MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Doutorado no Programa de Pós-Graduação em Biologia Computacional e Sistemas

CARACTERIZAÇÃO DA DIVERSIDADE MICROBIANA DA PRAIA DOS ANJOS – ARRAIAL DO CABO – RJ ATRAVÉS DE METAGENÔMICA

RAFAEL RICARDO DE CASTRO CUADRAT

Rio de Janeiro
Maio de 2014

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

Rafael Ricardo de Castro Cuadrat

Caracterização da diversidade microbiana da Praia dos Anjos – Arraial do Cabo – RJ através de metagenômica

Tese apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em Biologia Computacional e Sistemas

Orientador: Prof. Dr. Alberto Martin Rivera Dávila

RIO DE JANEIRO
Maio de 2014

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

C961 Cuadrat, Rafael Ricardo de Castro

Caracterização da diversidade mibrobiana da Praia dos Anjos – Arraial do Cabo – RJ através de metagenômica/ Rafael Ricardo de Castro Cuadrat. – Rio de Janeiro, 2014.

xiv, 134 f. : il. ; 30 cm.

Tese (Doutorado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2014.

Bibliografia: f. 117-127

1. Metagenômica. 2. NGS. 3. Ressurgência. 4. PKS. 5. Fotossíntese anoxigênica. I. Título.

CDD 579.177

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: RAFAEL RICARDO DE CASTRO CUADRAT

CARACTERIZAÇÃO DA DIVERSIDADE MICROBIANA DA PRAIA DOS ANJOS – ARRAIAL DO CABO – RJ ATRAVÉS DE METAGENÔMICA

ORIENTADOR: **Prof. Dr. Alberto Martin Rivera Dávila**

Aprovada em: 13/05/2014

EXAMINADORES:

Prof. Dr. Floriano Paes Silva Junior **– Presidente** (IOC/FIOCRUZ)
Prof. Dr. Marcos Paulo Catanho de Souza (IOC/FIOCRUZ)
Prof. Dr. Juliano de Carvalho Cury (UFSJ)
Prof. Dr. Maria Claudia Reis Cavalcanti (IME)
Prof. Dr. Ana Carolina Paulo Vicente (IOC/FIOCRUZ)

Rio de Janeiro, 13 de maio de 2014

AGRADECIMENTOS

Agradeço primeiramente aos meus pais e ao meu irmão, por todo apoio incondicional que me prestam desde sempre. À Carol, por ser nesse tempo todo a minha maior amiga, companheira, cúmplice e colaboradora. Ao meu orientador, a quem hoje acredito poder chamar de amigo, Alberto, por toda a confiança depositada em mim desde o começo de minha vida acadêmica, ainda na graduação, e por todas as conversas, conselhos e até mesmo as poucas broncas, que sempre me fizeram refletir sobre qual direção tomar. Aos amigos do LBCS e de toda a BCS (prefiro não citar os nomes para não cometer injustiças por esquecimento). À doutora Yara e toda a equipe de seu laboratório, por ceder mais uma vez o espaço e os equipamentos necessários no começo do desenvolvimento deste trabalho. Aos amigos da Toca, pelos momentos de descontração, sem os quais seria impossível manter a cabeça no lugar durante esses longos 4 anos. Aos colegas do BeGenDiv em Berlim, principalmente a Camila pela indicação de orientadores e apoio prestado enquanto estive em solo germânico, e ao grande amigo Kiko, por todos os bons momentos de descontração, e por me ajudar tanto no momento de “crise” na troca de orientação no exterior. Ao Dr. Hans-Peter Grossart, por me receber tão bem como orientado na Alemanha, assumindo o lugar do Dr. Martin Allgaier (a quem também agradeço pela oportunidade) e me dando sugestões e dicas tão importantes para o sucesso desse trabalho. Às agências de fomento CAPES e CNPq, pelas bolsas fornecidas e financiamento do projeto. Ao Dr. Juliano Cury, por aceitar colaborar nos artigos dessa tese, além de assumir o papel de revisor da mesma. Aos membros da banca, por dispor de precioso tempo na leitura e revisão deste trabalho. Ao Dr. Jefferson Melges (*in memoriam*) pelo grande incentivo ao mundo digital, especialmente por me introduzir no sistema operacional “Linux”, ainda quando eu era adolescente. Meus sinceros agradecimentos.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Caracterização da diversidade microbiana da Praia dos Anjos – Arraial do Cabo – RJ através de metagenômica

RESUMO

TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Rafael Ricardo de Castro Cuadrat

Os ambientes marinhos cobrem cerca de 70% da superfície do planeta. Esses habitats apresentam uma grande variabilidade de temperatura, pressão e salinidade, abrigando uma vasta biodiversidade microbiana, pertencentes aos 3 domínios da vida (Archaea, Bacteria e Eukarya). Estes microrganismos representam até 98% da produção primária destes ambientes representando grande potencial para exploração e descoberta de novos compostos naturais de interesse para a indústria farmacêutica e da biotecnologia.

Entretanto, apenas cerca de 1% dos microrganismos podem ser cultivados com as técnicas atuais utilizando-se meios de cultura em laboratório. Com objetivo de contornar esta limitação, estudos de metagenômica vem sendo conduzidos para analisar amostras de diferentes ambientes, incluindo ambientes aquáticos como rios, lagos e regiões costeiras ou de oceano aberto.

No presente estudo, foram avaliados a diversidade taxonômica e o potencial metabólico de uma amostra (fracionada em duas por filtração, nomeadas: amostra E – retida na membrana de 0,8 µm e amostra P – retida na membrana de 0,22 µm) coletada na Praia dos Anjos (Arraial do Cabo – RJ), um ambiente de grande interesse por ser afetado pelo fenômeno da ressurgência, além de sofrer impacto antrópico (turismo e pesca). Foram também triados genes do metabolismo secundário (PKS e NRPS) de microrganismos através de pirosequenciamento do DNA total da comunidade.

Um total de 651.083 e 542.647 sequências de nucleotídeos (*reads*) foram obtidas das amostras P e E, respectivamente. As sequências obtidas foram analisadas através de similaridade com bancos de sequências públicas (Genbank) utilizando o pacote BLAST e o programa MEGAN para classificação taxonômica baseada no algoritmo do Último Ancestral Comum (LCA).

O filo mais abundante nas duas amostras foi Proteobacteria, seguido por Bacteroidetes e Cyanobacteria (este último principalmente na amostra E). Membros do clado Roseobacter (principalmente gêneros *Roseobacter* e *Ruegeria*) foram encontrados em alta abundância nas duas amostras, porém a dominância é maior na amostra P (representando até 29% dos gêneros identificados).

Através de modelos HMM (do inglês “Hidden Markov Models”), foram triadas sequências de domínios conservados ceto-acil sintase - KS e domínio de condensação – C, de enzimas PKS e NRPS, respectivamente. Um total de 84 sequências de KS e 46 sequências de domínio C foram encontradas nas duas amostras, mostrando o potencial deste ambiente para a descoberta de novos compostos de interesse para a indústria.

Adicionalmente, a abundância e diversidade de bactérias aeróbias fotossintetizantes anoxigênicas (AAP) no metagenoma de Arraial do Cabo e em metagenomas públicos do projeto GOS foi investigada através de uma metodologia *in silico* utilizando perfis de modelos ocultos de markov (pHMM) para triar os genes do núcleo de reação da fotossíntese anoxigênica (*puf*M e *puf*L), além do gene *bch*X, e através destes estimar a abundância e diversidade de AAPs em metagenomas. A amostra de maior abundância em AAPs foi a amostra P de Arraial do Cabo, com aproximadamente 23,88% do total de células presentes na amostra. Das 10 amostras do GOS mais abundantes em AAPs, 8 (80%) foram obtidas de regiões próximas a linha do equador. Foi possível classificar as sequências de *puf*M em filogrupos, mostrando alta abundância do filogrupo G (clado das *Roseobacter*) em Arraial do Cabo. Este filogrupo se mostrou o mais cosmopolita, presente em 11 das 12 (91,66%) amostras analisadas. Os resultados nos permitiram concluir que o ambiente estudado foi afetado pelo fenômeno da ressurgência, e a amostra foi coletada após o *bloom* do fictoplanton.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

MICROBIAL DIVERSITY CARACTERIZATION OF ANGEL´S BEACH – ARRAIAL DO CABO – RJ USING METAGENOMIC APPROACH

ABSTRACT

PHD THESIS IN COMPUTATIONAL AND SYSTEM BIOLOGY

Rafael Ricardo de Castro Cuadrat

The marine environments cover ~70% of the Earth's surface. These habitats present great variability of temperature, pressure and salinity, harboring a wide range of microorganisms from the three domains of life (Archaea, Bacteria and Eukarya) which are responsible for ~98% of the marine primary production. This huge biodiversity represents a great potential, as its exploration allows us to discover new enzymes for industrial use. However, only ~1% of the environmental microorganism can be cultivated using culture-dependent approaches. To overcome this limitation, metagenomics studies have been conducted using samples for different environments, including aquatic oneslike costal seawater, deep seawater, open ocean waters and freshwater from rivers and lagoons. In this work, we explore the taxonomic diversity and the metabolic potential (to find new natural compounds produced by PKS and NRPS enzymes) of the Praia dos Anjos (Angel's Beach), in Arraial do Cabo, Rio de Janeiro, Brazil by pyrosequencing its metagenome. The sample was fractionated by filtration in 2 membranes to separate the eukaryotic (sample E) from prokaryotic communities (sample P). A total of 651,083 and 542,647 reads were obtained for samples P and E, respectively. The obtained sequences were analyzed by similarity using the BLAST package and the MEGAN program, applying the Last Common Ancestral (LCA) algorithm. The MG-RAST pipeline was used to annotate the genes from the community.

The most abundant bacterial phylum present in both samples was Proteobacteria, followed by Bacteroidetes and Cyanobacteria (mainly on sample E). Members of the Roseobacter clade (*Roseobacter* and *Ruegeria* genus) was abundant in both samples, but in larger abundance on sample P (up to 29% of identified genus).

The keto-acyl synthase (KS) domains from PKSs and condensation domain (C) from NRPS were screened using pHMMs approach. A total of 84 KS sequences and 46 C sequences were obtained from both samples, showing the potential of this environment for the discovery of new compounds. The aerobic anoxygenic phototrophs bacteria (AAP) are photoheterotrophic microorganisms that play important roles on biogeochemical cycles. In oceans, this group is widely distributed, however, its abundance and relevance in carbon fixation is poorly understood. In the present work, with the aim to estimate the abundance and diversity of AAPs in the metagenome from Arraial do Cabo, an *in silico* approach using Hiden Markov Models profiles (pHMM) was developed to screen core genes of anoxygenic photosynthesis (*pufM* and *pufL*), in addition to chlorophyllide reductase subunit X gene (bchX). The metagenomes from Global Ocean Sample Expedition (GOS) was also screened with comparative purposes. The most abundant sample was sample P from Arraial do Cabo, with ~23.88% of total cells in the sample. The 10 most abundant samples from GOS, in addition to the 2 samples from Arraial do Cabo, were selected to assembly, ORF extraction and phylogenetic analysis of *pufM* genes. From the 10 GOS samples, 80% were collected in sites close to the Equador. It was possible to classify the most of sequences in phylogroups, showing a high abundance of phylogroup G (*Roseobacter* clade) in Arraial do Cabo samples. This phylogroup was the most ubiquitous, present on 11 from 12 (91.66%) assembled samples.

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

PKS	do inglês “poliketyde syntase” ou policetídeo sintase
NRPS	do inglês “non-ribosomal peptide syntase” ou peptídeo sintase não ribossomal
Pb	pares de base (nitrogenadas)
Nt	nucleotídeo
RNA	do inglês “ribonucleic acid” ou ácido ribonucleico
DNA	do inglês “deoxyribonucleic acid” ou ácido desoxirribonucleico
IBGE	Instituto Brasileiro de Geografia e Estatistica
NGS	do inglês “Next generation sequencing” ou sequenciamento de nova geração
PCR	do inglês “Polymerase Chain Reaction”ou reação em cadeia da polimerase
FAS	do inglês “Fatty Acid Synthase” ou Ácido Graxo Sintase
ATP	do inglês “Adenosine triphosphate” ou adenosina trifosfato
µm	micrômetro
PSU	do ingles “Practical Salinity Units”
HMM	do inglês “Hidden Markov Models” ou modelos ocultos de Markov
AAP	do inglês “Aerobic Anoxygenic Phototrophs” ou aeróbias fotossintetizantes anoxigênicas
ACP	proteína carreadora do grupamento acil
KS	ceto-acil sintase
AT	Acil transferase
DH	β-hidróxi dehidratase
KR	β-ceto-acil redutase
ER	enoil redutase
A	domínio de adenilação
C	domínio de condensação
PCP	proteína carreadora do grupamento peptidil
TE	domínio tioesterase
MT	domínio metiltrasnferase
E	domínio de epimerização
PEP	domínio fosfoenolpiruvato

PCP	proteína carreadora de grupo pepitdil
GOS	Global Ocean Sampling Expedition
NADPH	do ingles “nicotinamide adenine dinucleotide phosphate-oxidase” ou nicotinamida adenina dinucleótido fosfato
RC	do ingles “Reaction Center” ou núcleo de reação
CON	domínio de condensação
CYC	domínio de ciclização

1 INTRODUÇÃO

Biodiversidade marinha e a biotecnologia industrial

Estudos demonstram que em ambientes marinhos existem aproximadamente $3{,}67\times10^{30}$ células microbianas (Whitman *et al.* 1998). Estima-se que a abundância de bactérias seja de até 10^{6} células por mililitro de água na zona pelágica marinha, representando a maior parte da biomassa oceânica (Azam *et al.* 1998). Esta gigantesca biodiversidade possui grande potencial biotecnológico, pois seu estudo permite a descoberta de novas enzimas de interesse para a indústria.

Os ambientes marinhos são muitos diversos e os microrganismos que os habitam são expostos a extremos de pressão, temperatura, salinidade e disponibilidade de nutrientes. Os diferentes nichos marinhos possuem comunidades bacterianas únicas e muito distintas, adaptadas às mais diferentes situações. Isto leva a uma grande diversidade bioquímica, ainda pouco explorada (Kennedy *et al.* 2008).

O estudo dessa diversidade bioquímica, principalmente originada do metabolismo secundário de microrganismos, é de extrema importância para diversos tipos de indústria, como por exemplo, a farmacêutica e a alimentícia. Uma grande quantidade de metabólitos secundários vem sendo estudados, e muitos já estão no mercado, como antibióticos, antitumorais, imunossupressores, pigmentos alimentícios, etc. (Castoe *et al.* 2007).

Porém, por limitações técnicas, a maior parte dos compostos descobertos são originados de microrganismos cultiváveis, que compõem a minoria da biodiversidade microbiana existente, pois estima-se que apenas entre 0,001 a 0,1% dos microrganismos presentes nos oceanos, seja atualmente cultivável em laboratório (Pace *et al*., 1997; Tringe *et al.,* 2005).

Para tentar superar esta limitação, foram desenvolvidas técnicas de estudo dos ácidos nucléicos independente de cultivo dos organismos (chamados coletivamente de metagenômica ou genômica ambiental), como o estudo de biodiversidade baseado no gene codificador da menor subunidade ribossomal do RNAr (16S nos procariotos e 18S nos eucariotos), a construção e sequenciamento de grandes bibliotecas de DNA ambiental ou ainda o sequenciamento direto do DNA ambiental utilizando tecnologias de sequenciamento de alta vazão como Roche 454 ou Illumina HiSeq.

Através destas técnicas é possível não apenas realizar inferências sobre quais são os microrganismos existentes em diversos ambientes, como também estudar o metabolismo primário e secundário dos organismos de um determinado ambiente. Através do estudo do metabolismo primário dos microrganismos é possível entender o funcionamento e a participação dos mesmos nos ciclos biogeoquímicos como o ciclo do carbono, enxofre ou nitrogênio (Balvanera *et al*., 2006; Zarraonaindia *et al*., 2013). Ainda, estudando o metabolismo secundário, é possível descobrir novas enzimas e compostos com potencial biotecnológico e farmacêutico como policetídeos e peptídeos não-ribossomais (Kennedy *et al.* 2008; Schirmer *et al.,* 2005).

Com o crescente número de estirpes microbianas resistentes às drogas existentes no mercado, cada vez mais se faz necessária a descoberta de novos fármacos (Tillotson *et al.,* 2013). De acordo com a Sociedade de Doenças Infecciosas da América (IDSA, 2010), pelo menos 10 novos antibióticos são necessários para contornar as bactérias super-resistentes que vêm causando novas epidemias de infecção hospitalar.

Entretanto, apesar de terem sido desenvolvidos mais de 20 novas classes de antibióticos entre 1930 e 1962 (Coates *et al.,* 2002; Coates *et al.*, 2011), desde então apenas mais duas classes foram descobertas (Zappia *et al.*, 2007).

A maneira mais tradicional, e de maior sucesso, empregada para o desenvolvimento de novos antibióticos tem sido a descoberta de novos produtos naturais (de plantas, bactérias, fungos, etc.) e em muitos casos, a modificação química dos mesmos (Singh & Barret, 2006). Em poucos casos, foram desenvolvidos compostos sintéticos (Fernandes, 2006).

Dentre os organismos marinhos, aqueles que compõem o bacterioplâncton são os mais abundantes em compostos de interesse (Desriac *et al*., 2013). Na maior parte dos ambientes marinhos, o grupo taxonômico mais abundante no bacterioplanctôn são as bactérias do filo Proteobacteria (classe alfa-proteobacteria) (Joint *et al.,* 2010*)*. Entretanto, a maior parte dos novos compostos vem sendo descobertos no filo Actionobacteria, que compõe em média apenas 5-10% do bacterioplanctôn (King *et al*., 2012; Lau *et al*., 2013).

O litoral brasileiro e a Região dos Lagos

A costa do Brasil estende-se por 7.491 quilômetros e é influenciada tanto pela corrente quente do Norte do país quanto pela corrente fria das Ilhas Malvinas, na região sul. Possui uma grande diversidade de ecossistemas, como mangues, praias arenosas, recifes de corais, dunas, lagunas e estuários (Prates *et al*., 2007).

No litoral do estado do Rio de Janeiro (região sudeste do Brasil), está localizada a região conhecida como Região dos Lagos, localizada ao norte do município do Rio de Janeiro, na mesorregião das Baixadas Litorâneas. É formada por 7 municípios em torno das lagoas de Araruama e de Saquarema. Esta região é de grande importância econômica para o Estado do Rio de Janeiro, por sua intensa atividade turística e pesqueira.

Um destes 7 municípios, chamado de Arraial do Cabo, foi fundado em 1985 após a emancipação de Cabo Frio e possui cerca de 27 mil habitantes, segundo dados de 2008 do IBGE. Esta região, além de influenciada por atividade antrópica, é também influenciada por correntes frias, oriundas do fenômeno da ressurgência.

A região portuária deste município, localizada na Praia dos Anjos, atende tanto a barcos de pescadores quanto a barcos de turismo, utilizados para passeios pelas diversas praias do município, algumas acessíveis apenas por via aquática ou pequenas trilhas.

O presente estudo centra-se especificamente nesta região da bacia da Praia dos Anjos, previamente estudada também por Cury e colaboradores (Cury *et al*., 2011), utilizando técnicas moleculares (amplificação de genes ribossomais) para caracterizar a biodiversidade da região. No estudo de Cury *et al.* (2011), além da região portuária (referida pelo autor como PO), foi também estudada uma região em alto mar, influenciada diretamente pela ressurgência (referida pelo autor como RE), para fins comparativos e caracterização das mudanças provocadas pelo fenômeno.

Figura 1.1 mostra o mapa da região portuária de Arraial do Cabo, obtida no Google Maps, com o ponto de coleta utilizado no presente estudo (PO).

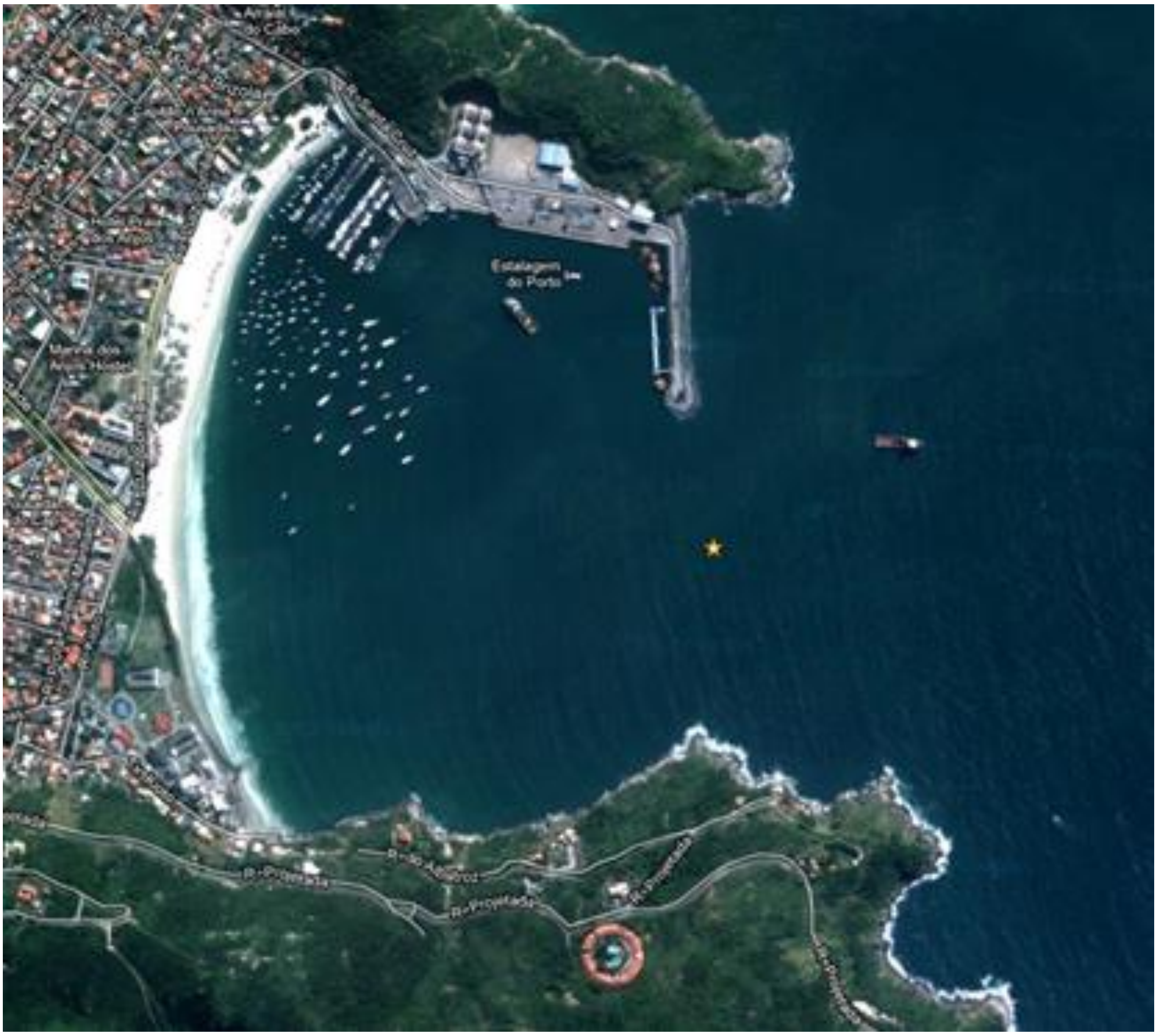


Figura 1**.1 - Mapa da Praia dos Anjos (Arraial do Cabo - RJ).** A região onde a amostra foi coletada está marcada com uma estrela (Fonte: Google Maps)

Toda esta região é afetada pelo fenômeno da ressurgência (Valentin *et al*. 1984; Rodrigues *et al*., 2001), em que as águas frias do fundo sobem para a superfície, levando nutrientes para a zona eufótica e por consequência, induzindo um *bloom* do fitoplanctôn, com consequente aumento da biomassa (Cury *et al.,* 2011). O *bloom* consiste em grande multiplicação dos organismos do fitoplanctôn (micro-organismos com capacidade fotossintética), que aproveitam a abundância de nutrientes oriundos das águas profundas e da luz abundante na zona eufótica.

Na região de Cabo Frio e Arraial do Cabo, a Água Central do Atlântico Sul (ACAS) sobe para a superfície graças a ação dos ventos leste-nordeste, que ocorrem mais frequentemente entre primavera e o verão (Valentin *et* al., 1987; Campos *et al*., 2000; Castelão *et al*., 2006; Pereira *et al*., 2008).

As regiões afetadas por ressurgência são de grande importância econômica, pois correspondem a cerca de 25% da produção mundial de pescado, mesmo representando apenas cerca de 5% da área dos oceanos (Jennings *et al*., 2001). Por este motivo, estas regiões, especialmente onde pode ocorrer poluição antrópica, devem ser monitoradas visando a sua conservação.

Metagenômica

A metagenômica é uma abordagem que surgiu na década de 90, como forma de estudar os ácidos nucléicos de organismos não cultiváveis. Diversas estratégias vêm sendo desenvolvidas, com diferentes objetivos, como por exemplo, o estudo da biodiversidade de um determinado ambiente, ou a localização de genes ou *clusters* metabólicos responsáveis pela síntese de compostos de interesse biotecnológico.

As estratégias que mais vem sendo utilizadas são as seguintes: (i) estudo dos genes RNAr: 16S para os procariotos e 18S para os eucariotos; (ii) clonagem de DNA ambiental em bibliotecas: grandes fragmentos com óperons inteiros podem ser clonados e expressados em vetores como fosmídeos; (iii) sequenciamento direto do DNA ambiental (Figura 1.2).

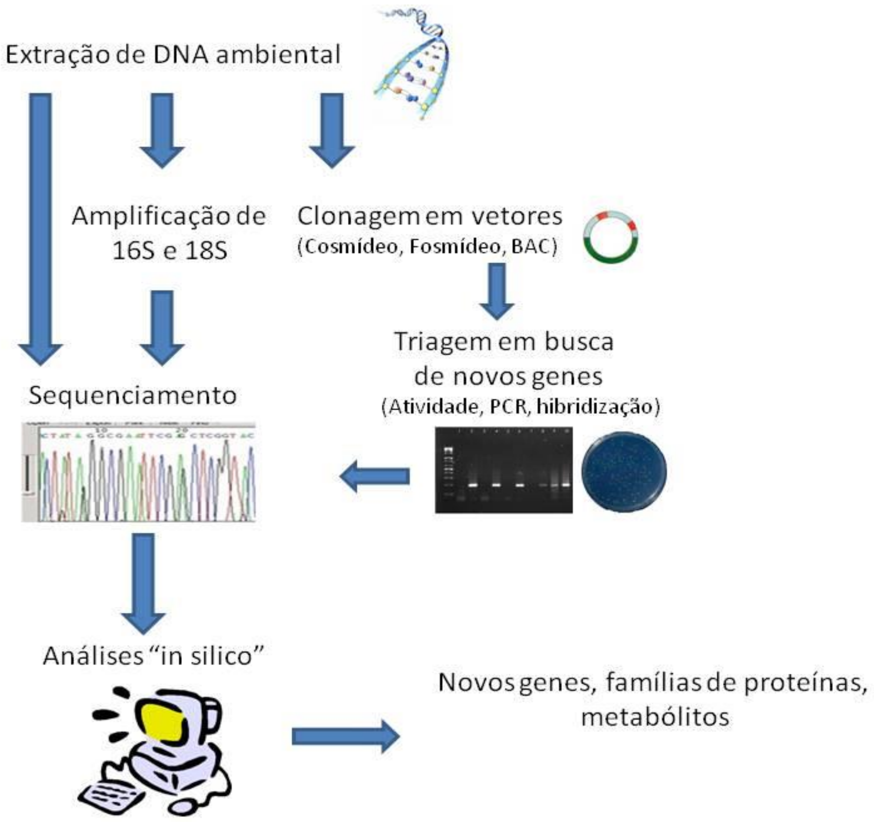

Extração de DNA ambiental
Amplificação de 16S e 18S
Clonagem em vetores (Cosmídeo, Fosmídeo, BAC)
Triagem em busca de novos genes (Atividade, PCR, hibridização)
Sequenciamento
Análises "in silico"
Novos genes, famílias de proteínas, metabólitos

Figura 1.2 - Fluxograma ilustrando algumas estratégias utilizadas em metagenômica.
Fonte: Dissertação de Mestrado de Rafael Cuadrat, 2010

Diversos estudos metagenomicos ao redor do mundo foram realizados e grandes volumes de dados foram gerados a partir do sequenciamento de DNA ambiental, primeiro utilizando sequenciamento com a técnica de Sanger, como por exemplo o metagenoma do intestino humano publicado por Gill *et al.* (2006) e o projeto "*Global Ocean Sampling Expedition (GOS)*" publicado por Rusch *et al.* (2007). Com esta técnica era possível gerar sequências com tamanho entre 400 e 700 pares de base, porém o alto custo e a pequena vazão de dados se mostrou inadequada para avaliar ambientes com grande diversidade. Com o surgimento de tecnologias de nova geração para sequenciamento (NGS) estes estudos se tornaram mais viáveis. As tecnologias NGS incluem Roche/454 Life Sciences (GS20, FLX, LXR); Illumina/Solexa (Illumina G2) e Applied Biosystems (SOLiD). Diferentes gerações destas tecnologias surgiram com diferentes performances, variando desde a vazão e o número de bases geradas por rodada, até o comprimento total da sequência gerada (tamanho do *read*). A tecnologia que atinge o maior comprimento de *read* e que vem sendo mais utilizada para metagenomas é a da Roche (454). Na geração atual (GS FLX+), o tamanho total do *read* pode atingir 1.000 pares de base (média de 700 pares de base) e a vazão total pode atingir 700 Mb (milhões de pares de base) por corrida (http://454.com/products/gs-flx-system/index.asp). Com *reads* grandes se torna mais fácil a montagem de genes e predição de regiões codificantes, facilitando a triagem dos dados, por exemplo, em busca de genes que codifiquem enzimas de interesse biotecnológico.

Existem diversos bancos de dados públicos que disponibilizam sequências e metadados de metagenômica em diversos ambientes como por exemplo o IMG/M (http://img.jgi.doe.gov/cgi-bin/m/main.cgi) que na versão atual (4.0) (Markowitz *et al*., 2014) possui 3328 datasets de 460 projetos de metagenomica de ambientes em várias partes do mundo e de diferentes tipos de ambientes como água do mar, solo, ar, bioreatores etc. Já o CAMERA (http://camera.calit2.net/) (Sun*et al*., 2011) e o MG-RAST (http://metagenomics.anl.gov/) (Meyer *et al*., 2008), além de banco de dados, oferecem também *pipelines* para análises de metagenomas *online*.

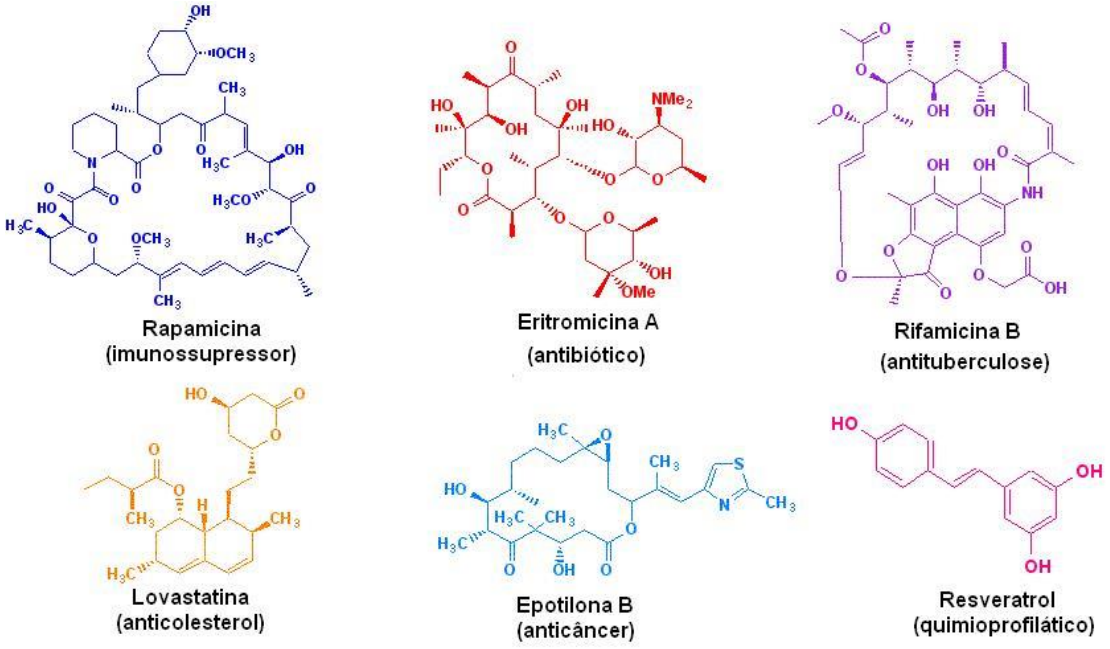
No presente estudo foi utilizada a abordagem de sequenciamento direto do DNA ambiental, através de pirosequenciamento (454 ROCHE FLX+), com objetivo de caracterizar a diversidade microbiana e avaliar o potencial biotecnológico da Praia dos Anjos, em Arraial do Cabo – RJ. A partir dos resultados prévios obtidos por Cury e colaboradores, onde pode-se observar maior abundância de Actionobactérias em PO do que em RE, foi decidido estudar o ponto PO, pois este filo bacteriano é conhecidamente muito rico em metabólitos secundários com atividades de interesse para a indústria farmacêutica (Policetídeos e Peptídeos Não Ribossomais).

