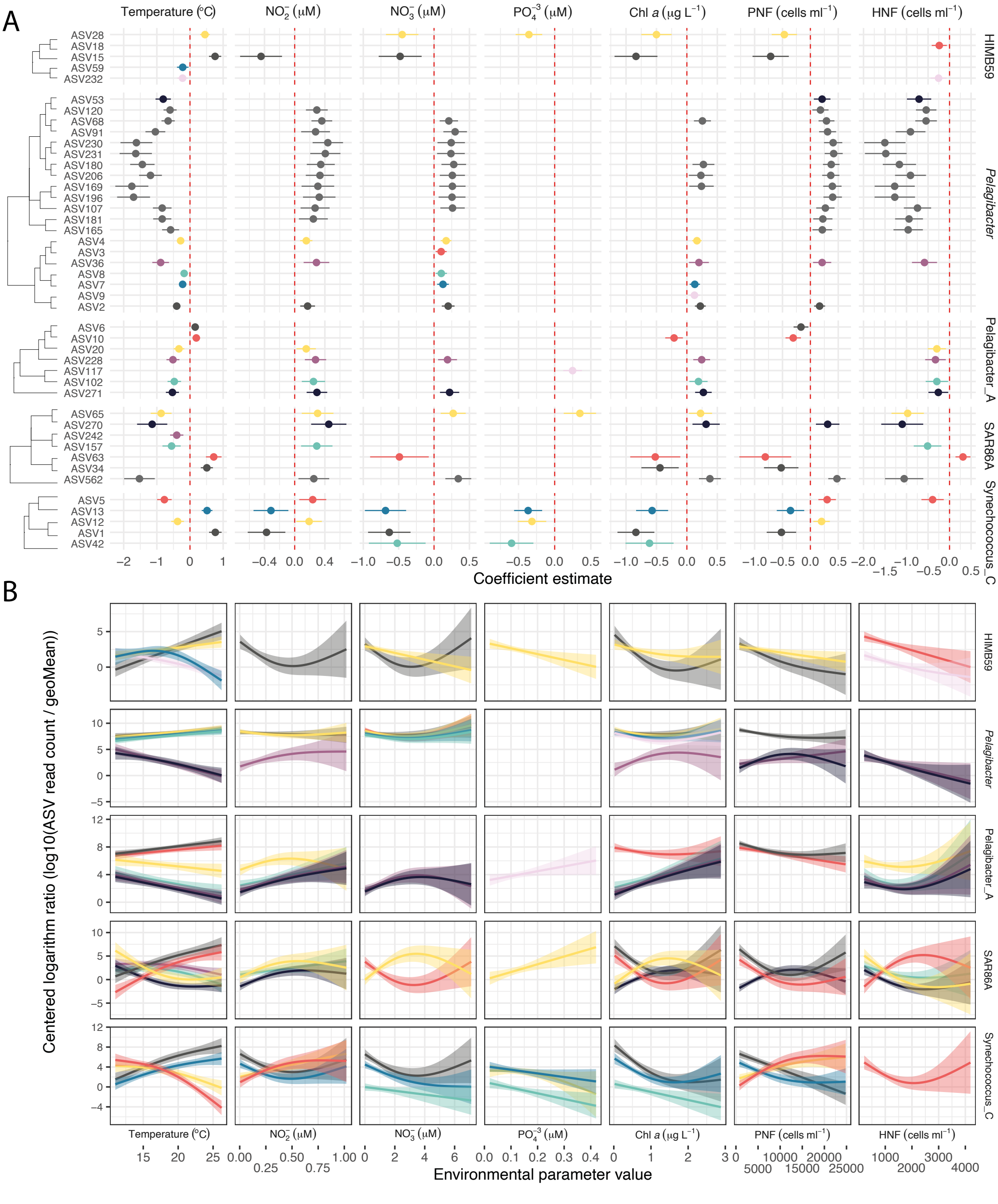


Figure 4: A) Significant models among ASVs from HIMB59, *Pelagibacter*, Pelagibacter_A, SAR86 and *Synechococcus* genera (rows) and various environmental parameters (columns). The coefficient estimate indicates positive or negative responses to the parameter and is shown with a 95% confidence interval. The color corresponds to the different ASVs within a genus (only the top 8 more abundant ASVs are colored, the other ASVs are shown in grey). ASVs are ordered through a hierarchical clustering based on nucleotide divergence. B) Generalized additive model fits between the ASV centered logarithm ratio abundances and the parameter value distribution for the significant ASVs in the upper plot. Panels and ASV colors shown as in (A). PNF: Phototrophic nanoflagellates; HNF: Heterotrophic nanoflagellates.

AG-337-I02 (AEGEAN-169 marine group), D2472 (SAR86), and *Luminiphilus* had ASVs that responded cohesively to a given parameter, displaying the same response sign (Supplementary Figure 6). Most of these bacterial genera showed a negative relative abundance response to temperature and a positive relationship with the concentration of inorganic nitrogen compounds. The exception to this trend was *Luminiphilus*, showing the opposite coefficient sign for all parameters. HIMB59 (former SAR11 clade V), Pelagibacter_A, SAR86A, and *Synechococcus* showed differences in the ecological patterns within each genus (Figure 4A). Within SAR86A, two contrasting patterns could be observed; ASV34 and ASV63 (nucleotide divergence of 1; Supplementary Figure 7) presented a positive relationship to temperature and a negative one to nitrate and chlorophyll *a* concentration, while ASV562, ASV270, ASV65, and ASV157 presented the opposite responses (these ASVs had nucleotide distances ranging from 1 to 9; Figure 4A). In the case of *Synechococcus*, a similar trend was observed (ASV5 and ASV12 vs. ASV1 and ASV13, Figure 4) but the phylogenetic distance does not hint to a possible explanation, as seen in the previous section (Figure 3). Between ASV1 and ASV5 there was only a 3-nucleotide divergence (99.26% identity), but their seasonality was clearly different (Supplementary Figure 8). We checked the *Synechococcus* ASVs taxonomy at a finer resolution using a picocyanobacterial-specific database, Cyanorak (Garczarek *et al.*, 2020). In particular, ASV5 presented a 100% identity match with strain PROS-9-1 belonging to Clade Ib, found in cold or temperate waters (Farrant *et al.*, 2016). ASV1, on the other hand, resulted in a 100% match with members from multiple clades (Clades I, II, and III). In our long-term dataset, we found that the ASV5 peaks corresponded to the recurrent yet temporally restricted *Synechococcus* blooms observed during spring with flow cytometry (Supplementary Figure 8). Pelagibacter_A also presented two specific responses with ASV6 and ASV10 (1 nucleotide divergence) responding similarly, in contrast to the other ASVs presenting a significant change within the genus (Figure 4). Finally, the different ASVs belonging to HIMB59 (former SAR11 clade V) presented multiple responses (Figure 4).

Seasonality at broad taxonomical levels.

Having delineated how the ASVs behave seasonally and what are the drivers of these differences, we tested whether synchronized responses at higher taxonomic levels existed. When we analyzed the general distribution across ranks, we found that the class rank was mostly non-seasonal (98.9% peak normalized power – PN values, $p \leq 0.01$, PN ≤ 10; Figure 5). Both the order and family ranks displayed a similar distribution with ~50% of the results being seasonal, while this value increased up to ~60% at the genus rank. These distributions were different for each class; Alphaproteobacteria presented a clear bimodality while Gammaproteobacteria values were evenly distributed across the PN statistic (Figure 5). By checking each level separately, the bulk Alphaproteobacteria class distribution (Supplementary Figure 9, PN mean = 5.3) could be linked directly to that of the Pelagibacterales order, since this was the most abundant group (Supplementary Figure 9B) and appeared as non-seasonal (PN mean = 5.7, Supplementary Figure 9A). Observing the other prevalent orders (Rhodobacterales, Puniceispirillales –SAR116 clade– and HIMB59), the seasonality statistic was quite robust when randomly removing different ASVs (Supplementary Figure 9). Puniceispirillales for example appeared mostly during summer. This observation was different for the Gammaproteobacteria orders (Supplementary Figure 10A); the SAR86 and Pseudomonadales orders were close to the seasonality threshold resulting in half of the randomizations as non-seasonal. Moreover, for the Pseudomonadales order, we observed that it was composed of various families, each with different seasonality (Supplementary Figure 10B). The Bacteroidia class only showed seasonality at the genus level for UBA7446, an uncultured genus within the family *Flavobacteriaceae* (Supplementary Figure 11). Thus, the distributions at the order level were diametrically different, with Alphaproteobacteria including some seasonal orders, Gammaproteobacteria orders presenting a peak in the limit of seasonality, and all orders of Bacteroidia presenting a non-seasonal trend. Nevertheless, for most groups, the family and genus ranks presented similar seasonal trends to those displayed by the order to which they belonged.

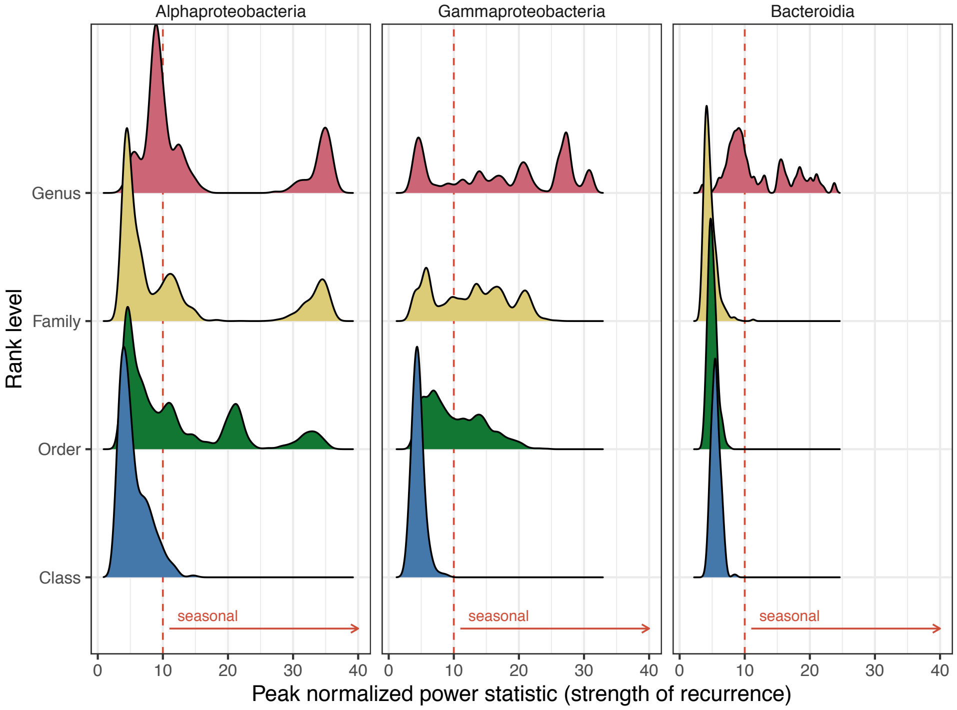


Figure 5: Density distribution of the peak normalized power statistic (as proxy for seasonality) for each rank level in the Alphaproteobacteria, Gammaproteobacteria and Bacteroidia classes. The red lines indicate the used threshold for seasonality ($q \leq 0.05$ and PN ≥ 10).

1.4 Discussion

We explored how marine bacterial communities are structured seasonally at fine taxonomical levels and whether the structure is maintained at higher ranks through long-term sampling and amplicon sequencing of the 16S rRNA gene in a temperate coastal environment. Specifically, we investigated how closely related ASVs responded to the environmental conditions that appeared recurrently at the coastal site. Overall, we found that around half of the total relative abundance of the community displayed seasonality at the ASV level. Within the genus level, we showed how niche similarity decreased with increasing nucleotide divergence for at least 3 genera. We then checked how various environmental parameters define the niche for the components of various genera. Finally, we analyzed how the patterns of seasonality aggregate at broader taxonomic ranks, proving that, in our dataset, the class level was non-seasonal and that the other ranks tested (i.e. order and family) presented a variety of trends.

As discussed above, the use of 16S rRNA gene amplicons has its limitations for the delineation of biological units (VanInsberghe *et al.*, 2020). The power of this genetic marker to resolve closely related taxa changes for different bacterial clades, but various studies have shown that species delineation is not always achievable by sequencing a region of this phylogenetic marker (Johnson *et al.*, 2019; VanInsberghe *et al.*, 2020). Despite this limitation, amplicon marker gene sequencing still represents the fastest and most comprehensive approach for studying ecological patterns through identifying robust trends in large datasets. To stay on the conservative side in our interpretations, we set the within-genus level as the one for which we can assign patterns with certainty.

Contrasting environmental conditions throughout the year.
The environmental parameters displayed a clear seasonal pattern, with the highest rates of change between summer and winter, and the bacterial community mirrored these changes as observed in alpha diversity and community similarity (beta diversity). Patterns of alpha and beta diversity had been studied before at our study site but in much shorter surveys (1-2 years, Alonso-Sáez *et al.*, 2007; Mestre *et al.*, 2020). The analysis of eleven years of data unveiled that the highest differences in community structure also occurred between summer and winter, while the highest variability was found between spring and winter, which could be related to the recurrent phytoplankton blooms that occur during these periods, with differing intensity over the decade (see also the abundance of phototrophic nanoflagellates, PNF, in Supplementary Figure 1) (Nunes *et al.*, 2018).

Patterns of community structure have been largely studied in different temperate coastal environments accurately describing yearly successions (Bunse and Pinhassi, 2017). The community composition however can be driven by regional differences, such as the recurrence of phytoplankton blooms (Lindh *et al.*, 2015) or nutrient fluxes (Lambert *et al.*, 2021), modifying the bacterioplankton patterns from site to site. In the nearby long-term microbial station SOLA (Banyuls-sur-Mer, France), a seven-year seasonal study compared the bacterial, eukaryotic and archaeal community through ASV delineation (Lambert *et al.*, 2018). The number of ASVs in the bacterial community was similar to that observed here (6825 ASVs in this study vs 6242 at SOLA) and a similar community composition was observed, e.g. *Pelagibacteraceae* and Synechococcales dominated the communities at both sites. However, some differences were detected between our study and that of Lambert *et al.* (2018); a relevant group in Blanes Bay was the HIMB59 order, initially considered part of the SAR11 clade V (Viklund *et al.*, 2013; Martijn *et al.*, 2018), which was absent from the SOLA station dataset (Salter *et al.*, 2015; Lambert *et al.*, 2018). This result could either be related to primer biases or to differences in the taxonomic assignation. This group has been assigned a variety of names and phylogenetic positions; as an example, MAGs from the HIMB59 order were identical to the AEGEAN-169 marine group at the 16S rRNA gene comparison. This group, found in multiple surface and deep waters sites (Alonso-Sáez *et al.*, 2007; Cram *et al.*, 2015), appears in the SILVA classification within the Rhodospirillales order. Martijn *et al.* (2018) however concluded that the HIMB59 and other relevant MAGs conform a separate clade neither within the Pelagibacterales nor the Rhodospirillales, in agreement with the Genome Taxonomy Database assignation used here.

Half of the total community is seasonal.

Determining seasonality is not trivial, as it implies taking a binary decision for a trait that is likely continuous in a gradient rather than into two discrete states. In our analysis, we found a total of 297 seasonal ASVs (34% of the evaluated ASVs), which made up a total of 47% of the sequence relative abundance. This number of seasonal bacterial ASVs triplicates the results found by Lambert *et al.* (2018) (89 ASVs), and the total relative abundance of seasonal bacteria was also higher in our study compared to that observed at the SOLA station (47% vs 31.3%). Since we followed the same statistical methodologies, the observed differences were somehow surprising. Differences in the length of the time series (7 years at SOLA vs 11 years at Blanes Bay) and the sampling scheme, with biweekly sampling at SOLA and monthly at Blanes Bay, could result to a certain degree in the observed disparities. Another explanation could derive from the presence of more irregular perturbations, such as river discharge in the Banyuls basin affecting the recurrence of the community through for example more variable salinity levels (Guizien *et al.*, 2007). Further studies would be needed to find a possible explanation for these discrepancies.

The seasonal patterns observed in our time series varied among different taxonomic groups (Supplementary Figure 5). Pelagibacter_A (SAR11 clade II) did not present seasonal ASVs; this result contrasts with what was observed in the Bermuda Atlantic Time series (BATS), where this group is present mostly during spring (Giovannoni, 2017). On the other hand, AG-337-I02 (order HIMB59) peaked during winter in Blanes, coinciding with what was observed at BATS (using SAR11 clade V as the group's nomenclature). Nevertheless, the biogeochemical setting, physical forcing and other environmental factors that could control the temporal dynamics at BATS (Steinberg *et al.*, 2001) are quite different from those of the coastal NW Mediterranean. Besides, HIMB114 (SAR11 clade III) presented peak abundances during summer in Blanes, a result also observed in Banyuls-sur-Mer (Salter *et al.*, 2015). Our study thus complements the data existing from previous long-term datasets. A direct comparison of data from distinct sites would help understand these differences but this comparison is constrained by the different methodologies used (i.e. hypervariable region amplified or primer set used). When the sequencing of the complete 16S rRNA gene becomes a common practice, comparisons across microbial observatories will be easier to conduct (Johnson *et al.*, 2019).

Niche similarity decreases with genetic distance in the 16S rRNA gene.

Temporal distributions can inform on niche relatedness among closely related taxa. Specifically, cooccurrence and covariance could point to niche similarity. In this study, we found a clear trend between niche similarity and nucleotide divergence for *Pelagibacter*, Pelagibacter_A (i.e. SAR11 clade I and II), and less clearly for SAR86A. The pattern is consistent with environmental filtering, in which similar niches are occupied by closely related taxa sharing similar traits or adaptations, as seen previously for other taxonomic groups in environments such as lakes (Horner-Devine and Bohannan, 2006; Tromas *et al.*, 2018). Environmental filtering would include both abiotic (environmental filtering *sensu stricto*) and biotic factors such as ecological interactions (Cadotte and Tucker, 2017; Tromas *et al.*, 2020). For most genera, however, there was no clear pattern. Since the 16S rRNA gene is very conserved, comparing niche similarity among ASVs could imply comparisons at broader level that that of strains. Each change in this marker gene can represent multiple changes at the genomic level, which could involve a change in niche distribution (Grote *et al.*, 2012; VanInsberghe *et al.*, 2020). In fact, even when merging the results for all the genera (excluding the SAR11 groups), there was no clear decrease in Rho with increasing nucleotide divergence. Nevertheless, as stated before, we observed a pattern for *Pelagibacter* and Pelagibacter_A. A possible reason for that observation is that these are the only groups presenting enough ASVs to result in a clear trend. Besides these two genera, others presenting a similar decrease pattern were SAR86A and *Luminiphilus*, which are the subsequent groups in number of ASVs per genera (22 and 26 respectively; Figure 3). Another possible explanation is that the 16S rRNA gene could reflect in a greater way the genomic differences for *Pelagibacter* than for other groups, possibly due to the special evolutionary history of this group (López-Pérez *et al.*, 2020). Both an increase in sequencing depth and an improvement of the resolution for the marker gene by sequencing a larger fragment could help to obtain a clearer picture (Ladau and Eloe-Fadrosh, 2019).

When we checked how the individual ASVs responded to the measured environmental variables, we found two types of responses at the genus level: groups in which all the ASVs displayed a similar response, such as *Pelagibacter*, AG-337-I02 (AEGEAN-169), D2472 (SAR86) and *Luminiphilus*, and groups with ASVs presenting temporal differentiation, such as *Synechococcus* and SAR86A. The groups presenting the same patterns varied in their response; in the case of *Pelagibacter*, there was a clear distinction between the seasonal ASVs and the ones appearing all year round (e.g. in Figure 4, see the two clusters in the *Pelagibacter* dendrogram). *Pelagibacter* therefore presented multiple variants with similar responses to the studied environmental changes (Larkin and Martiny, 2017). On the other hand, different *Synechococcus* ASVs presented completely different adaptations –e.g. ASV1 and ASV5– in an example of a clear niche switch by a previous ecotype differentiation. In the latter case, ASV1 presented multiple matches in the Cyanorak database, which exemplifies the problems with the limited power of the 16S rRNA gene V3-V4 regions to resolve species for certain groups (Johnson *et al.*, 2019). This could reflect that there are many clades considered as the same ASV, which could explain that this variant dominates all year round. Summing up, these results illustrate the diversity of ecological trends within genera, which would have been hidden using sequence clustering methods.

Lack of seasonality at the class level.

It has been hypothesized that phylogenetic related taxa could share ecological traits and respond similarly to environmental changes (Philippot *et al.*, 2010; Martiny *et al.*, 2015) but it is unclear whether bacteria from the same genus, family, order or class phylogenetic ranks are ecologically cohesive (Philippot *et al.*, 2010). These ecological traits could be determined by phylogenetic history, as seems to be the case of particle-attached vs free-living lifestyle (Salazar *et al.*, 2016; Mestre *et al.*, 2017). In the case of surface coastal waters, periodic changes in environmental conditions should promote recurrent niches. By randomly aggregating the ASVs at different ranks, broad patterns of abundance could emerge coming from cohesive seasonal responses. Our results were opposite to those observed in the English Channel, with the Alphaproteobacteria and Gammaproteobacteria classes presenting a high autocorrelation driven by a strong seasonal pattern (Gilbert *et al.*, 2012; Faust *et al.*, 2015). The higher annual temperature range in the English Channel could explain the observed differences compared with Blanes Bay, with less temperature variability. By facing a stronger environmental gradient, the whole community composition could consequentially change at a higher taxonomic rank. Bimodal distributions (seasonal and non-seasonal results) originate in groups containing ASVs that have strong seasonal trends and other non-seasonal ASVs, as is the case for Rhodobacterales and Pseudomonadales, copiotrophic groups occupying many different niches. *Rhodobacteraceae*, for example, includes ASVs with seasonal peaks in every season (Supplementary Figure 5). Finally, the groups with all ASVs being seasonal could present more constrained optimal conditions of growth than those groups that appear randomly or all year-round. Examples of this behavior are the Puniceispirillales (SAR116 clade), a group harboring proteorhodopsin for which most of the ASVs were seasonal and peaked during summer (Lee *et al.*, 2019). Metagenomic and genome-centric approaches as well as physiological experimentation with available isolates would help shedding light on the traits that determine the niche for these cohesive groups and the differences with the more diverse groups.

1.5 Conclusions

The use of a long-term time series and fine-grained taxonomic resolution through the use of ASVs allowed to compare within-genus ecological distributions in a coastal site. Specifically, we could prove that for certain genera niche similarity decreased with 16S rRNA gene nucleotide divergence, indicating that more similar variants coexist. Our results thus point to environmental selection as an important process structuring the seasonal dynamics of the studied microbiota. Both abiotic conditions and biotic processes (e.g. competition and other interactions) would exert selection in the analyzed community. Additionally, through modeling of the differential abundance with a variety of environmental parameters, we unveiled the presence of different ecological patterns spanning different seasons. Finally, the analysis of different seasonality distributions for each phylogenetic rank indicated that the class rank was non-seasonal for the groups analyzed, being thus ecologically non-coherent. Contrarily, some groups at the family and genera ranks presented cohesive responses. Overall, this study sheds light on the niche specialization of relevant genera in marine coastal microbial communities.

1.6 Acknowledgments

We thank all the people involved in operating the BBMO, especially Clara Cardelús and Captain Anselm for facilitating sampling, Vanessa Balagué for laboratory procedures and Ramon Massana for sharing the HNF data. We also thank the MarBits Bioinformatics platform of the Institut de Ciències del Mar, and in particular Pablo Sánchez, for computing support. We additionally thank Maria Touceda-Suárez, Gabriele Schiro and Yongjian Chen from the Barberán Lab for insightful discussions. We would also like to thank three anonymous reviewers for the useful suggestions that helped substantially improve previous versions of this manuscript. This research was funded by grants REMEI (CTM2015-70340-R), MIAU (RTI2018-101025-B-I00) and ECLIPSE (PID2019-110128RB-I00) from the Spanish Ministry of Science and Innovation, by the Grup de Recerca de la Generalitat de Catalunya (2017SGR/1568) and received support from the Spanish government through the 'Severo Ochoa Centre of Excellence' accreditation (CEX2019-000928-S). Adrià Auladell was supported by a Spanish FPI grant and IF received the support of the Fundación BBVA through the 'Becas Leonardo a Investigadores y Creadores Culturales' 2019 Program.

1.7 References

Alonso-Sáez, L., Balagué, V., Sà, E.L., Sánchez, O., González, J.M., Pinhassi, J., *et al.* (2007) Seasonality in bacterial diversity in north-west Mediterranean coastal waters: Assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiol Ecol* 60: 98–112.

Alonso-Sáez, L., Díaz-Pérez, L., and Morán, X.A.G. (2015) The hidden seasonality of the rare biosphere in coastal marine bacterioplankton. *Environ Microbiol* 17: 3766–3780.

Apprill, A., McNally, S., Parsons, R., and Weber, L. (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat Microb Ecol* 75: 129–137.

Auladell, A., Sánchez, P., Sánchez, O., Gasol, J.M., and Ferrera, I. (2019) Long-term seasonal and interannual variability of marine aerobic anoxygenic photoheterotrophic bacteria. *ISME J* 13: 1975–1987.

Bunse, C. and Pinhassi, J. (2017) Marine Bacterioplankton Seasonal Succession Dynamics. *Trends Microbiol* 25: 1–12.

Buttigieg, P.L., Fadeev, E., Bienhold, C., Hehemann, L., Offre, P., and Boetius, A. (2018) Marine microbes in 4D-using time series observation to assess the dynamics of the ocean microbiome and its links to ocean health. *Curr Opin Microbiol* 43: 169–185.

Cadotte, M.W. and Tucker, C.M. (2017) Should Environmental Filtering be Abandoned? *Trends Ecol Evol* 32: 429–437.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13: 581.

Campbell, B.J., Yu, L., Heidelberg, J.F., and Kirchman, D.L. (2011) Activity of abundant and rare bacteria in a coastal ocean. PNAS 108: 12776–12781.

Chafee, M., Fernàndez-Guerra, A., Buttigieg, P.L., Gerdts, G., Eren, A.M., Teeling, H., and Amann, R.I. (2018) Recurrent patterns of microdiversity in a temperate coastal marine environment. *ISME J* 12: 237–252.

Cordero, O.X. and Polz, M.F. (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* 12: 263–273.

Cram, J.A., Chow, C.-E.T., Sachdeva, R., Needham, D.M., Parada, A.E., Steele, J.A., and Fuhrman, J.A. (2015) Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *ISME J* 9: 563–580.

Eiler, A., Hayakawa, D.H., and Rappé, M.S. (2011) Non-Random Assembly of Bacterioplankton Communities in the Subtropical North Pacific Ocean. *Front Microbio*l 2:.

Falkowski, P. (2012) Ocean Science: The power of plankton. *Nature* 483: S17–S20.

Farrant, G.K., Doré, H., Cornejo-Castillo, F.M., Partensky, F., Ratin, M., Ostrowski, M., *et al.* (2016) Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. PNAS 113: E3365–E3374.

Faust, K., Lahti, L., Gonze, D., de Vos, W.M., and Raes, J. (2015) Metagenomics meets time series analysis: Unraveling microbial community dynamics. *Curr Opin Microbiol* 25: 56–66.

Fuhrman, J.A., Cram, J.A., and Needham, D.M. (2015) Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol* 13: 133–146.

Garczarek, L., Guyet, U., Doré, H., Farrant, G.K., Hoebeke, M., Brillet-Guéguen, L., *et al.* (2020) Cyanorak v2.1: a scalable information system dedicated to the visualization and expert curation of marine and brackish picocyanobacteria genomes. *Nucleic Acids Res* 49: gkaa958.

Gasol, J.M., Cardelús, C., Morán, X.A.G., Balagué, V., Forn, I., Marrasé, C., *et al.* (2016) Seasonal patterns in phytoplankton photosynthetic parameters and primary production at a coastal NW Mediterranean site. *Sci Mar* 80: 63–77.

Gasol, J.M. and Morán, X.A.G. (2016) Flow Cytometric Determination of Microbial Abundances and Its Use to Obtain Indices of Community Structure and Relative Activity. In Hydrocarbon and Lipid Microbiology Protocols: Single-Cell and Single-Molecule Methods. Springer Protocols Handbooks. McGenity, T.J., Timmis, K.N., and Nogales, B. (eds). Berlin, Heidelberg: Springer, pp. 159–187.

Gilbert, J.A., Steele, J.A., Caporaso, J.G., Steinbrück, L., Reeder, J., Temperton, B., *et al.* (2012) Defining seasonal marine microbial community dynamics. *ISME J* 6: 298–308.

Giner, C.R., Balagué, V., Krabberød, A.K., Ferrera, I., Reñé, A., Garcés, E., *et al.* (2019) Quantifying long-term recurrence in planktonic microbial eukaryotes. *Mol Ecol* 28: 923–935.

Giovannoni, S.J. (2017) SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Annu Rev Mar Sci* 9: 231–255.

Grasshoff, K., Ehrhardt, M., and Kremling, K. (1983) Methods of seawater analysis, 2nd ed. Verlag Chemie, Weinheim.

Grote, J., Thrash, J.C., Huggett, M.J., Landry, Z.C., Carini, P., Giovannoni, S.J., and Rappé, M.S. (2012) Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade. *mBio* 3: e00252-12.

Guizien, K., Charles, F., Lantoine, F., and Naudin, J.-J. (2007) Nearshore dynamics of nutrients and chlorophyll during Mediterranean-type flash-floods. *Aquat Living Resour* 20: 3–14.

Hammarlund, S.P., Gedeon, T., Carlson, R.P., and Harcombe, W.R. (2021) Limitation by a shared mutualist promotes coexistence of multiple competing partners. *Nat Commun* 12: 619.

Haro-Moreno, J.M., Rodriguez-Valera, F., Rosselli, R., Martinez-Hernandez, F., Roda-Garcia, J.J., Gomez, M.L., *et al.* (2020) Ecogenomics of the SAR11 clade. *Environ Microbiol* 22: 1748–1763.

Hastie, T. and Tibshirani, R. (1986) Generalized Additive Models. *Statist Sci* 1: 297–310.

Heiberger, R.M. (2020) HH: Statistical Analysis and Data Display: Heiberger and Holland.

Herlemann, D.P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J.J., and Andersson, A.F. (2011) Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J* 5: 1571–1579.

Horner-Devine, M.C. and Bohannan, B.J.M. (2006) Phylogenetic clustering and overdispersion in bacterial communities. *Ecology* 87: S100–S108.

Hutchinson, G.E. (1957) Concluding Remarks. *Cold Spring Harb Sym* 22: 415–427.

Johnson, J.S., Spakowicz, D.J., Hong, B.-Y., Petersen, L.M., Demkowicz, P., Chen, L., *et al.* (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 10: 5029.

Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., *et al.* (2014) Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science* 344: 416–420.

Ladau, J. and Eloe-Fadrosh, E.A. (2019) Spatial, Temporal, and Phylogenetic Scales of Microbial Ecology. *Trends Microbiol* 27: 662–669.

Lambert, S., Lozano, J.-C., Bouget, F.-Y., and Galand, P.E. (2021) Seasonal marine microorganisms change neighbours under contrasting environmental conditions. *Environ Microbiol* 23: 2592-604

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., and Galand, P.E. (2018) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *ISME J*. 13: 388-401

Larkin, A.A. and Martiny, A.C. (2017) Microdiversity shapes the traits, niche space, and biogeography of microbial taxa: The ecological function of microdiversity. *Envir Microbiol Rep* 9: 55–70.

Lee, J., Kwon, K.K., Lim, S.-I., Song, J., Choi, A.R., Yang, S.-H., *et al.* (2019) Isolation, cultivation, and genome analysis of proteorhodopsin-containing SAR116-clade strain Candidatus Puniceispirillum marinum IMCC1322. *J Microbiol* 57: 676–687.

Lemonnier, C., Perennou, M., Eveillard, D., Fernandez-Guerra, A., Leynaert, A., Marié, L., *et al.* (2020) Linking Spatial and Temporal Dynamic of Bacterioplankton Communities With Ecological Strategies Across a Coastal Frontal Area. *Front Mar Sci* 7: 376.

Lindh, M.V., Sjöstedt, J., Andersson, A.F., Baltar, F., Hugerth, L.W., Lundin, D., *et al.* (2015) Disentangling seasonal bacterioplankton population dynamics by high-frequency sampling: High-resolution temporal dynamics of marine bacteria. *Environ Microbiol* 17: 2459–2476.

López-Pérez, M., Haro-Moreno, J.M., Coutinho, F.H., Martinez-Garcia, M., and Rodriguez-Valera, F. (2020) The Evolutionary Success of the Marine Bacterium SAR11 Analyzed through a Metagenomic Perspective. *mSystems* 5: e00605-20.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S., and Bähler, J. (2015) Proportionality: A Valid Alternative to Correlation for Relative Data. *PLoS Comput Biol* 11: e1004075.

Martijn, J., Vosseberg, J., Guy, L., Offre, P., and Ettema, T.J.G. (2018) Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* 557: 101–105.

Martin, B.D., Witten, D., and Willis, A.D. (2020) Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann Appl Stat* 14: 94–115.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17: 10.

Martin-Platero, A.M., Cleary, B., Kauffman, K., Preheim, S.P., McGillicuddy, D.J., Alm, E.J., and Polz, M.F. (2018) High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nat Commun* 9: 266.

Martiny, J.B.H., Jones, S.E., Lennon, J.T., and Martiny, A.C. (2015) Microbiomes in light of traits: A phylogenetic perspective. *Science* 350: aac9323.

Massana, R., Murray, A.E., Preston, C.M., and Delong, E.F. (1997) Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Appl Environ Microbiol* 63: 50–56.

McMurdie, P.J. and Holmes, S. (2013) Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8: e61217.

Mestre, M., Borrull, E., Sala, M.M., and Gasol, J.M. (2017) Patterns of bacterial diversity in the marine planktonic particulate matter continuum. *ISME J* 11: 999–1010.

Mestre, M., Höfer, J., Sala, M.M., and Gasol, J.M. (2020) Seasonal Variation of Bacterial Diversity Along the Marine Particulate Matter Continuum. *Front Microbiol* 11: 1590.

Needham, D.M., Fichot, E.B., Wang, E., Berdjeb, L., Cram, J.A., Fichot, C.G., and Fuhrman, J.A. (2018) Dynamics and interactions of highly resolved marine plankton via automated high-frequency sampling. *ISME J* 12: 2417–2432.

Needham, D.M. and Fuhrman, J.A. (2016) Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol* 1: 16005.

Nunes, S., Latasa, M., Gasol, J.M., and Estrada, M. (2018) Seasonal and interannual variability of phytoplankton community structure in a Mediterranean coastal site. *Mar Ecol Prog Ser* 592: 57–75.

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

Pedersen, E.J., Miller, D.L., Simpson, G.L., and Ross, N. (2019) Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ* 7: e6876.

Philippot, L., Andersson, S.G.E., Battin, T.J., Prosser, J.I., Schimel, J.P., Whitman, W.B., and Hallin, S. (2010) The ecological coherence of high bacterial taxonomic ranks. *Nat Rev Microbiol* 8: 523–529.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41: D590–D596.

Quinn, T.P., Richardson, M.F., Lovell, D., and Crowley, T.M. (2017) propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci Rep* 7: 1–9.

R Core Team (2014) R: A language and environment for statistical computing.

Ruf, T. (1999) The Lomb-Scargle Periodogram in Biological Rhythm Research: Analysis of Incomplete and Unequally Spaced Time-Series. *Biol Rhythm Res* 30: 178–201.

Salazar, G., Cornejo-Castillo, F.M., Benítez-Barrios, V., Fraile-Nuez, E., Álvarez-Salgado, X.A., Duarte, C.M., *et al.* (2016) Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J* 10: 596–608.

Salter, I., Galand, P.E., Fagervold, S.K., Lebaron, P., Obernosterer, I., Oliver, M.J., *et al.* (2015) Seasonal dynamics of active SAR11 ecotypes in the oligotrophic Northwest Mediterranean Sea. *ISME J* 9: 347–360.

Schloss, P.D. (2020) Reintroducing mothur: 10 Years Later. *Appl Environ Microbiol* 86: 13.

Shade, A., Jones, S.E., Caporaso, J.G., Handelsman, J., Knight, R., Fierer, N., and Gilbert, A. (2014) Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio* 5: 1–9.

Steinberg, D.K., Carlson, C.A., Bates, N.R., Johnson, R.J., Michaels, A.F., and Knap, A.H. (2001) Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep Sea Res Part II Top Stud Oceanogr* 48: 1405–1447.

Tibshirani, R., Walther, G., and Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Ser C-Appl Stat* 63: 411–423.

Tromas, N., Taranu, Z.E., Castelli, M., Pimentel, J.S.M., Pereira, D.A., Marcoz, R., *et al.* (2020) The evolution of realized niches within freshwater Synechococcus. *Environ Microbiol* 22: 1238–1250.

Tromas, N., Taranu, Z.E., Martin, B.D., Willis, A., Fortin, N., Greer, C.W., and Shapiro, B.J. (2018) Niche Separation Increases With Genetic Distance Among Bloom-Forming Cyanobacteria. *Front Microbiol* 9: 438.

VanInsberghe, D., Arevalo, P., Chien, D., and Polz, M.F. (2020) How can microbial population genomics inform community ecology? *Phil Trans R Soc B* 375: 20190253.

Viklund, J., Martijn, J., Ettema, T.J.G., and Andersson, S.G.E. (2013) Comparative and Phylogenomic Evidence That the Alphaproteobacterium HIMB59 Is Not a Member of the Oceanic SAR11 Clade. *PLoS ONE* 8: e78858.

Wickham, H. (2016) ggplot2: Elegant graphics for data analysis, Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., *et al.* (2019) Welcome to the tidyverse. *J Open Source Softw*. 4: 1686.

Willis, A., Bunge, J., and Whitman, T. (2017) Improved detection of changes in species richness in high diversity microbial communities. *J R Stat Soc Ser C-Appl Stat* 66: 963–977.

Willis, A.D. and Martin, B.D. (2020) Estimating diversity in networked ecological communities. Biostatistics kxaa015.

Wright, E., S. (2016) Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. T*he R Journal* 8: 352.

Yentsch, C.S. and Menzel, D.W. (1963) A method for the determination of phytoplankton chlorophyll and phaeophytin by fluorescence. *Deep-Sea Res Oceanogr Abstr* 10: 221–231.

1.8 Supplementary figures

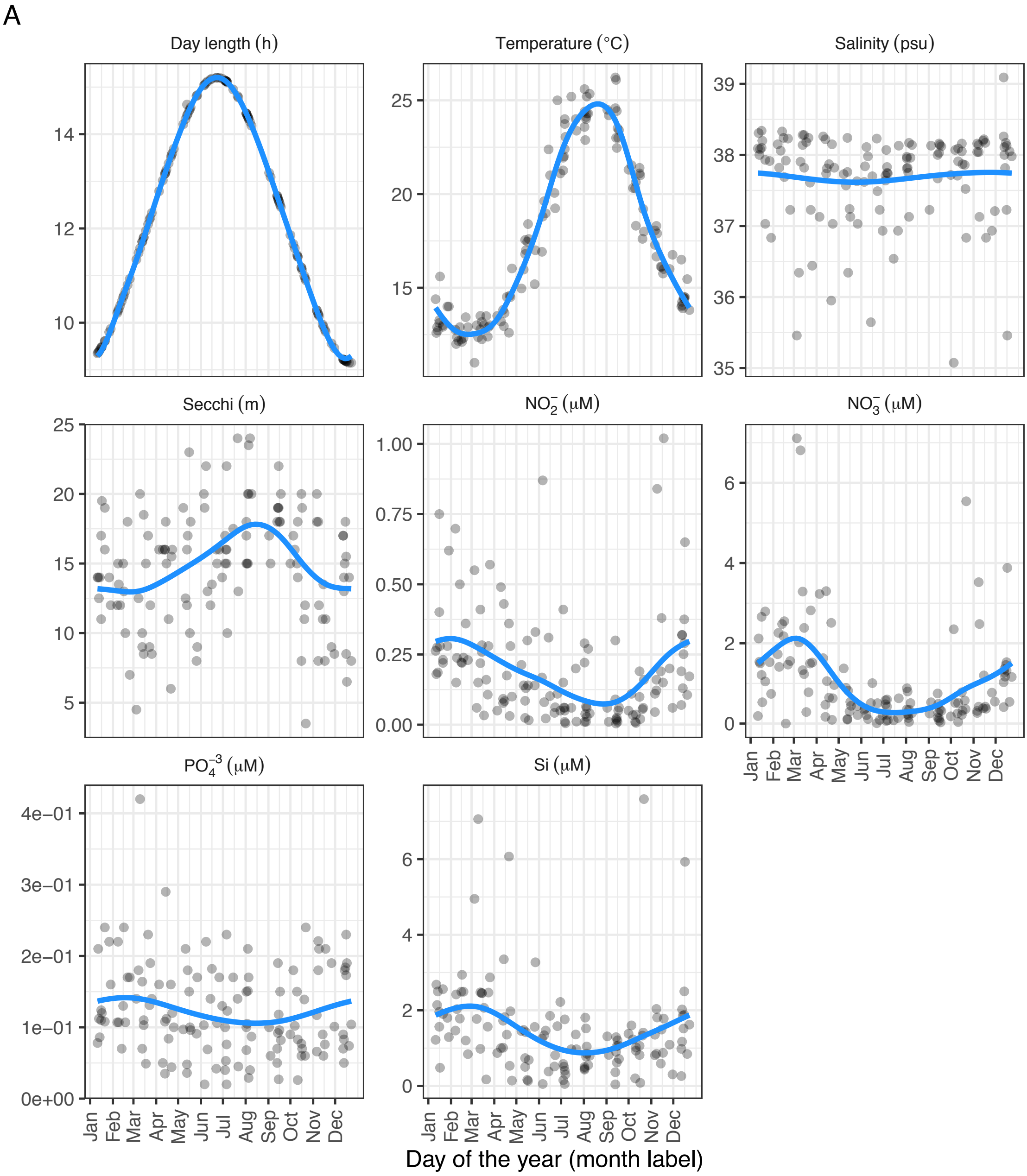
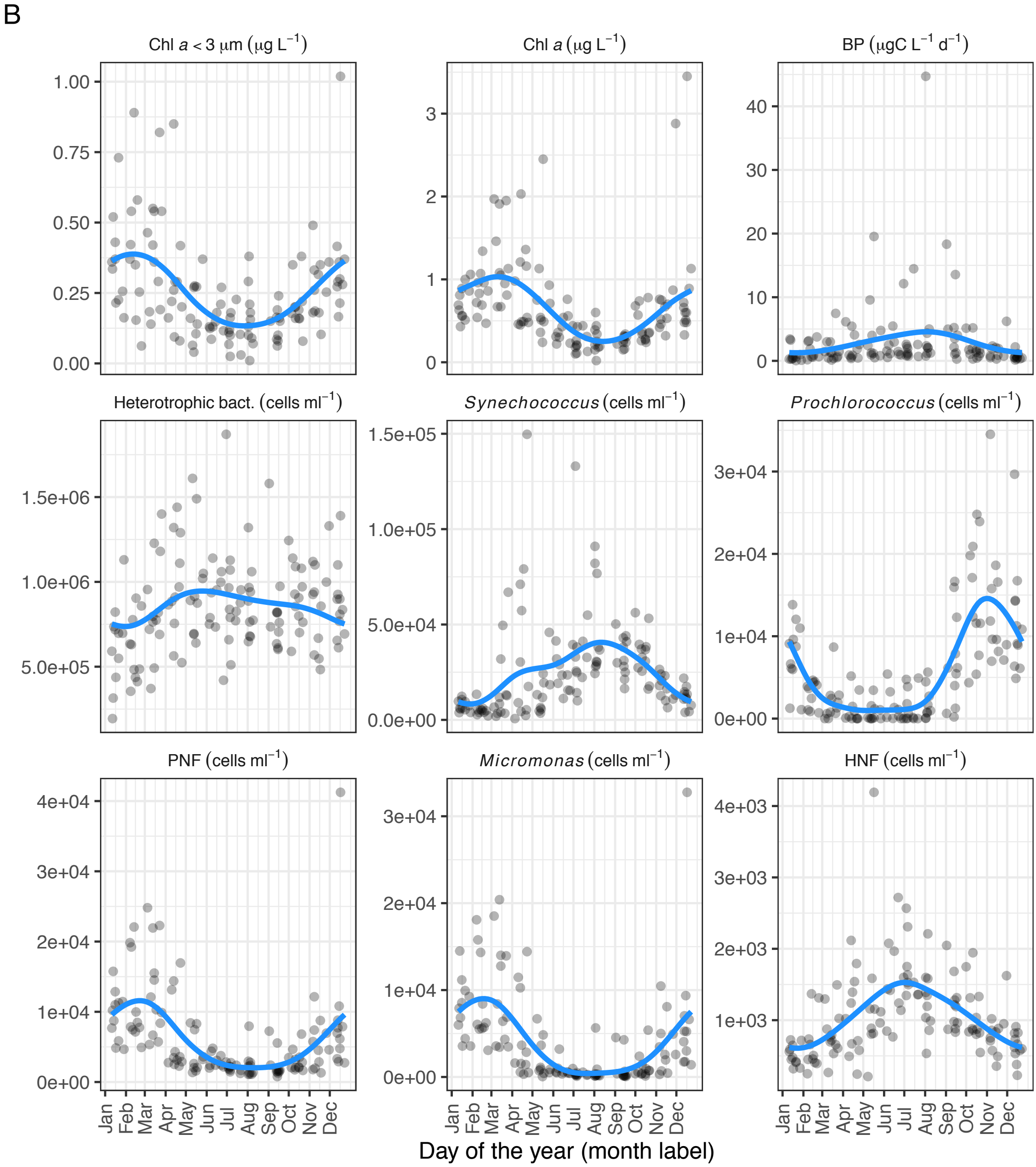


Figure S1: Distribution of the physicochemical (A) and biological (B) environmental variables measured in the Blanes Bay Microbial Observatory during the 11 years of this study. The Y axis corresponds to the parameter value (units indicated in the plot title) and the X axis correponds to the day of the year (month is shown for orientation, with the line ticks for the first day). A generalized additive model was fitted to the data. BP: Bacterial production; PNF: Phototrophic nanoflagellates; Cha<3 μ m: Chlorophyll *a* from the fraction smaller than 3 μ m; HNF: Heterotrophic nanoflagellates.


B
Chl *a* < 3 μm (μg L^{-1})
1.00
0.75
0.50
0.25
0.00
Chl *a* (μg L^{-1})
3
2
1
0
BP (μgC L^{-1} d^{-1})
40
30
20
10
0
Heterotrophic bact. (cells ml^{-1})
1.5e+06
1.0e+06
5.0e+05
Synechococcus (cells ml^{-1})
1.5e+05
1.0e+05
5.0e+04
0.0e+00
Prochlorococcus (cells ml^{-1})
3e+04
2e+04
1e+04
0e+00
PNF (cells ml^{-1})
4e+04
3e+04
2e+04
1e+04
0e+00
Micromonas (cells ml^{-1})
3e+04
2e+04
1e+04
0e+00
HNF (cells ml^{-1})
4e+03
3e+03
2e+03
1e+03
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Day of the year (month label)

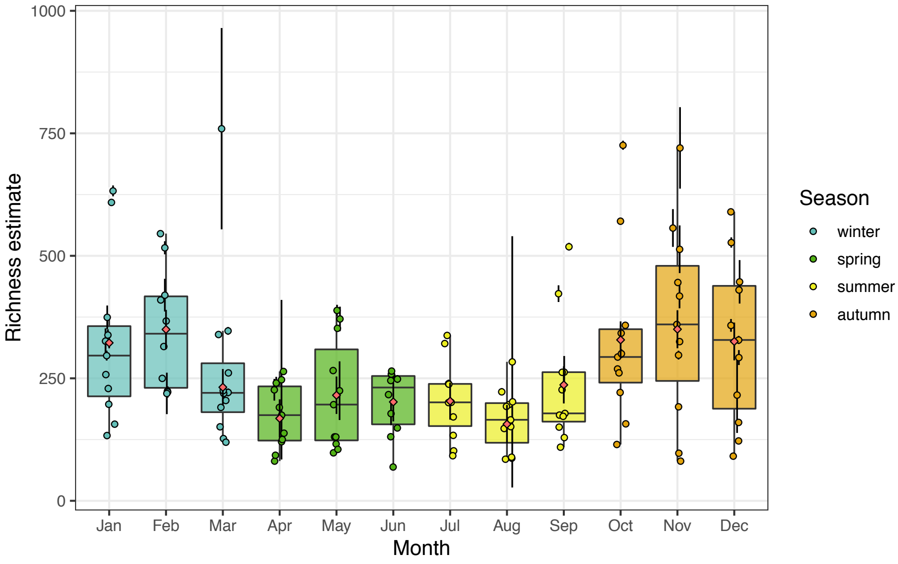


Figure S2: Alpha diversity (richness estimate) for the whole time series (11 years). Dots colored by season correspond to the sample estimates with the confidence interval at 95%; red rhomboids correspond to the mean month richness estimate by the *breakaway* package (with confidence interval 95%). Each boxplot presents the median and the 25% and 75% limits with the distribution of 11 points, and whiskers represent 1.5 times the interquartile range.

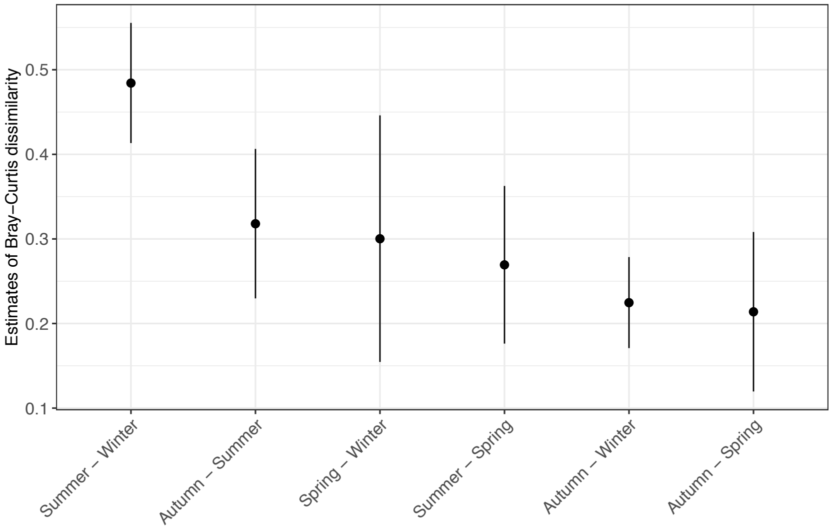


Figure S3: Comparisons of beta diversity estimates of Bray Curtis dissimilarity between the different seasons. The estimates and the 95% confidence intervals are displayed.

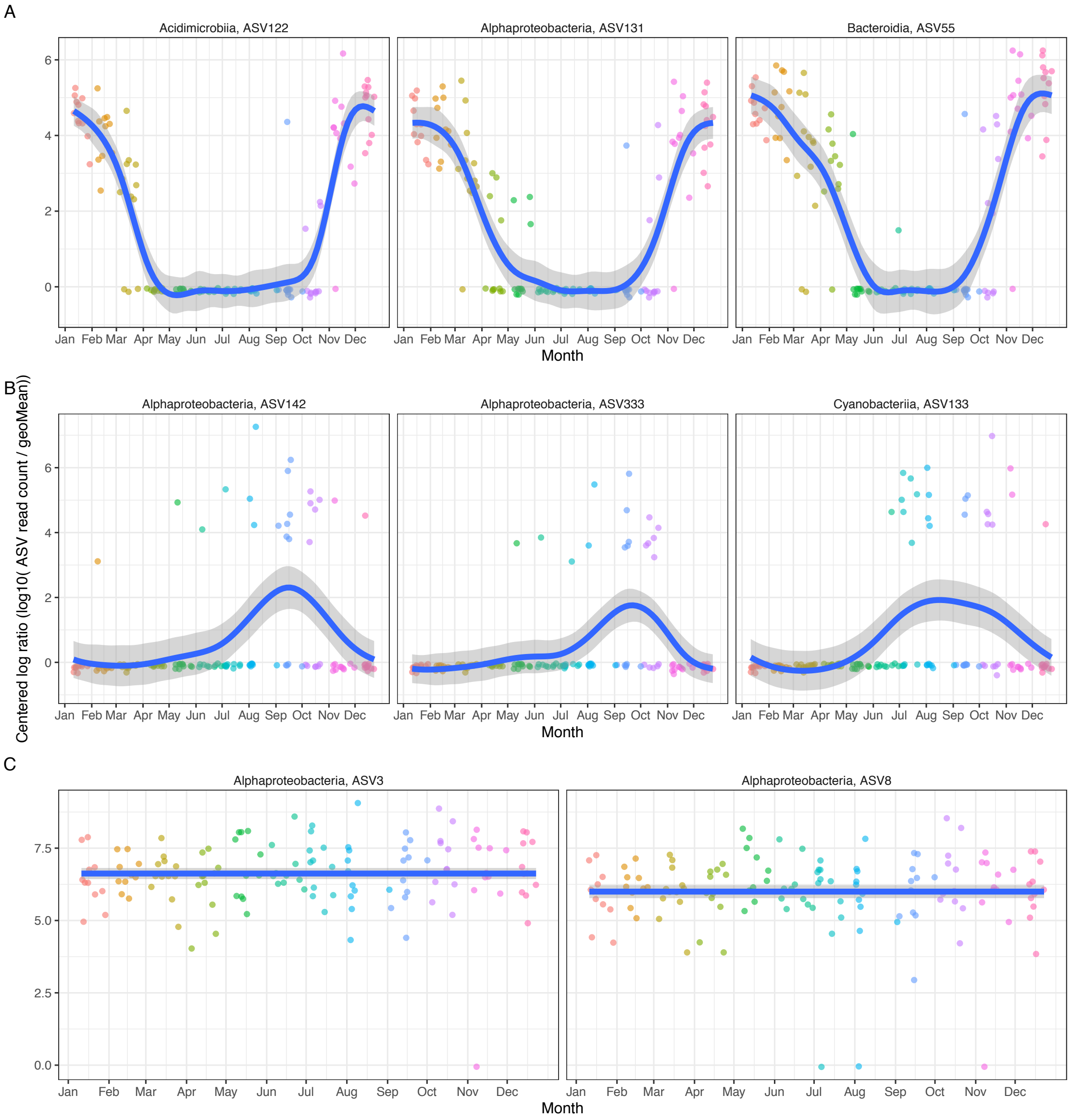


Figure S4: A) Distribution of ASVs with strong autumn-winter seasonality; B) ASVs with seasonality appearing only during specific years, and C) ASVs with no seasonality. The X axis corresponds to the day of the year (month is shown for orientation, with the line ticks for the first day) and the Y axis presents the read count transformed through the centered logarithm ratio abundance. A generalized additive model smooth is adjusted to the data points. For ASVs in panel A, taxonomic classification reached the class level only.

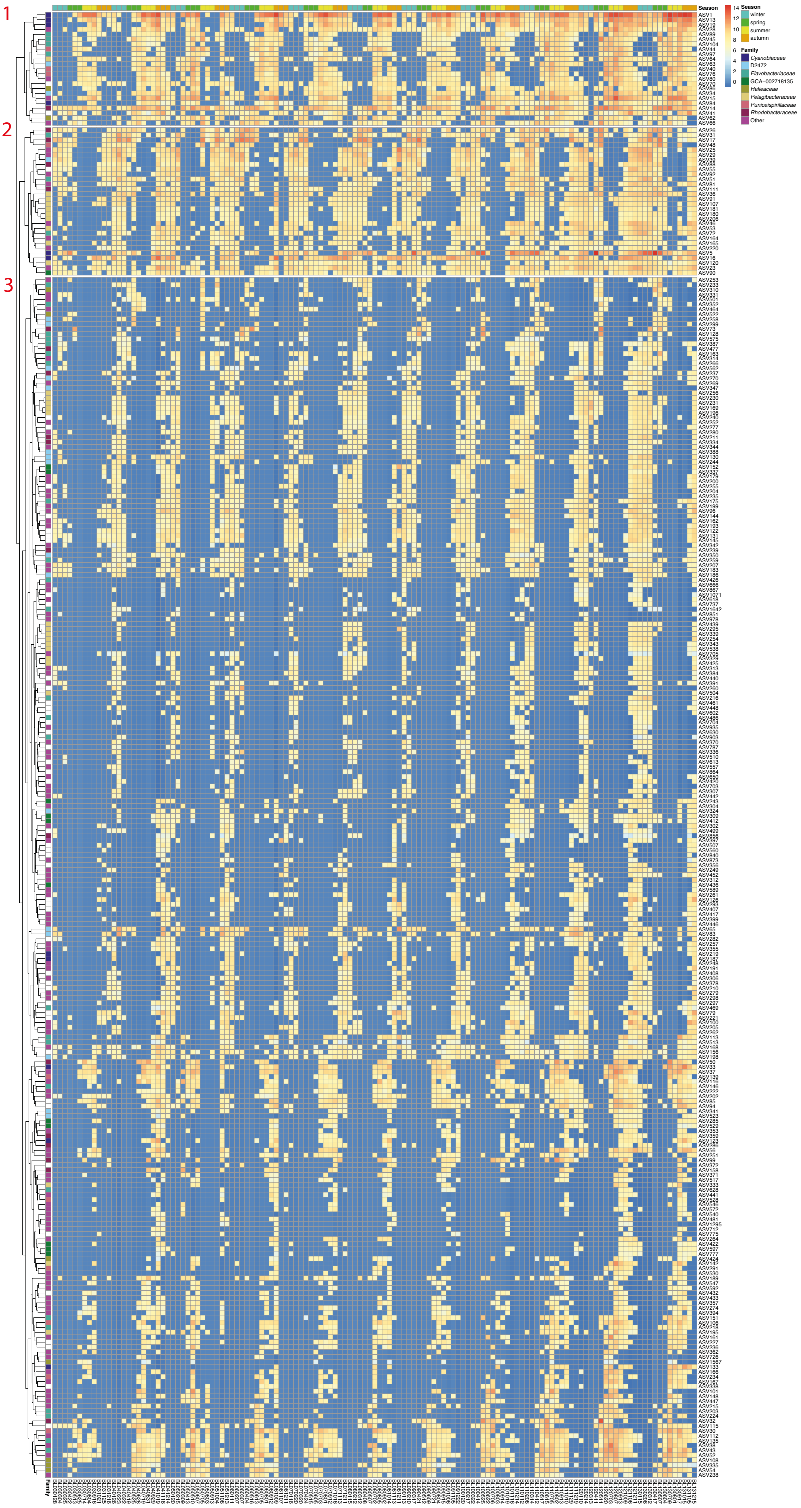


Figure S5: Heatmap displaying all the seasonal ASVs with a hierarchical clustering of the distributions. The 3 main clusters identified are presented and indicated with a red number. Each ASV is color coded by taxonomy and each sample by the season.

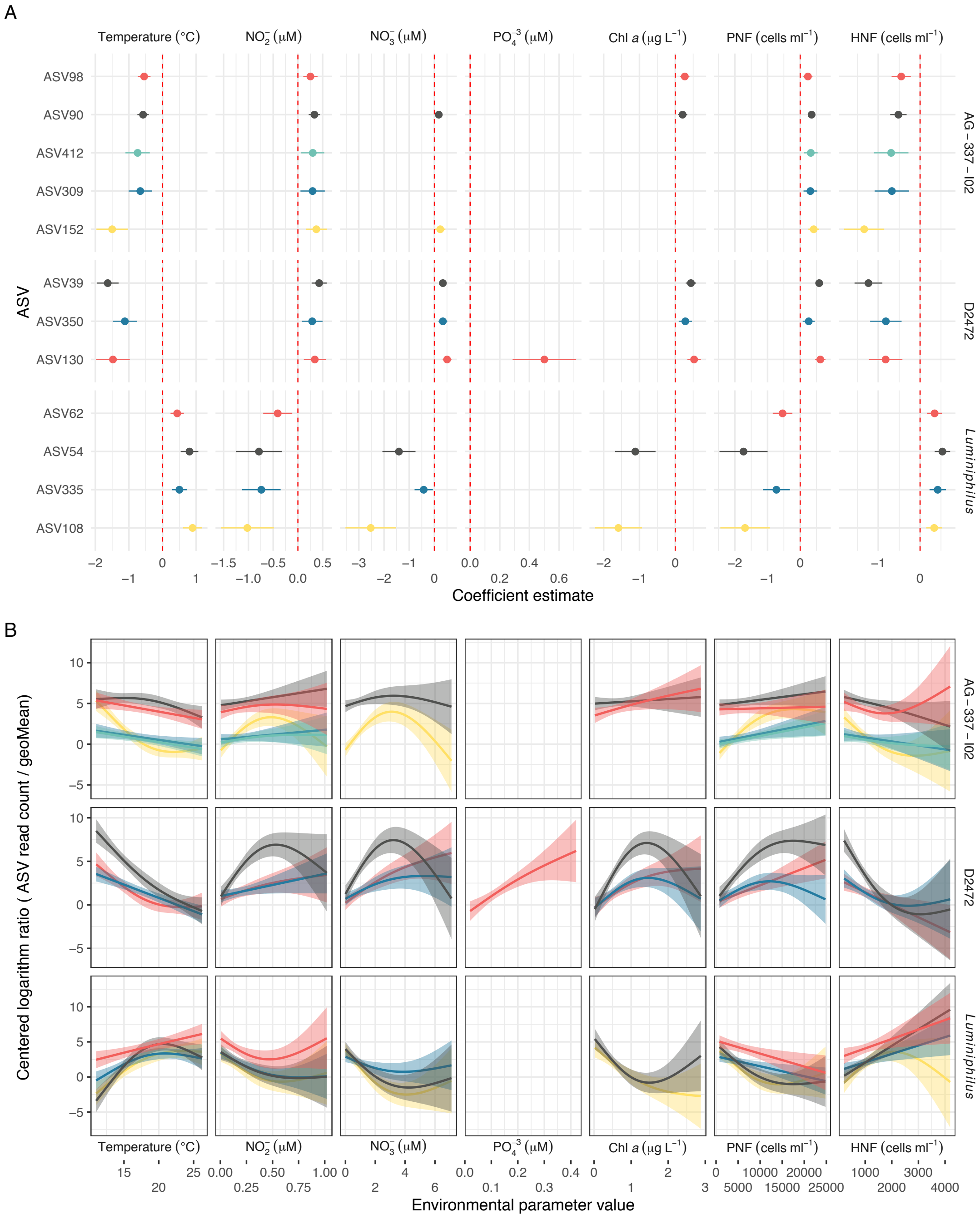


Figure S6: A) Significant models between ASVs from AG-337-I02 (Alphaproteobacteria), D2472 (Gammaproteobacteria) and *Luminiphilus* (Gammaproteobacteria) genera (rows) and environmental parameters (columns). The coefficient estimate indicates positive or negative responses to the parameter and is shown with a 95% confidence interval. The colors correspond to the different ASVs within a genus (only top 8 abundant ASVs are colored, the other ASVs are shown in grey). ASVs are ordered through a hierarchical clustering based on nucleotide divergence. B) Generalized additive model fits between the ASV centered logarithm ratio abundances and the parameter value distribution for the significant ASVs indicated in the upper plot. Panels and ASV colors are distributed as in the upper panel. PNF: Phototrophic nanoflagelates; HNF: Heterotrophic nanoflagelates.

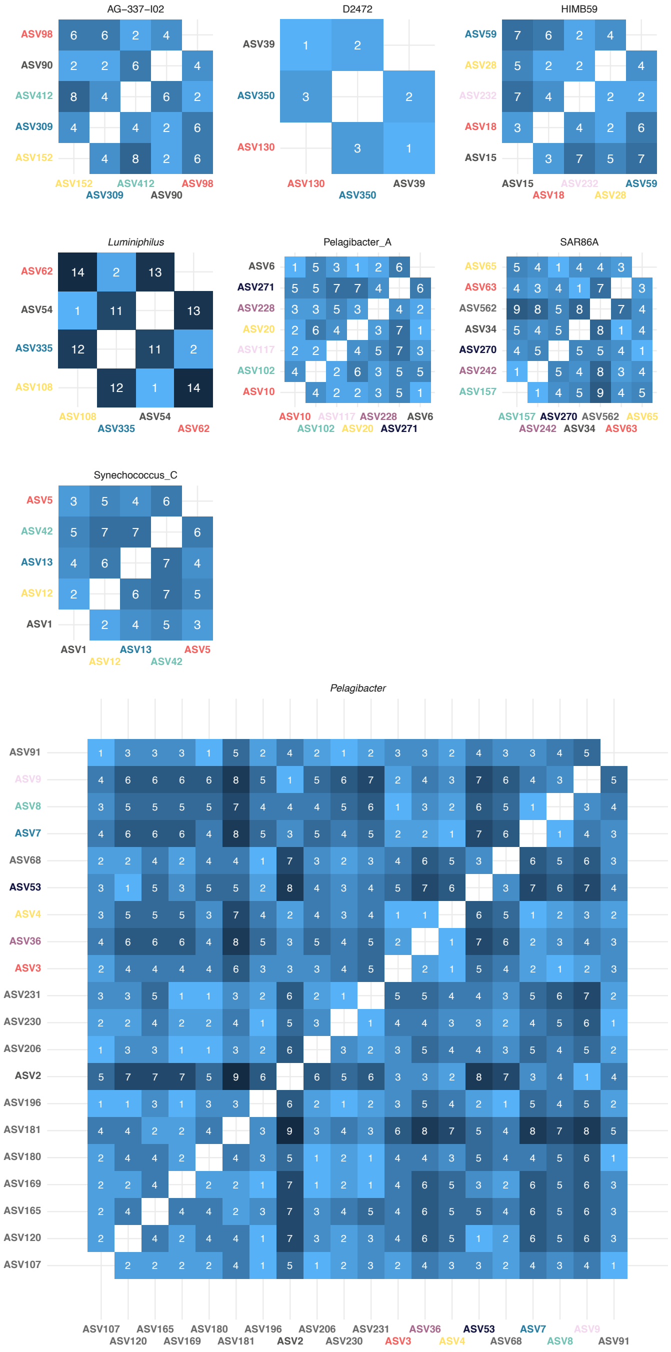


Figure S7: Nucleotide divergence heatmap among groups of ASVs presenting a significant response to environmental parameters. The genera included are: AG-337-I02, HIMB59, Pelagibacter_A and Pelagibacter (Alphaproteobacteria); D2472, SAR86 and *Luminiphilus* (Gammaproteobacteria); and Synechococcus_C (Cyanobacteria). The color corresponds to the different ASVs within a genus (only the top 8 abundant ASVs are colored, the other ASVs are shown in grey). Five nucleotide divergence equals to a median sequence identity of 98.8% in the 16S rRNA gene.

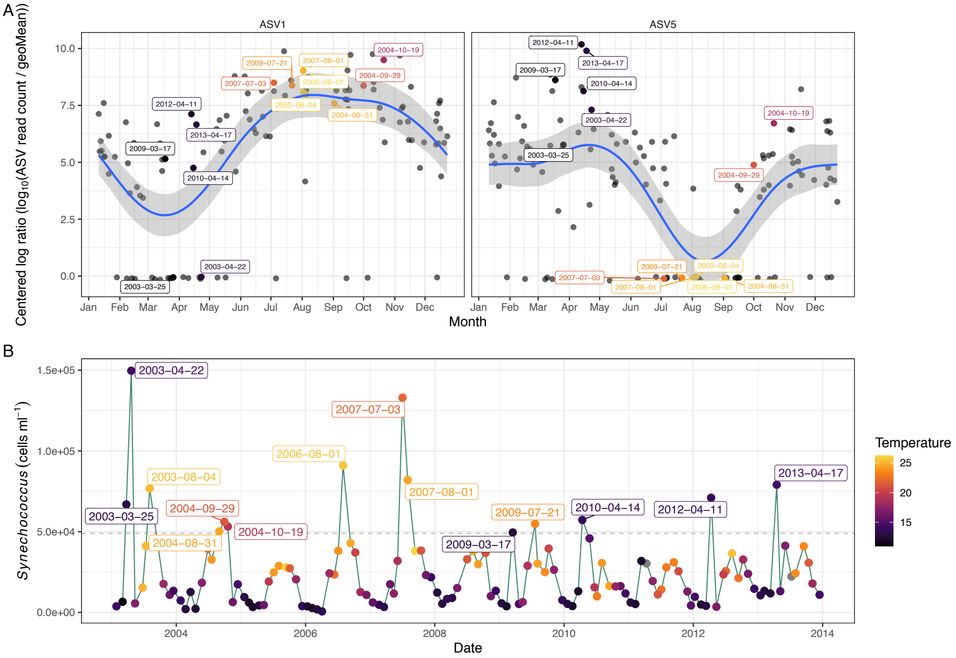


Figure S8: A) Monthly distribution of ASV1 and ASV5, both belonging to the Synechococcus_C genus. The Y axis corresponds to the centered logarithm ratio (with a pseudocount of 1) and the X axis corresponds to the day of the year (month is shown for orientation, with the line ticks for the first day). Dates when a bloom of *Synechococcus* was detected through flow cytometry counts are labelled. B) Time series of *Synechococcus* abundance (cells ml⁻¹) during the 11 years. The data points are colored by water temperature (C). The grey dashed line indicates the samples presenting more than 50000 cells ml⁻¹.

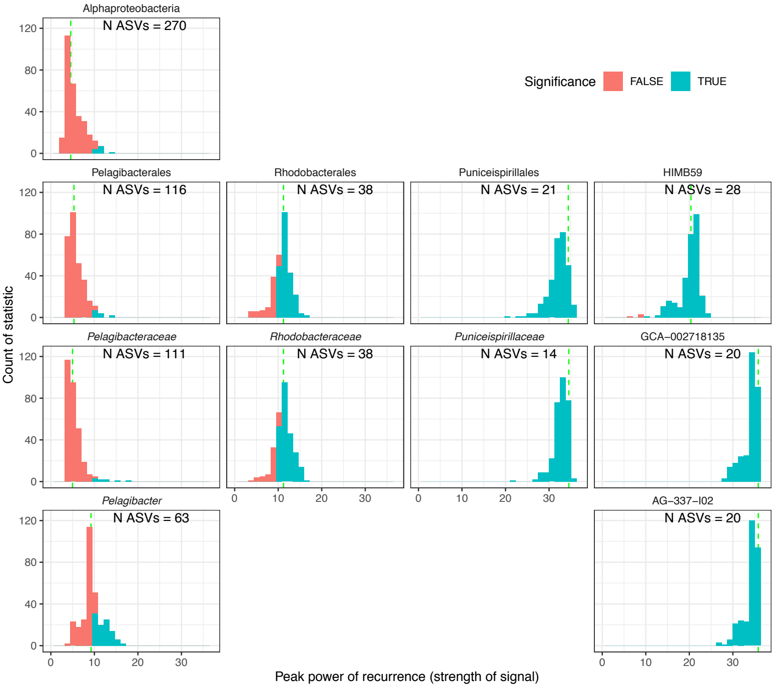


Figure S9: A) Histograms of the peak normalized power statistic at the class, order, family and genus level (from top to bottom, each line is a rank) of 80% of the ASVs conforming the rank for class Alphaproteobacteria. The red bins indicate the non-significant results (PN10, *q*≤0.01) and blue bins the significant ones. The dashed green line represents the statistic including all ASVs.

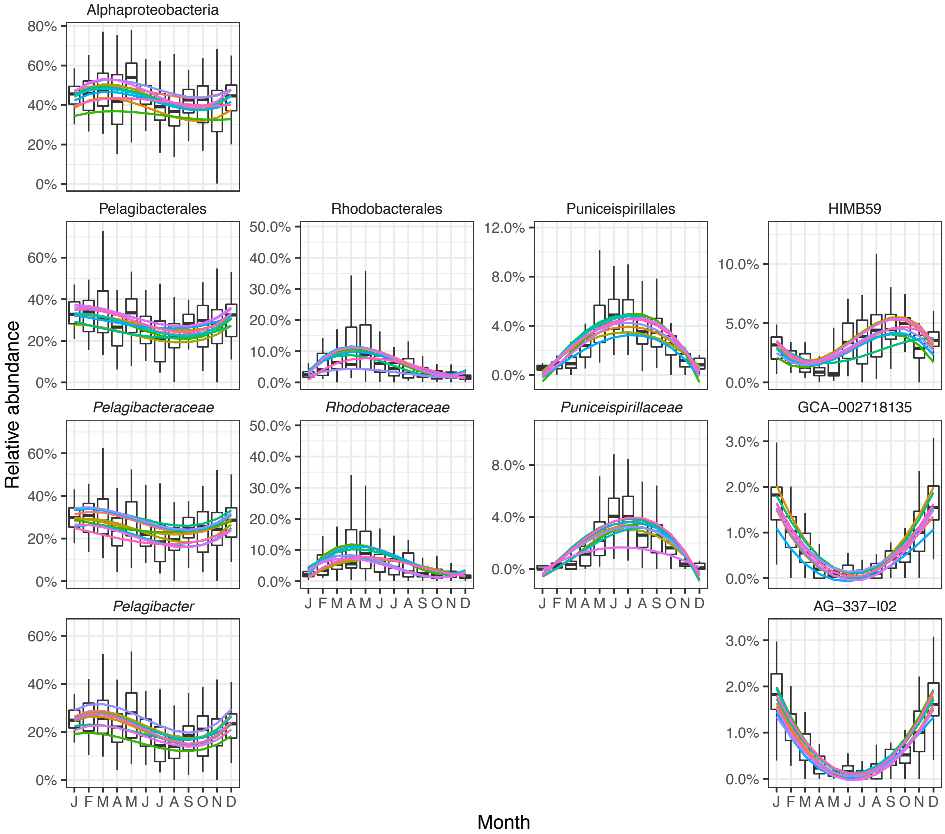


Figure S9: B) Relative abundance distribution of a random selection of 80% of the ASVs calculated 10 times (each line in a different color). Each boxplot presents the median and the 25% and 75% limits of the distribution of 110 points, and whiskers represent 1.5 times the interquartile range. The line is a smooth fitting of the change over time, with a color for each randomization.

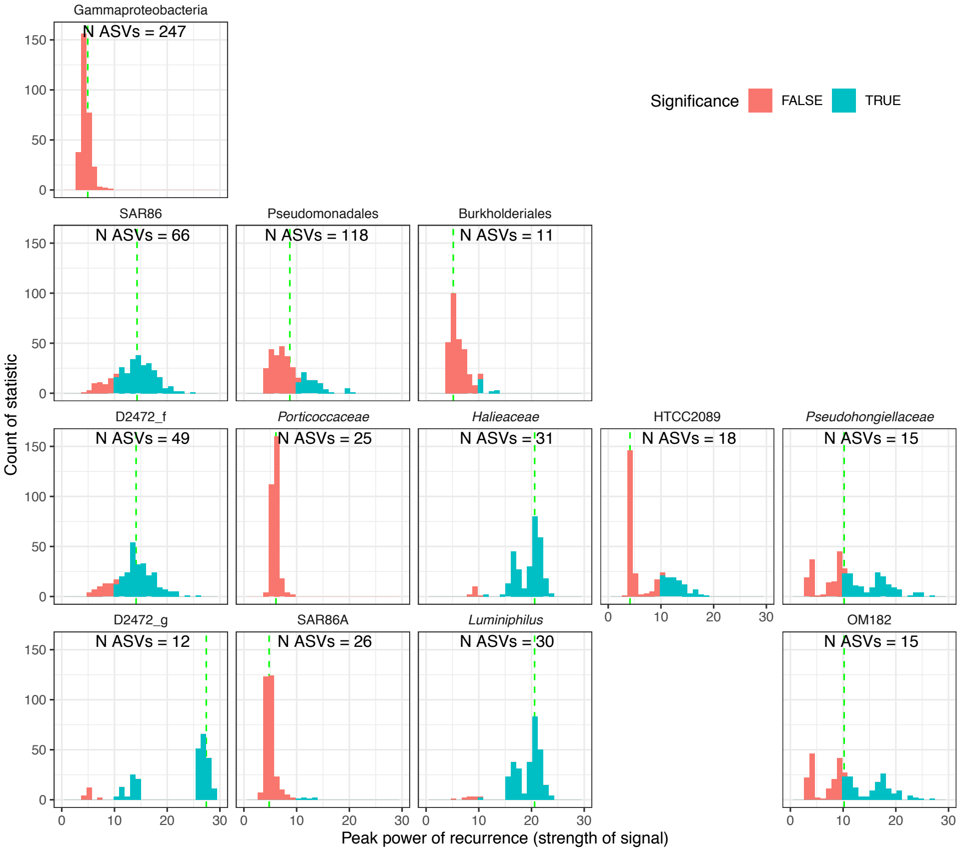


Figure S10: A) Histograms of the peak normalized power statistic at the class, order, family and genus level (from top to bottom, each line is a rank) of 80% of the ASVs conforming the rank for class Gammaproteobacteria. The red bins indicate the non-significant results (PN10, *q*≤0.01) and blue bins the significant ones. The dashed green line represents the statistic including all ASVs.

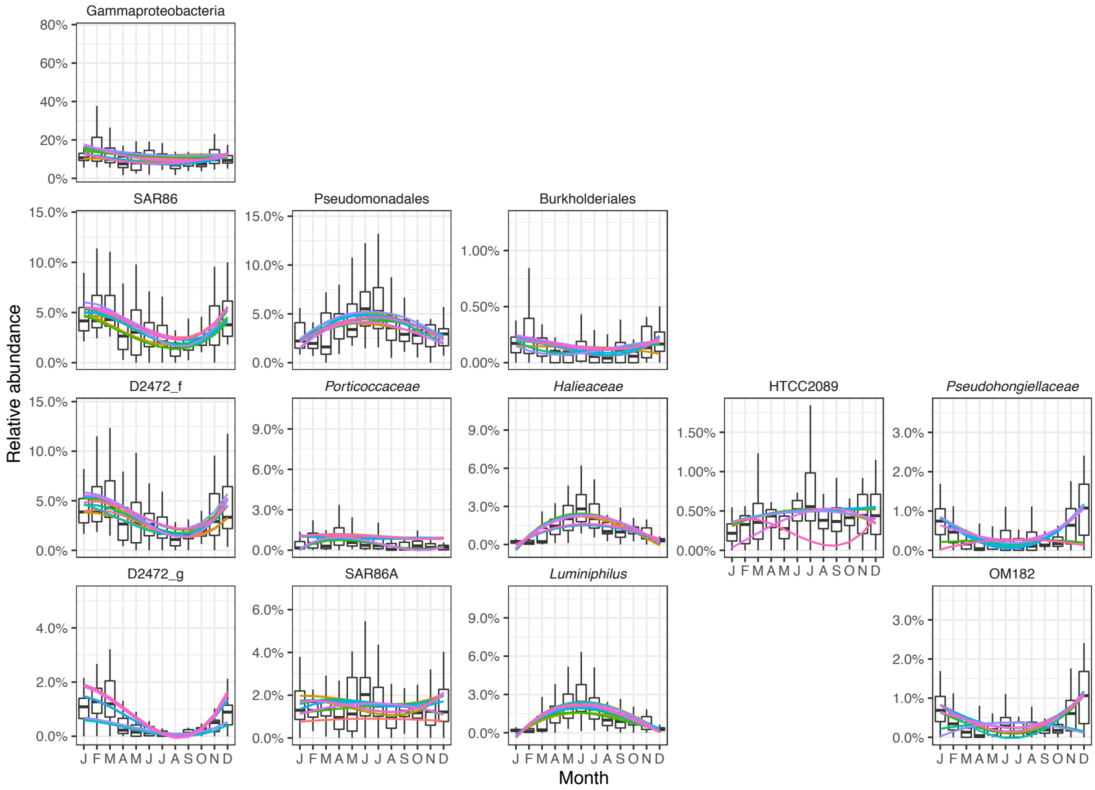


Figure S10: B) Relative abundance distribution of a random selection of 80% of the ASVs calculated 10 times (each line in a different color). Each boxplot presents the median and the 25% and 75% limits of the distribution of 110 points, and whiskers represent 1.5 times the interquartile range. The line is a smooth fitting of the change over time, with a color for each randomization.

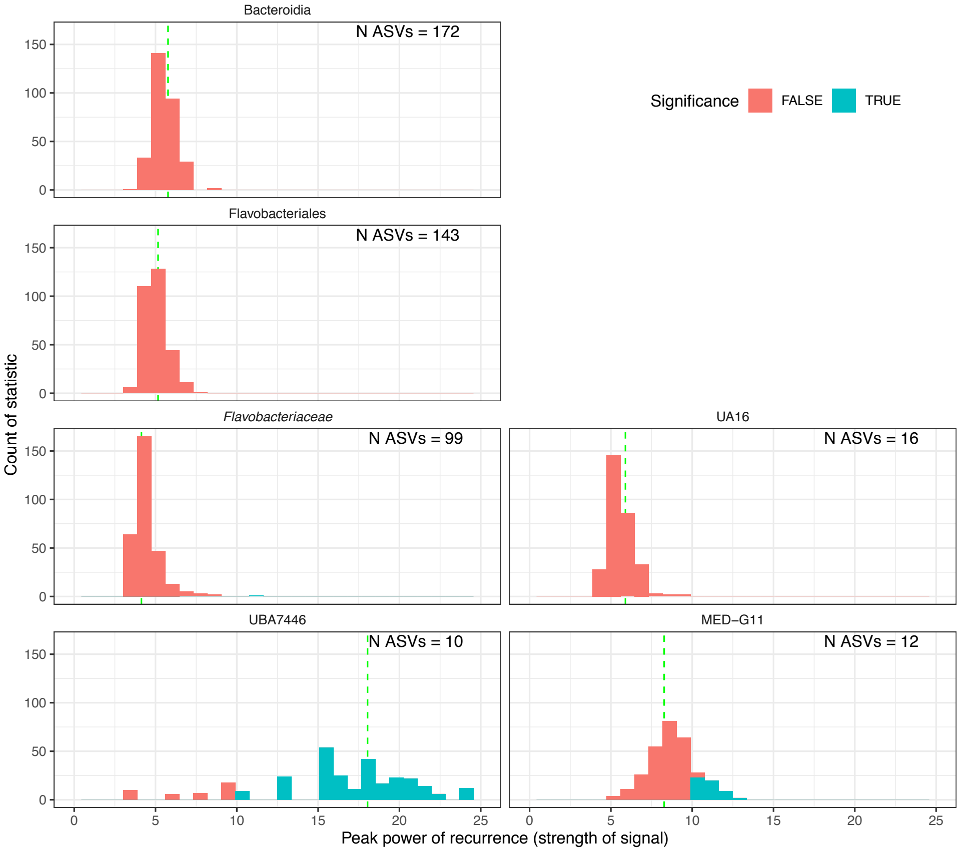


Figure S11: A) Histograms of the peak normalized power statistic at the class, order, family and genus level (from top to bottom, each line is a rank) of 80% of the ASVs conforming the rank for class Bacteroidia. The red bins indicate the non-significant results (PN10, *q*≤0.01) and blue bins the significant ones. The dashed green line represents the statistic including all ASVs.

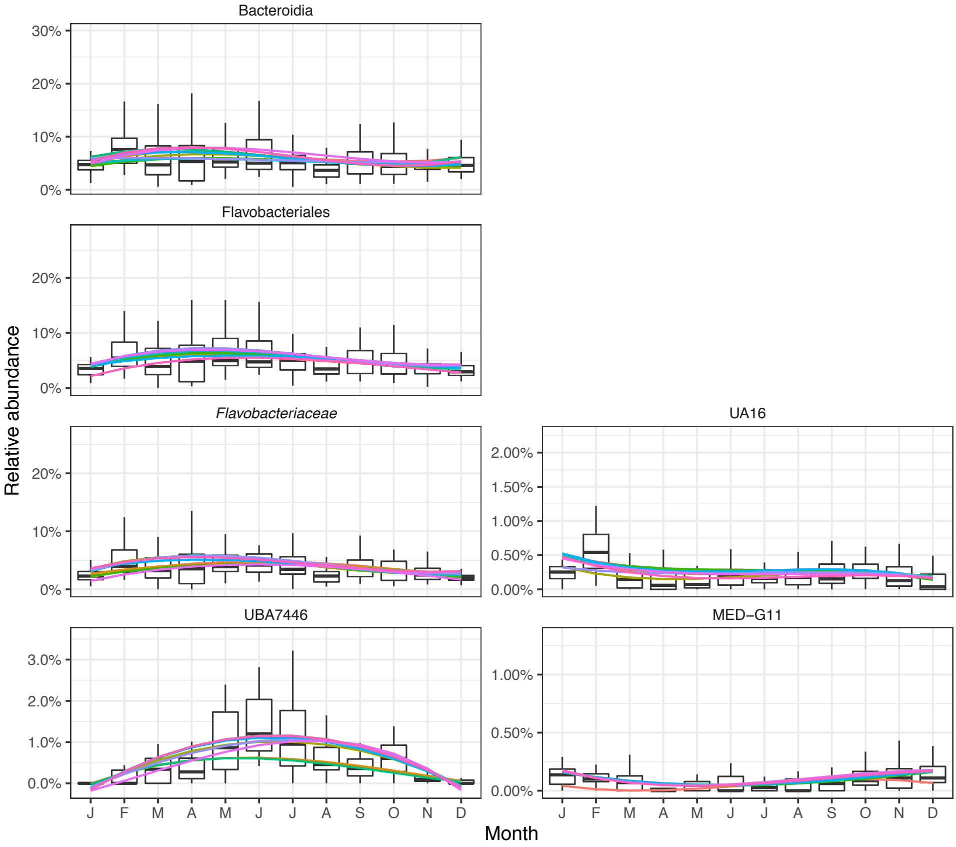
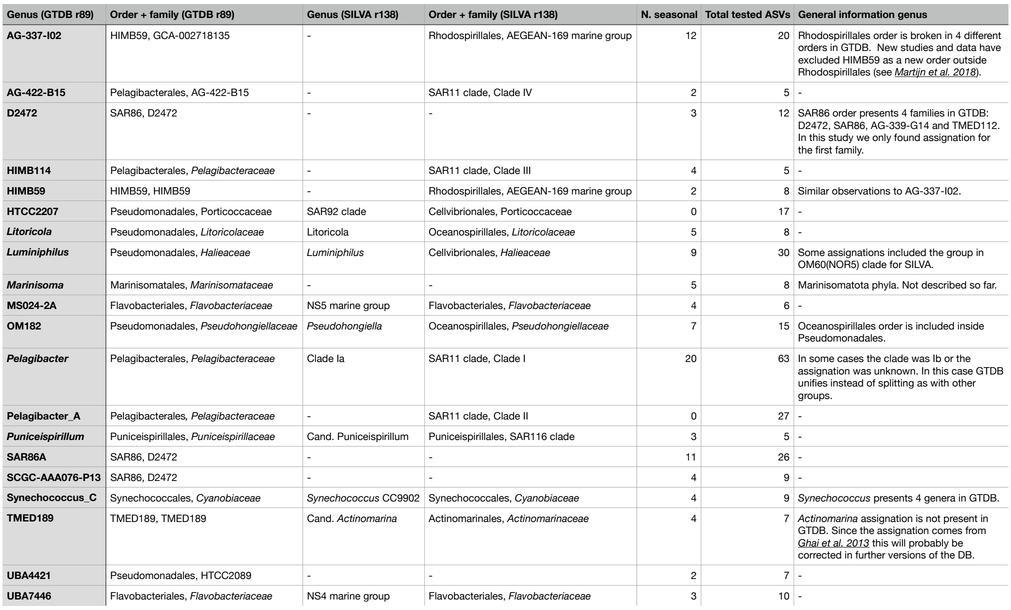


Figure S11: B) Relative abundance distribution of a random selection of 80% of the ASVs calculated 10 times (each line in a different color). Each boxplot presents the median and the 25% and 75% limits of the distribution of 110 points, and whiskers represent 1.5 times the interquartile range. The line is a smooth fitting of the change over time, with a color for each randomization.

1.9 Supplementary tables

link Supplementary Table 1

Supplementary Table 1: Taxonomy and occurrence distribution of each ASV. ASV name, taxonomy (from domain to genus), abundance category (abundant or rare), distribution (broad, intermediate or narrow), Conditionally Rare taxa (CRT) and ASV seasonality.



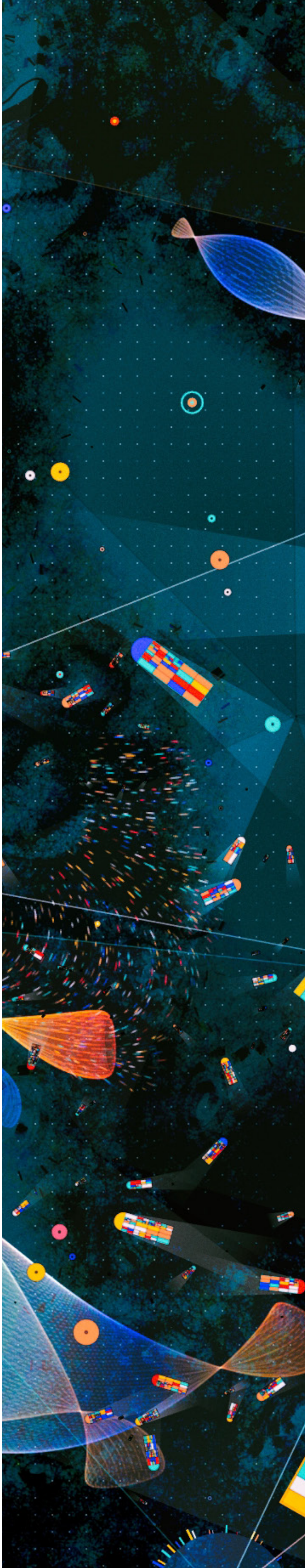
Genus (GTDB r89)	Order + family (GTDB r89)	Genus (SILVA r138)	Order + family (SILVA r138)	N. seasonal	Total tested ASVs	General information genus
AG-337-I02	HIMB59, GCA-002718135	-	Rhodospirillales, AEGEAN-169 marine group	12	20	Rhodospirillales order is broken in 4 different orders in GTDB. New studies and data have excluded HIMB59 as a new order outside Rhodospirillales (see *Martijn et al. 2018*).
AG-422-B15	Pelagibacterales, AG-422-B15	-	SAR11 clade, Clade IV	2	5	-
D2472	SAR86, D2472	-	-	3	12	SAR86 order presents 4 families in GTDB: D2472, SAR86, AG-339-G14 and TMED112. In this study we only found assignation for the first family.
HIMB114	Pelagibacterales, *Pelagibacteraceae*	-	SAR11 clade, Clade III	4	5	-
HIMB59	HIMB59, HIMB59	-	Rhodospirillales, AEGEAN-169 marine group	2	8	Similar observations to AG-337-I02.
HTCC2207	Pseudomonadales, Porticoccaceae	SAR92 clade	Cellvibrionales, Porticoccaceae	0	17	-
Litoricola	Pseudomonadales, *Litoricolaceae*	Litoricola	Oceanospirillales, *Litoricolaceae*	5	8	-
Luminiphilus	Pseudomonadales, *Halieaceae*	*Luminiphilus*	Cellvibrionales, *Halieaceae*	9	30	Some assignations included the group in OM60(NOR5) clade for SILVA.
Marinisoma	Marinisomatales, *Marinisomataceae*	-	-	5	8	Marinisomatota phyla. Not described so far.
MS024-2A	Flavobacteriales, *Flavobacteriaceae*	NS5 marine group	Flavobacteriales, *Flavobacteriaceae*	4	6	-
OM182	Pseudomonadales, *Pseudohongiellaceae*	*Pseudohongiella*	Oceanospirillales, *Pseudohongiellaceae*	7	15	Oceanospirillales order is included inside Pseudomonadales.
Pelagibacter	Pelagibacterales, *Pelagibacteraceae*	Clade Ia	SAR11 clade, Clade I	20	63	In some cases the clade was Ib or the assignation was unknown. In this case GTDB unifies instead of splitting as with other groups.
Pelagibacter_A	Pelagibacterales, *Pelagibacteraceae*	-	SAR11 clade, Clade II	0	27	-
Puniceispirillum	Puniceispirillales, *Puniceispirillaceae*	Cand. Puniceispirillum	Puniceispirillales, SAR116 clade	3	5	-
SAR86A	SAR86, D2472	-	-	11	26	-
SCGC-AAA076-P13	SAR86, D2472	-	-	4	9	-
Synechococcus_C	Synechococcales, *Cyanobiaceae*	*Synechococcus* CC9902	Synechococcales, *Cyanobiaceae*	4	9	*Synechococcus* presents 4 genera in GTDB.
TMED189	TMED189, TMED189	Cand. *Actinomarina*	Actinomarinales, *Actinomarinaceae*	4	7	*Actinomarina* assignation is not present in GTDB. Since the assignation comes from *Ghai et al. 2013* this will probably be corrected in further versions of the DB.
UBA4421	Pseudomonadales, HTCC2089	-	-	2	7	-
UBA7446	Flavobacteriales, *Flavobacteriaceae*	NS4 marine group	Flavobacteriales, *Flavobacteriaceae*	3	10	-

Supplementary Table 2: Correspondence between the GTDB and SILVA nomenclature. The first two columns correspond to the genus, family and order from the GTDB r89, and the next two provide the same information in SILVA DB r138. The column "N. seasonal" indicates the number of seasonal ASVs from the total of ASVs tested. Finally, the column "General Information Genus" provides useful information behind some of the changes in the nomenclature.

Genus	*df*	logLik	AIC	BIC	deviance	*df* (residual)	*p*	R^2
Pelagibacter	2	171.5	−337.1	−325.2	9.1	380	<0.0001	0.126
SAR86A	2	15.2	−24.3	−21.0	0.3	20	0.052	0.135
Litoricola	2	14.3	−22.6	−19.9	0.2	16	0.683	−0.051
Pelagibacter_A	2	18.7	−31.5	−25.2	1.8	57	0.003	0.130
Synechococcus_C	2	2.6	0.8	2.7	0.6	12	0.89	−0.082
Luminiphilus	2	18.2	−30.4	−26.6	0.4	25	0.13	0.053
AG-337-I02	2	11.3	−16.6	−13.4	0.5	20	0.19	0.038

Supplementary Table 3: Linear regression coefficients for each genus between Rho proportionality values and nucleotide divergence. Df, degrees of freedom; logLik, log likelihood; AIC, Akaike Information Criterion; BIC Bayesian Information Criterion; deviance; residual degrees of freedom; *p* values of the coefficient; R^2 values.

CHAPTER II



Chapter II

Long-term seasonal and interannual variability of marine aerobic anoxygenic photoheterotrophic bacteria

Adrià Auladell, Pablo Sánchez, Olga Sánchez, Josep M. Gasol, Isabel Ferrera

ISME Journal (2019). DOI: 10.1038/s41396-019-0401-4

Abstract

We studied the long-term temporal dynamics of the aerobic anoxygenic phototrophic (AAP) bacteria, a relevant functional group in the coastal marine microbial food web, using high-throughput sequencing of the *pufM* gene coupled with multivariate, time series and co-occurrence analyses at the Blanes Bay Microbial Observatory (NW Mediterranean). Additionally, using metagenomics, we tested whether the used primers captured accurately the seasonality of the most relevant AAP groups. Phylogroup K (Gammaproteobacteria) was the greatest contributor to community structure over all seasons, with phylogroups E and G (Alphaproteobacteria) being prevalent in spring. Diversity indices showed a clear seasonal trend, with maximum values in winter, which was inverse to that of AAP abundance. Multivariate analyses revealed sample clustering by season, with a relevant proportion of the variance explained by day length, temperature, salinity, phototrophic nanoflagellate abundance, chlorophyll *a* and silicate concentration. Time series analysis showed robust rhythmic patterns of co-occurrence, but distinct seasonal behaviors within the same phylogroup, and even within different Amplicon Sequence Variants (ASVs) conforming the same Operational Taxonomic Unit (OTU). Altogether, our results picture the AAP assemblage as highly seasonal and recurrent but containing ecotypes showing distinctive temporal niche partitioning, rather than being a cohesive functional group.

2.1 Introduction

Bacteria are extremely abundant and diverse in the ocean where they drive most biogeochemical cycles. Recent developments in sequencing technologies have allowed studying microbial diversity at unprecedented scales. Mapping microbial communities in hundreds of samples from recent global expeditions has resulted in a comprehensive picture of how they vary across space (Yooseph *et al.*, 2007; Salazar *et al.*, 2015; Sunagawa *et al.*, 2015). Likewise, long-term microbial observatories are key to understand microbial variation over time, particularly in temperate zones encompassing contrasting meteorological seasons (Kane, 2004; Buttigieg *et al.*, 2018). To date, different seasonal studies conducted in fixed stations in the Atlantic (Bermuda Atlantic Time Series Study, Western English Channel Time Series) and Pacific (Hawaii Ocean Time Series, San Pedro Ocean Time Series (SPOT)) Oceans, and in the Mediterranean Sea (Service d'Observation du Laboratoire Arago Station) concur that plankton turnover is mostly driven by the environment, and that the seasonal patterns are repeatable over time (Gilbert *et al.*, 2012; Fuhrman *et al.*, 2015; Galand *et al.*, 2018).

Thus far, most seasonal studies have focused on determining the variation of phylogenetic groups based on 16S or 18S rRNA gene sequencing for bacterioplankton and eukaryotic plankton respectively (Kim *et al.*, 2014; Fuhrman *et al.*, 2015; Martin-Platero *et al.*, 2018; Giner *et al.*, 2019). However, these phylogenetic units may include different ecotypes given that closely related or even identical rRNA gene-identified species can possess different functional traits (Martiny *et al.*, 2013) as a result of processes such as horizontal gene transfer (HGT) that can disconnect functional from phylogenetic diversity (Louca *et al.*, 2016). While a considerable amount of information on the seasonality of bulk microbial communities and of some particular phylogroups exists (i.e. Galand *et al.*, 2010), the seasonality of individual functional groups is barely known.

A functional guild of particular interest is the polyphyletic (i.e., derived from more than one common ancestor through HGT) aerobic anoxygenic phototrophic (AAP) bacteria. These organisms have the ability of photoheterotrophy, that is, they are capable of using both organic matter and light as energy sources (Koblížek, 2015). Their discovery challenged previous simplistic views of the structure of ocean microbial food webs (Fenchel, 2001). AAP bacteria are relatively common in the euphotic zone of the oceans (Schwalbach and Fuhrman, 2005; Jiao *et al.*, 2007; Lami *et al.*, 2007; Cottrell and Kirchman, 2009; Ritchie and Johnson, 2012), exhibit faster growth rates than other bacterioplankton groups (Ferrera *et al.*, 2011; Ferrera, Sánchez, *et al.*, 2017) and their cells are in general larger than most marine heterotrophic bacteria (Sieracki *et al.*, 2006). Altogether, these characteristics make them relevant in the ecosystem by processing a large amount of organic matter (see review by Koblížek, 2015).

Phylogenetically, the AAPs belong to the Alphaproteobacteria, Betaproteobacteria and Gammaproteobacteria. However, since these organisms acquired the ability of photoheterotrophy through HGT, the 16S rRNA gene, typically used for identifying prokaryotes, cannot be used as a genetic marker of AAPs in environmental studies. Alternatively, the *pufM* gene, present in all anoxygenic phototrophs containing type-2 reaction centers, is routinely used in AAP diversity surveys. Based on the phylogeny of this gene and the structure of the *puf* operon, Yutin *et al.* (2007) defined 12 distinct phylogroups (named from A to L) using metagenomic data from the Global Ocean Survey. Currently, the taxonomic assignation of short environmental sequences of the *pufM* gene is commonly done using this 12-phylogroup classification. In recent years, several authors have investigated their diversity and community structure in relation to environmental gradients across spatial scales using the variability of this marker gene (Jiao *et al.*, 2007; Yutin *et al.*, 2007; Ritchie and Johnson, 2012; Boeuf *et al.*, 2013; Bibiloni-Isaksson *et al.*, 2016; Lehours *et al.*, 2018) but much less is known about their temporal dynamics. Two independent studies conducted in the NW Mediterranean (Ferrera *et al.*, 2014) and the East coast of Australia (Bibiloni-Isaksson *et al.*, 2016) examined the variability of AAPs using *pufM* amplicon sequencing and showed that these assemblages seem to be highly dynamic. These two studies analyzed only one year of samples but long-term surveys are necessary to understand their seasonal and interannual patterns of biodiversity, stability, predictability, interactions between species, and responses to environmental changes.

We present here the first long-term exploration of marine AAP assemblages using Illumina sequencing of the amplified *pufM* gene from monthly samples taken over 10 years at the coastal Blanes Bay Microbial Observatory (BBMO) in the NW Mediterranean Sea. We define the temporal patterns and unveil their recurrence, explore the long-term interactions between the different phylogroups, and identify the main environmental drivers acting upon the observed patterns. Taking advantage of the recent appearance of threshold-free algorithms for Amplicon Sequence Variants (ASV) analysis, which surpass the clustering of sequences based on similarity cutoffs (Eren *et al.*, 2015; Callahan *et al.*, 2016), we have gone one step beyond previous studies and explored the seasonality of ASVs potentially representing different AAP ecotypes at a more fine-grained level. These analyses ultimately allow us to explore the level of ecological consistency within the different phylogenetic clades, i.e. whether the different AAP phylotypes are ecologically cohesive or, contrarily, each phylogroup includes organisms presenting temporal niche partitioning. Additionally, by comparing the sequences recovered through amplicon sequencing to those extracted from metagenomes, we test whether the used primers are adequate to evaluate the seasonality of the dominant AAP groups.

2.2 Material and Methods

Location and sample collection

Surface water was collected monthly as described elsewhere (Ferrera *et al.*, 2014) from the Blanes Bay Microbial Observatory (41°40'N, 2°48'E), a shallow (~20 m) coastal site about 1 km offshore in the NW Mediterranean coast. A total of 120 samples, from January 2004 to December 2013 were collected and in situ prefiltered through a 200-µm mesh. Several environmental parameters were measured alongside sample collection as described in Supplementary Information 1. The measured variables as well as day length were included in an environmental data table containing a total of 23 biotic and abiotic variables that was used for statistical analysis. The environmental data are shown in Figures S1 and S2. The astronomical seasons (based on equinoxes and solstices) were used for establishing spring, summer, autumn and winter periods. Additionally, the mixing layer depth (MLD) was obtained for the first months of 2004 and from 2008 to 2010 as defined in Galí *et al.*, 2013.

DNA extraction, *pufM* amplification, quantification, sequencing and sequence processing

About 6 L of 200 µm pre-filtered surface seawater were sequentially filtered through a 20 µm mesh, a 3 µm pore-size polycarbonate filter (Poretics) and a 0.2 µm Sterivex Millipore filter using a peristaltic pump. Sterivex units were filled with 1.8 mL of lysis buffer (50 mM Tris-HCl pH 8.3, 40 mM EDTA pH 8.0 and 0.75 M sucrose), kept at -80°C and extracted using the phenol-chloroform protocol as in Massana *et al.* (1997). Note that AAP bacteria attached to particles larger than 3 µm were not the subject of this study.

Partial amplification of the *pufM* gene (~245 bp fragments) was done in 50 µl reactions using primers *pufM* forward (5'-TACGGSAACCTGTWCTAC-3', Béjà *et al.*, 2002) and *puf*_WAW reverse (5'-AYNGCRAACCACCANGCCCA-3', Yutin *et al.*, 2005), each at 0.2 µM final concentration. The final concentration of MgCl2 was 2 mM. PCR conditions were as follows: an initial denaturation step at 95°C for 5 min and 35 cycles at 95°C (30s), 58°C (30s), 72°C (40s) and a final elongation step at 72°C for 10 min. Sequencing was performed in an Illumina MiSeq sequencer (2 x 250 bp, Research and Testing Laboratory; http://rtlgenomics.com/). Primers and spurious sequences were trimmed using cutadapt v1.14 (Martin, 2011). DADA2 v1.4 was used to differentiate exact sequence variants and remove chimeras (parameters: maxN= 0, maxEE= c(2,4), trunclen= c(200,200)) (Callahan *et al.*, 2016). DADA2 resolves Amplicon Sequence Variants (ASVs) by modeling the errors in Illumina-sequenced amplicon reads. The approach is threshold-free, inferring exact variants up to 1 nucleotide of difference using the quality score distribution in a probability model. For comparison purposes, the ASVs were clustered using UCLUST (Edgar, 2010) at 94% of nucleic acid sequence similarity, a threshold typically used for the *pufM* gene (Zeng *et al.*, 2007). After filtering for chimeras and spurious sequences with DADA2, 74% of the initial reads (mean 25692, min 4172, max 135331) were kept for downstream analyses. Sample BL120313 (13 March 2012) was discarded due to low read counts (836 reads). DADA2 read filtering details can be found in Supplementary Table 1. Moreover, in order to determine whether the primers used in this PCR-based approach captured the seasonality patterns accurately, we used 35 metagenomes generated from the same time-series (samples from years 2011 to 2013; see a detailed explanation in Supplementary Information 2) for comparison. Copy numbers of the marker gene *pufM* were estimated by quantitative polymerase chain reaction (qPCR) as described in Ferrera *et al.*, (2017b) (see Supplementary Information 3 for details).

Phylogenetic classification

A custom-made database was generated combining sequences from previous AAP studies (Cuadrat *et al.*, 2016; Graham *et al.*, 2018; Lehours *et al.*, 2018; Yutin *et al.*, 2007), variants present in the Integrated Microbial Genomes system (Markowitz *et al.*, 2006) and other *pufM* sequences from the GenBank database. Additionally, the predicted sequences from the BBMO metagenomes were included in this database. The nucleotide sequences were aligned with the guidance of amino acid translations using TranslatorX (Abascal *et al.*, 2010), with a posterior manual curation after filtering the sequences by length (>600 bp). From the alignment, a phylogenetic tree was constructed with RAxML v8.2 (Stamatakis, 2014) (GTRGAMMA model, 1000 bootstraps), and the phylogroups were delimited in the resulting tree using iTOL (Letunic and Bork, 2016) (Supplementary Figure 3). Afterwards, the phylogenetic placement of the amplicon nucleotide sequences was performed with the Evolutionary Placement Algorithm v0.2 (Barbera *et al.*, 2018) to establish their phylogroup classification. Finally, to determine potential primer biases, the forward and reverse primers were contrasted against the nucleotide alignment.

Statistical analyses

All analyses were performed using the R language, with phyloseq and vegan packages (McMurdie and Holmes, 2013; R Core Team, 2014; Oksanen *et al.*, 2018). Alphadiversity was analyzed using the Chao1 and Shannon indices (Smith and Wilson, 1996). Betadiversity was analyzed using a Bray-Curtis dissimilarity matrix with a previous normalization through rarefying to 4172 reads per sample (Faith *et al.*, 1987; Weiss *et al.*, 2017). We used distance-based Redundancy Analysis (dbRDA, Legendre and Legendre, 1988) to find the environmental predictors (scaled to the mean) that best explained the patterns of community structure and diversity of AAPs over time, with a previous multivariate non-parametric ANOVA for selecting significant variables ($p \leq 0.01$). A time-decay analysis of the assemblage was computed excluding rare ASVs as recommended elsewhere (Faust *et al.*, 2015). ASVs were considered rare when presenting less than 1% of relative abundance from the rarefied dataset following Alonso-Sáez *et al.* (2015) criterion.

Time series analysis

Fourier time series analysis was performed to study the AAP assemblage dynamics over a decade. An interpolation of the discarded sample (BL120313) was used to maintain equidistant time points. Values were normalized through the Aitchison log-centered ratio transformation (CLR), adequate for compositional data (Gloor *et al.*, 2017). A Fisher G-test with the R package *GeneCycle* was used to determine the significance ($p \leq 0.01$) of the periodic components (Ahdesmaki *et al.*, 2015). The time series was decomposed in three components - seasonal periodicity (oscillation inside each period), trend (evolution over time) and residuals -through local regression by the stl function. Additionally, the autocorrelogram was calculated using the *acf* function.

Network construction

We used Local Similarity Analysis (LSA) (Xia *et al.*, 2011; Durno *et al.*, 2013) with a previous CLR transformation for network construction. Briefly, given a time series data and a delay limit, LSA finds the configuration that yields the highest local similarity (LS) score. Only the ASVs present in >5 samples and the environmental variables presenting <5% of missing values were used. The remaining missing values for the variables after filtering were estimated by imputation with the *mice* package (Azur *et al.*, 2012). Only Interactions with LS >=0.5, $p \leq 0.001$ and 1-month delay were considered. The network was plotted using the ggraph package (Pedersen, 2017).

Reproducibility

The code for preprocessing and statistical analyses along with package versions is available in the following repository: https://gitlab.com/aauladell/AAP_time_series. Sequence data has been deposited in Genbank under accession number PRJNA449272.

2.3 Results and discussion

Patterns of community composition and structure.

The amplicon sequences retrieved from 10 years of sampling resulted in a total of 820 ASVs whereas the number of OTUs was 406 (94% similarity cutoff). Of the total ASVs, 276 presented only 1 nucleotide variation between sequences. In comparison with previous temporal studies (82 OTUs detected in Ferrera *et al.*, 2014, 89 in Bibiloni-Isaksson *et al.*, 2016), our study presents a more complete picture of *pufM* diversity and is the largest dataset of AAP diversity reported to date. Estimates of richness were higher during winter (mean 51, max 126 observed ASVs), decreasing to minimum values in the spring-summer period, precisely during May-August (mean 35, max 77) (Figure 1). The differences between winter and spring/summer were statistically significant (ANOVA, $p \leq 0.05$). A similar trend was observed when computing the Shannon index (Figure 1). Comparing the amplicon with the metagenomic data from 2011 to 2013, we observed that whereas 188 OTUs and 357 ASVs were present in the amplicons for that period, only a total of 84 different *pufM* sequences were recovered from the metagenomes. However, the Shannon diversity index for the two datasets presented a positive correlation (Pearson R=0.81, p=0.001, N=35) and the followed the same trend of increasing values in winter (Figure 1).

A notable negative correlation between day length and the Shannon index was observed (Pearson R=-0.57, $p \leq 0.01$, N=119). That relationship of diversity with day length had previously been observed in long-term bulk bacterioplankton community studies (Gilbert *et al.*, 2012), as well as with specific phylogenetic groups such as the SAR11 (Salter *et al.*, 2015). A possible explanation is that the deep winter mixing allows the development of high diversity assemblages in contrast to the selection of specific oligotrophic ecotypes occurring during the stratified summer season (Salter *et al.*, 2015). In fact, mixed layer depth was a significant predictor of the Shannon diversity index (Spearman R=0.56, $p \leq 0.001$, N=35, Supplementary Figure 4). Interestingly, this trend of higher alphadiversity in winter is opposed to that of AAP abundance (Supplementary Figure 5); higher abundances of the *pufM* gene during spring and summer were measured by qPCR as compared to winter and fall ($p \leq 0.01$). We also found a positive correlation between the qPCR data and the abundance of *pufM* sequences retrieved in the 3 years of metagenomes (Pearson R=0.77, $p \leq 0.001$, N=12) (Supplementary Figure 6), in which the higher abundances were found in spring followed by summer. These results support previous observations obtained through various methodological approaches (Ferrera *et al.*, 2014 using microscopy counts; Bibiloni-Isaksson *et al.*, 2016 using qPCR and Galand *et al.*, 2018 using metagenomics) and confirm that there is a clear inverse relationship between AAPs abundance and diversity.

Regarding community composition across the decadal period (Figure 2, Supplementary Figure 7), phylogroup K (*Gammaproteobacteria*) affiliated to the NOR5/OM60 clade) was the most prevalent and dominant over the years (83.8% ± SE 2.3, mean relative abundance), in agreement with previous reports for this station (Ferrera *et al.*, 2014) and for other regions such as the Baltic Sea (Mašín *et al.*, 2006). Yet, a decrease in their contribution was observed during February and March (down to 59.6% and 52% on average respectively). During these months, the contribution of phylogroup E (*Rhodobacter*-like) to community structure was greater, albeit with a high variation over the decade (±26% SD). The previous 1-year study of AAP diversity conducted by Ferrera *et al.* (2014) reported a

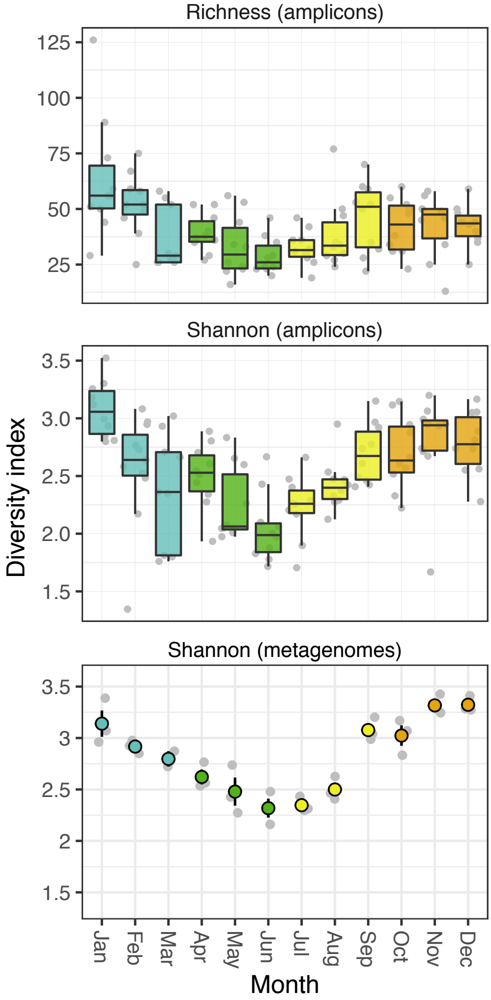


Figure 1. Alphadiversity distribution of the AAP community for each month colored by season. Richness (number of observed ASVs) and Shannon indexes obtained through amplicon sequencing over a decade (2004-2013) are shown in the top and middle panels respectively. Each boxplot presents the median and interquartile range of the distribution of 10 data points shown in grey (with the exception of March, with 9 data points). Whiskers represent 1.5 times the interquartile range. The bottom panel shows the Shannon index values obtained from the metagenomic dataset (2011 to 2013). The colored dots represent the mean monthly values and the bars the standard error of the mean for the 3-year period.

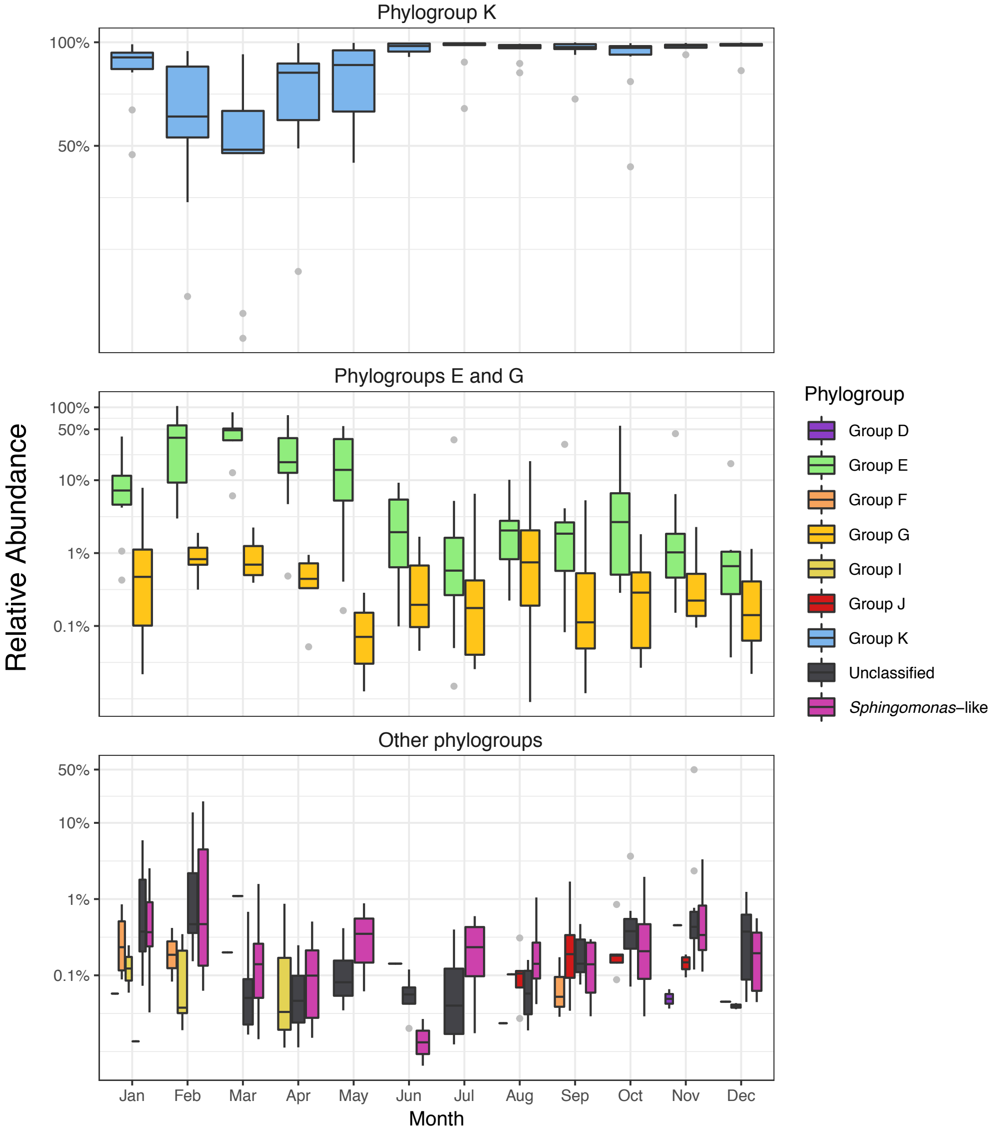


Figure 2. Variation in the relative contributions of AAP phylogroups K (top panel), G, E (middle panel), phylogroups D, F, I, J, *Sphingomonas*-like and the unclassified group (bottom panel) for each month over the studied decade (2004-2013). Each boxplot presents the median and the 25% and 75% limits with the distribution of 10 data points in grey (with the exception of March, with 9 data points), and whiskers represent 1.5 times the interquartile range.

similar observation and moreover, a study of the 16S rRNA gene diversity from the same location also suggested that Alphaproteobacteria dominate the bacterial assemblages during the local spring bloom (Alonso-Sáez *et al.*, 2007). Regarding phylogroups D, F (Rhodobacterales-like), H (uncultured), J (*Rhodospirillales*-like), I (*Betaproteobacteria*), *Sphingomonas*-like and the unclassified ASVs, these presented a mean relative contribution below 1% in the amplicon dataset. These groups displayed occasional peaks (>1% relative abundance) with no clear periodic trend, for example, *Sphingomonas*-like AAPs showed a contribution of 14% in February 2012 (Supplementary Figure 7). Overall, these observations are similar to those obtained from the metagenomic distribution (3 years instead of 10) with a good agreement in the relative abundance recovered for the prevailing phylogroups E and K as well as for the less common Sphingomonas-like sequences (correlation values: 0.9, 0.69 and 0.91 respectively; Table S2). Contrarily, phylogroup G seems to be underrepresented in the amplicon dataset likely due to the presence of 3 to 5 mismatches in the forward primer with the most abundant metagenomic variants.

Metagenomics is often considered the least biased approach for functional gene analysis since the method is PCR-independent and does not suffer from amplification biases that could result in misrepresentation of the relative abundances of certain populations. Nonetheless, for a given time and money investment, metagenomes retrieve less copies of specific marker genes, offering thus less inquiry potential if the main purpose of our study is the barcoding of a particular group of organisms. We found that richness estimates were higher using amplicon data since more variants of the *pufM* gene were recovered with that approach than from metagenomes, but the seasonal trends in diversity identified by both methodologies were remarkably similar (Figure 1). Likewise, in terms of community structure there was a good agreement for the most prevailing groups with the exception of phylogroup G. Moreover, the seasonal trends observed at the phylogroup and even at the sequence variant level recovered using these two distinct methodological approaches, presented a close resemblance (Figures S8 and S9, see *Seasonality at the fine scale* section below).

Noteworthy, the amplicon approach allowed identifying the seasonal tendencies of many more individual ecotypes that what would have been possible through metagenomics, while metagenomics captured some low abundance groups missing in the amplicon dataset. In particular, phylogroups A, B, C and L, accounting for a total relative abundance of 7.0, 3.1, 3.9 and 0.2% respectively, were only retrieved through metagenomics. Primer coverage analysis revealed that the forward primer contains between 3 and 8 mismatches with the metagenomic sequences from these phylogroups (details not shown), which could explain their absence in the amplicon dataset and why these groups are rarely reported in AAP surveys based on amplicon sequencing (Bibiloni-Isaksson *et al.*, 2016; Lehours *et al.*, 2018). Exceptionally, Ferrera *et al.*, 2014 reported the presence of one single OTU of group C contributing substantially (13% relative abundance) to the community during winter in Blanes Bay, which differs from the present results. To investigate this discrepancy, we carefully compared the sequence of this OTU to our updated database and found that it had been misclassified and it belongs to phylogroup K while does not show any significant similarity to the new phylogroup C sequences retrieved from the metagenomes. These observations highlight the need to increase the information present in databases to obtain accurate taxonomic assignations. In fact, only a few isolates from phylogroup K exist and none is available for phylogroup C, hampering the classification of these groups as discussed elsewhere (Caliz and Casamayor, 2014). In contrast, phylogroups F, H, I and J (<1% total relative abundance) were recovered only when using amplicons. Their low relative abundance possibly explains their absence in the metagenomic dataset. Overall, these results remark the need to undertake a revision of the primers typically used for high-throughput sequencing of *pufM* in order to increase their phylogenetic recovery but, at the same time, demonstrate that PCR-free metagenomics and amplicon-based approaches perform in a comparable fashion in recovering major AAP groups and, most importantly, that the seasonal patterns observed through amplicon sequencing are robust.

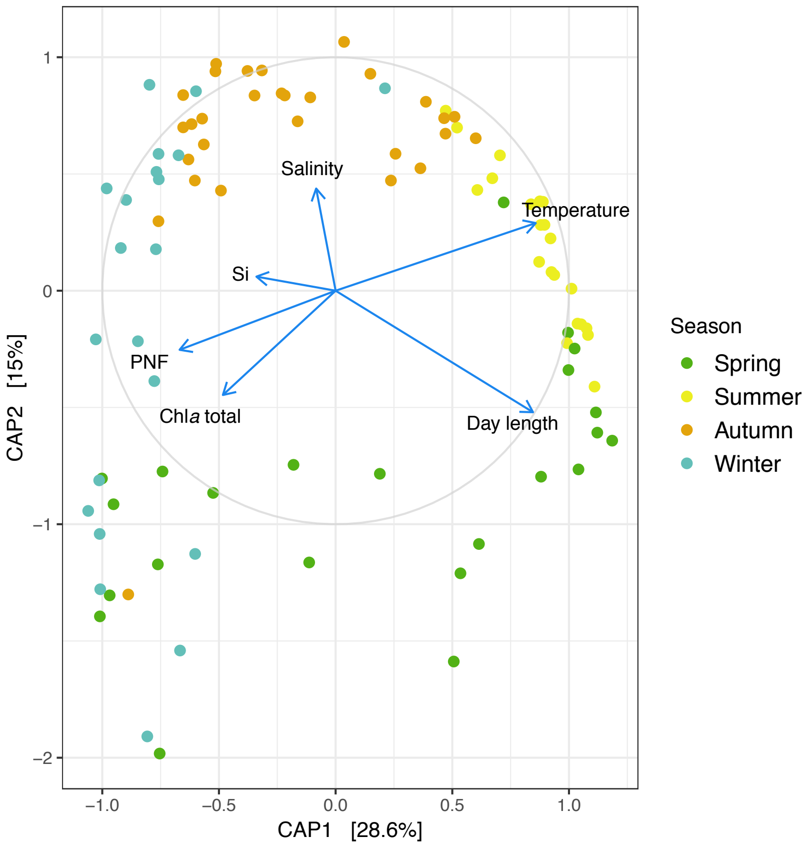


Figure 3. Distance based redundancy analysis of the samples (dots) with the 5 explanatory variables (arrows) influencing the distribution (PERMANOVA *p*≤0.01; day length, temperature, salinity, silicate concentration (Si), Chlorophyll *a* (Chla) and phototrophic nanoflagellate abundance (PNF)). The ordination was performed on the Bray-Curtis dissimilarity of log10 transformed data (with a pseudocount of 1) matrix (after rarefying). Samples are colored by season.

Patterns of betadiversity and recurrence.

Non-metric multidimensional scaling (nMDS) using various distance measurements indicated a clear separation of the samples at different temporal scales: by month (Bray Curtis, PERMANOVA R^2=0.51, *p*≤0.001) and by season (Bray Curtis, PERMANOVA R^2=0.31, *p*≤0.001, Supplementary Figure 10). Spring and winter samples were more dissimilar than those of summer or autumn. The reasons for this pattern are uncertain but could be related to higher date to date environmental variability or to the mixing of the water column that occurs during winter in this station (Gasol *et al.*, 2016).

Community structure was strongly linked to day length, temperature, salinity, phototrophic nanoflagellate abundance, chlorophyll *a* and silicate concentration, as revealed by distance-based redundancy analysis (dbRDA; Figure 3, PERMANOVA *p*≤0.01, Supplementary Information 4), which explained 51.4% of the variation with the first 2 axis explaining 43.6%. In particular, late spring and early summer samples were mostly influenced by day length and temperature, whereas autumn samples were partially influenced by salinity (Figure 3). Day length has previously been shown to explain the seasonal variability of the bulk bacterioplankton (Gilbert *et al.*, 2012) and AAP community structure (Ferrera *et al.*, 2014), but the mechanisms underlying this relationship are unclear. Interestingly, a group of samples from winter and spring appeared to be heavily influenced by the presence of ASVs (ASV8, ASV14 and ASV46) belonging to phylogroup E (*Rhodobacter*-like, Supplementary Figure 11), to the abundance of phototrophic nanoflagellate and the concentration of chlorophyll *a* (Figure 3), which could be related to the phytoplankton spring bloom that typically occurs in February-March in Blanes (Nunes *et al.*, 2017). The summer samples were associated to the high contribution of gammaproteobacterial ASV1 and the fall/early-winter cluster to more diverse communities of other gammaproteobacterial ASVs (Supplementary Figure 11).

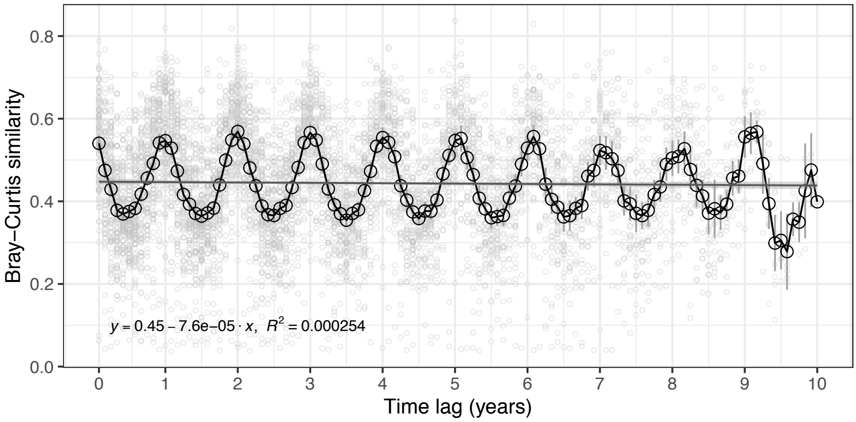


Figure 4. Bray-Curtis similarity between samples plotted against the time lag between each of them (time-decay plot). Mean similarity values for each time lag are plotted in an empty black dot with standard error bars (background grey filled dots show each comparison). A linear regression is plotted with 95% confidence intervals shown.

Finally, to explore the recurrence of the communities, the Bray-Curtis similarity between samples was plotted against the time lag resulting in the so-called time-decay curve (Figure 4) (Shade *et al.*, 2013; Fuhrman *et al.*, 2015). In our study, the assemblage was maintained over time with a median similarity of 0.45, with 6-month oscillations from the yearly maximum (~0.55) to the minimum (~0.39) values. These results indicate that AAP communities are under strong environmental selection that leads to a high seasonal behavior and translates into yearly repeatable communities. To our knowledge, this is the first time that the recurrence of a functional group of planktonic organisms, defined by a marker gene, has been demonstrated. Comparing the results to the 16S rRNA data from the SPOT and the Western Channel time series results we observe that the seasonal turnover at SPOT is less clear than in our location, and an initial decay of similarity is observed reaching a later plateau over time. In the Western Channel, the seasonality is equally marked but the initial decay is even more pronounced than in SPOT (see Figure 2 in Hatosy *et al.*, 2013). A possible explanation to these differences is that our comparison accounts only for a highly seasonal sub-community (as these organisms are able to use light, and light varies seasonally) while the overall bacterial/prokaryotic community responds to more variables. Further analyses with other functional genes should help understand whether these patterns are robust for distinct groups.

Patterns of co-occurrence.

A co-occurrence network was built with 127 ASVs present in >5 samples and 14 environmental variables, presenting 70 nodes and 142 edges after filtering by local similarity and significance (LS >= 0.5, $p \leq 0.001$,) (Figure 5). Noteworthy, most of the ASVs retained in the network were seasonal (46 out of 61) (see below). In terms of topology, the network presents one large cluster and other four minor clusters, being the largest one formed by 54 nodes mainly containing ASVs from phylogroups K and E, displaying multiple interactions with various ecosystem variables (temperature, day length and the abundance of phototrophic picoeukaryotes and nanoflagelates). Temperature was the variable presenting the largest number of the interactions (14), most of them being delayed one month. Out of these, many were of negative nature with *Gammaproteobacteria*-like ASVs that lower their relative abundance during summer (for example ASV26, 10, 11) while others were positive with ASVs that dominate the AAP community during this season (ASV1). Interestingly, many positive and negative interactions exist between various ASVs of phylogroup K and G and the abundance of phototrophic eukaryotes but none with other phylogroups such as phylogroup E or *Sphingomonas*-like. Strong biotic relationships between AAP species and phytoplankton have been reported, particularly with dinoflagellates (Biebl *et al.*, 2005; Yang *et al.*, 2018) and large fractions of particle-attached AAP bacteria have been observed in various marine environments (Waidner and Kirchman, 2007; Lami *et al.*, 2009). Here, we focused on the free-living fraction of AAPs but further interaction network analyses using both free-living and particle-attached AAP bacteria in combination with phototrophic eukaryotic species data would allow to investigate deeper these biotic relationships.

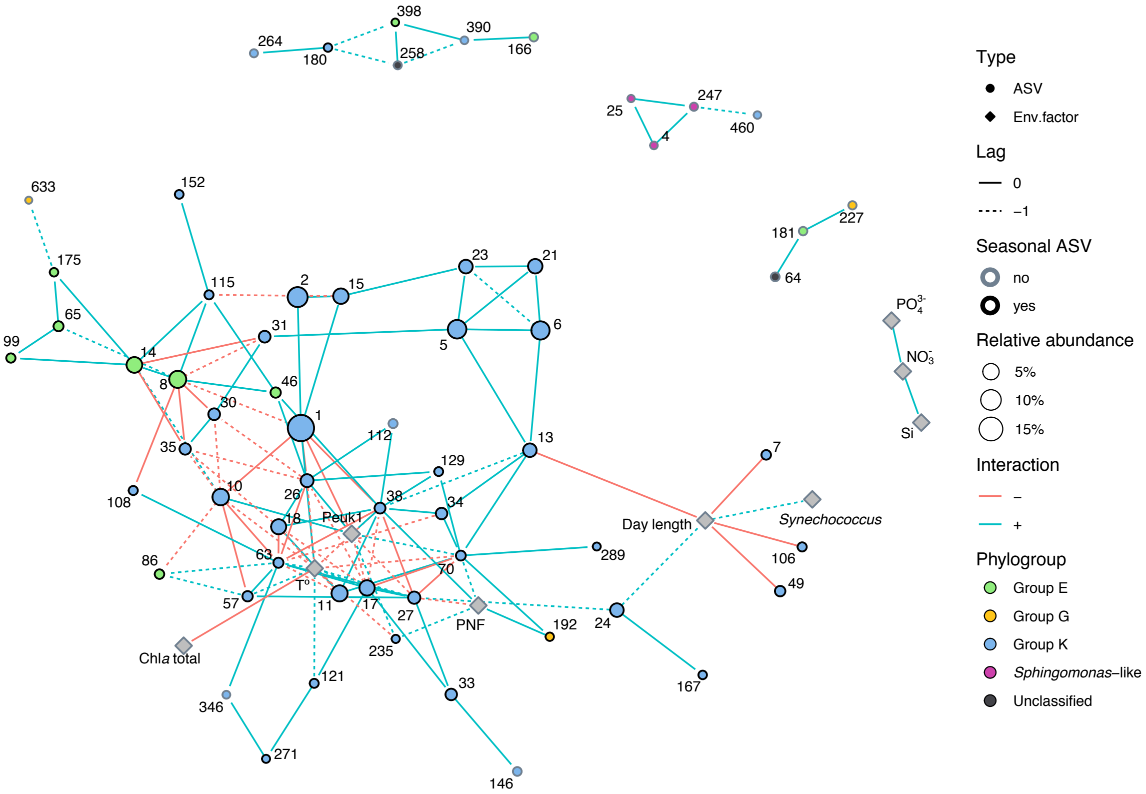


Figure 5. Fast local similarity network showing clusters with >=3 nodes. Node shape designates the type of variable, with the filling specifying the phylogroup, the size the total relative contribution and the stroke color if the ASV displays a seasonal behavior. Edges can be lagged (discontinuous line) or direct and have negative (i.e., anticorrelation) or positive local scores (LS). The label on the nodes indicates the ASV number. T°: temperature; PNF: abundance of phototrophic nanoflagellates; Peuk1: abundance of picoeukaryotes group I (see Supplementary Information 1 for details).

The majority of correlations occurred within rather than between phylogroups as previously observed (Bibiloni-Isaksson *et al.*, 2016); yet, whereas some groups presented mainly positive intergroup interactions (phylogroup E or *Sphingomonas*-like), phylogroup K showed positive and negative interactions between its ASVs. As an example, a clear negative correlation between ASV1-ASV26 and ASV10-ASV35, all of them being part of phylogroup K, can be observed in Figure 5. We also noticed that various ASVs of phylogroup E, such as ASV14 and ASV8 (*Alphaproteobacteria*-like), were positively related among them while presenting negative associations with ASVs from phylogroup K (*Gammaproteobacteria*-like ASV30 and ASV35). Negative correlations between phylogroup K and phylogroup G (also *Alphaproteobacteria*-like) had previously been reported (Ferrera *et al.*, 2014; Bibiloni-Isaksson *et al.*, 2016). These data thus point towards intergroup competition between members of the *Alpha*- and *Gammaproteobacteria*-like AAPs.