Enzimas de interesse biotecnológico

Policetídeo Sintases (PKSs)

PKSs são enzimas que produzem um grande grupo de metabólitos secundários chamados de policetídeos. Esses metabólitos possuem diversas aplicações na indústria, sendo muitos de importância médica. Dentre os principais, podemos citar compostos com atividade antimicrobiana como, por exemplo, a

Eritromicina; imunossupressora como a Rapamicina; antiparasitária como Avermectina; e até mesmo toxinas prejudiciais à saúde humana, como a Aflatoxina (Casto *et al.,* 2007). Na Figura 1.3 pode-se observar a estrutura bioquímica de alguns policetídeos.



Rapamicina (imunossupressor)

Eritromicina A (antibiótico)

Rifamicina B (antituberculose)

Lovastatina (anticolesterol)

Epotilona B (anticâncer)

Resveratrol (quimioprofilático)

Figura 1.3 - Estrutura química de alguns policetídeos com atividade de interesse medico.
Fonte: Adaptado de http://linux1.nii.res.in/~pksdb/polyketide.html

Estes metabólitos tem sido encontrados em diversos organismos como bactérias, fungos, plantas, insetos, dinoflagelados, moluscos e esponjas (Gokhale*et al.,* 2007).

As PKSs possuem similaridade (tanto em sequência quanto em estrutura) com as enzimas responsáveis pela produção de ácido graxo, denominadas Ácido Graxo Sintases (FASs). Ambas tipicamente catalisam sucessivas condensações de unidades simples de carbono (grupos acil-coA, geralmente acetil-coA e malonil-coA), para construir uma cadeia cetônica**.** Contudo, na biossíntese de ácidos graxos acontece a completa redução dos grupos cetônicos, com a produção de cadeias de carbono completamente reduzidas, enquanto nos policetídios a cadeia permanece parcialmente ou não reduzida (Castoe*et al.,* 2007).

**As PKSs podem ser classificadas quanto ao tipo (
Tabela 1.1):**

Tipo I - grandes enzimas multifuncionais, multidomínios, que possuem todas as atividades enzimáticas necessárias para o alongamento e processamento da cadeia policetídica. Em alguns casos a biossíntese de policetídeos por PKSs tipo I é realizada por mais de uma proteína e os genes codificantes estão organizados em grupos (*clusters*) como, por exemplo, o grupo de três genes responsáveis pela produção da PKS que sintetiza a Eritromicina (Cane *et al.,* 1998; Lal *et al.,* 2000).

As PKSs tipo I podem ser modulares (geralmente em bactérias) ou iterativas (geralmente em fungos, porém presentes em algumas bactérias).

Nas modulares, cada enzima inclui um ou mais módulos, e cada módulo é responsável por um turno de condensação e processamento da cadeia. Cada domínio catalítico nas PKSs modulares é utilizado apenas uma vez na biossíntese do policetídeo. Estas PKSs sintetizam policetídeos macrocíclicos, através da condensação de acetatos, propionatos e butiratos. O nível de redução dos grupos beta-carbonil realizado em cada ciclo de condensação é variável. Os policetídeos macrocíclicos são os de maior importância clínica. Avanços nos estudos sobre as PKSs modulares realizados nas últimas décadas vêm demonstrando ser possível realizar recombinações entre os módulos destas enzimas, modificando a estrutura e função do policetídeo, o que gera um grande potencial de produção de novos compostos (Rup *et al.,* 2000).

As PKSs tipo I iterativas possuem um único módulo, que realiza vários turnos de alongamento da cadeia, utilizando cada domínio várias vezes durante a biossíntese. Estas PKSs catalisam a formação de policetídeos aromáticos, como por exemplo, o ácido 6-metilsalicíclico (Shen, 2003).

Tipo II - Ao contrário das tipo I, a atividade enzimática para o alongamento e processamento da cadeia é realizada por enzimas separadas, codificadas por diferentes ORFs (cada uma homologa à um domínio das PKS do tipo I) e cada ORF é utilizada de maneira iterativa (Castoe *et al.,* 2007).

Tipo III - Responsáveis pela produção de Chalcona em plantas e polihidroxi-fenois em bactérias. Diferentemente dos outros tipos de PKS, na tipo III a cadeia é alongada e processada em um único e sítio ativo multifuncional. Não existe domínio cerreador do grupo acil (ACP) neste tipo de PKS, atuando a mesma diretamente nos grupos acil-coA (Castoe *et al.,* 2007).

Os domínios catalíticos (no caso das tipo I) ou enzimas (no caso das tipo II) presentes nas PKSs são: ceto-acil sintase (KS), proteína carreadora do grupo acil (ACP), acil transferase (AT), cetoredutase (KR), deidratase (DH), tioesterase (TE), enoil redutase (ER), metil trasferase (MT), claisen ciclase (CYC) e domínio de condensação (CON). Os domínios essenciais para uma PKS modular mínima são KS, ACP e AT. Estes, além do domínio TE, realizam reações de condensação da cadeia. Já os domínios KR, DH e ER são responsáveis por reações de redução enquanto MT, CYC e CON realizam modificações pós-condensação (www.rasmusfrandsen.dk/ny_side_8.htm).

Tabela 1.1 - Resumo dos tipos de PKSs existentes com suas respectivas estruturas, mecanismos e distribuição nos organismos.
Adaptado de Watanabe & Ebizuka 2004

Tipo I modular	**Proteína única com múltiplos módulos e múltiplos domínios**	**Linear, cada sítio ativo utilizado uma única vez**	**Bactérias**
Tipo I iterativa	Proteína única com único módulo e múltiplos domínios	Iterativo, cada sítio ativo utilizado várias vezes	Fungos e bactérias
Tipo II	Múltiplas proteínas, cada uma com um domínio ativo	Iterativo, cada sítio ativo utilizado uma ou mais vezes	Bactérias
Tipo III	Proteína única com múltiplos módulos	Iterativo, cada sítio ativo utilizado várias vezes	Plantas e bactérias

Alguns autores sugerem que a diversidade de PKSs é muito maior em termos de mecanismo e estrutura do que se pode classificar com o sistema de tipagem atual. Um exemplo é a atividade não iterativa exibida durante a catálise de um policetídeo por enzimas codificadas por um grupo de genes para PKS clonado a partir de *Streptomyces griseus*. Esta, além de não possuir atividade iterativa, como as PKS tipo II (apesar de ser composta por múltiplas enzimas de domínio único), não utiliza domínio ACP, sendo desta forma, similar as PKS tipo III. Todavia, seus domínios KS são claramente homólogos aos de tipo I e II, não podendo assim ser classificada como tipo III (Shen, 2003).

A maioria dos genes da família das PKSs e NRPSs foi encontrada em genomas de Actinobacteria, que, como dito anteriormente, constituem apenas cerca de 5 a 10% do microbioma total dos oceanos (Jamieson *et al*., 2012; King *et al*., 2012; Lau *et al., 2013*). Entretanto, o filo das proteobactérias, que constitui a maior parte do microbioma destes ambientes, possui também pontencial de fornecer compostos de interesse biotecnológico, produzidos por PKSs e NRPSs (Milne *et al*., 1998; Grossart *et al.,* 2004; Slightom *et al*., 2009; Cude *et al.,* 2012; Desriac *et al*., 2013). A famiília das Rhodobacteraceae (alfa-proteobactérias) constitui o principal grupo com potencial de fornecer novos compostos, e no estudo de Cury e colaboradores (2011), esta foi a família de maior abundância encontrada no mesmo local estudado nesta tese.

Diversos estudos vem sendo realizados buscando novas PKSs em diferentes ambientes com ajuda da metagenômica, como por exemplo em esponjas marinhas (Kennedy *et al.* 2008, Schirmer *et al.* 2005) e em solos (Courtois *et al.* 2003, Wawrik *et al.* 2005). A maior parte dos estudos foca-se nos domínios KS, pois os mesmos são considerados conservados, e por isso é mais fácil o desenho de iniciadores de reação em cadeia da polimerase (PCR) e também a realização de estudos filogenéticos (Parsley *et al.,* 2011;Trindade-Silva *et al*., 2013). Apenas um estudo foi conduzido até hoje com objetivo de triar *in silico* metagenomas sequenciados por *shotgun*, realizado por Foerstner e colaboradores (2008).

Através de estudos filogenéticos dos domínios KS, como por exemplo o conduzido por Jenke-Kodama e colaboradores em 2005, é possível determinar o tipo de PKS (I ou II), se a PKS é modular ou iterativa, se pertence a uma enzima híbrida (PKS-NRPS) e também separá-las das homologas FAS.

Diversas ferramentas foram desenvolvidas para triagem e classificação dos domínios KS, sendo a mais recente, desenvolvida por Ziemert e colaboradores, chamada de *Natural Product Domain Seeker* - NapDos (http://napdos.ucsd.edu/)

(Ziemert *et al*., 2012). Esta ferramenta se mostrou a mais eficiente na classificação de sequencias incompletas (muito encontradas em metagenomas).

Porém, em nosso conhecimento, até o presente momento, este é o primeiro estudo a explorar a diversidade destas enzimas em metagenoma aquático de um ambiente afetado por ressurgência através do pirosequenciamento total do metagenoma.

Peptídeo Sintase Não Ribossomal (NRPS)

Peptídeos não ribossomais são metabólitos secundários produzidos por uma ampla gama de microrganismos como cianobactérias e actinobactérias. Estes metabólitos possuem atividades diversas como antimicrobiana, anti-viral e citotóxica (Nagle & Gerwick, 1995). São produzidos por enzimas modulares chamadas peptídeo sintase não ribossomal (NRPS). Cada módulo é dividido em domínios (Desriac *et al*., 2013) envolvidos na ativação e condensação dos aminoácidos (além de funções acessórias) formando peptídeos (Silva-Stenico *et al.,* 2010). O núcleo proteico mínimo de uma NRPS consiste em um domínio de adenilação (A), para seleção e ativação dos aminoácidos utilizados como monômeros na formação do peptídeo, um domínio de condensação (C), que catalisa a formação de ligações peptídicas entre os aminoácidos e por último um domínio de tiolização (T) ou proteína carreadora de grupo pepitdil (PCP) que transfere os monômeros para a cadeia em formação (Schwarzer *et al.,* 2001, Mootz *et al.,* 2002). Além destes, o domínio TE, que é responsável pela liberação da cadeia peptídica em formação, está presente no módulo de terminação. Muitos peptídeos não ribossomais são sintetizados seguindo a regra de colinearidade, onde o número e a ordem dos módulos no genoma representam a ordem do encadeamento dos aminoácidos no produto final (Figura 1.4) (Schwarzer *et al*., 2001). Porém, alguns peptídeos fogem dessa regra, seguindo o modelo de iteratividade visto também em PKSs, onde um único módulo é utilizado várias vezes durante a síntese, ou o modelo de não-linearidade, onde a organização tradicional dos domínios (C-A-PCP) pode variar. Nos dois últimos casos, a predição do produto final a partir de dados genômicos se torna mais complicada (Desriac *et al*., 2013).

Além dos domínios do núcleo proteico mínimo, podem estar presentes domínios adicionais como o de Epimerização (E), Heterociclização (Cy), Oxidação

(Ox), Metilação (M) e Formilização (F). Estes domínios acessórios contribuem para a diversificação estrutural dos produtos obtidos (Caboche *et al*., 2008).

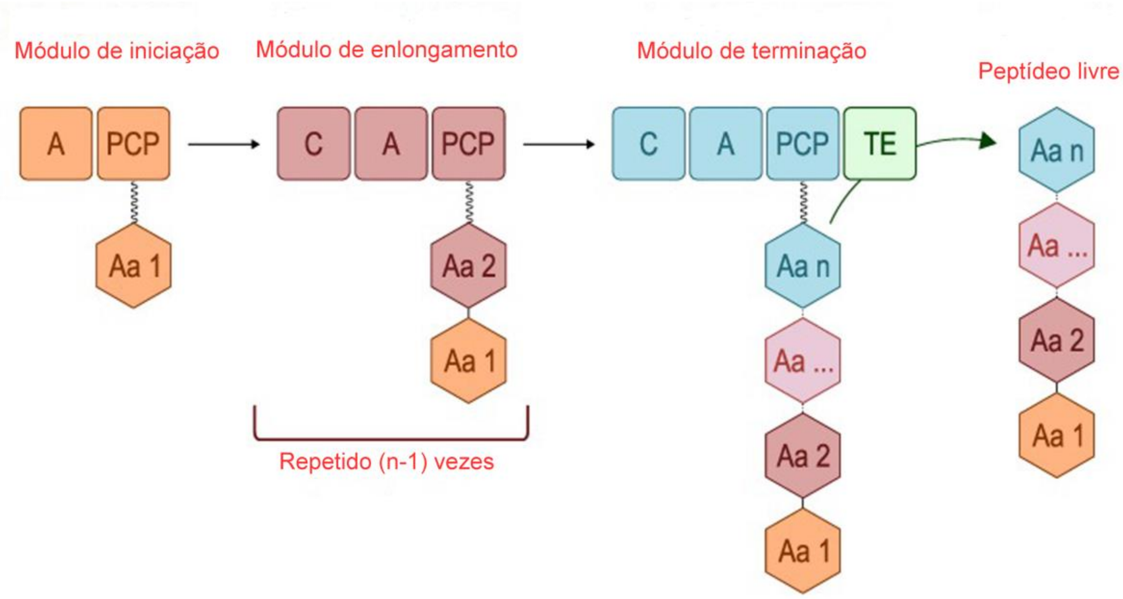


Figura 1.4 - Diagrama representando a síntese do peptídeo não ribossomal. Aa: aminoácido; A: domínio de adenilação; C: domínio de condensação; PCP: proteína carreadora do grupamento peptidil; e TE: domínio tioesterase (Adaptado de Desriac *et al*., 2013).

Existem ainda enzimas híbridas, formadas por domínios de NRPS e PKS em uma mesma janela aberta de leitura (ORF). Um exemplo de produto sintetizado por enzima híbrida é o antibiótico Rapamicina (Schwarzer *et al.* 2001). A Figura 1.5 ilustra a organização de uma enzima híbrida.

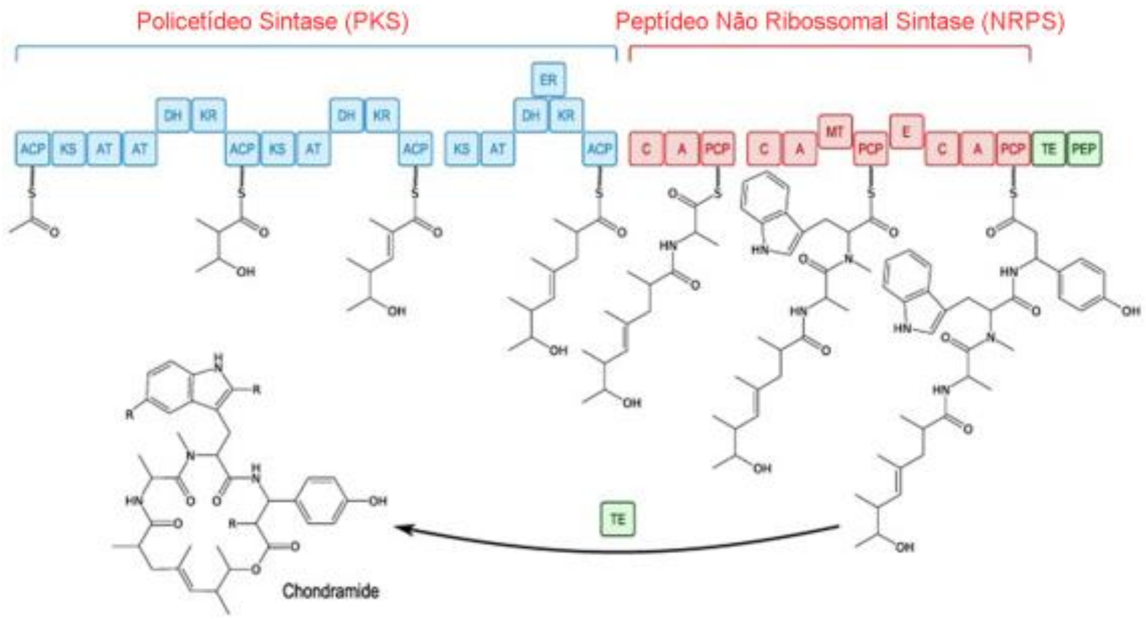


Figura 1.5 - Diagrama representando a produção de um metabólito híbrido por uma enzima NRPS-PKS.
ACP: proteína carreadora do grupamento acil; KS: ceto-acil sintase; AT: Acil transferase; DH: β-hidróxi dehidratase; KR: β-ceto-acil redutase; ER: enoil redutase; A: domínio de adenilação; C: domínio de condensação; PCP: proteína carreadora do grupamento peptidil; TE: domínio tioesterase; MT: domínio metiltrasnferase; E: domínio de epimerização; PEP: domínio fosfoenolpiruvato (Adaptado de Desriac *et al.*, 2013).

Assim como o domínio KS em PKSs, o domínio C de NRPS é o mais utilizado para triagens e estudos filogenéticos por possuir um grau mais alto de conservação que os demais domínios (Ziemert et al., 2012).

Fotossíntese Bacteriana

A fotossíntese pode ser definida como o processo de redução de gás carbônico (CO_2) em biomassa usando energia provida pela luz. A redução biológica de CO_2 requer adenosina trifosfato (ATP) e elétrons, que podem ser providos na forma de NADPH ou ferrodoxina reduzida. Entretanto, a fonte de elétrons é específica de cada organismo e pode ser água (H_2O), gás sulfídrico (H_2S), hidrogênio (H_2) ou outros compostos inorgânicos reduzidos. Neste processo, a luz inicia a transferência de elétrons através da oxidação da clorofila e da redução de um aceptor de elétrons. Uma transferência secundária de elétrons se inicia (sem necessidade de luz) levando a produção de uma força motriz protônica que é utilizada na produção de ATP (Donald *et al*., 2006).

O processo de fotossíntese utilizando clorofila muito possivelmente se originou em bactérias. Até o fim da década de 1970, eram conhecidos seis grupos bacterianos capazes de realizar fotossíntese baseada em clorofila: Cianobactérias, Bactérias Púrpuras Sulfurosas (Gama-proteobacteria da ordem Chromatiales) e Não Sulfurosas (Alpha- e Beta-proteobacteria), bactérias verdes sulfurosas (Chlorobi), bactérias fototróficas anoxigênicas filamentosas (Chloroflexi), heliobacteria e acidobacteria (Hohmann-Marriott & Blankenship, 2011).

Existem 2 tipos de fotossíntese: oxigênica, realizada por cianobactérias, algas e plantas (que resulta na produção de oxigênio) e anôxigênica, sendo esta última a forma mais ancestral de fotossíntese descrita (estima-se ter surgido aproximadamente 3 bilhões de anos atrás, quando a atmosfera terrestre não possuía oxigênio) (Rye & Holland, 1998; Xiong & Bauer, 2002; Raymond & Blankenship, 2004).

A fotossíntese anoxigênica é exclusiva de bactérias e se baseia no pigmento bacterioclorofila. Inicialmente acreditava-se que todas as bactérias fotossintetizantes anoxigênicas estavam presentes apenas em ambientes anaeróbicos, como na atmosfera primitiva terrestre, porém em 1979 foi descrita a primeira bactéria fotossintetizante anoxigênica aeróbia (AAP) (Shiba *et al*., 1979), constituindo este o sétimo grupo de bactérias fotossintéticas. A tabela 1.2 mostra os 7 grupos de bactérias fotossintetizantes conhecidos, com seu tipo de fotossíntese e respiração, e sua classificação taxonômica.

Tabela1.2 - Resumo dos grupos de bactérias fotossintetizantes

Grupo	Tipo de fotossíntese	Respiração	Filo
Cianobactéria	Oxigênica	Aeróbia	Cianobactéria
Bactérias púrpuras	Anoxigênica	Anaeróbia	Proteobactéria (alfa, beta e gama)
Bactérias verdes sulfurosas	Anoxigênica	Anaeróbia	Chlorobi
Bactérias anoxigênicas filamentosas	Anoxigênica	Anaeróbia	Chloroflexi
Heliobacteria	Anoxigênica	Anaeróbia	Firmicutes
Acidobacteria	Anoxigênica	Anaeróbia	Acidobacteria
AAPs	Anoxigênica	Aeróbia	Protobactéria (alfa, beta e gama)

Bactérias fotossintetizantes anoxigênicas aeróbias (AAPs)

As AAPs são bactérias que crescem fotoheterotroficamente e necessitam de oxigênio não apenas para crescer, como para sintetizar seu aparato fotossintético. O pigmento utilizado por elas é exclusivamente a bacterioclorofila A (bchl A). Em sua maioria, as AAPs pertencem ao grupo das Alpha-proteobacterias, com poucas espécies de Beta e Gama-proteobacteria (Csotonyi *et al*., 2001, Hunter *et al*., 2009). Alguns estudos especulam a hipótese de que elas se originaram evolutivamente das bactérias púrpuras não sulfurosas (Hunter *et al*., 2009), sendo muitas vezes difícil de separa-las das mesmas, e até mesmo de espécies não fotossintéticas, por marcadores filogenéticos como os genes do RNAr (Yurkov & Hughes, 2013).

O aparato fotossintético das AAPs permanece bastante similar ao das bactérias púrpuras, com um centro de reação (RC) ligado a um complexo de captação de luz (LH1), podendo ainda ter um segundo complexo opcional (LH2).

As AAPs são consideradas de grande interesse por estarem envolvidas em ciclos de carbono e energia. Evidencias genômicas e caracterizações fisiológicas tem demonstrado que as AAPs possuem um grande potencial metabólico, incluindo nitrificação, fixação de dióxido de carbono, produção de carotenóides, utilização de carbono de baixo peso molecular como fonte de energia, etc. Elas foram descritas inicialmente em águas costeiras de oceanos, porém hoje se sabe que elas habitam diversos tipos de ambientes terrestres e aquáticos (água doce e salgada) (Shiba *et al*., 1979; Beja *et al*., 2002; Csotonyi *et al.*, 2010; Atamna-Ismaeel *et al*., 2012). Em oceanos, este grupo taxonômico está amplamente distribuído (Kolber *et al*., 2000; 2001), entretanto, a abundância deste grupo nestes ambientes e a importância dos mesmos nos ciclos biogeoquímicos é ainda pouco compreendida (Goericke, 2002; Schwalbach & Fuhrman, 2005).

Diversos estudos foram realizados com intuito de estimar a abundância de AAPs em ambientes marinhos, utilizando diferentes técnicas, como detecção de Bchl A por fluorescência (Kolber *et al*.,2000; 2001), amplificação de genes marcadores por PCR em tempo real (Schwalbach & Fuhrman, 2005) e metagenômica (Beja *et al*., 2002; Oz *et al*., 2005; Waidner & Kirchman, 2005; Yutin *et al*., 2005), porém com resultados contraditórios (Yutin *et al.*, 2007). No estudo realizado por Yutin e colaboradores, os metagenomas do *GOS* (Rusch *et al*., 2007) foram triados para caracterização da abundância e distribuição das AAPs em ambientes marinhos, encontrando grande variabilidade de AAPs e de abundância relativa das mesmas nos diversos ambientes (entre menos de 1% a 10% do total de células do ambiente estudado). Entretanto, em um estudo de um ambiente muito oligotrófico no pacífico sul, as AAPs constituem cerca de 24% do total de células procarióticas (Lami *et al*., 2007), sugerindo que diversas variáveis do ambiente podem influenciar na abundância e diversidade destas bactérias.

Os genes marcadores mais utilizados nas triagens em busca de AAPs são o *puf*M e *puf*L, que estão localizados no operon *puf* e codificam subunidades do centro de reação (RC). Porém, estes genes são bem conservados entre AAPs e bactérias púrpuras anaeróbias, necessitando assim análises filogenéticas para diferencia-los. Além destes, o gene *bch*X (codificador de "*Chlorophyllide*" reductase) também já foi utilizado, em adição os genes do operon *puf*, por Yutin e colaboradores em seu estudo para estimar a abundância e diversidade de AAPs nos metagenomas do GOS (Yutin *et al*., 2007).

No presente estudo, os 3 genes supracitados foram utilizados como marcadores para abundância e diversidade de AAPs no metagenoma de Arraial do Cabo, e nos metagenomas do GOS, com fins comparativos.

2 OBJETIVOS

Objetivo Geral

Explorar a diversidade taxonômica e metabólica microbiana da Praia dos Anjos, Arraial do Cabo – Rio de Janeiro através de metagenômica (pirosequenciamento do DNA ambiental).

Objetivos Específicos

- Caracterizar taxonomicamente e funcionalmente a comunidade microbiana da superfície aquática da Praia dos Anjos (Arraial do Cabo – RJ) através de metagenomica.

- Triar genes do metabolismo secundário, PKS e NRPS, mostrando o potencial da comunidade para prover novos genes de interesse biotecnológico na comunidade microbiana da Praia dos Anjos.

- Estimar a abundância e diversidade de bactérias fotossintetizantes anoxigenicas aeróbias (AAP) no metagenoma estudado e em metagenomas públicos para fins comparativos.

3 TRABALHO 1: “METABOLIC AND MICROBIAL DIVERSITY EXPLORED BY METAGENOMIC ANALYSIS OF UPWELLING AFFECTED BRAZILIAN COASTAL SEAWATER REVEALS SEQUENCE DOMAINS OF TYPE I PKS AND NRPS”

O primeiro trabalho desenvolvido para esta tese é intitulado “Microbial, metabolic diversity and genes of PKS and NRPS revealed by metagenomic analysis of Brazilian coastal seawater“

Este trabalho foi submetido à revista Plos one no dia 26 de março de 2014.

Neste estudo exploramos a diversidade de micro-organismos presentes em uma amostra superficial da Praia dos Anjos – Arraial do Cabo coletada no verão, após o fenômeno da ressurgência, atráves de pirosequenciamento e análises de bioinformática. Adicionalmente, exploramos a diversidade dos domínios mais conservados das famílias de enzimas Policetídeo Sintases (PKS) e Peptídeo Sintases Não Ribossomais (NRPS), mostrando o potencial metabólico da comunidade microbiana deste ambiente.

Metabolic and microbial diversity explored by metagenomic analysis of upwelling affected Brazilian coastal seawater reveals sequence domains of type I PKS and NRPS

Rafael R. C. Cuadrat[1], Juliano C. Cury[1,2], Alberto M. R. Davila[1]

1- Computational and Systems Biology Laboratory, Computational and Systems Biology Pole, Oswaldo Cruz Institute, Fiocruz, Avenida Brasil, 4365. Rio de Janeiro, Brazil. CEP 21040-360. Phone: +55-21-3865-8132. E-mail: davila@fiocruz.br

2 – Molecular Microbiology Laboratory, Federal University of São João del-Rei, Sete Lagoas Campus. Rod. MG 424, Km 47, Sete Lagoas – MG, Brazil. CEP 35701970. CP 56. Phone: +55-31-92144198. Email: jccury@hotmail.com

ABSTRACT

Marine environments harbor a wide range of microorganisms from the three domains of life. These microorganisms belie great potential, as their exploration should enable discovery of new enzymes and bioactive compounds for industrial use. Unfortunately, only ~1% of microorganisms from the environment can currently be identified following culture, limiting the discovery of new compounds. To overcome this limitation, a metagenomics approach has been widely adopted for the biodiversity studies on samples from marine environments. In this study, pyrosequencing of marine metagenomes afforded examination of the potential for new natural compound synthesis mediated by polymorphism in the Polyketide

Synthase (PKS) and Nonribosomal Peptide Synthetase (NRPS) genes. The samples were isolated from Praia dos Anjos (Angel's Beach) water, in Arraial do Cabo, Rio de Janeiro, Brazil, an environment particularly affected by upwelling. The water sample was fractionated by filtration through two membranes enriching for the prokaryotic (sample P) and eukaryotic (sample E) communities. A total of 651,083 and 542,647 reads were obtained for samples P and E, respectively. The most abundant bacterial phylum present in both samples was Proteobacteria, followed by Bacteroidetes and Cyanobacteria. Members of the Roseobacter clade (*Roseobacter* and *Ruegeria* genus) were abundant in both samples, but with larger abundance on sample P (up to 29% of identified genus).

The high abundance of *Roseobacter* clade and *Synechococcus* genus plus the nutrients abundance in the sample enforce the hypothesis that the environment was affected by upwelling with subsequently phytoplankton bloom.

Using Hidden Markov Models (HMM) facilitated screens of KS (keto-synthase) and C (condensation) domains from PKS and NRPS, respectively. A total of 84 KS and 46 C domain new sequences from both samples were obtained, showing the biotechnological potential of this environment. This was the first study conducted to screen PKS and NRPS genes in an upwelling affected sample.

INTRODUCTION

Marine environments cover ~70% of the Earth's surface. These habitats show great variation in temperature, pressure and salinity. They harbor a wide range of microorganisms from the three domains of life (Archaea, Bacteria and Eukarya) which are responsible for ~98% of marine primary production [1][2]. This huge biodiversity has great potential, as its exploration affords discovery of new enzymes for industrial use. However, only from 0.001% to 1% of environmental microorganism can be identified using culture-dependent approaches [3][4][5]. To overcome this limitation, metagenomic studies have been conducted using samples from a variety of aquatic environments from costal seawater, deep seawater and open ocean waters, to freshwater from rivers and lagoons [6][7][8].

The pioneering metagenomic study of marine planktonic microbiota, the Global Ocean Sampling Expedition, generated 7.7 million sequencing reads (6.3 billion base pairs) from water samples collected across a several-thousand kilometer transect from the North Atlantic through the Panama Canal and ending in the South Pacific [9]. After the advent of next generation technologies for DNA sequencing (NGS) such as ROCHE 454 and Illumina, many other studies have been performed around the world, from Artic to Antarctic [10][11]. However, the microbial diversity in the marine waters of Brazilian coast remains poorly characterized. The Brazilian coast extends for 7,491 km, and is influenced by the warm North Brazilian Current in the northern portion, the cold Malvinas/Falklands Current in the southern portion and to a lesser extent, by river mouths and upwelling regions [12].

Only a few studies have explored the marine waters of the Brazilian Coast using metagenomic approaches. Among them, Gregoracci *et al.* [13] studied the bacterioplankton of Guanabara Bay (the second largest bay of Brazil), in Rio de Janeiro state; Trindade-Silva and colleagues [14] characterized the microbial diversity associated to the marine sponge *Arenosclera brasiliensis* from water of João Fernandinho beach (Rio de Janeiro State) [14] and Cury and colleagues studied the taxonomic diversity of coastal seawater from Arraial do Cabo (Rio de Janeiro State, Cabo Frio region) [15], an important fishing and tourism region influenced by an upwelling system and anthropogenic activity, with sporadic sewage emissions [16] [17]. Upwelling is characterized by the up-flow of cold and nutrient-rich waters and it disturbs ecosystem dynamics and increases biomass and primary production of these environments. From the Brazilian coastal waters, it is at Arraial do Cabo that this upwelling effect is most intense [18]. Thus, Arraial do Cabo is the preferred site for detailed study of the effects of upwelling on the composition of the marine life and for, consideration of any detrimental impact to these upwelling effects from sewage emission which may affect local fishing.

As previous, SSU rDNA amplification approach used in Arraial do Cabo [15] was limited to the general taxonomic assembly of the microorganisms, rather than providing information about the metabolic potential of the community [19]. By using a whole metagenome pyrosequencing approachit should be possible not only to estimate the biodiversity, but also to explore functional gene diversity and select genes of biotechnological interest [20].

The Polyketide Synthases (PKSs) and Nonribosomal Peptide Synthetases (NRPS) encode two families of secondary metabolite enzymes from microorganisms which are of great interest to the biotechnological industry. They are responsible for the production of a variety of compounds from antibiotics to pigments and from, antitumor agents to immunosuppressant [21][22]. There are 3 types of PKS. The type I PKS genes encode large multi-domain enzymes with all the necessary components for elongating and processing the polyketide chain of the same protein. They can be classified as modular (the most from bacteria) or iterative (from fungi and bacteria) [23][24]. The type II PKS genes encode multi enzyme complexes (with three or more enzymes) acting in an iterative manner [25]. The type III PKS genes encode enzymes responsible for the production of Chalcones in plants and polyhydroxy phenol in bacteria [26]. In comparison, the NRPS genes encode modular enzymes that can activate and condense amino acids to produce small peptides (nonribosomal peptides) [27].

Most of new genes of these families have been discovered in genomes from the phylum Actinobacteria, that accounts for 5% to 10% of total bacteria on marine water [28][29][30]. However, the natural products from the most abundant phylum from marine oligotrophic environments (Proteobacteria) are poorly studied [31]. Most of the previous studies conducted to screen for PKS and NRPSs diversity were performed in soil or host associated (marine invertebrates) microbiomes [32][33]. However, Grossart *et al.* in 2004 showed many members of Rhodobacteraceae family, isolated from organic aggregates in German Wadden Sea, producing secondary metabolites (with bacterial inhibitory activity encoded by PKS and NRPS genes) [34][35]. Thus, many antimicrobial peptides were found in marine proteobacteria [36][37][38]. In this study, we explore the taxonomic diversity and the metabolic potential of microbes from the seawater of the Praia dos Anjos (Angel's Beach), in Arraial do Cabo, Brazil by pyrosequencing its metagenome. The sample site is close to the one sampled by Cury and colleagues, called POS sample [15] and it was collected at the summer season. We choose this site due its high abundance of Proteobacteria from Rhodobacteraceae family (with great metabolic potential), and the presence of Actinobacteria which were in greater abundance than the other sites analyzed in that previous study [15]. We choose the summer season because the upwelling phenomena in this region occurs with higher frequency and intensity during this season [15][18]. At the time of writing, this is the first study aiming to find novel PKS and NRPS genes in upwelling affected environment.