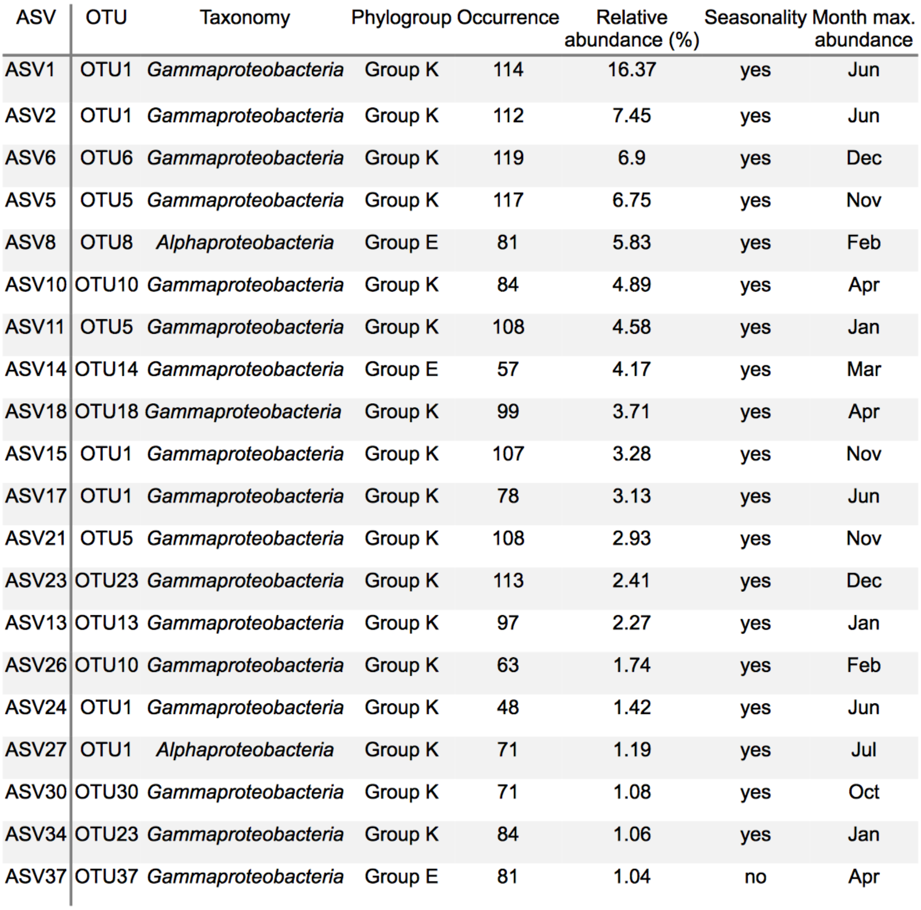
Looking at the interactions within closely related ASVs, i.e. those forming the same OTU, we observed multiple connections between them, for example, among ASV1, ASV2 and ASV15, all belonging to OTU1 or ASV14, ASV65 and ASV175, all forming OTU14. Nevertheless, network analysis revealed that sometimes there is a dissociation of these closely related ASVs, as seen for ASV17 and ASV27, belonging to OTU1 which do not present connections with other ASVs from the same OTU. These observations support the idea that ASVs may represent individual AAP ecotypes encompassing distinct ecological patterns and reflects the usefulness of breaking apart sequence clusters into variants in order to dig into the ecology of these organisms.

Seasonality at the fine scale.

The seasonality of each ASV was measured by evaluating if their relative abundance distribution presented a significant periodicity (Fisher G-test) through the long-term time series, and if so, by comparing them at different levels of resolution: across closely related sequences (ASVs) and across sequence clusters (OTUs and phylogroups). Seasonal patterns (*p*≤0.01) were present in 58 out of 127 ASVs analyzed (those ASVs present in >5 samples), affiliated to phylogroups K (44 ASVs), E (9) and G (2), J (1) and the unclassified group (2) (Table 1, Supplementary Table 3). In order to discard that potential amplification artifacts could influence the observed ASV seasonal trends, we mapped representative ASVs from the prevailing phylogroups E, G and K to the metagenomic sequences and compared the seasonal behavior in both datasets, obtaining a remarkable good concordance (Supplementary Figure 9).

The seasonal ASVs corresponded to 92% of the total read counts, and 83.4% of the counts corresponded to phylogroup K (*Gammaproteobacteria*). All periodicities found were of 1 year, with the exception of ASV152 (*Gammaproteobacteria*), that presented a periodicity of 2 years. Some of these ASVs always presented relative contributions above 1% regardless of season (all from phylogroup K), some presented values above 1% in a specific season (seasonal contributors), and other ASVs peaked (>1%) only occasionally (herein referred as opportunistic; see examples in Supplementary Figures 12-15). In fact, most ASVs presented an opportunistic behavior, with low contribution to total community composition during the decade and peaking occasionally. Various studies of the whole bacterioplankton community have observed this variety of strategies coexisting within a given clade (Shade *et al.*, 2014; Alonso-Sáez *et al.*, 2015; Fuhrman *et al.*, 2015). Our results reveal that this trend is maintained for this specific functional assemblage, with a few prevalent ecotypes and a larger pool of specialized ASVs, i.e. appearing within specific environmental conditions, within each phylum.

Comparing among the seasonal ASVs, we distinguished different behaviors. For example, ASVs divergent enough to form distinct OTUs but belonging to the same phylogroup did not always follow the same distribution (Figure 6A); e.g. for phylogroup K, the annual maxima of ASV1 occurred during June and July with a minimum in February/March, whereas ASV10 presents the



ASV	OTU	Taxonomy	Phylogroup	Occurrence	Relative abundance (%)	Seasonality	Month max. abundance
ASV1	OTU1	*Gammaproteobacteria*	Group K	114	16.37	yes	Jun
ASV2	OTU1	*Gammaproteobacteria*	Group K	112	7.45	yes	Jun
ASV6	OTU6	*Gammaproteobacteria*	Group K	119	6.9	yes	Dec
ASV5	OTU5	*Gammaproteobacteria*	Group K	117	6.75	yes	Nov
ASV8	OTU8	*Alphaproteobacteria*	Group E	81	5.83	yes	Feb
ASV10	OTU10	*Gammaproteobacteria*	Group K	84	4.89	yes	Apr
ASV11	OTU5	*Gammaproteobacteria*	Group K	108	4.58	yes	Jan
ASV14	OTU14	*Gammaproteobacteria*	Group E	57	4.17	yes	Mar
ASV18	OTU18	*Gammaproteobacteria*	Group K	99	3.71	yes	Apr
ASV15	OTU1	*Gammaproteobacteria*	Group K	107	3.28	yes	Nov
ASV17	OTU1	*Gammaproteobacteria*	Group K	78	3.13	yes	Jun
ASV21	OTU5	*Gammaproteobacteria*	Group K	108	2.93	yes	Nov
ASV23	OTU23	*Gammaproteobacteria*	Group K	113	2.41	yes	Dec
ASV13	OTU13	*Gammaproteobacteria*	Group K	97	2.27	yes	Jan
ASV26	OTU10	*Gammaproteobacteria*	Group K	63	1.74	yes	Feb
ASV24	OTU1	*Gammaproteobacteria*	Group K	48	1.42	yes	Jun
ASV27	OTU1	*Alphaproteobacteria*	Group K	71	1.19	yes	Jul
ASV30	OTU30	*Gammaproteobacteria*	Group K	71	1.08	yes	Oct
ASV34	OTU23	*Gammaproteobacteria*	Group K	84	1.06	yes	Jan
ASV37	OTU37	*Gammaproteobacteria*	Group E	81	1.04	no	Apr

Table 1. Summary information for the top 20 ASVs. The following columns are listed: ASV name, OTU correspondence, phylogroup correspondence, taxonomy, occurrence, relative abundance, seasonality behavior, and month of maximum mean relative abundance.

opposite distribution. Contrarily, most ASVs belonging to phylogroups G and E followed a similar trend among them, with their maxima in March, being ASV86 an exception presenting a maximum in September. Looking at a further level of resolution, i.e., comparing the seasonality of closely related ASVs (that would form the same OTU), we observed that these generally displayed similar temporal patterns although some notable exceptions existed. An example is represented in Figure 6B in which the seasonal periodicities of 5 closely related ASVs – all corresponding to OTU1- are plotted together. In this figure, a slight succession of the summer maxima can be observed (ASV2 peaking before ASVs 33 and 1, with ASV57 afterwards), being all these only 1 nucleotide different among them. Yet, ASV128 (presenting a distance of 4 nucleotides to OTU1) displays a different distribution peaking during winter. The existence of divergent distributions of ASVs composing the same OTU demonstrates the need to break apart the clusters of related sequences, since these can hide distinct ecological patterns. Furthermore, while the previous AAP temporal studies provided insights of the inter-annual community structure, this is the first study that identifies the long-term tendencies of individual ecotypes.

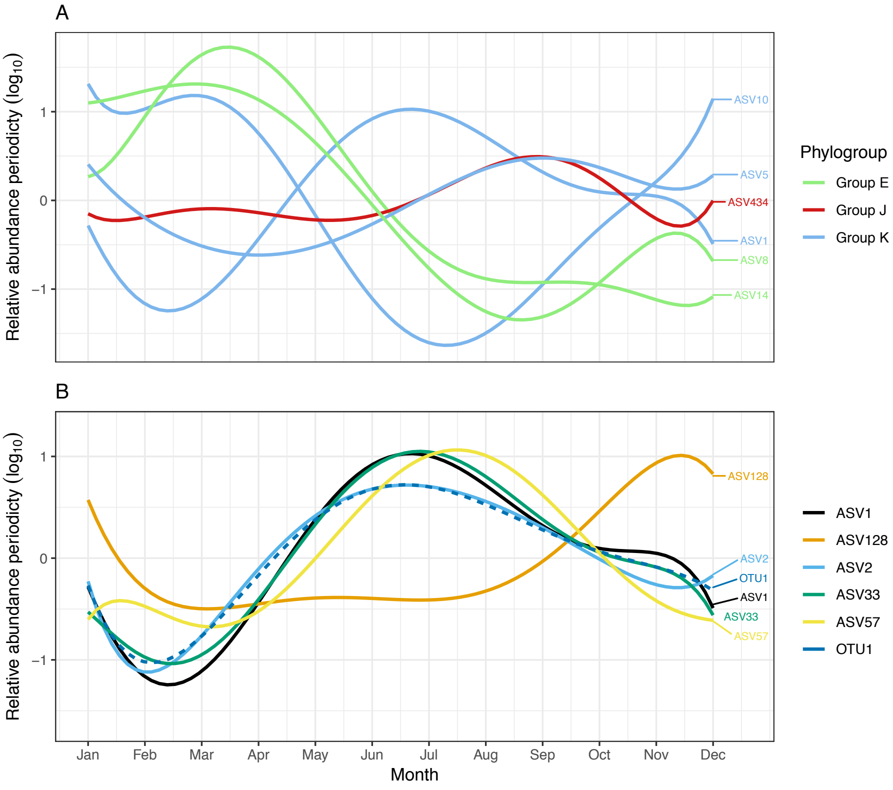


Figure 6. Seasonal component of the relative abundance distribution ($\log_{10}$ +1 transformed) for some remarked ASVs fitted with a polynomial function. (A) Various ASVs with distant nucleotide similarity, colored by phylogroup assignation. (B) Various ASVs belonging to OTU1 (dashed line corresponds to OTU1). The patterns were defined based on the relative abundance dynamics of 10 years by time series analysis.

At the other end, when we explored the seasonality at the phylogroup level, we found that phylogroup K as a whole did not present a statistically significant seasonal pattern (*p*>0.01) (Supplementary Figures 16-17). The disparity of distributions of the various sequences within may be the reason of the loss of a significant signal when computing seasonality at the group level. Contrarily, the autocorrelograms showed phylogroup E presenting a high value (max 0.34 over a year), followed by phylogroup J (Supplementary Figure 16B). These results could indicate a higher degree of ecotype differentiation in gammaproteobacterial phylogroup K as compared to alphaproteobacterial phylogroups E and G. A possible explanation is that phylogroup K is phylogenetically broader (based on the 16S rRNA gene sequences) as compared to phylogroups E and G, resulting in more variable tendencies within it. Further analyses including the genomic context with the assignation of sequence variants to Metagenome Assembled Genomes (MAGs) or genome sequencing of new isolates of phylogroup K, would help splitting this phylogroup into smaller phylogenetic clusters, perhaps showing ecological coherence. Lehours *et al.* (2018) recently tested the ecological consistency of the AAP across different oceanic regions and, interestingly, identified clades with good ecological and phylogenetic coherence. Our temporal analyses add a new level of complexity by showing that, despite a certain degree of consistency exists, highly similar ASVs can present very different seasonal distributions that could translate into different ecology.

Concluding remarks.

This work shows that the AAPs present a peak of diversity during winter, contrary to their abundance, and that gammaproteobacterial AAPs are the prevalent members of the community in the Mediterranean Sea year-round. Our results also evidence that the AAP assemblages show seasonal patterns repeatable over long periods of time. This study also demonstrates that PCR-free metagenomics and amplicon-based approaches perform in a comparable fashion in recovering major AAP groups and that the seasonal patterns observed through amplicon sequencing are robust. Interestingly, distinct seasonal behaviors were observed within the same phylogroup and even within different ASVs conforming the same OTU. In contrast to the recent spatial study of Lehours *et al.* (2018), in which they reported ecological cohesiveness when comparing contrasting biomes, we found that the different AAP phylotypes do not appear as coherent when studying their seasonal behavior and seem to be rather composed of different ecotypes with distinctive temporal niche partitioning. Overall, these results show that the analysis of long time series allows exploring in-depth patterns of a highly dynamic microbial group and provides a framework for modeling their ecological role in relation to seasonality of marine carbon cycling.

2.4 Acknowledgements

We thank the many people involved in maintaining the BBMO and those taking care of sampling, particularly Clara Cardelús and Vanessa Balagué. We would also like to thank Ramon Massana and Irene Forn for providing microscopy counts. We thank the Marbits bioinformatics platform at ICM-CSIC, particularly Ramiro Logares and Célio Dias for providing the metagenomic data and Anders Kristian Krabberød (University of Oslo) for help with network analyses. Access to the MareNostrum Supercomputer was granted to Ramiro Logares under agreement BCV-2017-3-0001. qPCR analysis was done at the Institute of Evolutionary Biology (Barcelona) thanks to José Luís Maestro. This research was funded by grant REMEI (CTM2015-70340-R) from the Spanish Ministry of Economy, Industry and Competitiveness and Grant 2017SGR/1568 from CIRIT-Generalitat de Catalunya to Consolidated Research Groups.

2.5 References

Abascal, F., Zardoya, R., and Telford, M.J. (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38: W7–W13.

Ahdesmaki, M., Fokianos, K., and Strimmer, K. (2015) Package 'GeneCycle'.

Alonso-Sáez, L., Balagué, V., Sà, E.L., Sánchez, O., González, J.M., Pinhassi, J., *et al.* (2007) Seasonality in bacterial diversity in north-west Mediterranean coastal waters: Assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiol. Ecol.* 60: 98–112.

Alonso-Sáez, L., Díaz-Pérez, L., and Morán, X.A.G. (2015) The hidden seasonality of the rare biosphere in coastal marine bacterioplankton. *Environ. Microbiol.* 17: 3766–3780.

Azur, M.J., Stuart, E.A., Frangakis, C., and Leaf, P.J. (2012) Multiple Imputation by Chained Equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res*. 20: 40–49.

Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis, A. (2018) EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. Syst. Biol. syy054-syy054.

Béjà, O., Suzuki, M.T., Heidelberg, J.F., Nelson, W.C., Preston, C.M., Hamada, T., *et al.* (2002) Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* 415: 630–633.

Bibiloni-Isaksson, J., Seymour, J.R., Ingleton, T., van de Kamp, J., Bodrossy, L., and Brown, M. V. (2016) Spatial and temporal variability of aerobic anoxygenic photoheterotrophic bacteria along the east coast of Australia. *Environ. Microbiol.* 18: 4485–4500.

Biebl, H., Allgaier, M., Tindall, B.J., Koblizek, M., Lünsdorf, H., Pukall, R., and Wagner-Döbler, I. (2005) Dinoroseobacter shibae gen. nov., sp. nov., a new aerobic phototrophic bacterium isolated from dinoflagellates. *Int. J. Syst. Evol. Microbiol.* 55: 1089–1096.

Boeuf, D., Cottrell, M.T., Kirchman, D.L., Lebaron, P., and Jeanthon, C. (2013) Summer community structure of aerobic anoxygenic phototrophic bacteria in the western Arctic Ocean. *FEMS Microbiol. Ecol.* 85: 417–432.

Buttigieg, P.L., Fadeev, E., Bienhold, C., Hehemann, L., Offre, P., and Boetius, A. (2018) Marine microbes in 4D — using time series observation to assess the dynamics of the ocean microbiome and its links to ocean health. *Curr. Opin. Microbiol.* 43: 169–185.

Caliz, J. and Casamayor, E.O. (2014) Environmental controls and composition of anoxygenic photoheterotrophs in ultraoligotrophic high-altitude lakes (Central Pyrenees). *Environ. Microbiol. Rep.* 6: 145–151.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13: 581.

Cottrell, M.T. and Kirchman, D.L. (2009) Photoheterotrophic microbes in the arctic ocean in summer and winter. *Appl. Environ. Microbiol.* 75: 4958–4966.

Cuadrat, Rafael R. C., Ferrera, Isabel, Grossart Hans-Peter, Dávila, A.M.R. (2016) Picoplankton Bloom in Global South? A High Fraction of Aerobic Anoxygenic Phototrophic Bacteria in Metagenomes from a Coastal Bay (Arraial do Cabo—Brazil). *OMICS*. 20: 76-87

Durno, W.E., Hanson, N.W., Konwar, K.M., and Hallam, S.J. (2013) Expanding the boundaries of local similarity analysis. *BMC Genomics* 14: 1–14.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.

Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2015) Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J*. 9: 968–979.

Faith, D.P., Minchin, P.R., and Belbin, L. (1987) Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* 69: 57–68.

Faust, K., Lahti, L., Gonze, D., de Vos, W.M., and Raes, J. (2015) Metagenomics meets time series analysis: Unraveling microbial community dynamics. Curr. Opin. Microbiol. 25: 56–66.

Fenchel, T. (2001) Marine bugs and carbon flow. *Science* 292: 2444 LP-2445.

Ferrera, I., Borrego, C.M., Salazar, G., and Gasol, J.M. (2014) Marked seasonality of aerobic anoxygenic phototrophic bacteria in the coastal NW Mediterranean Sea as revealed by cell abundance, pigment concentration and pyrosequencing of *pufM* gene. *Environ. Microbiol.* 16: 2953–2965.

Ferrera, I., Gasol, J.M., Sebastián, M., Hojerová, E., and Kobížek, M. (2011) Comparison of growth rates of aerobic anoxygenic phototrophic bacteria and other bacterioplankton groups in coastal mediterranean waters. *Appl. Environ. Microbiol.* 77: 7451–7458.

Ferrera, I., Sánchez, O., Kolářová, E., Koblížek, M., and Gasol, J.M. (2017) Light enhances the growth rates of natural populations of aerobic anoxygenic phototrophic bacteria. ISME J. 11: 2391–2393.

Ferrera, I., Sarmento, H., Priscu, J., Chiuchiolo, A., Gonzalez, J.M., and Grossart, H.P. (2017) Diversity and distribution of freshwater aerobic anoxygenic phototrophic bacteria across a wide latitudinal gradient. *Front. Microbiol.* 8: 175.

Fuhrman, J.A., Cram, J.A., and Needham, D.M. (2015) Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Microbiol.* 13: 133–146.

Galand, P.E., Gutiérrez-Provecho, C., Massana, R., Gasol, J.M., and Casamayor, E.O. (2010) Inter-annual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea (Blanes Bay Microbial Observatory). *Limnol. Oceanogr.* 55: 2117–2125.

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

Galí, M., Simó, R., Vila-Costa, M., Ruiz-González, C., Gasol, J.M., and Matrai, P. (2013) Diel patterns of oceanic dimethylsulfide (DMS) cycling: Microbial and physical drivers. *Global Biogeochem. Cycles* 27: 620–636.

Gasol, J.M., Cardelús, C., Morán, X.A.G., Balagué, V., Forn, I., Marrasé, C., *et al.* (2016) Seasonal patterns in phytoplankton photosynthetic parameters and primary production at a coastal NW Mediterranean site. *Sci. Mar.* 80S1: 63–77.

Gilbert, J.A., Steele, J.A., Caporaso, J.G., Steinbrück, L., Reeder, J., Temperton, B., *et al.* (2012) Defining seasonal marine microbial community dynamics. *ISME J.* 6: 298–308.

Giner, C.R., Balagué, V., Krabberød, A.K., Ferrera, I., Reñé, A., Garcés, E., *et al.* (2018) Quantifying long-term recurrence in planktonic microbial eukaryotes. *Mol. Ecol.* 28: 923-935

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017) Microbiome datasets are compositional: And this is not optional. *Front. Microbiol*. 8: 1–6.

Graham, E.D., Heidelberg, J.F., and Tully, B.J. (2018) Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J.* 12: 1861

Hatosy, S.M., Martiny, J.B.H., Sachdeva, R., STeele, J., and Fuhrman, J.A. (2013) Beta diversity of marine bacteria depends on temporal scale. *Ecology* 94: 1898–1904.

Jiao, N., Zhang, Y., Zeng, Y., Hong, N., Liu, R., Chen, F., and Wang, P. (2007) Distinct distribution pattern of abundance and diversity of aerobic anoxygenic phototrophic bacteria in the global ocean. *Environ. Microbiol.* 9: 3091–3099.

Kane, M.D. (2004) Microbial Observatories: Exploring and Discovering Microbial Diversity in the 21st Century. *Microb. Ecol.* 48: 447–448.

Kim, D.Y., Countway, P.D., Jones, A.C., Schnetzer, A., Yamashita, W., Tung, C., and Caron, D. a (2014) Monthly to interannual variability of microbial eukaryote assemblages at four depths in the eastern North Pacific. *ISME J*. 8: 515–30.

Koblížek, M. (2015) Ecology of aerobic anoxygenic phototrophs in aquatic environments. *FEMS Microbiol. Rev.* 39: 854–870.

Lami, R., Cottrell, M.T., Ras, J., Ulloa, O., Obernosterer, I., Claustre, H., *et al.* (2007) High abundances of aerobic anoxygenic photosynthetic bacteria in the South Pacific Ocean. *Appl. Environ. Microbiol.* 73: 4198–4205.

Lami, R., Ras, J., Lebaron, P., and Koblí, M. (2009) Distribution of free-living and particle-attached aerobic anoxygenic phototrophic bacteria in marine environments. *Aquat. Microb. Ecol.* 55: 31–38.

Legendre, P. and Legendre, L. (1988) Numerical Ecology, Volume 24. *Developments Environ. Model.* 24: 870.

Lehours, A.-C., Enault, F., Boeuf, D., and Jeanthon, C. (2018) Biogeographic patterns of aerobic anoxygenic phototrophic bacteria reveal an ecological consistency of phylogenetic clades in different oceanic biomes. *Sci. Rep.* 8: 4105.

Letunic, I. and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44: W242–W245.

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

Markowitz, V.M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., *et al.* (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* 34: D344–D348.

Martin-Platero, A.M., Cleary, B., Kauffman, K., Preheim, S.P., McGillicuddy, D.J., Alm, E.J., and Polz, M.F. (2018) High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nat. Commun.* 9: 266.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17: 10.

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

Mašín, M., Zdun, A., Stoń-Egiert, J., Nausch, M., Labrenz, M., Moulisová, V., and Koblížek, M. (2006) Seasonal changes and diversity of aerobic anoxygenic phototrophs in the Baltic Sea. *Aquat. Microb. Ecol.* 45: 247–254.

Massana, R., Murray, A.E., Preston, C.M., Delong, E.F., Massana, R., Murray, A.E., and Preston, C.M. (1997) Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel . Vertical Distribution and Phylogenetic Characterization of Marine Planktonic Archaea in the Santa Barbara Channel. *Appl Environ Microbiol* 63: 50–56.

McMurdie, P.J. and Holmes, S. (2013) Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* 8: e61217.

Nunes, S., Latasa, M., Gasol, J.M., and Estrada, M. (2017) Seasonal and interannual variability of phytoplankton community structure in a Mediterranean coastal site. *Mar. Ecol. Prog. Ser.* 592: 57–75.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., *et al.* (2018) vegan: Community Ecology Package.

Pedersen, T.L. (2017) ggraph: An Implementation of Grammar of Graphics for Graphs and Networks.

R Core Team (2014) R: A language and environment for statistical computing.

Ritchie, A.E. and Johnson, Z.I. (2012) Abundance and genetic diversity of aerobic anoxygenic phototrophic bacteria of coastal regions of the pacific ocean. *Appl. Environ. Microbiol.* 78: 2858–2866.

Salazar, G., Cornejo-Castillo, F.M., Benítez-Barrios, V., Fraile-Nuez, E., Álvarez-Salgado, X.A., Duarte, C.M., *et al.* (2015) Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J.* 10: 596–608.

Salter, I., Galand, P.E., Fagervold, S.K., Lebaron, P., Obernosterer, I., Oliver, M.J., *et al.* (2015) Seasonal dynamics of active SAR11 ecotypes in the oligotrophic Northwest Mediterranean Sea. *ISME J.* 9: 347–360.

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

Shade, A., Caporaso, J.G., Handelsman, J., Knight, R., and Fierer, N. (2013) A meta-analysis of changes in bacterial and archaeal communities with time. *ISME J.* 754: 1493–1506.

Shade, A., Jones, S.E., Caporaso, J.G., Handelsman, J., Knight, R., Fierer, N., and Gilbert, A. (2014) Conditionally Rare Taxa Disproportionately Contribute to Temporal Changes in Microbial Diversity. *mBio* 5: 1–9.

Sieracki, M.E., Gilg, I.C., Thier, E.C., Poulton, N.J., and Goericke, R. (2006) Distribution of planktonic aerobic anoxygenic photoheterotrophic bacteria in the northwest Atlantic. *Limnol. Oceanogr.* 51: 38–46.

Smith, B. and Wilson, J.B. (1996) A Consumer's Guide to Evenness Indices. *Oikos* 76: 70–82.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., *et al.* (2015) Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348: 1261359.

Waidner, L. a and Kirchman, D.L. (2007) Aerobic anoxygenic phototrophic bacteria attached to particles in turbid waters of the Delaware and Chesapeake estuaries. *Appl. Environ. Microbiol.* 73: 3936–44.

Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., *et al.* (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5: 27.

Xia, L.C., Steele, J.A., Cram, J.A., Cardon, Z.G., Simmons, S.L., Vallino, J.J., *et al.* (2011) Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst. Biol*. 5: S15.

Yang, Q., Jiang, Z.-W., Huang, C.-H., Zhang, R.-N., Li, L.-Z., Yang, G., *et al.* (2018) Hoeflea prorocentri sp. nov., isolated from a culture of the marine dinoflagellate Prorocentrum mexicanum PM01. *Antonie Van Leeuwenhoek* 111: 1845–1853.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., *et al.* (2007) The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biol.* 5: 0432–0466.

Yutin, N., Suzuki, M.T., and Be, O. (2005) Novel Primers Reveal Wider Diversity among Marine Aerobic Anoxygenic Phototrophs. *Appl. Environ. Microbiol.* 71: 8958–8962.

Yutin, N., Suzuki, M.T., Teeling, H., Weber, M., Venter, J.C., Rusch, D.B., and Béjà, O. (2007) Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. *Environ. Microbiol.* 9: 1464–1475.

Zeng, Y.H., Chen, X.H., and Jiao, N.Z. (2007) Genetic diversity assessment of anoxygenic photosynthetic bacteria by distance-based grouping analysis of *pufM* sequences. *Lett. Appl. Microbiol.* 45: 639–645.

2.6 Supplementary figures

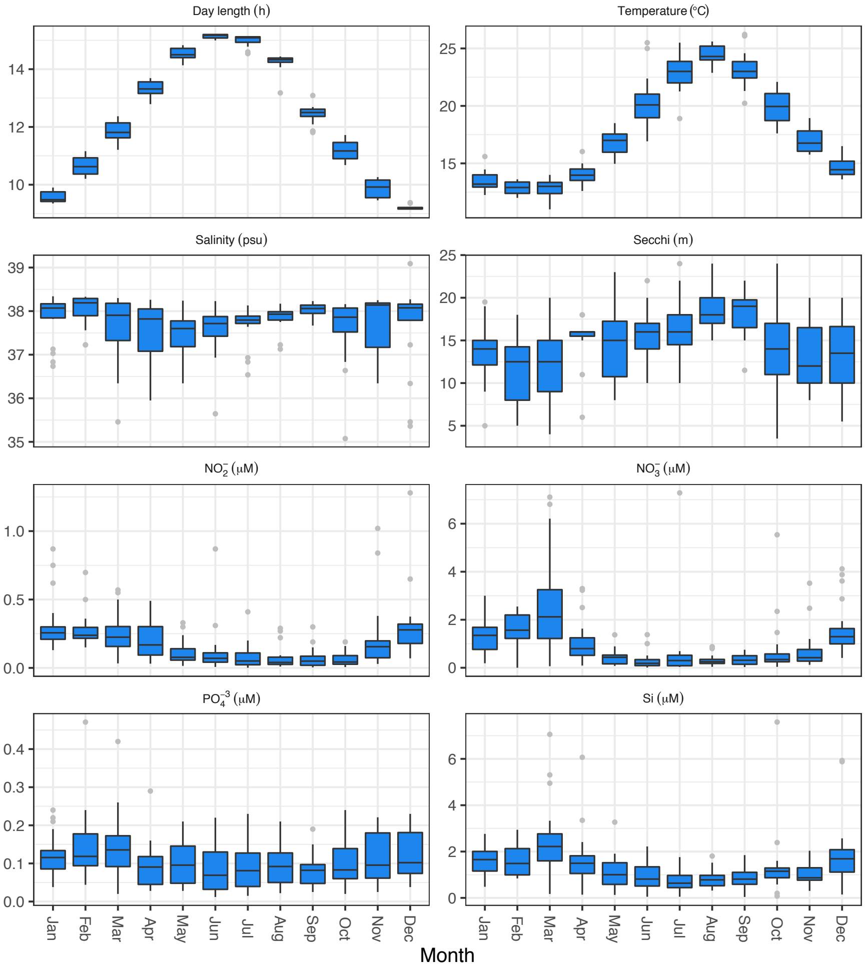


Figure S1. Boxplots showing the value distribution of the abiotic variables across the decade. The bottom and top of the boxes represent the first and third quartiles, with the central band representing the median, and whiskers represent 1.5 times the interquartile range.

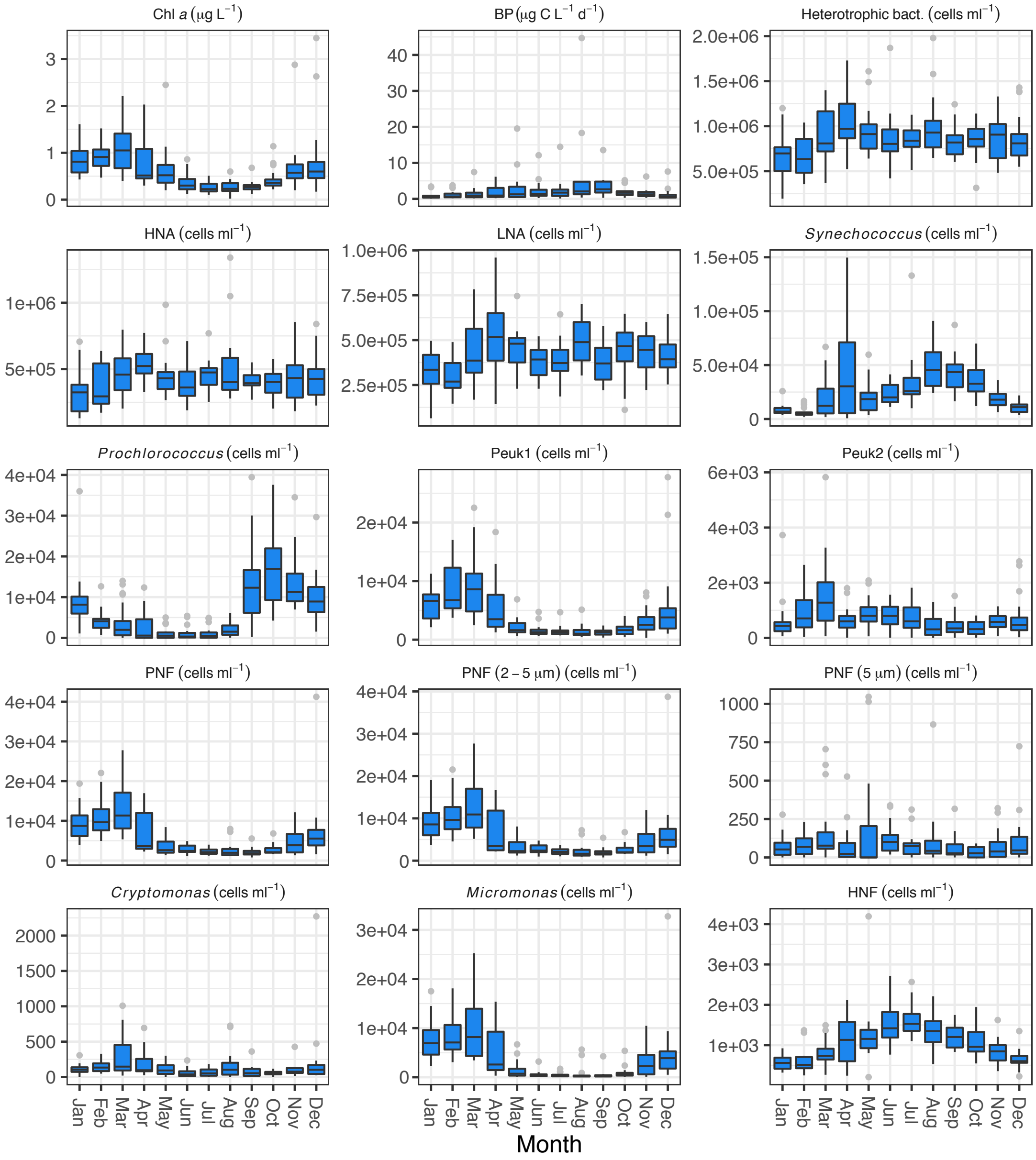


Figure S2. Boxplots showing the value distribution of the biotic variables across the decade. The bottom and top of the boxes represent the first and third quartiles, with the central band representing the median, and whiskers represent 1.5 times the interquartile range. Chl *a*: Chlorophyll *a*; BP: bacterial production; HNA: high nucleic acid content; LNA: low nucleic acid content; Peuk1: picoeukaryotes group I; Peuk2: picoeukaryotes group II; PNF phototrophic nanoflagellates; HNF: heterotrophic nanoflagellates. See Supplementary Information 1 for details.

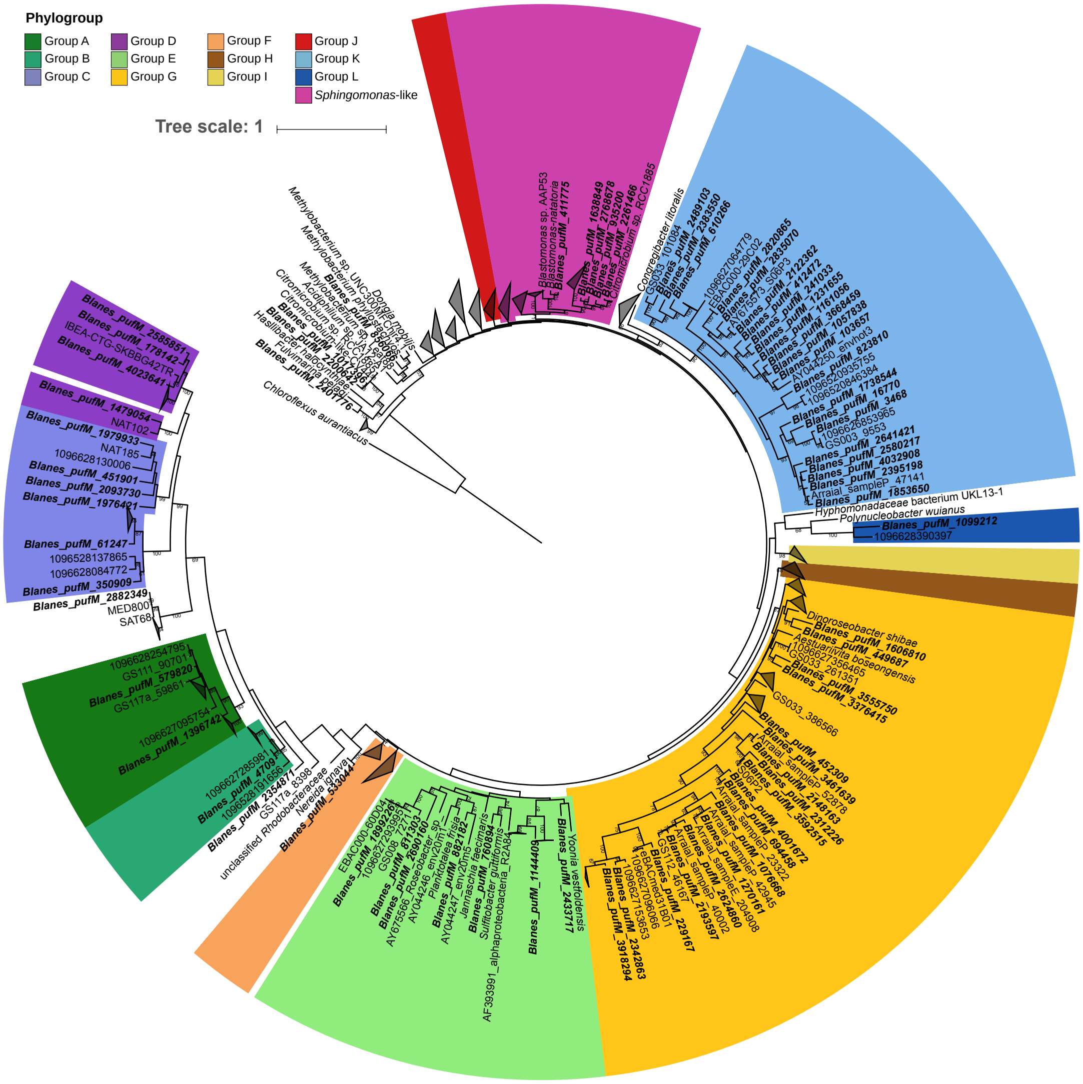


Figure S3. Maximum likelihood phylogenetic tree of the *pufM* metagenomic gene sequences. Color highlights the different phylogroups defined by Yutin *et al.* (2007), and *Sphingomonas*-like AAPs. The sequences from this study are named Blanes_*pufM*_ plus an identifier, and are highlighted in bold.

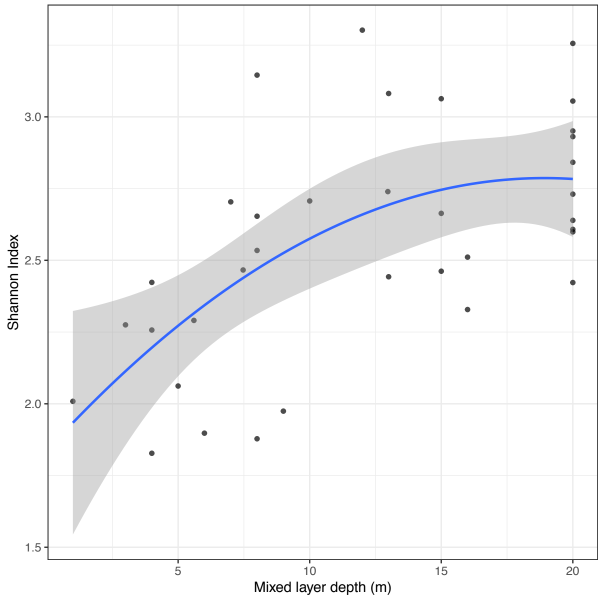


Figure S4. Polynomial regression between the Shannon index values and the mixing layer depth (MLD). The regression curve is plotted together with the 95% confidence intervals of the prediction.

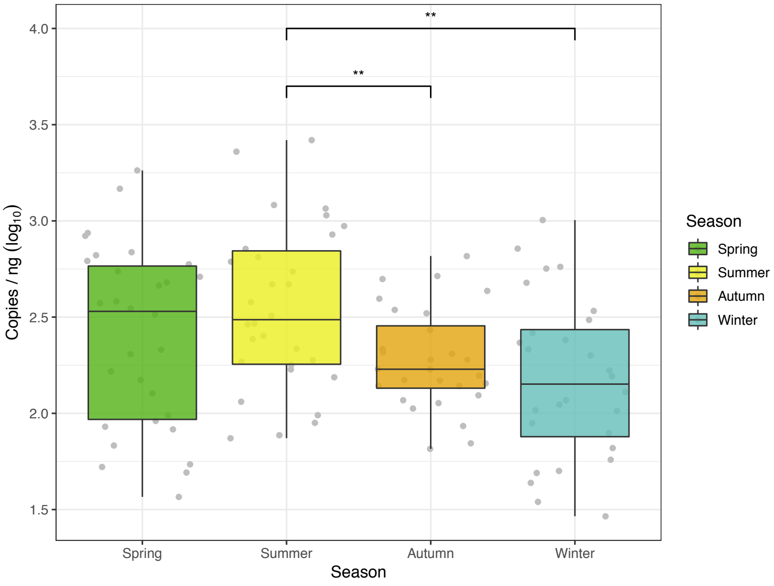


Figure S5. Seasonal distribution of *pufM* gene abundance measured by qPCR. Data has been pooled in boxplots for each season. The measurements are *log*₁₀ transformed, and the asterisks indicate significant differences between groups (*p*≤0.01). The values shown are the average of three replicates in 1 ng of genomic DNA. The bottom and top of the boxes represent the first and third quartiles, with the central band representing the median, and whiskers represent 1.5 times the interquartile range.

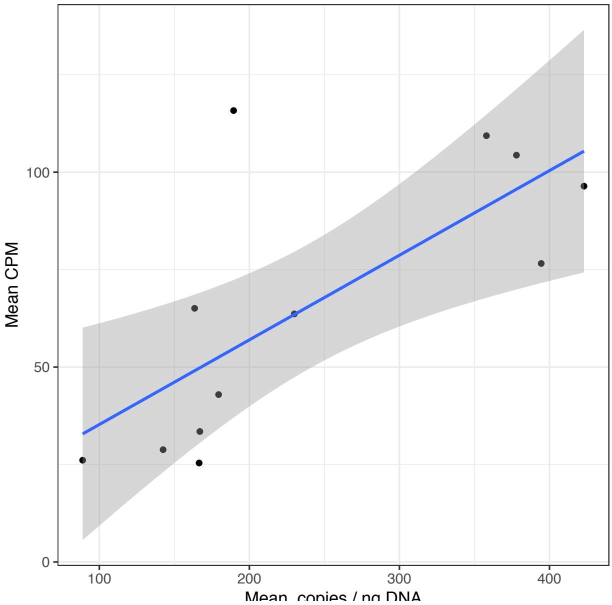


Figure S6. Correlation between the mean abundances for each month through qPCR (*pufM* copies / ng gDNA, x axis) and the mean abundances for each month through metagenomics (copies per million reads, CPM, y axis). A linear regression with the 95% confidence bounds of the prediction is plotted.

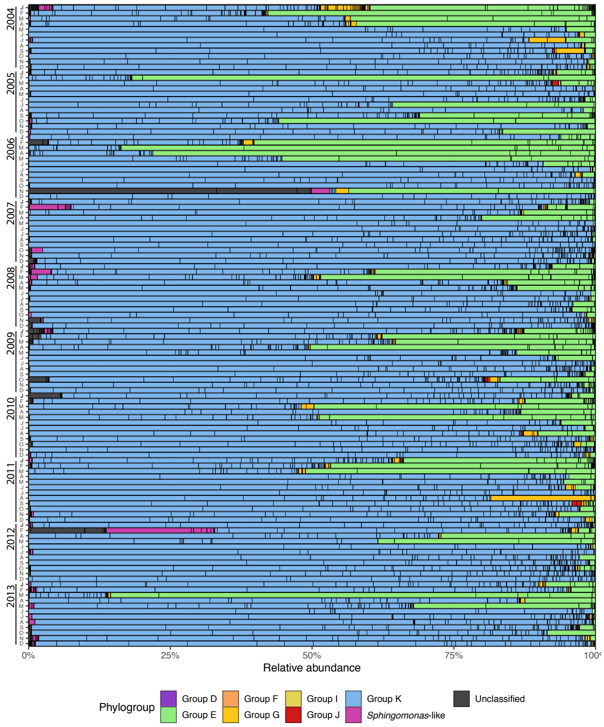


Figure S7. Monthly relative abundance of the taxonomic composition of AAPs assemblage at the phylogroup level. Each sample is sorted by date from 2004 to 2013, and each level (line section in black) corresponds to a particular ASV.

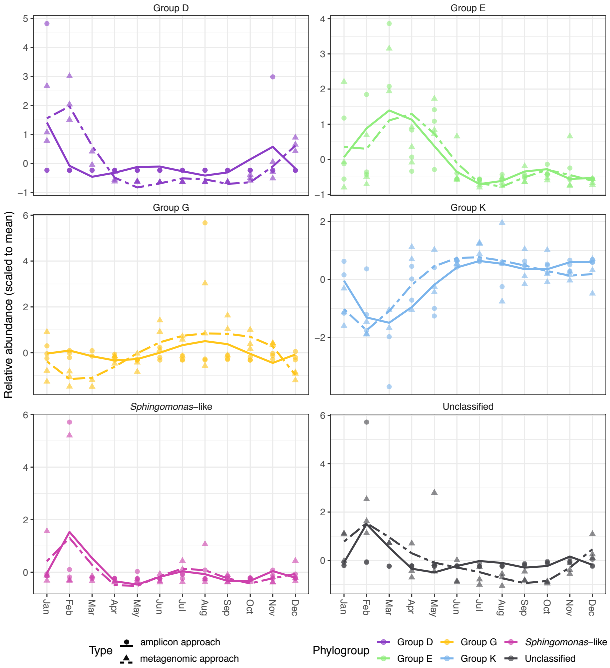


Figure S8. Trends of the relative abundance (scaled to mean) for each phylogroup comparing the amplicon (circles, solid line) and metagenomic (triangles, dashed line) datasets.

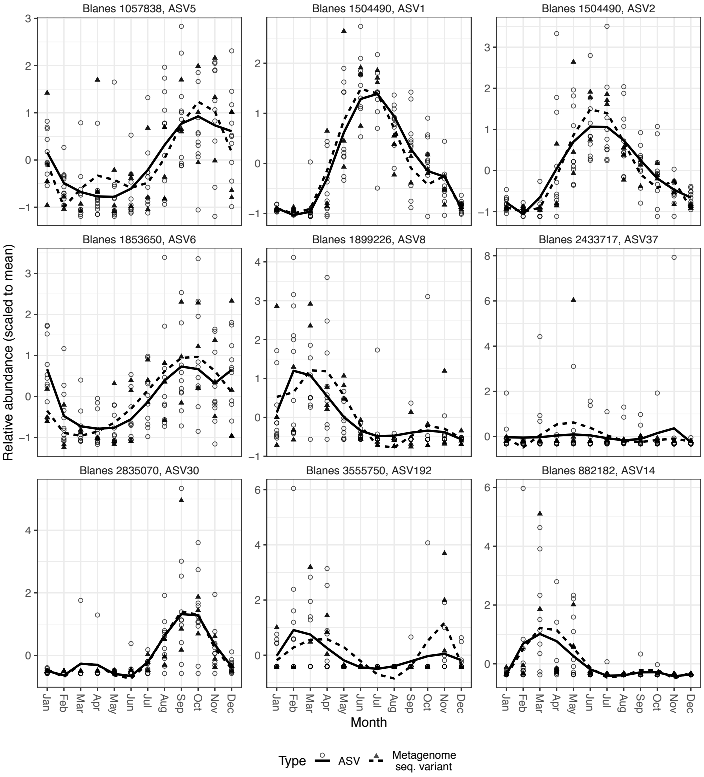


Figure S9. Comparison of relative abundance trends for some seasonal ASVs (circles, solid line) and their corresponding metagenomic variant (triangles, dashed line). Values are scaled to mean.

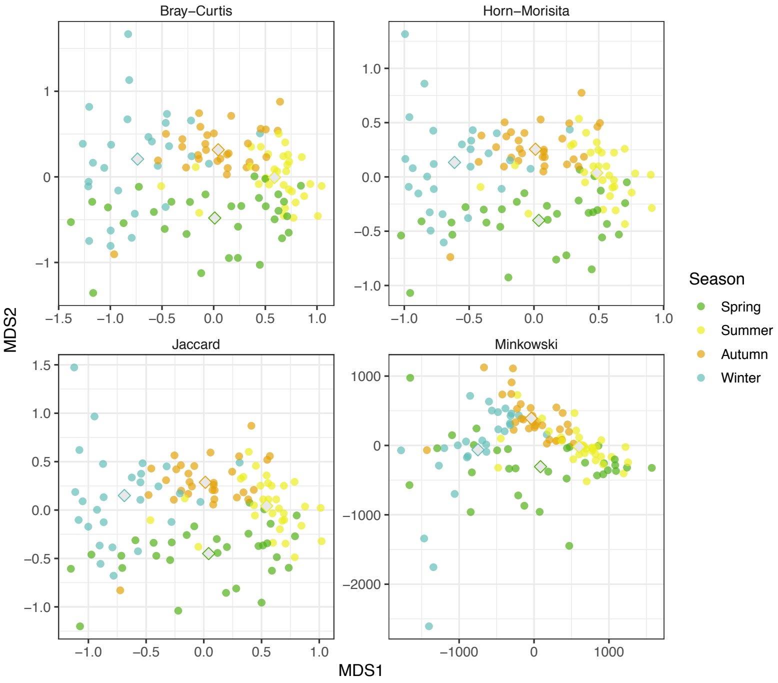


Figure S10. Non-metric multidimensional scaling of the samples with implementation of various ecological relevant distances: Bray-Curtis, Horn-Morisita, Jaccard, and Minkowski. Diamonds display the barycenter for each category.

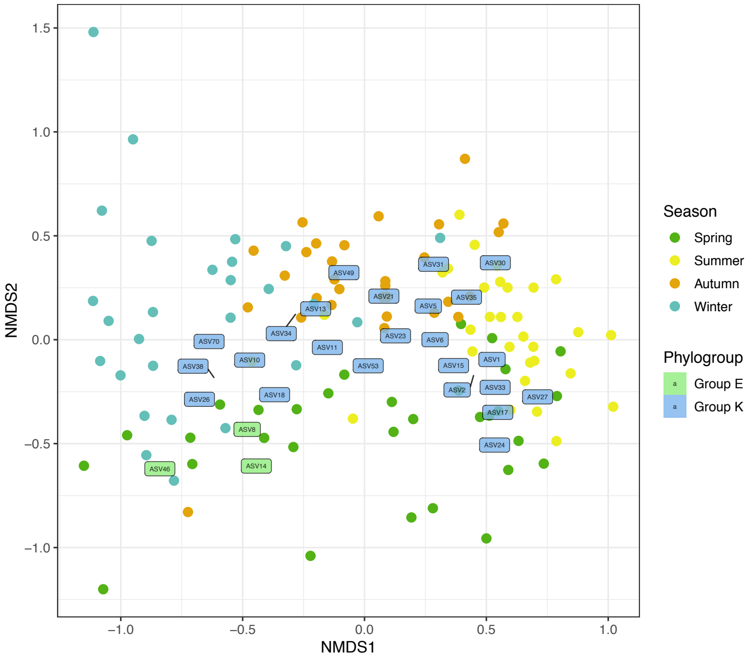


Figure S11. Non-metric multidimensional scaling biplot showing both sample (colored by season) and taxa distribution (colored by phylogroup). All ASVs were considered, but only the most abundant are shown (27 ASVs) to ease visualization.

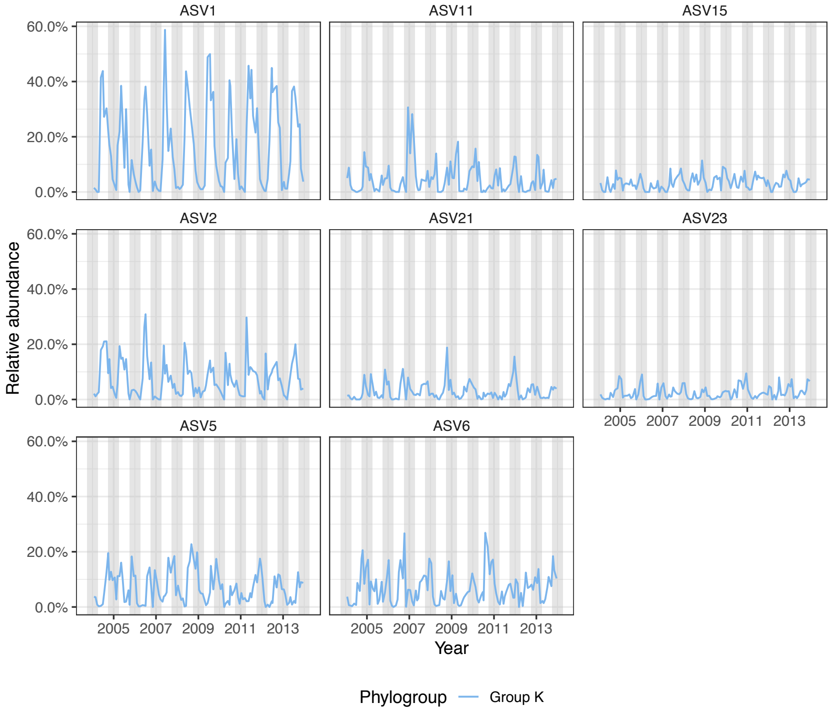


Figure S12. Relative abundance of seasonal ASVs that were always abundant throughout the decade. The background differentiates autumn and winter (grey) from spring and summer (white).

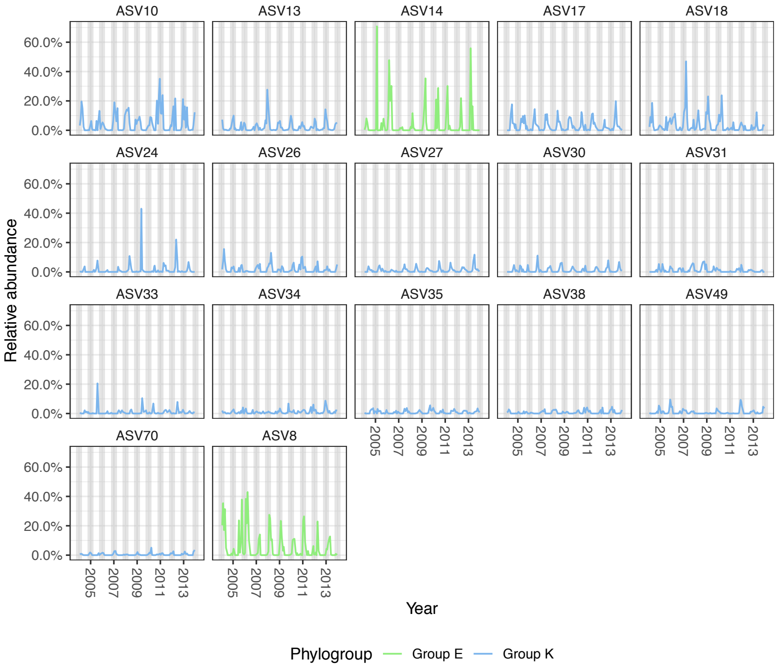


Figure S13. Relative abundance of seasonal ASVs being abundant in a specific season over the decade. The background differentiates autumn and winter (grey) from spring and summer (white).

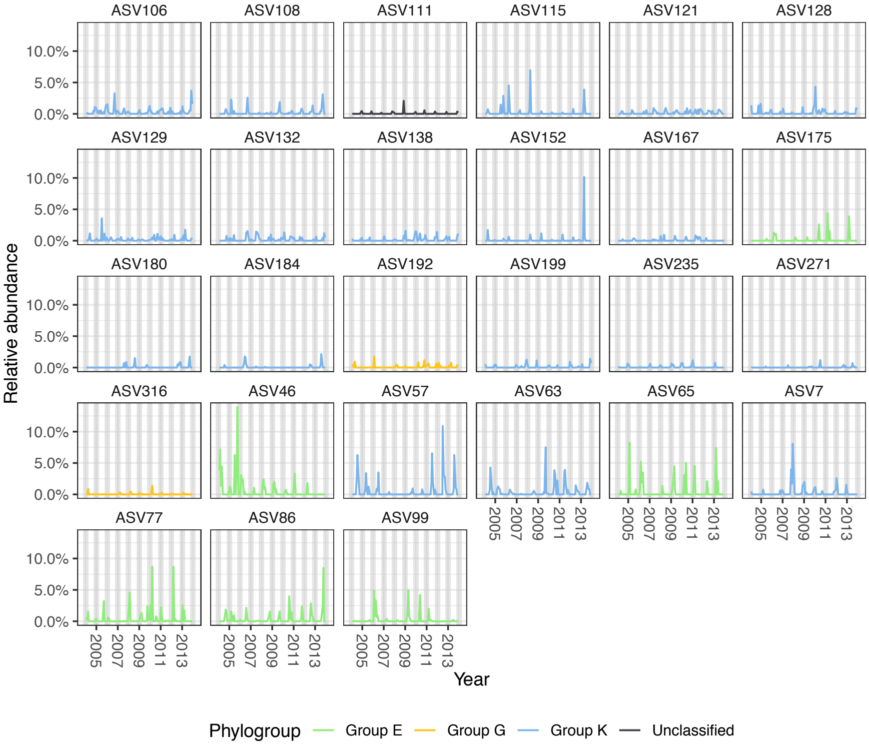


Figure S14. Relative abundance of seasonal ASVs with peak behavior (i.e. ASVs becoming >1% on specific dates). The background differentiates autumn and winter (grey) from spring and summer (white).

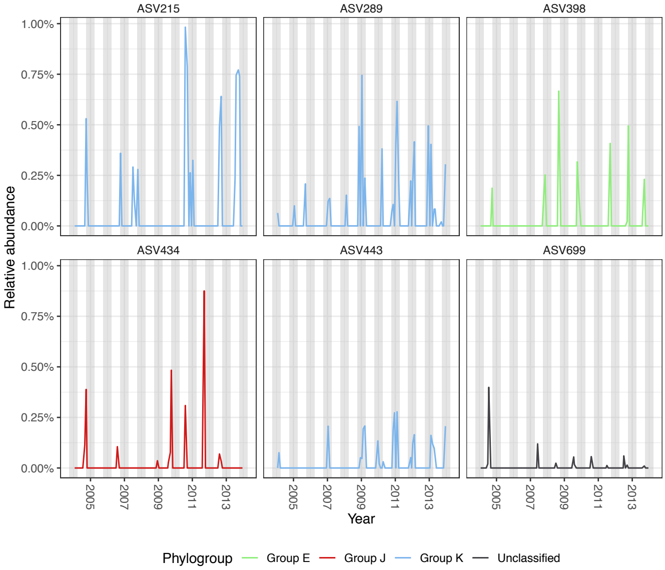


Figure S15. Relative abundance of seasonal ASVs that were permanently rare (<1% of relative abundance) ASVs over the decade. The background differentiates autumn and winter (grey) from spring and summer (white).

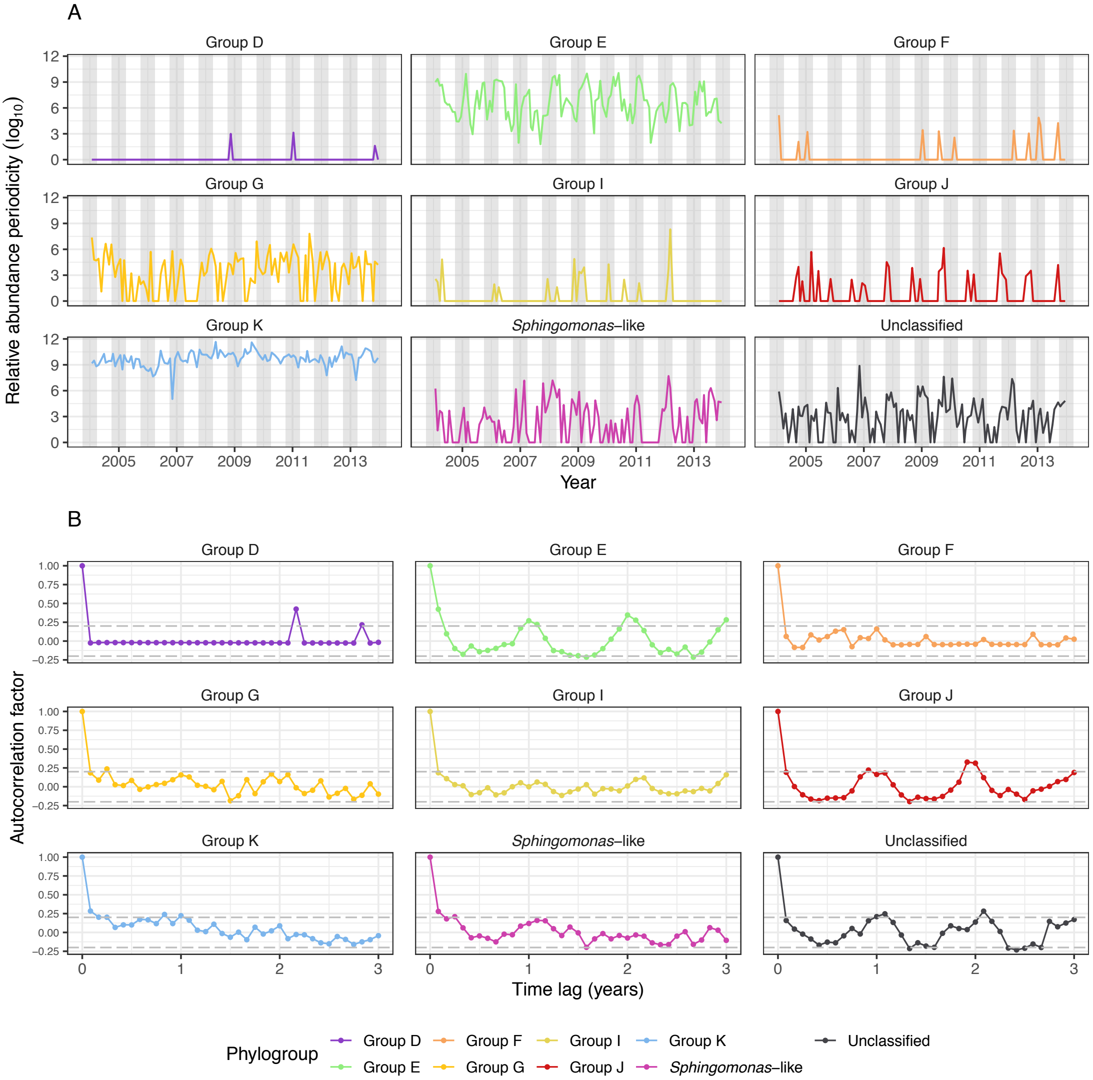


Figure S16. (A) Logarithmic transformed relative abundances of each phylogroup during 10 years (the background differentiates autumn and winter (grey) from spring and summer (white)). (B) Autocorrelogram (correlations of the whole time series at different time lags) of each phylogroup (grey lines at 0.2 and -0.2 threshold for differentiating random correlations from noise).

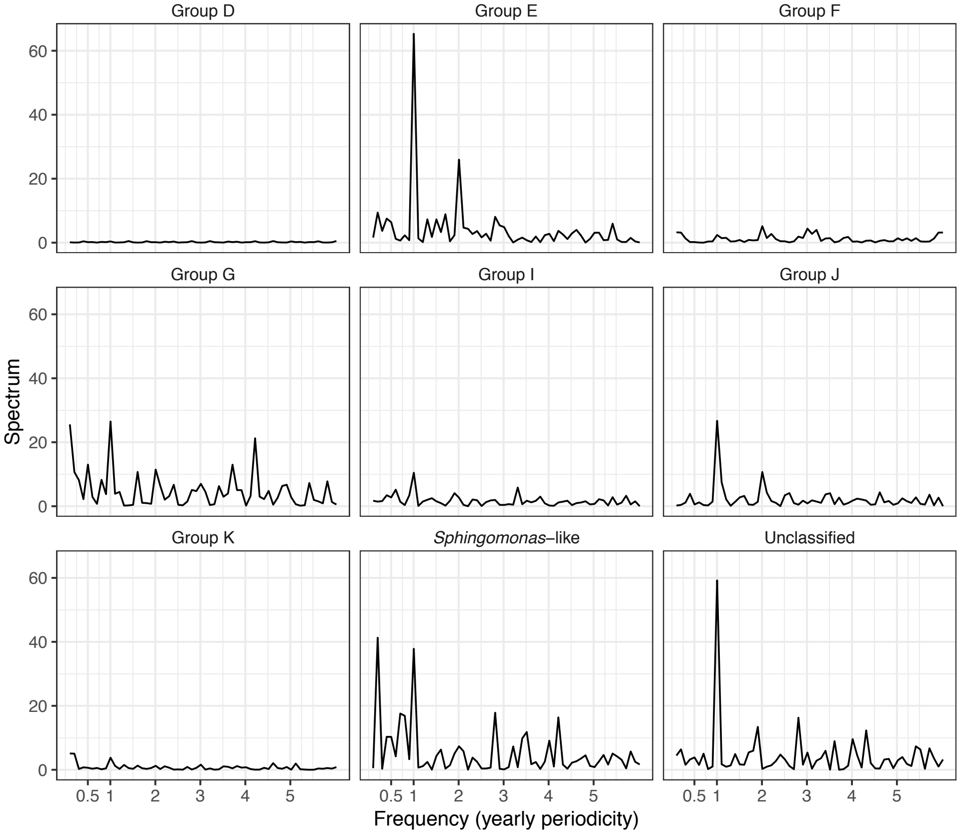
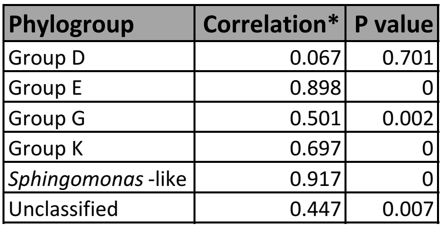


Figure S16. Periodogram distribution of the phylogroup relative abundances, with the spectral density against each frequency. The highest value corresponds to 0.083, which equals to 1/12 (annual frequency).

2.7 Supplementary tables

link Supplementary Table 1

Supplementary Table 1: Track analysis through DADA2 processing for each sample. The values (in number of reads) are the following: input (raw data), denoised (after the filtering of maximum expected error), merged (F and R reads merging), tabled (abundance table created through the DADA2 model) and nonchim (chimera removal). Additionally, the count of ASVs (count_asv), singletons or doubletons (ASVs represented by one or two reads, respectively) and OTUs (count_otu) for each sample is displayed.