MATERIALS AND METHODS

Sampling

A total of 300 L of superficial (< 2 m) water was collected from the Praia dos Anjos (Angel's Beach), Arraial do Cabo, Rio de Janeiro, Brazil (-22°58'31.33", -42°0'46.84"). No specific permits were required for the described field studies.

The water pH, temperature and salinity were measured *in situ* and 1 L was used for determination of Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total N, Nitrate and Ammonium.

The COD test was performed by the closed reflux method followed by photometric determination, using a COD reactor (Hach Company, Loveland, CO, USA) and visible spectrophotometer (model DR-2500; Hach Company). BOD5, nitrate and ammonium were determined using the potentiometric method with selective electrodes Orion 081010MD, Orion 9707BNWP and Orion 9512HPBNWP, respectively (Hach Company). The methodologies used to assess the physicochemical parameters were consistent with the methods described in the Standard Methods for Examination of Water and Wastewater (APHA 1998).

The 300 L sample was filtered first through 0.8 µm membranes (aiming to hold eukaryotes and particle-associated prokaryotes – named Sample E) and then through 0.22 µm membranes (aiming to hold free living prokaryotes – named sample P) using a vacuum filtration system.

DNA extraction and quantification

The DNA was extracted from membranes using the Meta-G-Nome™ DNA Isolation Kit (EPICENTRE). In order to obtain 20 µg of DNA, a total of 20 membranes of each sample were used in extractions. This large amount of DNA is necessary to avoid bias. The extracted DNA samples were verified by agarose gel (1%) eletrophoresis (100V) and quantified using ImageJ software, NanoDrop (Thermo Scientific) and Qubit (Life Technologies).

DNA pyrosequencing and sequences pre-processing

A total of 2 µg DNA from each sample (P and E) was sent to LNCC (Laboratorio Nacional de Computação Científica, Petrópolis, Rio de Janeiro, Brazil) in order to pyrosequencing on a 454 (ROCHE) sequencer using the GS FLX+ System. One 454 plate was used, and DNA of each sample constituted half of the plate.

The SFF files generated were analyzed on Stingray (stingray.biowebdb.org) [39] to generate the clipped FASTA and QUAL files. CD-HIT-454 [40] program was used to remove artificial duplicates (artifacts) using default parameters and LUCY v1.20 [41] (default parameters) was used to remove low quality and small sequences (< 20 Phred score, < 100 base pairs [bp]).

Reads analysis

The metagenomic sequences (reads) were analyzed using programs BLAST 2.2.21 (BLASTN and BLASTX) [42], MEGAN 4.0 [43] and MG-RAST [44] (MG-RAST ID: 4539290.3 for sample P; 4539291.3 for sample E).

Analysis of the SSU rDNA sequences

The SSU rDNA sequences (16S for prokaryotes and 18S for eukaryotes) were extracted using the INFERNAL program [45] with covariance models (CM) generated by 16S sequences (Archaea and Bacteria) and 18S sequences (Eukarya). These sequences were submitted to BLASTN 2.2.21 (e-value cutoff e^{-5}) against the SILVA SSU Database release 108 (http://www.arb-silva.de/) [46]. The BLASTN results were loaded in MEGAN 4.0 [43] in order to perform the taxonomic characterization using the LCA algorithm (maximum number of matches per read: 5, min support: 5, min score: 35, top percent: 10).

Analysis of the whole metagenomic sequences

All the sequences (reads) were submitted to BLASTN 2.2.21 (e-value cutoff e^{-5}) against GenBank (NT– from NCBI) and the results were loaded in MEGAN v4 [43] in order to perform the taxonomic characterization using the LCA algorithm (maximum number of matches per read: 5, min support: 5, min score: 75, top percent: 10).

After the preprocessing step, all reads obtained were submitted to MG-RAST [44] for the analysis using the full pipeline and annotated using the M5NR database.

Metagenomic reads assembly

The metagenomic reads from both samples were assembled using CAP3 [47] with default parameters. The contigs and singlets were concatenated and the METAGENMARK [48] was used to extract the metagenomic Open Reading Frames (ORFs) and to translate those ORFs to protein sequences using the Transeq program from the EMBOSS package [49].

Screening for genes of NRPS C domain and type I PKS KS domains

A HMM profile (pHMM) approach was used to screen the genes of type I PKS (KS domains) and NRPS (C domains). The pipeline used in this work was adapted from a previous work developed by Dumaresq *et al.* [50]. The KS domains of type I PKS (modular and iterative) were obtained from MAPSIDB (http://gate.smallsoft.co.kr:8080/pks/mapsidb) [51]. The NRPS C domains were obtained from NRPSDB (http://linux1.nii.res.in/~zeeshan/webpages/home.html) [52]. For each domain, a multiple alignment was generated by MAFFT [53] and a HMM profile was built using the hmmbuild software from the HMMER v3.0 [54] package. Those profiles (PKS-pHMM and NRPS-pHMM) were then used to screen the translated ORFs of both samples from Arraial do Cabo using hmmsearch with e-value cutoff 0.1.

The translated ORFs that showed hits with PKS-pHMM and NRPS-pHMM were extracted from our metagenome dataset using FASTACMD (from BLAST 2.2.21 package [42]). A reverse search of the extracted translated ORFs against the PKS-pHMM and NRPS-pHMM was conducted using hmmscan (e-value cutoff 0.1), in order to classify the type of PKS from metagenomic sequences. In order to confirm and validate the results, all the extracted translated ORFs similar to profiles were submitted to BLASTP 2.2.21 (e-value cutoff e^{-5}) against RefSeq protein database release 61. The annotation of the best five hits of each environmental sequence was used to annotate them. Finally, the confirmed KS and C domain sequences were submitted to Natural Product Domain Seeker - NapDoS (http://npdomainseeker.ucsd.edu/) [55] in order to classify the domains and to carry out phylogenetic analysis.

RESULTS AND DISCUSSION

Sampling Characteristics

At the time of sampling, the temperature of the water was 26°C, the pH was 7.5 and the salinity was 33%. An analysis of the water sample showed BOD was 1 mg/L, COD was 60 mg/L, and that the concentration of nitrate, ammonium and total nitrogen were 0.9 mg/L (or ~15.51 µM), 0 mg/L (< 0,5 µM) and 0.4 mg/L (or ~28.57 µM), respectively. The concentration of Nitrate (15.51 µM) was similar to the measured by Cury *et al.*[15] in the previous study (13.00 µM). The very low concentration of Ammonium (measured as NH_3) was not expected, due to the proximity to the coast with high anthropogenic activity and because in the previous study conducted by Cury *et al.*[15], the concentration at the same site was ~ 0.9 µM. These results may suggest that at the time-point used, the samples collected were unaffected by sewage disposal since the study conducted by Coelho-Souza *et al.* [18] showed high concentrations (above 2 µM) of ammonium in the same site, when the sewage disposal was visible. However, this result may also indicates that the region was affected by the upwelling and subsequently phytoplankton bloom), depleting the free ammonium in the environment, once it was demonstrated the higher rate of ammonium uptake occur during upwelling season [56] with strong preference for ammonium rather nitrate uptake [57]. In addition, natural eutrophication by upwelling events is usually associated with nitrate inputs [58]. Moreover, Albuquerque *et a*l., 2014 [59] have showed the occurrence of upwelling in the same month and year (January of 2012) of our study through the analysis of regional winds vectors (ASCAT program) and temporal variability of the water column thermal structure.

DNA pyrosequencing and sequences pre-processing

A total of 651,083 and 542,647 reads were obtained for samples P and E, respectively. After pre-processing, this number decreased to 595,534 and 469,354, respectively. The average size of the reads was 588 and 595 bp for P and E samples, respectively.

SSU rDNA sequence analysis and taxonomic assignment

Using INFERNAL with all Covariance Models (CMs) (after removing redundancy) 1,501/595,534 putative rDNA reads for sample P (0.25% of the total reads) and 672/469,354 putative rDNA reads for sample E (0.14% of the total reads) were obtained. These sequences were submitted to BLASTN against SILVA SSU database and a total of 1,210/1,501 (80.61% of the rDNA sequences) and 502/672 (74.70% of the rDNA sequences) hits were obtained for samples P and E, respectively.

Using MEGAN, 91.65% of the rDNA sequences from sample P were classified as Bacteria, whereas 7.27% were classified as Eukarya and 0.82% as Archaea (Figure 1). For sample E, 48% of rDNA sequences were classified as Bacteria and 51.59% as Eukarya (Figure 1). Archaeal sequences were not detected in sample E.

Analyzing all pre-processed reads from both samples using BLASTN algorithm against Genbank database (NT) and MEGAN 4.0 (LCA algorithm), 35.63% and 16.53% of the reads from P and E samples, respectively, were classified. For sample P, 72.64% of those sequences were classified as Bacteria, whereas 0.03% as Archaea, 2.52% as Eukarya, and 0.17% as Viruses (Figure 1). For sample E, 63.02% of these sequences were classified as Bacteria, 0.07% as Archaea, 12.73% as Eukarya, and 0.81% as Viruses (Figure 1).

The low abundance of eukaryotes found in sample P can be explained by the fact that most eukaryotes are retained by the 0.8 µm pore membrane. The relative high abundance of prokaryotes found in sample E is probably due to: (i) the fact that a portion of prokaryotic cells may be directly retained by the 0.8 µm membranes, being larger than 0.8 µm; (ii) a portion of prokaryotic cells may be associated with organic matter fragments, and consequently retained by the 0.8 µm membranes, and (iii) a portion of prokaryotic cells may be retained by the membrane that already presents their pores blocked by the eukaryotic cells, larger bacterial cells and organic matter.

The high percentage of reads classified as Bacteria (~70%) and the low abundance of Eukarya group (<10%) on total-reads analysis for sample P agree with its SSU rDNA analysis. On the other hand, the results of the same analysis for sample E do not corroborate the results obtained using SSU rDNA analysis, as Bacteria is the larger domain, with >60% of the reads classified as belonging to that group, whereas only a small fraction of reads were classified as Eukarya when total reads are analyzed (Figure 1). A possible explanation is that GenBank (NT) database contains much more sequences of Bacteria than Eukarya (on current release, ~8.5 billion bp originated from Bacteria, whereas ~2.5 billion bp originated from Eukarya) [60]. This can explain the fact that so many sequences of eukaryotes may not have presented hit in BLASTN results. This also explains the fact that sample E exhibit larger percentage of sequences with no hit when compared with sample P (83.35% on sample E and 64.17% on sample P).

Diversity of Bacteria

The most abundant bacterial phylum present in both samples is Proteobacteria (~90% for P sample and ~45% for E sample) (Figure 2), as expected for a marine environment [33].

The Alphaproteobacteria is the most abundant class on sample E considering the both used approaches (Figure S1). This was also the case for sample P, but only using total reads analysis. For the analysis using SSU rDNA approach, Gammaproteobacteria was the most abundant class (Figure S1). Beta-, Delta- and Epsilonproteobacteria classes could be observed at low abundance using the BLASTN approach, but absent using SSU rDNA approach (Figure S1). The previously study conducted on the same environment by Cury *et al*. [15] showed similar high abundance of Alphaproteobacteria for the same point of sample collection (Praia dos Anjos harbor). On the other hand, the abundance of Gammaproteobacteria was more than 40% for sample P using both approaches and sample E using SSU rDNA approach, whereas Cury *et al*. found less than 1% of Gammaproteobacteria [15]. However, it is important to note that the filtration and method of analysis of Cury's study was different filtering directly on a single 0.22 µm membrane and then using a PCR amplification approach of SSU rDNA [15].

The second most abundant bacterial phylum in both samples was Bacteroidetes (Figure 2). Members of this group are present in many different ecological niches, including soil, ocean, freshwater, and the gastrointestinal tract of animals, playing many biological functions, including degradation of organic matter [61].

The most abundant class of Bacteroidetes in both samples was Flavobacteria, with more than 90% of the sequences classified in this group (Figure S2); agreeing with the study of Cury *et al*. [15] that found Flavobacteria as the major class of Bacteroidetes in the same region.

In sample E we observed a high abundance of sequences classified as Cyanobacteria (17.35% of all sequences of bacteria) (Figure 2). The higher abundance of Cyanobacteria on sample E when compared with the sample P may be explained by the size of the cyanobacterial cells that can be up to 3 µm for some species [62][63]. Unlike most of the oligotrophic marine environments, where normally the *Prochlorococcus* is the dominant genus of Cyanobacteria [64], the most abundant genera found in sample E were *Synechococcus* (49.82%) and *Synechocystis* (35.95%), both belonging to the order Chroococcales.

Many species from genus *Synechococcus* can assimilate nitrogen compounds (ammonium and nitrate), and reside in relatively nutrient-rich waters at coastal sites throughout the world [65]. Preston *et al.* (2011) [66] showed increases for *Synechococcus* in coastal waters at the relaxation of upwelling.

It was possible to detect 1,114 reads (13.12% of total Cyanobacteria reads) classified as *Synechococcus* CC9311. This strain was isolated from the edge of California Current [67] and many related strains have been isolated from costal environments, displaying a coastal type chlorophyll profile [68]. A comparative genomic study was conducted by Palenik *et al*. [68], showing many adaptations in the genome of the CC9311 strain, as metallo enzymes and light apparatus adaptations absent in the strains from open ocean, strongly suggesting a high level of adaptation of the CC9311 strain to costal environments.

The order Rhodobacterales (Alphaproteobacteria) was the most abundant on sample P, with 42.23% of the reads, followed by the Alteromonadales order (Gammaproteobacteria), with 17.47% (Figure S3). On the other hand, the most abundant order of sample E was Flavobacteriales (Bacteroidetes), with 25.45% of the reads, followed by Rhodobacterales (Alphaproteobacteria), with 25.03%). These results also corroborate with the previous work of Cury *et al*. [15], where the most abundant order was Rhodobacterales (more than 90% of the Alphaproteobacteria) in samples from the same place.

Unlike the SSU rDNA-based study performed by Cury and colleagues [15], we used shotgun approach in the present work. This allowed us to study more deeply the taxonomic groups present in this environment, like family and genus. The most abundant family in sample P was Rhodobacteraceae (43.58%), followed by Alteromonadaceae (8.89%). For the sample E, the most abundant families were Flavobacteriaceae (25.20%) and Rhodobacteraceae (25.01%) (Figure S4). The Rhodobacteraceae is a family of Alphaproteobacteria that contains chemoorganotrophs and photoheterotrophs members [69]. Alteromonadaceae is a family of Gammaproteobacteria class [70][71] and most of the members are found in marine environments [72].

The family *Flavobacteriaceae* is the larger group of the phylum Bacteroidetes. Many species of this family inhabit marine environments and play important roles on the mineralization of organic matter in these ecosystems [73][74][75]. This may explain why sample E has the highest abundance of this family, since it was filtered by a 0.8 µm membrane, where the most of organic matter particles were attached.

The most abundant genera detected in sample P were *Ruegeria* and *Roseobacter*, whereas for sample E, the most abundant genera were *Synechococcus*, *Lacinutrix* and *Ruegeria* (Fugures S5 and S6). The bacteria from the *Roseobacter* genus are aerobic anoxygenic phototrophic bacteria (AAP) [76] that belongs to Rhodobacteraceae family of the Alphaproteobacteria phylum. They represent one of the most abundant groups of Bacteria in oceans, typically comprising upwards of 20% of coastal and 15% of mixed-layer ocean bacterioplankton communities [77]. Members of *Roseobacter* genus plays important roles on sulfur cycle, and some isolates from this group were the first marine strains found that simultaneously possess two key pathways for the degradation of the osmolyte dimethylsulfoniopropionate (DMSP), an organosulfur compound (secondary metabolite) found in marine phytoplankton, seaweeds, and some species of terrestrial and aquatic vascular plants [78][79]. *Ruegeria* is a genus of the same family of *Roseobacter*, but unlike *Roseobacter*, they cannot do anoxygenic photosynthesis (do not produce bacteriochlorophyll) [80]. The *Lacinutrix* is a genus of the Flavobacteriaceae family that is commonly isolated from algae and calanoid copepods [81], and this may explain the fact that this genus was detected only in sample E. It is likely that the detected DNA of *Lacinutrix* remained attached to the 0.8 µm membrane along with algae and copepods.

The genus *Pelagibacter* also was found with high abundance (6% on sample P and 5% on sample E). This genus is a member of the SAR11 clade, a ubiquitous group in the world's oceans [82], and dominant on surface bacterioplankton [83]. However, unlike in this study, in many oligotrophic waters this genus is normally the most abundant Bacteria, representing more than 10% of the total bacteria from these environments [14][84].

Members of Roseobacter clade were also detected (using FISH and SSU rDNA approaches) as the most abundant group (overcoming SAR11 clade) in a upwelling affected coastal environment during all seasons (“Ría de Vigo” NW, Spain), but its more abundant in the bloom season [85]. The same results were also found (using DAPI approach) in the estuarine environment of “Ría de Aveiro” (Portugal), also impacted by upwelling system [86].

In oligotrophic coastal waters, like Arraial do Cabo, high abundance of *Roseobacter* have been observed to thrive only during chlorophyll a-rich periods (upwelling seasons) [87].

The possible explanation to the high abundance of *Roseobacter* and *Reugeria* in upwelling affected environment may be the bloom of picophytoplankton with high production of DMSP, since many of these species can degrade this compound and its degradation product, dimethylsulphide (DMS) [88][89].

The nutrients concentration plus the high abundance of many species of *Roseobacter* clade and *Synechococcus* genus suggest that the upwelling phenomenon and subsequently phytoplankton bloom were affecting the collected sample, as expected in the summer season at this place [18].

Diversity of Eukarya

Considering the results of SSU rDNA analysis, the most abundant group of the Eukarya domain is Metazoa for both samples (Figure 2). Considering the total read analysis, Viriplantae and Stramenopiles were the most abundant groups for samples P and E, respectively (Figure 2). The most abundant group of Viriplantae, in both samples, is the green algae from Chlorophyta group (more than 85% from Viriplantae reads), mainly from the pikophytoplankton genera *Micromonas* and *Ostreococcus*.

From the Stramenopiles, the most abundant phylum is Bacillariophyta (diatoms). They are the most important members of pikophytoplankton, among the most diversified groups of photosynthetic eukaryotes, with possibly over 100,000 extant species, contributing with around 40% of the marine primary productivity [90][91][92].

Together, the photosynthetic members of pikophytoplankton play major role in primary production in oligotrophic environments (up to 80% of the autotrophic biomass) [93] [94].

Figure S7 shows the reads classification on phylum level of each of three main groups of Eukarya domain (Metazoa, Viridiplantae and Stramenopiles). The results of SSU rDNA analysis corroborate the previous study performed by Cury *et al*. [15]. However, the results from total reads analysis were dissimilar, suggesting that the results from SSU analysis may be biased.

Diversity of Archaea

The abundance of Archaea sequences was very low (Figure 1). Using rDNA approach, we detected only 10 sequences, that were belonging to sample P. Using the total reads analysis, it was possible to detect 62 archaeal sequences in sample P, which 30 were classified as Euryarchaeota, 29 as Thaumarchaeota and 3 are unclassified Archaea; and 60 sequences in sample E, which 28 were classified as Euryarchaeota, 24 as Thaumarchaeota and 8 are unclassified Archaea (Figure 2). This very low abundance of Archaea was not expected, because in many other tropical costal metagenomes, the abundance of Archaea was from 1.1% in João Fernandinho beach, in Rio de Janeiro State (geographically close to Arraial do Cabo) [14] to 2.9% on coast of Galapagos Island [9]. In our study, considering the BLASTN analysis, we found 0.04% of archaeal sequences, considering all classified sequences. The total absence of Crenarchaeota group may be explained by the very low concentration of ammonium, once this group use ammonia as its sole source of energy [95].

Analysis of metagenomic sequences using MG-RAST pipeline

In order to obtain a functional annotation of metagenomic DNA from Arraial do Cabo waters, the reads obtained from samples P and E were submitted to MG-RAST after pre-processing on Stingray. The Table 1 summarizes the statistics of the sequences from both samples.

The alpha diversity from samples was 240.06 and 464.02 species for samples P and E, respectively. The higher diversity value for sample E may reflect the size selection exclusion arising from filtration. Figure S8 shows the rarefaction curve from both samples.The reads of the two samples were compared to M5NR using a maximum e-value of 1e^{-5}, a minimum identity of 60%, and a minimum alignment length of 15 measured in amino acid for protein and base pairs for RNA databases. Figure S9 shows the number of hits obtained for each database screened.

Metagenomic reads assembly

With the aim of obtaining contigs from the metagenomic reads, we used the CAP3 program [47]. The assembly of environmental sequences is a complex problem and so many algorithms were proposed to address it [96][97][98][99]. The main problems of the assemblage are the low coverage and the possible formation of chimeras (especially on environments with high diversity) [100] [101]. Using the CAP3 with the very stringent default parameters we tried to minimize the problem of chimeras, but the low coverage only can be outlined with more deep sequencing effort. We obtained a total of 29,074 contigs and 269,587 singlets for sample P. For sample E, we obtained 20,792 contigs and 396,371 singlets. From these sequences (singlet + contigs), were also obtained 409,111 and 451,722 ORFs for sample P and E, respectively, using METAGENMARK.

Screening genes of NRPS and PKS

Many studies have been conducted for functional or PCR-based screening of PKS and NRPS in diverse environments [32][33][102][103]. However, despite the growth of metagenome databases, to the date, only few studies were performed using computational approaches to screen PKS in whole shotgun sequenced metagenomes. One of these was conducted by Foerstner *et al* [104] where six natural environment datasets were screened using a HMM profile approach. Moreover, no study has been performed to screen secondary metabolites genes in water from upwelling affected coastal environment.

Because of the high abundance of *Roseobacter* clade organisms and Cyanobacteria in the samples from Arraial do Cabo (see *Diversity of Bacteria*), we decided to conduct an *in silico* screening of PKS and NRPS genes in order to evaluate the potential of the environment to provide new secondary metabolites, since many studies have showed the presence of these genes in the genomes of these very abundant taxonomical groups [34][36][37][38].

PKS KS domain

In this study, two KS pHMM were built: from sequences of modular KS and iterative KS. These pHHM were used to retrieve putative KS domains in metagenome from Arraial de Cabo and the NapDos was used because this system has the capacity to classify KS and C sequences from poor assembled genomes and metagenomes [55]. Using the KS modular pHMM, we found a total of 28 hits in sample P and 37 in sample E. These sequences were submitted to BLASTP against RefSeq protein database (e-value cutoff e^{-5}). Only 7 sequences returned no hits, one from sample P and 6 from sample E. Using the annotation of the five best hits from RefSeq database it was possible to confirm the PKS function of 46.42% and 72.97% of the KS sequences of P and E samples, respectively.

All KS sequences were submitted to NapDoS [55]. For sample P, it was possible to classify 78.57% of the sequences, whereas for sample E, it was possible to classify 91.89% of the sequences.

Using the KS iterative, we obtained 21 and 16 sequences from sample P and sample E, respectively. These sequences were submitted to BLASTP against RefSeq Protein (e-value cutoff e^{-5}). For sample P, all the sequences showed hits on blast, and for sample E, only 2 sequences don't show hits. Using the annotation of the five best hits from RefSeq database was possible to confirm the PKS function of 28.57%

(6/21) and 56.25% (9/16) of the sequences from samples P and E, respectively. These sequences were also submitted to NapDoS. For sample P, it was possible to classify 71.42% (15/21) of the sequences, whereas for sample E, it was possible to classify 75% of the sequences. From the total KS domains obtained by both pHMM (102), 38 sequences from sample P and 46 sequences from sample E (totaling 82.35%) were confirmed in silico as KS (by blastP and/or NapDos results), respectively. The false positives found (17.65%) were expected, because the HMM approach is very sensitive to detecting distant homologues [54] and the Fatty Acid Synthase (FAS) is homologous to PKS [105]. The advantage of the use of pHMM to screen type I PKS in metagenomic shotgun data and the possible recovery of false positives was discussed on a study performed by Foerstner *et al* [104].

From the 84 type I KS, 4 were similar to Rhodobacteraceae organisms in the blastP results (against RefSeq protein). From these 4 KS sequences, two were classified in the phylogenetic tree as hybrid KS/PKS enzymes, agreeing with results obtained by Martens *et al* [35], showing many hybrids enzymes in isolates from *Roseobacter* clade.

The relative abundance of KS domain present in water was 0.0092% (38/409,111) from sample P and 0.0101% (46/451,722) from sample E. In the study of Foerstner *et al.* [104], the environment with the higher KS domain abundance was Minnesota farm soil [106], where 52 type I KS were found in 183,536 ORFs (0.0283%), 2.8 times more abundant than sample E of the present study. Soil environments commonly possess a high diversity of secondary metabolites because the microorganisms compete intensely with each other [107]. In addition, in the same study, the samples from an open ocean oligotrophic region (Sargasso Sea) [108] (ranging from 0.1-0.8 µm like sample P), were screened, showing 69 type I KS sequences in 1,214,207 ORFs (0.0056%), a relative smaller abundance than our sample P. These results confirm the potential of the costal upwelling affected metagenome for the screening of secondary metabolites.

Figure S10 shows the classification of KS domains obtained from both samples using both pHMMs.

Most of the sequences retrieved with modular KS pHMM were classified as modular by NapDoS (46% from both samples) and Hybrid KS (18% from sample E and 24% from sample P). These hybrid domains can be modular or iterative and are present on hybrid PKS/NRPS enzymes [109]. The Trans-AT domain was also present (11% on sample E and 14% on sample P) and, unexpectedly, using this pHMM it proved possible to retrieve few iterative KS domains. On the other hand, the iterative KS pHMM was able to retrieve only a few iterative sequences (only the iterative Enediyne sequences were retrieved). Most of the sequences retrieved with this pHMM on sample E were Trans-AT domains (31%). However, on sample P, the most abundant type of KS domains retrieved with the iterative KS pHMM was modular. All the sequences classified as Polyunsaturated fatty acids **(**PUFA) by NapDos were manually verified and the best blast hit (against RefSeq) show similarity with PKS domain (with more significant e-value than the NapDos result). The high abundance of KS modular can be explained by the fact that in modular PKS the number of copy of each domain is so much higher than in iterative PKS [105].

Additionally, aiming a most accurate classification of the KS sequences, a phylogenetic analysis was performed using the 10 KS sequences from sample P and 15 from sample E (larger than 200 amino acids) to generate a phylogenetic tree on NapDos with the reference sequences (similar to the environmental sequences) (Figure 3).

The topology of the tree corroborates the results from many phylogenetic studies of KS domain [56][100]. It was possible to separate the homologous Fab (from Fatty acid synthase), the type II KS alpha and beta domains, the Polyunsaturated fatty acids **(**PUFA) domains (with one sequence from sample E in this clade), two clades from the Modular Trans-AT domains (where 7 sequences from sample P and two from sample E were classified), the Iterative KS (with two sequences from sample E), the Hybrid KS (PKS/NRPS hybrid enzymes), with one sequence from sample P and 6 from sample E, the KS1 domains (Typical starter KS KSQ) domains and KS domains in Curacin and Salinisporamide biosynthesis pathways, with one sequence from sample E, and finally the Modular KS domains (where one sequence from sample P and 3 from sample E were classified). Like the results showed in Figure S10, the phylogenetic tree shows the most of sequences close to modular (Cis, Trans and Hybrid) KS sequences.

NRPS C domain

Using the C domain pHMM, a total of 50 hits were obtained, 14 from sample P and 36 from sample E. These sequences were submitted to BLASTP against RefSeq protein (e-value cutoff e^{-5}). Sample P yielded 11 hits (78.57%) with confirmed annotations. Sample E yielded 32 hits (88.88%) and was possible to confirm the annotation from 31 of these. All 43 annotated sequences were submitted to NapDos and classified. It was not possible to confirm the annotation of one sequence (7.14%) from sample P and 3 sequences (8.33%) from sample E by these methods.

The environmental C domain sequences larger than 200 amino acids were also submitted to phylogenetic analysis using the NapDoS pipeline, with the reference sequences (similar to the environmental sequences). Figure 4 shows the result of this analysis.

The topology of the inferred tree was expected, separating the types of C domains as showed in previous studies [56]. Most of the sequences were from LCL type (4 from sample E and 1 from sample P) and Epimerization (4 from sample E).

As in PKS screening, sample E showed higher abundance of NRPS C domain than sample P. This may be due to the presence of marine snow in sample E, with competition for space and nutrients in the particle associated bacteria as a selective force [110].

Conclusions

In this work, two fractions of the seawater metagenome from Praia dos Anjos (Angel´s Beach) a coastal environment, affected by upwelling, were pyrosequenced.

The results show that the sample was likely to have been collected after upwelling and subsequent phytoplankton bloom and demonstrate a high abundance of Proteobacteria in both fractions (89.1% for sample P and 48.5% for Sample E), with genera *Ruegeria*, *Roseobacter*, *Synechococcus* and *Lacinutrix* the most abundant. Screening of the metagenome revealed 84 KS domain ORFs and 46 C domain ORFs (from PKS and NRPS). These sequences were manually verified and classified by NapDoS system and BLAST, showing a close enough similarity to curated sequences to confirm the biological function. However, the degree of divergence suggests that they are probably new alleles. Based on these results we will now prepare an environmental DNA (fosmid) library from which to clone full sequences of PKS and NRPS, in order to conduct functional and sequence screening, using the sequences generated in this study as probes. In addition, a time-series study may be conducted in the future, to better understand the main differences between the microbial communities from each season in this region.

Acknowledgments

We thank to the sequencing group of LNCC for performing the DNA pyrosequencing. To Dr. Adriana M. Froes, Dr. Aline Dumaresq, Dr. Gisele Lopes Nunes, Fábio Bernardo da Silva and Bruno Manoel da Silva for their help in collecting and pre-processing the samples. To Dr. Rodrigo Jardim for all the bioinformatics support. To Dr. Hans-Peter Grossart for valuable suggestions and to Dr Kevin M. Tyler for the help with the English of the MS. To Dr. Yara Traub-Cseko for her willingness and kindess allowing us to use the Molecular Biology of Parasites and Vectors Lab´s facilities. To CNPq, FIOCRUZ and CAPES for financial support.

Author Contributions

Conceived and designed the experiments: RRCC, JCC and AMRD. Performed the experiments: RRCC. Analyzed the data: RRCC and JCC. Contributed reagents/materials/analysis tools: RRC, JCC and AMRD. Wrote the paper: RRCC, JCC and AMRD.

References

1 - Kennedy J, Flemer B, Jackson SA, Lejon DPH, Morrissey JP, et al. (2010) Marine Metagenomics: New Tools for the Study and Exploitation of Marine Microbial Metabolism. Marine Drugs 8: 608–628. doi:10.3390/md8030608.

2 - Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere." Proceedings of the National Academy of Sciences 103: 12115–12120.

3 - Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. Nature Reviews Genetics 6: 805–814. doi:10.1038/nrg1709.

4 - Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276: 734–740.

5 - Kennedy J, Flemer B, Jackson SA, Lejon DPH, Morrissey JP, et al. (2010) Marine Metagenomics: New Tools for the Study and Exploitation of Marine Microbial Metabolism. Marine Drugs 8: 608–628. doi:10.3390/md8030608.

6- Ghai R, Rodŕíguez-Valera F, McMahon KD, Toyama D, Rinke R, et al. (2011) Metagenomics of the Water Column in the Pristine Upper Course of the Amazon River. PLoS ONE 6: e23785. doi:10.1371/journal.pone.0023785.

7- Ghai R, Hernandez CM, Picazo A, Mizuno CM, Ininbergs K, et al. (2012) Metagenomes of Mediterranean Coastal Lagoons. Scientific Reports 2.

8- Konstantinidis KT, Braff J, Karl DM, DeLong EF (2009) Comparative Metagenomic Analysis of a Microbial Community Residing at a Depth of 4,000 Meters at Station ALOHA in the North Pacific Subtropical Gyre. Applied and Environmental Microbiology 75: 5345–5355. doi:10.1128/AEM.00473-09.

9- Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biology 5: e77. doi:10.1371/journal.pbio.0050077.

10 - Hamdan LJ, Coffin RB, Sikaroodi M, Greinert J, Treude T, et al. (2013) Ocean currents shape the microbiome of Arctic marine sediments. ISME J 7: 685–696. doi:10.1038/ismej.2012.143.

11 - Yau S, Lauro FM, Williams TJ, Demaere MZ, Brown MV, et al. (2013) Metagenomic insights into strategies of carbon conservation and unusual sulfur biogeochemistry in a hypersaline Antarctic lake. ISME J 7: 1944–1961. doi:10.1038/ismej.2013.69.

12 - Prates AP, Henrique De Lima L, Chatwin A (2007) Coastal and marine conservation priorities in Brazil. In: Chatwin A, editor. Priorities for coastal and marine conservation in South America. Arlington, Virginia. USA: The Nature Conservancy. pp. 15–23.

13 - Gregoracci GB, Nascimento JR, Cabral AS, Paranhos R, Valentin JL, et al. (2012) Structuring of Bacterioplankton Diversity in a Large Tropical Bay. PLoS ONE 7: e31408. doi:10.1371/journal.pone.0031408.

14 - Trindade-Silva AE, Rua C, Silva GGZ, Dutilh BE, Moreira APB, et al. (2012) Taxonomic and Functional Microbial Signatures of the Endemic Marine Sponge Arenosclera brasiliensis. PLoS ONE 7: e39905. doi:10.1371/journal.pone.0039905.

15- Cury JC, Araujo FV, Coelho-Souza SA, Peixoto RS, Oliveira JAL, et al. (2011) Microbial Diversity of a Brazilian Coastal Region Influenced by an Upwelling System and Anthropogenic Activity. PLoS ONE 6: e16553. doi:10.1371/journal.pone.0016553.

16- Ferreira CEL, Gonçalves JEA, Coutinho R (2006) Ship hulls and oil platforms as potential vectors to marine species introduction. Journal of Coastal Research 39: 1341–1346.

17- López MS, Coutinho R (2010) Positive interaction between the native macroalgae Sargassum sp. and the exotic bivalve Isognomon bicolor? Brazilian Journal of Oceanography 58: 69–72.

18- Coelho-Souza SA, Pereira GC, Coutinho R, Guimarães JR (2013) Yearly variation of bacterial production in the Arraial do Cabo protection area (Cabo Frio upwelling region): an evidence of anthropogenic pressure. Brazilian Journal of Microbiology 44: 1349–1357.

19 – Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M (2011) Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC genomics 12: S4.

20 – Jiang C-J, Hao Z-Y, Zeng R, Shen P-H, Li J-F, et al. (2011) Characterization of a Novel Serine Protease Inhibitor Gene from a Marine Metagenome. Marine Drugs 9: 1487–1501. doi:10.3390/md9091487.

21- Gokhale RS, Sankaranarayanan R, Mohanty D (2007) Versatility of polyketide synthases in generating metabolic diversity. Curr Opin Struct Biol 17: 736–743. doi:10.1016/j.sbi.2007.08.021.

22 - Koglin A, Walsh CT (2009) Structural insights into nonribosomal peptide enzymatic assembly lines. Natural Product Reports 26: 987. doi:10.1039/b904543k.

23 - Lal R, Kumari R, Kaur H, Khanna R, Dhingra N, et al. (2000) Regulation and manipulation of the gene clusters encoding type-I PKSs. Trends Biotechnol 18: 264–274.

24 - Cane DE (1998) Harnessing the Biosynthetic Code: Combinations, Permutations, and Mutations. Science 282: 63–68. doi:10.1126/science.282.5386.63.

25 – Sun W, Peng C, Zhao Y, Li Z (2012) Functional Gene-Guided Discovery of Type II Polyketides from Culturable Actinomycetes Associated with Soft Coral Scleronephthya sp. PLoS ONE 7: e42847. doi:10.1371/journal.pone.0042847.

26 - Castoe TA, Stephens T, Noonan BP, Calestani C (2007) A novel group of type I polyketide synthases (PKS) in animals and the complex phylogenomics of PKSs. Gene 392: 47–58. doi:10.1016/j.gene.2006.11.005.

27 - Silva-Stenico ME, Silva CSP, Lorenzi AS, Shishido TK, Etchegaray A, et al. (2011) Non-ribosomal peptides produced by Brazilian cyanobacterial isolates with antimicrobial activity. Microbiol Res 166: 161–175. doi:10.1016/j.micres.2010.04.002.

28 - King GM, Smith CB, Tolar B, Hollibaugh JT (2012) Analysis of composition and structure of coastal to mesopelagic bacterioplankton communities in the northern gulf of Mexico. Front Microbiol 3. doi:10.3389/fmicb.2012.00438.

29- Jamieson RE, Rogers AD, Billett DSM, Smale DA, Pearce DA (2012) Patterns of marine bacterioplankton biodiversity in the surface waters of the Scotia Arc, Southern Ocean. FEMS Microbiol Ecol 80: 452–468. doi:10.1111/j.1574-6941.2012.01313.x.

30 - Lau SCK, Zhang R, Brodie EL, Piceno YM, Andersen G, et al. (2013) Biogeography of bacterioplankton in the tropical seawaters of Singapore. FEMS Microbiol Ecol 84: 259–269. doi:10.1111/1574-6941.12057.

31 - Desriac F, Jégou C, Balnois E, Brillet B, Chevalier P, et al. (2013) Antimicrobial Peptides from Marine Proteobacteria. Marine Drugs 11: 3632–3660. doi:10.3390/md11103632.

32 - Graça AP, Bondoso J, Gaspar H, Xavier JR, Monteiro MC, et al. (2013) Antimicrobial Activity of Heterotrophic Bacterial Communities from the Marine Sponge Erylus discophorus (Astrophorida, Geodiidae). PLoS ONE 8: e78992. doi:10.1371/journal.pone.0078992.

33 - Schneemann I, Nagel K, Kajahn I, Labes A, Wiese J, et al. (2010) Comprehensive Investigation of Marine Actinobacteria Associated with the Sponge Halichondria panicea. Applied and Environmental Microbiology 76: 3702–3714. doi:10.1128/AEM.00780-10.

34 - Grossart H-P, Schlingloff A, Bernhard M, Simon M, Brinkhoff T (2004) Antagonistic activity of bacteria isolated from organic aggregates of the German Wadden Sea. FEMS Microbiol Ecol 47: 387–396. doi:10.1016/S0168-6496(03)00305-2.

35 - Martens T, Gram L, Grossart H-P, Kessler D, Muller R, et al. (2007) Bacteria of the Roseobacter clade show potential for secondary metabolite production. Microb Ecol 54. doi:10.1007/s00248-006-9165-2.

36 - Milne PJ, Hunt AL, Rostoll K, Van Der Walt JJ, Graz CJ (1998) The biological activity of selected cyclic dipeptides. J Pharm Pharmacol 50: 1331–1337.

37 - Slightom RN, Buchan A (2009) Surface colonization by marine roseobacters: integrating genotype and phenotype. Appl Environ Microbiol 75: 6027–6037. doi:10.1128/AEM.01508-09.

38 - Cude WN, Mooney J, Tavanaei AA, Hadden MK, Frank AM, et al. (2012) Production of the antimicrobial secondary metabolite indigoidine contributes to competitive surface colonization by the marine roseobacter Phaeobacter sp. strain Y4I. Appl Environ Microbiol 78: 4771–4780. doi:10.1128/AEM.00297-12.

39- Wagner G, Jardim R, Tschoeke DA, Loureiro DR, Ocaña KA, et al. (2014) Stingray: System for integrated genomic resources and analysis. BMC Research Notes 7: 132.

40- Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC bioinformatics 11: 187.

41 – Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics 11. doi:10.1186/1471-2105-11-187.

42 - Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410. doi:10.1016/S0022-2836(05)80360-2.

43 – Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Research 17: 377–386. doi:10.1101/gr.5969107.

44 – Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, et al. (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386. doi:10.1186/1471-2105-9-386.

45 – Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29: 2933–2935. doi:10.1093/bioinformatics/btt509.

46 – Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41: D590–D596. doi:10.1093/nar/gks1219.

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

48 - Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Research 38: e132–e132. doi:10.1093/nar/gkq275.

49 – Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276–277.

50 - Romão-Dumaresq AS, Fróes AM, Cuadrat RRC, Silva FP, Dávila AMR (2014) Towards a Comprehensive Search of Putative Chitinases Sequences in Environmental Metagenomic Databases. Natural Science 06: 323–337. doi:10.4236/ns.2014.65034.

51 - Tae H, Sohng JK, Park K (2009) MapsiDB: an integrated web database for type I polyketide synthases. Bioprocess Biosyst Eng 32: 723–727. doi:10.1007/s00449-008-0296-3.

52 – Ansari MZ, Yadav G, Gokhale RS, Mohanty D (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Research 32: W405–W413. doi:10.1093/nar/gkh359.

53 – Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059–3066.

54 – Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Computational Biology 7: e1002195. doi:10.1371/journal.pcbi.1002195.

55 – Ziemert N, Podell S, Penn K, Badger JH, Allen E, et al. (2012) The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. PLoS ONE 7: e34064. doi:10.1371/journal.pone.0034064.

56- Dickson M-L, Wheeler PA (1995) Ammonium uptake and regeneration rates in a coastal upwelling regime. Available: http://ir.library.oregonstate.edu/xmlui/handle/1957/13387.

57 - Seeyave S, Probyn T, Álvarez-Salgado XA, Figueiras FG, Purdie DA, et al. (2013) Nitrogen uptake of phytoplankton assemblages under contrasting upwelling and downwelling conditions: The Ría de Vigo, NW Iberia. Estuarine, Coastal and Shelf Science 124: 1–12. doi:10.1016/j.ecss.2013.03.004.

58 - Guenther M, Gonzalez-Rodriguez E, Carvalho W, Rezende C, Mugrabe G, et al. (2008) Plankton trophic structure and particulate organic carbon production during a

coastal downwelling-upwelling cycle. Marine Ecology Progress Series 363: 109–119. doi:10.3354/meps07458.

59 - Albuquerque ALS, Belem AL, Zuluaga FJB, Cordeiro LGM, Mendoza U, et al. (2014) Particle Fluxes and Bulk Geochemical Characterization of the Cabo Frio Upwelling System in Southeastern Brazil: Sediment Trap Experiments between Spring 2010 and Summer 2012. An Acad Bras Cienc.

60 - Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, et al. (2013) GenBank. Nucleic Acids Research 41: D36–D42. doi:10.1093/nar/gks1195.

61- Thomas F, Hehemann J-H, Rebuffet E, Czjzek M, Michel G (2011) Environmental and Gut Bacteroidetes: The Food Connection. Frontiers in Microbiology 2.

62 - Morel A, Ahn Y-H, Partensky F, Vaulot D, Claustre H (1993) Prochlorococcus and Synechococcus: A comparative study of their optical properties in relation to their size and pigmentation. Journal of Marine Research 51: 617–649.

63- Yew, S. P., Jau, M. H., Yong, K. H., Abed, R. M. M. and Sudesh, K (2005) Morphological Studies of Synechocystis sp. UNIWG under Polyhydroxyalkanoate Accumulating Conditions. Malaysian Journal of Microbiology 1:48-52.

64 – Partensky F, Blanchot J, Vaulot D (1999) Differential distribution and ecology of Prochlorococcus and Synechococcus in oceanic waters : a review. Monaco, MONACO: Musée océanographique 19:457-475.

65 - Scanlan DJ, West NJ (2002) Molecular ecology of the marine cyanobacterial genera Prochlorococcus and Synechococcus. FEMS Microbiol Ecol 40. doi:10.1111/j.1574-6941.2002.tb00930.x.

66 - Preston C, Harris A, Ryan JP, Roman B, Marin R, Jensen S et al. (2011). Application of quantitative PCR on a coastal mooring. PLOS One 6: e22522.

67 - Palenik B, Ren Q, Dupont CL, Myers GS, Heidelberg JF, et al. (2006) Genome sequence of Synechococcus CC9311: insights into adaptation to a coastal environment. Proceedings of the National Academy of Sciences 103: 13555–13559.

68 - Palenik B (2001) Chromatic Adaptation in Marine Synechococcus Strains. Applied and Environmental Microbiology 67: 991–994. doi:10.1128/AEM.67.2.991-994.2001.

69 – Garrity GM, Bell JA Lilburn T (2005) Family I. Rhodobacteraceae fam. nov. In: Brenner DJ, Krieg NR, Staley JT, Garrity M ,editors. Bergey's Manual of Systematic Bacteriology. (The Proteobacteria), part C (The Alpha-, Beta-, Delta-, and Epsilonproteobacteria), Springer 2:161.

70 **-** Ivanova EP, Mikhailov VV (2001) A new family, Alteromonadaceae fam. nov., including marine proteobacteria of the genera Alteromonas, Pseudoalteromonas, Idiomarina, and Colwellia. Microbiology 70: 10–17.

71 - Ivanova EP (2004) Phylogenetic relationships among marine Alteromonas-like proteobacteria: emended description of the family Alteromonadaceae and proposal of Pseudoalteromonadaceae fam. nov., Colwelliaceae fam. nov., Shewanellaceae fam. nov., Moritellaceae fam. nov., Ferrimonadaceae fam. nov., Idiomarinaceae fam. nov. and Psychromonadaceae fam. nov. International Journal of Systematic and Evolutionary Microbiology 54: 1773–1788. doi:10.1099/ijs.0.02997-0.

72 - Kwak M-J, Song JY, Kim BK, Chi W-J, Kwon S-K, et al. (2012) Genome Sequence of the Agar-Degrading Marine Bacterium Alteromonadaceae sp. Strain G7. Journal of Bacteriology 194: 6961–6962. doi:10.1128/JB.01931-12.

73 - Bernardet J-F, Nakagawa Y (2006) An Introduction to the Family Flavobacteriaceae. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E, editors. The Prokaryotes. Springer New York. pp. 455–480. Available: http://dx.doi.org/10.1007/0-387-30747-8_16.

74 - Cottrell MT, Kirchman DL (2000) Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low-and high-molecular-weight dissolved organic matter. Applied and Environmental Microbiology 66: 1692–1697.

75 - Zhang X-Y, Xie B-B, Qin Q-L, Liu A, Chen X-L, et al. (2012) Draft Genome Sequence of Strain P7-3-5, a New Flavobacteriaceae Bacterium Isolated from Intertidal Sand. Journal of Bacteriology 194: 6632–6632. doi:10.1128/JB.01748-12.

76 - Yurkov VV, Beatty JT (1998) Aerobic anoxygenic phototrophic bacteria. Microbiology and Molecular Biology Reviews 62: 695–724.

77 - Buchan A, Gonzalez JM, Moran MA (2005) Overview of the Marine Roseobacter Lineage. Applied and Environmental Microbiology 71: 5665–5677. doi:10.1128/AEM.71.10.5665-5677. 2005.

78 - González JM, Kiene RP, Moran MA (1999) Transformation of Sulfur Compounds by an Abundant Lineage of Marine Bacteria in the α-Subclass of the ClassProteobacteria. Applied and environmental microbiology 65: 3810–3819.

79 - Vila-Costa M, Simo R, Harada H, Gasol JM, Slezak D, et al. (2006) Dimethylsulfoniopropionate uptake by marine phytoplankton. Science 314: 652–654. doi:10.1126/science.1131043.

80 - Uchino Y, Hirata A, Yokota A, Sugiyama J (1998) Reclassification of marine *Agrobacterium* species: Proposals of *Stappia stellulata* gen. nov., comb. nov., *Stappia aggregata* sp. nov., nom. rev., *Ruegeria atlantica* gen. nov., comb. nov., *Ruegeria gelatinovora* comb. nov., *Ruegeria algicola* comb. nov., and *Ahrensia kieliense* gen. nov., sp. nov., nom. rev. The Journal of General and Applied Microbiology 44: 201–210.

81 - Nedashkovskaya OI, Kwon KK, Yang S-H, Lee H-S, Chung KH, et al. (2008) Lacinutrix algicola sp. nov. and Lacinutrix mariniflava sp. nov., two novel marine alga-associated bacteria and emended description of the genus Lacinutrix. INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY 58: 2694–2698. doi:10.1099/ijs.0.65799-0.

82 - Rappé, Michael S.; Connon, Stephanie A.; Vergin, Kevin L.; Giovannoni SJ (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418. Available: http://libgen.org/scimag/index.php?doi=10.1038/nature00917.

83 - Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, et al. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420: 806–810. doi:10.1038/nature01240.

84 - Allen LZ, Allen EE, Badger JH, McCrow JP, Paulsen IT, et al. (2012) Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. The ISME journal 6: 1403–1414.

85 - Alonso-Gutiérrez J, Lekunberri I, Teira E, Gasol JM, Figueras A, et al. (2009) Bacterioplankton composition of the coastal upwelling system of “Ría de Vigo”, NW Spain. FEMS Microbiology Ecology 70: 493–505. doi:10.1111/j.1574-6941.2009.00766.x.

86 - Henriques IS, Almeida A, Cunha A, Correia A (2004) Molecular sequence analysis of prokaryotic diversity in the middle and outer sections of the Portuguese estuary Ria de Aveiro. FEMS Microbiol Ecol 49: 269–279. doi:10.1016/j.femsec.2004.04.003.

87- Alonso-Saez L, Balague V, Sa EL, Sanchez O, Gonzalez JM, et al. (2007) Seasonality in bacterial diversity in north-west Mediterranean coastal waters: assessment through clone libraries, fingerprinting and FISH. FEMS Microbiol Ecol 60. doi:10.1111/j.1574-6941.2006.00276.x.

88 - Gonzalez JM, Simo R, Massana R, Covert JS, Casamayor EO, et al. (2000) Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. Appl Environ Microbiol 66: 4237–4246.

89 - Zubkov MV, Fuchs BM, Archer SD, Kiene RP, Amann R, et al. (2002) Rapid turnover of dissolved {DMS} and {DMSP} by defined bacterioplankton communities in the stratified euphotic zone of the North Sea. Deep Sea Research Part II: Topical Studies in Oceanography 49: 3017 – 3038. doi:http://dx.doi.org/10.1016/S0967-0645(02)00069-3.

90 **-** Nelson DM, Tréguer P, Brzezinski MA, Leynaert A, Quéguiner B (1995) Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem Cycles 9: 359–372. doi:10.1029/95GB01070.

91 - Raven JA, Waite AM (2004) The evolution of silicification in diatoms: inescapable sinking and sinking as escape? New Phytologist 162: 45–61. doi:10.1111/j.1469-8137.2004.01022.x.

92 - Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, et al. (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC Genomics 10: 624. doi:10.1186/1471-2164-10-624.

93 - Worden AZ, Nolan JK, Palenik B (2004) Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. Limnology and Oceanography 49: 168–179.

94 - Piganeau G, Moreau H (2007) Screening the Sargasso Sea metagenome for data to investigate genome evolution in Ostreococcus (Prasinophyceae, Chlorophyta). Gene 406: 184–190. doi:10.1016/j.gene.2007.09.015.

95 - Fuhrman JA, Hangstro¨m A (2008) Bacterial and Archaeal community structureand its patterns. In: Kirchman DL, ed. Microbial ecology of the oceans. New York: Wiley. pp 45–90.

96 - Lai B, Ding R, Li Y, Duan L, Zhu H (2012) A de novo metagenomic assembly program forshotgun DNA reads. Bioinformatics 28: 1455–1462. doi:10.1093/bioinformatics/bts162.

97 - Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Research 40: e155–e155. doi:10.1093/nar/gks678.

98 - Afiahayati, Sato K, Sakakibara Y (2013) An extended genovo metagenomic assembler by incorporating paired-end information. PeerJ 1: e196. doi:10.7717/peerj.196.

99 - Reddy RM, Mohammed MH, Mande SS (2014) MetaCAA: A clustering-aided methodology for efficient assembly of metagenomic datasets. Genomics. doi:10.1016/j.ygeno.2014.02.007.

100 - Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods 4. doi:10.1038/nmeth1043.

101 - Pignatelli M, Moya A (2011) Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data. PLoS ONE 6: e19984. doi:10.1371/journal.pone.0019984.

102 - Kennedy J, Codling CE, Jones BV, Dobson ADW, Marchesi JR (2008) Diversity of microbes associated with the marine sponge, Haliclona simulans, isolated from Irish waters and identification of polyketide synthase genes from the sponge metagenome. Environ Microbiol 10: 1888–1902. doi:10.1111/j.1462-2920.2008.01614.x.

103 - Trindade-Silva AE, Rua CPJ, Andrade BGN, Vicente ACP, Silva GGZ, et al. (2013) Polyketide synthase gene diversity within the microbiome of the sponge Arenosclera brasiliensis, endemic to the Southern Atlantic Ocean. Appl Environ Microbiol 79: 1598–1605. doi:10.1128/AEM.03354-12.

104 - Foerstner KU, Doerks T, Creevey CJ, Doerks A, Bork P (2008) A Computational Screen for Type I Polyketide Synthases in Metagenomics Shotgun Data. PLoS ONE 3: e3515. doi:10.1371/journal.pone.0003515.

105 - Jenke-Kodama H, Sandmann A, Muller R, Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. Mol Biol Evol 22: 2027–2039. doi:10.1093/molbev/msi193.

106 **–** Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554–557. doi:10.1126/science.1107851.

107 **-** Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5: R245–249.

108 - Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304. doi:10.1126/science.1093857.

109 - Fisch KM (2013) Biosynthesis of natural products by microbial iterative hybrid PKS–NRPS. RSC Advances 3: 18228. doi:10.1039/c3ra42661k.

110 - Slattery M, Rajbhandari I, Wesson K (2001) Competition-mediated antibiotic induction in the marine bacterium Streptomyces tenjimariensis. Microb Ecol 41: 90–96. doi:10.1007/s002480000084.

TABLES

Table 1: Statistics of the sequences from sample P and sample E.

Analysis	Sample P	Sample E
Upload: bp Count	**350,520,599 bp**	**279,401,007 bp**
Upload: Sequences Count	**595,497**	**468,832**
Upload: Mean Sequence Length	**588 ± 142 bp**	**595 ± 171 bp**
Upload: Mean GC percent	**46 ± 7 %**	**43 ± 10 %**
Post QC: bp Count	**350,507,486 bp**	**279,391,534 bp**
Post QC: Sequences Count	**595,464**	**468,792**
Processed: Predicted Protein Features	**500,671**	**515,393**
Processed: Predicted rRNA Features	**33,395**	**58,226**
Alignment: Identified Protein Features	**333,934**	**181,618**
Alignment: Identified rRNA Features	**1,409**	**3,035**
Annotation: Identified Functional Categories	**280,938**	**127,741**

FIGURES

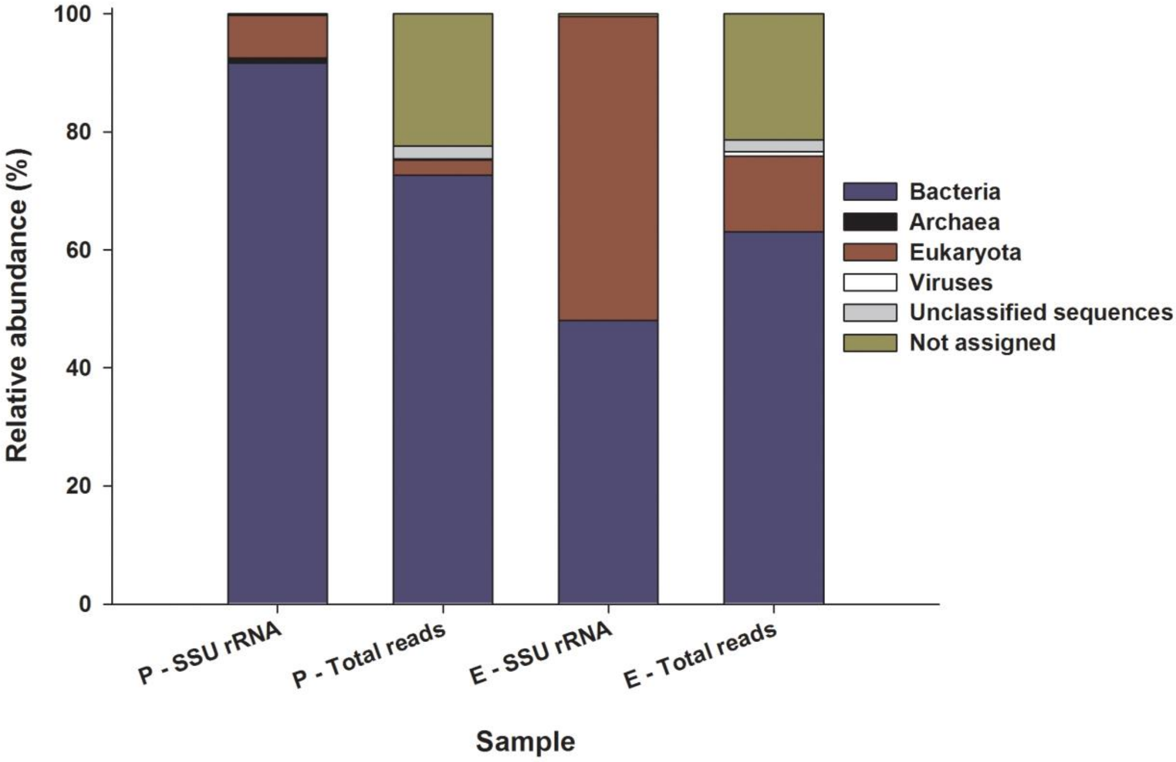


Figure 1. Taxonomic affiliation of SSU rDNA and total reads using MEGAN. SSU rDNA sequences from both samples (P and E) extracted using INFERNAL and classified by MEGAN at domain of life level using BLASTN result against SILVA SSU database and LCA algorithm. Total of reads from both samples (P and E) was classified by MEGAN at domain of life level using BLASTN result against GenBank (NT) and LCA algorithm.

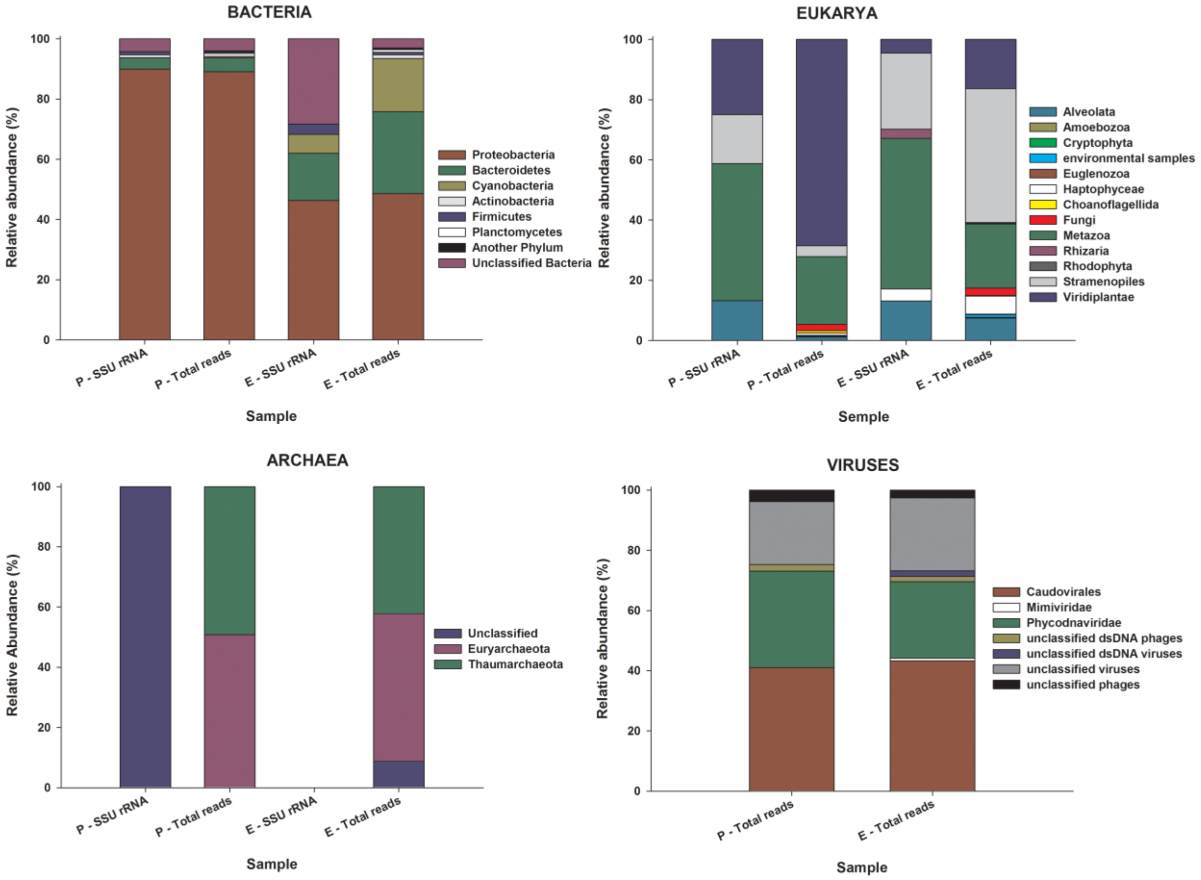


Figure 2. Comparative taxonomic affiliation of Eukarya, Archaea, Bacteria and Viruses reads using MEGAN. Reads from the three domains of life viruses and from both samples (P and E) were classified by MEGAN (Phylum level for bacteria and Archaea, family of viruses and Kingdom level of Eukaryota) using BLASTN result against GenBank (NT) and using BLASTN results from INFERNAL extracted reads (SSU rDNA) against SILVA SSU. In both analyses the algorithm used was LCA.

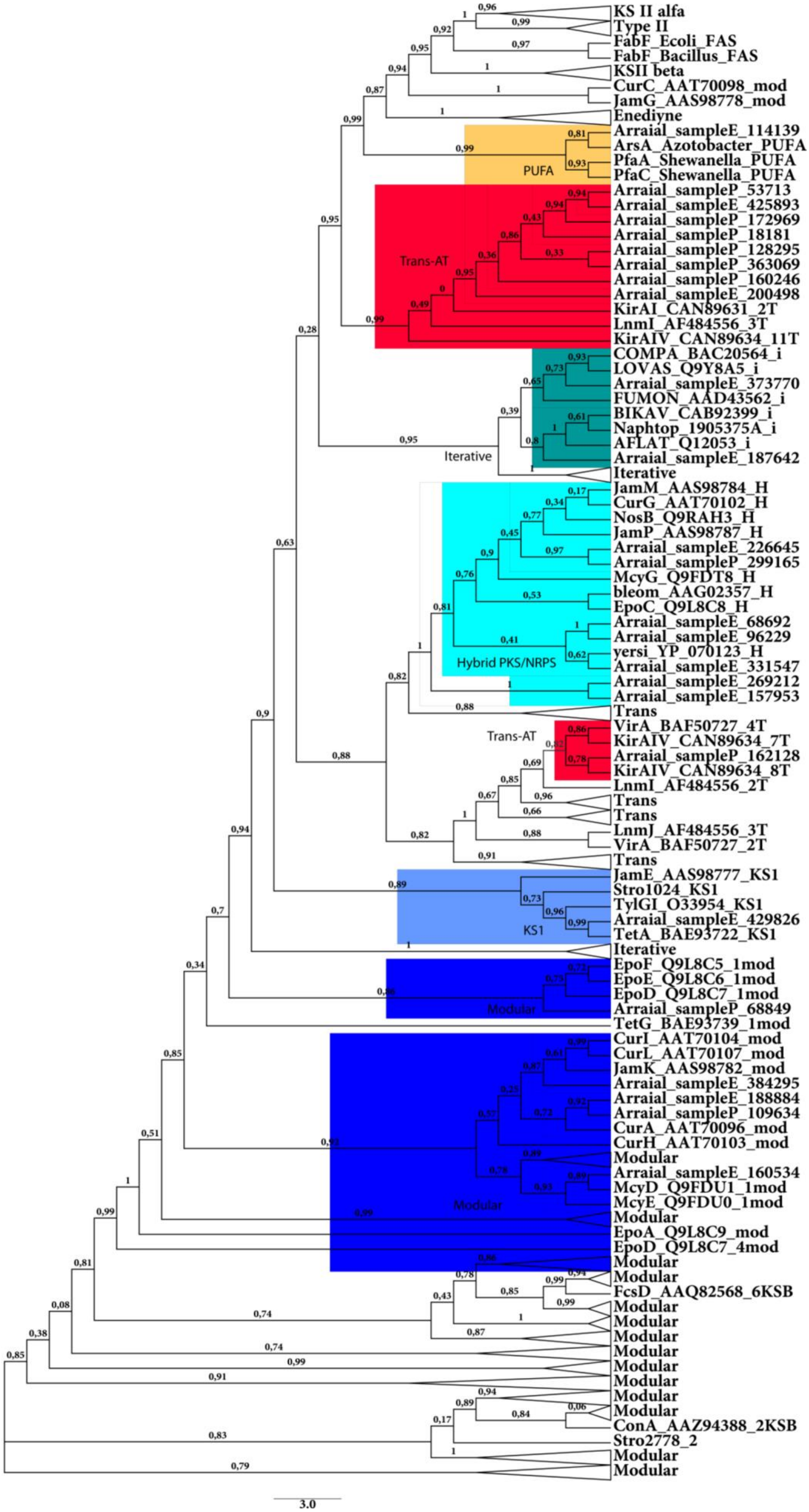

KS II alfa
Type II
FabF_Ecoli_FAS
FabF_Bacillus_FAS
KSII beta
CurC_AAT70098_mod
JamG_AAS98778_mod
Enediyne
Arraial_sampleE_114139
ArsA_Azotobacter_PUFA
PfaA_Shewanella_PUFA
PfaC_Shewanella_PUFA
PUFA
Arraial_sampleP_53713
Arraial_sampleE_425893
Arraial_sampleP_172969
Arraial_sampleP_18181
Arraial_sampleP_128295
Arraial_sampleP_363069
Arraial_sampleP_160246
Arraial_sampleE_200498
KirAI_CAN89631_2T
LnmI_AF484556_3T
KirAIV_CAN89634_11T
Trans-AT
COMPA_BAC20564_i
LOVAS_Q9Y8A5_i
Arraial_sampleE_373770
FUMON_AAD43562_i
BIKAV_CAB92399_i
Naphtop_1905375A_i
AFLAT_Q12053_i
Arraial_sampleE_187642
Iterative
Iterative
JamM_AAS98784_H
CurG_AAT70102_H
NosB_Q9RAH3_H
JamP_AAS98787_H
Arraial_sampleE_226645
Arraial_sampleP_299165
McyG_Q9FDT8_H
bleom_AAG02357_H
EpoC_Q9L8C8_H
Arraial_sampleE_68692
Arraial_sampleE_96229
yersi_YP_070123_H
Arraial_sampleE_331547
Arraial_sampleE_269212
Arraial_sampleE_157953
Hybrid PKS/NRPS
Trans
Trans-AT
VirA_BAF50727_4T
KirAIV_CAN89634_7T
Arraial_sampleP_162128
KirAIV_CAN89634_8T
LnmI_AF484556_2T
Trans
Trans
LnmJ_AF484556_3T
VirA_BAF50727_2T
Trans
JamE_AAS98777_KS1
Stro1024_KS1
TylGI_O33954_KS1
Arraial_sampleE_429826
TetA_BAE93722_KS1
KS1
Iterative
EpoF_Q9L8C5_1mod
EpoE_Q9L8C6_1mod
EpoD_Q9L8C7_1mod
Arraial_sampleP_68849
Modular
TetG_BAE93739_1mod
CurI_AAT70104_mod
CurL_AAT70107_mod
JamK_AAS98782_mod
Arraial_sampleE_384295
Arraial_sampleE_188884
Arraial_sampleP_109634
CurA_AAT70096_mod
CurH_AAT70103_mod
Modular
Arraial_sampleE_160534
McyD_Q9FDU1_1mod
McyE_Q9FDU0_1mod
Modular
Modular
EpoA_Q9L8C9_mod
EpoD_Q9L8C7_4mod
Modular
Modular
FcsD_AAQ82568_6KSB
Modular
Modular
Modular
Modular
Modular
Modular
Modular
Modular
ConA_AAZ94388_2KSB
Stro2778_2
Modular
Modular
3.0

Figure 3. Phylogenetic tree of environmental KS domains (larger than 200 amino acids) obtained from both sample (by KS modular and iterative pHMMs) with reference NapDoS sequences. The tree was generated by NapDos pipeline (using FASTTREE and Maximum Likelihood algorithm). Confidence values are showed on nodes.