Phylogroup	Correlation*	P value
Group D	0.067	0.701
Group E	0.898	0
Group G	0.501	0.002
Group K	0.697	0
Sphingomonas -like	0.917	0
Unclassified	0.447	0.007

Supplementary Table 2: Summary of the correlations between amplicon and metagenomic relative abundance and *p* value for each phylogroup recovered through both approaches.
* Pearson correlation values between the relative abundance of the metagenomic approach and the amplicon approach.

link Supplementary Table 3

Supplementary Table 3: Summary information of the sequence variants. The following columns are listed: ASV, nucleotide sequence, phylogroup, OTU correspondence, nucleotide differences, occurrence (number of samples present), total relative abundance (%), seasonality, month of maximum relative abundance, median (% for the specific maximum of relative abundance), and standard error (SE).

2.8 Supplementary information

Supplementary Information 1.
Environmental parameters measured alongside sample collection.

Several environmental parameters were measured alongside sample collection: temperature and salinity (measured with a CTD probe model SAIV A/S SD204); Secchi depth; the concentration of inorganic nutrients (determined spectrophotometrically using an Alliance Evolution II autoanalyzer according to standard procedures (Grasshoff *et al.*, 1983)); chlorophyll *a* (Chl *a*) concentration (measured in acetone extracts by fluorometry); the abundances of heterotrophic prokaryotes, phototrophic prokaryotes (*Prochlorococcus* and *Synechococcus*) and phototrophic picoeukaryotes (measured by flow cytometry as described in Gasol and Morán, 2015); the abundance of *Cryptomonas*, *Micromonas*, phototrophic and heterotrophic nanoflagellates (PNF and HNF) (which were enumerated by epifluorescence microscopy from 4,6-diamidino-2-phenylindole (DAPI) stained samples); and bacterial heterotrophic activity (estimated from the incorporation of tritiated leucine method (Kirchman *et al.* 1985) modified by Smith and Azam, 1992). The abovementioned variables, subdividing some into groups (i.e., three different subgroups of PNF -the total, and fractions 2-5 μm and >5 μm-, two of heterotrophic prokaryotes -high nucleic acid content, HNA, and low nucleic-acid content, LNA-, and flow cytometrically determined picoeukaryotes group I and group II -Peuk1 and Peuk2; populations characterized by distinct scatter and fluorescence-) as well as day length were included in an environmental data table that contained a total of 23 biotic and abiotic variables and was used for statistical analysis. The sample code is the following: BL + year (2 digits) + month (2 digits) + day (2 digits), e.g BL110607 is the 7th of June of 2011.

More details about methods used for obtaining these variables can be found in Alonso-Sáez *et al.*, 2008 and Gasol *et al.*, 2016.

References

Alonso-Sáez L, Vázquez-Domínguez E, Cardelús C, Pinhassi J, Sala MM, Lekunberri I, *et al.* (2008). Factors Controlling the Year-Round Variability in Carbon Flux Through Bacteria in a Coastal Marine System. *Ecosystems* 11: 397–409.

Gasol JM, Morán XAG. (2015). Flow cytometric determination of microbial abundances and its use to obtain indices of community structure and relative activity. Hydrocarb Lipid Microbiol Protoc - Springer Protoc Handbooks 1–29.

Gasol JM, Cardelús C, Morán XAG, Balagué V, Forn I, Marrasé C, *et al.* (2016). Seasonal patterns in phytoplankton photosynthetic parameters and primary production at a coastal NW Mediterranean site. *Sci Mar* 80S1: 63–77.

Grasshoff K, Ehrhardt M, Kremling K. (1983). Methods of seawater analysis. 2nd ed. Verlag Chemie, Weinheim.

Kirchman D, K'nees E, Hodson R. (1985). Leucine incorporation and its potential as a measure of protein synthesis by bacteria in natural aquatic systems. *Appl Environ Microbiol* 49: 599–607.

Smith DC, Azam F. (1992). A simple, economical method for measuring bacterial protein synthesis rates in seawater using tritiated-leucine. Mar Microb Food Webs 6: 107–114.

Supplementary Information 2.
Metagenomic analyses: Summary and references

In order to test whether the used primers are adequate to evaluate the seasonality of the dominant AAP groups, we used metagenomes generated from the same time-series (35 samples from 2011 to 2013). Sequencing was performed in an Illumina HiSeq4000 sequencer (2 x 150 bp) at the Centre Nacional d'Anàlisi Genòmica (CNAG, http://www.cnag.crg.eu/) with a yield of mean 40 Gbp per sample. The predicted gene sequences and the abundances across samples (transcripts per million normalization, TPM, in the results referred as copies per million, CPM) were obtained through a custom pipeline.

Intially, the samples were trimmed with TRIMMOMMATIC v0.38, using default settings. From the clean reads, an assembly of the whole dataset (35 samples, mean 40 Gbp each) was performed with Ray v2.3 (Boisvert *et al.*, 2012) at kmer 55 in the Marenostrum supercomputer (Barcelona, Spain, https://www.bsc.es/). Using this specific program in the Marenostrum computer allowed to work in paralel in multiple nodes, a necessary condition to process such a large dataset (1.4 Tbp of sequences). Only contigs > 1000 bp were considered in the downstream analyses.

The contigs were then subjected to a gene prediction using both Prodigal v2.6.3 (Hyatt *et al.*, 2010) and Meta-GeneMark v3.38 (Zhu *et al.*, 2010), and the resulting predicitions were filtred through a minimum length of 100 pb. Both predictions were clustered at 95% identity and 90% overlap using CDHIT v4.6 (Li and Godzik, 2006).

Finally, BWA v0.4 (Li and Durbin, 2009) was used for mapping the reads against the genes, and SAMBAMBA (Tarasov *et al.*, 2015) was used for parsing the SAM and BAM files. Afterwards, the abundance values were normalized using the calculation of TPM (transcripts per million, herein in this paper CPM, copies per million) using eXpress v1.5.1 (Roberts and Pachter, 2013).

The predicted genes were searched against a custom *pufM* database through nBLAST v2.7 (Altschul *et al.*, 1990), identifying possible *pufM* variants (filtering options: >75% identity, >200 bp of alignment, 50% coverage and <0.001 *e-value*). Predicted *pufM* metagenomic sequences were deposited in Genbank under accesion numbers MK548413 to MK548496.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* 215: 403–410.

Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol* 13: R122.

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119.

Li H, Durbin R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25: 1754–1760.

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

Roberts A, Pachter L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods* 10: 71–73.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 31:2032-4.

Zhu W, Lomsadze A, Borodovsky M. (2010). Ab initio gene identification in metagenomic sequences . *Nucleic Acids Res* 38: e132–e132.

Supplementary Information 3. qPCR analyses

Copy numbers of the marker gene *pufM* were estimated by quantitative polymerase chain reaction (qPCR) using primer pair *pufM*557F (3'-CGCACCTGGACTGGAC-5') and *pufM*_WAWR (3'-AYNGCRAACCACCANGCC-CA-5') following the protocol described by Waidner and Kirchman (2008) with slight modifications. Amplifications were performed in 25 µl reactions using Maxima SYBR Green qPCR Master Mix (2X; Fermentas) and primers at a final concentration of 0.08 µM each. Genomic DNA from each sample was diluted (5 ng µl^{-1}) so that the total amount of DNA in each reaction was constant. Gene abundances in 2 µl (10 ng) were measured on a MyiQ™ Single-Color Real-Time PCR Detection System (Bio-Rad). Copy number per ng of gDNA was quantified by comparing the cycle at which fluorescence crossed a threshold to a standard curve constructed using a serial dilution generated from the amplification of the *pufM* gene from Roseobacter sp. COL2P. Assays were performed in triplicate for each sample, with standards and negative controls included in each run.

References

Waidner LA, Kirchman DL. Diversity and distribution of ecotypes of the aerobic anoxygenic phototrophy gene *pufM* in the Delaware estuary. Appl Environ Microbiol. 2008 Jul;74(13):4012-21. doi: 10.1128/AEM.02324-07.

Supplementary Information 4.
PERMANOVA statistic for the of environmental variables tested.

link Supplementary Information 4

CHAPTER III

Chapter III

Seasonality of biogeochemically relevant microbial genes in a coastal ocean microbiome

Adrià Auladell, Lidia Montiel Fontanet, Célio Dias Santos Júnior, Marta Sebastián, Ramiro Logares, Isabel Ferrera, Josep M Gasol

A manuscript in preparation for *Environmental Microbiology*

Abstract

Microbes drive the biogeochemical cycles of marine ecosystems through their vast metabolic functional diversity. While we have a fairly good understanding of the spatial distribution of these metabolic processes in various ecosystems achieved through the determination of the presence of the responsible key genes, not much is known about their seasonal dynamics. We analyzed the annual patterns of 21 biogeochemical relevant functions by analyzing the presence of key functional genes in a coastal ocean environment, unveiling the main taxonomic groups harboring the studied genes and analyzing their single variant dynamics. Gene richness of most functional genes followed that of the whole community richness, decreasing during summer and reaching maximal values during autumn and winter, with the exception of *dmdA*, *psbA*, *narB* and *nasA* genes that presented departures from that trend. The majority of genes presented a seasonal abundance trend; photoheterotrophic processes were enriched during spring, phosphorous-related genes were dominant during summer coinciding with phosphate limitation conditions, and assimilatory nitrate reductases correlated negatively with nitrate availability. Additionally, we identified the main taxa driving each function in each season and described the role of underrecognized taxa such as *Litoricolaceae* in carbon fixation (*rbcL*), urease (*ureC*) and CO oxidation (*coxL*). Finally, we show that for some groups, the seasonality of bacterial families is different than that of its gene repertoire, so that different genera within the same group present different functional specialization. Our study unveils the seasonality of key biogeochemical functions and the main taxonomic groups that present these relevant functions each season in a coastal ocean ecosystem.

3.1 Introduction

Microbes account for ~70% of the total marine biomass, playing key roles in ocean biogeochemical processes (Falkowski, 2012; Bar-On *et al.*, 2018). Bacteria and archaea represent a large fraction of this biomass and hold a tremendous metabolic variability (Falkowski *et al.*, 2008). The introduction of molecular biology techniques in the late 80's allowed for the first time to distinguish the major taxonomic groups developing in the seas (Giovannoni *et al.*, 1990), but with the rapid expansion of omics technologies, we are transitioning from 'who is there' to 'what are they doing', unveiling the repertoire of functional genes and their impact in environmental processes (Gasol and Kirchman, 2018). Some relevant findings obtained from such technologies have been, among others, the discovery of novel metabolisms such as photoheterotrophy (Béjà *et al.*, 2000), the role of urea in nitrification by marine archaea (Alonso-Sáez *et al.*, 2012), or the importance of heterotrophs in nitrogen fixation in the ocean surface (Delmont *et al.*, 2018). In turn, the discovery of new metabolisms increases our understanding on how marine biogeochemical cycles operate (Grossart *et al.*, 2020). An illustrative case is that of the proteorhodopsin (PR) gene; from its initial discovery in 2000, it has been proven to be present in ~80% of the microbial community members in the sunlit ocean (Yooseph *et al.*, 2007; DeLong and Béjà, 2010). Experimental and field studies have shown that this protein supports the survival and growth of various taxonomic groups through light utilization (Gómez-Consarnau *et al.*, 2007, 2010; Steindler *et al.*, 2011), and it has been suggested that is one of the major energy-transducing mechanisms harvesting solar energy in the surface ocean (Gómez-Consarnau *et al.*, 2019).

The study of the abundance, taxonomic diversity and geographic distribution of key marker genes has allowed to investigate various relevant metabolisms in the main biogeochemical cycles (reviewed in Ferrera *et al.*, 2015). Some examples are the study of photoheterotrophy by PR-containing bacteria and aerobic anoxygenic phototrophic (AAP) bacteria (by means of PR and *pufM* genes, Koblížek, 2015; Pinhassi *et al.*, 2016), carbon monoxide (CO) oxidation and uptake (cox gene, Moran and Miller, 2007; Cordero *et al.*, 2019), nitrate assimilation (*narB* and *nasA* genes, Martiny *et al.*, 2009) and phosphorus utilization (*phoX* and *ppx* genes, Dyhrman *et al.*, 2007). Nowadays, both amplicon (rRNA or functional marker genes) and metagenomic approaches are being used to unveil their environmental distribution. As an example, AAPs have been found in multiple marine biomes through amplicon approaches, and it has been found that globally the most prevalent groups are *Rhodobacteraceae* (Alphaproteobacteria) and *Haliaceae* (Gammaproteobacteria, Lehours *et al.*, 2018; Gazulla *et al.*, 2021). However, the metagenomic approaches have allowed to find previously missed taxa for this functional group, such as *Candidatus* Luxescamonaceae (Alphaproteobacteria), which presents potential for carbon fixation (Graham *et al.*, 2018). Other study cases are the genes to avoid phosphorous limitation. Phosphorous deficiency in the ocean exerts strong selective pressure on organisms and several taxonomic groups have developed strategies to prevent this depletion (Martiny *et al.*, 2006). Examples range from lipid remodeling expressing a phospholipase (plcP, Carini *et al.*, 2015; Sebastián *et al.*, 2016) to the expression of alkaline phosphatases to exploit alternative phosphorous sources (*phoX* and *phoD*, Sebastián and Ammerman, 2009). Biogeographically, the presence of phosphorous genes is linked to the specific nutrient stress levels of the ocean basin for some taxonomic groups such as *Prochloroccocus* and Pelagibacterales (Haro-Moreno *et al.*, 2020; Ustick *et al.*, 2021).

Despite amplicon analysis allows the screening of hundreds of samples at a reduced cost, many proteins are highly variable, difficulting the stablishing of a region good enough to amplify and delineate the whole taxonomic diversity. On the contrary, metagenomic approaches avoid this limitation, and nowadays their costs have decreased significantly. In fact, the global expeditions from the last decade have revealed an unprecedented amount of functional gene data using metagenomics, providing valuable information about their diversity and biogeography (Yooseph *et al.*, 2007; Sunagawa *et al.*, 2015; Salazar *et al.*, 2019; Acinas *et al.*, 2021; Ustick *et al.*, 2021). Still, knowledge on how the seasonal environmental changes influence biogeochemical functions for key enzymes is growing at a slow pace. The seasonal trends of specific groups such as photosynthetic bacteria (Paerl *et al.*, 2012), ammonia oxidizing bacteria (Galand *et al.*, 2010), and photoheterotrophic groups (Ferrera *et al.*, 2014; Nguyen *et al.*, 2015; Auladell *et al.*, 2019) have been described through amplicon analysis. Besides, some temporal metagenomic analyses have focused on studying specific taxonomic groups through the generation of metagenome assembled genomes (MAGs, Kashtan *et al.*, 2014; Pereira *et al.*, 2021). Despite the potential of MAGs in the analysis of functional groups, they can miss the contribution of rare groups not presenting a good genome recovery. Following the recent discussions regarding functional redundancy (Louca *et al.*, 2018), Galand *et al.*, (2018) found a link between temporal community turnover and the functional repertoire, hinting that functional redundancy in marine waters was rather low. Processes such as photoheterophy (*pufM* gene) or carbon fixation (bacterial and archaeal RuBisCO) change between seasons, and functional richness correlate with taxonomic richness. Another recent study looked at microbial trait variability using multiple metagenomic time series data (Beier *et al.*, 2020), and found that the metacommunity size (i.e. number of species carrying a specific function) translates into a large temporal variability of gene alleles. These temporal changes were also analyzed through metatranscriptomics during two years, albeit with a small sample number (Alonso-Sáez *et al.*, 2020). The temporal expression patterns of several genes for energy conservation such as carbon monoxide oxidation (*coxL*), reduced sulfur (*soxB*) and the oxidation of ammonia (*amoA*) differed temporally. Additionally, they detected that the expression of alkaline phosphatases was promoted during periods of phosphorous limitation, whereas other phosphorous transporters such as pit were only detected in post-bloom conditions. Although metatranscriptomic analyses allow obtaining relevant insights, transcription changes at a faster pace than community composition (Moran *et al.*, 2013), and long-term multiannual data can offer a more robust picture of the seasonal patterns of functional groups. Through metagenomics, it is possible to quantify the enrichment of specific functions following the community turnover observed in temperate locations (Fuhrman *et al.*, 2015; Lambert *et al.*, 2019; Auladell *et al.*, 2021), and furthermore, through multiyear analyses we could corroborate whether the pattern is recurrent. We present here a 7-year metagenomic analysis of monthly samples in a microbial observatory in the North-Western Mediterranean coastal sea (Blanes Bay Microbial Observatory, BBMO), focusing on 21 functional genes coding for key biogeochemical functions. Through the analysis of these genes and using information on the environmental variables defining seasonality of the area, we have: (i) determined when each function prevails in the ecosystem, (ii) obtained a detailed picture of the main taxonomic groups explaining each of the selected functions, and (iii) explored whether the distribution of these functions change among genera within the same taxonomic family on a seasonal basis.

3.2 Results and discussion

Environmental setting

The BBMO is a well studied temperate shallow coastal site subjected to strong seasonal forcing in the NW Mediterranean. Its environmental characteristics have been studied for more than 25 years, providing a rather complete understanding of the main biotic and abiotic processes determining its ecosystem's ecology (Gasol *et al.*, 2016). The environmental seasonal pattern is typical for a temperate coastal system, as seen by the recurrent environmental patterns (Supplementary Figure 1). The summer season presents low dissolved inorganic nutrients (mean of 0.6 and 0.08 µM for NO_3- and PO_4^{3-} respectively), being strongly limited by phosphorous during this season (Pinhassi *et al.*, 2006; Sebastián *et al.*, 2016). With the start of autumn, the increase of precipitation, changes in wind regime and water column mixing in nearby oceanic waters facilitate the entry of inorganic nutrients, thus enhancing the growth of several different bacterial groups with the result of higher richness (Mestre *et al.*, 2020; Auladell *et al.*, 2021). In late winter, the ecosystem reaches the highest values of phytoplankton biomass (Chlorophyll *a*, average 0.88 µg L^{-1}) due to the increased availability of nutrients (mean 1.64 and 0.11 µM for NO_3- and PO_4^3- respectively), and the increase in day length and light irradiance. During this season, phytoplankton is dominated by photosynthetic nanoflagellate (mainly haptophytes) and diatom blooms (Nunes *et al.*, 2018). Finally, during spring the continued growth of phytoplankton and heterotrophic bacteria (~9 x 105 cells ml−1) depletes most of the dissolved nutrients. Gradually, eukaryotic phytoplankton start to decrease while the number of heterotrophic nanoflagellates increases (1.3 x 103 cells ml−1, R. Massana, unpublished data). Day length is maximal by the end of the spring (15 hours), preceding the start of summer, closing the seasonal cycle. These trends vastly influence bacterial community composition which presents a strong seasonality (Alonso-Sáez *et al.*, 2007; Mestre *et al.*, 2020; Auladell *et al.*, 2021).

process in which phages incorporate iron atoms into their tails to infect microbes (a theory known as 'Ferrojan Horse Hypothesis', Bonnain *et al.*, 2016) and, therefore, there is a selective pressure to promote gene variability to avoid phage entry. Contrarily, the *amoA* and *narB* genes presented the lowest total richness values (25 and 23 variants). Regarding proteorhodopsins, the blue-absorbing type was the most diverse (727 variants), while the two green-absorbing types presented around 300 variants each (Figure 1). The richness of most of these genes was highest during autumn and winter, with minimum values in summer, therefore following the pattern of the whole community species richness (Mestre *et al.*, 2020; Auladell *et al.*, 2021), and in agreement with the conclusions that linked taxonomic and functional richness in a nearby microbial observatory (Galand *et al.*, 2018). Nonetheless, a few genes presented differences from the general trend; the richness of *dmdA* variants reached maximum values during late winter, with a median of 70 variants. Additionaly, the trend was not exactly the same for all years, since in some years the richness reached 100 variants (Supplementary Figure 2). The gene encoding the photosystem II (*psbA*) presented bimodality in its richness distribution during the 7 years; we found up to 40 variants in some years whereas others had only 20 (Supplementary Figure 2). This pattern was the result of an increase of variants from multiple cyanobacterial groups during spring and summer (Supplementary Figure 2B). The presence of multiple *psbA* variants indicates multiple coexisting *Synechococcus* ecotypes during these seasons, as shown for *Prochlorococcus* elsewhere (Kashtan *et al.*, 2014). At a global scale, the maximal abundance of cyanobacteria is associated with low nutrient concentrations and high temperatures (Flombaum *et al.*, 2013; Hunter-Cevera *et al.*, 2016). Here, we observed that there are more variants during spring and summer, coinciding with the maximal abundances. Whether this is due to different species-level populations contributing to the spring-summer blooms remains unknown. Finally, both *narB* and *nasA* –encoding nitrate reductases– followed a similar pattern presenting the highest diversity during summer, although this pattern was based in a small number of variants and therefore could be uncertain.

Most genes present seasonal changes of abundance

To test whether the studied genes presented seasonal abundance changes, we used the ratio between the read counts of a particular gene and the geometric mean of the count of 8 single copy genes (see Experimental Procedures). The values are thus abundance ratios (therein 'abundance'), indicating how much represented this gene is in comparison to the selected single copy genes in terms of number of reads (see Experimental Procedures; Figure 2). Gene abundances can present interannual variations and to test for recurrence we used the Lomb-Scargle periodogram. A total of 12 out of the 21 tested genes presented a significant seasonal pattern ($q \leq 0.05$, $PN \geq 8$, Figure 2). For the rest of genes that were not statistically seasonal, we could differentiate between genes that presented a random pattern (e.g. *fecA*) from genes presenting temporal variations that were not strong enough to be detected by the Lomb-Scargle method (e.g. *psbA*). As an example, *tauA* displayed a high monthly variation but the ratio was higher during spring and summer than during

Marker gene richness follows whole community patterns

The marker genes chosen in this study belong to various functional categories and biogeochemical cycles: for carbon cycle we selected phototrophic processes (PR, *pufM*, *pufL* and *psbA*), carbon fixation (*rbcL*), oxidation of inorganic compounds such as carbon monoxide (*coxL*), and transport of taurine (*tauA*); for the nitrogen cycle, we chose nitrogen reductases (*narB*, *nasA*), the cleavage of urea (*ureC*) and ammonia oxidation (*amoA*); for phosphorous biogeochemistry we selected *phnD*, *phnM*, *pstS*, *phoD*, *phoX*, *ppx*, *ppk1*, and *plcP* genes involved in multiple processes to avoid phosphorous starvation (Table 1); for sulfur biogeochemistry we analysed the *dmdA* involved in demethylation of dimethylsulfoniopropionate or DMSP; and finally, for the iron cycle, we selecyed *fecA*, a ferric iron transmembrane transporter. Table 1 presents an overview of each of the selected genes. A total of 93750 gene variants of this set of genes were observed (Table 1). We calculated each gene's total abundance and the seasonal changes in richness (Figure 1). The genes with the highest number of variants were those related to phosphorous metabolism (min 2730, max 14683). Other genes such as *tauA* and *fecA* also presented a high number of variants (1392 and 10926 respectively). The high variability of *fecA* has been discussed in a recent study (Beier *et al.*, 2020), and linked to the

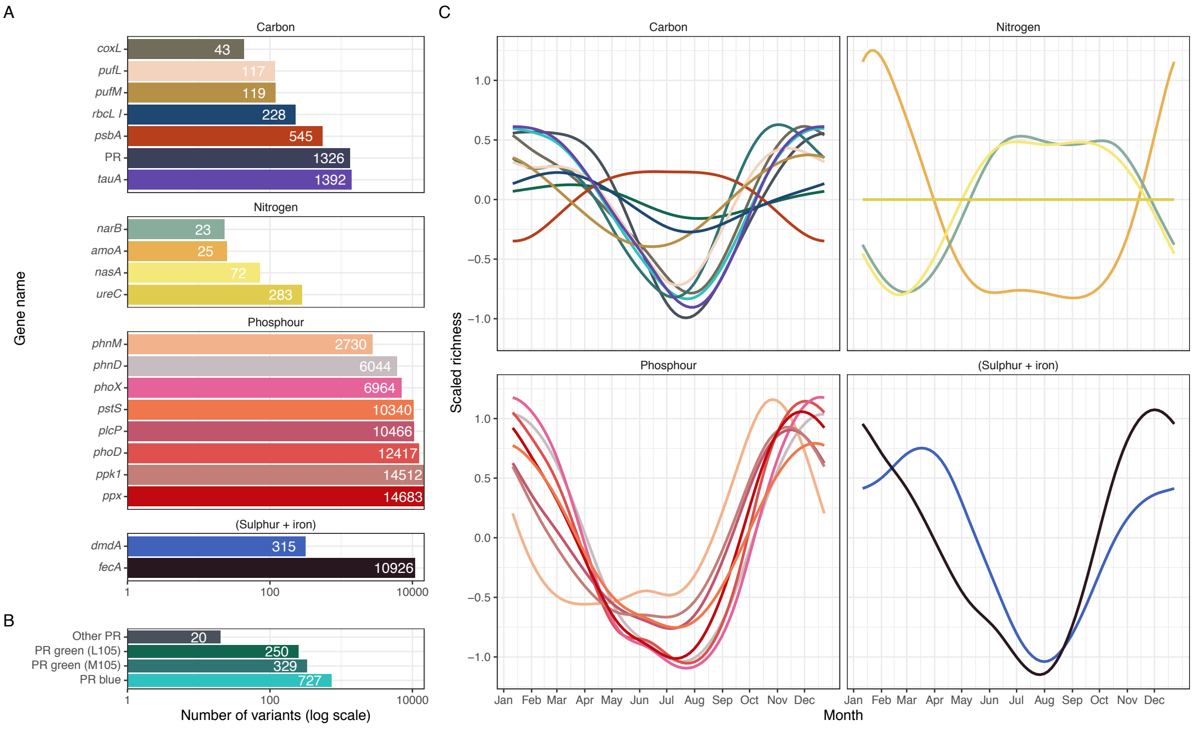


Figure 1: A) Total number of variants for each studied functional gene and B) the protheorhodopsin types. The X axis indicates the number of variants in logarithmic scale. Each panel specifies the corresponding main biogeochemical cycle to which the genes are associated with. The colors are specific for each gene. C) Seasonal trend of richness for each gene. The X axis displays the month and the Y axis the richness scaled to the mean. Each gene is colored following the color palette in panel A.

Name	Long Name	Basic Function	Annotation	Total Variants	Total Evaluated	Seasonal Variants	Percentage Seasonal (%)	Median Richness	Q25 Richness	Q75 Richness
Carbon cycle										
PR blue	proteorhodopsin blue spectral tuning (glutamine aa pos. 105)	light-driven electron transfer reactions	MicRhoDE db	727	313	143	45.7	114.5	86.8	132
PR green (M105)	proteorhodopsin green spectral tuning (methionine aa pos. 105)	light-driven electron transfer reactions	MicRhoDE db	329	156	60	38.5	55	46	62
PR green (L105)	proteorhodopsin green spectral tuning (leucine aa pos. 105)	light-driven electron transfer reactions	MicRhoDE db	250	79	32	40.5	26	22.8	30
pufM	photosyntetic reaction center (M subunit)	light-driven electron transfer reactions	K08929	119	33	8	24.2	10	7.8	12
pufL	photosyntetic reaction center (L subunit)	light-driven electron transfer reactions	K08928	117	40	11	27.5	14.5	12	17
rbcL I	Ribulose bisphosphate (large subunit) I	primary CO2 fixation	K01601	228	53	20	37.7	19	17	22
psbA	photosystem II P680 reaction center	light-driven electron transfer reactions	K02703	545	61	11	18.0	17.5	13	41
coxL	Carbon monoxide dehydrogenase (large subunit)	CO oxidation to CO2	K03520	43	19	11	57.9	7	6	8
tauA	taurine transport system substrate-binding protein	taurine transport	COG4521	1392	396	139	35.1	136	111.8	155.2

Table 1: Summary of the main properties of the selected functional genes. The genes are grouped by the biogeochemical cycles to which they are related to. Long name specifies the complete name of the gene; 'Annotation' specifies the database used for annotation; 'Total Evaluated' indicates the number of variants present in at least 8 samples; 'Seasonal variants' are those seasonal according to the Lomb-Scargle test ($q \leq 0.05$, PN ≥ 8); 'Percentage seasonal', the % of seasonal variants with respect of the evaluated variants; Median, Q25 and Q75 the number of variants (richness) for the median, first and third quantile of the distribution.

Name	Long Name	Basic Function	Annotation	Total Variants	Total Evaluated	Seasonal Variants	Percentage Seasonal (%)	Median Richness	Q25 Richness	Q75 Richness
Nitrogen cycle										
narB	nitrate reductase	reduction of nitrate to nitrite, dependent on Fd red	K00367	23	12	7	58.3	4.5	3	6
nasA	nitrate reductase	reduction of nitrate to nitrite, dependent on NADH	K00372	72	18	5	27.8	5	3.5	7
amoA	Ammonia monooxygenase subunit A	oxidation of ammonia to hydroxylamine	K10944	25	11	5	45.5	3	2	4.5
ureC	Urea degradation	hydrolisis of urea to ammonia and carbamate	K01428	283	69	22	31.9	19	16	28
Phosporous cycle										
phnD	Phosphonate ABC transporter	Phosphonate metabolism	COG3221	6044	1492	476	31.9	497.5	377.2	602.8
phnM	Alkylphosphonate utilization protein	Phosphonate metabolism	COG3454	2730	710	175	24.6	236	201.8	262.5
pstS	High affinity phosphate system	Phosphate uptake	COG0226	10340	2393	534	22.3	785	668.2	900.8
phoD	Alkaline phosphatase	Utilization of Phosphoesters	COG3540	12417	2973	853	28.7	945.5	746.5	1190
phoX	Alkaline phosphatase	Utilization of Phosphoesters	COG3211	6964	1656	504	30.4	524.5	387.2	706.8
ppx	Exopolyphosphatase	Polyphosphate metabolism	COG0248	14683	3450	728	21.1	1109	901.8	1263.2
ppk1	Polyphosphate kinase	Polyphosphate metabolism	COG0855	14512	2973	480	16.1	974.5	847.2	1082
plcP	Phospholipase C	Membrane phospholipid remodelling upon P starvation	COG2908	10466	2796	735	26.3	906.5	797	1026.2
Other										
dmdA	DMSP demethylase	removal of a methyl group from DMSP	K17486	315	168	74	44.0	66	52	73
fecA	Fe(3+) dicitrate transport protein	Fe(3+) dicitrate transport	COG4772	10926	3621	1360	37.6	1335	1080.8	1506.2
Total				93550	23492	6393	27.2			

winter (Figure 2). Along with the abundance ratios changes among seasons, we also determined the changes in taxonomic composition of each of the target genes to explore which groups encoded the different biogeochemical functions (Figure 3).

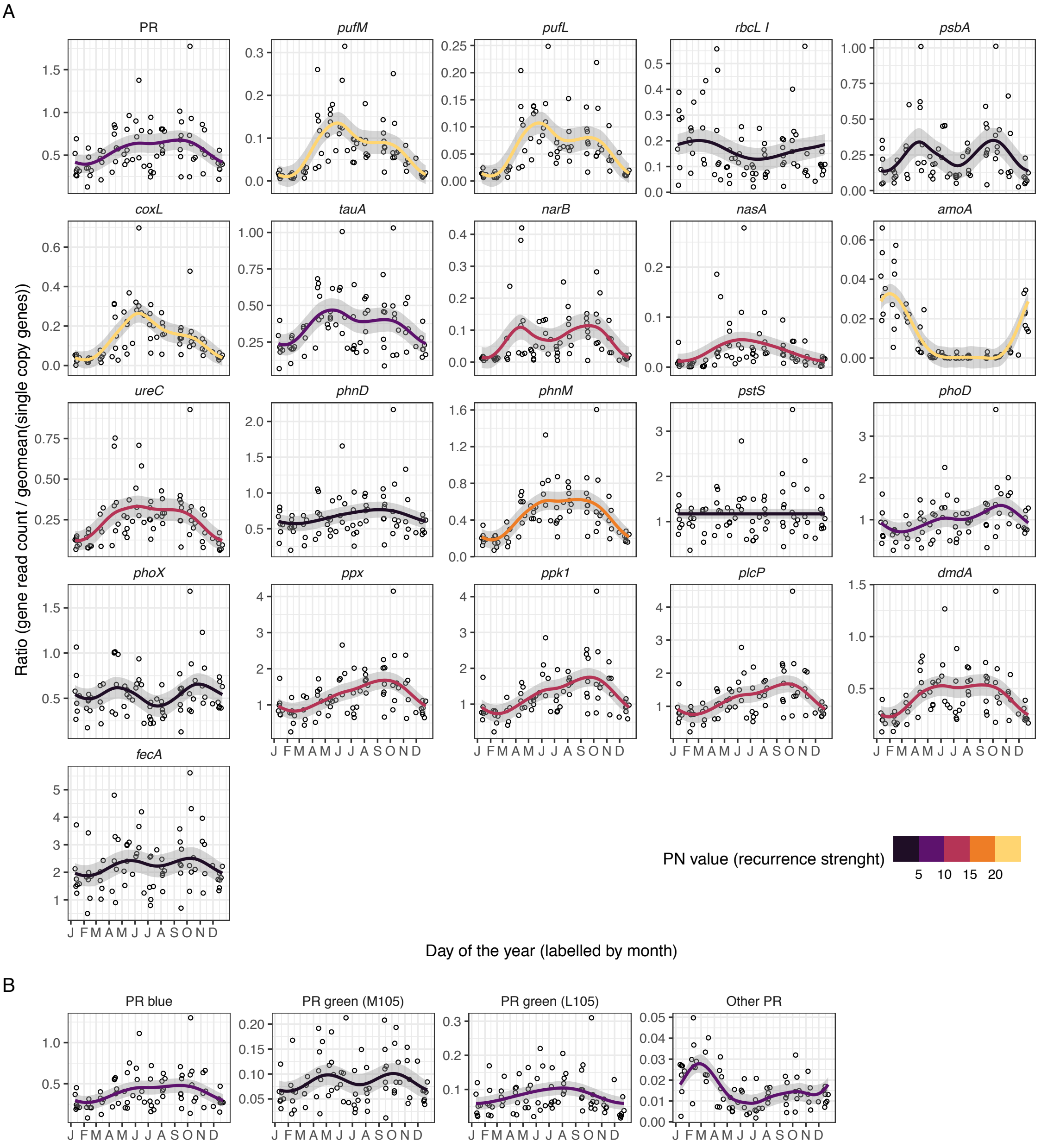


Figure 2: A) Temporal distribution of the read count ratio for each gene. The X axis indicates the day of the year (labelled by the month initials) and the Y axis the ratio between the read count of the gene divided by the geometric mean of a selection of 8 single copy genes (see Experimental Procedures for details). A generalized additive model is fitted to the data, colored based on the peak normalized value to show how strong is the seasonal signal (the PN value, based in the Lomb-Scargle test, q ≤ 0.05). B) Temporal distribution of the read count ratio of a selection of proteorhodopsin variants.

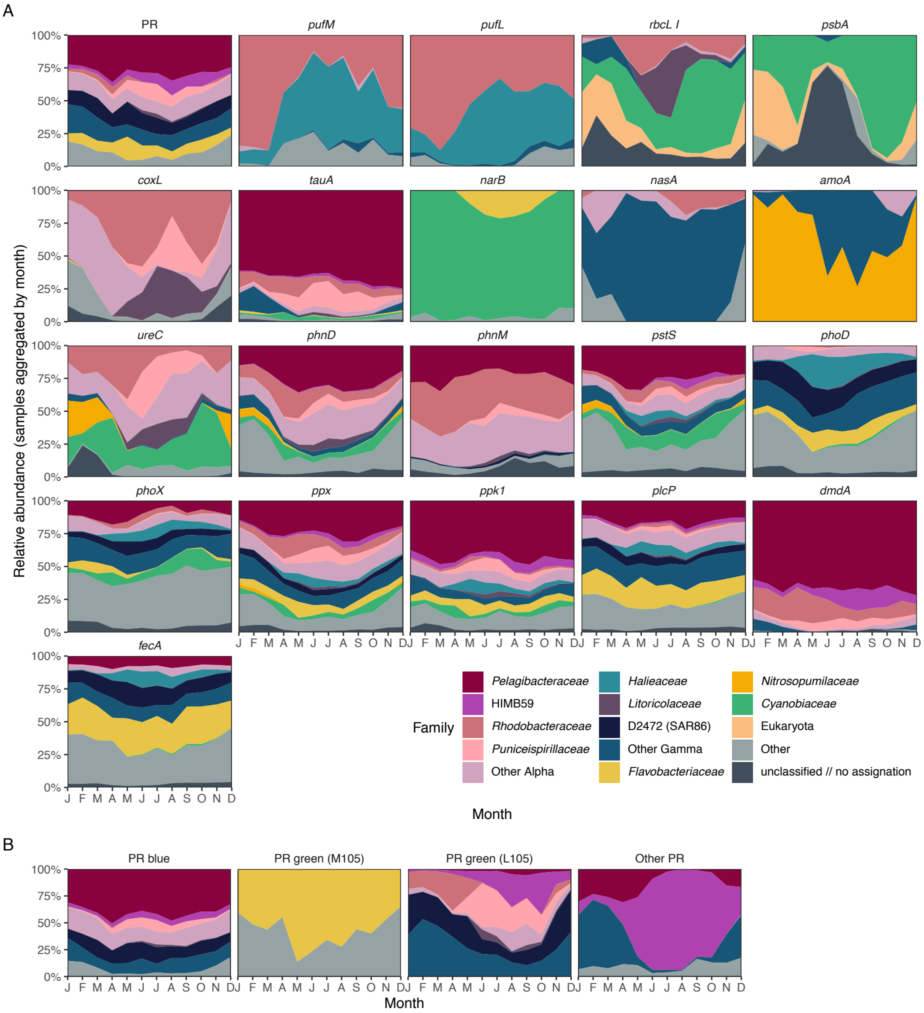


Figure 3: A) Relative distribution of each functional gene at the family level. The Y axis corresponds to the relative abundance and the X axis to the month. The colors differentiate the main family groups as shown in the legend. B) Relative taxonomic distribution of a selection of proteorhodopsin variants.

The genes related to phototrophic processes (*pufL* and *pufM* –together labelled as *pufLM*– PR and *psbA*) presented diverse patterns of abundance, with proteorhodopsin presenting the highest abundance (median ratio 0.5), followed by *psbA* (0.17) and *pufLM* (0.05) (Figure 2). The abundance order (PR > *psbA* > *pufLM*) is in agreement with a previous assessment of this distribution (Finkel *et al.*, 2013) and with the proportions observed through direct pigment estimation in the Mediterranean Sea (Gómez-Consarnau *et al.*, 2019). Both PR and *pufLM* presented a statistically significant seasonal distribution (recurrence strength = 6.65 and 22.1 respectively) whereas *psbA* was not recurrent, with lots of variance each month (recurrence strength = 1.6) but presented two peaks, one in spring and one in summer. The genes involved in the biosynthesis of the photosynthetic reaction center of AAPs (*pufLM*) peaked in spring, with the highest abundances associated to the

Rhodobacteraceae (Alphaproteobacteria) and *Luminiphilus* (Gammaproteobacteria) groups (Figure 3). The genes related to oxygenic photosynthesis and carbon fixation (*psbA* and *rbcL*) mimicked the known recurrences of the main photosynthetic populations, with eukaryotic groups dominating during winter (Nunes *et al.*, 2018; Giner *et al.*, 2019), *Synechococcus* blooming in spring and summer, and *Prochloroccus* during autumn (Gasol *et al.*, 2016; Auladell *et al.*, 2021). Notably, a single variant with unknown taxonomy dominated *psbA* abundances during late spring/early summer, appearing after the spring *Synechococcus* bloom (dark grey in Figure 3). This *psbA* variant did not match any bacterial or eukaryotic known sequence, yet it had multiple matches to cyanophages (details not shown). The appearance of this variant coupled with the decrease of the spring *Synechococcus* bloom could be indicative of a key role of this cyanophage in the bloom demise. Similarly, recent studies have shown that cyanophage *psbA* variants can outnumber the photosynthetic host gene copies (Sieradzki *et al.*, 2019). These observations deserve further analysis that go beyond the scope of this study.

The distribution of *rbcL* followed the patterns of photosynthetic bacteria and eukaryotes (Figure 3), and was also present in some heterotrophic groups thus able to fix carbon, mainly *Rhodobacteraceae* and Gammaproteobacteria (Badger and Bek, 2008). During summer, one of the most abundant groups harboring *rbcL* was the genus Litoricola (Gammaproteobacteria), which, together with Oceanospirillales, have recently been included within the Pseudomonadales order (Liao *et al.*, 2020). Furthermore, the ability to fixate carbon in this group has recently been confirmed by a study using single amplified genomes (SAGs, Pachiadaki *et al.*, 2019). Finally, the various PR genes presented divergent seasonal patterns (Figures 2 and 3). We observed a dominance of the blue type (abundance median = 0.32), in contrast with previous results, which showed the green types to be more typical of coastal waters, whereas the blue type dominated in open waters (Pinhassi *et al.*, 2016). The average chlorophyll *a* levels of Blanes Bay were 0.64 mg m^{-3}, and average water transparency is 14 m (Gasol *et al.*, 2016), corresponding to an oligotrophic coastal site and partially explaining this result. Whereas both blue and green L105 PRs were seasonal, appearing during summer and decreasing in winter (recurrence strength = 6.6), the green M105 PR did not present a clear seasonal pattern (recurrence strength = 3.3, p = 0.21). The blue PR type was mostly found in *Pelagibacteraceae*, SAR86 and other Alpha- and Gammaproteobacteria (Figure 3). Instead, the green L105 PR was more present in SAR86 in winter and in other Gammaproteobacteria groups such as *Thioglobaceae*, while during summer the gene was present in *Puniceispirillaceae* and HIMB59. These two types of PRs were mostly present in the typically 'oligotrophic' bacteria, i.e. those with small sizes and small genomes (Pachiadaki *et al.*, 2019; Spietz *et al.*, 2019). The M105 green PR on the other hand was present mainly in *Flavobacteriaceae* groups, and dominated almost entirely by Flavobacteriales. Although multiple single groups within the Flavobacteria presented a significant seasonal trend (Teeling *et al.*, 2016), its seasonality was not always in the same season, with different genera peaking at all seasons and masking a unified single seasonal pattern, which is reflected in the non-seasonality of the green M105 PR subtype.

We also inspected *coxL* and *tauA* genes, both involved in the carbon cycle. The former codifies for the carbon monoxide dehydrogenase that oxidizes carbon monoxide (CO) to CO_2 as a supplemental energy source to survive carbon limitation, a process that has been suggested to be relevant in the coastal ocean (Moran and Miller, 2007), while *tauA* codes a transporter to incorporate taurine –an amino acid-like compound– to cells, one of the main contributors of carbon and energy source in the epipelagic waters (Clifford *et al.*, 2019). The seasonal pattern of *coxL* had its maximum in late spring, reaching a median abundance ratio of 0.3, and nearly disappearing during winter (Figure 2). The higher values were linked to *Rhodobacteraceae*, particularly to a single gene variant matching an uncultured genus –named LFER01– that was incorporated to the GTDB in a recent study in the Caspian Sea (Mehrshad *et al.*, 2016) and that belongs to the *Roseobacter* clade (Luo and Moran, 2014). During summer, *Puniceispirillales* (SAR116 clade) and *Litoricola* (Gammaproteobacteria) were also the main groups containing *coxL*. The highest values during July coincided with the maximum *coxL* transcript abundance of Rhodobacterales in a previous study (Alonso-Sáez *et al.*, 2020). With a similar abundance pattern, *tauA* reached maximum values during spring, albeit with large variability (recurrence index = 6.1). Taxonomically, *tauA* was dominated by *Pelagibacteraceae* all year around.

Focusing on the nitrogen cycle, *narB* –a gene coding a subunit of the nitrate reductase known in *Cyanobiaceae*– presented two abundance peaks matching the recurrent *Synechococcus* blooms of spring and summer (Auladell *et al.*, 2021). Taxonomically, the spring bloom presented two main *Synechococcus* variants, whereas the summer bloom was formed by a single 16S rRNA gene variant. During summer, *Flavobacteriaceae* also contained this gene, although information regarding assimilatory nitrate reductase activity of this group in seawater is lacking. A KEGG search against GTDB shows that *narB* presents 314 hits in *Flavobacteriaceae*, a similar number of matches to the ones found in *Cyanobiaceae*. The groups harboring *nasA* (nitrate reductase) were not very abundant (median ratio = 0.02) and appeared mostly during summer and autumn (Figure 2), with a single Gammaproteobacteria variant dominating from April to November (Figure 3) that was related to *Pseudohongiella nitratireducens*. During summer, NO_3- concentration reached its lowest levels at Blanes Bay, and thus both *Pseudohongiella* and *Cyanobiaceae* could be involved in the NO_3- decrease alongside eukaryotes. Following the opposite seasonal pattern, the *amoA* gene –encoding ammonia monooxygenase– was present during winter and disappeared during summer (mean abundance ratio = 0.002). Previous studies using qPCR for both 16S rRNA and *amoA* genes in Blanes Bay, observed that the patterns could be linked to Crenarchaeota Group I (Galand *et al.*, 2010). Our study indentifies that the main contributor to *amoA* was *Candidatus* Nitrosopelagicus genus, a recently described archaeal group, that was within the previously named Thaumarchaeota group (Santoro *et al.*, 2015; Rinke *et al.*, 2021). Finally, *ureC* –encoding a urease degrading urea to ammonium– presented a seasonal pattern with two states: high abundance during spring and summer (mean abundance ratio = 0.27) and lower values in autumn and winter (0.14). During winter, *Nitrosopumilaceae* and *Synechococcus* were the most common groups, whereas *Puniceispirilalles*, *Rhodobacteraceae* and *Litoricola* dominated during summer. To our knowledge, the presence of the *ureC* gene in *Puniceispirilalles* and *Litoricola* has only been noticed in a recent SAGs sequencing study (Pachiadaki *et al.*, 2019, see Supplementary Table 5).

Regarding the phosphorous cycle, our results indicate a synchronized pattern for some genes and multiple different responses for others (Figure 2). The genes encoding for functions related to polyphosphate metabolism (*ppx*, an exopolyphospatase, and *ppk1*, a polyphosphate kinase), the enzyme remodeling membrane phospholipids (*plcP*) and a hypothetical alkaline phosphatase (*phoD*) presented some of the highest abundance ratios (median ratio = 1) and a seasonal pattern with the highest values at the end of summer (Figure 2). The particularly higher abundances at the end of summer could be a reflect of the phosphorous limitation conditions typically occurring in this coastal site (Pinhassi *et al.*, 2006). On the other hand, the *phoD* alkaline phosphatase was more abundant than *phoX* in our system (mean abundance ratio 0.92 vs 0.48). The difference between these genes was identified in a previous study using data from the Global Ocean Sampling expedition in the Sargasso Sea (Luo *et al.*, 2009). In our study, the taxonomic distribution was also different, with *phoD* being mostly associated to SAR86, *Haliaceae* and *Flavobacteriaceae*, whereas *phoX* was more widespread. The other studied genes (*ppx, ppk1, plcP*) were widespread, with *Pelagibacteraceae* dominating for *ppk1* and *Flavobacteriaceae* for plcP. The *phn* genes, that were initially described in a single operon, are currently known to show multiple different syntenies (Martínez *et al.*, 2012), which could explain the differences in the abundances between *phnD* and *phnM* (Figure 2). In our analysis, *phnD* –coding the phosphonate ABC transporter– was non-seasonal, whereas phnM –coding the alkylphosphonate utilization protein– presented maxima in spring and summer. Previous results have shown that *phnD* is not expressed under phosphorous limitation (Martínez *et al.*, 2012) pointing to a possible explanation for the lack of a seasonal pattern. Taxonomically, *phnM* was assigned to the *Rhodobacteraceae* and other alphaproteobacterial groups, whereas *phnD* was more widely distributed, a fact also reflected by the number of variants detected for each gene (2730 vs 6044, respectively). The *pstS* gene –encoding a phosphate transporter– did not show seasonality, and was present mainly in alphaprotebacterial groups, *Nitrosopumilaceae*, Gammaproteobacteria and *Cyanobiaceae*. Overall, 5 out of 8 of the analyzed phosphorous genes peaked at summer.

Lastly, we analyzed a gene related to the sulfur cycle (*dmdA*) and an iron transporter (*fecA*). The *dmdA* –coding a dimethylsulfonioproprionate demethylase– turns dimethylsulfonioproprionate (DMSP) to methyl-mercaptopropionate (MMPA) to incorporate it as a source of reduced sulfur and carbon. This gene presented the highest relative abundances during spring and summer, and was dominated by *Pelagibacteraceae* (Figure 2, 3). During these seasons, the DMSP assimilation ratios were the highest in our system (Simó *et al.*, 2009). Other relevant groups presenting the *dmdA* gene were *Puniceispirillaceae* and the HIMB59 family (previously considered part of SAR11 clade V). Regarding the *fecA* gene –encoding a dicitrate siderophore transporter (Schauer *et al.*, 2008)–, it presented the highest abundance (mean abundance ratio = 2) of the ones tested and did not display seasonality. Its high abundance could be linked to being present in multiple copies per genome with variable affinities to different siderophores (Tang *et al.*, 2012). Taxonomically, the main groups harboring *fecA* were Gammaproteobacteria and Flavobacteria. The low abundance of alphaproteobacterial groups could be explained by the different strategies they use for incorporating iron, since *Rhodobacteraceae* and *Pelagibacteraceae* are specialized in obtaining the inorganic iron through transporters (Tang *et al.*, 2012, Debeljack 2019). The lack of a specific seasonal pattern in this gene is in agreement with the fact that iron is not typically a limiting factor for microorganisms in the Mediterranean Sea (Sherrell and Boyle, 1988).

Overall, our results show that a majority (57% with Lomb-Scargle ≥8 PN; 71% considering all significant values) of the functional genes present variations in abundance at a yearly scale, while prokaryote abundance varies little seasonally (Supplementary Figure 1). These results describe for the first time the seasonal trends of multiple genes and reinforce patterns observed for some of them that had been presented before (Galand *et al.*, 2010, 2018; Auladell *et al.*, 2019). Multiple functions displayed patterns that were linked to the abiotic environmental conditions. The high presence of phoshorous genes in summer is likely linked to the selected pressure exerted by inorganic nutrient limitation in the study site (Pinhassi *et al.*, 2006). Similarly, the nitrate reductases appeared simultaneously with the nitrate decrease during summer. For other functions, such as the oxidation of carbon monoxide, we did not have any biogeochemical measurements to compare with the observed patterns. Additionally, our taxonomic analysis sheds light onto the key players for each biogeochemical process, unveiling relevant groups that had not been previously considered. Examples of those are the dominance of Flavobacteria and Gammaproteobacteria in functions such as *phoD*, *nasA* and *fecA*, or the dominance of *Rhodobacteraceae*, *Puniceispirillaceae* and *Litoricolaceae* in *rbcL*, *ureC* and *coxL* during summer. The importance of *Puniceispirillaceae* and *Litoricolaceae* in these processes was so far unknown. Focusing the analysis on specific functions has proven thus useful to unveil the most likely relevant players of prokaryote-driven biogeochemical processes. This type of data however reflects the effects of community turnover on the enrichment and selection of functions. To further investigate the actual relevance of these functions, the characterization of the relative importance of seasonally varying gene expression vs. community turnover would still be needed.

Contrasted gene repertoire in the *Pelagibacteraceae* and SAR86 families

Having obtained an overview of the aggregated seasonal pattern for each gene, and the assignment of these genes to broad taxonomic groups, we wanted to deepen our knowledge into the seasonality of each particular gene variant within each of the selected functional genes. We first tested whether each variant within each gene presented seasonality and then, for those that were seasonal, we identified the yearly seasonal maxima (e.g. 'for 5 of the 7 years the maximum is in winter'). Finally, we performed a Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) ordination of the seasonal variants to visualize groups of variants that clustered together (Figure 4). For the analyzed genes, we found 6359 out of 23503 seasonal variants (Lomb-Scargle test, $q \leq 0.05$, PN ≥ 8). Most of these variants reached their maximum values only in a particular season (4862, 76%), while some presented maxima in two seasons (1523, 23%). Specifically, most of the seasonal variants peaked in winter (2430 variants, 38%), followed by autumn (1945, 30%), summer (1282, 20%), and spring (728, 12%). Multiple variants assigned to autumn and spring seasons were grouped withint the summer and winter clusters in the UMAP ordination (Figure 4A), corresponding to the natural transition between seasons. Specifically, we observed a 'tail' connecting the winter and spring clusters, which corresponds to the variants presenting an abundance maximum during April, and also that the autumn cluster positioned on the left of summer was composed mostly by variants with a maximum in October (Figure 4A). In conclusion, the four seasons presented distinguishable clusters of variants, albeit with summer and winter being more explicit, and the spring and autumn clusters being more variable. The variability of these two seasons can be probably attributed to the meterological conditions and environmental changes that undergo year-to-year variations, not always coinciding with the equinoxes and solstices dates that were used here to define the seasons.

We observed taxonomy taxonomy rather than the specific function explained each variant's seasonal pattern (Figure 4B). For most metabolic functions, the season presenting most seasonal variants was winter. During this season the system reached its highest taxonomic richness, possibly due to the mixing of waters carrying more nutrients and the increased resuspension that likely triggered the growth of multiple "rare" bacteria (Auladell *et al.*, 2021). Additionally, the seasonal genes appearing tipically in spring were from specific taxonomic groups such as *Haliaceae*, *Rhodobacteraceae*, and *Synechococcus*, whereas autumn was dominated by other groups such as *Prochlorococcus* and HTCC2089 (Pseudomonadales). On the other hand, some phosphorous genes presented seasonal patterns linked to specific abiotic conditions rather than to taxonomic composition. Genes such as *phnM*, *phoD*, *ppx*, *ppk1* and *plcP* presented the most abundant seasonal variants during summer, matching the aggregated seasonal pattern observed discussed above (Figures 2 and 4C). We wanted to further test whether there were cases in which the phosphorous gene pattern deviated from the seasonal pattern of the taxonomic family, and thus if there was niche differentiation at lower taxa levels. At the family level, only 91 families presented at least 10 seasonal gene variants. Of them, 25 families (27%) presented a higher proportion of phosphorous genes during the summer season. To test for differences between the phosphorous functional pattern and the taxon abundance pattern, we compared the 16S rRNA gene data from our previous study, aggregated at the family level, to the seasonal variant patterns (Supplementary Figure 4; Auladell

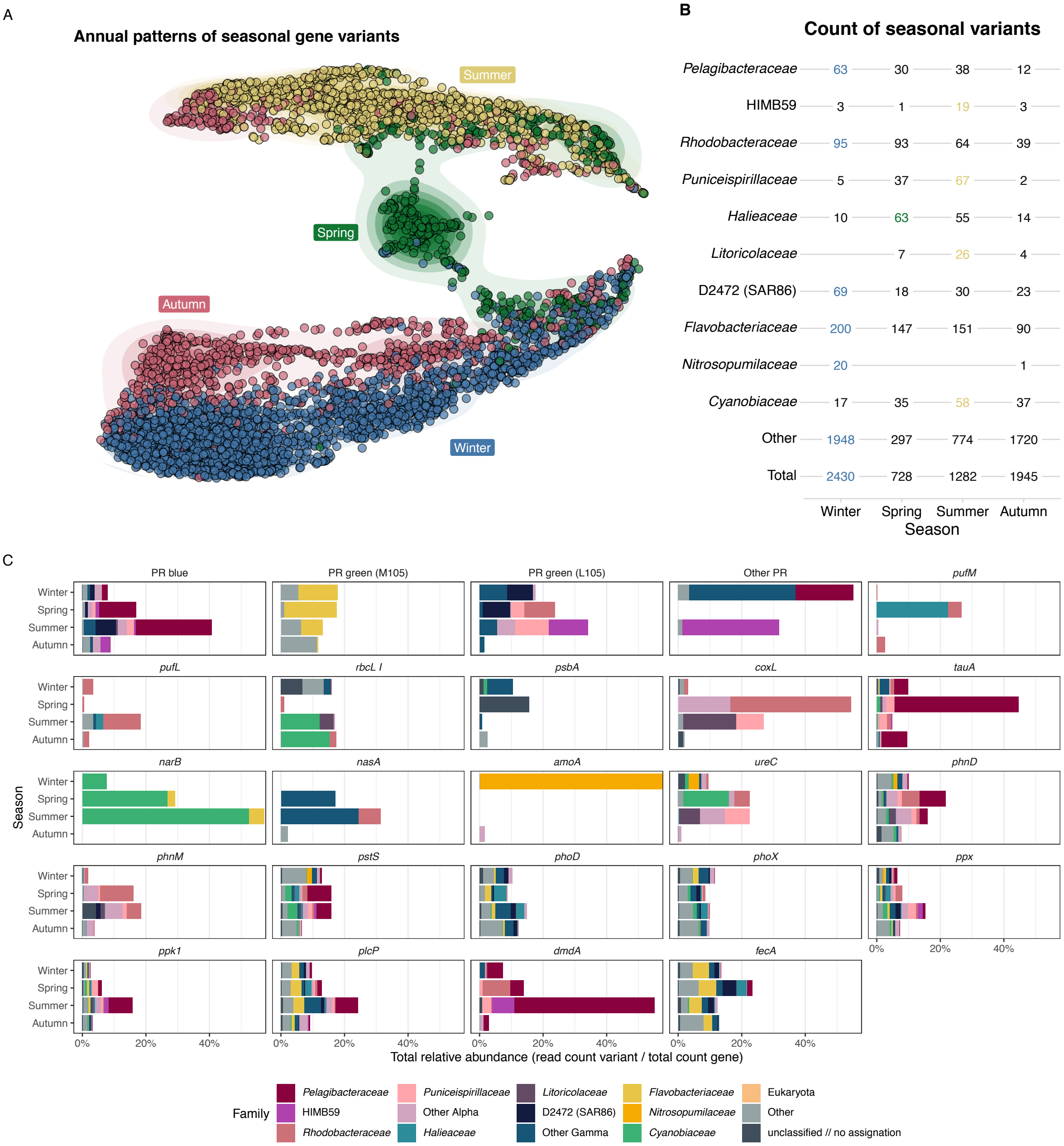


Figure 4: A) Uniform Manifold Approximation and Projection (UMAP) ordination of the gene variants presenting seasonality. The colors differentiate the seasons according to the day of the year (following the astronomical definition). B) Number of seasonal variants for each family (Y axis) and season (X axis). The colored values indicate the season in which each specific group had its the maximum. C) Seasonal distribution of the gene variants. The X axis represents the total relative abundance of each variant using the total read count of the specific gene as the denominator. The Y axis shows the season in which the gene reaches the maximum abundance ratio. The colors differentiate the main family groups as shown in the figure legend.

et al., 2021). Families such as *Puniceispirillaceae* and *Litoricolaceae* presented the same abundance maximum for the 16S rRNA gene than for the abundance of the phosphorous gene repertoire (Supplementary Figure 4). These taxa appeared to be adapted to oligotrophic summer conditions, and possibly presented a genomic repertoire that helped them to avoid phosphorous limitation. In contrast, based on the phosphorous repertoire, *Pelagibacteraceae* and SAR86 presented seasonal deviations from the 16S rRNA gene abundance patterns (Figure 5). These results indicate that at lower taxonomic levels, specific ecotypes could present a differentiated genomic repertoire adapted to the varying seasonal conditions.

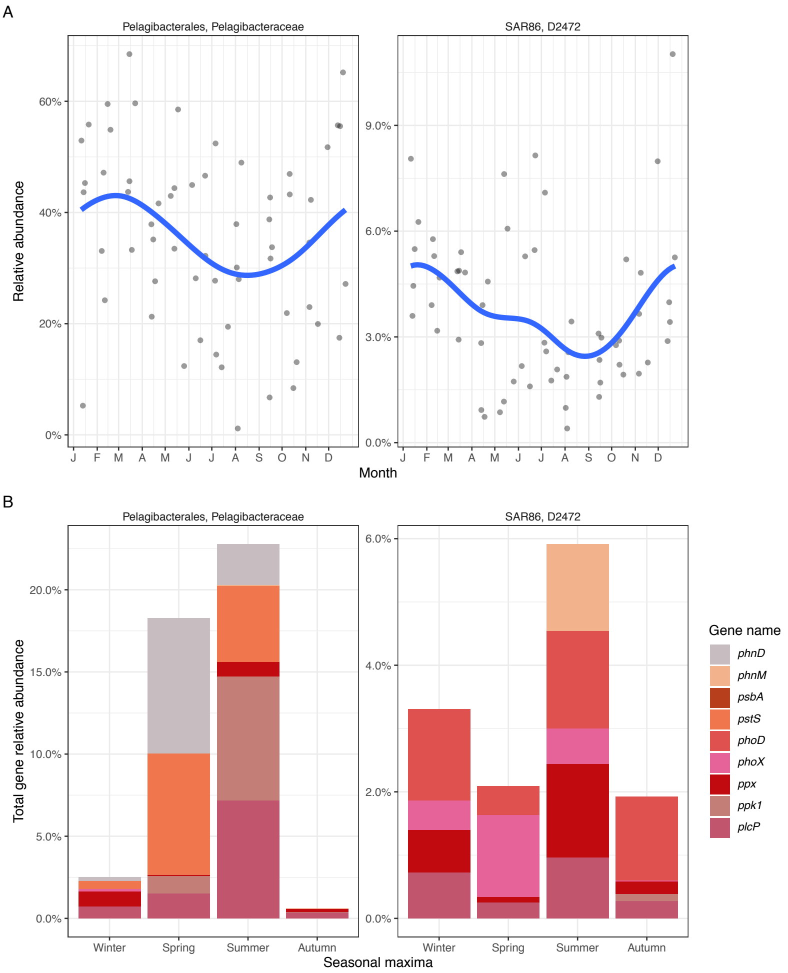


Figure 5: A) Seasonal patterns of the *Pelagibacteraceae* and D2472 (SAR86) families using the 16S rRNA gene. The X axis indicates the month and the Y axis the relative abundance of the family 16S rRNA gene read counts (obtained from Auladell *et al.* 2021). A generalized additive model smooth is adjusted to the data points. B) Seasonal distribution of the phosphorous gene variants for the selected families. The X axis is the total relative abundance of each variant using the total read count of the specific gene as the denominator. The Y axis is the season in which the gene reaches the maximum abundance ratio. The colors differentiate the phosphorous genes.

Given the observed differences at the family level for phosphorous genes, we wanted to compare the studied marker genes of these families at the genus level (Supplementary Figure 5). The most abundant seasonal genera within *Pelagibacteraceae* were *Pelagibacter* (SAR11 clade I), MED-G40 (SAR11 subclade IIa), Pelagibacter_A (SAR11 clade II), HIMB114 (SAR11 clade III) and AG-414-E02 (SAR11 subclade Ic). Within the *Pelagibacter* genus, we also differentiated one Mediterranean genomospecies (known as gMED as described in Haro-Moreno *et al.*, 2020) through a BLAST against SAGs reconstructed from our long-term station. Some variants within *Pelagibacter*, the gMED genomospecies, the MED-G40 genus and Pelagibacter_A presented the abundance pattern with a maximum during summer (Supplementary Figure 5). Most of the variants for these genera were phosphorous genes such as *phoD*, *pstS* and *plcP* in the case of the MED-G40, whereas Pelagibacter_A also presented *phoX*, *ppk1* and *ppx*. Sometimes, within each genus the seasonal variants presented a difference in the abundance ratio. As an example, the *dmdA* variant presented by the gMED genomospecies was more abundant than the rest of the genes such as *plcP* and *ppk1*. Through the BLAST matches against the SAGs, we observed that this variant was conserved among multiple SAGs, and therefore the 95% identity clustering is aggregating them, which is reflected by the abundance. In the abovementioned study on the ecogenomics of the SAR11 clade, Haro-Moreno *et al.* (2020) showed an increase in the phosphorous genes in the gMED genomospecies as compared to SAR11 genomes from other latitudes. Our results support this conclusion and extends the results to other *Pelagibacteraceae* genera such as HIMB114 and Pelagibacter_A.

A similar distribution of the phosphorous genes was observed within the D2472 family (SAR86 clade), in which the genera presenting seasonal variants were D2472, MED-G78, SAR86A and SCGC–AAA076–P13. Unfortunately, for the SAR86 clade there are no detailed phylogenies and genome descriptions as for those of the SAR11 clade. A recent study found 5 differentiated clusters for the clade, but not much discussion regarding the genomic repertoire of each cluster exists (Hoarfrost *et al.*, 2020). In Blanes Bay, both the SAR86A and SCGC–AAA076–P13 groups presented summer ecotypes containing a high proportion of phosphorous genes (Supplementary Figure 5). The latter of these genera was the only one containing *pstS* and *phnM*. A pangenomic analysis comparing the genomic repertoire of both Pelagibacterales and/or SAR86 using seasonal time-series data would help disentangle the complete genomic repertoire differentiation beyond their phosphorous gene differences. Overall, these results show how adaptation to nutrient limitations have occurred at multiple taxonomic levels, with some groups such as *Puniceisipirillaceae* presenting adaptations at the family level, whereas other groups such as *Pelagibacteraceae* and D2472 present specific genera adapted to the oligotrophic summer conditions. Furthermore, our results indicate that trait plasticity linked to the nutrient stress observed on a biogeographical dimension (Ustick *et al.*, 2021) can be also detected at the seasonal scale.

3.3 Conclusions

In this study, we unveiled the seasonal patterns of 21 key biogeochemical marker genes using a 7-year metagenomic time series from a coastal site. Our data show that the marker genes presenting the highest richness were related to phosphorous starvation and Fe3+ dicitrate transport, and that generally the patterns of gene richness followed the species richness of the whole community. Most of the studied genes presented recurrent seasonal dynamics with succession between the different taxonomic groups. Genes such as *pufLM*, *coxL*, *ureC* and *tauA* were predominant during spring, phosphorous cycling genes were enriched during summer while *amoA* presented its maximum during autumn and winter. We also identified the main taxonomic groups displaying these functions, and unveiled groups that previously had been not considered for certain functions, such as *Litoricola* for carbon fixation, CO oxidation and urease production. Finally, by analysing the seasonal patterns a fine scale (i.e., individual gene variants), we showed that the abundance patterns displayed by the phosphorous marker genes for *Pelagibacteraceae* and D2472 (SAR86 family), differed from the family abundance pattern. Our data provides a framework to understand the seaonality of key biogeochemical processes in the coastal ocean and to generate new hypotheses about the relevance of specific organisms for each of these processes.

3.4 Experimental procedures

Sampling and sequencing procedure

We collected surface water samples from the Blanes Bay Microbial Observatory (BBMO, 41°40'N, 2°48'E) as described in (Gasol *et al.*, 2016). This long-term station is a shallow (~20 m) coastal site about ~1 km offshore in the NW Mediterranean coast. We sampled every month, from January 2009 to December 2015 (7 years), and obtained a dataset of 80 samples. Several environmental parameters were collected simultaneously to generate an environmental data table with a total of 23 biotic and abiotic variables (see Auladell *et al.* (2021) for details). We used the astronomical equinoxes and solstices to define the seasons.

About 4 L of 200-µm pre-filtered surface seawater were sequentially filtered through a 20-µm mesh, a 3-µm pore- size polycarbonate filter (Poretics), and a 0.2-µm Sterivex Millipore filter using a peristaltic pump. Sterivex units were processed to obtain the genomic DNA (see Auladell *et al.*, 2021), which was stored at -80 °C. Sequencing of the samples was carried in two batches. An aliquot for the 0.2-3 µm fraction from each sample was processed using a Kapa Hyper kit, and quality control was done with an agarose gel and qubit. Afterwards, the first 3 years were sequenced using an Illumina Hiseq4000 and for the following 4 years using Illumina NovaSeq6000 (2 x 150 bp, Centre

Nacional d'Anàlisi Genòmica CNAG). The 80 samples generated a total of 22.5 billion sequences with an average 133 million reads (minimum = 2.3; maximum = 232 million reads).