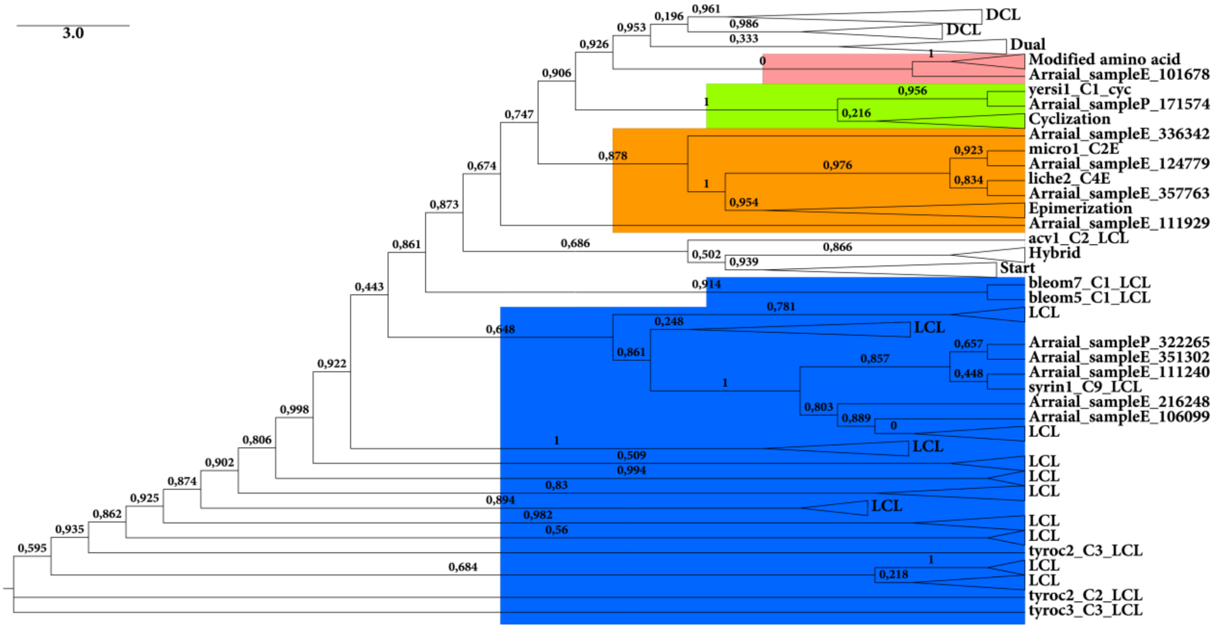


Figure 4. Phylogenetic tree of environmental C domains (larger than 200 amino acids) obtained from both sample (by C domain pHMM) with reference NapDoS sequences. The tree was generated by NapDos pipeline (using FASTTREE and Maximum Likelihood algorithm). Confidence values are showed on nodes.

SUPPLEMENTARY MATERIAL

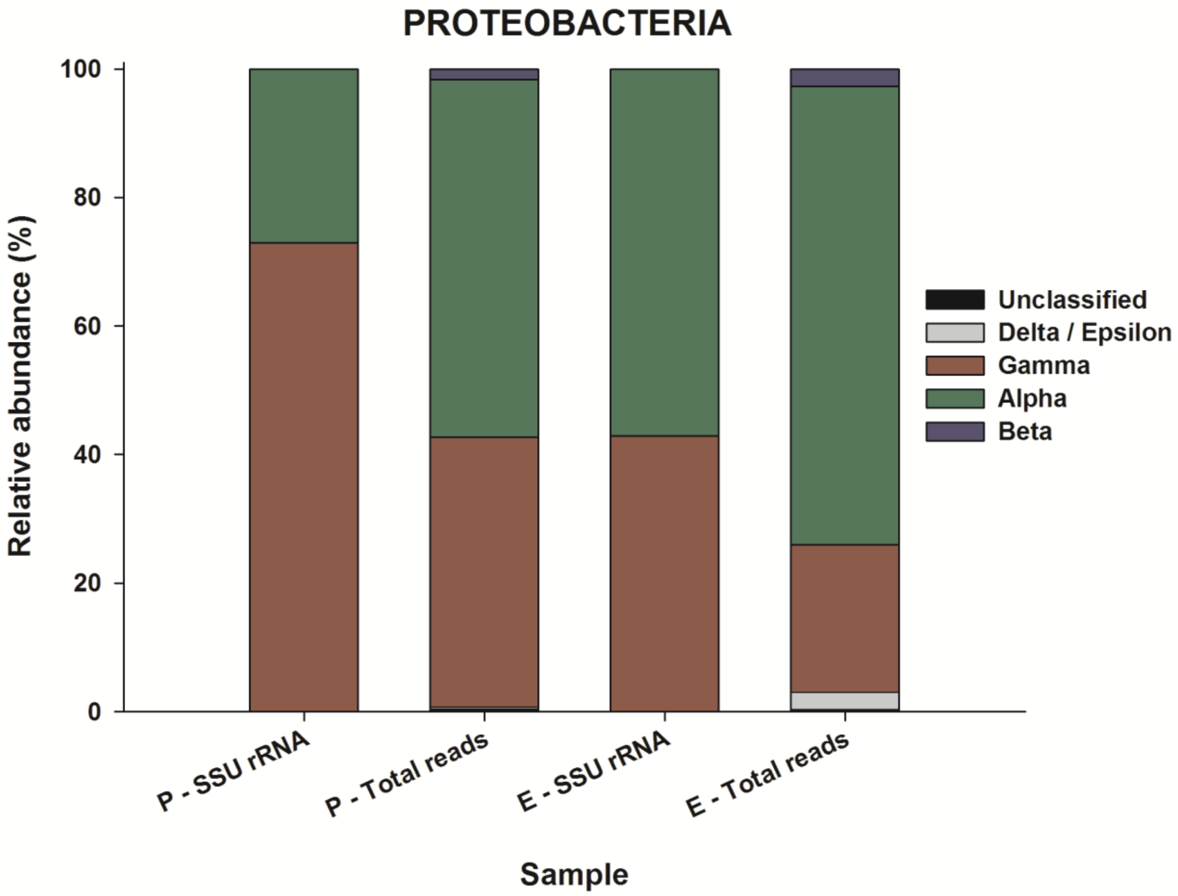


Figure S1. Comparative taxonomic affiliation of Proteobacteria reads at class level. Infernal SSU rDNA extracted reads and total reads from samples P and E were classified by MEGAN using BLASTN result against GenBank (NT) and SILVA SSU. In both analyses the algorithm used was LCA.

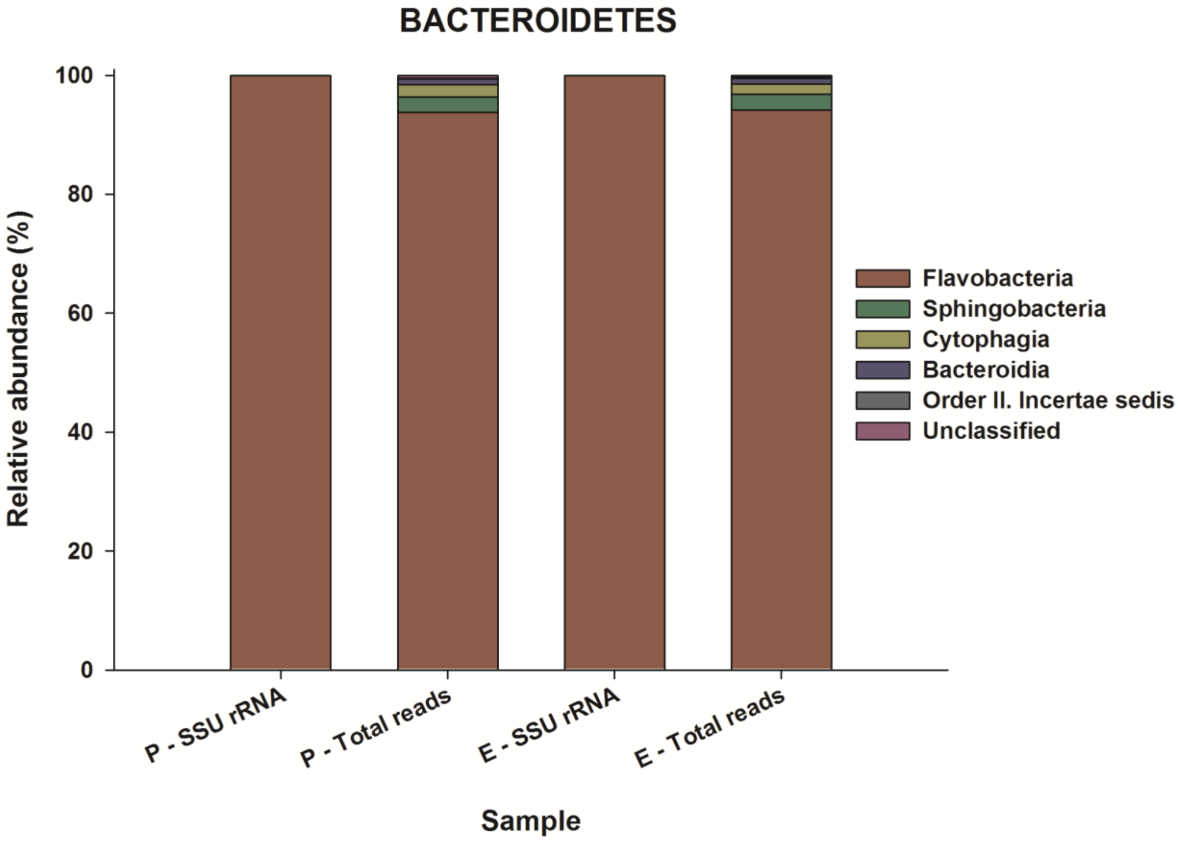


Figure S2. Comparative taxonomic affiliation of Bacteroidetes reads at class level. INFERNAL extracted reads (SSU rDNA) and total reads of samples P and E were classified by MEGAN using BLASTN result against GenBank (NT) and SILVA SSU. In both analyses the algorithm used was the LCA.

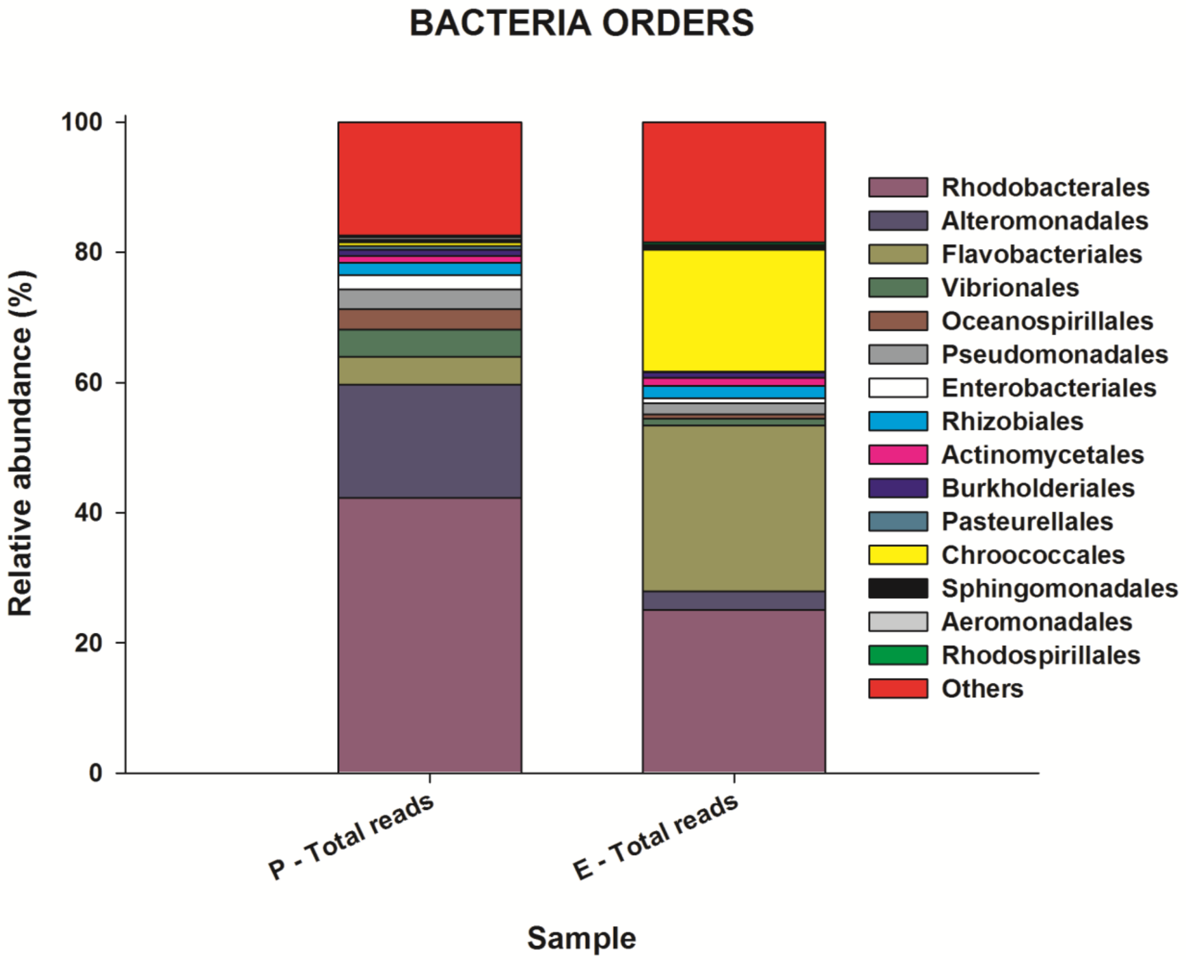


Figure S3.Comparative taxonomic affiliation of bacteria reads at order level. Total reads ofsamples P and E were classified by MEGAN using BLASTN results against GenBank (NT) using LCA algorithm. Only the 15 most abundant groups are showed.

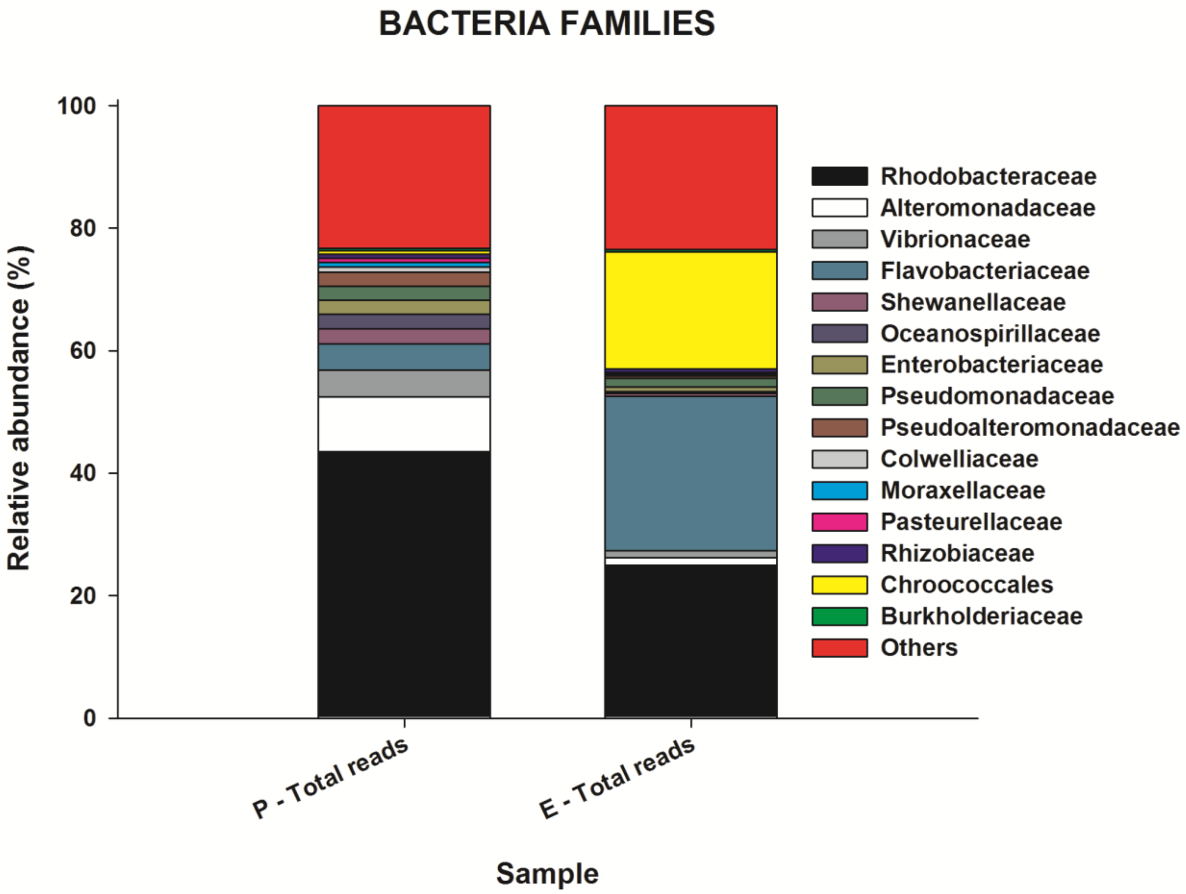


Figure S4.Comparative taxonomic affiliation of bacteria reads at family level. Total reads ofsamples P and E were classified by MEGAN using BLASTN results against GenBank (NT) using LCA algorithm. Only the 15 most abundant groups are showed

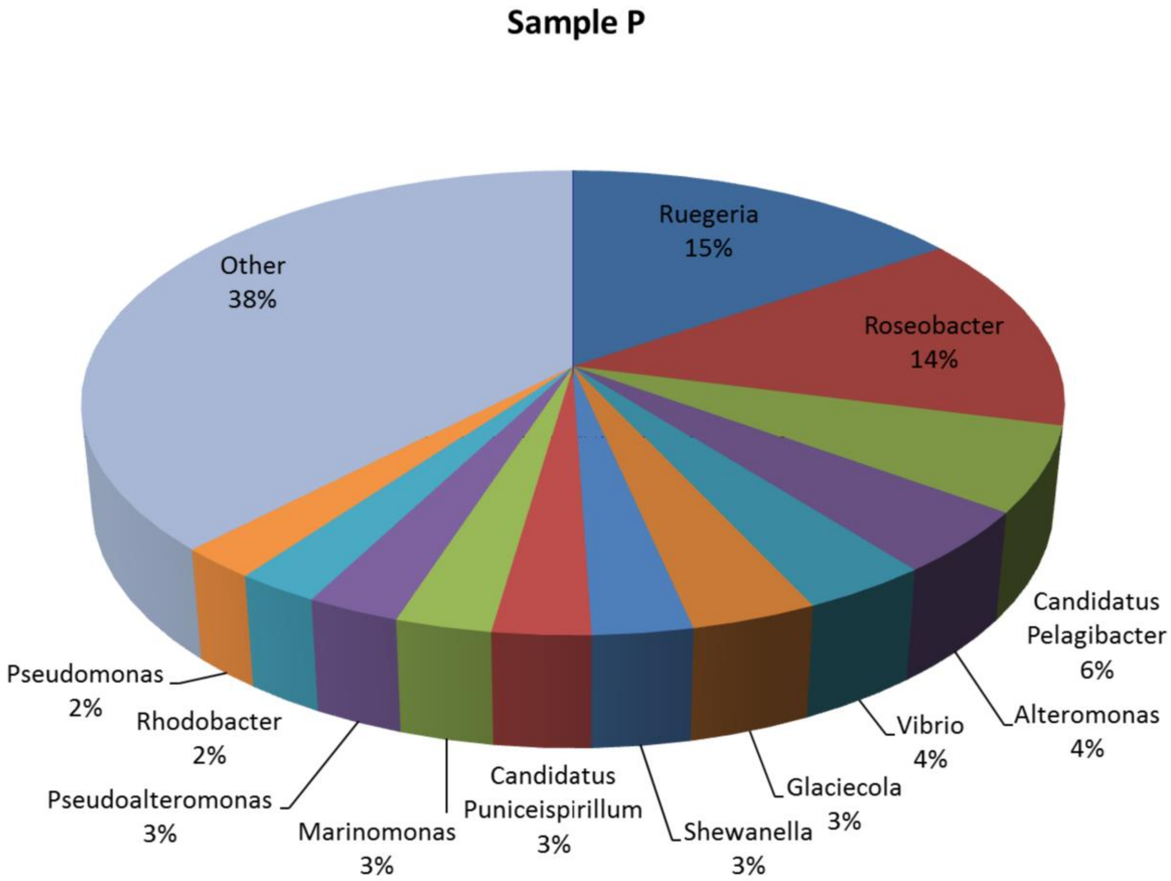


Figure S5. **Most abundant identified genera in sample P obtained analyzing the BlastN results in MEGAN 4.0 using LCA algorithm.**

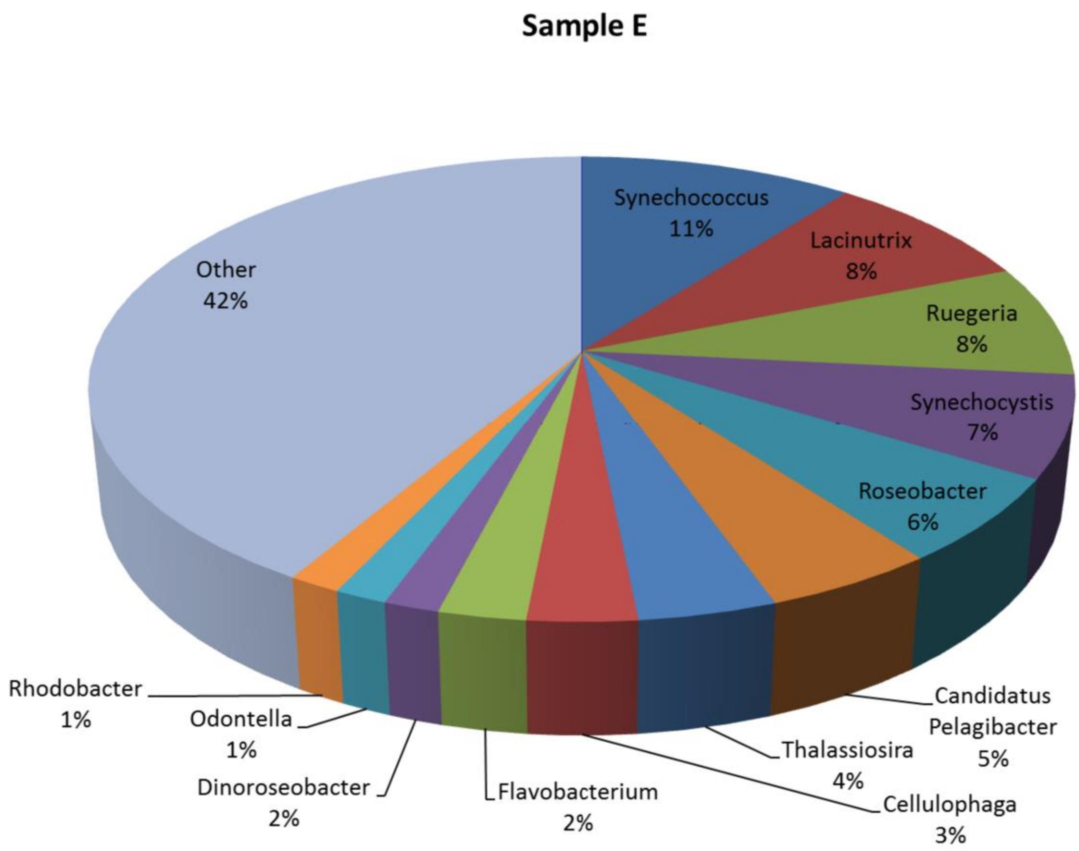


Figure S6. Most abundant identified genera in sample E obtained analyzing BlastN results in MEGAN 4.0 using LCA algorithm).

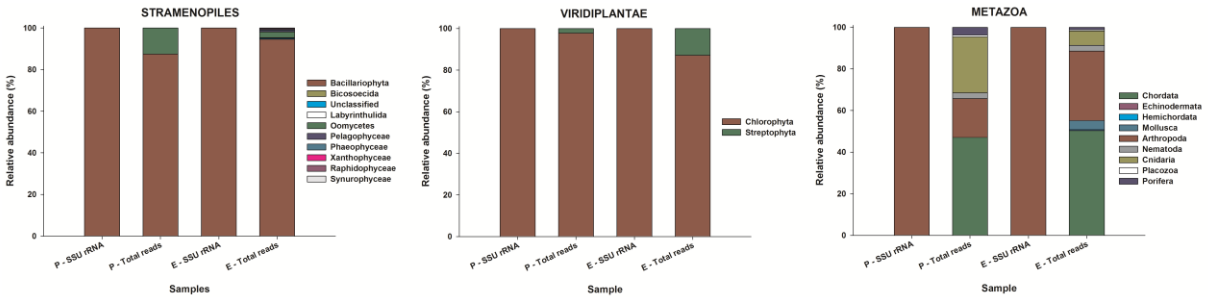


Figure S7: Comparative taxonomic affiliation of Stramenopiles, Metazoa and Viriplantae reads at class level. Infernal extracted reads (SSU rDNA) and total

reads ofsamples P and E were classified by MEGAN using BLASTN results against GenBank (NT) and SILVA SSU. In both analyses the algorithm used was LCA.

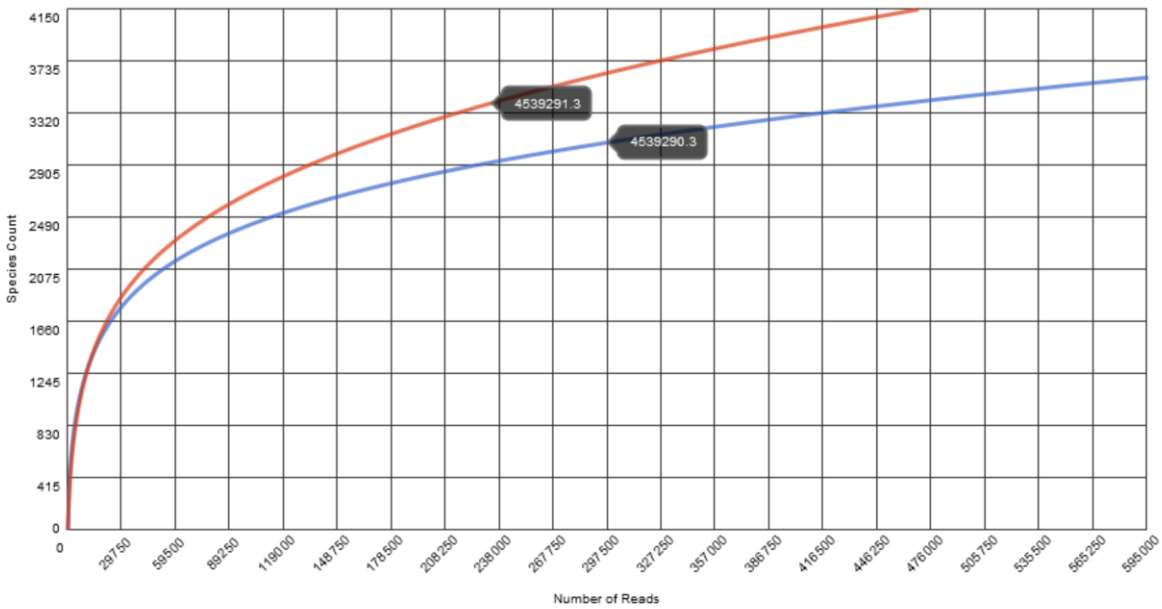


Figure S8. Rarefaction curve calculated by MG-RAST. Data calculated for metagenomes 4539291.3 (sample E) and 4539290.3 (sample P).

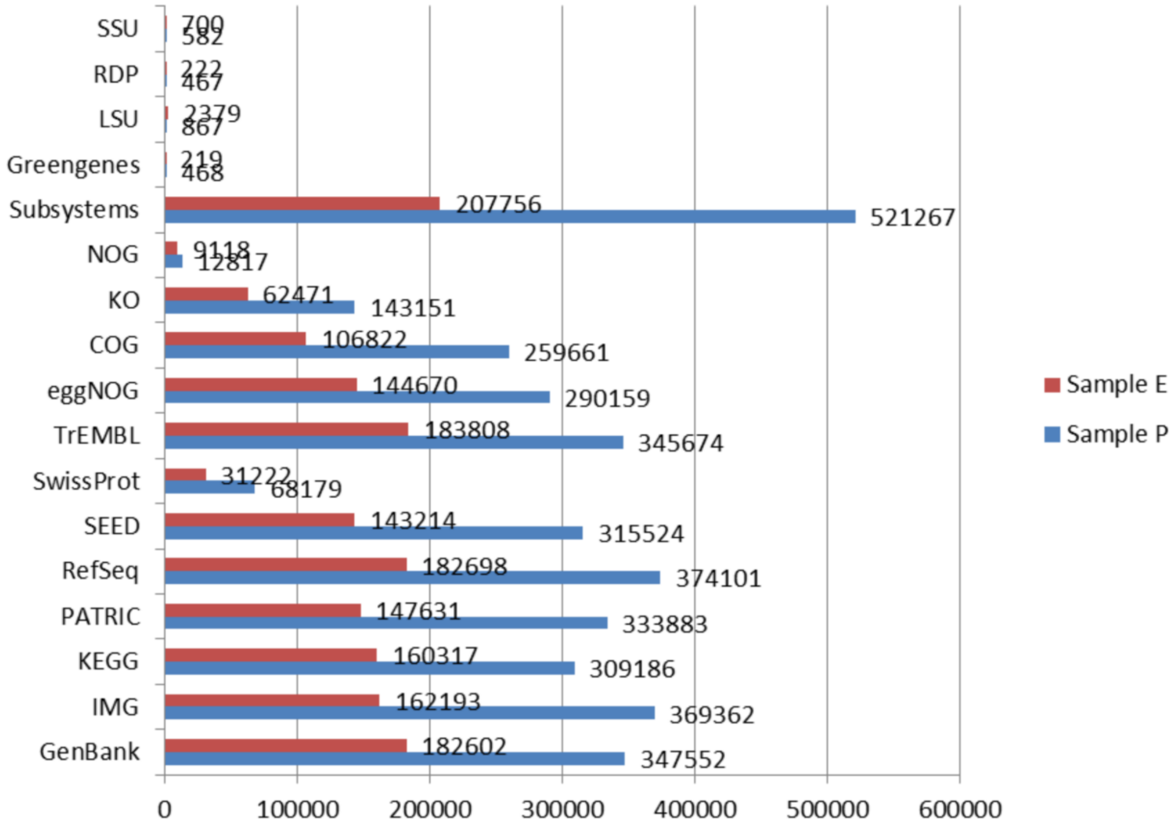
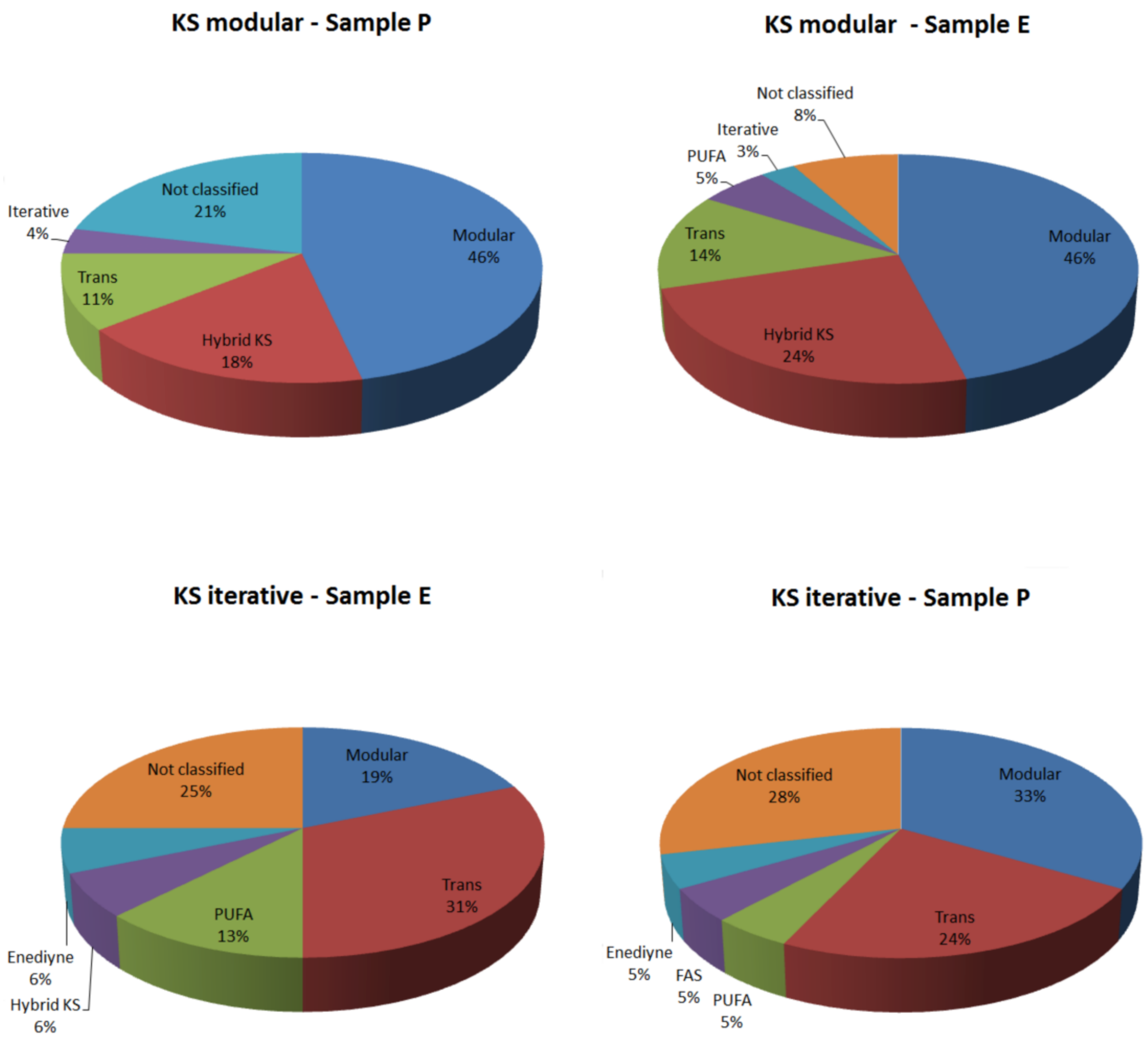


Figure S9. Number of hits obtained with each database from M5NR on MG-RAST for sample P and sample E.



KS modular - Sample P

KS modular - Sample E

KS iterative - Sample E

KS iterative - Sample P

Figure S10. The figure shows the classification of KS domain (obtained with pHMM KS modular and pHMM KS iterative, from both samples) by NapDos.

4 TRABALHO 2: “A NEW PROFILE HMM APPROACH REVEALS A HIGH FRACTION OF AEROBIC ANOXYGENIC PHOTOTROPHIC BACTERIA (AAP) IN METAGENOME FROM A TROPICAL OLIGOTROPHIC COASTAL BAY (ARRAIAL DO CABO – BRAZIL)”

Este trabalho foi desenvolvido durante o estágio realizado no exterior (doutorado sanduíche) com bolsa do programa “Ciência sem Fronteiras”. O laboratório onde o trabalho foi realizado fica localizado em Berlim, Alemanha. O orientador estrangeiro foi o Dr. Hans-Peter Grossart, do *Leibniz Institute of Freshwater Ecology and Inland Fisheries* (IGB), e participante do *Berlin Center for Genomics in Biodiversity Research* (BeGenDiv). O mesmo será submetido para a revista Applied Environmental Microbiology (AEM).

No estudo anterior notamos que a amostra estudada é dominada por organismos do gênero *Roseobacter* e outros grupos taxonômicos próximos.

Este gênero é conhecido por possuir diversas espécies capazes de realizar fotossíntese anoxigênica, conhecidos como AAPs. Por este motivo, foi desenvolvido neste trabalho um pipeline para estimar a abundância e diversidade de AAPs na amostra da Praia dos Anjos, através da triagem de genes marcadores, mostrando uma alta abundância desse tipo de bactéria neste ambiente, provavelmente associada ao fenômeno da ressurgência e a abundância de luz disponível durante o verão na região.

A new profile HMM approach reveals a high fraction of aerobic anoxygenic phototrophic bacteria (AAP) in metagenomes from a coastal bay (Arraial do Cabo – Brazil)

Rafael R. C. Cuadrat[1,2], Isabel Ferrera[2,4] , Hans-Peter Grossart[2,3], Alberto M. R. Davila[1]

[1] Computational and Systems Biology Laboratory, Computational and Systems Biology Pole, Oswaldo Cruz Institute, Fiocruz, E-mail: davila@fiocruz.br

[2] Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Germany

[3] Potsdam University, Institute for Biochemistry and Biology, Potsdam, Germany

[4]Institut de Ciències del Mar, CSIC, Barcelona, Spain

ABSTRACT

Aerobic anoxygenic phototrophic bacteria (AAP) play important roles in carbon and energy cycling in various aquatic systems. Several studies demonstrate the great metabolic versatility of these bacteria. In oceans, this specific bacterial group is widely distributed, however, the abundance and importance for aquatic carbon fixation and biomass production is still rather poorly understood. Therefore, we evaluated the abundance and diversity of AAPs in a metagenome from a tropical bay (Arraial do Cabo, Brazil) by developing Profile Hiden Markov Models (pHMM) as a new *in silico* approach to screen for core genes of anoxygenic photosynthesis (*pufM* and *pufL*), in addition to the chlorophyllide reductase subunit X gene (*bchX*). Metagenomes from the Global Ocean Sample Expedition (GOS) were additionally screened for comparative purposes. In the free-living bacterial fraction (<0.8-0.22 µm, sample P) AAPs were highly abundant in the coastal bay from Arraial do Cabo (~23.88% of total bacterial cells, whereas in the GOS the abundance was up to ~15%). Ten samples from the GOS dataset which showed the highest fraction of AAPs and our 2 samples from Arraial do Cabo were selected for assembly, ORF extraction and phylogenetic analysis of *pufM* genes. Interestingly, most selected GOS samples (80%) were originated from sites close to the equator line. We were able to assign most of the retrieved sequences to specific phylogroups with a particularly high abundance of phylogroup G (*Roseobacter* clade) in Arraial do Cabo samples.

INTRODUCTION

Aerobic anoxygenic phototrophic bacteria (AAPs) potentially play important roles in carbon and energy cycling in freshwater and marine systems. They require oxygen and reduced organic compounds to grow (1), but on the other hand they produce the pigment bacteriochlorophyll *a* (Bchla) and use it to generate additional ATP. Many studies have demonstrated the great metabolic potential of these bacteria, which includes nitrification, carbon dioxide fixation, carotenoids synthesis and the use of low-molecular-weight organic carbon as energy source (2, 3, 4). Therefore, they can inhabit a wide variety of different environments ranging from terrestrial to aquatic systems both marine and freshwater including extreme environments like Antarctic lakes (5, 6, 7). In oceans, this group is widely distributed (8, 9), however, their abundance and importance for carbon fixation and energy cycling is still poorly understood (9, 10).

Many studies were performed in order to estimate AAP abundance and diversity in marine environments, using many different approaches, e.g. fluorescence detection of Bchla (1, 11, 12, 13), qPCR (14, 15), pyrosequencing (8), and metagenomic approaches (16, 17, 18, 19). In the study performed by Yutin *et al* (19), the metagenomes from the Global Ocean Sampling Expedition (GOS) (20, 21) were screened for AAPs using specific marker genes revealing a relative AAP abundance of 1% to 10% of total bacteria, which are much lower than the values reported by Lami et al. (13) from the oligotrophic southern Ocean (~25%). Reported abundances of AAPs range between <1% up to 25% (7, 13, 14, 22) and despite initial reports support the hypothesis proposed by Kolber (1) that these organisms would have an advantage in oligotrophic conditions recent reports suggest that in fact they thrive better in more eutrophic environments (22, 23). Many environmental characteristics such as association to particles, temperature, light attenuation, nutrient limitation or vulnerability to predation have been proposed as factors that influence the abundance of AAP bacteria, but their role is still not well understood. According to the study by Yutin et al. (19), the AAPs can be classified into 12 phylogroups (from A to L) through *puf*-operon synteny analysis and *pufM* phylogeny.

The primary aim of the current study was to estimate abundance and diversity of AAPs in a metagenome from an upwelling affected coastal bay in the Southwestern Atlantic Ocean (Arraial do Cabo, Brazil) (24). We developed a new *in silico* approach by using Profile Hiden Markov Models (pHMM) to screen for two core genes of anoxygenic photosynthesis (*pufM* and *pufL*), distinguishing them from the oxygenic photosynthesis genes (*psbA* – D1 and *psbD* - D2) in addition to analyzing the chlorophyllide reductase subunit X gene (*bch*X). The *puf* genes have been used as AAP markers in many studies (8, 18, 19). This approach was used to screen for AAPs in the Arraial do Cabo metagenomes and were compared to those from the GOS datasets. Our analysis had the goal to reveal a deeper insight into AAP abundance and phylogeny in coastal marine waters.

MATERIALS AND METHODS

Metagenomic datasets and sequence pre-processing

The two samples from Arraial do Cabo (sample P and E) seawater were used in addition to all 82 samples from the GOS dataset. The samples from GOS dataset were collect around the world in diverse environments, from inside lagoons to open ocean regions (20, 21). Sample P accounts for free-living bacteria (0.2 to <0.8 µm) and sample E to particle-associated bacteria and Eukaryotes (>0.8 µm) collected from an upwelling affected coastal bay as described by Cuadrat et al. (24). The samples were collect during the summer, on the upwelling season. Two datasets were generated: (i) All reads were translated into 6 frames using the TRANSEQ (from EMBOSS 6.1.0 package, default parameters) (25); (ii) The reads from Arraial do Cabo and from 10 samples from GOS (highest in AAP abundance) were individually assembled using CAP3 (default parameters) (26) and the Open Read Frames (ORFs) were extracted from the contigs and singlets using the METAGENMARK (version 2.8, default parameters) (27).

Estimates of AAP abundance in metagenomes by screening for *pufM*, *pufL* and *bchX* gene frequencies

The sequences from groups of orthologs from marker AAP genes; the homologous from this genes (oxygenic photosynthesis), and the constitutive gene *rec*A were obtained (both nucleotide and amino acid sequence in fasta format) from KEGG Orthology (KO): K08929 (*pufM*), K08928 (*pufL*), K11333 (*bchX*), K03553 (*recA*), K02703 (*psbA* - D1 protein), K02706 (*psbD* - D2 protein), K04037 (*bchL*), K11334 (*bchY*), K11335 (*bchZ*), and K04038 (*bchN*). These groups were aligned with MAFFT (v7.029b) (28) and each alignment was converted to Stockholm format using a custom PERL script. The program HMMBUILD (from HMMER 3.0 package) (29) was used to build a pHMM from each alignment and each pHMM was used to search (using the HMMSEARCH from HMMER 3.0 package, e-value cutoff 0.1) the metagenomic datasets (translated reads and ORFs). The hits were extracted by using the FASTACMD program (from BLAST 2.2.21 package (30)) and obtained sequences were used in HMMSCAN (from HMMER 3.0, e-value cutoff 0.1) and compared against all pHMM (concatenated and submitted to the HMMPRESS). When using this approach, it is possible to avoid the false detection of distant homologs (i.e., *puf*M and *psb*A – D1), and classifying the environmental sequences by using the best hits of HMMSCAN.

The number of “read equivalents” of each environmental ORF (obtained by the screening with the pHMMs) was calculated using the approach adapted from that described by Yutin and colleagues (19) and a script developed in RUBY 1.9.3 and BIORUBY (31).

In order to estimating the frequency of each marker gene, the number of their reads (or “read equivalents” in ORFs analysis) was normalized to the number of reads of the house keeping gene *rec*A (coding a critical DNA repair enzyme). This gene represents a single-copy gene in the genome of all bacteria (similar to the *puf* genes in the AAP genomes), has the same mean size as the AAP marker genes and thus can be used to estimate the number of bacterial genomes present in the analyzed metagenomic samples (19, 32, 33). The percent fraction of the AAP marker gene was calculated as follows:

Percent fraction of the AAP maker gene = (number of reads from the marker gene / number of reads of recA) * 100

Additionally, mean and standard deviation of *pufM*, *pufL*, and *bchX* gene abundance were calculated to estimate AAP numbers in each analyzed sample.

Confirming sequence annotation, calculating its specificity and sensitivity of our approach

In order to confirm the annotation of the environmental sequences (ORFs) obtained by our newly developed approach, the program BLASTX (30) and the RefSeq database (release 61) (from NCBI) were used. The best hits were manually verified and the percentage of false positives was calculated for each gene. The specificity for each gene was inferred by the mean of false positives:

Specificity (%) = 100 – mean of percentage from false positives for all marker genes

The sensitivity was estimated running the pipeline against the KEGG Orthology (KO) reference sequences and calculating the percentage of sequences obtained from each pHMM.

Phylogenetic analysis of *puf*M genes

The environmental *pufM* ORFs with more than 700 nucleotides (nt) were extracted and concatenated with reference sequences (from KO and from NCBI). The software MEGA 5.1 (34) was used to (i) translate the nucleotide sequences to amino acids; (ii) align the amino acid sequences using the MUSCLE (35) program (default parameters) and reverse translation of the alignment to nucleotide; and (iii) calculate the best evolutionary and substitution model for sequence alignment and subsequent phylogenetic analysis.

The obtained alignment was exported as FASTA format and trimmed by the TRIMAL 1.2 (36) to remove alignment positions with more gaps than nucleotides before conversion of the final alignment to the NEXUS format.

The program Mr Bayes 3.2 (37) was used on CIPRES GATEWAY (http://www.phylo.org/portal2/) (38) together with the Generalized Time-Reversible (GTR) model and gamma distribution, to generate a phylogenetic tree using Bayesian analysis. A total of two analyses were carried out with four parallel chains and 10 millions of executions.

Later, the phylogenetic tree and the alignment were imported into the ARB 5.5 (39) program to generate a local ARB database. Environmental ORFs smaller than 700 nt were added to the custom database by the quick add tool of ARB and used to construct the phylogenetic tree by using the parsimony method.

RESULTS

Metagenomic datasets and sequence pre-processing

The total number of sequence reads from Arraial do Cabo was 1,064,888 (595,534 from sample P [free-living] and 469,354 from sample E [particle-associated]) and from the GOS (20) dataset is 12,672,518. Table S1 shows the number of reads for each sample from GOS.

By assembling the reads using the CAP3 program, a total of 29,074 contigs and 269,587 singlets from sample P were generated. From sample E, 20,792 contigs and 396,371 singlets were generated. Using the METAGENMARK program, the total number of ORFs obtained was 409,111 for sample P and 451,722 for sample E. Table S2 shows the total numbers of sequences (contigs, singlets and ORFs) obtained from our GOS dataset analysis.

AAP abundance in the metagenomes estimated by the mean of the ratio of *pufM, pufL* and *bchX* gene in relation to the housekeeping *recA* gene

The screening of the environmental reads revealed a total of 860 and 248 hits obtained from sample P and E, respectively. Table 1 shows the number of hits obtained for each gene screened in the Arraial do Cabo samples.

Table 1: Number of hits obtained for each AAP marker gene and the *rec*A gene in Arraial do Cabo samples.

Samples	***pufM***	***pufL***	***bchX***	***recA***
Sample P	106	117	136	501
Sample E	2	12	15	197

The number of hits obtained in the 82 samples from the GOS dataset is given in table S3. Additionally, AAP abundance was calculated using the mean of the ratio of each marker gene (*pufM*, *pufL* and *bchX*) in relation to the housekeeping *recA* gene. Highest abundance of AAPs was found in sample P (23.88%±3.02 of free-living bacteria).

Figure 1 shows the percentage of AAPs in the 10 GOS samples which had the highest AAP % fraction in addition to our two samples from Arraial do Cabo (Sample P [free-living bacteria] and E [particle-associated bacteria]).

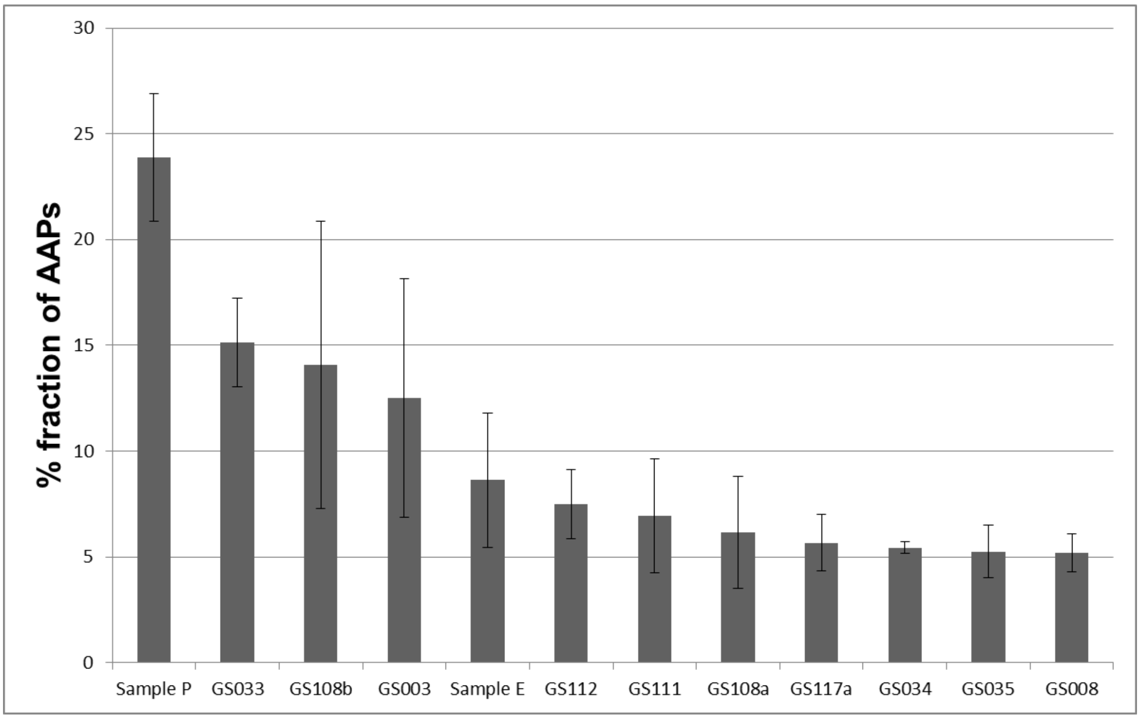


Figure 1: Percent fraction of AAPs in 10 samples of the GOS dataset with the highest AAP frequencies and our two samples from Arraial do Cabo (unassembled samples).

Sample P - Arraial do Cabo, GS033 - Punta Cormorant, Floreana Island (Hypersaline Lagoon), GS108b - Coccos Keeling, Inside Lagoon (>0.8 µm fraction), GS003 - Browns Bank, Gulf of Maine, Sample E – Arraial do Cabo (>0.8 µm fraction), GS112 - Indian Ocean, GS111 - Indian Ocean, GS108a - Coccos Keeling , Inside Lagoon, GS117a - St. Anne Island, Seychelles, GS034 - North Seamore Island (Galapagos), GS035 - Wolf Island (Galapagos), GS008 - Newport Harbor, RI.

After sequence assembly and ORF extraction from these 10 GOS samples (and our two samples from Arraial do Cabo), a total of 195, 155 and 200 putative *puf*M, *puf*L and *bch*X ORFs were found, respectively. Table 2 gives the number of ORFs for each gene obtained from the analyzed environmental samples.

Table 2: Number of putative *pufM*, *pufL*, *bchX* and *recA* ORFs obtained for each analyzed environmental sample.

Sample	***puf*M**	***puf*L**	***bch*X**	***rec*A**
GS111 - Indian Ocean	3	2	5	79
GS008 - Coccos Keeling , Inside Lagoon	7	5	6	137
GS112 - Indian ocean (454 FLX)	32	21	15	666
GS108b - Coccos Keeling , Inside Lagoon (>0.8 µm)	6	5	3	42
GS035 – Wolf Islands	6	3	8	174
GS003 - Browns Bank, Gulf of Maine	2	5	7	61
GS108a - Coccos Keeling , Inside Lagoon (>0.1 <0.8 µm)	5	2	1	64
GS117a - St. Anne Island, Seychelles	16	12	9	227
GS033 -Punta Cormorant, Floreana Island (Hypersaline Lagoon)	61	51	82	294

GS034 - North Seamore Island	9	8	6	178
Arraial do Cabo sample P	33	33	39	219
Arraial do Cabo sample E (>0.8 µm)	15	8	19	148
Total	195	155	200	2289

The % fraction of AAPs in ORFs was calculated using an approach adapted from a study of Yutin and colleagues (19), calculating the "read equivalents" of each gene in the detected ORFs.

Like for unassembled reads, in the ORFs, the highest numbers of AAPs were found in sample P from Arraial do Cabo (22.03% +-3.6). Figure 2 shows the % fraction of AAPs obtained from all 12 samples used for our analysis.

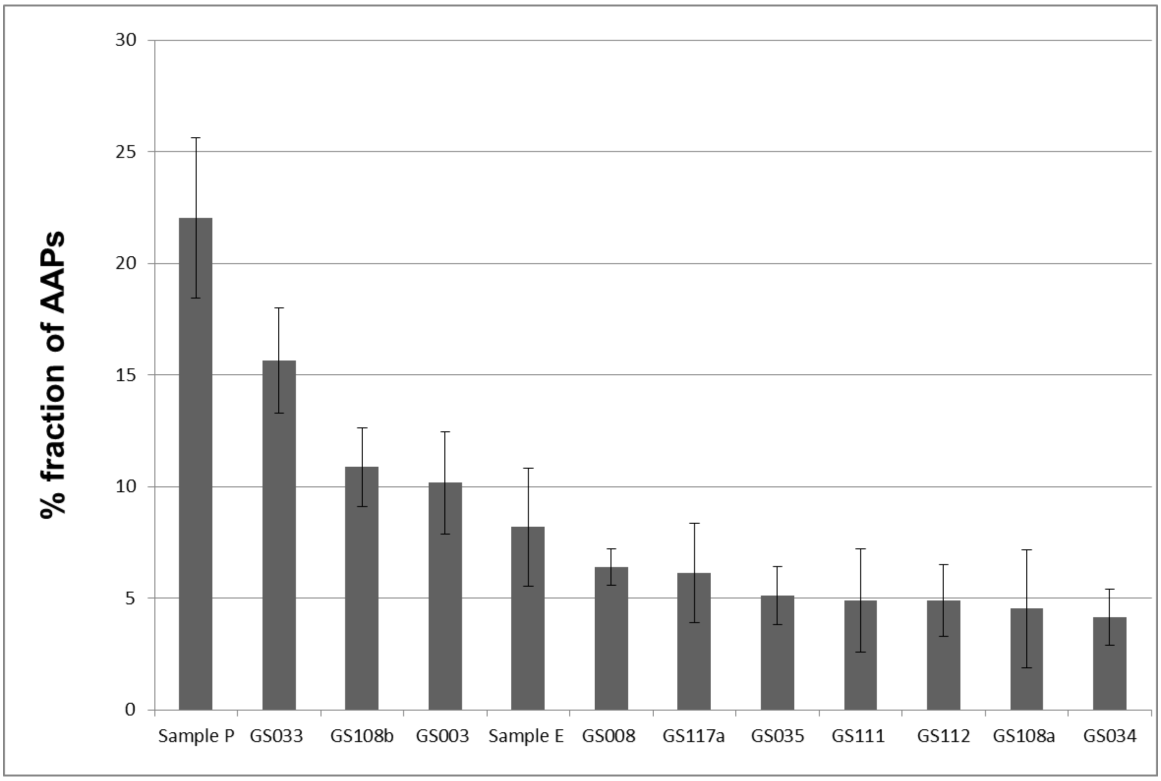


Figure 2: Percent fraction of AAPs in ORFs (calculated from "reads equivalents" of each ORF) from the selected 10 samples of the GOS dataset and our two samples from Arraial do Cabo. Sample P (free-living) - Arraial do Cabo, GS033 - Punta Cormorant, Floreana Island (Hypersaline Lagoon), GS108b - Coccos Keeling , Inside Lagoon (0.8 µm fraction), GS003 - Browns Bank, Gulf of Maine, Sample E (particle-associated) – Arraial do

Cabo (0.8 µm fraction), GS008 - Newport Harbor, RI, GS117a - St. Anne Island, Seychelles, GS035 - Wolf Island (Galapagos), GS112 - Indian Ocean, GS111 - Indian Ocean, GS108a - Coccos Keeling , Inside Lagoon, , GS034 - North Seamore Island (Galapagos).

Figure 3 gives the worldwide distribution of the analyzed GOS samples and our two samples from Arraial do Cabo.

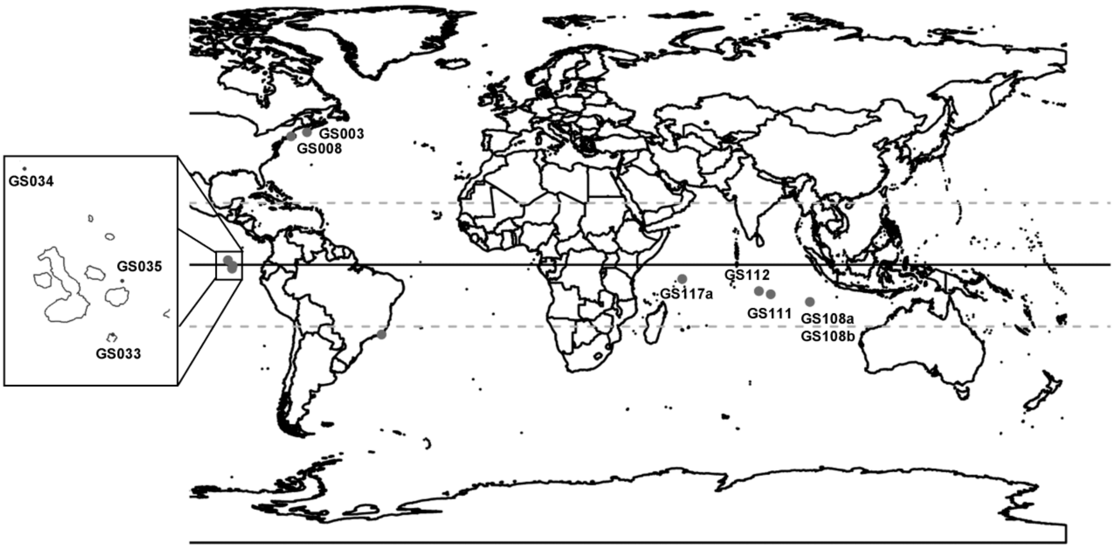


Figure 3: Worldwide distribution of sample sites of the 10 GOS data set with the highest AAP % fraction and the samples from Arraial do Cabo.

Sequence annotation, specificity and sensitivity of our analysis approach

In order to confirm the function of the obtained ORFs, all sequences were submitted to BLASTX and compared to the RefSeq protein database. All hits were manually checked.

From all obtained ORFs with the *pufM* gene pHMM, only 4 sequences showed a similarity with other genes and 3 showed no hits on RefSeq database. Sequences with no hit were submitted to BLASTX and compared to GenBank, whereby 2 of the 3 obtained sequences showed a similarity to environmental *pufM* genes. Therefore, a total of 5 sequences (four similar to different known genes and one without any hit) were classified as possible false positives (5/195 – 2.56%).

Analyzing all ORFs obtained from *pufL* gene pHMM, only 1.29 % (2 of 155) of sequences showed hits with different genes in the database and solely one sequence showed no hit. The last one was submitted to BLASTX and compared to GenBank and solely revealed a hit with hypothetical proteins. Thus, a total of three sequences were possible false positive (3/155 – 1.93%) and consequently removed from our phylogenetic analysis.

From all sequences obtained with *bchX* pHMM, 17 sequences revealed hits with other genes and solely two showed no hit with known sequences when using both databases (RefSeq and NCBI). Thus, in total, 19 possible false positive sequences were found (9.5%).

Finally, the specificity of our approach was calculated by using the mean of the 3 false positives genes, 4.66%, subtracting it from 100, which then resulted in a mean specificity of 95.34%.

Phylogenetic analysis of *pufM* genes

In order to classify our environmental *pufM* sequences, a Bayesian analysis (using Mr Bayes) was performed using the ORFs from our 10 analyzed GOS samples and our 2 samples from Arraial do Cabo, in addition to the 38 retrieved reference sequences from KEGG (KO) and NCBI. Figure 4 shows the phylogenetic tree.

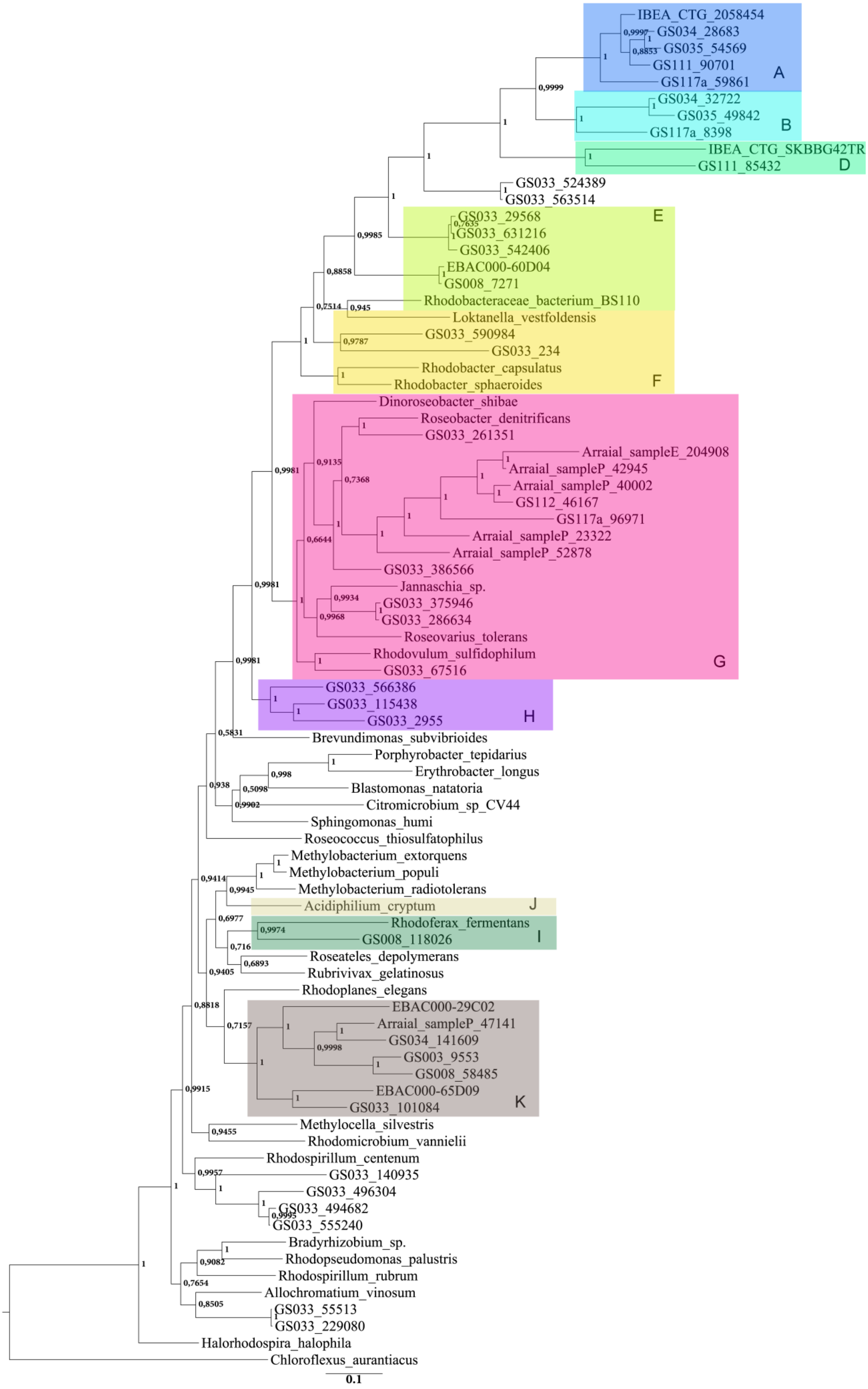

IBEA_CTG_2058454
0,9997
GS034_28683
1
0,8853
GS035_54569
1
GS111_90701
GS117a_59861
A
0,9999
GS034_32722
1
GS035_49842
1
GS117a_8398
B
1
IBEA_CTG_SKBBG42TR
1
GS111_85432
D
1
GS033_524389
1
GS033_563514
1
GS033_29568
0,7635
1
GS033_631216
GS033_542406
E
0,9985
EBAC000-60D04
1
GS008_7271
0,8858
Rhodobacteraceae_bacterium_BS110
0,7514
0,945
Loktanella_vestfoldensis
GS033_590984
0,9787
GS033_234
1
Rhodobacter_capsulatus
1
Rhodobacter_sphaeroides
F
Dinoroseobacter_shibae
Roseobacter_denitrificans
1
GS033_261351
0,9135
Arraial_sampleE_204908
1
Arraial_sampleP_42945
0,7368
1
Arraial_sampleP_40002
1
GS112_46167
1
1
GS117a_96971
0,9981
1
Arraial_sampleP_23322
1
Arraial_sampleP_52878
0,6644
GS033_386566
Jannaschia_sp.
0,9934
GS033_375946
1
0,9968
GS033_286634
0,9981
1
Roseovarius_tolerans
Rhodovulum_sulfidophilum
1
GS033_67516
G
0,9981
GS033_566386
1
GS033_115438
1
GS033_2955
H
0,5831
Brevundimonas_subvibrioides
Porphyrobacter_tepidarius
1
Erythrobacter_longus
0,998
0,938
0,5098
Blastomonas_natatoria
0,9902
Citromicrobium_sp_CV44
Sphingomonas_humi
Roseococcus_thiosulfatophilus
Methylobacterium_extorquens
1
Methylobacterium_populi
1
0,9414
0,9945
Methylobacterium_radiotolerans
Acidiphilium_cryptum
J
0,6977
Rhodoferax_fermentans
0,9974
GS008_118026
I
0,716
Roseateles_depolymerans
0,6893
0,9405
Rubrivivax_gelatinosus
Rhodoplanes_elegans
0,8818
EBAC000-29C02
0,7157
1
Arraial_sampleP_47141
1
GS034_141609
0,9998
GS003_9553
1
1
GS008_58485
EBAC000-65D09
1
GS033_101084
K
0,9915
Methylocella_silvestris
0,9455
Rhodomicrobium_vannielii
Rhodospirillum_centenum
0,9957
GS033_140935
1
1
GS033_496304
1
GS033_494682
0,9995
GS033_555240
1
Bradyrhizobium_sp.
0,9082
Rhodopseudomonas_palustris
1
Rhodospirillum_rubrum
0,7654
Allochromatium_vinosum
0,8505
GS033_55513
1
GS033_229080
Halorhodospira_halophila
Chloroflexus_aurantiacus
0.1

Fig 4: Phylogenetic tree of *pufM* genes from all 10 GOS samples, our 2 Arraial do Cabo samples, and all reference sequences retrieved from NCBI and KEEG (KO). Only sequences with more than 700 nucleotides were used. The tree was obtained by Bayesian analysis on Mr Bayes 3.2, using the GTR model and gamma distribution. Two executions were carried out with four parallel chains and 10 millions of executions. The highlighted clades refer to the different AAP phylogroups defined by Yutin *et al.* (19).