Trimming, assembly and gene prediction

Samples were trimmed with TRIMMOMMATIC v0.38 to remove low quality reads (Bolger *et al.*, 2014). Each sample was then assembled individually with MEGAHIT to obtain contigs (Li *et al.*, 2015). We predicted the protein coding regions in the contigs using Prodigal v2.6.3 and MetaGeneMark v3.38 (Hyatt *et al.*, 2010; Zhu *et al.*, 2010). We obtained a redundant dataset of 5 thousand million genes, and after clustering at 95% identity and 90% coverage through Linclust v10 (Steinegger and Söding, 2018), the final catalog consisted of 231 million genes.

Gene annotation

We focused on a specific subset of genes with known functions involved in relevant metabolic processes (reviewed in Ferrera *et al.*, 2015). In particular, from the whole gene catalog we selected and annotated 24 relevant genes for the major biogeochemical cycles: *coxL, rbcL subunit I, chiA, pufM, pufL, PR, psbA, tauA, phnD, phnM, pstS, phoX, phoD, ppx, ppk1, plcP, nifH, narB, nasA, hao, amoA, ureC, dmdA* and *fecA* (see Table 1). Most of the genes were exclusive from bacteria and archaea, but *rbcL* and *psbA* are also found in eukaryotes. These genes were selected based on two main criterions: they had to show high specificity for the function of interest (in order to avoid false positives and missassignations) and they had to be well characterized in the protein databases. For most genes, we used the Kyoto Encyclopedia of Genes based in HMM (KOFAM database) (Aramaki *et al.*, 2020). This database consists in an HMM for each specific KEGG ortholog (KO) and a score threshold for filtering unspecific results. For phosphorous-related genes, that are more taxonomically widespread and genetically diverse, we used a reverse PSI-BLAST v2.7 and a custom perl script for filtering multiple hits against the Cluster of Orthologous Groups (COGs) (-soft_masking true -evalue 0.1) (Altschul *et al.*, 1990; Galperin *et al.*, 2015). Proteorhodopsins were annotated using the MicRhoDE database through DIAMOND v2.0.7 (--id 70, --query-cover 80 --evalue 0.1) (Boeuf *et al.*, 2015; Buchfink *et al.*, 2015). The putative PRs were aligned with MAFFT v7.4 together with a set of reference sequences (Olson *et al.*, 2018). Afterwards, we looked for the presence of the aminoacids implicated in the PR ion pumping mechanism (residues 97, 101 and 108) and the variations in the aminoacid shown to be important for the spectral tuning of the molecule (residue 105) (Olson *et al.*, 2018). The most common aminoacid variants for residue 105 (Q, L and M) were analyzed separately (PR blue, PR green L105 and PR green M105) aggregating the other variants as 'Other PR'. Finally, we also differentiated between two types of *coxL* –CODHI and II– by checking for the presence of an aminoacidic signature distinguishing the variants (AYXCSFR, King and Weber, 2007). In this analysis we only kept CODHI since it is the only variant with proven oxidation potential (King and Weber, 2007). After generating the abundance table, we observed that the *chiA*, *hao*, and *nifH* genes presented a low number of detected variants (min = 1, max = 8 variants) with a small read count per sample (min = 1, max = 420 read counts). Specifically, there was only one variant of *chiA* detected, while *nifH* presented 4 variants with a total read count of 336 reads and hao presented 8 variants with a total read count of 2200 reads. As a comparison, when we observed the distribution of *amoA* the total read count was 36000 (min = 1, max = 2774 read counts), 16 times more than *hao*, the most abundant one. Given that these three genes did not present enough data to determine with precision their temporal trends on a multi-year basis, they were excluded from subsequent analyses, and the subsequent analysis were performed with 21 genes.

Read mapping

We used DIAMOND to match the raw reads from each sample to our gene database (--query-cover 90, --identity 95, --top 5 --min-score 20). The output presented the top 5 matches for each read. Since proteins present conserved regions that could recruit reads incorrectly, and to avoid missassignations, we filtered the 5 top matches through the Functional Analysis of Metagenomes by Likelihood Inference (FAMLI v1.2) algorithm (Golob and Minot, 2020). Briefly, FAMLI iteratively assigns multi-mapping reads to the most likely true peptide through checking the coverage evenness along the length of the sequence.

Taxonomy

To assign the taxonomy to each gene variant we used the last common ancestor (LCA) algorithm as implemented in MMSEQ2 v13 (Steinegger and Söding, 2017; Mirdita *et al.*, 2021). For each contig in the database, MMSEQ2 predicts the individual coding sequences, establishes the putative taxonomy of the genes through the LCA, and checks the whole contig taxonomy concordance. We used the Genome Taxonomy DataBase (GTDB, release 95), presenting 194.600 genomes in 31.910 species clusters (Parks *et al.*, 2018). The taxonomy was also assigned with UniRef90 to obtain matches for eukaryotic genes (Suzek *et al.*, 2007). For the variants matching *Pelagibacteraceae*, an additional step was performed to improve the taxonomy assignation. In a previous study, Haro-Moreno *et al.*, (2020) obtained SAGs from the SAR11 clade in the BBMO long-term station. Through a BLAST analysis (-perc_identity 95, -max_target_seqs 10, -cov 95) we differentiated the gMED subclade (Haro-Moreno *et al.*, 2020) by the matching between these variants and the SAG genomes. Supplementary Table 1 links the classic NCBI nomenclature with the GTDB nomenclature, providing references with the reasoning behind specific name changes.

Statistics

We performed all the analyses with the R v3.5 language (R Core Team, 2014). We used tidyverse v1.3 to process the data and ggplot2 v3.2 for all visualizations. For the analysis of seasonality, the gene variant read counts were transformed to ratios. Sample-wise, the gene read count is divided by the geometric mean of a set of 8 single copy gene (GTP1, *pheS, argS, serS, cysS, tsaD, ffh, ftsY*) read counts, obtaining a ratio instead of relative abundances. Mathematically, working with ratios instead of relative abundances avoids the proportion constraints; if the multiple samples are transformed to proportions with a total of 100%, when one gene increases substantially, the others are necessarily decreasing (Gloor *et al.*, 2017). Single copy gene abundances are used as a denominator to obtain a common scale. The single copy gene read count was correlated with total sample sequencing depth (see Supplementary Figure 6), removing the influence of sequencing depth. To test whether each of the genes displayed seasonality –that is, recurrent changes over time– we used the Lomb-Scargle periodogram (LSP) as implemented in the *lomb* package v1.2 (Ruf, 1999). Briefly, the LSP determines the spectrum of frequencies (the different sine waves with periods, for example half a year or one year) composing the dataset. Afterwards, through data randomizations, it tests whether the observed periods could occur by chance through a random distribution ($q \leq 0.01$, FDR correction). Through the peak normalized (PN) score we determine how strong is the recurrence of an analyzed gene. We considered the results as seasonal only if PN was above 8 and $q \leq 0.01$. In a previous study, we had used a threshold of PN ≥ 10, but we decided to decrease the threshold as it was considered too stringent based on an analyzis of the same dataset with an alternative methodology called recurrence index (Giner *et al.*, 2019). We found that a PN = 8 presented more concordance between methods (Supplementary Information 1). The seasonal test was only applied to gene variants present in at least 8 samples (which is 10% of the samples). Finally, we wanted to disentangle if the gene variants clustered by season. We performed an ordination analysis of all the seasonal gene variants using the Uniform Manifold Approximation and Projection (UMAP, McInnes *et al.*, 2020). UMAP is a novel dimension reduction technique used when the datasets are complex and large. Given that we have ~7000 seasonal gene variants, this approach is faster and more comprehensive than common ordinations such as non-metric multidimensional scaling (NMDS).

3.5 Acknowledgements

We thank all the people involved in operating the BBMO, especially Clara Cardelús and Captain Anselm for facilitating sampling, and Vanessa Balagué for laboratory procedures, including DNA extraction. Special thanks to Ramon Massana for sharing the unpublished flagellate data of the time series. We also thank the MarBits Bioinformatics platform of the Institut de Ciències del Mar, in particular Pablo Sánchez, for computing support. This research was funded by grants REMEI (CTM2015-70340-R), MIAU (RTI2018-101025-B-I00) and ECLIPSE (PID2019-110128RB-100) from the Spanish Ministry of Science and Innovation, and received support from the Spanish government through the 'Severo Ochoa Centre of Excellence' accreditation (CEX2019-000924-S). Adrià Auladell was supported by a Spanish FPI grant from the Ministry of Science and Innovation. Isabel Ferrera had the support of a '2019 Leonardo Grant for Researchers and Cultural Creators' from the BBVA Foundation. The Foundation takes no responsibility for the contents of this publication, which are entirely the responsibility of its authors.

3.6. References

Acinas, S.G., Sánchez, P., Salazar, G., Cornejo-Castillo, F.M., Sebastián, M., Logares, R., *et al.* (2021) Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun Biol* 4: 1–15.

Alonso-Sáez, L., Balagué, V., Sà, E.L., Sánchez, O., González, J.M., Pinhassi, J., *et al.* (2007) Seasonality in bacterial diversity in north-west Mediterranean coastal waters: Assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiol Ecol* 60: 98–112.

Alonso-Sáez, L., Morán, X.A.G., and González, J.M. (2020) Transcriptional Patterns of Biogeochemically Relevant Marker Genes by Temperate Marine Bacteria. *Front Microbiol* 11: 465.

Alonso-Sáez, L., Waller, A.S., Mende, D.R., Bakker, K., Farnelid, H., Yager, P.L., *et al.* (2012) Role for urea in nitrification by polar marine Archaea. *PNAS* 109: 17989–17994.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36: 2251–2252.

Auladell, A., Barberán, A., Logares, R., Garcés, E., Gasol, J.M., and Ferrera, I. (2021) Seasonal niche differentiation among closely related marine bacteria. *ISME J* 1–12.

Auladell, A., Sánchez, P., Sánchez, O., Gasol, J.M., and Ferrera, I. (2019) Long-term seasonal and interannual variability of marine aerobic anoxygenic photoheterotrophic bacteria. *ISME J* 13: 1975–1987.

Badger, M.R. and Bek, E.J. (2008) Multiple Rubisco forms in proteobacteria: their functional significance in relation to CO_2 acquisition by the CBB cycle. *J Exp Bot* 59: 1525–1541.

Bar-On, Y.M., Phillips, R., and Milo, R. (2018) The biomass distribution on Earth. *PNAS* 115: 6506–6511.

Beier, S., Andersson, A.F., Galand, P.E., Hochart, C., Logue, J.B., McMahon, K., and Bertilsson, S. (2020) The environment drives microbial trait variability in aquatic habitats. *Mol Ecol* 29: 4605–4617.

Béjà, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P., *et al.* (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* 289: 1902–1906.

Boeuf, D., Audic, S., Brillet-Guéguen, L., Caron, C., and Jeanthon, C. (2015) MicRhoDE: a curated database for the analysis of microbial rhodopsin diversity and evolution. Database (Oxford) 2015: bav080.

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

Bonnain, C., Breitbart, M., and Buck, K.N. (2016) The Ferrojan Horse Hypothesis: Iron-Virus Interactions in the Ocean. *Front Mar Sci* 3:82.

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

Carini, P., Mooy, B.A.S.V., Thrash, J.C., White, A., Zhao, Y., Campbell, E.O., *et al.* (2015) SAR11 lipid renovation in response to phosphate starvation. *PNAS* 112: 7767–7772.

Clifford, E.L., Varela, M.M., De Corte, D., Bode, A., Ortiz, V., Herndl, G.J., and Sintes, E. (2019) Taurine Is a Major Carbon and Energy Source for Marine Prokaryotes in the North Atlantic Ocean off the Iberian Peninsula. *Microb Ecol* 78: 299–312.

Cordero, P.R.F., Bayly, K., Man Leung, P., Huang, C., Islam, Z.F., Schittenhelm, R.B., *et al.* (2019) Atmospheric carbon monoxide oxidation is a widespread mechanism supporting microbial survival. *ISME J* 13: 2868–2881.

Delmont, T.O., Quince, C., Shaiber, A., Esen, Ö.C., Lee, S.T., Rappé, M.S., *et al.* (2018) Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* 3: 804–813.

DeLong, E.F. and Béjà, O. (2010) The Light-Driven Proton Pump Proteorhodopsin Enhances Bacterial Survival during Tough Times. *PLoS Biol* 8: e1000359.

Dyhrman, S., Ammerman, J., and Van Mooy, B. (2007) Microbes and the Marine Phosphorus Cycle. *Oceanog* 20: 110–116.

Falkowski, P. (2012) Ocean Science: The power of plankton. *Nature* 483: S17–S20.

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

Ferrera, I., Borrego, C.M., Salazar, G., and Gasol, J.M. (2014) Marked seasonality of aerobic anoxygenic phototrophic bacteria in the coastal NW Mediterranean Sea as revealed by cell abundance, pigment concentration and pyrosequencing of *pufM* gene. *Environ Microbiol* 16: 2953–2965.

Ferrera, I., Sebastian, M., Acinas, S.G., and Gasol, J.M. (2015) Prokaryotic functional gene diversity in the sunlit ocean: Stumbling in the dark. *Curr Opin Microbiol* 25: 33–39.

Finkel, O.M., Béjà, O., and Belkin, S. (2013) Global abundance of microbial rhodopsins. *ISME J* 7: 448–451.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincon, J., Zabala, L.L., Jiao, N., *et al.* (2013) Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *PNAS* 110: 9824–9829.

Fuhrman, J.A., Cram, J.A., and Needham, D.M. (2015) Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol* 13: 133–146.

Galand, P.E., Gutiérrez-Provecho, C., Massana, R., Gasol, J.M., and Casamayor, E.O. (2010) Inter-annual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea (Blanes Bay Microbial Observatory). *Limnol Oceanogr* 55: 2117–2125.

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

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* 43: D261-269.

Gasol, J.M., Cardelús, C., Morán, X.A.G., Balagué, V., Forn, I., Marrasé, C., *et al.* (2016) Seasonal patterns in phytoplankton photosynthetic parameters and primary production at a coastal NW Mediterranean site. *Sci Mar* 80S1: 63–77.

Gasol, J.M. and Kirchman, D.L. eds. (2018) Microbial ecology of the oceans, third edition. Hoboken, NJ: John Wiley & Sons/Blackwell.

Gazulla, C., Auladell, A., Junger, P.C., Royo-Llonch, M., Duarte, C.M., Gasol, J.M., *et al.* (2021) Global diversity and distribution of aerobic anoxygenic phototrophs in the tropical and subtropical oceans. *Environ Microbiol.*

Giner, C.R., Balagué, V., Krabberød, A.K., Ferrera, I., Reñé, A., Garcés, E., *et al.* (2019) Quantifying long-term recurrence in planktonic microbial eukaryotes. *Mol Ecol* 28: 923–935.

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

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017) Microbiome datasets are compositional: And this is not optional. *Front Microbiol* 8: 1–6.

Golob, J.L. and Minot, S.S. (2020) In silico benchmarking of metagenomic tools for coding sequence detection reveals the limits of sensitivity and precision. *BMC Bioinform* 21: 459.

Gómez-Consarnau, L., Akram, N., Lindell, K., Pedersen, A., Neutze, R., Milton, D.L., *et al.* (2010) Proteorhodopsin Phototrophy Promotes Survival of Marine Bacteria during Starvation. *PLoS Biol* 8: e1000358.

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

Gómez-Consarnau, L., Raven, J.A., Levine, N.M., Cutter, L.S., Wang, D., Seegers, B., *et al.* (2019) Microbial rhodopsins are major contributors to the solar energy captured in the sea. *Sci Adv* 5: eaaw8855.

Graham, E.D., Heidelberg, J.F., and Tully, B.J. (2018) Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J* 12: 1861.

Grossart, H., Massana, R., McMahon, K.D., and Walsh, D.A. (2020) Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnol Oceanogr* 65: S2–S20.

Haro-Moreno, J.M., Rodriguez-Valera, F., Rosselli, R., Martinez-Hernandez, F., Roda-Garcia, J.J., Gomez, M.L., *et al.* (2020) Ecogenomics of the SAR11 clade. *Environ Microbiol* 22: 1748–1763.

Hoarfrost, A., Nayfach, S., Ladau, J., Yooseph, S., Arnosti, C., Dupont, C.L., and Pollard, K.S. (2020) Global ecotypes in the ubiquitous marine clade SAR86. *ISME J* 14: 178–188.

Hunter-Cevera, K.R., Neubert, M.G., Olson, R.J., Solow, A.R., Shalapyonok, A., and Sosik, H.M. (2016) Physiological and ecological drivers of early spring blooms of a coastal phytoplankter. *Science* 354: 326–329.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform* 11: 119.

Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., *et al.* (2014) Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science* 344: 416–420.

King, G.M. and Weber, C.F. (2007) Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nat Rev Microbiol* 5: 107–118.

Koblížek, M. (2015) Ecology of aerobic anoxygenic phototrophs in aquatic environments. *FEMS Microbiol Rev* 39: 854–870.

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., and Galand, P.E. (2019) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *ISME J* 13: 388–401.

Lehours, A.C., Enault, F., Boeuf, D., and Jeanthon, C. (2018) Biogeographic patterns of aerobic anoxygenic phototrophic bacteria reveal an ecological consistency of phylogenetic clades in different oceanic biomes. *Sci Rep* 8: 4105.

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

Liao, H., Lin, X., Li, Y., Qu, M., and Tian, Y. (2020) Reclassification of the Taxonomic Framework of Orders Cellvibrionales, Oceanospirillales, Pseudomonadales, and Alteromonadales in Class Gammaproteobacteria through Phylogenomic Tree Analysis. *mSystems* 5: e00543-20.

Louca, S., Polz, M.F., Mazel, F., Albright, M.B.N., Huber, J.A., O'Connor, M.I., *et al.* (2018) Function and functional redundancy in microbial systems. *Nat Ecol Evol* 2: 936–943.

Luo, H., Benner, R., Long, R.A., and Hu, J. (2009) Subcellular localization of marine bacterial alkaline phosphatases. *PNAS* 106: 21219–21223.

Luo, H. and Moran, M.A. (2014) Evolutionary ecology of the marine Roseobacter clade. *Microbiol Mol Biol Rev* 78: 573-587.

Martínez, A., Osburne, M.S., Sharma, A.K., DeLong, E.F., and Chisholm, S.W. (2012) Phosphite utilization by the marine picocyanobacterium Prochlorococcus MIT9301. *Environ Microbiol* 14: 1363–1377.

Martiny, A.C., Coleman, M.L., and Chisholm, S.W. (2006) Phosphate acquisition genes in Prochlorococcus ecotypes: Evidence for genome-wide adaptation. *PNAS* 103: 12552–12557.

Martiny, A.C., Kathuria, S., and Berube, P.M. (2009) Widespread metabolic potential for nitrite and nitrate assimilation among Prochlorococcus ecotypes. *PNAS* 106: 10787–10792.

McInnes, L., Healy, J., and Melville, J. (2020) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:180203426 [cs, stat].

Mehrshad, M., Amoozegar, M.A., Ghai, R., Shahzadeh Fazeli, S.A., and Rodriguez-Valera, F. (2016) Genome Reconstruction from Metagenomic Data Sets Reveals Novel Microbes in the Brackish Waters of the Caspian Sea. *Appl Environ Microbiol* 82: 1599–1612.

Mestre, M., Höfer, J., Sala, M.M., and Gasol, J.M. (2020) Seasonal Variation of Bacterial Diversity Along the Marine Particulate Matter Continuum. *Front Microbiol* 11: 1590.

Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J., and Levy Karin, E. (2021) Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* 37: 3029–3031.

Moran, M.A. and Miller, W.L. (2007) Resourceful heterotrophs make the most of light in the coastal ocean. *Nat Rev Microbiol* 5: 792–800.

Moran, M.A., Satinsky, B., Gifford, S.M., Luo, H., Rivers, A., Chan, L.-K., *et al.* (2013) Sizing up metatranscriptomics. *ISME J* 7: 237–243.

Nguyen, D., Maranger, R., Balagué, V., Coll-Lladó, M., Lovejoy, C., and Pedrós-Alió, C. (2015) Winter diversity and expression of proteorhodopsin genes in a polar ocean. *ISME J* 9: 1835–1845.

Nunes, S., Latasa, M., Gasol, J.M., and Estrada, M. (2018) Seasonal and interannual variability of phytoplankton community structure in a Mediterranean coastal site. *Mar Ecol Prog Ser* 592: 57–75.

Olson, D.K., Yoshizawa, S., Boeuf, D., Iwasaki, W., and DeLong, E.F. (2018) Proteorhodopsin variability and distribution in the North Pacific Subtropical Gyre. *ISME J* 12: 1047–1060.

Pachiadaki, M.G., Brown, J.M., Brown, J., Bezuidt, O., Berube, P.M., Biller, S.J., *et al.* (2019) Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* 179: 1623-1635.e11.

Paerl, R.W., Turk, K.A., Beinart, R.A., Chavez, F.P., and Zehr, J.P. (2012) Seasonal change in the abundance of Synechococcus and multiple distinct phylotypes in Monterey Bay determined by rbcL and narB quantitative PCR: Seasonal Synechococcus abundances at Monterey Bay Mooring M0. *Environ Microbiol* 14: 580–593.

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

Pereira, O., Hochart, C., Boeuf, D., Auguet, J.C., Debroas, D., and Galand, P.E. (2021) Seasonality of archaeal proteorhodopsin and associated Marine Group IIb ecotypes (Ca. Poseidoniales) in the North Western Mediterranean Sea. *ISME J* 15: 1302–1316.

Pinhassi, J., DeLong, E.F., Béjà, O., González, J.M., and Pedrós-Alió, C. (2016) Marine Bacterial and Archaeal Ion-Pumping Rhodopsins: Genetic Diversity, Physiology, and Ecology. *Microbiol Mol Biol Rev* 80: 929–954.

Pinhassi, J., Gómez-Consarnau, L., Alonso-Sáez, L., Sala, M., Vidal, M., Pedrós-Alió, C., and Gasol, J. (2006) Seasonal changes in bacterioplankton nutrient limitation and their effects on bacterial community composition in the NW Mediterranean Sea. *Aquat Microb Ecol* 44: 241–252.

R Core Team (2014) R: A language and environment for statistical computing.

Rinke, C., Chuvochina, M., Mussig, A.J., Chaumeil, P.-A., Davín, A.A., Waite, D.W., *et al.* (2021) A standardized archaeal taxonomy for the Genome Taxonomy Database. Nat Microbiol 6: 946–959.

Ruf, T. (1999) The Lomb-Scargle Periodogram in Biological Rhythm Research: Analysis of Incomplete and Unequally Spaced Time-Series. *Biological Rhythm Research* 30: 178–201.

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., *et al.* (2019) Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* 179: 1068-1083.e21.

Santoro, A.E., Dupont, C.L., Richter, R.A., Craig, M.T., Carini, P., McIlvin, M.R., *et al.* (2015) Genomic and proteomic characterization of “Candidatus Nitrosopelagicus brevis”: An ammonia-oxidizing archaeon from the open ocean. *PNAS* 112: 1173–1178.

Schauer, K., Rodionov, D.A., and de Reuse, H. (2008) New substrates for TonB-dependent transport: do we only see the ‘tip of the iceberg’? *Trends Biochem Sci* 33: 330–338.

Sebastián, M. and Ammerman, J.W. (2009) The alkaline phosphatase PhoX is more widely distributed in marine bacteria than the classical PhoA. *ISME J* 3: 563–572.

Sebastián, M., Smith, A.F., González, J.M., Fredricks, H.F., Van Mooy, B., Koblížek, M., *et al.* (2016) Lipid remodelling is a widespread strategy in marine heterotrophic bacteria upon phosphorus deficiency. *ISME J* **10**: 968–978.

Sherrell, R.M. and Boyle, E.A. (1988) Zinc, chromium, vanadium and iron in the Mediterranean Sea. *Deep-Sea Res Oceanogr Res* **35**: 1319–1334.

Sieradzki, E.T., Ignacio-Espinoza, J.C., Needham, D.M., Fichot, E.B., and Fuhrman, J.A. (2019) Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nat Commun* **10**: 1169.

Simó, R., Vila-Costa, M., Alonso-Sáez, L., Cardelús, C., Guadayol, Ò., Vázquez-Domínguez, E., and Gasol, J. (2009) Annual DMSP contribution to S and C fluxes through phytoplankton and bacterioplankton in a NW Mediterranean coastal site. *Aquat Microb Ecol* **57**: 43–55.

Spietz, R.L., Marshall, K.T., Zhao, X., and Morris, R.M. (2019) Complete Genome Sequence of "Candidatus Thioglobus sp." Strain NP1, an Open-Ocean Isolate from the SUP05 Clade of Marine Gammaproteobacteria. Microbiol Resour Announc 8: e00097-19.

Steindler, L., Schwalbach, M.S., Smith, D.P., Chan, F., and Giovannoni, S.J. (2011) Energy Starved Candidatus Pelagibacter Ubique Substitutes Light-Mediated ATP Production for Endogenous Carbon Respiration. *PLoS ONE* **6**: e19725.

Steinegger, M. and Söding, J. (2018) Clustering huge protein sequence sets in linear time. Nature Communications 9: 2542-2542.

Steinegger, M. and Söding, J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**: 1026–1028.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., *et al.* (2015) Ocean plankton. Structure and function of the global ocean microbiome. Science 348: 1261359.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**: 1282–1288.

Tang, K., Jiao, N., Liu, K., Zhang, Y., and Li, S. (2012) Distribution and Functions of TonB-Dependent Transporters in Marine Bacteria and Environments: Implications for Dissolved Organic Matter Utilization. *PLoS ONE* **7**: e41204.

Teeling, H., Fuchs, B.M., Bennke, C.M., Krüger, K., Chafee, M., Kappelmann, L., *et al.* (2016) Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. *eLife* **5**: e11888.

Ustick, L.J., Larkin, A.A., Garcia, C.A., Garcia, N.S., Brock, M.L., Lee, J.A., *et al.* (2021) Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science* **372**: 287.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., *et al.* (2007) The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biology* **5**: 0432–0466.

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* **38**: e132–e132.

3.7 Supplementary figures

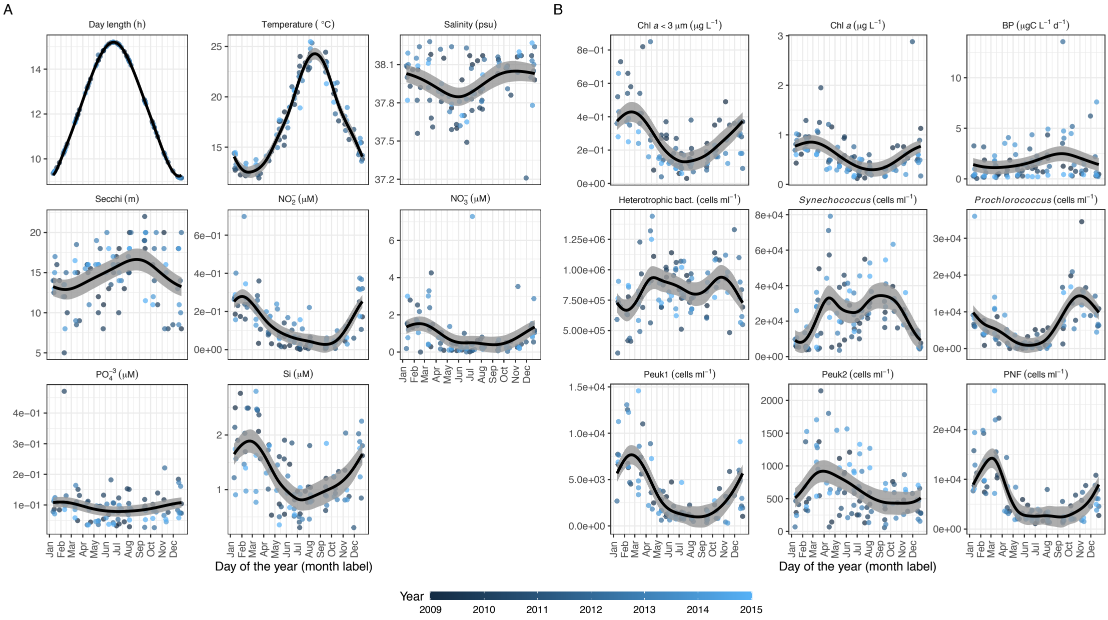


Figure S1: Distribution of the physicochemical (A) and biological (B) environmental variables measured in the Blanes Bay Microbial Observatory during the 7 years used in this study. The Y axis corresponds to the parameter value (units indicated in the plot title) and the X axis corresponds to the day of the year (month is shown for orientation, with the line ticks for the first day of each month). A generalized additive model is fitted to the data. Chla<3 µm: Chlorophyll *a* from the 3-µm fraction or smaller; BP: Bacterial production; Peuk1: small picoeukaryotes; Peuk2: large picoeukaryotes; PNF: Phototrophic nanoflagellates.

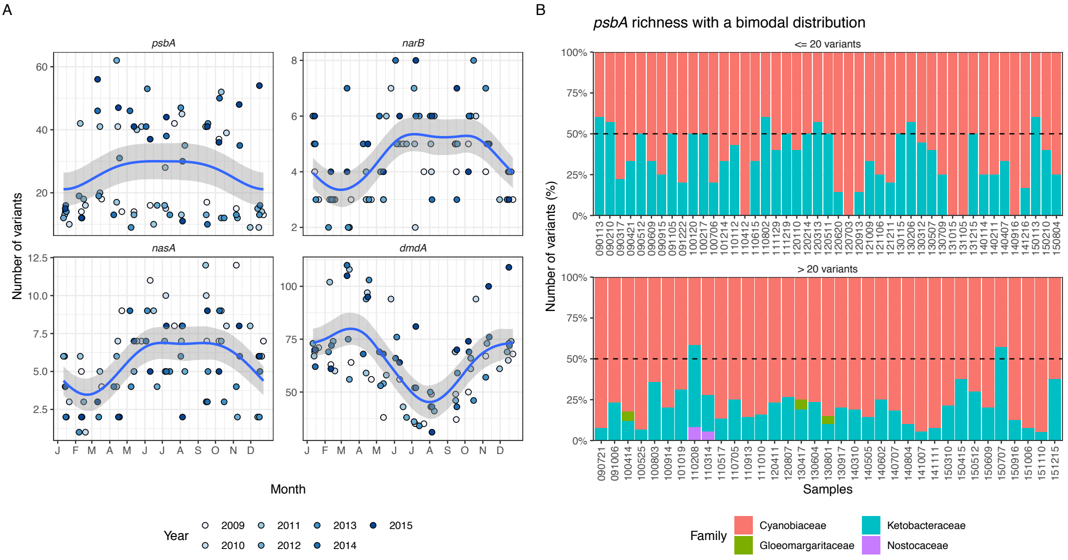


Figure S2: A) Temporal trend of richness (number of variants) for *psbA, narB, nasA* and *dmdA* genes. The X axis displays the month and the Y axis the number of variants for each sample. The points are colored by the year of sampling. B) Barplot differentiating the taxonomic origin of the psbA gene variants for the samples with less than 20 variants (upper panel) and the samples with more than 20 variants (lower panel). The bars are colored by the taxonomic family.

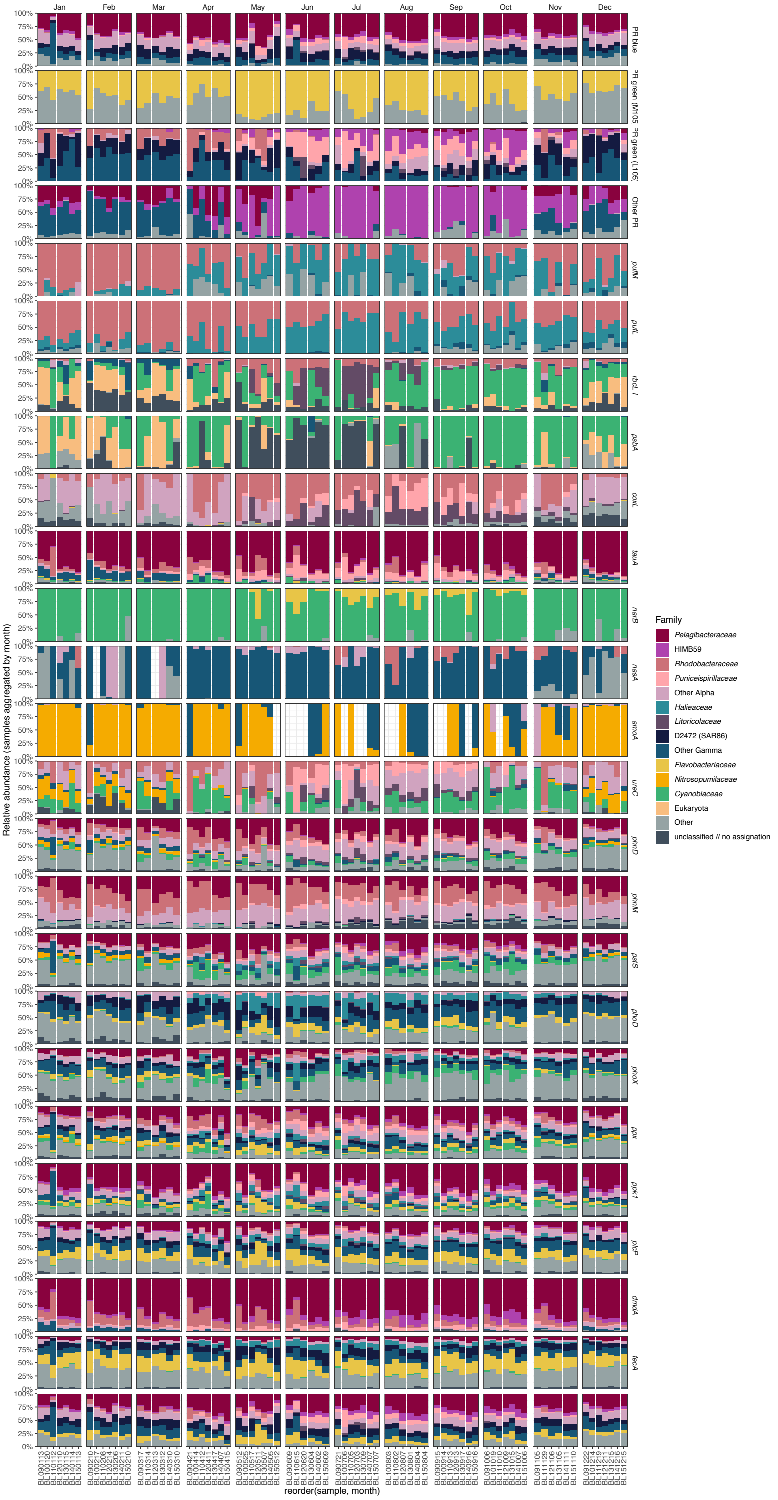


Figure S3: Selected taxonomic distribution for each individual sample. Each panel is a gene and each block presents the different samples in a specific month. The Y axis represents the relative abundance. The colors differentiate the main family groups.

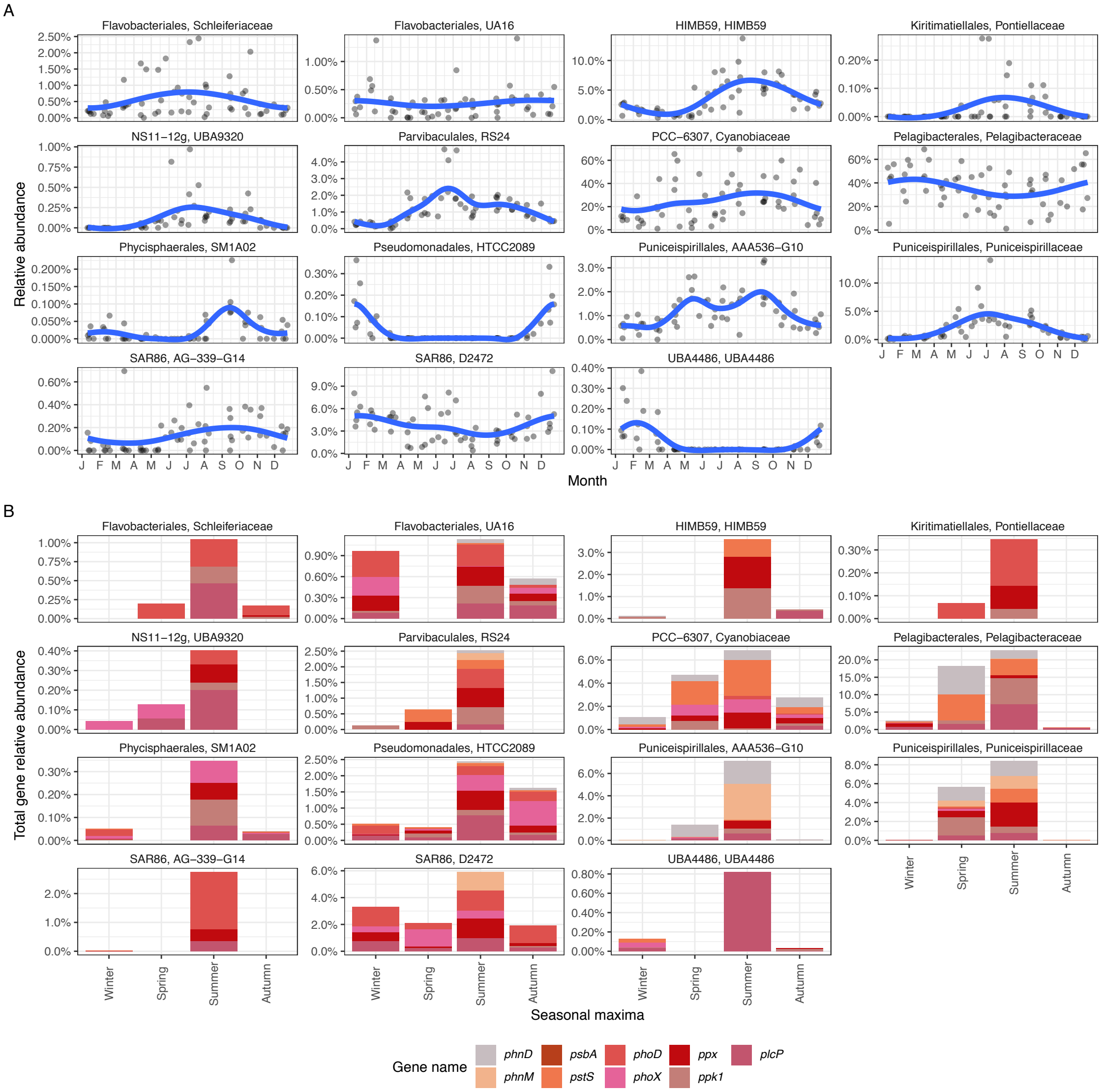


Figure S4: A) Seasonal pattern of the abundance of selected families presenting enrichment of phosphorous genes during summer. The X axis presents the month and the Y axis presents the centered logarithm ratio abundance of the family 16S rRNA gene read counts from Auladell *et al.* (2021). A generalized additive model smooth is adjusted to the data points. B) Seasonal distribution of the phosphorous gene variants for the selected families presented in panel A. The X axis is the season in which the gene reaches the maximum abundance ratio. The Y axis is the total relative abundance of each variant using the total read count of the specific gene as the denominator. The colors differentiate the different phosphorous genes.

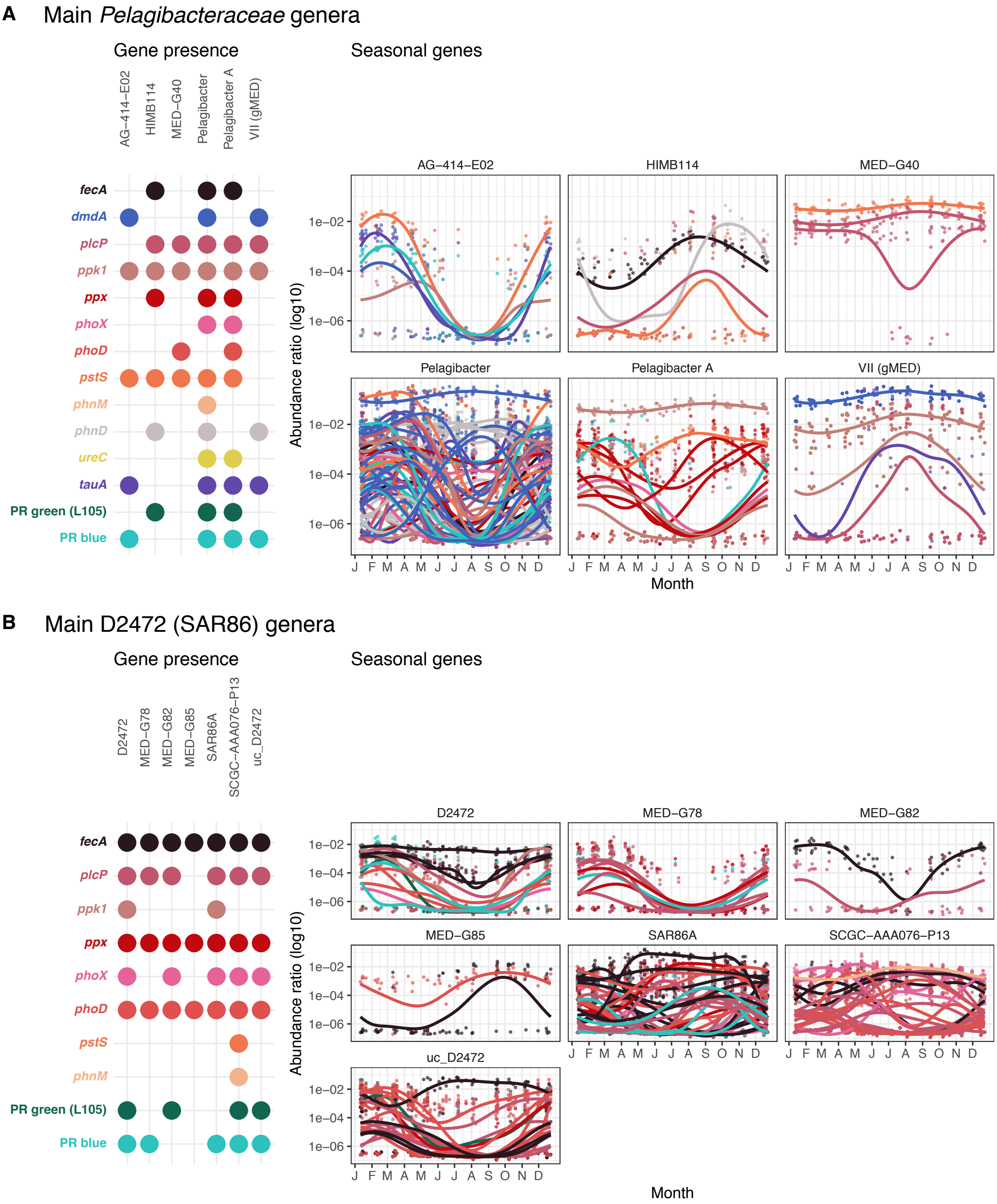


Figure S5: Temporal pattern of the abundance of seasonal gene variants for the main genera of *Pelagibacteraceae* (A) and D2472 (SAR86 family) (B). The plot on the left shows the presence of the genes (both seasonal and non-seasonal). The right plot shows the temporal distribution of the seasonal genes. The X axis indicates the day of the year (labelled by the month initials) and the Y axis presents the ratio between the gene read count divided by the geometric mean of a selection of single copy genes (see Experimental Procedures). Genes are colored following the palette in the previous panel.

3.7 Supplementary figures

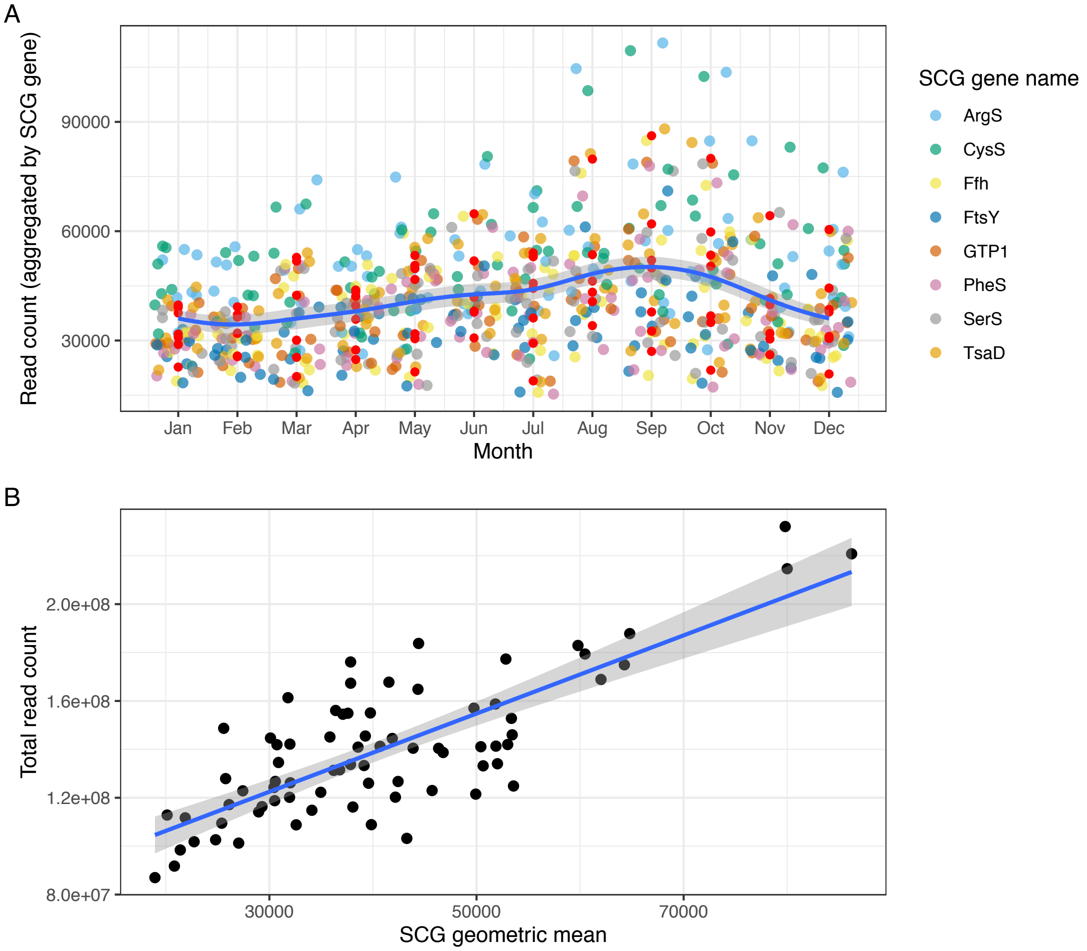


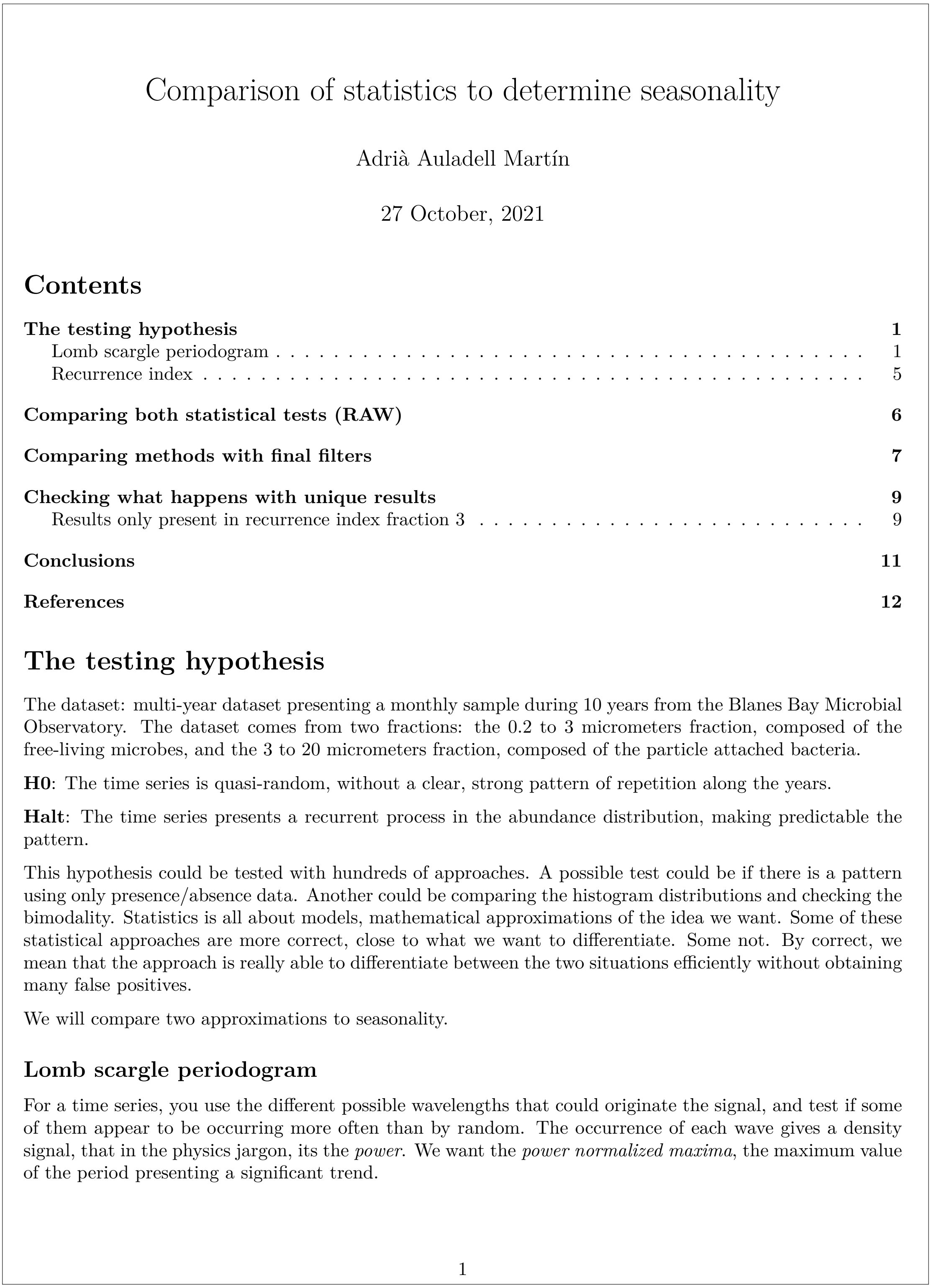
Figure S6: A) Read count change through the months. The X axis indicates the months and the Y axis the read counts. Each colored point represents the read count for each single copy gene and the red points show the sample median. A linear smooth is presented to better visualize the pattern. B) Relationship between the total sample read count and the single copy gene geometric mean.

3.8 Supplementary tables

Classic nomenclature	**GTDB nomenclature (r95)**	**Main change**	**References**
SAR116	Puniceispirilalles (order)	Isolation of *Puniceispirillum* and phylogenetic differentiation	Oh, H.-M., Kwon, K.K., Kang, I., Kang, S.G., Lee, J.-H., Kim, S.-J., and Cho, J.-C. (2010) Complete Genome Sequence of "Candidatus Puniceispirillum marinum" IMCC1322, a Representative of the SAR116 Clade in the Alphaproteobacteria. *JB* **192**: 3240–3241.
SAR86	D2472 (genus), SAR86ABCD (genera)	Standarization of the taxonomic ranks and nomenclature change.	Dupont, C.L., Rusch, D.B., Yooseph, S., Lombardo, M.-J., Alexander Richter, R., Valas, R., et al. (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J 6: 1186–1199.
Oceanospirillales	-	Disappareance of the order due to not being monophyletic, to be included in Pseudomonadales	Liao, H., Lin, X., Li, Y., Qu, M., and Tian, Y. (2020) Reclassification of the Taxonomic Framework of Orders *Cellvibrionales, Oceanospirillales, Pseudomonadales* , and *Alteromonadales* in Class *Gammaproteobacteria* through Phylogenomic Tree Analysis. *mSystems* **5**:.
SAR11 clade V	HIMB59 (order)	A phylogenetic rooting outside the Pelagibacterales order.	Viklund, J., Martijn, J., Ettema, T.J.G., and Andersson, S.G.E. (2013) Comparative and Phylogenomic Evidence That the Alphaproteobacterium HIMB59 Is Not a Member of the Oceanic SAR11 Clade. *PLoS ONE* **8**: e78858.
SAR11 clade II	Pelagibacter_A (genus)	Standarization of the taxonomic ranks and nomenclature change.	-
Euryarchaeota Marine Group II	Poseidoniaceales (order)	Name for the whole order and description of the two main families, *Poseidoniaceaceae* and *Thalassarchaeaceae*.	Rinke, C., Rubino, F., Messer, L.F., Youssef, N., Parks, D.H., Chuvochina, M., et al. (2019) A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.). *ISME J* **13**: 663–675.

Supplementary Table 1: Correspondence between the classical nomenclature and GTDB nomenclature (release r95). 'Classic nomenclature' is the original name for referring to that specific group, 'GTDB nomenclature' is the name used and/or an updated version of the name given the new information, 'Main change' specifies in which way the change happened and 'References' indicates the literature in which the change is based.

3.9 Supplementary information



Comparison of statistics to determine seasonality

Adrià Auladell Martín

27 October, 2021

Contents

The testing hypothesis

The dataset: multi-year dataset presenting a monthly sample during 10 years from the Blanes Bay Microbial Observatory. The dataset comes from two fractions: the 0.2 to 3 micrometers fraction, composed of the free-living microbes, and the 3 to 20 micrometers fraction, composed of the particle attached bacteria.

H0: The time series is quasi-random, without a clear, strong pattern of repetition along the years.

Halt: The time series presents a recurrent process in the abundance distribution, making predictable the pattern.

This hypothesis could be tested with hundreds of approaches. A possible test could be if there is a pattern using only presence/absence data. Another could be comparing the histogram distributions and checking the bimodality. Statistics is all about models, mathematical approximations of the idea we want. Some of these statistical approaches are more correct, close to what we want to differentiate. Some not. By correct, we mean that the approach is really able to differentiate between the two situations efficiently without obtaining many false positives.

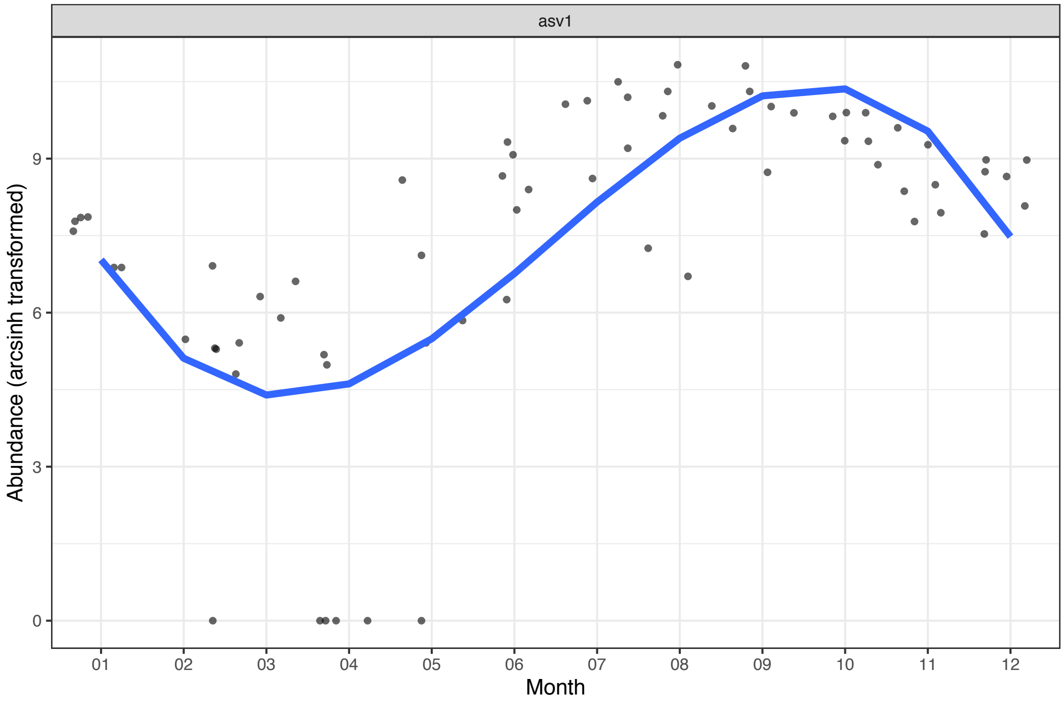
We will compare two approximations to seasonality.

Lomb scargle periodogram

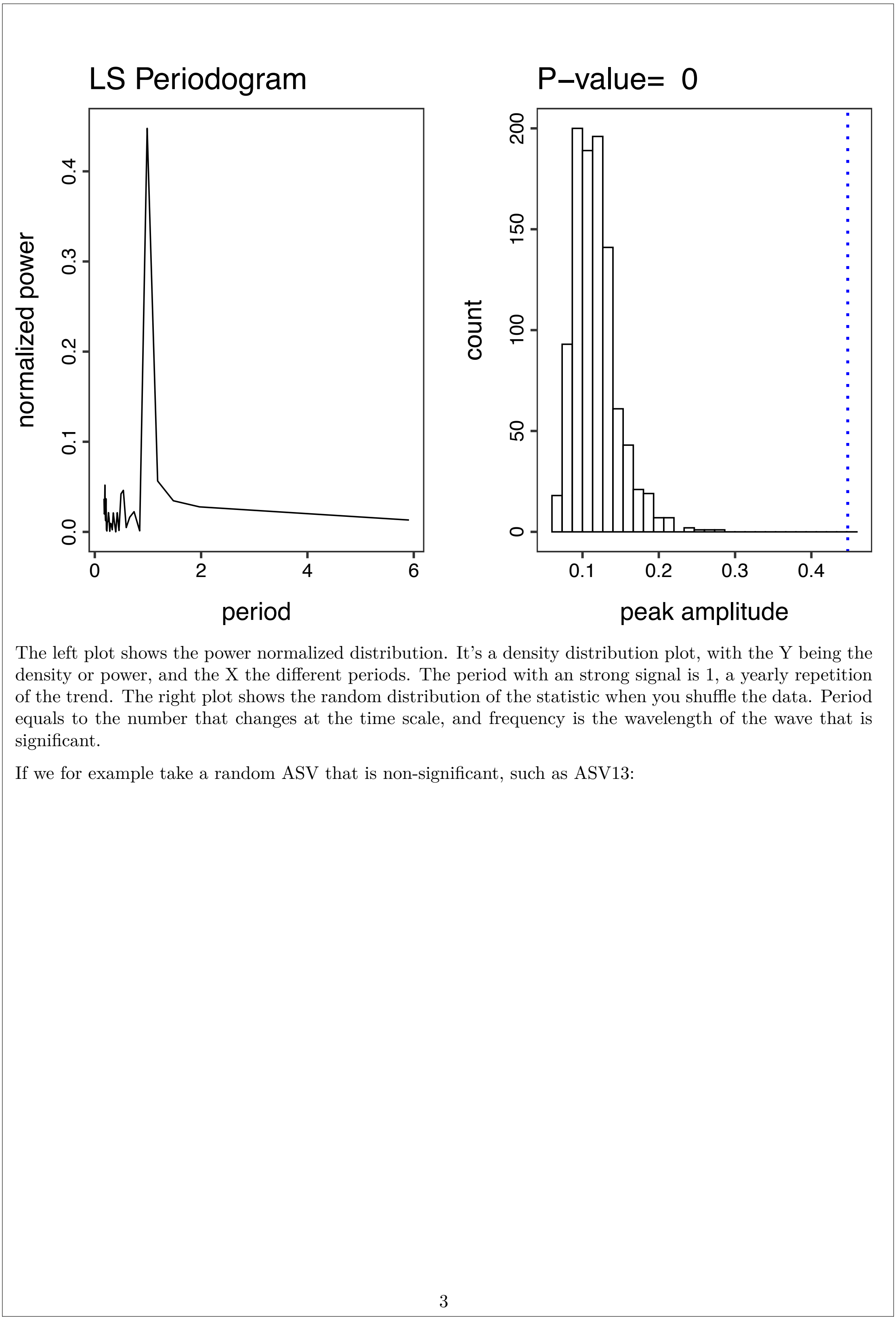
For a time series, you use the different possible wavelengths that could originate the signal, and test if some of them appear to be occurring more often than by random. The occurrence of each wave gives a density signal, that in the physics jargon, its the *power*. We want the *power normalized maxima*, the maximum value of the period presenting a significant trend.

Therefore, for each ASV you break down the wave in the different frequencies that could we forming the trend, and calculate which ones appear.

An example, with ASV1:

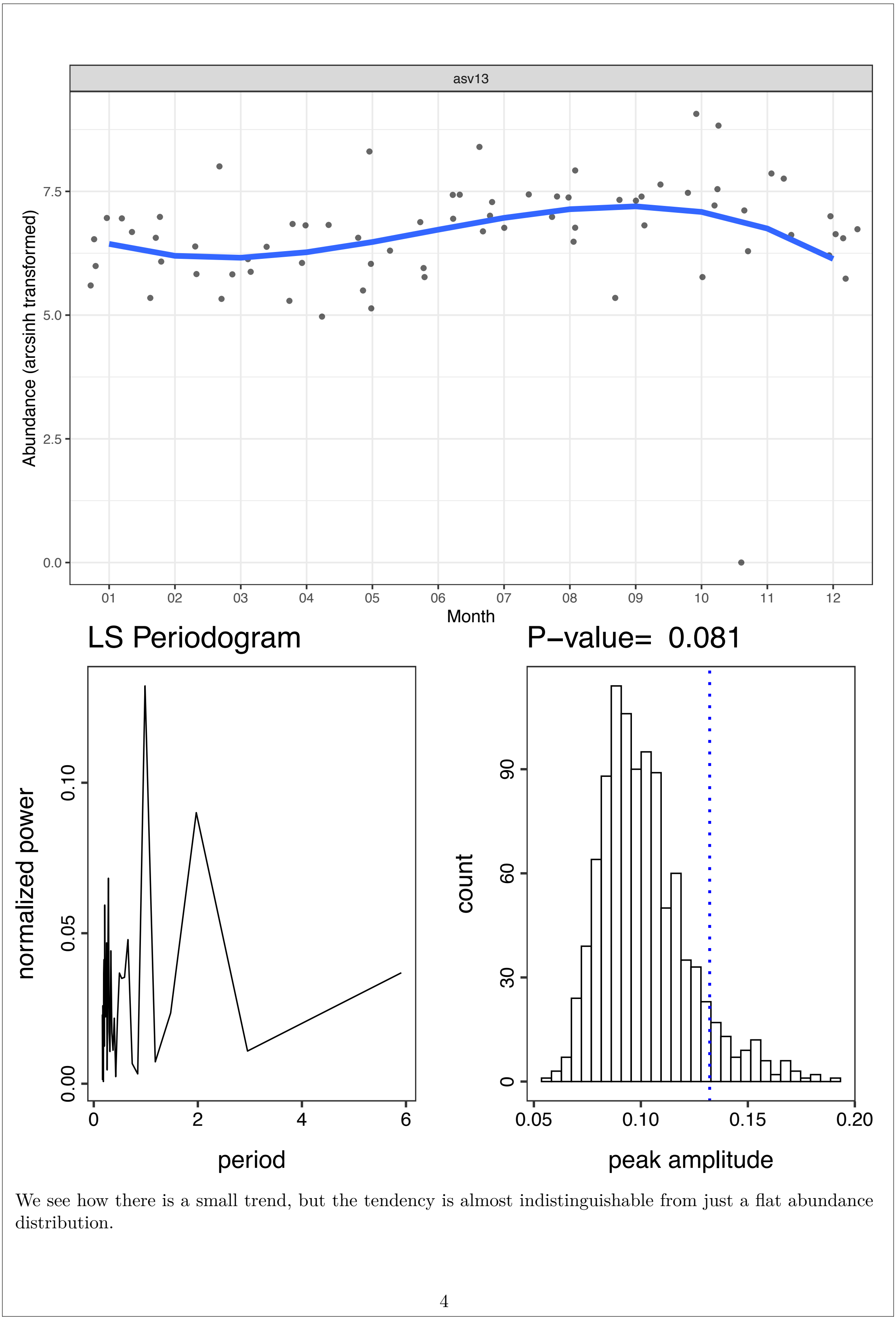


We see clearly a trend, and the blue line doing a good fitting. Good! Let's use the random lomb scargle method:



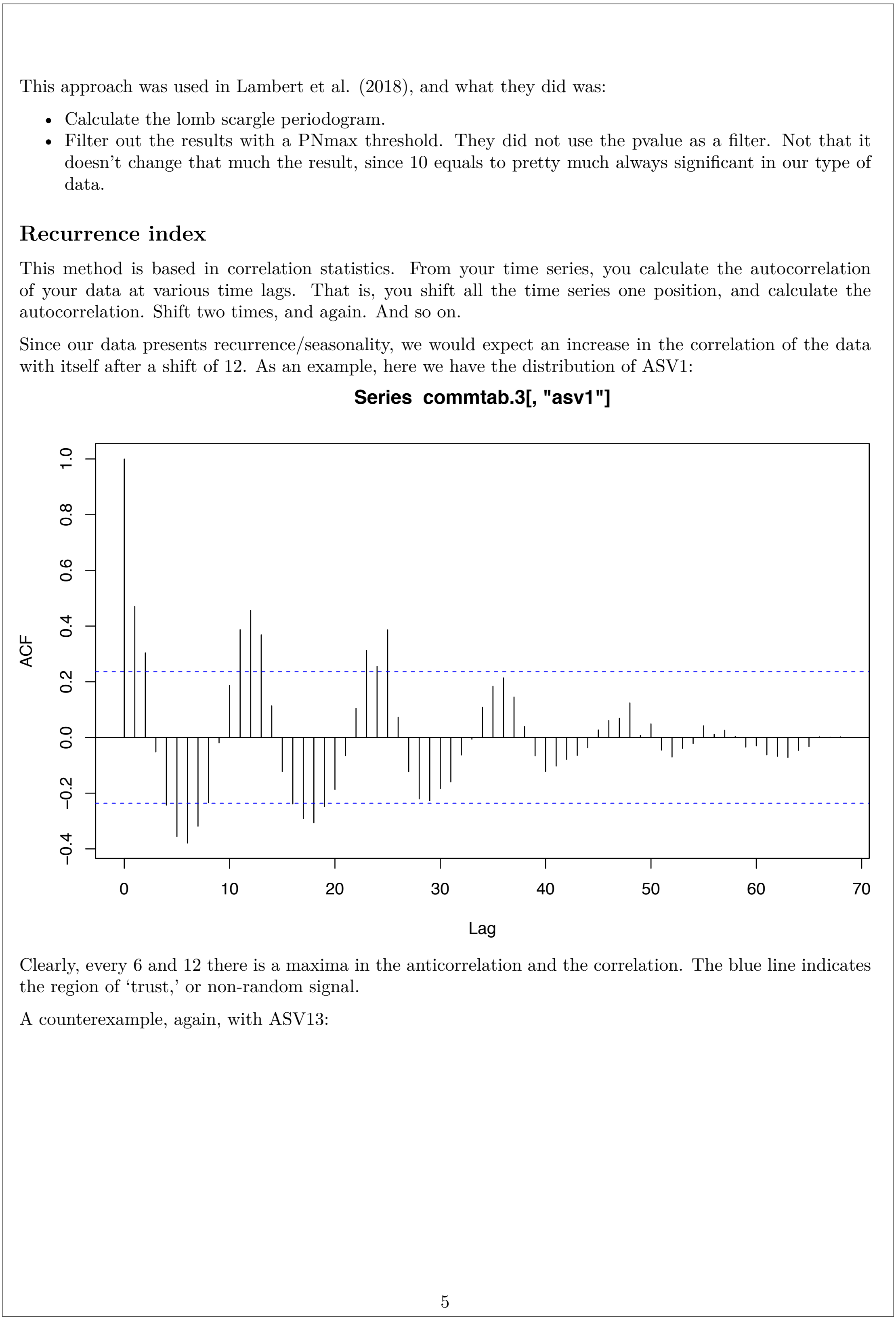
The left plot shows the power normalized distribution. It's a density distribution plot, with the Y being the density or power, and the X the different periods. The period with an strong signal is 1, a yearly repetition of the trend. The right plot shows the random distribution of the statistic when you shuffle the data. Period equals to the number that changes at the time scale, and frequency is the wavelength of the wave that is significant.

If we for example take a random ASV that is non-significant, such as ASV13:



We see how there is a small trend, but the tendency is almost indistinguishable from just a flat abundance distribution.

4



This approach was used in Lambert et al. (2018), and what they did was:

- Calculate the lomb scargle periodogram.
- Filter out the results with a PNmax threshold. They did not use the pvalue as a filter. Not that it doesn't change that much the result, since 10 equals to pretty much always significant in our type of data.

Recurrence index

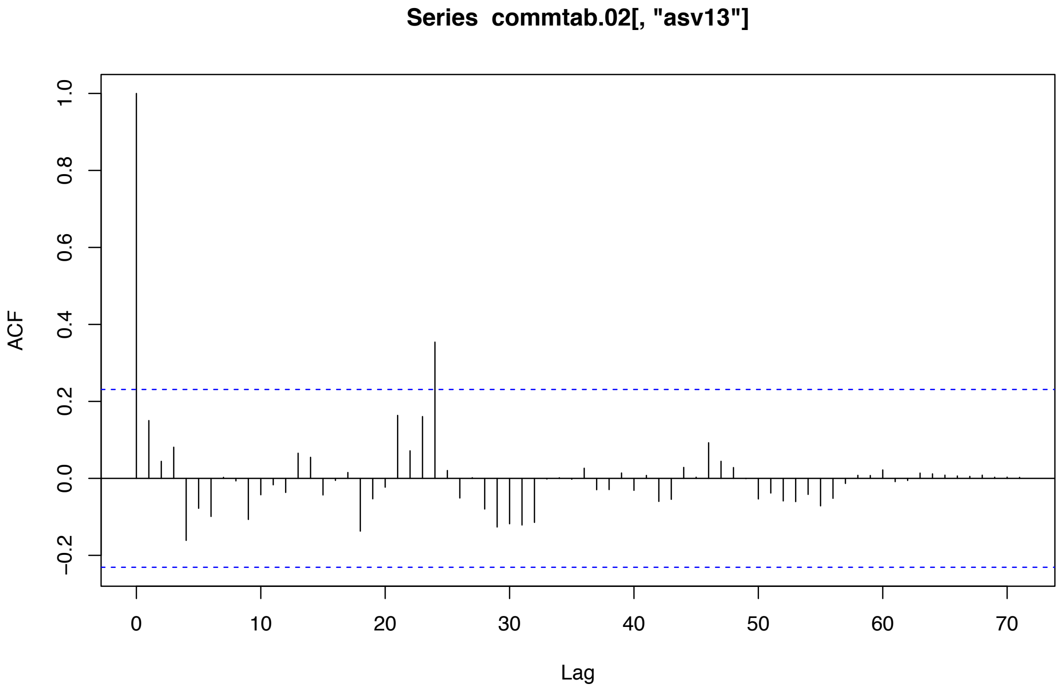
This method is based in correlation statistics. From your time series, you calculate the autocorrelation of your data at various time lags. That is, you shift all the time series one position, and calculate the autocorrelation. Shift two times, and again. And so on.

Since our data presents recurrence/seasonality, we would expect an increase in the correlation of the data with itself after a shift of 12. As an example, here we have the distribution of ASV1:

Series commtab.3[, "asv1"]

Clearly, every 6 and 12 there is a maxima in the anticorrelation and the correlation. The blue line indicates the region of 'trust,' or non-random signal.

A counterexample, again, with ASV13:

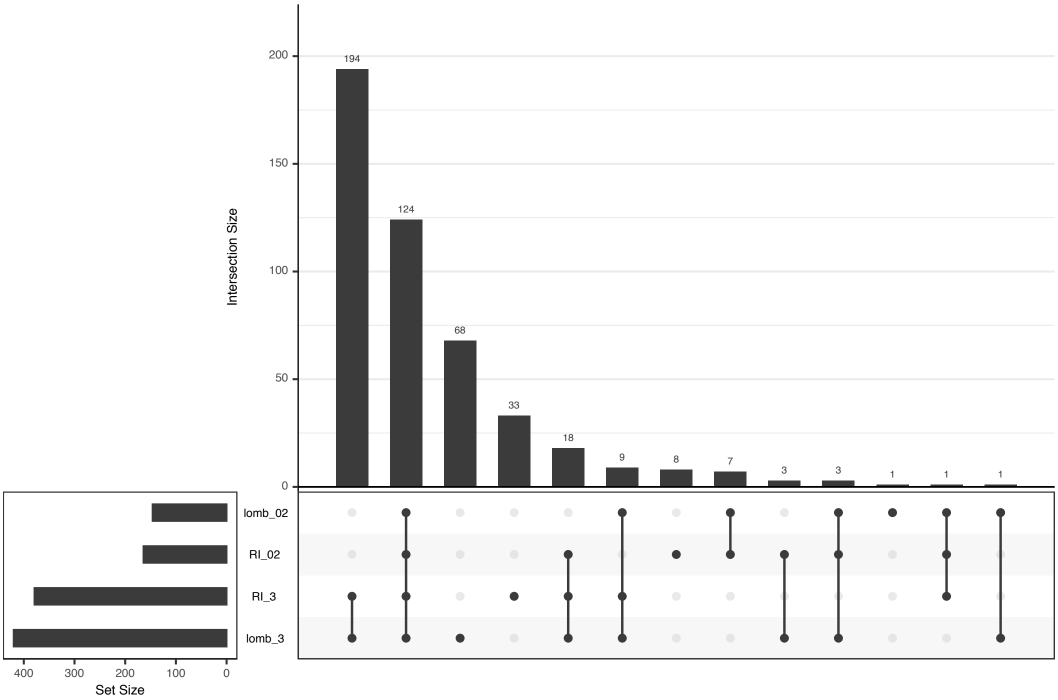


Giner et al. (2019) proposed a method to quantify seasonality using this method, performing the following:

- Sum the absolute number of all the ACF function, obtaining a recurrence value. The value is RF.
- Shuffle the sample values for each ASV, and calculate again many times the statistic RF(random). Obtain a median and confidence intervals of this median (usually 95%).
- Check if the value obtained initially, is inside the RF(random) region. If it's inside, the null hypothesis (the value could come from a random distribution) is confirmed. If not, we confirm that the alternative hypothesis, a seasonal signal, is the situation we have.
- Finally, in the paper the authors also applied another filter, in a similar fashion to filtering for a PNMax of 10. By dividing RF / RF(rand), if the threshold is above 1.15 was considered significant. They applied 1.2 for another group, but we will take into consideration the most generous application in this case.

Comparing both statistical tests (RAW)

First we will check the similarities between the models applying only the *adhoc* filters (non significance without filtering for the effect size, or how strong is the recurrence in this case.



This plot shows the number of values in each set (left barplot) and the intersections between the sets (bubble and line plot and the bars on top). We observe that the test obtaining the most results is the lomb scargle for the fraction 3, followed by the recurrence index, and then the fraction 2.

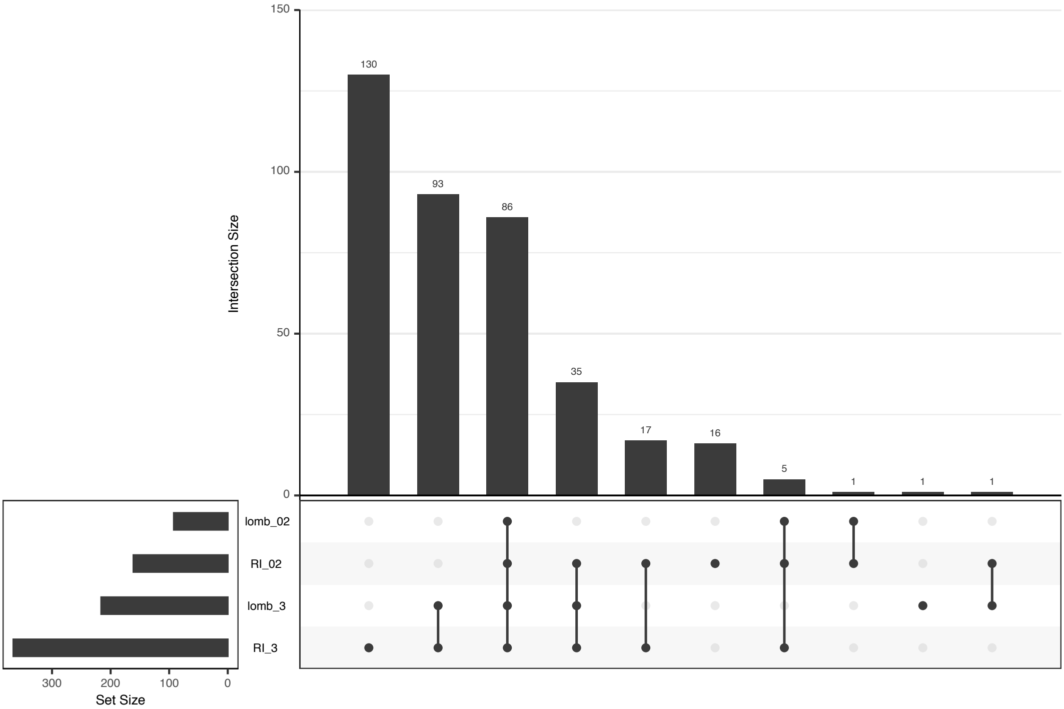
As highlights of what we see is:

- Fraction 3 presents 194 ASVs considered seasonal by both methods.
- Lomb scargle presents 68 values that only appear in this method for fraction 3.
- There are 124 shared seasonal ASVs, independently of the method or fraction.
- The recurrence index presents 33 ASVs that are only seasonal for this method. For the 2 fraction, this value is 8.

It's not surprising that the methods present some slight differences, but as we can see these differences are not trivial at all, and sometimes can change some of the conclusions we could obtain.

Comparing methods with final filters

Now let's see how the results compare when we apply the filters that both authors propose. That is, a PNmax >= 10 and RI >= 1.15 minimum.



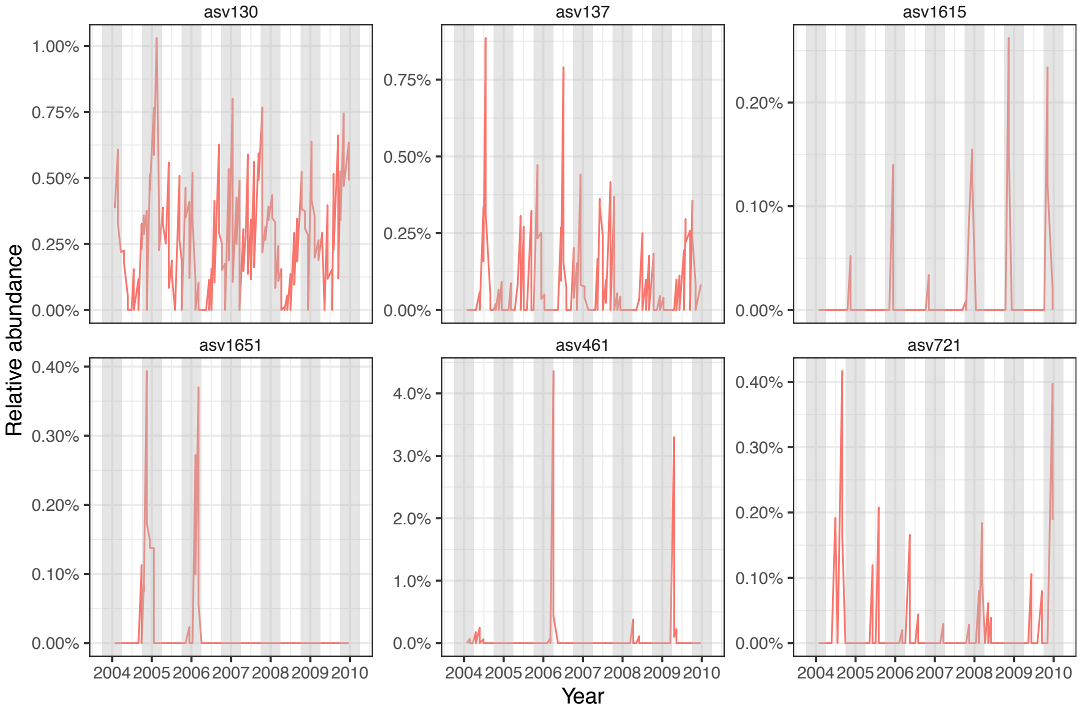
Some things have changed now:

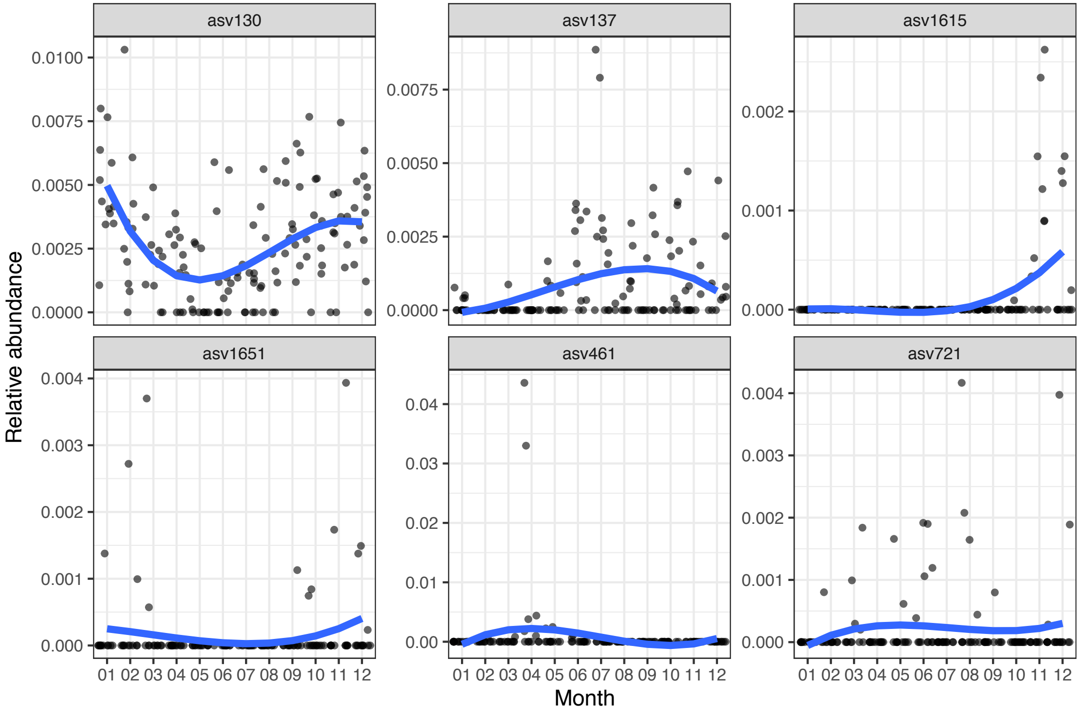
- The recurrence index is the statistic with most of the results, followed by lomb scargle (both fraction 3).
- Regrading unique results, now it's RI with 130 seasonal ASVs the one on top, followed by RI with 16 ASVs. This result made me stick with the lomb scargle, given that is more conservative to what we consider as seasonal. Initially, in my first approximation to seasonality (Auladell et al. (2019)), we used fourier approaches, and these two papers were published in the same year. To switch to one method or the other, I used this comparison.

Let's try to see what is happening here with some of these unique results for each method.

Checking what happens with unique results

Results only present in recurrence index fraction 3

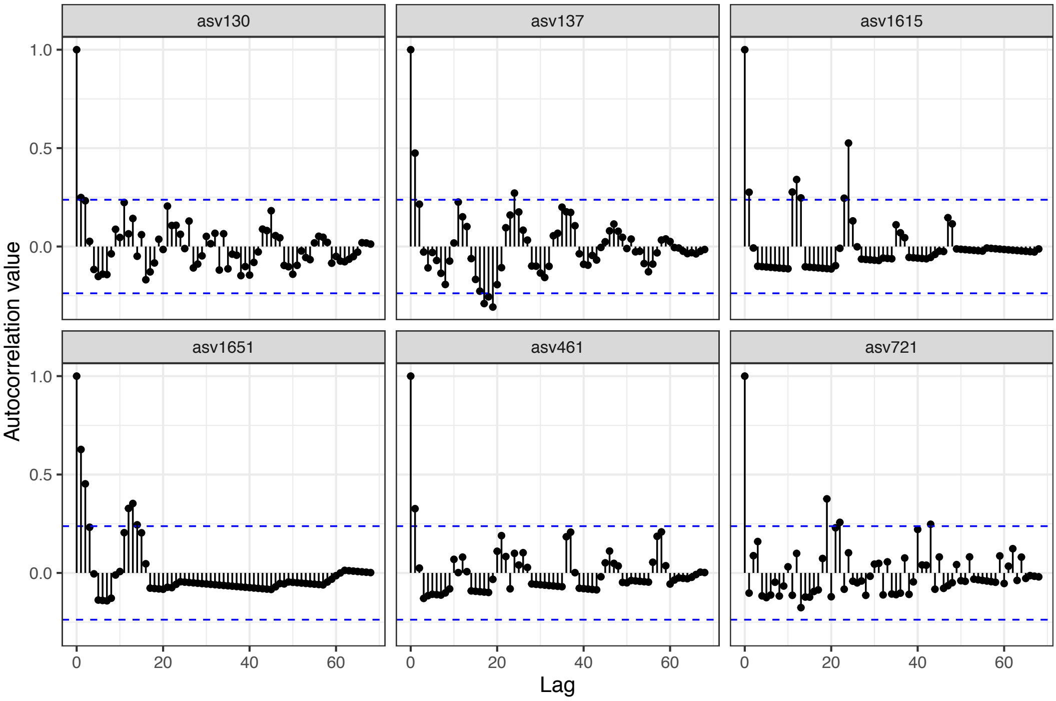




Most of the results have in common: - Either having a small read count abundance. - Having a lot of dispersion in the relative abundance.

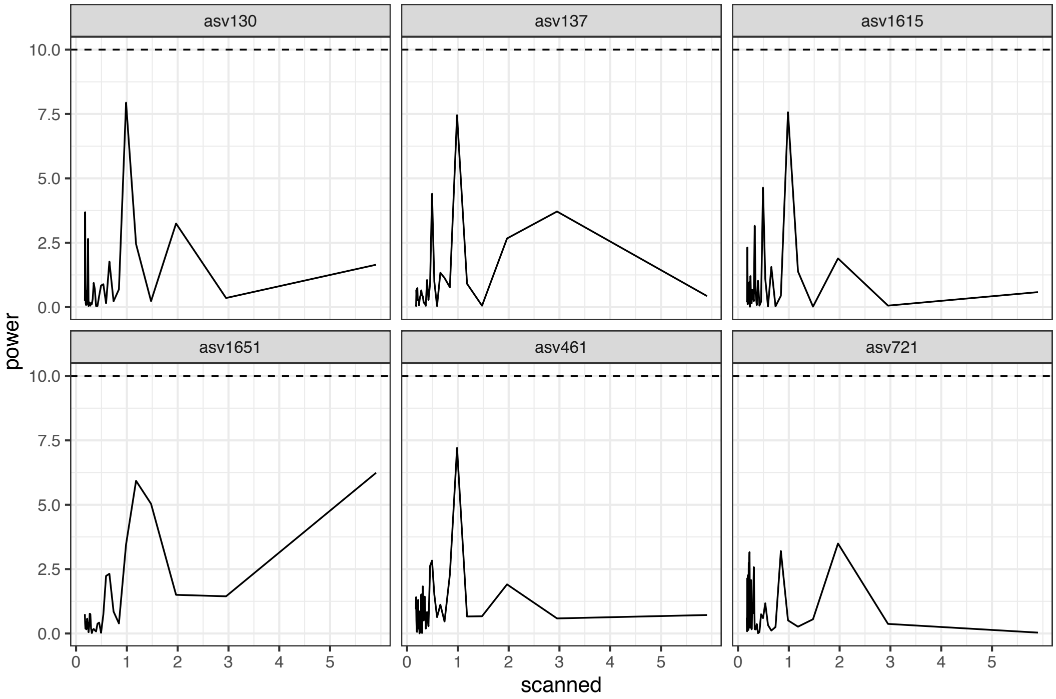
Watch out! We used ASV13 as an example of no seasonality, but this plots are coming from fraction 3, which may give in some cases different results than the fraction 2.

Let's look at the periodograms and the autocorrelation plots



- ASV130 presents a seasonal structure in the ACF, but with really small values too.
- ASV1615 is present only from time to time. There is autocorrelation, but the values are very low, and intermittent (because most of the distribution is basically 0).
- ASV721 clearly is not seasonal and therefore we could consider the result a false positive.

If we check the periodograms:



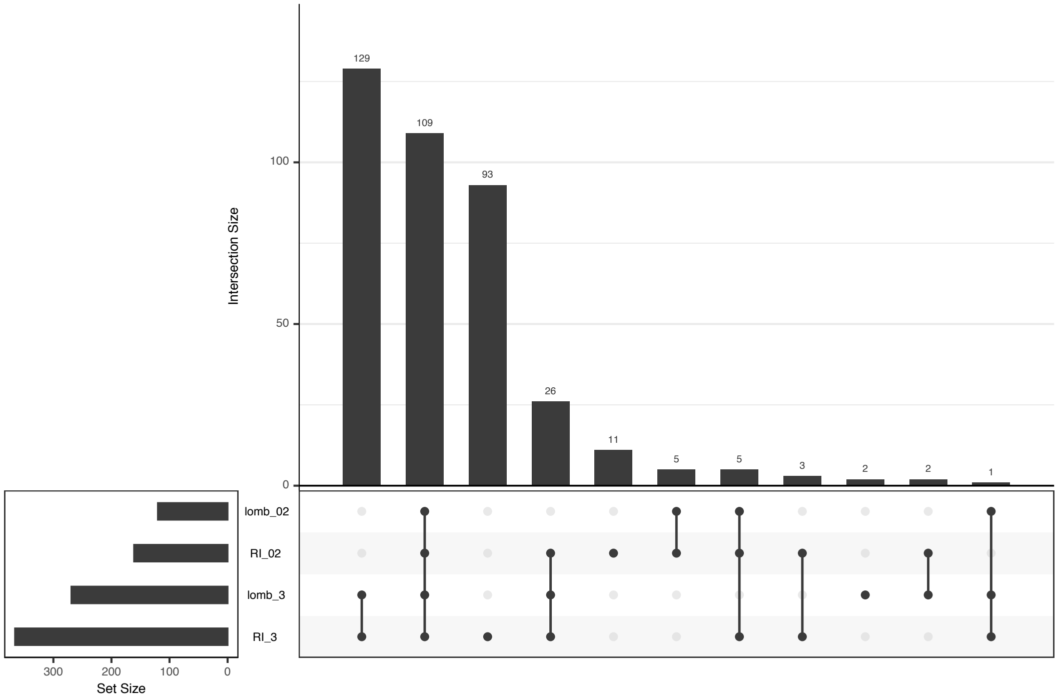
Conclusions

As a general conclusion: **with the RI method, we are gaining sensitivity at the cost of the specificity, since there are more false positives**.

In this random selection we have seen some values that we could believe as 'seasonal,' amid a really small signal, and some results that we clearly see that are not seasonal and therefore that the method, even with the filters applied in its manuscript are too permissive sometimes.

This is sometimes a property of the model you have chosen. Since the ACF can be big at some random spikes, and then the random model not present this spikes, sometimes the result could be significant without coming from a seasonal distribution. Obviously, since these models have to be quite general, sometimes you have to accept is as it is and go on with your life.

Another consideration we can do is that the use of 10 as a filter for PNMax is not a value written in stone. We could consider necessary a compromise and go down with the value! Let's try to see the distribution when we use 8.5:



- Now the biggest intersect is between lomb and RI for fraction 3.
- lomb applied to fraction 3 doesn't give too many results unique to the method (only 1). Recurrence index still gives too many to use it comfortably for me..
- And in total we get quite an increase in significant results. We increase the number of seasonal ASVs from 308 to 379, a 23% of increase.

Therefore another important conclusion: **maybe we can allow ourselves to be less restrictive with the lomb scargle method, since we have the knowledge and another method to compare the results**

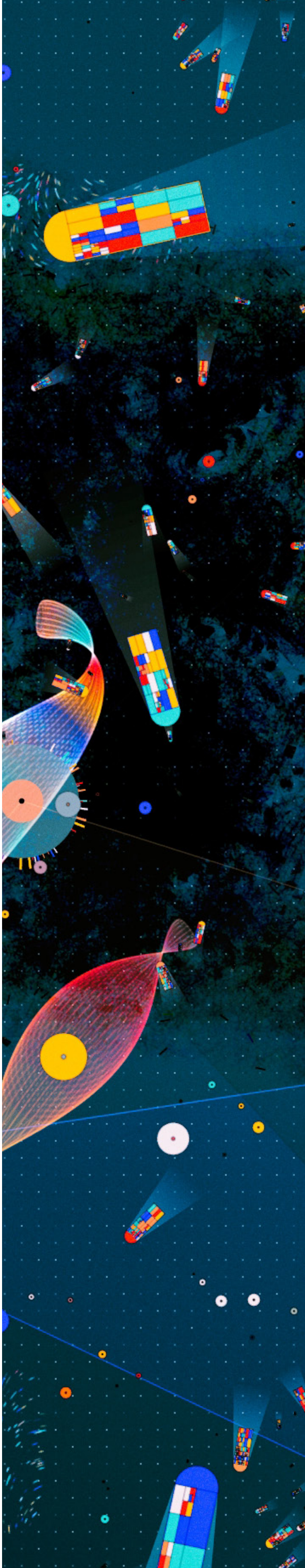
References

Auladell, Adrià, Pablo Sánchez, Olga Sánchez, Josep M. Gasol, and Isabel Ferrera. 2019. "Long-Term Seasonal and Interannual Variability of Marine Aerobic Anoxygenic Photoheterotrophic Bacteria." *The ISME Journal* 13 (8): 1975–87. https://doi.org/10.1038/s41396-019-0401-4.

Giner, Caterina R., Vanessa Balagué, Anders K. Krabberød, Isabel Ferrera, Albert Reñé, Esther Garcés, Josep M. Gasol, Ramiro Logares, and Ramon Massana. 2019. "Quantifying Long-Term Recurrence in Planktonic Microbial Eukaryotes." *Molecular Ecology* 28 (5): 923–35. https://doi.org/10.1111/mec.14929.

Lambert, Stefan, Margot Tragin, Jean-Claude Lozano, Jean-François Ghiglione, Daniel Vaulot, François-Yves Bouget, and Pierre E. Galand. 2018. "Rhythmicity of Coastal Marine Picoeukaryotes, Bacteria and Archaea Despite Irregular Environmental Perturbations." *The ISME Journal*, September. https://doi.org/10.1038/s41396-018-0281-z.

CHAPTER IV



Chapter IV

Seasonal influence of predation, viral mortality, light and nutrient limitation on a marine microbiome assessed through metagenome assembled genomes

Adrià Auladell, Isabel Ferrera, Olga Sánchez, Josep M Gasol

A manuscript in preparation

Abstract

Marine microbial communities are assembled through a complex array of interactions between their constituting members and the changing environment, with bottom-up and top-down factors exerting a strong selection that may vary seasonally. The influence of these factors has been assessed through fluorescent in situ hybridization and amplicon sequencing approaches, but how individual species respond to them and how selection operates at the strain level remains largely unknown. We experimentally manipulated seawater from a coastal marine community at different seasons by modifying the effect of predators, viruses, nutrient limitation and light availability, and assessed the growth of species using metagenome assembled genomes (MAGs). Overall, we recovered 262 MAGs mainly from the Rhodobacterales, Flavobacteriales and Alteromonadales classes. Season and treatment greatly influenced community composition, with 26% of the MAGs being indicative of the control treatments, 24% of both the control and predator-reduced treatments, 12.8% indicators of both the virus-reduced and the diluted treatments, and 7.3% of the predator-reduced treatment only. *Flavobacteriaceae* MAGs grew mostly in the predator-reduced treatment with distinct MAG-defined species in each season, whereas *Alteromonadaceae* and *Sphingomonadaceae* taxa developed preferably in the virus-reduced and diluted treatments indistinctively of season. The presence of specific functional groups, such as photoheterotrophs, was influenced by treatment and by whether the organism had typical oligotrophic or copiotrophic genomic properties (i.e. depending on genome size and codon usage). Strain delineation indicated that, generally in these experiments, one clonal strain dominated, suggesting that strain selection is driven mostly by competitive exclusion.

4.1 Introduction

Marine bacteria and archaea drive ocean biogeochemical cycles through their unique metabolic repertoire (Falkowski *et al.*, 2008). These communities are also highly diverse, harboring thousands of taxa, and presenting a complex array of interactions in which the coexisting individuals simultaneously compete and cooperate to sustain their life in the marine environment (Bergelson *et al.*, 2021). Given this large diversity, high throughput technologies have been key to disentangle how natural microbial communities are assembled.

The main community assembly processes –selection, drift, dispersal and speciation– act in a combined fashion in any biological system (Vellend, 2010). Disentangling which one is more important in a given ecosystem is one of the key questions in microbial ecology (Langenheder and Lindström, 2019). Amplification of 16S rRNA gene hypervariable regions has allowed tracking and unraveling these processes in the natural habitat (Zinger *et al.*, 2011; Logares *et al.*, 2020). Yet, the maximum taxonomic resolution of amplicon tagging approaches has been a matter of discussion since their introduction (Acinas *et al.*, 2004; Johnson *et al.*, 2019; Schloss, 2021). The hypervariable regions of the 16S rRNA gene present a resolution at the species/genus range, given that for many species these regions are identical (Johnson *et al.*, 2019; VanInsberghe *et al.*, 2020). This limited level of resolution coupled with the lack of knowledge on how the functional traits are distributed within species might mask important factors shaping taxa selection (Salazar and Sunagawa, 2017). Metagenomic approaches have in part overcomed these limitations since they allow obtaining metagenome assembled genomes (MAGs) from which we can inspect the specific functional repertoire of the dominant species in the environment (Rodríguez-Valera, 2002; Grossart *et al.*, 2020). To date, several studies described species distributions based on MAG reconstructions (Delmont and Eren, 2018; Graham *et al.*, 2018; Acinas *et al.*, 2021). As an example, they showed that the spatial distribution of the of the Pelagibacterales order was driven by selection even to single aminoacid variants in key genes (Delmont *et al.*, 2019). Metagenomics can thus be helpful to understand how selection by different ecological factors operates at strain and species levels. Using the Curtobacterium genus as example, McLaren and Callahan (2018) showed that different strains and species presented a specific genetic repertoire for humic compound degradation, yet the genus as a whole presented a similar response to one particular environmental factor, drought. These results suggest that phylogenetic trait conservation is variable between taxa when there are strong adaptative drivers of selection. How selection operates at the species level remains however a challenging problem for which metagenomics could provide some insights.

Annual changes in abiotic parameters in the coastal surface temperate ocean, mainly light and temperature, generate a marked seasonality in bacterial and archaeal community structures (Bunse and Pinhassi, 2017; Lambert *et al.*, 2018; Lemonnier *et al.*, 2020; Auladell *et al.*, 2021). In the coastal NW Mediterranean (e.g. Blanes Bay, Gasol *et al.* 2016), water mixing during fall and early winter substantially increases nutrient concentrations, in late winter and spring there is an enhanced growth of phytoplankton, and summer conditions are characterized by warm and commonly nutrient-deficient waters, which select for groups able to sustain growth under these conditions (Auladell *et al.*, 2021). Likewise, in this microbial observatory, predators and viruses present seasonality in bulk abundance and diversity (Unrein *et al.*, 2007; Boras *et al.*, 2009; Giner *et al.*, 2019). It has been shown that the growth of most microbial groups is generally regulated by predation, viruses and nutrient limitation (Ferrera *et al.*, 2011; Kirchman, 2016; Sánchez *et al.*, 2017). Taxonomic groups such as Alteromonadales, Bacteroidetes and Rhodobacterales show an increase in growth rates when these factors are manipulated (Yokokawa *et al.*, 2004; Ferrera *et al.*, 2011; Kirchman, 2016; Sánchez *et al.*, 2017). Similarly, functional groups such as the aerobic anoxygenic phototrophic bacteria are also tightly regulated by predation and nutrients (Koblízek *et al.*, 2007; Ferrera *et al.*, 2011), while light can also enhance their growth rates in nature (Ferrera *et al.*, 2017). A recent study in Blanes Bay showed that distinct seasons have different patterns of microbial growth: during winter the community presented the highest growth rate, and light had a greater influence during spring and summer (Sánchez *et al.*, 2020). However, most of these analyses were performed at the whole group (order, class) level, without differentiating the patterns of the different species within the group. While in recent years some studies have determined the effects of these ecological factors on operational taxonomic units (OTUs, Teira *et al.*, 2019) and amplicon sequence variants (ASVs, Fecskeová *et al.*, 2021), the link between taxa selection and the genomic repertoire of the selected taxa is largely unexplored. Another unresolved question is the phylogenetic level at which the various community assembly processes act. A study in soil and plant-associated microbiomes showed that community assembly converged at the family level, with genus, species or strain assembly being more random than at the family level (Goldford *et al.*, 2018). Contrarily, another study found that the presence of specific strains, not that of species, determined how the community was assembled, indicating that ecological dynamics at the strain level are relevant (Goyal *et al.*, 2021). Experimental manipulations of environmental conditions in combination with metagenomics could help to assess the influence of the environment on community taxonomy and function and, ultimately, to better understand these processes in the natural habitat.

Here, we report on a series of manipulation experiments in different seasons for which we obtained metagenomic assembled genomes (MAGs) in order to determine how the community was modulated by different top-down and bottom-up factors. In particular, we assessed the influence of predation removal, virus reduction, nutrient availability and light on the growth of microorganisms. Three experiments performed in contrasting seasons with different initial communities allowed us to test whether the communities converged between treatments. Through functional genomic annotation, we established links between the genetic repertoire of the selected organisms and the treatments, and assessed if these were maintained at different seasons. Finally, we also used population genomic techniques to explore strain diversity in each experiment to test if a single or rather various ecotypes can dominate under specific conditions.

4.2 Materials and Methods

Sample collection and environmental data

Samples were collected from the Blanes Bay Microbial Observatory (BBMO), a shallow (~20 m) coastal station located ~1 km offshore in the North Western Mediterranean Sea (41°40'N, 2°48'E), for which seasonal changes in environmental parameters have been extensively characterized (e.g. Gasol *et al.*, 2016). Three experiments were conducted using surface water collected on 21 February 2017 (winter), 26 April 2017 (spring) and 5 July 2017 (summer). Seawater was sieved through a 200-µm mesh and transported to the laboratory within 2 h. All the following measurements were performed in situ: water temperature and salinity were measured with a CTD (acronym for conductivity, temperature, and depth) SAIV SD204 probe; photosynthetically active radiation (PAR) at the sampling station was measured with a multichannel filter radiometer (PUV-2500; Biospherical Instruments Inc.), and light penetration was estimated using a Secchi disk. The concentration of inorganic nutrients was determined spectrophotometrically using an Alliance Evolution II autoanalyzer following standard procedures (Grasshoff *et al.*, 1983). Chlorophyll *a* (Chl a) concentration was measured from acetone extracts by fluorometry. Abundances of heterotrophic bacteria, photosynthetic picophytoplankton and viruses at the sampling sites were measured by flow cytometry with a FACSCalibur (BectonDickinson) flow cytometer (Gasol and Morán, 2016). Heterotrophic nanoflagellates (HNF) were filtered onto polycarbonate 0.6-µm filters and stained with 4', 6-diamidino-2-phenylin- dole (DAPI, final concentration 1 µg·mL^{-1}), and counted in an Olympus BX61 epifluorescence microscope (Porter and Feig, 1980). Further details can be found in Sánchez *et al.* (2020).

Experimental design

The same experimental design was used at each season; we exposed the collected seawater to six different treatments: CT (control), experiment with unfiltered seawater, both in natural light/dark cycles and in continuous dark (CL, control light, and CD, control dark treatments); PR (predator-reduced), seawater prefiltered with a 1-µm filter to remove predators while keeping most bacteria, both in natural light/dark cycles and in continuous dark (PL, predator-reduced light, and PD, predator-reduced dark treatments); DL (diluted light), a 1:4 dilution of whole seawater with 0.2-µm-filtered seawater to reduce both predation and competition for nutrient and carbon resources among bacteria exposed to natural light/dark cycles; VL (virus-reduced light), a 1:4 dilution of whole seawater with seawater filtered through a 30-kDa VivaFlow cartridge to reduce predation, viruses and resource competition, exposed to natural light/dark cycles. Samples were subjected to these manipulations and kept at in situ temperature until the start of the experiment (~20 h from sampling). Water was distributed into 9-L Nalgene bottles, which were incubated in triplicate for 1.5-2 days in large water baths (200 L) with circulating seawater to maintain the temperature close to in situ conditions. The light treatments were limited to photosynthetically active radiation (PAR) by maintaining the bottle incubations under natural light conditions with the exclusion of UV radiation, using two layers of an Ultraphan URUV Farblos Filter and a net that reduced light intensity to roughly mimic the light conditions of a water depth of 3 m, calculated from the transparency measures at sampling site. We monitored PAR continuously in the incubation water baths. The dark treatments bottles were completely covered with two layers of black plastic bags to prevent light exposure.

DNA extraction, sequencing and quality control

Samples were sieved through a 20 μm mesh to remove large particles and microbial biomass was concentrated onto 0.2 μm polycarbonate filters using a peristaltic pump. Large volume samples (~2-4 L) were filtered from each replicate of all treatments at the beginning and at the end of the experiment (36 h after the start in summer and winter, 48 h in spring). We extracted the DNA from the filters as described in Massana *et al.* (1997), purified and concentrated using Amicon 100 columns (Millipore) and quantified in a NanoDrop-1000 spectrophotometer (Thermo Scientific). We stored the DNA at −80°C and an aliquot from each sample was used for sequencing using a Novaseq6000 machine (Centre Nacional d'Anàlisi Genòmica, CNAG) with paired-end fragments of 150 bp. A total of 66 samples were sequenced with an average 115 million reads (min = 67M, max = 238M) each. The winter and summer experiments presented 2 replicates for the final times, whereas the spring experiment presented 3 replicates. We used illumina-utils (Eren *et al.*, 2013) for quality filtering the short reads from the metagenomes with the *iu-filter-quality-minoche* function (default parameters), which removes noisy reads following the method described in Minoche *et al.* (2011).

Metagenome assembled genome generation

We assembled each sample independently using MEGAHIT v1.2.8 (Li *et al.*, 2015) to obtain contigs. We recruited the short reads from each sample to the correspondent contigs using Bowtie2 v2.4.3 (Langmead and Salzberg, 2012), and used samtools v1.12 (Li *et al.*, 2009) to sort the output SAM files into BAM files. We used METABAT v2.12 (Kang *et al.*, 2019) to obtain a set of putative bins (minimum contig length = 1500 bp) based on the sequencing depth of each contig and its tetranucleotide frequency. Since each set of bins (uncurated genomes) came from one sample, many of them were identical. Thus, we used dRep v3.2.2 (Olm *et al.*, 2017) to dereplicate these bins (average nucleotide identity, ANI = 95%) into a single set of raw bins representative of all the samples. The 95% ANI threshold differentiates between bins at the species level, obtaining a single dataset for the whole study (Olm *et al.*, 2020). From this bin set, we used the contigs workflow implemented in anvi'o v7 (Eren *et al.*, 2015) based in snakemake (Mölder *et al.*, 2021), which (1) identifies open reading frames using Prodigal v2.60 (Hyatt *et al.*, 2010), (2) identifies single copy core genes using HMMER v3.2.1 (Eddy, 2011) and a collection of built-in HMM profiles for bacteria and archaea, (3) establishes a taxonomy based in these single copy genes based in the Genome Taxonomy Database (GTDB), and (4) maps again all the samples using Bowtie2 and samtools for BAM generation. For all analyses we used those bins presenting ≥ 70% completion and ≤ 10% contamination. We also analyzed all the bins presenting ≥ 40% completion since many bacterial groups can be difficult to recover due to high microdiversity. These last bins were not included in the subsequent functionality analysis since the low completion would impact the analysis based on gene presence. We checked these bins manually to refine them and remove possible incorrect binning, obtaining metagenome assembled genomes (MAGs). Specifically, through the anvi'o interface we displayed the contigs forming a MAG using multiple information (sequence composition, differential coverage for the 66 samples, taxonomic annotation) and curated the undesired contigs. The final dataset consisted of 262 MAGs, from which 175 presented a completion above 70% and 87 MAGs a completion between 70% and 40%. We recruited reads from all the samples to the 262 MAGs using Bowtie2 and samtools as explained above.

Functional analysis

We only annotated the 175 high-quality MAGs (≥ 70% completion). We used prokka v1.13 (Seemann, 2014) to obtain the coding DNA sequences (CDS) through prodigal and annotated the MAGs using the NCBI's Clusters of Orthologous Groups (COG, Tatusov *et al.*, 2003). We also annotated the genes using the KEGG database based in hidden markov models (KOFAM, Aramaki *et al.*, 2020). From all possible genes, we focused in detail on a set of key marker genes with relevance in the marine biogeochemical cycles, selected in a previous study (Chapter 3 thesis) and a recent metatranscriptome analysis (Alonso-Sáez *et al.*, 2020). Additionally, we obtained the KEGG module completion through the *anvi-estimate-metabolism* function (Eren *et al.*, 2015). The KEGG modules are complete metabolic pathways.

mOTU generation

To explore patterns of richness and dissimilarity between samples, we used the mOTUs2 v3 pipeline that allows obtaining species profiles of each sample and compute the dissimilarity between samples and species richness (Milanese *et al.*, 2019). Briefly, the method maps all the reads to a reference database based in the genome taxonomy database using a set of single copy genes (Parks *et al.*, 2018). Through the recovered counts, it calculates the relative abundance of each taxonomic group and determines the unassigned fraction.

Maximum growth rate prediction

We estimated the maximal growth rate of each MAG through the gRodon package (Weissman *et al.*, 2021). The method estimates maximal growth rates of prokaryotic organisms from genome-wide codon usage statistics. We used these maximal growth rates to differentiate oligotrophic and copiotrophic bacteria using the predicted minimal doubling time of 5 hours as the threshold of separation (as in Weissman *et al.*, 2021). The definition of these lifestyles is evolutionary, with an oligotroph being an organism for which selection for rapid maximal growth is weak enough so that translation efficiency is not optimized by selection on codon usage. This definition contrasts to the most common one for these terms that is based on resource use and specific growth rate (discussed in Giovannoni *et al.*, 2014).

Strain delineation

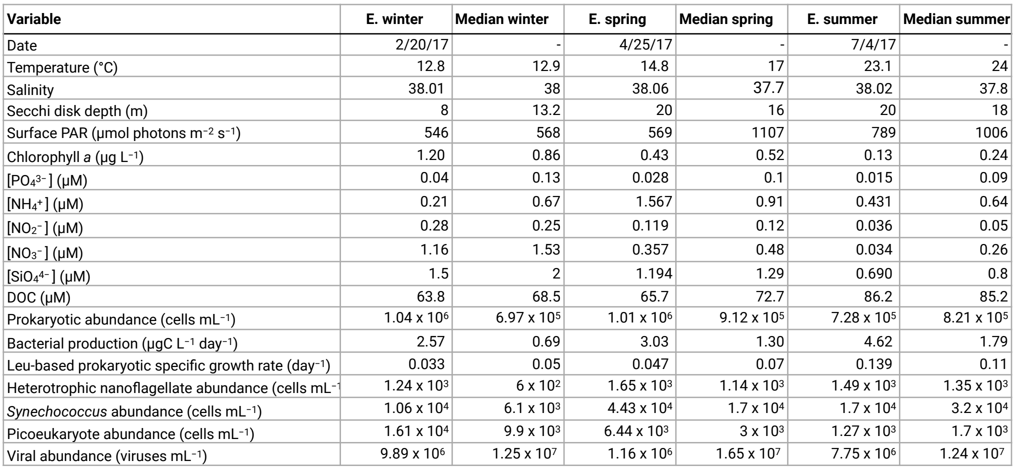
By dereplicating the genomes at 95% average nucleotide identity (ANI) –a proxy for species (Olm *et al.*, 2017)– each of the MAGs could represent the genome of a population composed by different strains. We used InStrain v1.3.1 (Olm *et al.*, 2021) to identify the strains present within a MAG. Briefly, InStrain measures the genetic heterogeneity of a microbial population and performs comparisons between organisms in different samples. The program compares the BAM files with read information against the representative genome and generates microdiversity statistics. Specifically, we used here the consensus ANI and the population ANI. The first metric accounts for both major (all reads different than the reference) and minor (a significant number of reads do not match) allele mismatches. The population ANI is more restrictive; given multiple mismatches in a nucleotide position (for example, an 'A' in the reference), if that allele is actually present in the aligned reads (e.g. 'A' is in 40% of the reads), it is considered as intraspecific genetic variation and the base is not considered a substitution (Olm *et al.*, 2021). We applied strain analysis only to the 175 high-quality MAGs (≥ 70% completion).

Statistical analyses

We performed all analyses using the R v3.6.2 language (R Core Team, 2014). For data processing we used the tidyverse v1.3 package (Wickham *et al.*, 2019), and ggplot2 v3.3 for data visualization (Wickham, 2016). We performed a principal component analysis (PCA) of the mOTU distribution with the *prcomp* function from base R. We tested the significance of season, treatment and time as factors structuring the samples with a PERMANOVA, implemented in the *adonis* function of the vegan v2.5 package (Oksanen *et al.*, 2013). Treatment and time were tested considering the season as block strata to take into account the effect of the season factor, using 999 permutations.

We also determined the MAGs distribution among experiments. All the MAG based analyses rely on a detection above 0.5, meaning that at least 50% of the genome has to be covered 1X to consider the MAG as present. The coverage of a MAG divided by the mean coverage of all the MAGs in that sample was considered as the MAG's abundance measurement. These relative abundance values are ratios, with similar mathematical properties to proportions, but avoiding some typical problems of compositional data (see Gloor *et al.*, 2016 for an in-depth explanation). Additionally, for each MAG we calculated the ratio between the initial and final time abundance as fold change and thus, growth. For each MAG, we used the indicator species statistic coded in the indicspecies

R package to differentiate those MAGs indicative of treatment or season (permutations = 999, *p* ≤ 0.01) using the r.g statistic, a point biserial correlation coefficient which takes into account both presence and abundance information. For this statistic we only used the final times. Using this indicator statistic, we also computed whether some functions were enriched in specific treatments and/or seasons. Specifically, we calculated the enrichment in the virus-reduced/diluted, control and seasons of modules, functions that present the whole metabolic pathway in a genome using the *anvi-compute-functional-enrichment* function, which determines the module enrichment through a generalized linear model with a logit linkage function (*q* ≤ 0.01, Shaiber *et al.*, 2020). We considered both the control (light and dark) and the virus-reduced/diluted as a single group to simplify the analysis, since we observed similar patterns between these pairs.



Variable	E. winter	Median winter	E. spring	Median spring	E. summer	Median summer
Date	2/20/17	-	4/25/17	-	7/4/17	-
Temperature (°C)	12.8	12.9	14.8	17	23.1	24
Salinity	38.01	38	38.06	37.7	38.02	37.8
Secchi disk depth (m)	8	13.2	20	16	20	18
Surface PAR (µmol photons m^{-2} s^{-1})	546	568	569	1107	789	1006
Chlorophyll *a* (µg L^{-1})	1.20	0.86	0.43	0.52	0.13	0.24
[PO_4^{3-}] (µM)	0.04	0.13	0.028	0.1	0.015	0.09
[NH_4^+] (µM)	0.21	0.67	1.567	0.91	0.431	0.64
[NO_2^-] (µM)	0.28	0.25	0.119	0.12	0.036	0.05
[NO_3^-] (µM)	1.16	1.53	0.357	0.48	0.034	0.26
[SiO_4^{4-}] (µM)	1.5	2	1.194	1.29	0.690	0.8
DOC (µM)	63.8	68.5	65.7	72.7	86.2	85.2
Prokaryotic abundance (cells mL^{-1})	1.04×10^6	6.97×10^5	1.01×10^6	9.12×10^5	7.28×10^5	8.21×10^5
Bacterial production (µgC L^{-1} day^{-1})	2.57	0.69	3.03	1.30	4.62	1.79
Leu-based prokaryotic specific growth rate (day^{-1})	0.033	0.05	0.047	0.07	0.139	0.11
Heterotrophic nanoflagellate abundance (cells mL^{-1}	1.24×10^3	6×10^2	1.65×10^3	1.14×10^3	1.49×10^3	1.35×10^3
Synechococcus abundance (cells mL^{-1})	1.06×10^4	6.1×10^3	4.43×10^4	1.7×10^4	1.7×10^4	3.2×10^4
Picoeukaryote abundance (cells mL^{-1})	1.61×10^4	9.9×10^3	6.44×10^3	3×10^3	1.27×10^3	1.7×10^3
Viral abundance (viruses mL^{-1})	9.89×10^6	1.25×10^7	1.16×10^6	1.65×10^7	7.75×10^6	1.24×10^7

Table 1: Environmental data for the dates in which the experiments were performed compared to the median values for each season in the Blanes Bay Microbial Observatory. The columns starting with 'E.' denote the Experiment values, and the columns starting with 'Median' indicate the season median values in the observatory.

4.3 Results

The physicochemical and biological values at the start of the experiments were representative of the overall characteristics of the different seasons at the site (Table 1), and very contrasting among them. As we detailed in a previous study (Sánchez *et al.*, 2020), at the time of sampling chlorophyll *a* and inorganic nutrient concentrations (PO_4^{3-}, NH_4+, NO_{2-}^{-} NO_3^{-}, SiO_4^{4-}) were relatively high in winter, whereas ammonium was exceptionally high in spring. Eukaryotic picophytoplankton abundance was higher in winter than in the other seasons following the pattern of chlorophyll *a*, while *Synechococcus* dominated in spring. Regarding top-down variables, heterotrophic nanoflagellates were similarly abundant among seasons, and viral abundance was lower in spring than in the other seasons.

The prokaryotic community also presented variations. Alpha diversity differed among experiments and between the final and initial times of each experiment (Figure 1). Winter had the highest richness (mean mOTUs t_0 = 1464 ± 90), followed by spring (mean mOTUs t_0 = 1200 ± 110) and summer (mean mOTUs t_0 = 695 ± 51). The final times of the experiments were less rich in most cases, with the exception of spring, which had similar final diversity values (mean mOTUs t_0 = 1155 ± 83). The treatments also generated differences, with VL and DL being the least rich in winter and summer at the end of the experiment (Figure 1). The global dissimilarity between samples also presented a strong seasonal signal (Figure 2). Most of the variation was explained by season (Euclidean, PERMANOVA $R^2 = 0.33$, $p \leq 0.001$), followed by treatment ($R^2 = 0.14$, $p \leq 0.001$). The initial and final times were non-significant as a whole ($R^2 = 0.01$, $p = 0.23$), probably because spring was the season presenting less variation between treatments and sampling time, with the control treatment at the final time clustering with the initial times. Winter and summer presented a clear differentiation between the initial and final times, with the diluted and virus-reduced treatments presenting the largest differentiation. Overall, season was the most relevant structuring factor and, in most cases, treatment also induced a change in community structure.

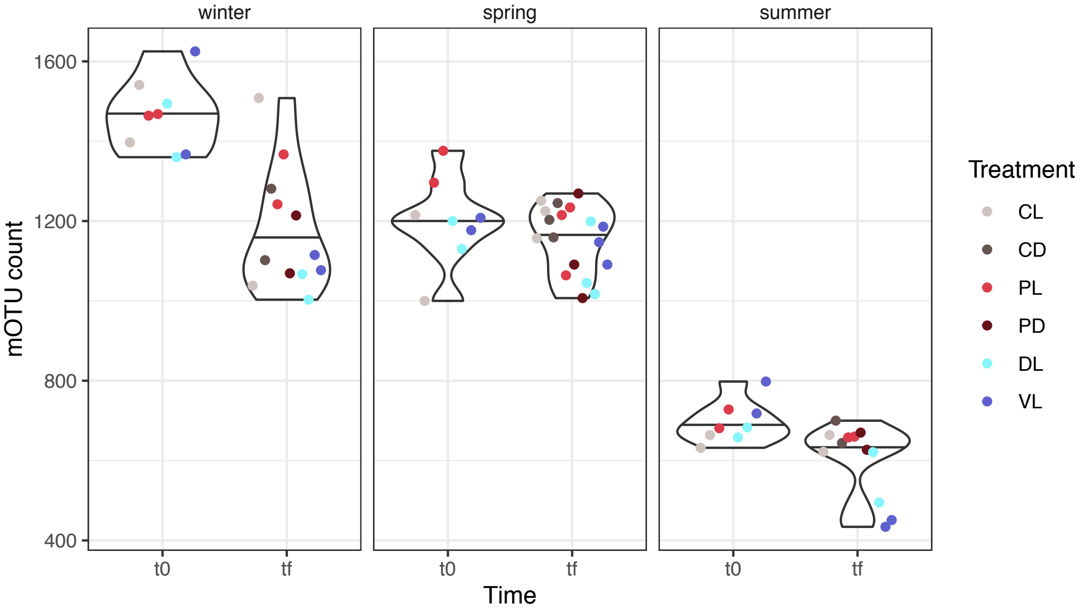


Figure 1: mOTU community richness at the start and the end of each experiment. The X axis indicates the initial (t_0) and final (t_f) time of the experiments and the Y axis indicates the number of different mOTUs observed in each sample (each point is one sample), and colored by treatment. The violin plot indicates the density distribution and the horizontal line the median value.

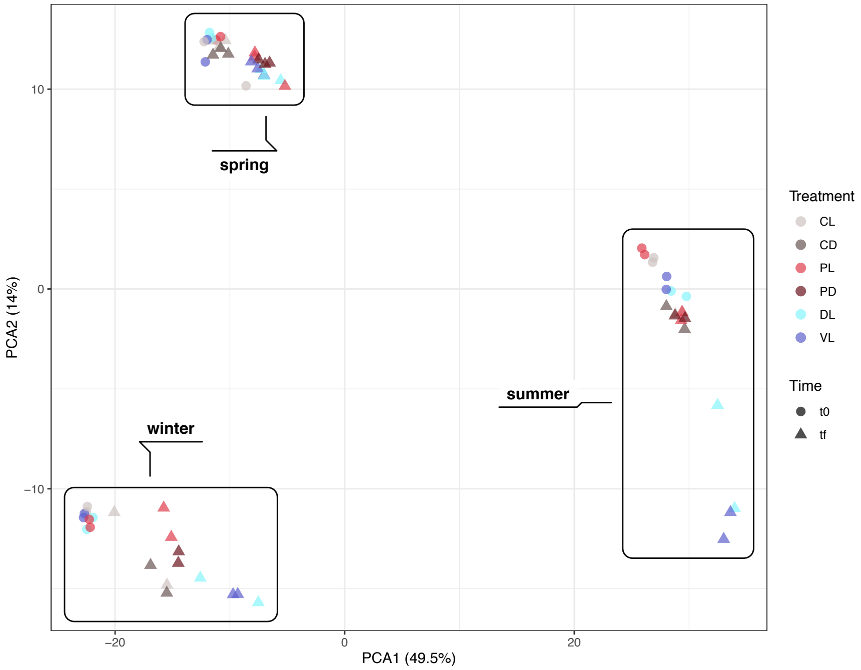


Figure 2: Principal component analysis displaying the Euclidean distance between communities. Each point is a sample, colored by treatment, and shaped differentiating the initial (t_0) and final (t_f) times of the experiments. The axes are the two main principal components, together explaining 63.5% of the dataset variability.

Selection by treatment and season

We obtained a total of 262 MAGs, with a mean completion score of 64% and a mean contamination score of 1.1% (Supplementary Table 1). These MAGs recruited on average 60% of the total reads, with a minimum sample recruitment of 33% and a maximum of 93%. The most common prokaryotic classes in our MAG collection were Alphaproteobacteria (71 MAGs, 27.8%), Bacteroidia (27.5%), Gammaproteobacteria (25.1%) and Verrucomicrobia (5.49%). Other relevant groups were Cyanobacteriia (3.1%), Acidimicrobiia (2.7%), Actinomycetia (1.9%), Planctomycetes (1.18%), and Gemmatimonadetes (0.7%). Additionally, MAG202 was affiliated to the Patescibacteria phylum, part of the candidate phyla radiation (Brown *et al.*, 2015). The most common families were *Flavobacteriaceae* (16.9%), *Rhodobacteraceae* (10.2%), *Sphingomonadaceae* (4.3%), and *Alteromonadaceae* (4.3%). A total of 70 MAGs were present in the 3 experiments, whereas 78 of them were specific of a single season/experiment, and 114 were present in two seasons (Supplementary Figure 1). A total of 80 MAGs were shared between winter and spring, whereas 26 MAGs were shared between spring and summer.

The abundance of each MAG –coverage divided by the sample mean MAGs coverage– was used to determine the distribution of each MAG in the experiments and treatments (Table 2, Supplemen-

Top5 MAGs presenting a X1.5 fold change in at least one treatment

	CL			CD			PL			PD			DL			VL		
MAGs	t0	tf	Ratio	t0	tf	Ratio	t0	tf	Ratio	t0	tf	Ratio	t0	tf	Ratio	t0	tf	Ratio
winter																		
MAG_00292 (O) Enterobacterales (F) *Alteromonadaceae* (G) *Glaciecola*	-	5.20	100.0	-	1.50	100.0	0.25	28.63	114.1	0.25	29.00	115.5	0.14	20.84	151.5	0.11	6.46	57.8
MAG_00204 (O) Enterobacterales (F) *Alteromonadaceae* (G) *Alteromonas*	-	0.10	100.0	-	-	-	-	0.38	100.0	-	0.26	100.0	-	1.32	100.0	-	1.02	100.0
MAG_00257 (O) Enterobacterales (F) *Alteromonadaceae* (G) *Alteromonas*	-	-	-	-	-	-	-	0.19	100.0	-	0.13	100.0	-	0.73	100.0	-	0.56	100.0
MAG_00170 (O) PCC-6307 (F) *Cyanobiaceae* (G) *Synechococcus*	2.41	10.44	4.3	2.41	5.72	2.4	3.50	1.11	0.3	3.50	0.76	0.2	4.31	3.15	0.7	4.67	8.95	1.9
MAG_00290 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) SCGC-AAA160-P02	11.35	44.41	3.9	11.35	60.95	5.4	16.36	123.02	7.5	16.36	127.16	7.8	10.84	99.08	9.1	13.02	107.71	8.3
MAG_00186 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) *Polaribacter*	1.41	6.84	4.8	1.41	7.75	5.5	1.79	68.69	38.4	1.79	69.37	38.8	1.10	79.60	72.4	1.33	21.21	15.9
MAG_00026 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) UBA3537	0.74	3.66	5.0	0.74	4.71	6.4	1.05	2.37	2.3	1.05	2.92	2.8	0.70	1.98	2.8	0.82	0.61	0.7
MAG_00193 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) HC6-5	0.48	2.09	4.3	0.48	2.58	5.4	0.57	0.95	1.7	0.57	1.23	2.2	0.42	1.76	4.2	0.51	0.22	0.4
MAG_00168 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) Hel1-33-131	0.23	0.89	3.9	0.23	0.82	3.6	0.20	0.80	4.0	0.20	0.92	4.5	0.15	0.78	5.3	0.19	0.55	3.0
MAG_00206 (O) Pseudomonadales (F) *Halieaceae* (G) *Luminiphilus*	0.64	0.71	1.1	0.64	0.69	1.1	1.00	0.51	0.5	1.00	0.42	0.4	0.79	0.11	0.1	0.67	-	0.0
MAG_00291 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) *Planktomarina*	2.76	16.84	6.1	2.76	18.67	6.8	5.64	4.45	0.8	5.64	3.92	0.7	3.34	5.92	1.8	3.42	12.51	3.7
MAG_00244 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) NA	0.17	3.49	20.2	0.17	2.15	12.4	0.47	9.79	20.8	0.47	10.37	22.0	0.22	30.48	135.5	0.24	69.49	292.2
MAG_00243 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) NA	0.17	2.97	17.7	0.17	1.83	10.9	0.40	8.24	20.5	0.40	8.85	22.0	0.20	24.70	122.5	0.21	57.75	274.7
MAG_00028 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) *Nereida*	-	0.15	100.0	-	0.13	100.0	-	0.06	100.0	-	0.06	100.0	-	0.20	100.0	-	0.38	100.0
MAG_00083 (O) Flavobacteriales (F) *Schleiferiaceae* (G) UBA10364	1.80	1.31	0.7	1.80	1.50	0.8	2.17	1.82	0.8	2.17	2.67	1.2	1.47	0.43	0.3	1.69	0.04	0.0
MAG_00214 (O) Flavobacteriales (F) *Schleiferiaceae* (G) UBA10364	0.22	0.15	0.7	0.22	0.10	0.5	0.29	0.06	0.2	0.29	0.08	0.3	0.20	-	0.0	0.21	-	0.0
MAG_00014 (O) Flavobacteriales (F) UA16 (G) UBA8752	3.35	10.34	3.1	3.35	12.30	3.7	0.93	0.47	0.5	0.93	0.46	0.5	3.79	1.65	0.4	3.73	0.23	0.1
MAG_00180 (O) Pirellulales (F) UBA1268 (G) UBA1268	0.46	1.04	2.3	0.46	1.67	3.6	1.46	0.30	0.2	1.46	0.35	0.2	1.13	0.51	0.4	1.09	1.19	1.1
MAG_00027 (O) NS11-12g (F) UBA9320 (G) UBA9320	-	-	-	-	0.05	100.0	-	-	-	-	-	-	-	-	-	-	-	-
spring																		
MAG_00292 (O) Enterobacterales (F) *Alteromonadaceae* (G) *Glaciecola*	0.51	1.29	2.5	0.51	1.04	2.0	0.25	1.95	7.6	0.25	3.24	12.7	0.21	26.39	124.0	0.30	36.83	122.9
MAG_00204 (O) Enterobacterales (F) *Alteromonadaceae* (G) *Alteromonas*	-	-	-	-	-	-	-	0.04	100.0	-	0.09	100.0	-	0.81	100.0	-	1.23	100.0
MAG_00257 (O) Enterobacterales (F) *Alteromonadaceae* (G) *Alteromonas*	-	-	-	-	-	-	-	-	-	-	0.05	100.0	-	0.40	100.0	-	0.59	100.0
MAG_00170 (O) PCC-6307 (F) *Cyanobiaceae* (G) *Synechococcus*	13.31	16.00	1.2	13.31	14.39	1.1	23.12	23.71	1.0	23.12	15.21	0.7	20.57	15.40	0.7	6.91	15.48	2.2
MAG_00168 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) Hel1-33-131	6.46	10.82	1.7	6.46	16.32	2.5	6.50	15.33	2.4	6.50	15.64	2.4	8.08	8.54	1.1	6.58	8.63	1.3
MAG_00026 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) UBA3537	6.45	10.19	1.6	6.45	18.06	2.8	7.82	12.05	1.5	7.82	10.95	1.4	8.11	6.07	0.7	6.09	5.17	0.9
MAG_00193 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) HC6-5	3.07	3.78	1.2	3.07	7.67	2.5	2.14	6.82	3.2	2.14	7.84	3.7	2.39	4.90	2.0	1.79	1.74	1.0
MAG_00290 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) SCGC-AAA160-P02	3.38	3.13	0.9	3.38	5.72	1.7	3.84	23.91	6.2	3.84	28.86	7.5	3.37	20.07	6.0	2.88	14.50	5.0
MAG_00062 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) NA	0.79	1.19	1.5	0.79	1.65	2.1	0.98	2.26	2.3	0.98	1.94	2.0	1.20	0.75	0.6	1.08	0.56	0.5
MAG_00186 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) *Polaribacter*	0.70	0.89	1.3	0.70	1.22	1.7	0.47	7.47	16.0	0.47	6.56	14.1	0.32	5.42	16.8	0.38	2.50	6.6
MAG_00206 (O) Pseudomonadales (F) *Halieaceae* (G) *Luminiphilus*	3.28	3.09	0.9	3.28	2.99	0.9	2.37	1.59	0.7	2.37	1.11	0.5	1.86	1.01	0.5	3.13	1.26	0.4
MAG_00002 (O) Puniceispirillales (F) *Puniceispirillaceae* (G) HIMB100	0.81	1.01	1.2	0.81	0.57	0.7	0.84	0.33	0.4	0.84	0.33	0.4	0.66	0.31	0.5	0.98	0.50	0.5
MAG_00291 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) *Planktomarina*	3.89	3.99	1.0	3.89	5.11	1.3	3.83	6.21	1.6	3.83	8.97	2.3	3.15	10.81	3.4	1.88	7.67	4.1
MAG_00028 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) *Nereida*	1.72	2.41	1.4	1.72	3.88	2.3	1.34	5.28	3.9	1.34	9.49	7.1	1.08	39.65	36.6	0.35	29.45	85.1
MAG_00244 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) NA	0.88	1.97	2.2	0.88	2.97	3.4	0.67	4.33	6.5	0.67	5.62	8.4	0.55	12.93	23.5	0.22	11.65	53.6
MAG_00243 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) NA	0.78	1.76	2.3	0.78	2.63	3.4	0.57	3.74	6.5	0.57	4.85	8.5	0.49	10.63	21.8	0.24	9.82	40.9
MAG_00083 (O) Flavobacteriales (F) *Schleiferiaceae* (G) UBA10364	9.47	7.35	0.8	9.47	8.12	0.9	6.95	14.09	2.0	6.95	14.48	2.1	5.73	2.30	0.4	7.71	2.44	0.3
MAG_00014 (O) Flavobacteriales (F) UA16 (G) UBA8752	0.17	0.13	0.8	0.17	0.16	1.0	-	-	-	-	-	-	0.12	0.07	0.6	-	-	-
MAG_00180 (O) Pirellulales (F) UBA1268 (G) UBA1268	2.37	2.56	1.1	2.37	4.10	1.7	5.30	4.75	0.9	5.30	4.38	0.8	3.92	2.42	0.6	1.20	2.49	2.1
MAG_00027 (O) NS11-12g (F) UBA9320 (G) UBA9320	1.38	2.19	1.6	1.38	3.40	2.5	1.41	1.52	1.1	1.41	1.45	1.0	1.58	0.57	0.4	1.36	0.89	0.7
summer																		
MAG_00204 (O) Enterobacterales (F) *Alteromonadaceae* (G) *Alteromonas*	-	1.41	100.0	-	3.38	100.0	-	1.67	100.0	-	5.51	100.0	2.36	31.89	13.5	2.12	69.66	32.9
MAG_00292 (O) Enterobacterales (F) *Alteromonadaceae* (G) *Glaciecola*	-	1.37	100.0	-	5.87	100.0	0.14	3.46	24.0	0.14	3.86	26.8	0.63	47.03	75.1	0.53	24.10	45.8
MAG_00257 (O) Enterobacterales (F) *Alteromonadaceae* (G) *Alteromonas*	-	0.83	100.0	-	1.88	100.0	-	1.17	100.0	-	3.23	100.0	1.39	19.58	14.1	1.25	38.52	30.9
MAG_00062 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) NA	1.39	3.32	2.4	1.39	3.13	2.3	1.76	3.95	2.2	1.76	4.50	2.6	1.39	1.20	0.9	1.36	0.51	0.4
MAG_00013 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) UBA8316	1.91	3.06	1.6	1.91	2.41	1.3	3.59	6.88	1.9	3.59	7.10	2.0	2.25	1.71	0.8	2.42	1.13	0.5
MAG_00026 (O) Flavobacteriales (F) *Flavobacteriaceae* (G) UBA3537	0.43	1.85	4.3	0.43	3.02	7.0	1.00	2.39	2.4	1.00	2.77	2.8	0.78	1.77	2.3	0.78	1.76	2.3
MAG_00206 (O) Pseudomonadales (F) *Halieaceae* (G) *Luminiphilus*	1.77	3.22	1.8	1.77	3.93	2.2	2.57	2.93	1.1	2.57	3.22	1.3	1.87	1.81	1.0	1.72	0.72	0.4
MAG_00172 (O) Actinomycetales (F) *Microbacteriaceae* (G) *Pontimonas*	2.40	1.56	0.6	2.40	1.38	0.6	2.39	4.16	1.7	2.39	2.50	1.0	3.24	2.95	0.9	2.32	1.67	0.7
MAG_00002 (O) Puniceispirillales (F) *Puniceispirillaceae* (G) HIMB100	3.01	5.36	1.8	3.01	5.57	1.9	3.21	4.02	1.3	3.21	3.42	1.1	3.61	1.10	0.3	4.18	0.22	0.1
MAG_00043 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) HIMB11	0.34	2.04	6.0	0.34	3.56	10.5	0.54	3.68	6.9	0.54	3.18	6.0	0.70	5.45	7.8	0.55	1.05	1.9
MAG_00243 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) NA	-	0.22	100.0	-	0.59	100.0	-	1.49	100.0	-	1.32	100.0	0.22	22.41	100.6	0.16	21.35	137.7
MAG_00244 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) NA	-	0.21	100.0	-	0.62	100.0	0.08	1.59	21.1	0.08	1.39	18.4	0.24	24.62	101.7	0.16	23.45	145.8
MAG_00028 (O) Rhodobacterales (F) *Rhodobacteraceae* (G) *Nereida*	-	-	-	-	0.21	100.0	-	0.20	100.0	-	0.19	100.0	0.14	2.91	21.1	-	3.83	100.0
MAG_00214 (O) Flavobacteriales (F) *Schleiferiaceae* (G) UBA10364	0.28	2.17	7.8	0.28	3.58	12.8	0.98	3.29	3.4	0.98	4.38	4.5	0.76	0.68	0.9	0.62	0.24	0.4
MAG_00083 (O) Flavobacteriales (F) *Schleiferiaceae* (G) UBA10364	0.64	1.71	2.7	0.64	2.30	3.6	1.80	2.78	1.5	1.80	3.63	2.0	1.15	0.75	0.6	0.97	0.62	0.6
MAG_00003 (O) Flavobacteriales (F) UA16 (G) UBA8752	1.57	3.49	2.2	1.57	3.18	2.0	0.86	2.86	3.3	0.86	3.55	4.1	1.42	1.07	0.8	1.20	0.51	0.4
MAG_00014 (O) Flavobacteriales (F) UA16 (G) UBA8752	0.10	0.11	1.1	0.10	0.09	0.9	-	0.10	100.0	-	0.12	100.0	-	-	-	0.08	-	0.0
MAG_00180 (O) Pirellulales (F) UBA1268 (G) UBA1268	-	-	-	-	-	-	0.08	0.09	1.1	0.08	0.07	0.8	-	-	-	-	0.10	100.0
MAG_00090 (O) Rhodobacterales (F) UBA8317 (G) UBA8317	0.47	1.26	2.7	0.47	1.57	3.4	0.87	4.95	5.7	0.87	2.91	3.3	0.98	2.60	2.7	0.62	1.26	2.0
MAG_00040 (O) NS11-12g (F) UBA9320 (G) MED-G17	2.65	4.17	1.6	2.65	3.33	1.3	4.25	7.47	1.8	4.25	8.12	1.9	3.45	1.95	0.6	3.16	1.74	0.6
MAG_00027 (O) NS11-12g (F) UBA9320 (G) UBA9320	0.52	0.94	1.8	0.52	0.93	1.8	1.07	1.32	1.2	1.07	1.50	1.4	0.87	0.37	0.4	0.83	0.39	0.5

The abundance is the MAG coverage divided by the overall sample mean coverage.