From the six environmental sequences obtained from Arraial do Cabo in our calculated tree, five (83.33%) were assigned to the *Roseobacter* clade (phylogroup G). The unique sequence is affiliated to phylogroup K.

Additionally, all short sequences were added to the Bayesian tree, using the ARB program ("quick add by parsimony" method) (39), in order to relate them to the retrieved phylotypes. The resulting tree is given in the supplementary material (figure S1). Figure 5 shows the relative abundance of each phylogroup in all 12 analyzed metagenome samples.

When adding all short sequences to the tree, solely one sequence from sample P (free-living) could be classified as phylogroup A (without representative cultivated organism), two sequences of sample P and one of sample E (particle-associated) were grouped into phylogroup F (*Rhodobacter* clade), 22 sequences of sample P and 10 of sample E were classified as phylogroup G (*Roseobacter* clade), 4 sequences of sample P and 4 of sample E were assigned to group H (uncultured), and just 3 sequences of sample P fell into phylogroup K (gammaproteobacterial clade).

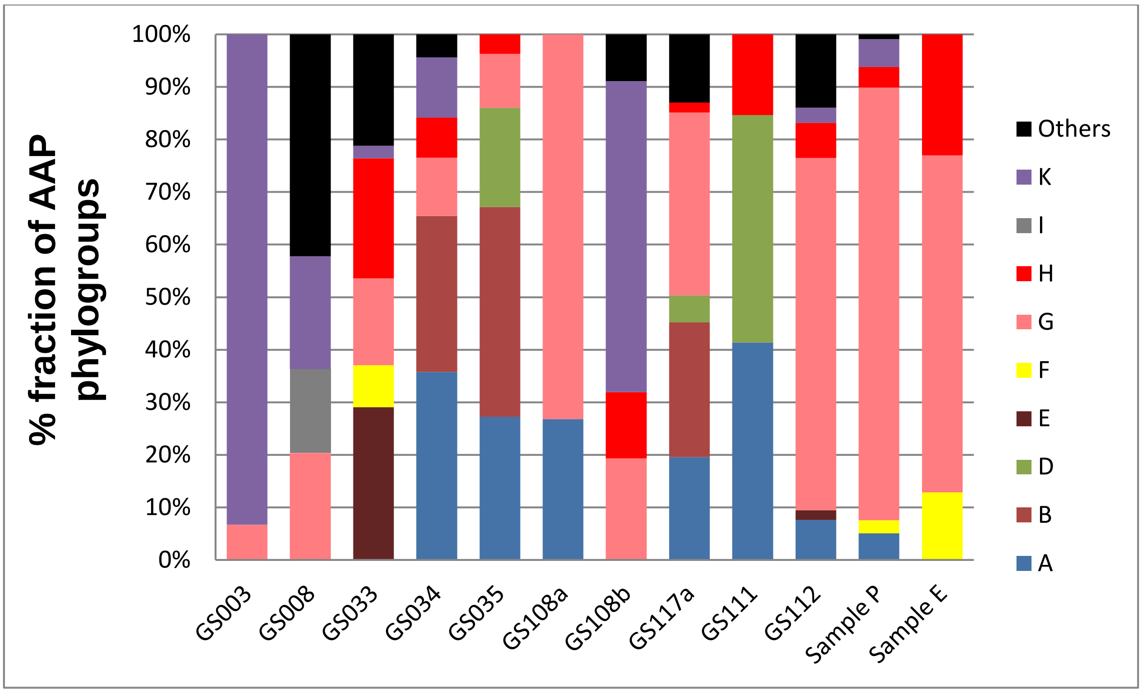


Figure 5: Relative abundance of each phylogroup retrieved from the different analyzed metagenomic samples. Number of read equivalents for each obtained ORF was counted and percentages were calculated by using the classification of the phylogenetic tree of figure S1.

Discussion

Our newly developed pipeline enabled us to rapidly screen a total of 12,672,518 reads from 82 GOS samples and 1,064,888 reads from our 2 Arraial do Cabo metagenomes. The pipeline was very sensitive and highly specific (95.34% of specificity and 100% of sensitivity), and proved to run on a simple desktop computer, even for such a large scale study including many metagenomic samples. The abundance of AAPs was evaluated on unassembled data (raw reads) allowing to determinate environments with the highest AAP fraction for targeted assembly selection, ORF extraction and AAP screening (using a similar approach of the unassembled screening). To our knowledge, this is the first study estimating AAP abundance on unassembled metagenomic samples, since the study of Yutin *et al.* (19) was performed for assembled samples. Moreover, Yutin and colleagues used the cross assembly (20) (contigs were obtained from all concatenated samples) which significantly increases the likelihood to generate chimeric sequences (40). Major advantages of using raw reads are (i) preventing the assembly step for all samples, which is computationally expensive and slow and (ii) avoiding chimeric sequences obtained by metagenomic assembly (41). However, due to the limitation of read size assembly, the assemble step is required for both the phylogenetic and *puf* operon synteny analyses.

The present study adds to the in depth analysis of the GOS samples (20) to the updated dataset, introducing an additional of 38 new samples (the initial version used by Yutin included 44 samples (19), and the current version contains 82 samples), plus our two samples from Arraial do Cabo. The main reason to study AAPs in the Arraial do Cabo samples is the high abundance of the genus *Roseobacter* (15% of identified genera on sample P (free-living bacteria) and a number of other known AAP genera (e.g., *Jannaschia* and *Dinoroseobacter*) found in the previous exploratory work performed to characterize these samples (24).

To minimize the chimeric sequence formation, all samples were assembled individually (our 2 samples from Arraial do Cabo and the 10 GOS samples with the highest number of AAP reads), using the CAP3 program with specific default parameters (26). Results of the AAP screening of unassembled samples were compared to those of assembled samples, showing a good consistency between the obtained results. Another advantage of the individual sample assembly is the possibility of further phylogeographic sequence analysis.

In addition to the 2 samples from Arraial do Cabo, 82 metagenomes from the GOS dataset were screened (unassembled reads). Interestingly, AAPs were particularly abundant in the metagenomes (figure 1) from 8 tropical sites (80%) close to the equator (figure 3). These results can be explained by the fact that all tropical sites are characterized by a high light availability and water transparency allowing for light harvesting even at greater depth (>100 m) and consequently positive AAP growth throughout an extended part of the water column compared to other marine sites (42).

By using the raw and assembled reads, sample P (free-living bacteria) from Arraial do Cabo revealed the highest % fraction of AAPs (up to 23.88% +- 3.02%). This result shows that Arraial do Cabo can be regarded as a marine environment with one of the highest so far known AAP abundance worldwide. However, although the GOS and Arraial do Cabo samples were fractionated, we need calculate the mean of the size-fractionated samples in order to compare with samples from other studies. Thus, the total abundance of AAPs in the samples of Arraial do Cabo was 16.07%.

The high AAP abundance in Arraial do Cabo is comparable to the most abundant in the Waidner *et al.* study, in turbid waters from estuaries (from 12% to 17% of the community), (43) and higher than Cottrell and Kirchman study in the temperate and polar Artic Ocean (from 5% to 8%) (7), and Ritchie *et al*., in coastal regions of the Pacific Ocean (1.2%, on average) (44). This fact may be explained by the fact of Arraial do Cabo was being affected by upwelling (24), once this phenomenon can induce a phytoplankton bloom, and many other studies showed high correlation between blooms and high abundance of AAPs from *Roseobacter* clade (45)(46)

The marine environment with higher AAP fraction described on literature was the very oligotrophic Southern Pacific Ocean AAPs (24%) (13). However, this number can be overestimated by the method used in this study (microscope cell count) that does not subtract picophytoplankton. On the other hand, in this study, we used metagenomic approach, targeting genes only from anoxygenic photosynthesis.

The sample with the second highest AAP abundance was GS33 (Browns Bank, Gulf of Maine), an anoxic hyper saline lagoon (63.4 PSU, dissolved oxygen, 0.06 mg l-1) with 15.64% +- 2.36 (20). This sample was discussed separately by Yutin *at al*. (19), because it is likely that due to the anoxic environment the detected anoxygenic phototrophic bacteria are anaerobic photoautotrophs and not AAPs. However, our phylogenetic results (corroborating the results from Yutin´s work (19)) reveal many *pufM* sequences clustering within the phylogroup G (16.49%% of total read equivalents), suggesting the presence of an active AAP community also in this environment.

It is important to note that the comparisons between different samples may be biased by differences in methods used for sampling, filtration, DNA extraction and sequencing. In addition, the timing of collection should be considered since many aquatic systems are characterized by seasonal variance in their AAP community (47)(48). Furthermore, Ferrera *et al.* (48) showed a high AAP abundance in summer but a low richness compared to the winter situation, corroborating many previous studies (49)(50)(51). Other studies have revealed many different environmental variables such as light, nutrient availability, temperature, vulnerability to predation and Chl *a* concentration influencing AAP abundance and diversity (50)(51)(52). However, Ferrera *et al.* (48) showed a tight correlation between day length and AAP abundance, corroborating data from AAP culture studies which suggest that light enhances organic carbon utilization efficiency, energy cycling and hence growth (53)(54).

The sample from Arraial do Cabo was collected in summer around noon when light irradiation was highest. Thus for future, seasonal studies which aim to better understand the variation of AAP light availability and the lack of organic carbon may be an important factor to explain the extraordinary high abundance of AAP in this environment.

Our AAP marker genes were the *pufM*, *pufL* and *bchX* genes, and thus were also used in many other studies using PCR, qPCR (mainly *pufM*) (15) or *in silico* analyses (16)(17)(18)(19). The *pufM* environmental sequences obtained in all of our 12 samples (>700 nucleotides) were used together with reference sequences in a Bayesian analysis (figure 4). To the obtained phylogenetic tree, small sequences were added by the ARB parsimony method (figure S1). The topology of the resulting tree corroborates previous studies, e.g. that of Yutin *et al*, 2007 (19) and of Lehours *et al*., 2010 (55), and related the reference sequences to specific phylogroups as expected. Further, these results were confirming by analyzing the distribution and phylogenetic relatedness of the *puf* operon, as discussed by Yutin *et al*. (19).

Our results show the dominance of phylogroup G (*Roseobacter* clade) in both Arraial do Cabo samples, with 82.36% (sample P[free-living bacteria]) and 64.05% (sample E [particle-associated bacteria]) of total AAP in this environment. Our results contrast with the Ferrera *et al.* study in the coastal Mediterranean (48), in which alphaproteobacterial groups E, F and G only outnumbered the gammaproteobacterial groups during winter (when nutrient concentrations were higher). However, our samples from Arraial do Cabo were collected in the summer at extremely low nutrient concentrations.

In our samples, *Roseobacter* clade was the most ubiquitous, present in 11 of 12 of the analyzed samples, corroborating the results of other studies (56) including the previous GOS study of Yutin et al. (19). However, in the analyzed GOS samples, we generally detected a higher abundance of the phylogroup G (*Roseobacter* clade) in samples from the Indian Ocean (less the GS111) when comparing with samples from the Eastern Pacific Ocean (Galapagos) or the Atlantic West Coast (USA).

AAP life styles and phylogroups

The relative abundance of the free-living AAP (samples P and GS108a) was higher than of particle-associated bacteria (samples E and GS108b), suggesting that the phylogroup G refers mainly to a free living lifestyle.

In addition, phylogroup A may also represent mainly the free living lifestyle since this group is absent samples E and GS108b (>0.8 µm). Interestingly, this group is also absent in the anoxic sample (GS033), suggesting a dependency of this group on oxygen availability.

In contrast, phylogroup H is more abundant in the >0.8 µm size fraction (samples E and GS108b) and the anoxic GS033 sample, but phylogroup E was exclusively found in the anoxic GS033 sample and at very low abundance in the GS112 sample. The phylogroup F (*Rhodobacter* clade) was found in both size fractions of Arraial do Cabo (although with a higher abundance in sample E), but in the GOS samples it was exclusively present in the anoxic GS033 sample.

The correlation between AAP abundance of the assigned groups in the anoxic GS033 sample, but also in the >0.8 µm fraction of samples E and GS108b can be explained by the formation of potentially anoxic microenvironments, e.g. on macroscopic organic aggregates even in an oxygenated water column. Such aggregates can be normally trapped on the 0.8 µm membranes. The specific AAP groups are abundant in these samples and seem to be well adapted to harvest light on the organic matter rich particles which also provide an excellent organic substrate for these photoheterotrophic bacteria (57)(58)(59).

In the study performed by Yutin and colleagues (2007) (19), *Rhodoplanes* (alfaproteobacteria) and *Rosealetes* (betaproteobacteria) genera clustered together. However, in our work, the *Rosealetes* clustered with other betaproteobacteria genera: *Rubrivivax* and *Rhodoferax,* separating them well from alfaproteobacteria. In our study, just a single environmental sequence was affiliated to this clade (GS008_118026).

Alphaproteobacteria of the genus *Rhodoplanes* clustered together with the phylogroup K (gamma-proteobacteria). This fact may be explained by possible horizontal gene transfer (HGT) of the photosynthetic apparatus, as proposed by several previous studies [46] [47] [48]. Some AAP strains (as *Roseobacter litoralis* Och 149), however, contain a plasmid with all genes from the anoxygenic photosynthesis (60)(61) and the presence of phage DNA which is directly associated with the photosynthesis operons (62)(63) and which may corroborate the hypothesis of HGT among AAPs.

Noteworthy, similarly to the study of Yutin et al, (19), no α-4 subclass AAP was detected in our extensive phylogenetic analysis. This group, normally present in diverse marine environments (64), forms a separate clade (*Erythrobacter*, *Blastomonas, Sphingomonas* and *Porphyrobacter*), without any known environmental sequence.

Conclusions

The present study presents a new approach for the fast screening of anoxygenic photosynthesis genes and to evaluate abundance and diversity of AAPs in environmental samples (raw and/or assembled reads).

Our results obtained from 84 unassembled and 12 assembled metagenome samples reveal that our newly developed approach leads to consistent results for both types of datasets. When using the unassembled samples it was possible to even screen large datasets and to select samples with the highest AAP abundance for further phylogenetic analysis. Free-living bacteria (sample P) from Arraial do Cabo showed an extremely high AAP abundance, which was even higher than of any other GOS sample analyzed and similar to another very oligotrophic marine environment studied by Lami *et al.* [6].

The environmental *pufM* ORFs obtained from the assembled samples were subjected to a phylogenetic analysis which enabled us to classify specific phylogroups of AAPs present in these environments. Thereby, *Roseobacter* clade turned out to be the most dominant AAP group in the Arraial do Cabo environment and the most ubiquitous AAP group of all 12 assembled metagenome samples.

These exciting results encourage us to perform a more intense time series study in the Arraial do Cabo region in the near future to investigating the dynamics of AAP at different locations and seasons and to better understand the ecological role of these unique bacteria for biogeochemical and energy cycling in such environments.

References

1. Kolber ZS (2001) Contribution of Aerobic Photoheterotrophic Bacteria to the Carbon Cycle in the Ocean. Science 292: 2492–2495. doi:10.1126/science.1059707.

2. Denner EBM, et al. 2002. Erythrobacter citreus sp. nov., a yellow pigmented bacterium that lacks bacteriochlorophyll a, isolated from the western Mediterranean Sea. Int. J. Syst. Evol. Microbiol. **52**:1655–1661.

3**.** Fuchs BM, et al. 2007. Characterization of a marine gammaproteobacterium capable of aerobic anoxygenic photosynthesis. Proc. Natl. Acad. Sci. U. S. A. **104**:2891–2896.

4**.** Gich F, Overmann J. 2006. Sandarakinorhabdus limnophila gen. nov., sp. nov., a novel bacteriochlorophyll a-containing, obligately aerobic bacterium isolated from freshwater lakes. Int. J. Syst. Evol. Microbiol. **56**:847–854.

5. Csotonyi JT, Swiderski J, Stackebrandt E, Yurkov V. 2010. A new environment for aerobic anoxygenic phototrophic bacteria: biological soil crusts. Adv. Exp. Med. Biol. 675:3-14.

6. Labrenz M, Lawson PA, Tindal BJ, Collins MD, Hirsch P. 2005. Rosei-salinus antarcticus gen. nov., sp. nov., a novel aerobic bacteriochlorophyll a-producing alpha-proteobacterium isolated from hypersaline Ekho Lake, Antarctica. Int. J. Syst. Evol. Microbiol. 55:41- 47.

7. Cottrell, M.T. & Kirchman, D.L., 2009, Photoheterotrophic microbes in the Arctic Ocean in summer and winter, *Applied and environmental microbiology*, 75(15), pp. 4958-66.

8. Ferrera I, Borrego CM, Salazar G, Gasol JM. (2013) 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. Environmental Microbiology doi: 10.1111/1462-2920.12278

9. Koblízek, M., 2011, Role of photoheterotrophic bacteria in the marine carbon cycle, *Microbial Carbon Pump in the Ocean. Jiao, N., Azam, F., and Sanders, S.(eds). Washington, DC, USA: Science/AAAS*, pp. 49-51.

10. Goericke R (2002) Bacteriochlorophyll a in the ocean: Is anoxygenic bacterial photosynthesis important? Limnology and oceanography 47: 290–295.

11. Kolber ZS, Van Dover CL, Niederman RA, Falkowski PG (2000) Bacterial photosynthesis in surface waters of the open ocean. Nature 407: 177–179. doi:10.1038/35025044.

12. Cottrell MT, Mannino A, Kirchman DL (2006) Aerobic Anoxygenic Phototrophic Bacteria in the Mid-Atlantic Bight and the North Pacific Gyre. Applied and Environmental Microbiology 72: 557–564. doi:10.1128/AEM.72.1.557-564.2006.

13. Lami R, Cottrell MT, Ras J, Ulloa O, Obernosterer I, et al. (2007) High Abundances of Aerobic Anoxygenic Photosynthetic Bacteria in the South Pacific Ocean. Applied and Environmental Microbiology 73: 4198–4205. doi:10.1128/AEM.02652-06

14. Schwalbach MS, Fuhrman JA (2005) Wide-ranging abundances of aerobic anoxygenic phototrophic bacteria in the world ocean revealed by epifluorescence microscopy and quantitative PCR. Limnology and oceanography 50: 620–628.

15. Waidner LA, Kirchman DL (2008) Diversity and Distribution of Ecotypes of the Aerobic Anoxygenic Phototrophy Gene pufM in the Delaware Estuary. Applied and Environmental Microbiology 74: 4012–4021. doi:10.1128/AEM.02324-07

16. Béjà O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, et al. (2002) Unsuspected diversity among marine aerobic anoxygenic phototrophs. Nature 415: 630–633.

17. Oz A, Sabehi G, Koblizek M, Massana R, Beja O (2005) Roseobacter-Like Bacteria in Red and Mediterranean Sea Aerobic Anoxygenic Photosynthetic Populations. Applied and Environmental Microbiology 71: 344–353. doi:10.1128/AEM.71.1.344-353.2005.

18. Waidner LA, Kirchman DL (2005) Aerobic anoxygenic photosynthesis genes and operons in uncultured bacteria in the Delaware River. Environ Microbiol 7: 1896–1908. doi:10.1111/j.1462-2920.2005.00883.x.

19. Yutin N, Suzuki MT, Teeling H, Weber M, Venter JC, et al. (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. Environmental Microbiology 9: 1464–1475. doi:10.1111/j.1462-2920.2007.01265.x.

20. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biology 5: e77. doi:10.1371/journal.pbio.0050077.

21. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. PLoS Biology 5: e16. doi:10.1371/journal.pbio.0050016

22. Hojerová E, Mašín M, Brunet C, Ferrera I, Gasol JM, et al. (2011) Distribution and Growth of Aerobic Anoxygenic Phototrophs in the Mediterranean Sea: AAP bacteria in the Mediterranean Sea. Environmental Microbiology 13: 2717–2725. doi:10.1111/j.1462-2920.2011.02540.x.

23. Cottrell, M.T., Ras, J. & Kirchman, D.L., 2010, Bacteriochlorophyll and community structure of aerobic anoxygenic phototrophic bacteria in a particle-rich estuary, *The ISME journal,* 4(7), pp. 945-54.

24. Cuadrat, RC, Cury, JC, Dávila, MR. (2014). Microbial, metabolic diversity and genes of type I PKS and NRPS revealed by metagenomic analysis of Brazilian coastal seawater. Plos One, In Press.

25. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276–277.

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

27. Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Research 38: e132–e132. doi:10.1093/nar/gkq275

28. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059–3066.

29. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Computational Biology 7: e1002195. doi:10.1371/journal.pcbi.1002195.

30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410. doi:10.1016/S0022-2836(05)80360-2

31. Goto N, Prins P, Nakao M, Bonnal R, Aerts J, et al. (2010) BioRuby: bioinformatics software for the Ruby programming language. Bioinformatics 26: 2617–2619. doi:10.1093/bioinformatics/btq475.

32. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304. doi:10.1126/science.1093857.

33. Howard, E.C., Henriksen, J.R., Buchan, A., Reisch, C.R., Burgmann, H., Welsh, R., et al. (2006) Bacterial taxa that limit sulfur flux from the ocean. Science 314: 649–652

34. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731–2739. doi:10.1093/molbev/msr121.

35. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797. doi:10.1093/nar/gkh340.

36. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973. doi:10.1093/bioinformatics/btp348.

37. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574. doi:10.1093/bioinformatics/btg180.

38. Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Gateway Computing Environments Workshop (GCE), 2010. IEEE. pp. 1–8.

39. Ludwig W (2004) ARB: a software environment for sequence data. Nucleic Acids Research 32: 1363–1371. doi:10.1093/nar/gkh293.

40. Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: Read Length Matters. Applied and Environmental Microbiology 74: 1453–1463. doi:10.1128/AEM.02181-07.

41. Pignatelli M, Moya A (2011) Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data. PLoS ONE 6: e19984. doi:10.1371/journal.pone.0019984.

42. Hauruseu D, Koblizek M (2012) Influence of Light on Carbon Utilization in Aerobic Anoxygenic Phototrophs. Applied and Environmental Microbiology 78: 7414–7419. doi:10.1128/AEM.01747-12.

43. Waidner LA, Kirchman DL (2007) Aerobic Anoxygenic Phototrophic Bacteria Attached to Particles in Turbid Waters of the Delaware and Chesapeake Estuaries. Applied and Environmental Microbiology 73: 3936–3944. doi:10.1128/AEM.00592-07

44. Ritchie AE, Johnson ZI (2012) Abundance and Genetic Diversity of Aerobic Anoxygenic Phototrophic Bacteria of Coastal Regions of the Pacific Ocean. Applied and Environmental Microbiology 78: 2858–2866. doi:10.1128/AEM.06268-11.

45. Wemheuer B, Gullert S, Billerbeck S, Giebel H-A, Voget S, et al. (2014) Impact of a phytoplankton bloom on the diversity of the active bacterial community in the southern North Sea as revealed by metatranscriptomic approaches. FEMS Microbiol Ecol 87: 378–389. doi:10.1111/1574-6941.12230.

46. Alonso-Gutiérrez J, Lekunberri I, Teira E, Gasol JM, Figueras A, et al. (2009) Bacterioplankton composition of the coastal upwelling system of “Ría de Vigo”, NW Spain. FEMS Microbiology Ecology 70: 493–505. doi:10.1111/j.1574-6941.2009.00766.x.

47. Cottrell MT, Kirchman DL (2009) Photoheterotrophic Microbes in the Arctic Ocean in Summer and Winter. Applied and Environmental Microbiology 75: 4958–4966. doi:10.1128/AEM.00117-09.

48. Ferrera I, Borrego CM, Salazar G, Gasol JM (2013) 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: Marine AAP dynamics in coastal sea. Environmental Microbiology: n/a–n/a. doi:10.1111/1462-2920.12278.

49. Zhang Y, Jiao N (2007) Dynamics of aerobic anoxygenic phototrophic bacteria in the East China Sea: AAPB in the East China Sea. FEMS Microbiology Ecology 61: 459–469. doi:10.1111/j.1574-6941.2007.00355.x.

50. Masín M, Zdun A, Sto´n-Egiert J, Nausch M, Labrenz M, Moulisová V, et al. (2006) Seasonal changes and diversity of aerobic anoxygenic phototrophs in the Baltic Sea. Aquat Microb Ecol.:45: 247–254.

51. Lamy D, De Carvalho-Maalouf P, Cottrell MT, Lami R, Catala P, et al. (2011) Seasonal dynamics of aerobic anoxygenic phototrophs in a Mediterranean coastal lagoon. Aquat Microb Ecol 62: 153–163.

52. Hojerová E, Mašín M, Brunet C, Ferrera I, Gasol JM, et al. (2011) Distribution and Growth of Aerobic Anoxygenic Phototrophs in the Mediterranean Sea: AAP bacteria in the Mediterranean Sea. Environmental Microbiology 13: 2717–2725. doi:10.1111/j.1462-2920.2011.02540.x.

53. Hauruseu D, Koblizek M (2012) Influence of Light on Carbon Utilization in Aerobic Anoxygenic Phototrophs. Applied and Environmental Microbiology 78: 7414–7419. doi:10.1128/AEM.01747-12.

54 - Spring, S. & 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 microbiology*, 13(1), p. 117

55. Lehours A-C, Cottrell MT, Dahan O, Kirchman DL, Jeanthon C (2010) Summer distribution and diversity of aerobic anoxygenic phototrophic bacteria in the Mediterranean Sea in relation to environmental variables. FEMS Microbiol Ecol 74. doi:10.1111/j.1574-6941.2010.00954.x.

56. Buchan A, Gonzalez JM, Moran MA (2005) Overview of the Marine Roseobacter Lineage. Applied and Environmental Microbiology 71: 5665–5677. doi:10.1128/AEM.71.10.5665-5677.2005.

57. Cottrell MT, Ras J, Kirchman DL (2010) Bacteriochlorophyll and community structure of aerobic anoxygenic phototrophic bacteria in a particle-rich estuary. The ISME journal 4: 945–954.

58. Mašín, M., Cuperová, Z., Hojerová, E., Salka, I., Grossart, H.P., Koblížek, M.
(2012) Distribution of aerobic anoxygenic phototrophic bacteria in glacial lakes of
northern Europe. Aquat. Microb. Ecol. 66: 77–86, doi: 10.3354/ame01558

59. Salka, I., Cuperová, Z., Mašín, M., Koblížek, M., Grossart, H.-P. (2011) Rhodoferax-related pufM gene cluster dominates the aerobic anoxygenic phototrophic communities in German freshwater lakes. Environ. Microbiol. 13(11), 2865–2875, doi:10.1111/j.1462-2920.2011.02562.x

60. Petersen, J., Brinkmann, H., & Pradella, S. (2009). Diversity and evolution of rep ABC type plasmids in Rhodobacterales. Environmental Microbiology, 11, 2627–2638.

61. Kalhoefer, D., Thole, S., Voget, S., Lehmann, R., Liesegang, H., Wollher, A., et al. (2011).Comparative genome analysis and genome-guided physiological analysis of Roseobacter litoralis. BMC Genomics, 12, 324.

62. Jiao, N., Zhang, R., & Zheng, Q. (2010). Coexistence of two different photosynthetic operons in Citromicrobium bathyomarinum JL354 as revealed by whole-genome sequencing. Journal of Bacteriology, 192(4), 1169–1170.

63. Yurkov V, Hughes E (2013) Chapter Eleven - Genes Associated with the Peculiar Phenotypes of the Aerobic Anoxygenic Phototrophs. In: J. Thomas Beatty, editor. Advances in Botanical Research. Academic Press, Vol. Volume 66. pp. 327–358.

64. Yurkov V, Csotonyi J (2009) New Light on Aerobic Anoxygenic Phototrophs. In: Hunter CN, Daldal F, Thurnauer M, Beatty JT, editors. The Purple Phototrophic Bacteria. Advances in Photosynthesis and Respiration. Springer Netherlands, Vol. 28. pp. 31–55. Available: http://dx.doi.org/10.1007/978-1-4020-8815-5_3.

5 DISCUSSÃO

O presente trabalho mostra a exploração da diversidade taxonômica e metabólica de uma amostra fracionada do ambiente costeiro marinho de Arraial do Cabo, no Estado do Rio de Janeiro (Praia dos Anjos). Esta região vem sendo estudada por nosso grupo, primeiro, através de amplificação e sequenciamento de genes de RNAr (Cury *et al*., 2011) e agora através de pirosequenciamento do DNA total ambiental (*whole shotgun*).

No estudo realizado por Cury e colaboradores, o ambiente costeiro da Praia dos Anjos (afetada por atividades antropogênicas e despejo de esgoto ocasional) foi comparado com uma região de mar aberto, afetada diretamente pelo fenômeno da ressurgência, mostrando a diversidade microbiana dos dois ambientes.

Neste estudo, os parâmetros físico-químicos da amostra foram medidos no local, e a baixa concentração de amônia mostrou que o ambiente costeiro não estava sobre impacto direto de despejo de esgoto no momento da coleta, uma vez que o estudo de Coelho-Souza e colaboradores (Coelho-Souza *et al.,* 2013) na região demonstrou que, quando havia despejo visível na Praia dos Anjos, a concentração de amônia era acima de 2 µM. A baixa concentração de amônia (abaixo do limite detectável) indica ainda, a possibilidade de o ambiente estar sendo afetado pelo fenômeno da ressurgência, provavelmente tendo sido coletada após o *bloom* do fitoplanctôn, que utiliza amônia em seu metabolismo, causando depleção deste nutriente durante o *bloom*. A presença de nitrato em concentração de 15,51 µM, similar a encontrada por Cury e colaboradores em amostra afetada pela ressurgência, também corrobora a hipótese de que a amostra analizada neste estudo estava sendo afetada pelo fenômeno.

Adicionalmente, Albuquerque e colaboradores (2014) demonstraram que em Janeiro de 2012, mês em que a amostra do presente estudo foi coletada, ocorreu a subida de águas frias para a zona eufótica, corroborando a hipótese de ressurgência na ocasião.