MAGs not observed only in tf but not in t0 present an hypothethical X100 ratio

Table 2: Top 5 MAGs displaying a positive fold change (i.e. they grew) between the initial and final times in each season. The t_0 column indicates the initial mean abundance value, the t_f indicates the abundance at the final time, and the fold change is the division of these two numbers (in green when the value is over 1.2). Each MAG is labelled specifying the order (O), family (F) and genus (G). When the initial values were below detection, the fold change is set arbitrarily to 100.

tary Table 2). We also calculated the fold change of this value between the initial and final time to establish whether each species increased or decreased. The species indicator statistic (indval test, r.g ≥ 0.5, p ≤ 0.01) was used to differentiate which MAGs were indicator species for each experiment and treatment. We inspected the top 5 most abundant MAGs (Figure 3A) and those MAGs presenting a high abundance and a high fold change (Figure 3B). Many of the most abundant MAGs did not present similarities with cultured species, highlighted by their nomenclature obtained from the Genome Taxonomy Database (GTDB), such as MAG156 (TMED189 species, Acidimicrobiia), MAG51 (MED-G52, *Rhodobacteraceae*), or MAG153 (UBA7446, *Flavobacteriaceae*). Several of the most abundant MAGs did not present a high rate of growth in the experiments, or decreased slightly (Figure 3A). As an example, in spring, MAG170 (*Synechococccus*) was one of the most abundant MAGs with a fold change of ca. 1 between the initial and final times (i.e. no change), and similar patterns could be observed in winter in some specific *Flavobacteriaceae* and *Rhodobacteraceae* MAGs (MAG221, 189), and in summer in *Litoricola*, Verrucomicrobia and *Puniceispirillaceae* (MAG247, 223, 1). Other MAGs reached the top 5 increasing substantially between the initial and the final times: MAG290 (*Flavobacteriaceae*) during winter and spring was the most dominant species, *Alteromonadaceae* MAGs (MAG257, 204, 292) increased in abundance in all seasons together with two *Rhodobacteraceae* MAGs (MAG243, 244). Overall, these results show that 11 (42%) of the most abundant MAGs present in the experiments maintained their relative abundance from the intial conditions, whereas 15 (57%) MAGs presented large abundance increases (Figure 3A).

Alongside season, at the end of the incubations (t_f), different treatments had specific indicator MAGs (Figure 3A, 3B, Table 2). The treatment indicator MAGs were usually the same between both controls (CL and CD), both predator-reduced treatments (PL and PD), and between the diluted and the virus-reduced (VL and DL) treatments. Specifically, 26% of the MAGs were indicative of CL and CD, 24% appeared both in the control and predator-reduced treatments, 18% did not had any specific preference, 12.8% were indicatice of the virus-reduced and diluted treatments, 7.3% of the predator-reduced treatment, and some indicated only one treatment, such as control light (3%) or virus-reduced (1.9%). In the controls and predator-reduced treatments where we incubated under PAR or in the dark, there was not strong indicative species growing under light and not under dark conditions. The only clear example of this were 9 MAGs indicative for the control light experiment each in one or various of the experiments (Supplementary Figure 2), including MAGs of *Prochlorococcus* (MAG174), SAR86 (MAG129), and *Porticoccus* (MAG8). Generally, the control treatments included groups with a fold change around 1 but presenting specificity to these treatments, given that its abundance lowered in the other treatments (see Figure 3A). The predator-reduced treatment had mostly *Flavobacteriaceae* MAGs as indicator species (Figure 3B). Each season included specific species: *Flavicella*, *Polaribacter*, and *Winogradskyella* species (MAG60, 154, 232) in the spring experiment; three other *Flavobacteriaceae* MAGs, two from the the *Polaribacter* genus (MAG198, 229) and another without genus assignation (MAG185) in winter, and finally two species without

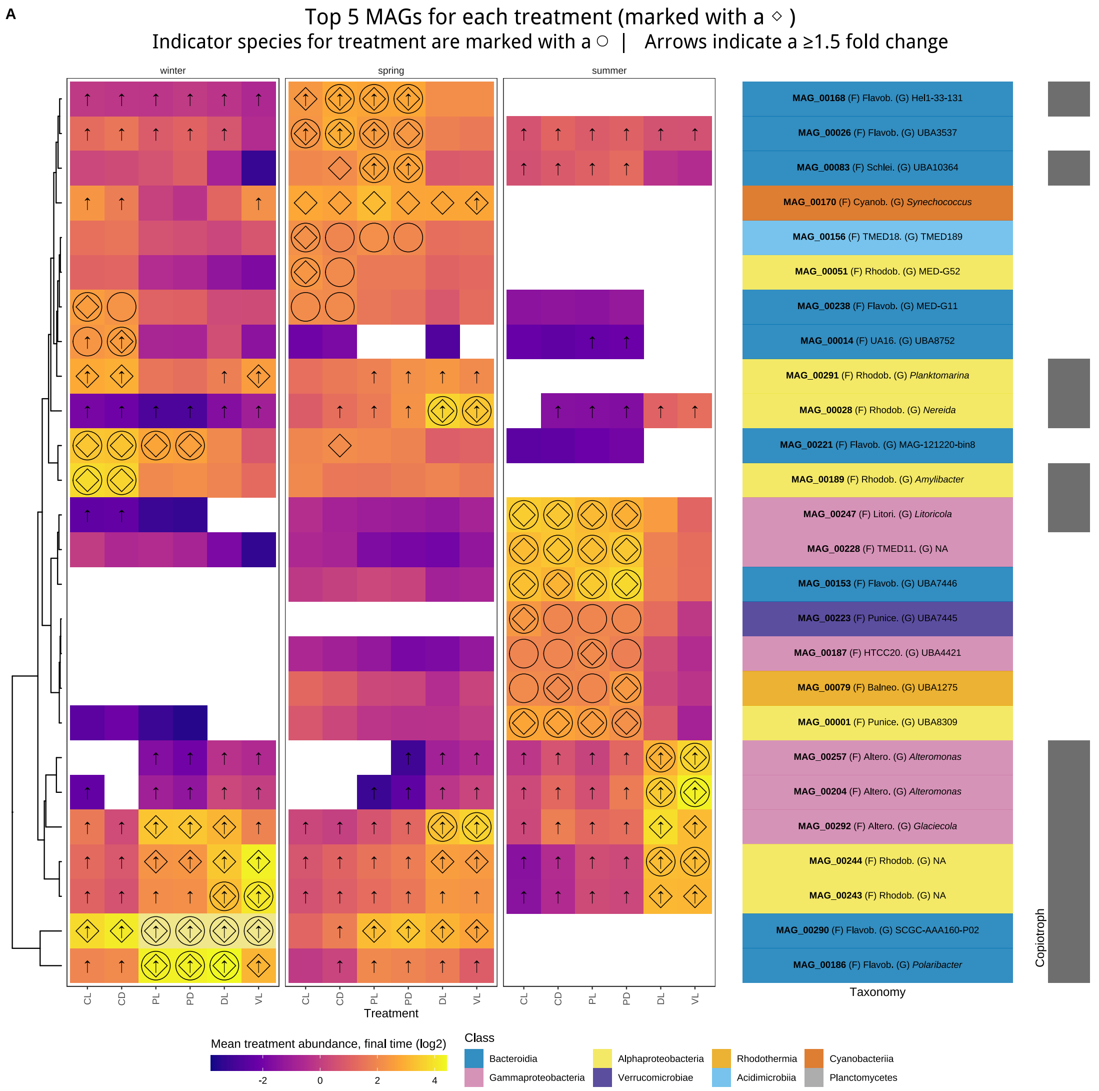
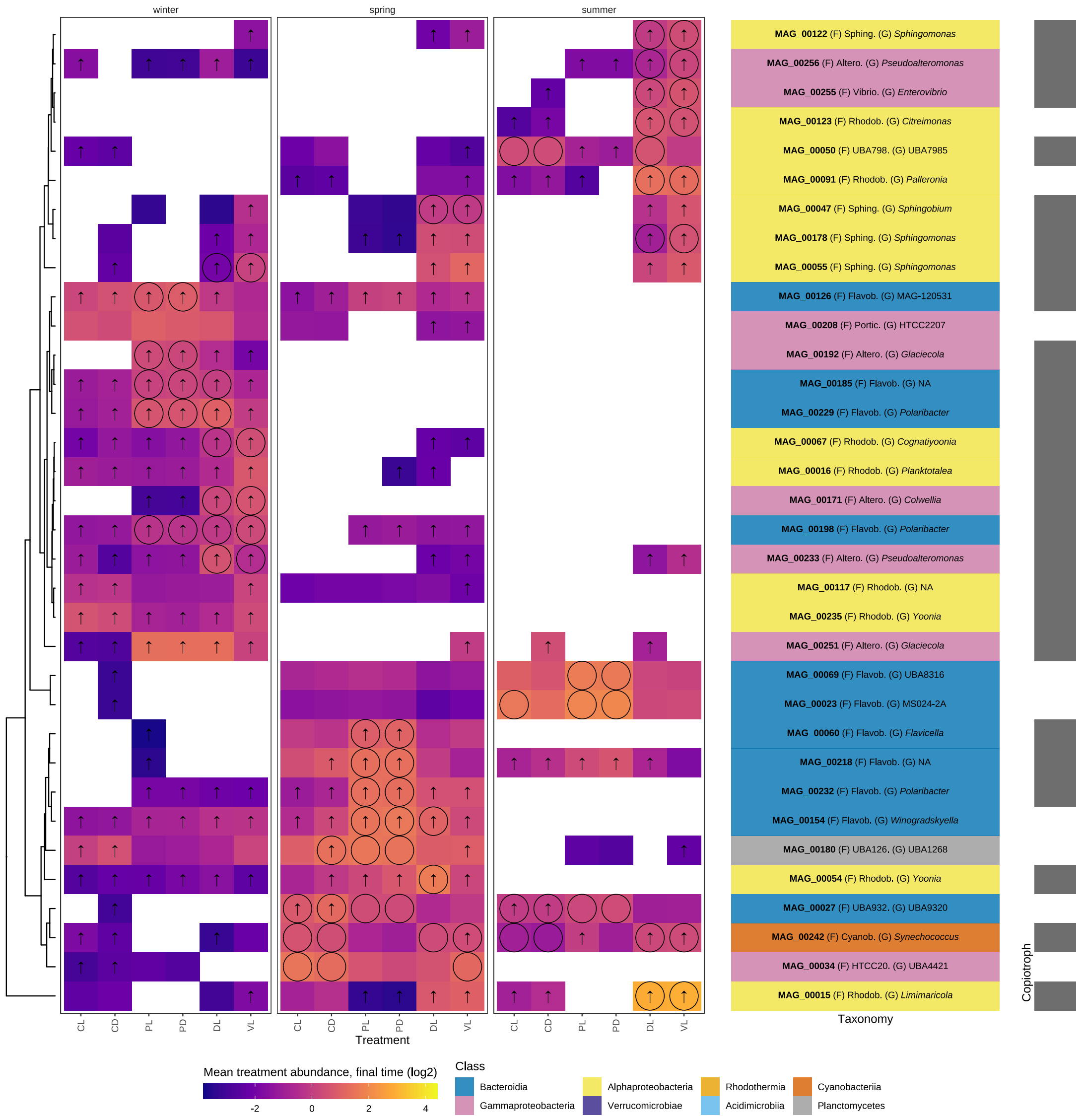


Figure 3: Heatmap with the mean abundance of A) the top 5 MAGs for each treatment, and B) the MAGs presenting the highest fold change and highest abundance from start to end of the experiment. The abundance is computed as the mean MAG coverage divided by the mean coverage of all MAGs in the sample, displayed in the plot in a log2 scale. MAG290 had a log₂ with a median of 7, out of the range of the other MAGs, and it was not included in the legend for a better visualization of the other MAGs. We colored these values with a light yellow. Taxonomy is indicated in the right (label indicating the family and the genus, colored by class) together with the growth category (copiotrophy in grey, oligotrophy in white). The MAGs are clustered based on abundance. A diamond in panel A indicates that the MAG is one of the top 5, and a circle in panels A and B indicates that the MAG is an indicator species for that season and treatment. The arrow indicates a fold change between the initial and final time higher or equal than 1.5.

B

Most abundant MAGs presenting the highest increase (↑)



a cultured relative (MAG23, 69) in summer. Another MAG indicative of the predator-reduced treatment was affiliated to *Glaciecola* (MAG192). The virus-reduced and diluted treatments on the other hand presented mostly *Sphingomonadaceae*, *Alteromonadaceae* and *Rhodobacteraceae* MAGs (Figure 3A, 3B). Many of these MAGs presented a similar increase in the different experiments, such as *Alteromononas* (MAG204, 257), *Glaciecola* (MAG292) and uncultured *Rhodobacteraceae* MAGs (MAG243, 244; Figure 3A). Although less abundant, many *Sphingomonadaceae* MAGs (MAG55, 122, 178) increased in all experiments in the diluted and in the virus-reduced treatments. Other species increasing in these treatments were *Colwellia*, increasing in winter (*Alteromonadaceae*, MAG187) and *Citerimonas* and *Palleronia* in the summer experiments (*Rhodobacteraceae*, MAG123, 193).

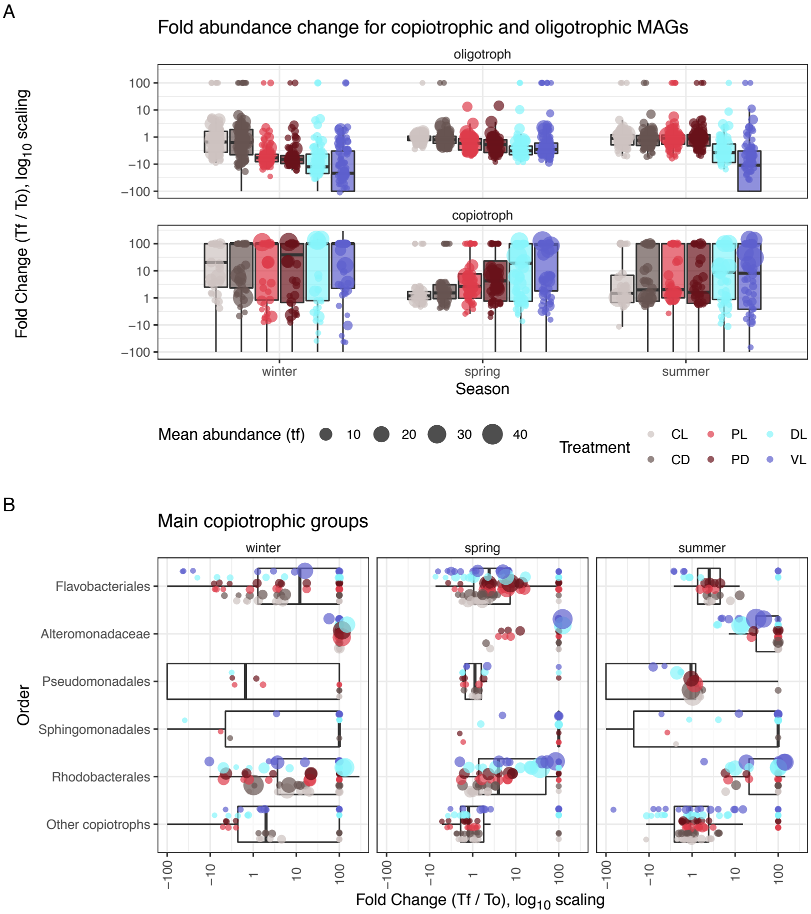


Figure 4: Fold change comparison between copiotrophic and oligotrophic bacteria. A) Bulk fold change comparison between copiotrophic and oligotrophic MAGs at each season. The X axis indicates the seasons and the Y axis indicates the fold change between the initial and the final times on a log10 scale. Each point and boxplots are colored by treatment, and the size of the points is related to the MAG abundance. B) Fold change of the main copiotrophic groups. Each point is colored by treatment and point size is related to the MAG abundance. Each boxplot presents the median and the 25% and 75% limits, and whiskers represent 1.5 times the interquartile range. When the initial values were below detection, the fold change is set arbitrarily to 100.

Besides describing the patterns of individual MAGs, we evaluated the trends between oligotrophic and copiotrophic bacteria (Figure 4, Supplementary Table 3). We assumed all bacteria presenting a maximal doubling time of under 5 hours to be copiotrophs (based on their genomic properties), and the rest as oligotrophs (see Materials and Methods and Weissman *et al.*, 2021). The different treatments presented variations in the fold change (FC) of these two types of organisms between the start and end of the experiment (Figure 4A). As a generality, oligotrophs presented a smaller fold change (median FC = 0.4, which imply a decrease in relative abundance from the start to the end of the experiment) than that of the copiotrophs (FC = 4.7, which indicates a high growth). For oligotrophs, generally the control treatment was the one with the highest fold change (FC = 0.8), with exception of the predator-reduced treatment in summer, presenting a similar value to the control (FC = 0.94). Some oligotrophs had a large FC of 100 in multiple treatments: *Winogradskyella* (Bacteroidia) and *Citreimonas* (Rhodobacterales) presented a predicted doubling rate close to the threshold that separates oligotrophs from copiotrophs.

Copiotrophs generally presented high growth in most treatments and seasons, with many taxa being non-detectable at the initial time, yet becoming dominant in the final time (FC > 100). Alteromonadales and Rhodobacterales were those presenting the highest fold change and relative abundance in most treatments, albeit with Rhodobacterales presenting more variance in the response (Figure 4B). Within Rhodobacterales, *Amylibacter* (Figure 3A) decreased (FC=0.5) while *Planktomarina* only grew slightly (median FC = 2), whereas other initially less abundant groups increased substantially, such as *Lentibacter* (FC = 9.7), *Yoonia* (FC >100), *Planktotaela* (FC > 100) or *Nereida* (FC > 100), and a few MAGs without assignation also presented a high increase (FC = 53). The copiotrophic Flavobacteriales MAGs also had a high FC, although these MAGs were mainly present in winter (*Polaribacter, Dokdonia, Aurantibacter*) and spring (*Polaribacter*, multiple uncultured MAGs), as explained above. Overall, our results illustrate the differential response of oligotrophs and copiotrophs in the various treatments and seasons. Likewise, our experimental approach shows a similar response of Alteromonadales to all treatment and seasons, and the variability in the response within the Rhodobacterales and Flavobacteriales orders.

Functional analysis

Having differentiated the growth and preference for treatments and seasons of each MAG, we looked for the presence of biogeochemically relevant functions in all the MAGs presenting high genome completeness (≥ 70%). We focused on the presence of key biogeochemically relevant genes that could associate the MAGs with key roles in ecosystem functioning. Several biogeochemically relevant genes were affiliated to specific taxonomic groups (Figure 5, Supplementary Figure 3), whereas others were widespread among multiple phyla, such as proteorhodopsin and the *amt* gene (coding an ammonia transporter). The *plcP* gene –coding an enzyme able to remodel membrane lipids to overcome phosphorous starvation– was mostly linked to the Bacteroidia and

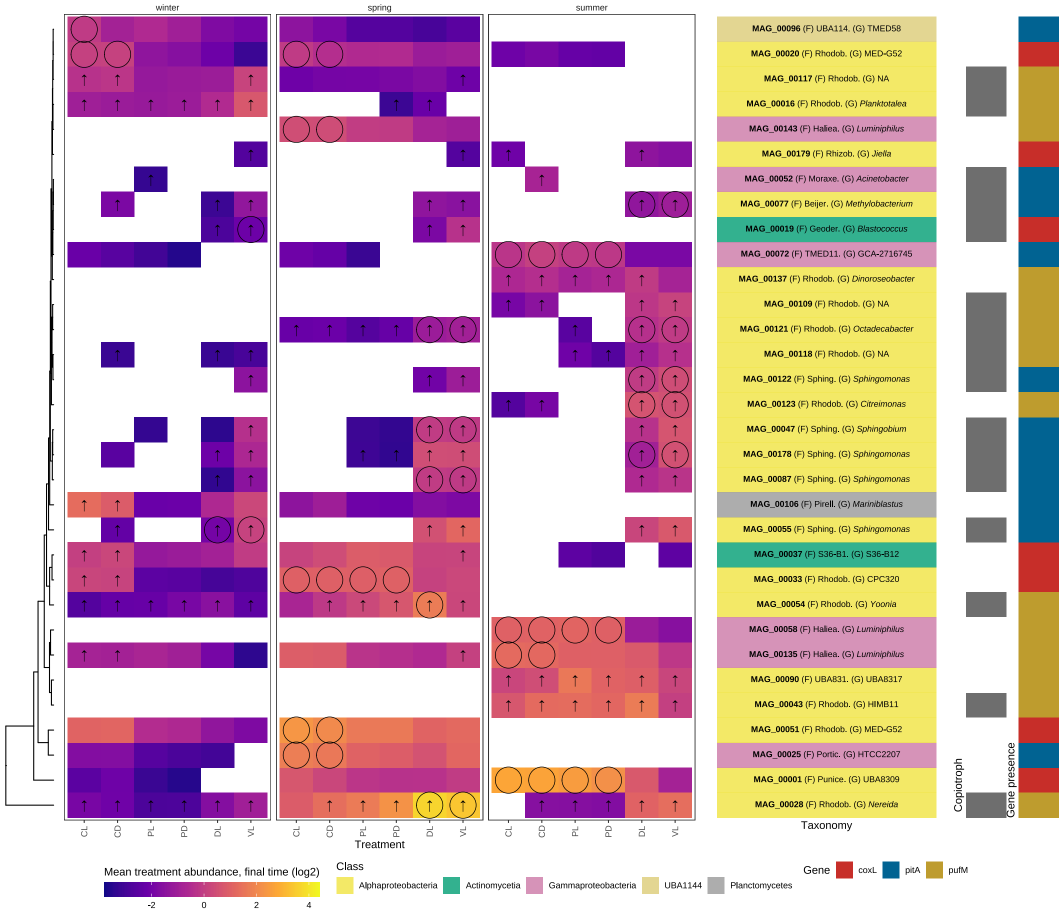


Figure 5: Heatmap of the MAGs containing *pufM*, *pitA* and *coxL* genes in their genetic repertoire. The abundance indicates the mean MAG coverage divided by the mean coverage of all MAGs in the sample, in a $\log_2$ scale. The columns on the right indicate taxonomy (label indicating the family and the genus, colored by class) and growth behavior (copiotrophy in grey, oligotrophy in white). The MAGs are clustered based on their abundance. A diamond indicates that the MAG is one of the top 5 in panel A, and a circle indicates that the MAG is an indicator species for that season and treatment. The arrow indicates a fold change between the initial and final time higher or equal than 1.5.

Gammaproteobacteria groups. The *pitA* gene –encoding a low affinitiy phosphorous transporter expressed under non-limiting phosphorous conditions– was enriched in the virus-reduced and diluted treatments (54% of the variants were harbored by MAGs growing only under these conditions), specially in *Sphingomonadaceae* MAGs. The *coxL* gene –coding a carbon monoxide oxidase– was enriched in groups that were abundant in the initial conditions but did not grow in the treatments, such as *Puniceispirillaceae* (MAG1) and some *Rhodobacteaceae* species (MAG20, 33). The MAGs harboring *pufM* –aerobic anoxygenic phototrophic (AAPs) bacteria– presented different responses to the treatments (Figure 5): *Luminiphilus* AAPs did not increase in abundance in most of the treatments and seasons while the *Rhodobacteraceae* copiotrophic AAPs grew in many treatments;

in the virus-reduced and diluted treatment the indicator species were *Octadecabacter* (MAG121), *Citerimonas* (MAG123), and an uncultured MAG (MAG109); other *Rhodobacteraceae* AAPs grew more indiscriminately, such as HIMB11 in summer (MAG43), *Nereida* in all seasons but specially during spring (MAG28), and *Yoonia* during winter and spring (MAG54). Finally, 4 out of the 9 MAGs being indicative of the control light treatment (Supplementary Figure 2), presented a proteorhodopsin, such as *Porticoccus* (MAG8), an uncultured *Woeseiaceae* (MAG21), SAR86 (MAG129), and an uncultured *Marinoscillaceae* (MAG149). Another of the species was a *Prochlorococcus* (MAG174).

Additionally, we inspected the enrichment of complete KEGG metabolic pathways (modules) between the virus-reduced/diluted and the control treatments in the different seasons (Figure 6), since these treatments presented the most specific response by specific taxons (Figure 3). In these treatments, in spring, the pathways to synthetize polyamines and glycogen were enriched (Figure 6A). Glycogen is a macromolecule used by many bacterial groups to store carbon (Sekar *et al.*, 2020) and was mainly enriched in Opitutales and Sphingomonadales MAGs (Figure 6B). The degradation of D-galactonate was also enriched in these conditions. This molecule is produced mostly by algae, and both Rhodobacterales and Sphingomonadales presented the enriched module for its degradation (Ficko-Blean *et al.*, 2017). The summer season also presented specific modules that appeared in these treatments. There were modules linked to the degradation of sugars, such as the Entner-Doudoroff pathway and the glyoxylate cycle, and modules linked to the synthesis of molecules such as the lysine aminoacid, nucleotides through the pentose phosphate pathway, NAD, and obtention of sulfates by assimilatory sulfate reduction. Two functions were linked to only a small fraction of MAGs. The biosynthesis of ectoine–an osmolyte produced to prevent osmotic stress (Widderich *et al.*, 2014)– was only associated to some specific Sphingomonadales MAGs. The degradation of tyrosine, possibly linked to the production of pigments, was found in one Puniceispirillales genome and multiple Rhodobacterales and Sphingomonadales genomes (Coon *et al.*, 1994). Multiple modules were enriched in the virus-reduced and diluted conditions irrespective of season. Most of them associated to the Sphingomonadales groups, although in some cases also to Caulobacterales (Figure 6B). The enriched modules ranged from the thiamine (vitamin B1) and trehalose (carbon storage, osmotic stress) biosynthesis to the creation of beta-lactamases. Sphingomonadales also presented variants of the typical cytochrome (variants o and bd) which present different affinities for O_2 (Gong *et al.*, 2018). Finally, many Flavobacterial groups presented the complete pathway for histidine degradation, appearing in the winter control treatments (Bender, 2012). Although many groups presented this module, most of them did not present the complete pathway. These analyses indicate that the selection of specific species in each treatment might have functional implications at the community level, such as the presence of functions to synthesize macromolecules during spring, and the degradation of glycogen and tyrosine coupled with synthesizing aminoacids and nucleotides in summer, indicating than under the virus-reduced and diluted treatments different processes are stimulated at each season.

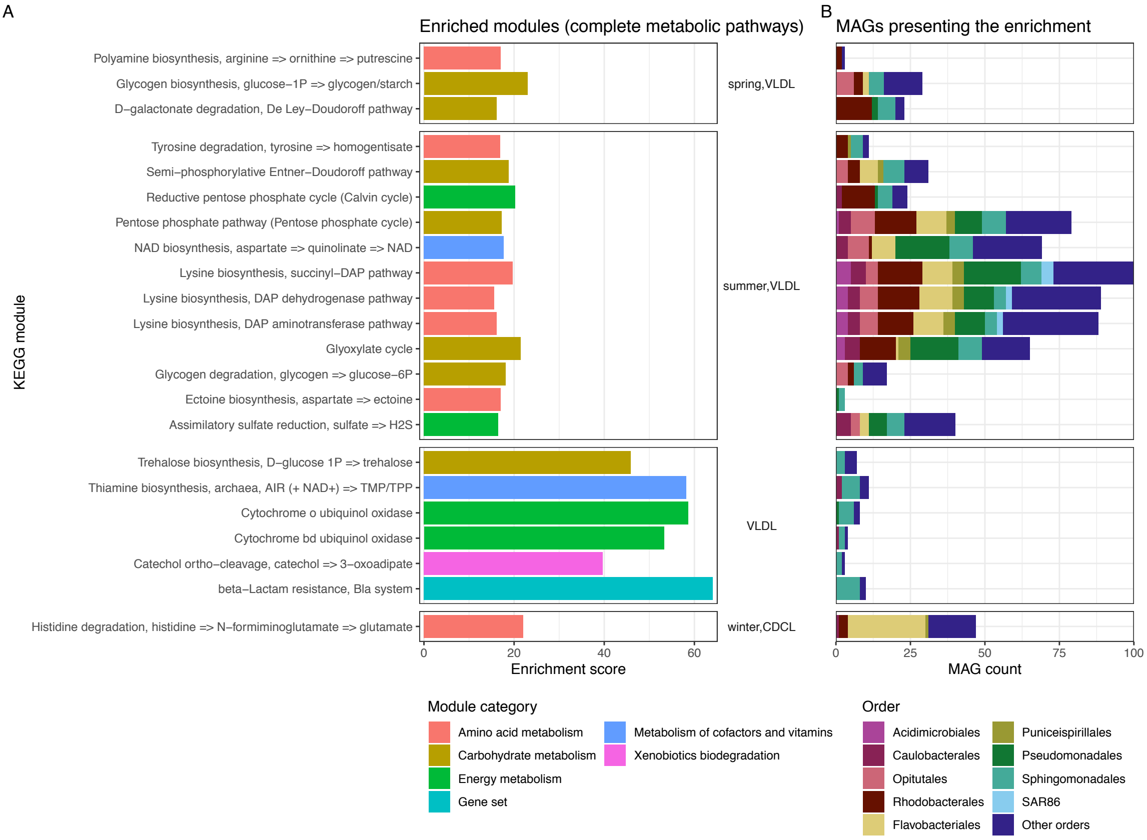


Figure 6: Functional enrichment of KEGG modules across the MAGs appearing in the virus-reduced, diluted and control treatments. A) The X axis presents the enrichment score (generalized linear model, $q \leq 0.01$) and the Y axis indicates the KEGG module definition. The facets indicate in which experiment or treatment the KEGG is enriched. The color code indicates the general module category. B) Taxonomy of the MAGs presenting the enriched module. The Y axis follows the order of panel A and the X axis specifies the number of MAGs. The bars are colored based on order taxonomy. Only the MAGs presenting ≥ 70% completion are shown.

Strain delineation

Finally, we evaluated the presence of different strains within each of the high-quality reconstructed MAGs (≥ 70% completion) to test the presence and distribution of strains among treatments, experiments, taxonomies and growth category (Figure 7). Microdiversity –the presence of intrapopulation genetic diversity– would indicate different adaptations to the conditions at the subspecies level. The metrics we used are based on average nucleotide identity (ANI) between the MAGs and the mapped reads. We used the consensus average ANI, which takes into account any mismatches (or alleles), and the population ANI, which only takes into account complete alleles and is therefore more conservative (see Materials and Methods). Summer and winter presented higher microdiversity than spring (Kruskal Wallis test, p ≤ 0.001, Figure 7A), whereas between the initial and final experimental times there was no difference. The most abundant MAGs in each experiment presented mostly a clonal compositon (>99% ANI; Supplementary Figure 4). In fact, only 12%

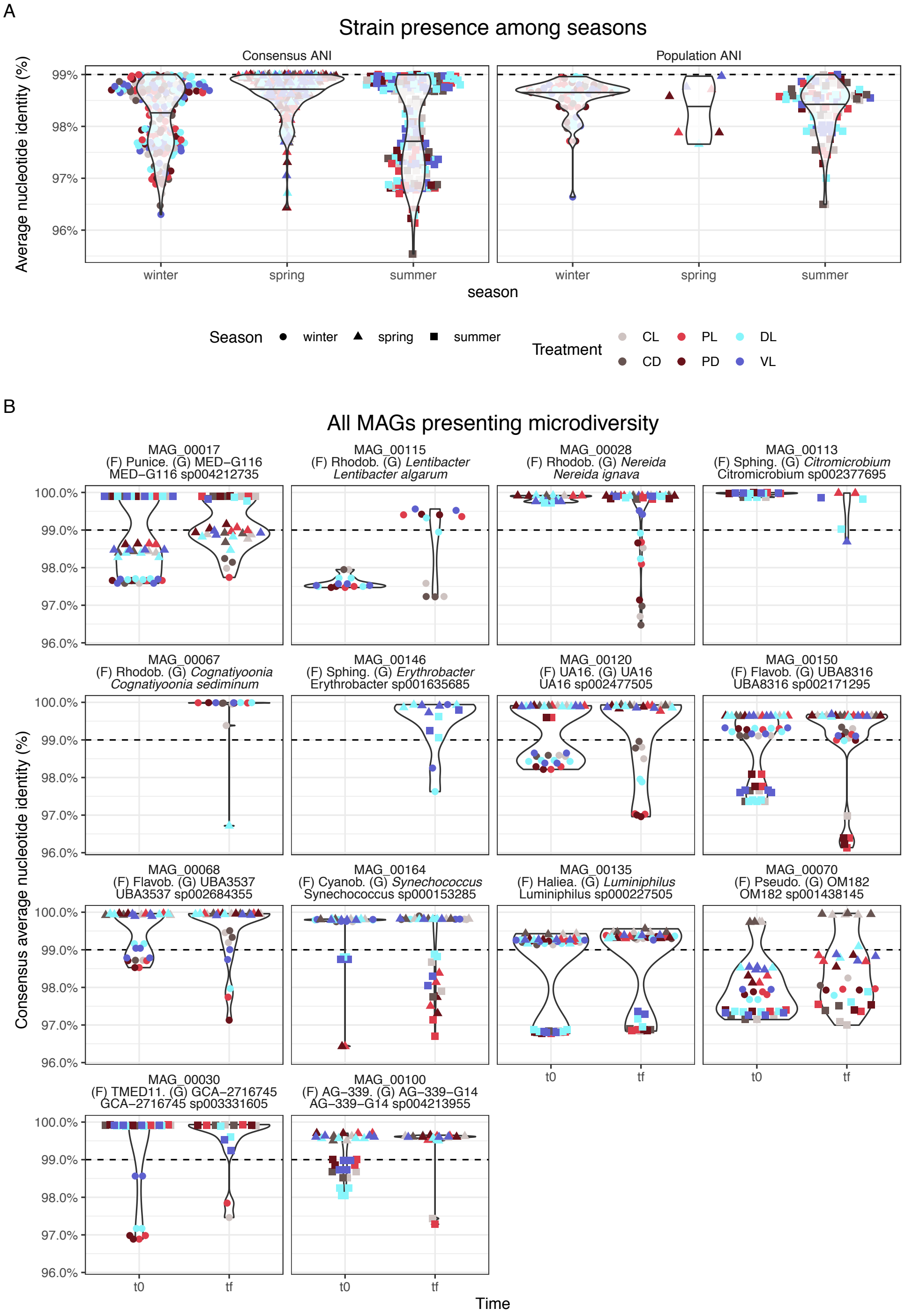


Figure 7: Microdiversity within the studied MAGs. A) Distribution of average nucleotide similitudes (with both consensus and population approaches, being the population approach more conservative) among all MAGs between seasons. The X axis shows the the time of the experiment, and the Y axis represents the average nucleotide identity (ANI) of each MAG. A violin plot is presented for each season, with the points colored by treatment and shaped by season. The dashed black lines indicate 99% identity, a threshold usually used for strain delineation. B) MAGs presenting microdiversity as shown by values below 99% identity (dashed black line). The X axis indicates the time of the experiment, and the Y axis the consensus ANI values. Each panel shows a MAG with its taxonomic assignation. The points are color-coded by treatment and shaped by season.

of the 175 MAGs presented differences between strains ≤ 99%. The exceptions to this trend were observed in MAGs of Nereida ignava, Lentibacter algarum, a MAG from the *Schleiferiaceae* family with an uncultured species (UBA10364), and an Sphingomonas MAG also without an assigned species (Supplementary Figure 4). The groups presenting the highest microdiversity belonged both to copiotrophs and oligotrophs (Figure 7B, Supplementary Table 3). Some of the MAGs presented clear differences in the ANI values in different seasons. MAGs 17, 28, 30, 68, 115, 120 presented the highest microdiversity during winter, whereas MAGs 70, 100, 135, 150 presented the highest values during summer. In most cases clonality was higher when the MAGs presented the highest abundances (Figure 8). That is, in each of these species, the highest abundance was reached when a single clonal population dominated, whereas at lower abundances there was more microdiversity. This microdiversity could be the result of multiple strains coexisting or that the most abundant strain was not the one from which we obtained the reference MAG. This general trend, however, had some exceptions. Luminiphilus MAG135 had high abundances both in spring and summer,

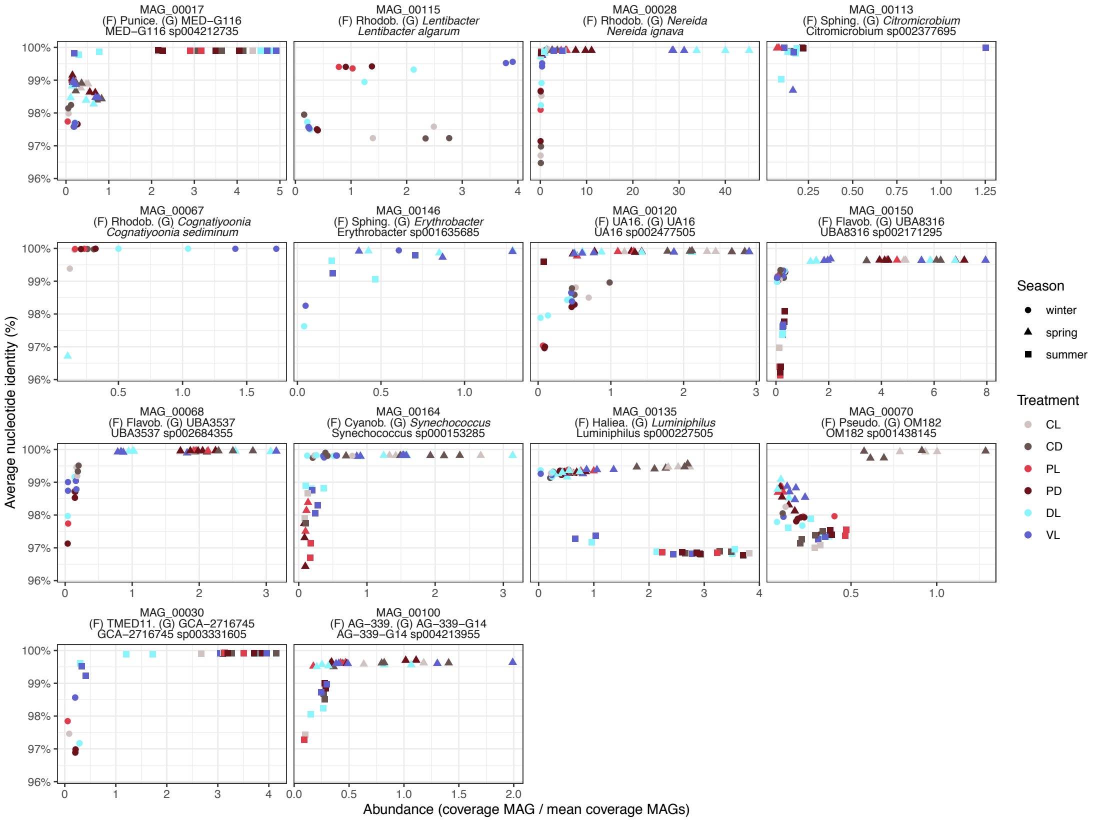


Figure 8: Relationship between the consensus average nucleotide identity (ANI, Y axis) and MAG abundance (X axis, coverage of each MAG divided by the mean coverage of all the MAGs). Each panel presents a single MAG, with each point representing a sample colored by treatment and shape by season.

but the high abundances in summer were linked to a high microdiversity (~97% ANI), whereas in spring it was mostly clonal (~99% ANI); *Lentibacter* MAG115 also reached a high abundance in the winter control (light and dark) treatment only, and kept the microdiversity level from the intial time (~97.5% ANI; Figure 8). Not all the MAGs presented variations in microdiversity. As an example, of the 4 *Luminiphilus* MAGs, only MAG135 had different strains, whereas the others were clonal in all the experiments. Overall, these results indicate that, as a general rule, when a specific organism dominates in the system, at the strain level clonality prevails.

4.4 Discussion

We obtained metagenomes from three experiments at different times of the year to disentangle the effects of predation by grazers, viral mortality, nutrient limitation and light, factors that interplay in nature influencing community selection processes. From these metagenomes, we recovered 262 MAGs for which we described their abundances after each treatment, the functional genomic potential, and we differentiated whether each growing MAG included multiple similar strains or not. Given that our experiments entailed manipulation of the seawater, a containment effect could have occured (Ferrera *et al.*, 2011; Baltar *et al.*, 2015; Ionescu *et al.*, 2015; Haro-Moreno *et al.*, 2019), yet we used large bottles to minimize severe changes due to water manipulation. The fact that our initial times clustered together in most cases indicates that the effect was not modifying substantially the community structure. The results presented advance beyond our previous analysis focused on the growth rates of the bulk community and those of large phylogenetic groups (Sánchez *et al.*, 2020), offering much more taxonomic detail. Yet, some of the main groups, particularly the Pelagibacterales order –present at the start of our experiments (Sánchez *et al.*, 2020)– were not recovered as MAGs due to their known high genomic microdiversity that difficults correctly assembling them (Haro-Moreno *et al.*, 2020). In these experiments, however, Pelagibacterales had relatively low growth rates (0 to 0.9 d^{-1}, Sánchez *et al.*, 2020), as typically occurs with these manipulations (Ferrera *et al.*, 2011; Sánchez *et al.* 2017; Teira *et al.*, 2019) and thus, since we aimed at characterizing those organisms growing under the experimental conditions assayed, this methodological constrain should impact little our observations.

Both richness and sample dissimilarity were highly influenced by season in our experiments. Richness was highly variable between treatments, being the values in the control and predator-reduced treatments –both in light and dark conditions– higher than in the virus-reduced and diluted treatments. These results contrast to what was observed by Teira *et al.* (2019) in similar experiments in the tropical and subtropical open ocean, that found a negative relationship between richness and light, and a positive one with predation. It is possible that light and predation regulate differently community diversity in the open ocean than in a coastal site. Likewise, differences in in situ temprerature (the open ocean experiments ranged between 21.7 and 28.7°C) may have also influenced the contrasting results. In terms of community compostion, the treatments had a large influence, shown by the higher number of indicator species in the control treatment, followed by the diluted and virus-reduced treatments and finally the predator-reduced condition. The control treatment included mostly oligotrophic groups that did not grow in the experiments, and specific groups such as Verrucomicrobia and Plantomycetota MAGs, that are often found in the particle attached fraction of plankton (Mestre *et al.*, 2020) and that could have been removed by the filtration in the other treatments (the control treatment was untreated). Light selected also specific MAGs in the control treatment, which became more resilient than in the dark condition (i.e. they did not decrease during the experiment). Most of these MAGs were either photosynthetic, such as *Prochlorococcus*, or interestingly had proteorhodopsin, such as *Woeseiaceae*, *Marinoscillaceae* and SAR86 (Supplementary Figure 3). Proteorhodopsin has been hypothesized to help bacteria avoid starvation (Gómez-Consarnau *et al.*, 2010).

The predator-reduced treatment favoured mainly *Flavobacteriaceae* MAGs. This result contrasts with a recent analysis using a similar experimental approach to remove predators based on amplicon sequencing (Fecskeová *et al.*, 2021) that did not find high growth rates for this group. Within these *Flavobacteriaceae* MAGs, we observed growth of different species at each season in the predator-reduced treatment (Figure 3). The preferences of these *Flavobacteriaceae* species complement the bulk growth rate calculated by FISH in our previous analysis (Sánchez *et al.*, 2020), that indicate that the large growth rate observed for this group can be explained by different species presenting possibly a differentiated trait repertoire adapted to each season. Seasonality within *Flavobacteriaceae* has been described before particularly in the Mediterrenean Sea (Díez-Vives *et al.*, 2019; Mena *et al.*, 2020; Auladell *et al.*, 2021), and our results suggest that predation is a main factor regulating the presence of these species. A similar result was observed with *Rhodobacteraceae* MAGs, most of them indicator species of the virus-reduced and diluted conditions and sometimes growing in the predator-reduced treatment. Generally, the presence of many specific lineages when grazing is reduced agrees with recent results (Fecskeová *et al.*, 2021), as opposed to other experiments also performed in Blanes Bay that did not show this predator-specific effect (Baltar *et al.*, 2016).

Other abundant groups presented growth indistinctly of season, such the *Alteromonadaceae* and *Sphingomonadaceae* MAGs growing in the virus-reduced and diluted treatments. The *Sphingomonadaceae* taxa are a physiologically versatile with a high genomic variability that are adapted to live in oligotrophic environments (Aylward *et al.*, 2013). Their growth in the diluted treatment is opposite to what was observed in Teira *et al.* (2019), which found a negative effect of dilution on this group. A possible explanation could be that most *Sphingomonadaceae* species are generally adapted to oligotrophic oceanic waters such as the open ocean sites studied by Teira *et al.* (2019), whereas in coastal waters the species present could have a more copiotrophic lifestyle (Figure 3). The growth of multiple species of *Alteromonadaceae* and *Sphingomonadaceae* in the same treatments could imply that these MAGs share similar traits and were favoured under similar conditions. This contrasts with the responses of *Flavobacteriaceae* and Rhodobacteraceae, that presented MAGs growing selectively in one season only and in a specific treatment. Selection of Alteromonadaceae and Sphingomonadaceae species therefore is likely to happen at a higher taxonomical level (i.e. family level), coexisting due to niche partitioning processes (Haro-Moreno *et al.*, 2019; Langenheder and Lindström, 2019). Overall, the differences observed in the diluted treatment (a treatment that overcomes substrate limitation) in the winter and summer experiments triggered a large change in community composition, whereas in spring this change was less obvious. Summer in Blanes Bay is severely limited by inorganic nutrients (Pinhassi *et al.*, 2006), which could explain the shift induced by the higher nutrient availability. The upshift in winter was however unexpected, given that inorganic nutrients are maximal during this season (Table 1). A possible explanation is that during winter all prokaryotic groups compete for the available nutrients but the relaxation of nutrient stress allows the least competitive groups to also develop.

On a functional basis, we inspected the presence of key biogeochemically relevant genes to understand how treatments and seasonality might potentially affect these functional groups (Figure 5, Supplementary Figure 3). From the subset of studied genes, we observed that pitA was enriched in the diluted and virus-reduced treatment, but overall, many genes were present in specific taxonomies rather than in specific treatments. The AAPs, being in the long-term time series system mainly dominated by *Haliaceae* and *Rhodobacteraceae* groups (Auladell *et al.*, 2019), differed in their responses to treatments. The high growth that we usually observe through direct counts (Ferrera *et al.*, 2017; Sánchez *et al.*, 2017) could be linked to the *Rhodobacteraceae* copiotrophs rather than to the oligotrophic *Haliaceae* AAPs, since these are the ones presenting a high FC (Figure 5), as also found in a recent study (Fecskeová *et al.*, 2021). Additionally, the enrichment analysis of specific modules identified specially the differentiated metabolism from the *Sphingomonadaceae* family (Figure 6). Almost all of the retrieved *Sphingomonadaceae* MAGs contained the KEGG module for thiamine production, special terminal oxidases and production of beta-lactamases. The presence of the complete pathway for thiamine production could point to a free-living lifestyle, allowing the group to present short growth pulses without co-dependences (auxotrophies) with other organisms such as the ones observed by oligotrophs (Johnson *et al.*, 2020). A surprising trait was the production of beta-lactamases, known in biology for their antibiotic resistance properties. Recent results have linked the production of these enzymes to the disruption of the quorum sensing of other bacterial groups (Selleck *et al.*, 2020), an strategy that could be useful to exclude possible competitors for nutrients. Given that these functional results only considered complete enriched modules as identified in the KEGG database, they must be considered as a first general overview.

Further analyses comparing the genomic repertoire within families would allow to better understand differences between treatments and seasons.

Finally, the strain analysis within the high-quality MAGs allowed us to have an idea of how species microdiversity is distributed among different treatments and seasons (Figure 7). This information is crucial to determine if different strains differ in their abundance patterns, between treatments and seasons. The lack of strain differentiation in spring compared to the winter and summer experiments was unexpected. Compiling some of the results of this study, the spring season had similar richness values between the final and initial times (Figure 1), changed the least between treatments when compared to the other experiments (Figure 2), and did not present strain diversity (Figure 7). Similarly, bulk community estimates for the spring experiment also showed low production and growth (Sánchez *et al.*, 2020). A possible explanation to that observation could be that during spring the lack of nutrient limitation coupled to the development of phytoplankton groups during winter might have created close-to-optimal conditions for the development of heterotrophic bacterial blooms with few clonal populations dominating the system. We in fact observed a relationship between microdiversity and abundance, with clonality at the highest MAG abundance, and microdiversity at lower abundance values in most cases (Figure 8). From a genomic population perspective, a single clonal genome is usually selected when there are strong selective pressures (Haro-Moreno *et al.*, 2019). The fact that only one strain dominanted in the final time for multiple treatments and seasons could imply that competitive exclusion could be playing an important role at the strain level (Cohan, 2017). A result to be further explored is why for some MAGs (such as MAG17, 70) the ANI values presented different microdiversity levels linked to specific seasons. As an example, MAG17 presented an ANI of 100% during summer, 99% in spring and ~98% in winter (Figure 7). This pattern is specific for each MAG, with the high microdiversity not always present in the winter season, but rather being driven by each MAG seasonal preference. Further analyses focusing on these specific microdiversity values will shed light on how these ecological selective processes operate.

Overall, our study contributes to understand how biotic factors, such as the presence of predators and viruses, and abiotic conditions, such as nutrient limitation and light, influence microbial species at different seasons of the year. Specifically, we observed that different *Flavobacteriaceae* MAGs were indicative species for the predator-reduced treatment depending on the season, and the *Rhodobacteraceae* species, indicative mostly for the diluted and virus-reduced treatments. Further, the metagenomic approach has allowed to link specific genetic repertoires to some conditions and disentangle the variability among similar strains, indicating that clonality dominates when a species becomes more abundant. Further analyses could focus on the pangenome of the summer *Flavobacteriaceae* groups, less known that the typical bloom-responding ones, and the differentiation of the genetic repertoire between strains of the same species.

4.5 Acknowledgements

We thank all the people involved in operating the BBMO and those involved in performing the experiments, especially Clara Cardelús and Captain Anselm for facilitating the sampling. We also thank the MarBits Bioinformatics platform at the Institut de Ciències del Mar, in particular Pablo Sánchez, for computing support. This research was funded by grants REMEI (CTM2015-70340-R), MIAU (RTI2018-101025-B-I00) and ECLIPSE (PID2019-110128RB-100) from the Spanish Ministry of Science and Innovation. Additional support was obtained through the 'Severo Ochoa Centre of Excellence' accreditation (CEX2019-000924-S) from the Spanish government, and by the Grup de Recerca de la Generalitat de Catalunya (2017SGR/1568). AA was supported by an FPI grant associated to the REMEI project.

4.6 References

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

Acinas, S.G., Sánchez, P., Salazar, G., Cornejo-Castillo, F.M., Sebastián, M., Logares, R., *et al.* (2021) Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun Biol* 4: 1–15.

Alonso-Sáez, L., Morán, X.A.G., and González, J.M. (2020) Transcriptional Patterns of Biogeochemically Relevant Marker Genes by Temperate Marine Bacteria. *Front Microbiol* 11: 465.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36: 2251–2252.

Auladell, A., Barberán, A., Logares, R., Garcés, E., Gasol, J.M., and Ferrera, I. (2021) Seasonal niche differentiation among closely related marine bacteria. *ISME J.*

Auladell, A., Sánchez, P., Sánchez, O., Gasol, J.M., and Ferrera, I. (2019) Long-term seasonal and interannual variability of marine aerobic anoxygenic photoheterotrophic bacteria. *ISME J* 13: 1975–1987.

Aylward, F.O., McDonald, B.R., Adams, S.M., Valenzuela, A., Schmidt, R.A., Goodwin, L.A., *et al.* (2013) Comparison of 26 Sphingomonad Genomes Reveals Diverse Environmental Adaptations and Biodegradative Capabilities. *Appl Environ Microbiol* 79: 3724–3733.

Baltar, F., Palovaara, J., Unrein, F., Catala, P., Horňák, K., Šimek, K., *et al.* (2016) Marine bacterial community structure resilience to changes in protist predation under phytoplankton bloom conditions. *ISME J* 10: 568–581.

Baltar, F., Palovaara, J., Vila-Costa, M., Salazar, G., Calvo, E., Pelejero, C., *et al.* (2015) Response of rare, common and abundant bacterioplankton to anthropogenic perturbations in a Mediterranean coastal site. *FEMS Microbiol Ecol* 91: fiv058.

Bender, R.A. (2012) Regulation of the Histidine Utilization (Hut) System in Bacteria. *Microbiol Mol Biol Rev* 76: 565–584.

Bergelson, J., Kreitman, M., Petrov, D.A., Sanchez, A., and Tikhonov, M. (2021) Functional biology in its natural context: A search for emergent simplicity. *eLife* 10: e67646.

Boras, J.A., Sala, M.M., Vázquez-Domínguez, E., Weinbauer, M.G., and Vaqué, D. (2009) Annual changes of bacterial mortality due to viruses and protists in an oligotrophic coastal environment (NW Mediterranean). *Envir Microbiol* 11: 1181–1193.

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., *et al.* (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523: 208–211.

Bunse, C. and Pinhassi, J. (2017) Marine Bacterioplankton Seasonal Succession Dynamics. *Trends Microbiol* 25: 1–12.

Cohan, F.M. (2017) Transmission in the Origins of Bacterial Diversity, From Ecotypes to Phyla. *Microbiol Spec* 5: 5.

Coon, S.L., Fuqua, W.C., and Weiner, R.M. (1994) Homogentisic Acid is the Product of MelA, Which Mediates Melanogenesis in the Marine Bacterium Shewanella colwelliana Dt. *Appl Environ Microbiol* 60: 5.

Delmont, T.O. and Eren, A.M. (2018) Linking Pangenomes and Metagenomes: The Prochlorococcus Metapangenome. *PeerJ* 6: e4320.

Delmont, T.O., Kiefl, E., Kilinc, O., Esen, O.C., Uysal, I., Rappé, M.S., *et al.* (2019) Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *eLife* 8: e46497.

Díez-Vives, C., Nielsen, S., Sánchez, P., Palenzuela, O., Ferrera, I., Sebastián, M., *et al.* (2019) Delineation of ecologically distinct units of marine Bacteroidetes in the Northwestern Mediterranean Sea. *Mol Ecol* 28: 2846–2859.

Eddy, S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* 7: e1002195.

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

Eren, A.M., Vineis, J.H., Morrison, H.G., and Sogin, M.L. (2013) A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. *PLoS ONE* 8: e66643.

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

Fecskeová, L.K., Piwosz, K., Šantić, D., Šestanović, S., Tomaš, A.V., Hanusová, M., *et al.* (2021) Lineage-Specific Growth Curves Document Large Differences in Response of Individual Groups of Marine Bacteria to the Top-Down and Bottom-Up Controls. *mSystems* 6: e00934-21.

Ferrera, I., Gasol, J.M., Sebastián, M., Hojerová, E., and Kobížek, M. (2011) Comparison of growth rates of aerobic anoxygenic phototrophic bacteria and other bacterioplankton groups in coastal mediterranean waters. *Appl Environ Microbiol* 77: 7451–7458.

Ferrera, I., Sánchez, O., Kolářová, E., Koblížek, M., and Gasol, J.M. (2017) Light enhances the growth rates of natural populations of aerobic anoxygenic phototrophic bacteria. *ISME J* 11: 2391–2393.

Ficko-Blean, E., Préchoux, A., Thomas, F., Rochat, T., Larocque, R., Zhu, Y., *et al.* (2017) Carrageenan catabolism is encoded by a complex regulon in marine heterotrophic bacteria. *Nat Commun* 8: 1685.

Gasol, J.M., Cardelús, C., Morán, X.A.G., Balagué, V., Forn, I., Marrasé, C., *et al.* (2016) Seasonal patterns in phytoplankton photosynthetic parameters and primary production at a coastal NW Mediterranean site. *Sci Mar* 80S1: 63–77.

Gasol, J.M. and Morán, X.A.G. (2016) Flow Cytometric Determination of Microbial Abundances and Its Use to Obtain Indices of Community Structure and Relative Activity. In Hydrocarbon and Lipid Microbiology Protocols: Single-Cell and Single-Molecule Methods. Springer Protocols Handbooks. McGenity, T.J., Timmis, K.N., and Nogales, B. (eds). Berlin, Heidelberg: Springer, pp. 159–187.

Giner, C.R., Balagué, V., Krabberød, A.K., Ferrera, I., Reñé, A., Garcés, E., *et al.* (2019) Quantifying long-term recurrence in planktonic microbial eukaryotes. *Mol Ecol* 28: 923–935.

Giovannoni, S.J., Cameron Thrash, J., and Temperton, B. (2014) Implications of streamlining theory for microbial ecology. *ISME J* 8: 1553–1565.

Gloor, G.B., Wu, J.R., Pawlowsky-Glahn, V., and Egozcue, J.J. (2016) It's all relative: analyzing microbiome data as compositions. *Ann Epidemiol* 26: 322–329.

Goldford, J.E., Lu, N., Baji, D., Sanchez-Gorostiaga, A., Segrè, D., Mehta, P., and Sanchez, A. (2018) Emergent simplicity in microbial community assembly. 7.

Gómez-Consarnau, L., Akram, N., Lindell, K., Pedersen, A., Neutze, R., Milton, D.L., *et al.* (2010) Proteorhodopsin Phototrophy Promotes Survival of Marine Bacteria during Starvation. *PLoS Biol* 8: e1000358.

Gong, X., Garcia-Robledo, E., Lund, M.B., Lehner, P., Borisov, S.M., Klimant, I., *et al.* (2018) Gene expression of terminal oxidases in two marine bacterial strains exposed to nanomolar oxygen concentrations. *FEMS Microbiol Ecol* 94: fiy072.

Goyal, A., Bittleston, L.S., Leventhal, G.E., Lu, L., and Cordero, O.X. (2021) Interactions between strains govern the eco-evolutionary dynamics of microbial communities, bioRxiv.

Graham, E.D., Heidelberg, J.F., and Tully, B.J. (2018) Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J* 12: 1861.

Grasshoff, K., Ehrhardt, M., and Kremling, K. (1983) Methods of seawater analysis, 2nd ed. Verlag Chemie, Weinheim.

Grossart, H., Massana, R., McMahon, K.D., and Walsh, D.A. (2020) Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnol Oceanogr* 65: S2–S20.

Haro-Moreno, J.M., Rodriguez-Valera, F., and López-Pérez, M. (2019) Prokaryotic Population Dynamics and Viral Predation in a Marine Succession Experiment Using Metagenomics. *Front Microbiol* 10: 2926.

Haro-Moreno, J.M., Rodriguez-Valera, F., Rosselli, R., Martinez-Hernandez, F., Roda-Garcia, J.J., Gomez, M.L., *et al.* (2020) Ecogenomics of the SAR11 clade. *Environ Microbiol* 22: 1748–1763.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119.

Ionescu, D., Bizic-Ionescu, M., Khalili, A., Malekmohammadi, R., Morad, M.R., de Beer, D., and Grossart, H.-P. (2015) A new tool for long-term studies of POM-bacteria interactions: overcoming the century-old Bottle Effect. *Sci Rep* 5: 14706.

Johnson, J.S., Spakowicz, D.J., Hong, B.-Y., Petersen, L.M., Demkowicz, P., Chen, L., *et al.* (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 10: 5029.

Johnson, W.M., Alexander, H., Bier, R.L., Miller, D.R., Muscarella, M.E., Pitz, K.J., and Smith, H. (2020) Auxotrophic interactions: a stabilizing attribute of aquatic microbial communities? F*EMS Microbiol Ecol* 96: fiaa115.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7: e7359.

Kirchman, D.L. (2016) Growth Rates of Microbes in the Oceans. *Ann Rev Mar Sci* 8: 285–309.

Koblízek, M., Masín, M., Ras, J., Poulton, A.J., and Prásil, O. (2007) Rapid growth rates of aerobic anoxygenic phototrophs in the ocean. *Environ Microbiol* 9: 2401–6.

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., and Galand, P.E. (2018) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *ISME J* 388–401.

Langenheder, S. and Lindström, E.S. (2019) Factors influencing aquatic and terrestrial bacterial community assembly. *Environ Microbiol Rep* 11: 306–315.

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

Lemonnier, C., Perennou, M., Eveillard, D., Fernandez-Guerra, A., Leynaert, A., Marié, L., *et al.* (2020) Linking Spatial and Temporal Dynamic of Bacterioplankton Communities With Ecological Strategies Across a Coastal Frontal Area. *Front Mar Sci* 7: 376.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Logares, R., Deutschmann, I.M., Junger, P.C., Giner, C.R., Krabberød, A.K., Schmidt, T.S.B., *et al.* (2020) Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome* 8: 55.

Massana, R., Murray, A.E., Preston, C.M., and Delong, E.F. (1997) Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Appl Environ Microbiol* 63: 50–56.

McLaren, M.R. and Callahan, B.J. (2018) In Nature, There Is Only Diversity. mBio 9: e02149-17.

Mena, C., Reglero, P., Balbín, R., Martín, M., Santiago, R., and Sintes, E. (2020) Seasonal Niche Partitioning of Surface Temperate Open Ocean Prokaryotic Communities. *Front Microbiol* 11: 1749.

Mestre, M., Höfer, J., Sala, M.M., and Gasol, J.M. (2020) Seasonal Variation of Bacterial Diversity Along the Marine Particulate Matter Continuum. *Front Microbiol* 11: 1590.

Milanese, A., Mende, D.R., Paoli, L., Salazar, G., Ruscheweyh, H.-J., Cuenca, M., *et al.* (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* 10: 1014.

Minoche, A.E., Dohm, J.C., and Himmelbauer, H. (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol* 12: R112.

Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., *et al.* (2021) Sustainable data analysis with Snakemake. F1000Research

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., *et al.* (2013) Package 'vegan.' Community ecology package

Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017) dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 11: 2864–2868.

Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J., and Banfield, J.F. (2021) inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* 39: 727–736.

Olm, M.R., Crits-Christoph, A., Diamond, S., Lavy, A., Matheus Carnevali, P.B., and Banfield, J.F. (2020) Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* 5: e00731-19

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

Pinhassi, J., Gómez-Consarnau, L., Alonso-Sáez, L., Sala, M., Vidal, M., Pedrós-Alió, C., and Gasol, J. (2006) Seasonal changes in bacterioplankton nutrient limitation and their effects on bacterial community composition in the NW Mediterranean Sea. *Aquat Microb Ecol* 44: 241–252.

Porter, K.G. and Feig, Y.S. (1980) The use of DAPI for identifying and counting aquatic microflora1. *Limnol Oceanogr* 25: 943–948.

R Core Team (2014) R: A language and environment for statistical computing.

Rodríguez-Valera, F. (2002) Approaches to prokaryotic biodiversity: a population genetics perspective: Population genetics and prokaryotic biodiversity. *Environ Microbiol* 4: 628–633.

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

Sánchez, O., Ferrera, I., Mabrito, I., Gazulla, C.R., Sebastián, M., Auladell, A., *et al.* (2020) Seasonal impact of grazing, viral mortality, resource availability and light on the group-specific growth rates of coastal Mediterranean bacterioplankton. *Sci Rep* 10: 19773.

Sánchez, O., Koblížek, M., Gasol, J.M., and Ferrera, I. (2017) Effects of grazing, phosphorus and light on the growth rates of major bacterioplankton taxa in the coastal NW Mediterranean: Growth rates of bacterioplankton. *Environ Microbiol Rep* 9: 300–309.

Schloss, P.D. (2021) Amplicon Sequence Variants Artificially Split Bacterial Genomes into Separate Clusters. *mSphere* 6: e00191-21.

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

Sekar, K., Linker, S.M., Nguyen, J., Grünhagen, A., Stocker, R., and Sauer, U. (2020) Bacterial Glycogen Provides Short-Term Benefits in Changing Environments. *Appl Environ Microbiol* 86: e00049-20.

Selleck, C., Pedroso, M.M., Wilson, L., Krco, S., Knaven, E.G., Miraula, M., *et al.* (2020) Structure and mechanism of potent bifunctional β-lactam- and homoserine lactone-degrading enzymes from marine microorganisms. *Sci Rep* 10: 12882.

Shaiber, A., Willis, A.D., Delmont, T.O., Roux, S., Chen, L.-X., Schmid, A.C., *et al.* (2020) Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol* 21: 292.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 14.

Teira, E., Logares, R., Gutiérrez-Barral, A., Ferrera, I., Varela, M.M., Morán, X.A.G., and Gasol, J.M. (2019) Impact of grazing, resource availability and light on prokaryotic growth and diversity in the oligotrophic surface global ocean. *Environ Microbiol* 21: 1482–1496.

Unrein, F., Massana, R., Alonso-Sáez, L., and Gasol, J.M. (2007) Significant year-round effect of small mixotrophic flagellates on bacterioplankton in an oligotrophic coastal system. *Limnol Oceanogr* 52: 456–469.

VanInsberghe, D., Arevalo, P., Chien, D., and Polz, M.F. (2020) How can microbial population genomics inform community ecology? *Phil Trans R Soc B* 375: 20190253.

Vellend, M. (2010) Conceptual Synthesis in Community Ecology. *Q Rev Biol* 85: 183–206.

Weissman, J.L., Hou, S., and Fuhrman, J.A. (2021) Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proc Natl Acad Sci USA* 118: e2016810118.

Wickham, H. (2016) ggplot2: Elegant graphics for data analysis, Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., *et al.* (2019) Welcome to the tidyverse. *J Open Source Soft* 4: 1686.

Widderich, N., Höppner, A., Pittelkow, M., Heider, J., Smits, S.H.J., and Bremer, E. (2014) Biochemical Properties of Ectoine Hydroxylases from Extremophiles and Their Wider Taxonomic Distribution among Microorganisms. *PLoS ONE* 9: e93809.

Yokokawa, T., Nagata, T., Cottrell, M.T., and Kirchman, D.L. (2004) Growth rate of the major phylogenetic bacterial groups in the Delaware estuary. *Limnol Oceanogr* 49: 1620–1629.

Zinger, L., Amaral-Zettler, L.A., Fuhrman, J.A., Horner-Devine, M.C., Huse, S.M., Welch, D.B.M., *et al.* (2011) Global Patterns of Bacterial Beta-Diversity in Seafloor and Seawater Ecosystems. *PLoS ONE* 6: e24570.

4.7 Supplementary figures

link Supplementary Figure 1

Figure S1: Heatmap displaying the presence of each MAG in the seasonal experiments. The presence is marked in grey and the absence in white. The columns indicate the seasonal experiments.

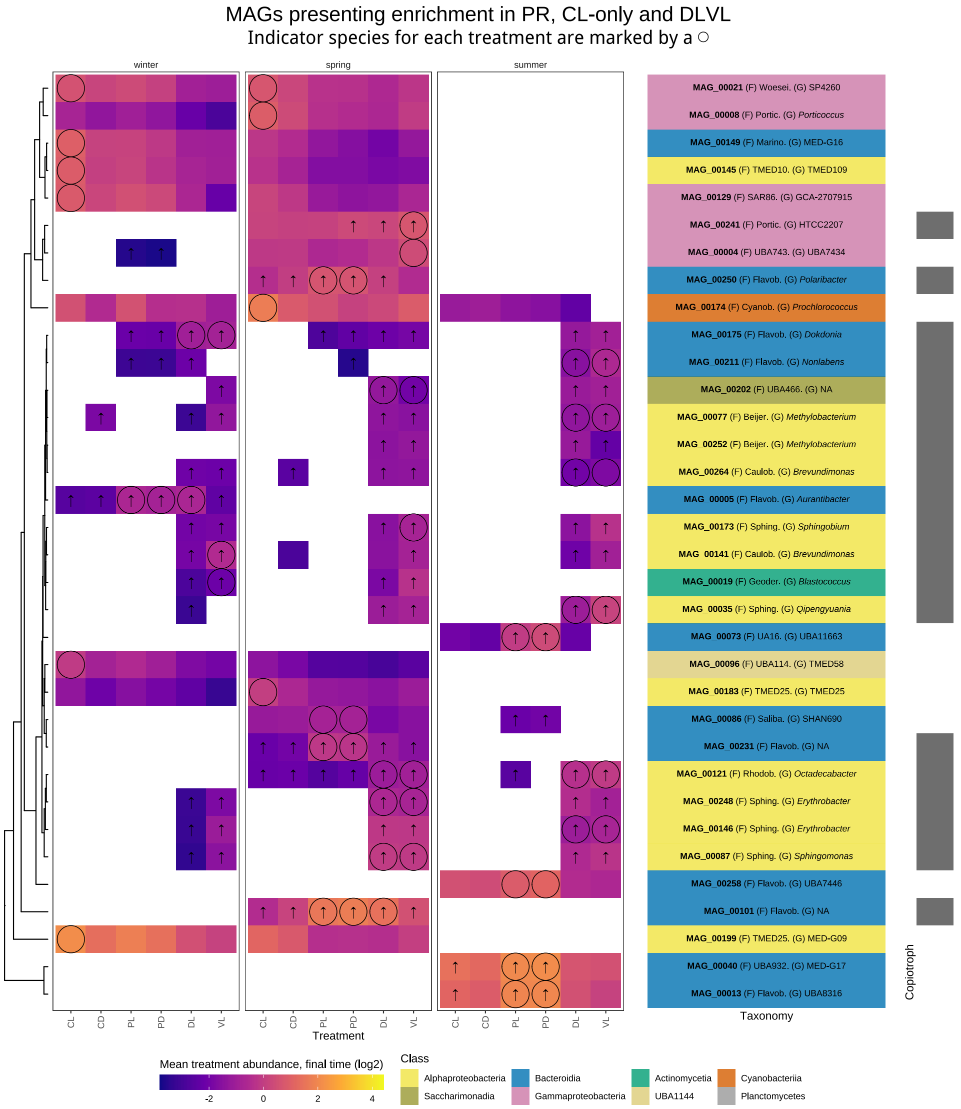


Figure S2: Heatmap presenting the MAGs that are indicator species in the predator-reduced, control light and virus-reduced and diluted treatment (not including those already in Figure 3). The abundance represents the mean MAG coverage divided by the mean coverage of all MAGs in the sample, and is displayed in a log2 scale in the plot. The columns on the right indicate taxonomy (colored by class) and growth behavior (copiotrophy in grey, oligotrophy in white). The MAGs are clustered based on their abundance. A diamond indicates that the MAG is one of the top 5 in panel A, and a circle indicates that the MAG is an indicator species for that season and treatment. The arrow indicates a fold change between the initial and final time higher or equal than 1.5.

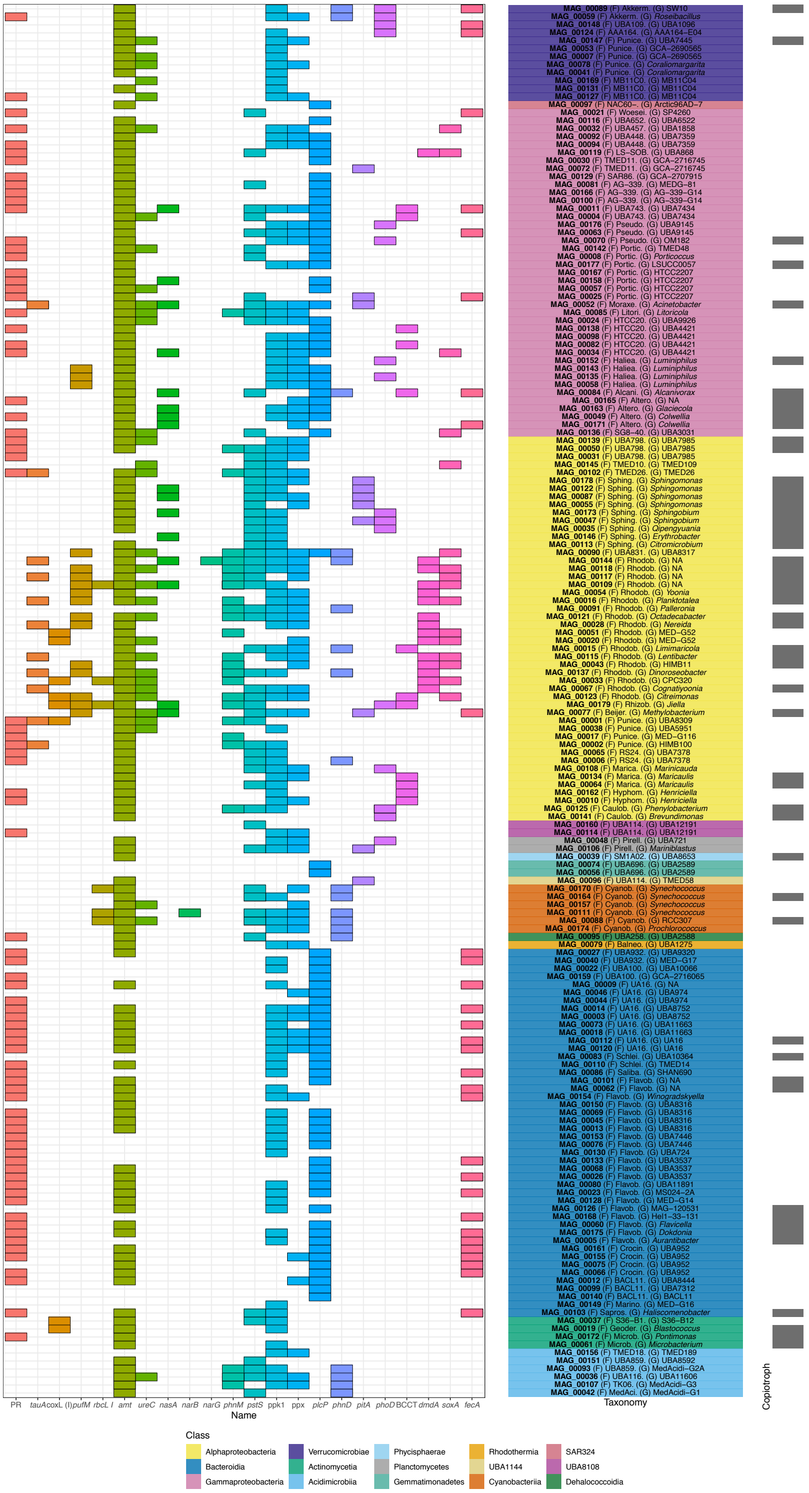


Figure S3: Presence of key relevant genes in all MAGs with ≥ 70% completion indicated by the heatmap. The panels on the right indicate the taxonomy (colored by the class), and the growth style (copiotroph in grey, white for oligotrophy). The MAGs are clustered based on taxonomy.

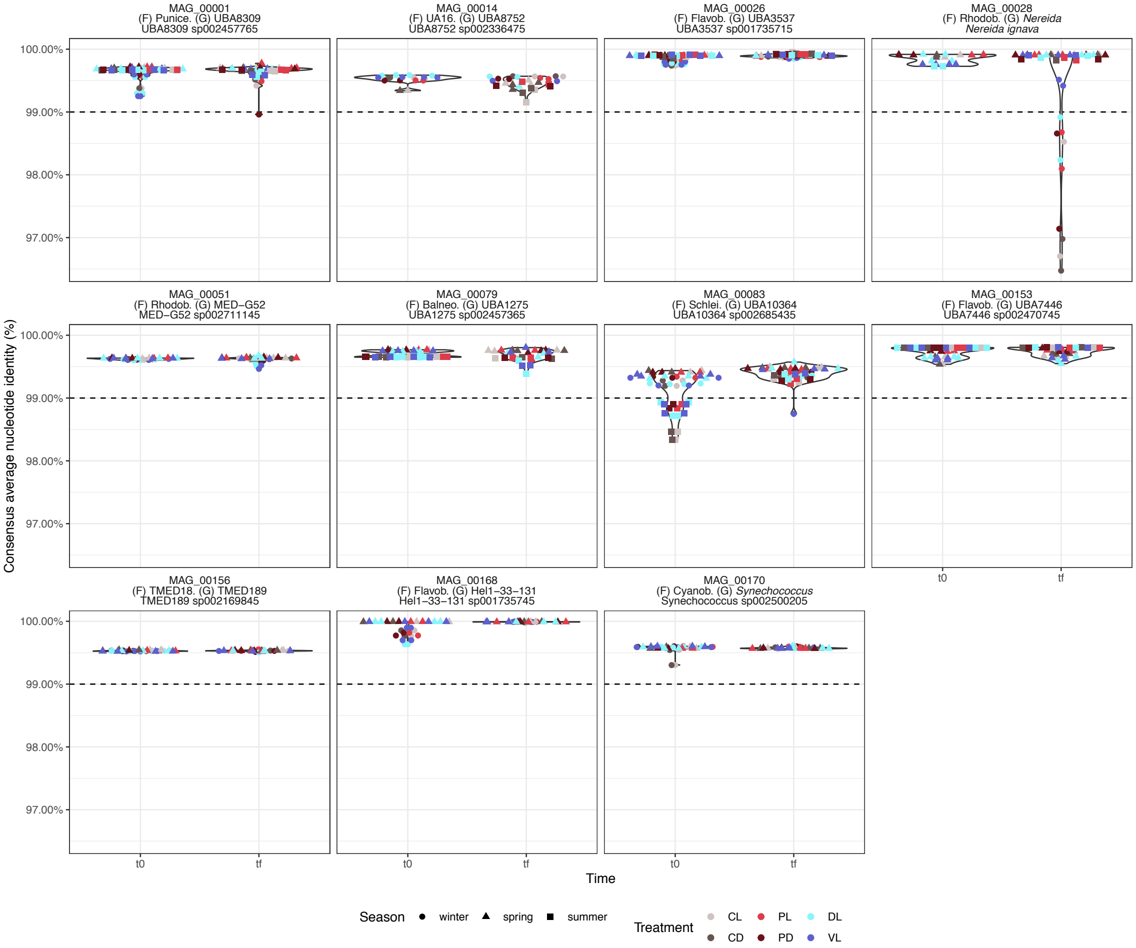


Figure S4: Average nucleotide identity of the top 5 most abundant MAGs presenting ≥ 70% completion. The X axis is the time of the experiment, and the Y axis the population ANI values. Each panel shows a MAG with taxonomic assignation. The points are colored by treatment and shaped by season. The dashed black lines indicate 99% identity, a possible threshold for strain delineation.

4.8 Supplementary tables

link Supplementary Table 1

Supplementary Table 1: Taxonomy and genomic statistics for each MAG. Columns named t_domain to t_species indicates the taxonomy based in the Genome Taxonomy Database; total length equals to the number of base pairs in the MAG; num_contigs indicate the number of contigs conforming the MAG; N50 the length of the contig when a 50% genome completion is reached by summing the contigs from the longest to the shortest; GC_content the % of GC of the MAG genome; Percent completion and percent redundancy indicate the results of the prediction of genome completion and how contaminated could be the genome respectively based on checkM.

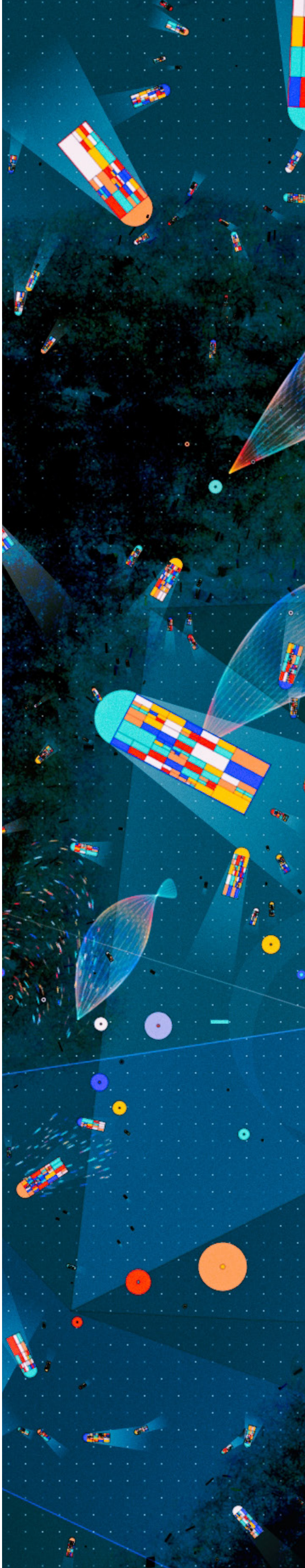
link Supplementary Table 2

Supplementary Table 2: Summary of the abundance distribution for each MAG in the experiments. The median and maximum abundance in all experiments is shown, together with the median and maximum fold change. The indicator species column shows the treatments for which this MAG is indicative (with the specific season in parenthesis).

link Supplementary Table 3

Supplementary Table 3: Predicted maximal growth rate for each MAG. The values are based in the gRodon R package, which estimates maximal growth rates of prokaryotic organisms from genome-wide codon usage statistics. The columns are the estimated maximal growth rate (in hours) together with the confidence interval (UpperCI and LowerCI), and the growth strategy based on whether the estimated growth rate is over 5, indicating a basically oligotrophic growth strategy, or below 5, indicating copiotrophic growth (see Weissman *et al.*, 2021).

GENERAL DISCUSSION



This thesis aimed at gaining insights into the seasonal patterns of the marine microbial community in the Blanes Bay Microbial Observatory (BBMO), from a taxonomic (**Chapter I**), functional (**Chapter II** and **III**) and genomic (**Chapter IV**) perspective. The thesis has been possible thanks to the notable efforts made to maintain for years this time series and collect monthly DNA samples and their environmental associated variables. These efforts allow to later conduct long-term microbial observation, which is of high relevance to understand the temporal dynamics of marine microbial communities in present and future environmental scenarios. In this general discussion, I integrate the results obtained from applying multiple approaches and perspectives, establishing links between taxa and functions, discussing the importance of the level of taxonomic resolution and finalizing with some future endeavors that I foresee for time series analysis and marine microbial ecology in general.

Seasonality of microbes in the coastal ocean

The structure and dynamics of ecosystems are largely influenced by environmental heterogeneity (Levin, 1992), a property present at multiple scales ranging from the microhabitat to global ocean patterns (Giller *et al.*, 1994; Pinel-Alloul, 1995). These scales are hierarchical, interrelated, and each drive the main processes governing community assembly (Vellend, 2010). One approach to assess the different scales in microbial ecology studies is to evaluate three main axes of variation: space, time, and phylogeny. All these scales can be measured using a different grain (the breadth of individual samples units, such as 10 L vs. 1 L) and extent (the breadth of the whole study, such as sampling the whole Pacific Ocean vs. sampling a single coastline; Ladau and Eloe-Fadrosh, 2019). The properties of our study are the following: we sample in a single fixed point –regional scale–, the sampling has been performed through several years following a monthly sampling scheme –long extent with coarse grain–, and the phylogenetic scale has been explored at multiple levels in each **chapter** (see the taxonomic resolution section below). Through such a long scale, the dominant variation factors in a temperate coastal ocean are expected to be the environmental heterogeneity generated by the change in abiotic processes (Hewson *et al.*, 2006). Overall, this thesis shows that the relatively large range of temperature, day length, and nutrient concentrations among other factors deeply influence the structure of the community down to closely related taxa (**Chapter I**), specific functional groups (**Chapter II** and **III**), and strains (**Chapter IV**).

In **Chapter I**, I show that 47% of the total relative abundance of bacteria had a significant seasonal pattern. Both, abundant (113 ASVs) and rare taxa (184 ASVs), were seasonal and the use of a large resolution through amplicon sequence variants (ASVs) unveiled new unexplored patterns of seasonality such as the seasonal differentiation of *Puniceispirillaceae* ASVs in summer. Although our results show a high proportion of the community as seasonal, these numbers are probably an underestimate. A case could be made that it would be more relevant to understand why some taxa are non-seasonal rather than the contrary, since with such range of environmental seasonal variability being seasonal should be the rule. The temperate sea presents a wide range of variation in key abiotic factors, and it is plausible that multiple ecotypes have evolved and adapted to each of these conditions. For abundant non-seasonal taxa, it is possible that the maximum resolution that the 16S rRNA gene offers hides the patterns by the aggregation of distinct ecotypes into a single one. The results observed here point towards this direction: Pelagibacterales are some of the most abundant organisms in Blanes Bay (Alonso-Sáez *et al.*, 2007; Mestre *et al.*, 2020) as well as elsewhere in the ocean (Giovannoni, 2017), and while in **Chapter I** most of the *Pelagibacteraceae* ASVs were shown to be non-seasonal, in **Chapter III** multiple species from this family were observed to present adaptations to the summer conditions by having a genetic repertoire specialized to prevent phosphorous limiting conditions. Additionally, we established the seasonality of rare taxa, confirming the results obtained in previous studies (Alonso-Sáez *et al.*, 2015). Improvements in the sequencing depth will allow to unveil significant patterns in the rare non-seasonal taxa, that nowadays still remain hidden. Finally, the measured fraction of the community that is seasonal is also influenced by the choice of the statistic method and the thresholds used. In **Chapter I** we used the Lomb Scargle periodogram and used the same threshold than in Lambert *et al.* (2018) to compare results obtained in similar conditions. I later lowered the threshold by comparing with other statistics (Giner *et al.*, 2019) given that a lower threshold generated similar results between the two methods (see Supplementary Information 1 from **Chapter III**). Summing up, these results confirm the seasonal patterns observed in previous works (reviewed in Bunse and Pinhassi, 2017) and shows the utility of ASV delineation algorithms able to distinguish the seasonal variations at a finer taxonomic resolution.

Another question posed in this thesis is how similar the niche of closely related taxa is (**Chapter I**). The most relevant result in this direction was the observed negative relationship between niche similarity and phylogenetic distance for the main SAR11 clades (and less clearly SAR86). This result is consistent with species sorting for which the coexistence between closely related taxa is facilitated through niche partitioning processes (Langenheder and Lindström, 2019). What are the differentiated traits that allow the coexistence of closely related taxa is another key issue to explore. Recent results have shown how *Pelagibacteraceae* have a high recombination capacity among distantly related members (López-Pérez *et al.*, 2020). It is possible that the large populations of this group coupled to the environmental seasonal dynamics have allowed differentiation of the accessory genetic repertoire, allowing that closely related strains to coexist without outcompeting each other. As shown in **Chapter III**, the SAR86 and SAR11 species present specific adaptations to the summer conditions, something that underlines the functional basis for ecotype differentiation. To further investigate this topic, the next step would be to study the seasonal pattern of the *Pelagibacteraceae* genomic species using MAGs retrieved both from the time series metagenomic data, and the SAGs obtained from Blanes Bay by Haro-Moreno *et al.*, (2020) to complete the information obtained in **Chapter I** and **Chapter III**.

One possible pattern, which we expected to be likely to occur, were situations in which two similar taxa presented alternances so that in some years a variant was predominant, whereas in another year the other would be the dominant one. Such a result would indicate random assembly processes occurring between similar taxa, caused by drift or by dispersion events. In 11 years of sampling, nevertheless, these events were uncommon or nonexistent. Besides these in situ observations, we observed in the experiments conducted in **Chapter IV** that the patterns of growth were similar between the replicates, with the same MAG growing in each, indicating therefore a strong important role of selection and the lower relevance of drift processes. Random alternances between closely related taxa could still occur at lower temporal resolutions, and we would have missed them. A multi-year weekly sampling found changing co-occurrence patterns between years (Lambert *et al.*, 2021), and daily measurements found that the date of the maximum abundance of close OTUs indicated ecological differentiation (Martin-Platero *et al.*, 2018). Another possibility is that these selection processes to grow and thrive are rather dominated by interactions between species from different families. Differentiating the importance of these biological interactions is difficult, since these processes are happening in the community as a whole instead of focusing in an specific particular group, finding the emergent properties shared between the taxa (Bergelson *et al.*, 2021). Trying therefore to replicate and understand if these interactions are predictable and robust and what is the casual properties that generate these repetitions is one of the key endeavors for seasonal time series analysis.

Seasonality of functions and their implications for the system

In **Chapter III** we focused the analysis on key functional groups (e.g. nitrate reducers, primary producers, etc.) and their seasonal patterns. Coinciding with the results of **Chapter I**, most of the functional groups presented a seasonal pattern as a whole. The use of the gene as a biological unit of study imply understanding what we are specifically measuring when we enumerate them, given that the whole biological unit is the genome of the species harboring the gene. As McMahon puts it, “*Genes are expressed within cells, not in a homogenized cytoplasmic soup*.” (McMahon, 2015). The results presented, therefore, have to be seen as the emergent properties shared between all the taxa harboring a particular gene. In some cases, the seasonal variants within a functional group are a result of a similar selection pressure which translates into a single synchronic and aggregated pattern. This is what we observed for multiple phosphorous genes during the summer season. In other functional groups, the selective forces acting on other genes induce multiple differentiated seasonal patterns, not resulting in a similar global seasonal signal. An example was observed for the *rbcL* gene, having many seasonal variants belonging to cyanobacterial taxa and eukaryotes each with its own seasonal trend, or *pstS*, presented by multiple groups and not converging in summer like other phosphorous genes. Understanding how these functional groups are affected by these community assembly processes is key to obtain an explanation for the synchronization or random distributions. The aggregated synchronized pattern of phosphorus genes is a clear example of the dominant effect of nutrient limitation, which has also been observed in other studies (Haro-Moreno *et al.*, 2020; Ustick *et al.*, 2021). For other genes, it would be necessary to study how each gene influences the direct fitness of the strain or species that contains it in contrast with the other organisms on each specific environmental context. Given the fact that the recovery of all MAGs in a system is nowadays difficult, approaches focused in the genes giving taxonomic context are still a good strategy to perform these community analyses.

Moreover, assessing how the seasonal fluctuations in the functional groups are translated into gene expression and ultimately enzymatic activity in the environment is still needed. On a spatial scale, this was recent assessed using the *Tara* Oceans expedition dataset, which found that in the poles community turnover dominated the changes in transcript levels, whereas in more temperate conditions changes in gene expression was more important than community turnover (Salazar *et al.*, 2019). Given the large seasonal changes observed in the microbial community structure of our study site (**Chapter I**), we hypothesize that community turnover will have a larger impact than the regulation of gene expression. The obtention of seasonal metatranscriptomic datasets at the BBMO would allow to test this hypothesis.

On a technical note, the focus on specific functions allowed to use more refined bioinformatic approaches than if we had analyzed the patterns of all the functional genes. Specifically, the use of DIAMOND (Buchfink *et al.*, 2015) to match the reads to the subset of genes selected allowed us to retrieve on average 10X more reads than approaches based in the Burrows-Wheeler aligner (Li and Durbin, 2009) or similar methods that are less computationally demanding. This last method tends to be preferred by many authors because it is efficient for millions of genes, but detects less reads than the former option. Although the overall gene pattern could be similar, the patterns of gene variants would differ significantly between methods. An option that could facilitate enormously both, the computational tractability and the robustness of the results, would be to focus the entire gene-centric analysis to contigs larger than 1000 base pairs only. This small change would lower substantially the computing needs, and would link directly gene-centric with MAG analyses, since this second approach use only contigs of this approximate size (Kang *et al.*, 2019). Although this would imply losing some really rare variants, the overall usability of the gene catalogue would be improved. A study comparing the results of the two approaches would allow to observe how much gene diversity is lost by focusing on the subset of larger contigs, asses if the gain on computational tractability is worth the change.

The aerobic anoxygenic photoheterotrophs: a case study

Photoheterotrophy has been found as an important bacterial trait that can substantially affect biogeochemical cycling (Gómez-Consarnau *et al.*, 2019; Piwosz *et al.*, 2021). This thesis has allowed to gain important insights on the ecology and diversity of aerobic anoxygenic photoheterotrophic bacteria (AAPs). Of all the functional groups studied in **Chapter III**, the AAPs were one of the most seasonal groups, only rivaled by organisms containing *coxL* and *amoA*. The 92% of the total relative abundance of AAPs presented seasonal patterns, as opposed to only half of the total relative abundance of the whole community. Initially, in **Chapter II**, based in the distinction of phylotypes proposed by Yutin *et al.* (2007), we found that phylogroup K (Gammaproteobacteria) was the most abundant group, followed by phylogroup G (Alphaproteobacteria), validating the previous studies performed in Blanes Bay (Ferrera *et al.*, 2014). Phylogroup K is mostly linked to bacterial members of the OM60/NOR5 clade, typical of coastal waters (Yan *et al.*, 2009). The nomenclature used for distinguishing taxonomically the AAPs was updated through the course of the thesis. Through the use of the Genome Taxonomy Database (GTDB) in **Chapter III** we found that the main phylogroup K taxon in our system was genus *Luminiphilus* (OM60/NOR5). Phylogroup G was mainly affiliated to MED-G52, an uncultured genus within *Rhodobacteraceae* (the genome was assembled in Haro-Moreno *et al.*, 2018).

In **Chapter IV** we were also able to obtain representative MAGs of both *Luminiphilus* and MED-G52 from the manipulation experiments. These genera did not present a particularly high growth in the incubations, and the genomic analyses indicated that they were oligotrophs (Weissman *et al.*, 2021). In these experiments however we found many other AAPs able to grow substantially, mainly taxa within the *Rhodobacteraceae* family, such as *Citerimonas*, *Octadecabacter*, *Yoonia*, and *Nereida*, most of them copiotrophic bacteria. Through the match of the *pufM* gene variants presented in **Chapter III** and **Chapter IV** (Table 1), we could distinguish two ecological lifestyles within the AAPs. The most abundant AAPs in the BBMO time series are oligotrophs from a genomic perspective, whereas groups not as abundant present mainly a copiotrophic growth able to respond to fast nutrient inputs and with their population controlled by predation. This comparison therefore indicates that previous data from manipulation experiments performed focusing on AAPs (Ferrera *et al.*, 2011, 2017; Sánchez *et al.*, 2020) using microscopy counts are measuring mainly the growth of these fast responders, and not of the in situ dominant groups. The isolation of the most abundant oligotrophic AAPs for experimentation would be relevant to understand their lifestyle traits. In fact, experimental data with Gammaproteobacteria AAPs found that the regulation of bacteriochlorophyll *a* expression greatly differed from the regulation mechanisms presented by the members of the *Roseobacter* clade (Spring and Riedel, 2013). Characterizing the Alphaproteobacteria MED-G52 would also be key to deepen our understanding of the ecology of coastal AAPs.

nBLAST match between the BBMO and REMEI *pufM* variants

Gene variant (BBMO)	Total relative abundance (BBMO)	MAGs	Family	Genus	Species	Identity	length	mismatch	Fold change (REMEI)		
									median	max	strategy
BL131105.k127_3241710.mgm_4091835	19.10%	REM_Bin_00314[1]	*Halieaceae*	*Luminiphilus*	-	99.59%	975	4	-	-	-
BL110913.k127_979604.mgm_1298918	12.60%	REMEI_MAG_00135	*Halieaceae*	*Luminiphilus*	Luminiphilus sp000227505	85.34%	771	113	1.16	1.93	oligotroph
BL151006.k127_1259246_3.pro_779870	4.21%	REMEI_MAG_00225	*Rhodobacteraceae*	MED-G52	MED-G52 sp001627375	88.42%	924	107	0.44	0.80	oligotroph
BL100413.k127_3387677.mgm_3695606	4.19%	REMEI_MAG_00206	*Halieaceae*	*Luminiphilus*	-	89.54%	975	102	0.79	2.22	oligotroph
BL110913.k127_559423_73.pro_415420	3.36%	REMEI_MAG_00028	*Rhodobacteraceae*	*Nereida*	*Nereida ignava*	100.00%	930	0	100.00	100.00	copiotroph
BL140505.k127_788618_6.pro_346555	3.33%	REMEI_MAG_00143	*Halieaceae*	*Luminiphilus*	Luminiphilus sp002390485	99.80%	975	2	0.73	0.89	oligotroph
BL121009.k127_397646.mgm_554314	2.15%	REMEI_MAG_00043	*Rhodobacteraceae*	HIMB11	HIMB11 sp000472185	86.54%	929	125	6.45	10.52	copiotroph
BL141007.k127_2678758.mgm_3270912	2.15%	REMEI_MAG_00222	*Halieaceae*	*Luminiphilus*	-	97.74%	972	22	0.51	1.98	oligotroph
BL141007.k127_1594682.mgm_2079416	1.38%	REMEI_MAG_00058	*Halieaceae*	*Luminiphilus*	Luminiphilus sp002456975	96.61%	975	33	0.73	0.80	oligotroph
BL090609.k127_385495_26.pro_264673	1.25%	REMEI_MAG_00043	*Rhodobacteraceae*	HIMB11	HIMB11 sp000472185	100.00%	1044	0	6.45	10.52	copiotroph
BL110704.k127_99977.mgm_134354	1.04%	REMEI_MAG_00227	*Halieaceae*	*Luminiphilus*	Luminiphilus sp002683395	85.55%	976	139	0.73	1.91	oligotroph
BL140310.k127_3062105_1.pro_1582482	1.00%	REMEI_MAG_00225	*Rhodobacteraceae*	MED-G52	MED-G52 sp001627375	92.42%	924	70	0.44	0.80	oligotroph
BL100803.k127_2250725.mgm_2500557	0.97%	REMEI_MAG_00043	*Rhodobacteraceae*	HIMB11	HIMB11 sp000472185	86.73%	987	131	6.45	10.52	copiotroph
BL090317.k127_946426_9.pro_398865	0.64%	REMEI_MAG_00054	*Rhodobacteraceae*	*Yoonia*	-	99.78%	927	2	56.38	100.00	copiotroph
BL090915.k127_669386_20.pro_277514	0.47%	REMEI_MAG_00090	UBA8317	UBA8317	UBA8317 sp002469865	99.35%	927	6	3.02	5.67	oligotroph
BL120518.k127_405583.mgm_535099	0.36%	REMEI_MAG_00121	*Rhodobacteraceae*	*Octadecabacter*	*Octadecabacter jejudonensis*	100.00%	921	0	100.00	100.00	copiotroph
BL141111.k127_10581_2.pro_7744	0.35%	REMEI_MAG_00043	*Rhodobacteraceae*	HIMB11	HIMB11 sp000472185	87.27%	1045	131	6.45	10.52	copiotroph
BL100413.k127_416450.mgm_502882	0.26%	REMEI_MAG_00246	*Halieaceae*	*Luminiphilus*	Luminiphilus sp000169115	96.51%	975	34	0.94	2.28	oligotroph
BL150609.k127_2143569.mgm_2536507	0.25%	REMEI_MAG_00206	*Halieaceae*	*Luminiphilus*	-	89.27%	783	84	0.79	2.22	oligotroph
BL110412.k127_736559.mgm_892350	0.25%	REMEI_MAG_00181	*Halieaceae*	*Luminiphilus*	Luminiphilus sp002683335	100.00%	975	0	0.86	100.00	oligotroph

[1] The Bin presented a completeness of 0% and a contamination of 0% and it was not included Chapter IV for this reason.

Table 1. Nucleotide BLAST between the *pufM* sequences from Chapter III and the sequences obtained from the MAGs appearing in the REMEI experiment (Chapter IV). Gene variant indicates the sequences from the Blanes Bay Microbial Observatory (Chapter III), and the second column indicates the Total relative abundance (%) in the whole dataset; 'MAGs' indicate the MAG name, together with the taxonomy, median and maximum fold change in the REMEI experiments, and 'strategy' indicates the type of growth behavior (oligotrophic or copiotrophic, see Chapter IV); 'identity', 'length' and 'mismatch' indicate the percent identity of thematches, the length of the match and the number of mismatches. Only values over 85% identity are presented.

Interplay between seasonality, niche and taxonomic resolution

The ecological niche is the set of environmental conditions which enable the persistence of a population (Fahimipour and Gross, 2020). On a seasonal scale, the heterogeneity of the environmental conditions provides estimates of the width of the niche of the different species. How we measure and distinguish biological units will have an impact on the question of which is the niche width of a given species (i.e., under which sets of environmental values the growth of the unit is favorable). In **Chapter I** and **II** we had the opportunity to test how the observed niche is influenced by the methods used to define ecological units, when switching from OTUs to ASVs, we could prove that the higher

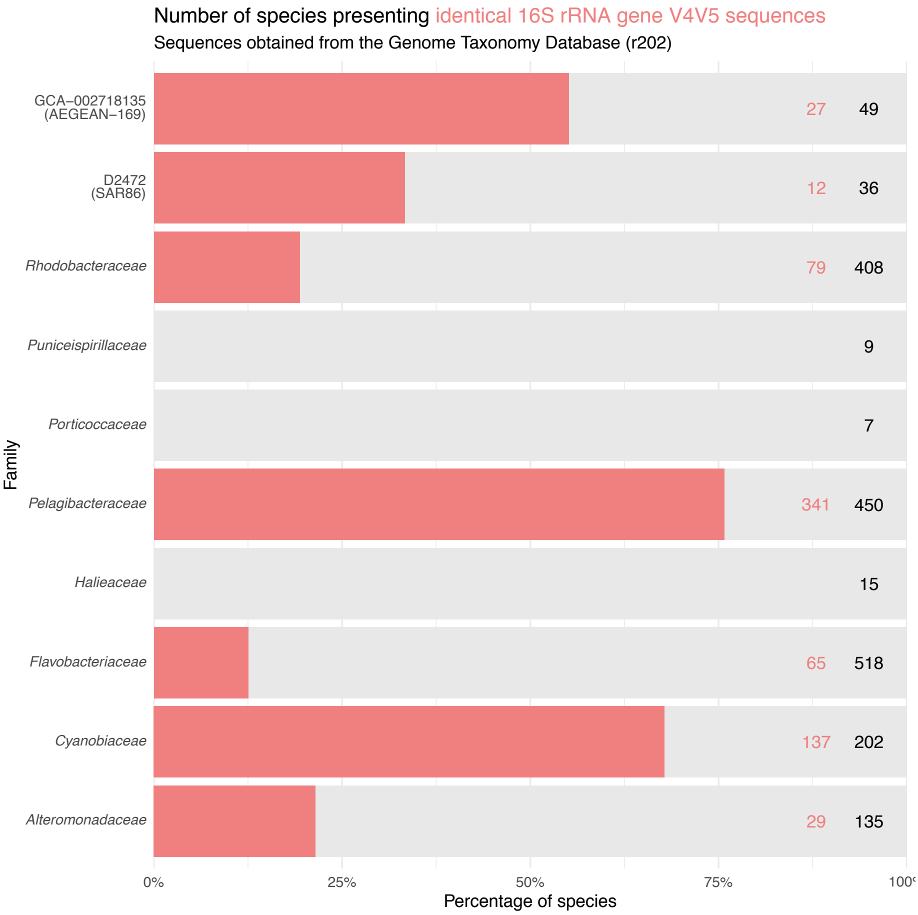


Figure 1. Percentage of sequences in the Genome Taxonomy Database (release 202) with an identical sequence (red) in key taxonomic families. All the 16S rRNA gene sequences were retrieved and processed using cutadapt (Martin, 2011) to only keep the V4-V5 region; an alignment of the trimmed sequences was used to cluster them at 100% and obtain the percentage of species with an identical sequence from the total of species present in GTDB for an specific genus. The Y axis indicates the various families and the X axis the percentage, with red for the number of identical regions and grey for the rest. The total number of species considered (black) and the number of species having an identical region (red) is displayed in each bar.

resolution unveiled hidden seasonal patterns within the same OTU, indicating different realized niches. Additionally, the niche width increases when moving from a strain, to a species, to a genus, or larger categories. Yet how the seasonal trends observed at the species level translates to the larger taxonomic ranks (genus, family) is still unknown. We have addressed the issue in **Chapter I** and partially in **Chapter IV**.

Most studies in microbial ecology do not bother questioning whether the 16S-based unit is a species or a strain. Most authors use whatever resolution without really thinking whether each OTU or ASV contains only one or more strains. To inspect how the resolution of the analysis might impact our understanding of the ecology of marine microbes, and particularly how the ASVs are related to the more commonly used taxonomic levels (strain, species), I extracted and compared the V4-V5 hypervariable regions of the 16S rRNA gene of several key GTDB marine taxa observed in **Chapter I**, aligned the sequences and evaluated which sequences are identical. These regions are what are amplified with primers typically used in the marine studies (Parada *et al.*, 2016). An overview of this relationship (Figure 1) indicates that in many cases, the resolution of the V4-V5 region splits units at a level in-between the species and genus (Figure 1, Table 2). This analysis shows how the % of species presenting an identical V4-V5 region is variable between taxa, with groups such as *Pelagibacteraceae*, D2472 (SAR86 family), AEGEAN-169, and *Cyanobiaceae* having multiple species with an identical V4-V5 region, whereas other groups, such as *Rhodobacteraceae*, *Haliaceae*, and *Flavobacteriaceae* where each sequence contains only one species. It is possible that the former groups, mostly composed of species with a streamlined genome and usually adapted to oligotrophy, present only one copy of the 16S rRNA gene, whereas the second group include mostly copiotrophic organisms, which usually present more than one operon copy (Vieira-Silva and Rocha, 2010). Through matching the ASVs of **Chapter I** (Table 2) to the V4-V5 16S rRNA gene regions obtained from GTDB we observed that some of the most abundant ASVs correspond to sequences shared between multiple species. An ASV then, is not always a species. Using the *Pelagibacter* genus as an example, ASV3 has a region that is shared between 41 GTDB species, ASV7 to 40 and ASV8 to 29. The fact that 41 species had the same identical sequence in the v4-V5 16S rRNA gene means that, when in **Chapter I** we describe the patterns of seasonality of ASV3, we cannot establish with certainty which of the 41 species exist in our system, or whether we are merging all or some of them in a single unit of diversity.

These results imply that the pattern of niche similarity and phylogenetic divergence observed in **Chapter I** to in-between the species and genus level, since the ASVs are aggregating species together but within each genus there are multiple ASVs. For the AAP (**Chapters II** and **III**), the relationship between each variant and species differentiation seems to be one to one: the most abundant *pufM* gene amplicons had only one 100% match with the metagenomic variants, and in **Chapter III** each variant was assigned mainly to different *Luminiphilus* species. In future endeavors these compari-

ASV in BBMO	Cluster genus	Example species inside	Number species identical region
Alteromonadaceae			
ASV404	*Alteromonas*	*Alteromonas abrolhosensis*	4
ASV267	*Pseudoalteromonas*	*Pseudoalteromonas atlantica*	6
Cyanobiaceae			
ASV16	Prochlorococcus_A	Prochlorococcus_A pastoris	10
ASV5	Synechococcus_C	Synechococcus_C sp002500205	2
ASV1	Synechococcus_E	Synechococcus_E sp002724845	7
ASV12	Synechococcus_E	Synechococcus_E sp000012505	3
D2472 (SAR86 family)			
ASV34	SAR86A	SAR86A sp000252525	4
GCA-002718135 (AEGEAN-169)			
ASV152	AG-337-I02	AG-337-I02 sp902591485	7
ASV285	AG-337-I02	AG-337-I02 sp902620425	5
ASV243	AG-337-I02	AG-337-I02 sp902584805	3
ASV98	AG-337-I02	AG-337-I02 sp902511405	2
Pelagibacteraceae			
ASV3	*Pelagibacter*	*Pelagibacter ubique*	41
ASV7	*Pelagibacter*	Pelagibacter sp902586125	40
ASV53	*Pelagibacter*	Pelagibacter sp902567045	30
ASV8	*Pelagibacter*	Pelagibacter sp902591025	29
ASV4	*Pelagibacter*	Pelagibacter sp902579105	23
ASV2	*Pelagibacter*	Pelagibacter sp003279685	13

ASV in BBMO	Cluster genus	Example species inside	Number species identical region
ASV60	*Pelagibacter*	Pelagibacter sp902622545	13
ASV9	*Pelagibacter*	Pelagibacter sp003279775	8
ASV87	*Pelagibacter*	Pelagibacter sp902627645	4
ASV36	*Pelagibacter*	Pelagibacter sp902514695	2
ASV68	*Pelagibacter*	Pelagibacter sp902611145	2
ASV10	Pelagibacter_A	Pelagibacter_A sp902624195	18
ASV6	Pelagibacter_A	Pelagibacter_A sp003213555	12
ASV35	Pelagibacter_A	Pelagibacter_A sp902593905	12
ASV20	Pelagibacter_A	Pelagibacter_A sp902569735	10
Porticoccaceae			
ASV21	TMED48	TMED48 sp002591625	2
Rhodobacteraceae			
ASV14	LGRT01	LGRT01 sp902539465	2

Table 2. Matching between the ASVs of Chapter I and the 100% species clusters (species presenting an identical region of the V4-V5 16S rRNA gene, see Figure 1). The first column indicates the ASV (following the nomenclature of Chapter I), the second the genus, the third the name of one of the species within the cluster, and the fourth the number of species having identical this region.

sons could be further explored to obtain the specific relationship between *pufM* gene variants and the taxonomic level these represent.

Chapter IV delves deeper into the taxonomic resolution. How the dynamics of the strains differentiate from that of the species or even the genera is still an unexplored issue (McLaren and Callahan, 2018). A recent study assessed the influence of phylogenetic and environmental constraints in the content and change of the pangenome, the entire set of genes from all strains in a species (Maistrenko *et al.*, 2020). The environmental preferences explained up to 49% of the variance in pangenome features, whereas the phylogenetic relationships explained 18%. Our experimental analysis in **Chapter IV** showed that microdiversity can vary between seasons and that the patterns of this variation are specific for each taxon, with some of them increasing in microdiversity in summer and others presenting this trend in winter. Reconstruction of MAGs allow these in-depth comparisons. Future analyses could be based on the evaluation of the microdiversity trends observed in the time series using the MAGs recovered in the experiment, assessing if we also observe a higher clonality when the abundance of the species is higher. Given the multiple studies pointing to the strain level as an extremely relevant level in explaining ecological dynamics (Alneberg *et al.*, 2020; Goyal *et al.*, 2021; Olm *et al.*, 2021), further analyses focused on delineating them could improve our ecosystem knowledge.

Future directions

Understanding microbial dynamics is a central question in the field of marine microbial ecology. Our future understanding of this topic will still be impacted by the need to obtain larger and more complete datasets. From my perspective, this trend should include the integration of data coming from multiple perspectives and previous studies, and the funding of these tasks to have stable and robust platforms.

Technologies able to sequence long-read DNA fragments have improved substantially in the last 4 years, and will be key for long-term time series analyses. Since the start of this thesis in 2017, the error rate of Nanopore and PacBio technologies has dropped significantly, to the point of becoming useful to even perform metagenomics (Haro-Moreno *et al.*, 2021). Moreover, new approaches that make use of the Illumina technologies allow the creation of synthetic long-reads (Callahan *et al.*, 2021), which could avoid the programmed obsolescence of the hundreds of Illumina machines producing short-read sequences. The obtention of long-read data –even if only for a small set of samples due to their still high prices– in time series could nicely complement the short-read data obtained before (**Chapter III** and IV), since it would allow obtaining more high quality circularized genomes in some cases (Haro-Moreno *et al.*, 2021). Given that in time series data all the information is collected on a similar spot, the compilation and merging of the long-read and short-read data would be an easy task.

The improvement in the genomic databases also calls for the generation of a representative genomic catalogue for the Mediterranean Sea. The creation of such datasets at local scales are key to improve the breadth and quality of information from the ecosystem. As an example, 15 out of the 262 MAGs obtained in **Chapter IV** present only a taxonomic match to the Mediterranean MAG dataset generated from 7 samples from the *Tara* Oceans expedition (Tully *et al.*, 2017), showing how even with a small sample number the amount of information that can be obtained is quite large. Nowadays, the creation of a standardized pipeline able to aggregate all the genomes in a high-quality dataset from the e.g. Service d'Observation du Laboratoire Arago, the BBMO and the Evolutionary Genomics Group from Universidad Miguel Hernández could be an initial ambitious effort for the North Western part of the Mediterranean Sea. New technologies, such as the nextflow and snakemake workflows coupled with Github would facilitate this process (Reiter *et al.*, 2021).

Despite its potential, the obtention of these high-quality genomes however will never completely substitute marker gene analysis, but rather enhance it. Nowadays, metagenomic approaches are preferred since the ratio between information obtained / price is likely much larger than with other techniques. From metagenomic samples we can obtain most of the functional traits of the species, avoiding biases linked to the amplicon sequencing approach, and we can differentiate down to strain level in some cases. The continuous generation of metagenomic data will characterize efficiently most of the abundant clades in the marine ecosystem, specially at regional levels that have long-term stations, and at that point the ratio information / price would favor other techniques, given that most of the information obtained by metagenomes would have already been characterized, and to just quantify spatial or temporal trends, techniques such as those based on marker gene would be sufficient (Schloss, 2020). When we reach this point, each technique would be best suited for different specific questions, and the marker gene analysis with long-read technologies would be key to answer many of the questions ecologists pose. Some examples of the questions that can be answered with marker genes, are the testing of macroecological questions down to the species/strain level (using long-reads), the tracking of the species abundance patterns in experiments, or the cheap screening of many environments to track the presence of a specific species that will be then sorted for single amplified genome sequencing. The use of long-reads for marker gene analysis will need the development of new algorithms since for the organisms presenting multiple 16S rRNA gene copies the obtention of larger sequences will artificially split genomes if approached through threshold-free methods (Schloss, 2021). In conclusion, the use of marker gene approaches still has a long-life expectancy in microbial ecology, especially in the future.

Another problem stemming from the usage of multiple techniques such as marker genes and metagenomics in microbial ecology is the integration of results coming from these multiple perspectives. Each technique deals with a specific taxonomic resolution as discussed before, which difficult the integration of these data and assessing the links between different studies. The creation of robust datasets such as the GTDB facilitates the connection between studies. Through the taxonomic assignation of the 16S rRNA gene, researchers can find information from the genomes of a specific group, and easily retrieve and consult the previous studies that generated these genomes. Websites to retrieve information from GTDB, such as Annotree (http://annotree.uwaterloo.ca/app/) facilitate the obtention and exploration of information from all the genomes compiled, such as the distribution of specific functions (Mendler *et al.*, 2019). As a representative example, from **Chapter II** (published in 2019) to **Chapter III** the use of GTDB allowed us to link most of the AAPs assigned just to phylogroup K in BBMO (Ferrera *et al.*, 2014) to the *Haliaceae* family, and most probably to the *Luminiphilus* species. One of the main concerns with GTDB is, however, that the new amount of information often forces the change of the common nomenclature used in previous studies. This problem is due to the exponential growth of deposited genomes, the advancement in phylogenetic approaches, and the fact that most novelty does not have isolates, which implies that the taxonomic names are also not fixed. This lack of a fixed and established nomenclature for key microbiological groups and the use of placeholders based on the obtention of genomes or 16S rRNA gene sequences can create "nomenclature wars" at worst, and at best non-robust traceable information (Hugenholtz *et al.*, 2021). One example of this is the disappearance of the Oceanospirillales order (Liao *et al.*, 2020). Although most of the changes in nomenclature occur for a good reason (e.g. new species and new phylogenies, or rebranching of groups thought to be monophyletic), there is a need to develop new naming protocols based on genomic data (reviewed in Murray *et al.*, 2020). Another important step towards facilitating the use of the adequate nomenclatures and understand the changes would be the creation and maintenance of websites linking old traditionally nomenclatures with the new ones (see Supplementary Table 2 in **Chapter II** as an example). Although the GTDB presents this functionality in the web, it can considerably be improved.

Overall, the improvements we are seeing in the omics field will contribute substantially to the gain in knowledge of the marine microbiome, creating a more mechanistic and predictive understanding of the microbial dynamics and their implications in the ecosystem biogeochemical cycles.

A final word

The evolution of the field of marine microbial ecology, however, depends entirely in our ability as scientists to reinvent ourselves in the ways in which we do science (Gardner *et al.*, 2021). The use of materials such as plastic, the improvement in computing capacity and the global expeditions has pushed the boundaries of knowledge at a pace never seen before. These activities, however, rely on large quantities of single-use plastic, fossil fuels, and large electricity usage. Although nowadays it is our normality, it is too high for the sustainable planetary boundaries, even if we consider science as a top priority endeavor. The climate crisis scenarios will force us to rethink how we perform marine science, since most of the science-as-usual practices nowadays depend on a too large carbon footprint and materials that will become scarce in the future. In less than 10 years we will have to rethink how we perform these activities, whether as a community decision or simply by waiting for the resource exhaustion. Stepping away actively from these commodities will necessarily mean a slower science. Albeit on a first thought this sounds like a drawback, if it is coupled with an active opposition against the precarity in academia, it could mean a substantial improvement on the quality of our work, staying away from the path towards quantity founded on the publish or perish reality. A switch from *publications* to *public actions* could be a good initial step towards the science we would like to do: a sustainable one, based on open, more equal values that allows to care for the environment we study.

References

Alneberg, J., Bennke, C., Beier, S., Bunse, C., Quince, C., Ininbergs, K., *et al.* (2020) Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Commun Biol* 3: 119.

Alonso-Sáez, L., Balagué, V., Sà, E.L., Sánchez, O., González, J.M., Pinhassi, J., *et al.* (2007) Seasonality in bacterial diversity in north-west Mediterranean coastal waters: Assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiol Ecol* 60: 98–112.

Alonso-Sáez, L., Díaz-Pérez, L., and Morán, X.A.G. (2015) The hidden seasonality of the rare biosphere in coastal marine bacterioplankton. *Environ Microbiol* 17: 3766–3780.

Bergelson, J., Kreitman, M., Petrov, D.A., Sanchez, A., and Tikhonov, M. (2021) Functional biology in its natural context: A search for emergent simplicity. *eLife* 10: e67646.

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

Bunse, C. and Pinhassi, J. (2017) Marine Bacterioplankton Seasonal Succession Dynamics. *Trends Microbiol* 25: 1–12.

Callahan, B.J., Grinevich, D., Thakur, S., Balamotis, M.A., and Yehezkel, T.B. (2021) Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome* 9: 130.

Fahimipour, A.K. and Gross, T. (2020) Mapping the bacterial metabolic niche space. *Nat Commun* 11: 4887.

Ferrera, I., Borrego, C.M., Salazar, G., and Gasol, J.M. (2014) Marked seasonality of aerobic anoxygenic phototrophic bacteria in the coastal NW Mediterranean Sea as revealed by cell abundance, pigment concentration and pyrosequencing of *pufM* gene. *Environ Microbiol* 16: 2953–2965.

Ferrera, I., Gasol, J.M., Sebastián, M., Hojerová, E., and Kobížek, M. (2011) Comparison of growth rates of aerobic anoxygenic phototrophic bacteria and other bacterioplankton groups in coastal mediterranean waters. *App Environ Microbiol* 77: 7451–7458.

Ferrera, I., Sánchez, O., Kolářová, E., Koblížek, M., and Gasol, J.M. (2017) Light enhances the growth rates of natural populations of aerobic anoxygenic phototrophic bacteria. *ISME J* 11: 2391–2393.

Gardner, C.J., Thierry, A., Rowlandson, W., and Steinberger, J.K. (2021) From Publications to Public Actions: The Role of Universities in Facilitating Academic Advocacy and Activism in the Climate and Ecological Emergency. *Front Sustain* 2: 679019.

Giller, P.S., Hildrew, A.G., and Raffaelli, D.G. (1994) Aquatic Ecology: Scale, Pattern and Process, Blackwell Scientific Publications.

Giner, C.R., Balagué, V., Krabberød, A.K., Ferrera, I., Reñé, A., Garcés, E., *et al.* (2019) Quantifying long-term recurrence in planktonic microbial eukaryotes. *Mol Ecol* 28: 923–935.

Giovannoni, S.J. (2017) SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Ann Rev Mar Sci* 9: 231–255.

Gómez-Consarnau, L., Raven, J.A., Levine, N.M., Cutter, L.S., Wang, D., Seegers, B., *et al.* (2019) Microbial rhodopsins are major contributors to the solar energy captured in the sea. *Sci Adv* 5: eaaw8855.

Goyal, A., Bittleston, L.S., Leventhal, G.E., Lu, L., and Cordero, O.X. (2021) Interactions between strains govern the eco-evolutionary dynamics of microbial communities, Microbiology. *bioRxiv*

Haro-Moreno, J.M., López-Pérez, M., and Rodriguez-Valera, F. (2021) Enhanced Recovery of Microbial Genes and Genomes From a Marine Water Column Using Long-Read Metagenomics. *Front Microbiol* 12: 708782.

Haro-Moreno, J.M., López-Pérez, M., de la Torre, J.R., Picazo, A., Camacho, A., and Rodriguez-Valera, F. (2018) Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. *Microbiome* 6: 128.

Haro-Moreno, J.M., Rodriguez-Valera, F., Rosselli, R., Martinez-Hernandez, F., Roda-Garcia, J.J., Gomez, M.L., *et al.* (2020) Ecogenomics of the SAR11 clade. *Environ Microbiol* 22: 1748–1763.

Hewson, I., Steele, J., Capone, D., and Fuhrman, J. (2006) Temporal and spatial scales of variation in bacterioplankton assemblages of oligotrophic surface waters. *Mar Ecol Prog Ser* 311: 67–77.

Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D.H., and Soo, R.M. (2021) Prokaryotic taxonomy and nomenclature in the age of big sequence data. *ISME J* 15: 1879–1892.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7: e7359.

Ladau, J. and Eloe-Fadrosh, E.A. (2019) Spatial, Temporal, and Phylogenetic Scales of Microbial Ecology. *Trends Microbiol* 27: 662–669.

Lambert, S., Lozano, J.-C., Bouget, F.-Y., and Galand, P.E. (2021) Seasonal marine microorganisms change neighbours under contrasting environmental conditions. *Environ Microbiol* 23: 2592–2604.

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., and Galand, P.E. (2018) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *ISME J* 388–401.

Langenheder, S. and Lindström, E.S. (2019) Factors influencing aquatic and terrestrial bacterial community assembly. *Environ Microbiol Rep* 11: 306–315.

Levin, S.A. (1992) The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology* 73: 1943–1967.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25: 1754–1760.

Liao, H., Lin, X., Li, Y., Qu, M., and Tian, Y. (2020) Reclassification of the Taxonomic Framework of Orders Cellvibrionales, Oceanospirillales, Pseudomonadales, and Alteromonadales in Class Gammaproteobacteria through Phylogenomic Tree Analysis. *mSystems* 5: e00543-20.

López-Pérez, M., Haro-Moreno, J.M., Coutinho, F.H., Martinez-Garcia, M., and Rodriguez-Valera, F. (2020) The Evolutionary Success of the Marine Bacterium SAR11 Analyzed through a Metagenomic Perspective. *mSystems* 5: e00605-20.

Maistrenko, O.M., Mende, D.R., Luetge, M., Hildebrand, F., Schmidt, T.S.B., Li, S.S., *et al.* (2020) Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J* 14: 1247–1259.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17: 10.

Martin-Platero, A.M., Cleary, B., Kauffman, K., Preheim, S.P., McGillicuddy, D.J., Alm, E.J., and Polz, M.F. (2018) High resolution time series reveals cohesive but short-lived communities in coastal plankton. *Nat Commun* 9: 266.

McLaren, M.R. and Callahan, B.J. (2018) In Nature, There Is Only Diversity. *mBio* 9: e02149-17.

McMahon, K. (2015) 'Metagenomics 2.0.' *Environ Microbiol Rep* 7: 38–39.

Mendler, K., Chen, H., Parks, D.H., Lobb, B., Hug, L.A., and Doxey, A.C. (2019) AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res* 47: 4442–4448.

Mestre, M., Höfer, J., Sala, M.M., and Gasol, J.M. (2020) Seasonal Variation of Bacterial Diversity Along the Marine Particulate Matter Continuum. *Front Microbiol* 11: 1590.

Murray, A.E., Freudenstein, J., Gribaldo, S., Hatzenpichler, R., Hugenholtz, P., Kämpfer, P., *et al.* (2020) Roadmap for naming uncultivated Archaea and Bacteria. *Nat Microbiol*.

Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J., and Banfield, J.F. (2021) inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* 39: 727–736.

Parada, A.E., Needham, D.M., and Fuhrman, J.A. (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* 18: 1403–1414.

Pinel-Alloul, P. (1995) Spatial heterogeneity as a multiscale characteristic of zooplankton community. *Hydrobiologia* 300: 17–42.

Piwosz, K., Villena-Alemany, C., and Mujakić, I. (2021) Photoheterotrophy by aerobic anoxygenic bacteria modulates carbon fluxes in a freshwater lake. *ISME J.*

Reiter, T., Brooks, P.T., Irber, L., Joslin, S.E.K., Reid, C.M., Scott, C., *et al.* (2021) Streamlining data-intensive biology with workflow systems. *GigaScience* 10: giaa140.

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., *et al.* (2019) Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* 179: 1068-1083.e21.

Sánchez, O., Ferrera, I., Mabrito, I., Gazulla, C.R., Sebastián, M., Auladell, A., *et al.* (2020) Seasonal impact of grazing, viral mortality, resource availability and light on the group-specific growth rates of coastal Mediterranean bacterioplankton. *Sci Rep* 10: 19773.

Schloss, P.D. (2021) Amplicon Sequence Variants Artificially Split Bacterial Genomes into Separate Clusters. *mSphere* 6: e00191-21.

Schloss, P.D. (2020) Reintroducing mothur: 10 Years Later. *Appl Environ Microbiol* 86: 13.

Spring, S. and Riedel, T. (2013) Mixotrophic growth of bacteriochlorophyll *a*-containing members of the OM60/NOR5 clade of marine gammaproteobacteria is carbon-starvation independent and correlates with the type of carbon source and oxygen availability. *BMC Microbiol* 13: 117

Tully, B.J., Sachdeva, R., Graham, E.D., and Heidelberg, J.F. (2017) 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. *PeerJ* 5: e3558.

Ustick, L.J., Larkin, A.A., Garcia, C.A., Garcia, N.S., Brock, M.L., Lee, J.A., *et al.* (2021) Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science* 372: 287.

Vellend, M. (2010) Conceptual Synthesis in Community Ecology. Q Rev Biol 85: 183–206.

Vieira-Silva, S. and Rocha, E.P.C. (2010) The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLoS Genet* 6: e1000808.

Weissman, J.L., Hou, S., and Fuhrman, J.A. (2021) Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *PNAS* 118: e2016810118.

Yan, S., Fuchs, B.M., Lenk, S., Harder, J., Jiao, N.-Z., and Amann, R. (2009) Biogeography and phylogeny of the NOR5/OM60 clade of Gammaproteobacteria. *Sys App Microbiol* 32: 124-139.

Yutin, N., Suzuki, M.T., Teeling, H., Weber, M., Venter, J.C., Rusch, D.B., and Béjà, O. (2007) Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. E*nviron Microbiol* 9: 1464–1475.

CONCLUSIONS

The main conclusions that arise from this thesis are:

i. The bacterial community of Blanes Bay presented recurrent patterns of seasonality. Based on 16S rRNA gene amplicons, 297 out of 6825 amplicon sequence variants (ASVs) were seasonal, which constituted almost half of the total relative abundance (47%).

ii. Niche similarity decreased as nucleotide divergence in the 16S rRNA gene increased for certain abundant taxa, such as Pelagibacter (SAR11 clade I) and Pelagibacter_A (SAR 11 clade II), pointing out that environmental selection is an important process structuring the seasonal dynamics of the members of these genera.

iii. The analysis of the seasonality distributions for each phylogenetic rank indicated that the class rank was non-seasonal for all the analyzed groups, being thus ecologically non-coherent. Contrarily, at the family and genera ranks, groups such as *Puniceispirillaceae* and *Haliaceae* presented cohesive responses.

iv. Aerobic anoxygenic photoheterotrophic bacteria (AAPs) showed repeatable long-term seasonal patterns, with several different phylotypes including ecotypes with distinctive temporal niche partitioning.

v. The AAP assemblages presented –during 10 years of analysis– a recurrent peak of diversity during winter, with gammaproteobacterial AAPs as the dominant members of the community year-round in Blanes Bay, while alphaproteobacterial taxa being prevalent in spring.

vi. Most of the 21 biogeochemically key marker genes studied in Blanes Bay presented recurrent seasonal dynamics, with multiple taxonomic groups contributing to each function. Genes such as *pufM*, coxL, ureC and tauA were enriched during spring, while phosphorous cycling genes were enriched during summer.

vii. The pattern of seasonal change of the phosphorous functional marker genes for *Pelagibacteraceae* and D2472 (SAR86 family) differed from the seasonal abundance of the corresponding family (based on 16S rRNA gene relative abundance), suggesting that each species presented specific adaptations, with some of them adapted to the summer conditions of phosphorous limitation.

viii. The analysis of experimental manipulations through a metagenomic perspective indicated that different *Flavobacteriaceae* MAGs were enriched in the predator-reduced treatment depending on the season, and the same occurred for several *Rhodobacteraceae* species in the diluted and virus-reduced treatments.

ix. The gene-content analysis of the dominant MAGs allowed to link specific genetic repertoires to some of the environmental conditions and to disentangle how strains are distributed through measurements of microdiversity, indicating that clonality dominates when the species become abundant.

x. The combination of multiple approaches in the study of time series –several molecular techniques and descriptive and experimental strategies– allows the obtention of stronger conclusions than the sum of each individual contribution.

Findability, Accessibility, Interoperability, and Reusability Data

This thesis follows the FAIR Guiding Principles for scientific data management and stewardship'*, allowing the scientific community to obtain the added-value gained by contemporary, formal scholarly digital publishing.

The code to produce all the analysis and visualizations for each chapter are available here:

Chapter I. https://github.com/adriaaulaICM/bbmo_niche_sea
Chapter II. https://gitlab.com/aauladell/AAP_time_series
Chapter III. https://github.com/adriaaulaICM/key_biogem_genes
Chapter IV. https://github.com/adriaaulaICM/remei_analysis_metaG

The data for Chapter I and II are found in the same repositories, with the location indicated in the repository manual.

The data for Chapter III and IV is still unpublished. Both chapters present the summary information (abundance tables, taxonomy, basic environmental data…) but the raw data is still unavailable.

* Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18