A amostra foi fracionada durante a filtragem, com objetivo de separar Eucariotos de Procariotos, e para tal foram utilizados dois tipos de membrana filtrante, a primeira, de poros com 0,8 µm de diâmetro, sendo a amostra nomeada a partir desse momento como amostra E, e a segunda com poros de 0,22 µm, sendo nomeada amostra P. Este procedimento foi adotado com o objetivo de enriquecer a amostra P, deixando o mínimo de eucariotos na amostra, para realização de estudos futuros envolvendo clonagem do DNA da mesma em vetores grandes (fosmídeos), triagens funcionais em busca de novos genes e expressão heteróloga dos mesmos. Separando a comunidade de eucariotos desta amostra, torna-se mais provável de se obter a clonagem e principalmente a expressão de genes de interesse, visto que genes de eucariotos são mais dificilmente expressos, envolvendo uma série de fatores de transcrição que podem estar fisicamente distantes dos genes regulados, impedindo assim a expressão dos mesmos em hospedeiros procarióticos (Agnan *et al*., 1997).

O sequenciamento das amostras foi realizado separadamente, cada um ocupando metade de uma placa de sequenciamento, gerando um total de 1.064.888 sequencias (*reads*). O tamanho médio das sequências foi de 588 e 595 pares de base (pb), na amostra P e E, respectivamente. Este resultado é compatível com o tamanho de reads gerados pelo método de sequenciamento Sanger, porém com um custo muito inferior e uma vazão muito maior. Este resultado foi possível graças à nova química utilizada pela ROCHE no sequenciador 454 (FLX+), sendo vantajoso para estudos de metagenômica, pois permite a anotação dos reads por similaridade sem montagem prévia, evitando risco de formações de sequencias quiméricas (Wommack *et al*., 2008). A maior parte dos metagenomas até hoje foi sequenciada utilizando a química antiga do 454, ou as novas gerações de Illumina (My-Seq ou Hi-Seq), com *reads* em torno de 300 pares de base. O maior tamanho de read obtido pode ser um dos fatores que explicam um maior percentual de sequências anotadas por similaridade no MG-RAST (91.7% na amostra P) do que em outros estudos de metagenômica, como o de Trindade-Silva e colaboradores (2012). Porém, vale ressaltar que o sequenciamento por Illumina de nova geração possui a vantagem de ter menor custo e mais alta vazão em relação ao 454.

As sequências obtidas das duas amostras foram submetidas à análise local de similaridade com o programa BLASTN do pacote BLAST, usando como referência o banco de sequências nucleotídicas GenBank (NT) do NCBI. A escolha do algoritmo de comparação de sequencias de nucleotídeos (BLASTN) em vez de proteínas (BLASTP) se deve ao fato de que para análises taxonômicas, as sequências nucleicas são mais informativas, por serem menos conservadas que as sequências proteicas. Adicionalmente, as sequências foram submetidas para análise na versão web do MG-RAST, que revelou valores de alpha-diversidade para amostra P (240,06) e amostra E (464,02). A possível explicação para a maior diversidade na amostra E reside no fato de que esta amostra sofreu menor seleção de tamanho de células (maior que 0,8 µm) do que a amostra P (entre 0,22 e 0,8 µm).

Os resultados do BLAST foram carregados no programa MEGAN, e analisados através de seu algoritmo de Último Ancestral em Comum (LCA). Foi possível observar a amostra P com o mínimo de eucariotos e ampla dominância de procariotos (até 96,37% do total de 159.962 *reads* anotados), porém como esperado, uma grande quantidade de procariotos (82,22% de 59.483 reads anotados) ficou também retida na membrana da amostra E. Este fato pode ser explicado pelas inúmeras simbioses encontradas entre cianobactérias e algas por exemplo, pela possível formação de colônias entre as bactérias, a associação com partículas orgânicas e ainda pela possibilidade de células eucarióticas bloquearem os poros antes da passagem das células bacterianas, que são menores, retendo parte delas na membrana. Porém, uma ampla parte do total de sequencias geradas na amostra E não foi anotada, e isso pode se dever ao fato de que existem muito mais sequências de procariotos do que eucariotos no banco de sequencias utilizado (GenBank), que na versão atual possui ~8.5 bilhões de pb de Bacteria e ~2.5 bilhões de pb de Eucariotos (Dennis *et al*., 2013).

Neste estudo, discutimos os resultados obtidos com os previamente publicados por Cury e colaboradores (2011), porém, deve-se ressaltar que existem viéses, como a metodologia de filtragem deste estudo que foi diferente da realizada previamente, onde a amostra de água foi filtrada diretamente em membranas de 0,22 µm, além do viés da técnica de amplificação (PCR) utilizada por Cury no estudo anterior.

Diversidade de Bactérias

Nas duas amostras as proteobactérias representam o filo bacteriano mais abundante (cerca de 90% na amostra P e 50% na amostra E), como esperado em ambientes marinhos. Porém, a abundância relativa na amostra P é muito mais alta que o normalmente encontrada em ambientes similares. No estudo realizado por Cury, na amostra obtida no mesmo local, o percentual de proteobactérias foi de aproximadamente 45%, comparável a amostra E do estudo atual, mas apenas metade do estimado na amostra P. Num estudo conduzido por Trindade-Silva e colaboradores, na praia João-Fernandinho, em Búzios, também na região dos lagos, próximo a Arraial do Cabo, a abundância de Proteobactérias foi também de aproximadamente 50% (Trindade-Silva *et al.,* 2012).

Dentre as classes de proteobactéria, a dominante é a Alfa-proteobactéria, seguida pela classe das Gamma-proteobactérias que, ao contrário do resultado obtido por Cury e colaboradores (onde a abundância foi abaixo de 1%), apresentou mais de 40% na amostra P e cerca de 25% na amostra E. No estudo realizado por Trindade-Silva em Búzios, o perfil de abundância das classes de proteobactéria é similar ao encontrado no presente estudo, com cerca de 35% de alfa-proteobactérias e aproximadamente 15% de gamma-proteobactérias.

O segundo filo bacteriano mais abundante foi Bacteroidetes, sendo o mesmo mais abundante na amostra E (cerca de 25% contra 5% na amostra P). Membros deste filo estão presentes em diferentes nichos ecológicos, incluindo solo, água do mar, água doce e trato gastrointestinal de animais, em geral exercendo diversas funções biológicas, incluindo degradação de matéria orgânica, o que pode explicar a maior abundância na amostra E (François *et al.,* 2011). No estudo realizado previamente em Arraial do Cabo, este filo é o terceiro mais abundante (atrás do filo das Cianobactérias), com aproximadamente 18%. Em Búzios, esta abundância foi ainda mais baixa, cerca de 10%, porém, mesmo assim ocupando a segunda posição em abundância.

A classe mais abundante de Bacteroidetes, nas duas amostras, foi a classe das Flavobacterias, com mais de 90% do total de Bacteroidetes. Este resultado corrobora o obtido por Cury, onde a classe Flavobacteria também foi a mais abundante deste filo.

O filo das cianobactérias aparece em terceiro lugar na abundância relativa, porém, está concentrado na amostra E, com aproximadamente 17% (na amostra P, cerca de 1%). O fato de as cianobactérias terem sido retidas praticamente em sua totalidade na membrana de 0,8 µm pode se dever ao fato de que suas células podem ter até 3 µm, além delas serem muito encontradas em relações de simbiose com algas (Morel *et al*., 1993; Yew *et al*., 2005). Quase a totalidade das sequencias de cianobactéria foram classificadas na classe Cyanophyceae, ordem Chroococcales. Podemos ressaltar que, ao contrário da maior parte dos ambientes marinhos oligotróficos, onde o gênero de Cianobactéria mais abundante costuma ser *Prochlorococcus* (Partensky *et al*., 1999), neste estudo os gêneros mais abundantes foram *Synechococcus* (49.82%) e *Synechocystis* (35.95%), enquanto apenas poucos reads (40 na amostra P e 53 na amostra E) foram classificados como *Prochlorococcus.* Este resultado corrobora os resultados das análises físico-químicas, indicando também que possivelmente a amostra coletada estava sob impacto do fenômeno da ressurgênica, pois em diversos estudos, nas regiões sob impacto deste fenômeno, o gênero *Synechococcus* foi dectectado em mais alta abundancia do que *Prochlorococcus* (Preston *et al*., 2011).

Um total de 1.114 *reads* (13,12%) foi classificado como *Synechococcus* CC9311. Esta estirpe foi inicialmente isolada da Corrente da Califórnia (Brian *et al*., 2006), e desde então diversas estirpes relacionadas foram encontradas em ambientes costeiros. Um estudo de genômica comparativa mostrou que a estirpe CC9311 possui diversas adaptações, como enzimas do metabolismo de metais e aparatos de captação de luz diferentes dos presentes em mar aberto, sugerindo fortemente uma adaptação desta estirpe a ambientes costeiros (Waterbury*et al*., 1989; Palenik *et al*., 2001).

Comparando todas as ordens bacterianas encontradas, podemos notar que a mais abundante na amostra P são as Rhodobacterales (Alphaproteobacteria), com 42,23%, seguido da ordem Alteromonadales (Gammaproteobacteria) com 17,47%. Por outro lado, na amostra E, nota-se um equilíbrio entre a ordem mais abundante, as Flavobacteriales (Bacteroidetes) com 25,45% e as Rhodobacterales (Alphaproteobacteria) com 25,03%. A alta abundância da ordem Rhodobacterales corrobora o estudo baseado em amplificação de genes de RNAr previamente realizado por Cury e colaboradores (Cury *et al*., 2011) no mesmo ambiente.

Porém, neste estudo foi possível classificar as sequências em um nível taxonômico mais profundo do que o estudo de Cury (Cury *et al*., 2011), devido a limitação da técnica de amplificação de um trecho do gene codificador do RNAr da menor subunidade ribossomal (16S/18S) utilizado no estudo prévio. Através da técnica de sequenciamento total sem amplificação, foi possível notar que as famílias mais abundantes na amostra P foram Rhodobacteraceae (43,58%) e Alteromonadaceae (8,89%), enquanto na amostra E foram Flavobacteriaceae (25,20%) e Rhodobacteraceae (25,01%) (Fig. 7, trabalho 1).

A família Rhodobacteraceae pertence à classe das Alfaproteobactérias e possui diversas espécies quimiotróficas e fototróficas (bactérias fotossintetizantes anoxigênicas) (Swingley*et al*.,2007). Já a família das Alteromonadaceae pertence à classe das gammaproteobactérias (Ivanova *et al*., 2001; Ivanova *et al*., 2004) e a maioria dos seus gêneros habita ambientes marinhos (Kwak *et al*., 2012). Dentre os membros desta família, podemos citar a estirpe *Alteromonadaceae* sp. G7, que recentemente teve seu genoma sequenciado (Kwak *et al*., 2012). Esta bactéria foi isolada de um ambiente costeiro, em um estudo direcionado a detectar organismos degradadores de ágar. Outro membro da família com interesse biotecnológico é o *Saccharophagus degradans*, que é capaz de degradar pelo menos 10 tipos de polissacarídeos complexos, como celulose, quitina, beta-glucana, laminarina, peptina, pululano, amido e xilana (Ekborg*et al*., 2006).

A família *Flavobacteriaceae* constitui o maior grupo de Bacteroidetes. Muitas espécies desta família habitam ambientes marinhos, contribuindo de maneira importante para a mineralização de matéria orgânica nesses ecossistemas (Cottrel *et al.,* 2000; Zhang *et al., 2012*). Talvez este fato explique o porquê destas bactérias estarem em maior abundância na membrana de 0,8 µm (amostra E), pois estariam associadas a partículas de matéria orgânica.

Diversidade de Eucariotos

Os grupos mais abundantes de eucariotos encontrados foram Viridiplantae e Stramenopiles, nas amostras P e E, respectivamente. Porém, analisando apenas os resultados das sequências de genes de RNAr 18S, o grupo mais abundante nas duas amostras foi Metazoa. Este resultado corrobora o resultado de Cury e colaboradores (2011) e é diferente do obtido com a análise do total de sequências, podendo indicar um viés quando se analisa apenas sequencias de RNAr 18S.

O grupo mais abundante de Viriplantae, nas duas amostras, é o das algas verdes da divisão Chorophyta (mais de 85% dos reads de Viriplantae), sendo os principais gêneros encontrados, os do picofitoplancton, *Micromonas* e *Ostreococcus.*

Do grupo Stramenopiles, o filo mais abundante é o Bacillariophyta (diatomáceas). Estes organismos são os membros mais importantes do picofitoplancton, e estão entre os mais diversificados grupos de eucariotos fotossintetizantes, com possivelmente mais de 100 mil espécies, contribuindo com cerca de 40% da produção primária marinha (Nelson *et al.,* 1995; Maumus *et al.,* 2009).

Juntos, os organismos que constituem o picofitoplancton possuem grande importância na produção primária em ambientes oligotróficos (responsáveis por até 80% da biomassa autotrófica) (Worden *et al.,* 2004; Piganeau*et al.,* 2007).

Diversidade de Arquéias

Nas duas amostras, a abundância de Archaea foi muito baixa, entre 0,04% (amostra P) e 0,10% (amostra E) (Figuras 1 e 2 do trabalho 1). Esta baixa abundância contrasta com ambientes similares, onde a mesma pode variar de 1,1% (Buzios, Praia João Fernandinho) (Trindade-Silva *et al*., 2012) a 2,9% (Costa da Ilha Galapagos) (Rusch *et al., 2007*).

Gêneros mais abundantes

Os gêneros mais abundantes foram *Ruegeria* e *Roseobacter* (amostra P) e *Synechococcus* e *Lacinutrix* (amostra E). Os 2 gêneros mais abundantes na amostra P pertencem à família Rhodobacteraceae (alpha-proteobacteria), já discutida anteriormente, sendo *Roseobacter* um gênero de bactérias fotossintetizantes anoxigênicas aeróbias (AAP) (Yurkov *et al*., 1998). Estas bactérias são muito abundantes em ambientes marinhos, podendo chegar a 20% do bacterioplancton em ambientes costeiros ou 15% em ambientes de mar aberto [60]. Elas parecem participar de maneira significativa no ciclo do enxofre e alguns isolados deste grupo foram os primeiros organismos a apresentarem simultaneamente duas vias para degradação de um composto organosulfurado, produzido no metabolismo secundário de alguns organismos do fitoplâncton, chamado dimetilsulfoniopropinato (DMSP) (González*et al*., 1999; Vila-Costa *et al.,* 2006).

Já o gênero *Ruegeria* não é capaz de realizar fotossíntese anoxigênica, apesar de ser filogeneticamente muito próximo às *Roseobacter*, mas também atua no ciclo do enxofre e algumas estirpes possuem também a capacidade de degradar DMSP. Este gênero necessita de NaCl para seu crescimento, habitando portanto apenas ambientes salinos (Uchino *et al*., 1998; Reisch *et al*., 2013). A alta abundância destes gêneros foi determinante para que o segundo trabalho desta tese fosse realizado, com intuito de investigar mais a fundo a diversidade e abundância dos gêneros de AAPs no ambiente estudado.

Ao contrário do esperado, os gêneros mais abundantes na amostra E (entre as sequências anotadas) não foram de eucariotos, e sim de bactérias. O gênero *Synechococcus* (Cianobactéria) já discutido anteriormente, foi o mais abundante, sendo o seguido por um gênero da família Flavobacteriaceae (*Lacinutrix*). Assim como muitas das Cianobactérias, as bactérias do gênero *Lacinutrix* vivem muitas vezes em simbiose com eucariotos, como algas e copépodes (Nedashkovskaya *et al*., 2008), o que pode explicar a retenção delas na membrana de 0,8 µm.

Vale ressaltar ainda a abundância do gênero *Pelagibacter* (6% na amostra P e 5% na amostra E). Este gênero pertence ao clado SAR11, distribuído cosmopolitamente nos oceanos e muitas vezes dominante no bacterioplancton da superfície marinha (Rappé*et al*., 2002; Morris*et al.,* 2002). Porém, na maioria dos ambientes oligotróficos, a abundância encontrada é maior do que 10%, acima da encontrada neste estudo (Trindade-Silva *et al.,* 2012; Allen *et al*., 2012).

Análises funcionais no MG-RAST

As sequências das duas amostras foram anotadas funcionalmente e categorizadas utilizando o *pipeline* do MG-RAST. A classificação funcional categorizada pelo SEED foi muito similar (ao nível mais alto) para as duas amostras, sendo a principal diferença nas categorias “*Phages, Prophages, Transposable* *elements and Plasmids"* e "*Photosynthesis"*, onde a abundância foi maior na amostra E. Esta maior abundância da categoria fotossíntese pode ser explicada pela presença de cianobactérias e também algas na amostra E (praticamente ausentes na amostra P).

Da mesma forma, na classificação funcional categorizada do KEEG (KO), as amostras foram muito parecidas (ao nível mais alto). As principais diferenças foram na categoria "*Metabolism and Environmental Information Processing",* mais abundante na amostra P e "*Genetic Information Processing*, *Cellular Processes*, *Human Diseases* and Organismal Systems" que, como esperado, foi mais abundante na amostra E. A possível explicação para este fato reside no fato de que a maquinaria celular e de processamento de informação genética é mais complexa em eucariotos, e portanto esta categoria está em maior abundância na amostra de maior riqueza em eucariotos.

Montagem das sequências ambientais (reads) e anotação das janelas abetas de leitura (ORFs)

A montagem de sequências ambientais é um problema complexo e muitos algorítmos foram propostos para este fim (Kultima *et al.,* 2012; Peer *et al.,* 2013). Os principais problemas da montagem são o fato de que em geral o sequenciamento de metagenomas possui baixa cobertura (muitos genomas para cobrir) e ainda, a possibilidade de montagem de sequências quiméricas (montagem inter-genômica). Optamos por utilizar o CAP3 com os parâmetros padrão, que são estringentes, para tentar contornar o problema da formação de quimeras, porém, o problema da baixa cobertura só poderia ser contornado com um maior esforço de sequenciamento. Foram obtidos um total de 29.074 *contigs* e 269.587 *singlets* (*reads* não montados) na amostra P. Na amostra E, foram obtidos 20.792 *contigs* e 396.371 *singlets*. O maior número de sequências não montadas na amostra E pode estar relacionado ao fato de esta amostra apresentar maior diversidade do que a amostra P, além da maior complexidade de genomas eucarióticos do que procarióticos, mais presentes na amostra E.

Utilizando o programa METAGENEMARK (Zhu *et al.,* 2010), foi possível extrair um total de 409.111 janelas abertas de leitura (ORFs) na amostra P e

451.722 na amostra E. Este alto número de ORFs encontrado (muito maior do que o número de *contigs* formados) sugere que o programa foi capaz de identificar muitas ORFs mesmo nas sequências não montadas (*reads*), podendo ser utilizado em metagenomas não montados.

Triagem de genes com interesse biotecnológico: Policetídeo Sintases (PKS) e Peptídeo Não-Ribossomal Sintases (NRPS)

Diversos estudos foram conduzidos com objetivo de triar novos genes das famílias PKS e NRPS em micro-organismos não cultiváveis. A maioria destes estudos se focam em amplificação de domínios conservados por PCR (Trindade-Silva *et al.,* 2013; Kennedy *et al*., 2008), e geralmente visam microbiota de solo ou de invertebrados marinhos.

Entretanto, apesar do crescimento exponencial dos bancos de dados de metagenomômas sequenciados por NGS, poucos estudos foram conduzidos visando a triagem de genes do metabolismo secundário nestes metagenomas sequenciados. O principal estudo deste tipo foi conduzido por Foerstener e colaboradores (2008), em que foram triados 6 metagenomas utilizando abordagem de Modelos Ocultos de Markov (pHMM). Porém, nenhum destes estudos foi conduzido em ambientes afetados por ressurgência.

Devido a grande abundância de organismos do clado *Roseobacter* e do filo das cianobactérias nas amostras estudadas, que já foram previamente descritos como micro-organismos com alto potencial de fornecer novos compostos do metabolismo secundário, uma abordagem utilizando perfis HMM foi utilizada não apenas para busca de genes de PKS (como conduzido por Foerstener e colaboradores), mas também na busca de genes da família NRPS.

Foram utilizados 2 pHMMs dos domínios KS de PKS tipo I (um construído com as sequências de KSs iterativas e outro com as KSs modulares). O programa HMMER 3.0 foi utilizado para buscar esses perfis nos metagenomas, e as sequências obtidas foram submetidas ao sistema NapDos (Ziemert *et al*., 2012) para classificação dos domínios e posterior analise filogenética. O programa HMMER 3.0 foi escolhido por utilizar a abordagem pHMM, mais sensível que as abordagens por similaridade (como o BLAST), além de ser menos oneroso computacionalmente para ser utilizado em grande volume de dados (Eddy, 2011). Já o sistema NapDos foi escolhido pois se mostrou mais eficiente na análise de sequências obtidas em genomas incompletos ou metagenomas, onde muitas vezes só é possível obter sequências incompletas (Ziemert*et al*., 2012).

A partir do pHHM de KS modular, foram obtidas 28 sequências na amostra P e 37 na amostra E. A função destas sequências foi verificada por similaridade. Como resultado, foi confirmada a anotação de 78,58% (amostra P) e 91,9% (amostra E) destas sequências.

Utilizando o pHMM de KS iterativa, foram obtidas 21 sequências na amostra P e 16 na amostra E. Foi possível confirmar a função de 76,20% das sequências da amostra P e 75% da amostra E.

Este resultado mostra o pHMM de KS modular com maior especificidade (menos falsos positivos) do que o pHMM de KS iterativa.

O total de sequências obtidas com os 2 pHMMs e anotadas como domínio KS de PKS foi de 84 (82,35% das 102 obtidas inicialmente). As demais sequências foram consideradas falsos positivos (17,65%), e isto pode se dever ao fato de a abordagem de pHMMs ser muito sensível, e detectar inicialmente sequências homólogas às PKSs, como por exemplo, Ácido Graxo Sintases (FAS) (Fisch, 2013).

As vantagens do uso de pHMMs para triagem de PKSs em metagenomas e os possíveis falsos positivos obtidos já foram discutidas anteriormente por Foerstner e colaboradores (Foerstner *et al*., 2008) e corroboram os resultados obtidos neste estudo.

A abundância relativa de domínios KS no metagenoma de Arraial do Cabo foi de 0,0092% (38 em 409.111 ORFs) na amostra P e 0,0101% (46 em 451.722 ORFs) na amostra E. No estudo de Foerstner e colaboradores (2008), a maior abundância obtida dentre os metagenomas triados foi na amostra de solo de uma fazenda de Minnesota (Tringe *et al.,* 2005), onde foram encontradas 52 sequências de KS tipo I em 183.536 ORFs (0,0283%), apenas 2,8 vezes maior que a abundância em Arraial do Cabo. Além disso, no mesmo estudo conduzido por Foerstner e colaboradores (2008), o metagenôma do Mar de Sargasso também foi triado em busca de PKS tipo I. Neste ambiente oligotrófico, foram encontradas 69 sequeências de KS tipo I em 1.214.207 ORFs (0,0056% do total), possuindo uma abundância relativa menor do que a encontrada em nossa amostra P (que foi filtrada de maneira similar à do Mar de Sargasso, que abrange organismos de 0,8 a 0,1 µm).

Este resultado mostra o potencial do ambiente estudado neste trabalho, pois sabemos que em solo a diversidade e riqueza de espécies é maior do que em ambientes marinhos, existindo portanto em solos uma espécie de "corrida armamentista" entre os organismos, com uma ampla produção de metabólitos secundários (Handelsman *et al.,* 1998).

De forma inesperada, das sequências obtidas com o pHMM de KS iterativa, apenas poucas sequências de fato iterativas foram obtidas (apenas sequências classificadas como "Enediyne").

Foi obtida uma alta abundância de KS modular (incluindo as de PKS híbridas) com os dois pHMMs, o que pode ser explicado pelo fato de que em PKSs modulares existem múltiplas cópias de cada domínio, enquanto nas iterativas, apenas uma (Jenke-Kodama *et al*., 2005).

Posteriormente, uma análise filogenética foi realizada no NapDos, com objetivo de classificar de maneira mais precisa as sequÊncias de domínio KS.

A topologia da árvore obtida corrobora diversos estudos filogenéticos anteriores (Shulse *et al*., 2011; Ziemert *et al*., 2012). Foi possível separar as sequências de: (i) FAS (fab) homólogas às PKSs, (ii) KS tipo II, (iii) PUFA (ácido graxo poliinsaturado), (iv) Trans-AT , (v) Iterativa, (vi) Híbrida PKS-NRPS, (vii) KS1 (sequências presentes nos módulos iniciadores da síntese de PKS) e (viii) Modular. Os resultados desta árvore confirmam a dominância de sequências modulares de PKS nas duas amostras (Cis, Trans e Híbridas).

Para triar genes NRPS no metagenoma de Arraial do Cabo, foi utilizado um pHMM do domínio C. Foi possível obter um total de 50 sequências (14 da amostra P e 36 da amostra E). Através das análises de similaridade com BlastP contra o RefSeq e da classificação pelo NapDos, foi possível confirmar a anotação de 92,83% (amostra P) e 91,67% (amostra E) das sequências obtidas, totalizando 46 sequências de domínio C de NRPS. Este resultado mostrou uma alta especificidade do pHMM de domínio C, maior do que os pHMMs de domínio KS de PKS.

Da mesma forma que para os domínios KS de PKS, os domínios C obtidos foram submetidos à analise filogenética (apenas sequências maiores que 200 aminoácidos). A topologia obtida corrobora estudos anteriores (Ziemert *et al*., 2012) e a maioria das sequências foram classificadas como do tipo LCL e Epimerização.

Estimativa da abundância e diversidade de AAPs

No primeiro trabalho desta tese, foi possível detectar uma alta abundância do gênero *Roseobacter* e outros gêneros próximos a este filogenéticamente.

Muitas espécies deste clado filogenético são conhecidas pela capacidade de realizar fotossíntese anoxigênica em ambiente aeróbico, sendo assim classificadas como AAPs (Swingley*et al*., 2007). Por este motivo, para confirmar a alta abundância de AAPs na amostra estudada, no segundo trabalho desta tese, foi desenvolvido um pipeline para estimar a abundância e estudar a diversidade de bactérias fotossintetizantes anoxigênicas anaeróbias (AAPs), através da triagem de genes marcadores *puf*M, *puf*L e *bch*X utilizando pHMMs dos mesmos. Estes genes são exclusivos de espécies AAPs, diferenciando assim, por exemplo, de espécies de *Roseobacter* exclusivamente heterotróficas.

Com o pipeline desenvolvido (em linguagem RUBY), foi possível triar um total de 12.672.518 *reads* de 82 metagenomas do GOS, além de 1.064.888 *reads* das duas amostras de Arraial do Cabo, utilizando um computador pessoal (PC) com recursos limitados.

A abordagem utilizada se mostrou sensível e específica (95,34% de especificidade e 100% de sensibilidade), mostrando-se eficaz e possível de se executar em um computador pessoal com poucos recursos, mesmo para estudos de larga escala.

Até o momento da escrita deste trabalho, este é o primeiro estudo a triar AAPs não apenas em metagenomas montados, como em não montados, uma vez que o estudo prévio realizado por Yutin e colaboradores (2007) foi realizado nos metagenomas montados da primeira verão do GOS.

Além disso, no estudo de Yutin e colaboradores (2007) os metagenomas foram concatenados e montados juntos, aumentando a chance de formação de sequencias quiméricas (Clark *et al*., 2012). As vantagens de se triar metagenomas não montados (*reads*) são: (i) evitar a montagem, que é lenta e custosa computacionalmente, principalmente em metagenomas grandes e complexos; (ii) evitar a formação de sequências quiméricas.

Porém, por causa da limitação de tamanho de reads obtidos, é necessário montar os metagenomas para estudos de filogenia ou de sintenia dos operons *puf*.

Os resultados de abundância obtidos nos metagenomas não montados foram consistentes com os obtidos nos montados, mostrando a viabilidade da utilização em dados brutos.

Analisando os 10 metagenomas que mostraram maior abundância de AAPs, podemos observar que 8 (80%) são geograficamente próximos da linha do Equador, o que é esperado, pela maior incidência de luz nestes locais (favorecendo o crescimento destes organismos fotossintetizantes).

Tanto nos resultados obtidos nos *reads* não montados, quanto nos montados, o metagenoma de maior abundância de AAPs foi o de Arraial do Cabo (cerca de 23,88% das células existentes no ambiente). Este resultado mostra o ambiente estudado nesta tese como um dos com maior abundância de AAP já descritos, similar ao ambiente muito oligotrófico do pacífico sul descrito por Lami e colaboradores (2007), com aproximadamente 24% de abundância.

A abundância de AAP neste ambiente é incomum e superior a encontrada em diversos ambientes (utilizando diferentes técnicas), como no trabalho de Waidner e colaboradores (2007) em estuários (de 12% a 17% de abundância), no Oceano Ártico (de 5% a 8% de abundância) (Cottrell & Kirchman, 2009) e em regiões costeiras do pacífico sul (em média 1,2% de abundância) (Ritchie *et al*., 2012).

O segundo ambiente mais abundante encontrado foi o GS33 (Browns Bank, Gulf of Maine), que é uma laguna hipersalina anóxica (63.4 PSU, oxigênio dissolvido: 0,06 mg l^{-1}) com abundância de AAPs em cerca de 15,64%. Este ambiente também se mostrou o mais abundante em bactérias fotossintetizantes anoxigênicas no estudo de Yutin e colaboradores (2007), porém foi discutido separadamente dos outros ambientes, por ser anóxico e, portanto, possuir bactérias anaeróbias fotossintetizantes anoxigênicas. Entretanto, nos estudos filogenéticos conduzidos por Yutin e colaboradores (2007) e também no presente estudo, a comunidade deste ambiente se mostrou mista, com bactérias do filogrupo G (aeróbias) estando presentes entre as mesmas (16,48% dos "*reads* equivalentes" nas ORFs).

É importante ressaltar que as comparações entre amostras podem apresentar viéses pelas diferenças entre métodos de filtragem, extração de DNA e sequenciamento. Além do mais, estudos já demostraram a sazonalidade da comunidade de AAP em alguns ambientes, mostrando que, como esperado, no verão a comunidade tende a estar em mais alta abundância do que no inverno (Cottrell & Kirchman, 2009; Ferrera *et al*., 2013). Alguns estudos tem também demostrado que outras variáveis ambientais parecem influenciar na abundância e diversidade de AAPs, como a quantidade de luz disponível, concentração de nutrientes, temperatura e concentração de Clorofila A (Masín *et al.,* 2006; Zhang & Jiao, 2007; Lamy *et al.,* 2011). Entretanto, Ferrera e colaboradores (2013) demostraram que a maior correlação é entre a quantidade de luz e a abundância de AAPs, sendo mais importante até mesmo do que a concentração de nutrientes, corroborando resultados obtidos em experimentos de cultura de AAPs, onde se mostrou que a luz é capaz de aumentar a eficiência na assimilação de carbono e do crescimento destas bactérias (Hauruseu & Koblízek, 2012).

As amostras de Arraial do Cabo foram obtidas no verão, este fato talvez possa explicar a alta abundância de AAPs encontrada neste ambiente, porém estudos futuros precisam ser realizados para estimar a abundância dos mesmos em outra estações do ano.

Analisando a árvore bayesiana gerada neste estudo, podemos perceber que 62,15% (amostra P) e 13,87% (amostra E) das sequências de *puf*M foram agrupadas com as sequencias do filogrupo G (Clado *Roseobacter*). Este resultado mostra a dominância do clado Roseobacter nestas amostras, corroborando os resultados do primeiro trabalho realizado para esta tese e também o estudo de Cury e colaboradores (2011).

A topologia obtida na árvore corrobora o estudo de Yutin e colaboradores (2007), separando os filogrupos da maneira esperada.

Estes resultados foram também confirmados pelas análises de sintenia do operon *puf*, como discutido por Yutin e colaboradores (2007) no mesmo estudo anterior.

No estudo de Yutin e colaboradores (2007), as bactérias *Rhodoplanes* (alfa-proteobacteria) e *Rosealetes* (beta-proteobacteria) agruparam no mesmo clado, apesar de serem de classes diferentes. Entretanto, no atual estudo, as bactérias do gênero *Rosealetes* ficaram no mesmo clado de outras beta-proteobacterias (*Rubravivax* e *Rhodoferax*), separando desta forma as beta-proteobactérias do clado das alfa-proteobactérias.

Já a alpha-proteobactéria do gênero *Rhodoplanes* agrupou com as gamma-proteobactérias do filogrupo K. Este fato possivelmente pode ser explicado por uma transferência horizontal dos genes do aparato fotossintético. Este tipo de transferência já foi inferida em estudos anteriores, além de já ter sido detectado uma espécie de *Roseobacter* com todos os genes do operon em um plasmídeo (Nagashima *et al.,* 2007; Igarashi *et al*., 2001; Swingley *et al*., 2009).

Surpreendentemente, nenhuma AAP da subclasse α-4 foi detectada neste estudo. Este grupo normalmente está presente em diversos ambientes marinhos (Yurkov & Csotonyi, 2009), e na árvore filogenética gerada neste estudo, as sequências de referencia (E*rythrobacter*, *Blastomonas, Sphingomonas* e *Porphyrobacter*) formam um clado a parte, sem nenhuma sequência ambiental.

Os resultados da classificação obtida pela análise filogenética (e corroborada pelas análises de sintenia do operon *puf*) mostram a dominância do filogrupo G nas amostras de Arraial do Cabo, com 82,36% (amostra P) e 64,05% (amostra E) do total de *reads* de AAP neste ambiente. Estes resultados não corroboram os obtidos por Ferrera e colaboradores (2013) em regiões costeiras do Mar Mediterrâneo, onde os grupos de alpha-proteobacteria E, F e G apenas estão em abundância maior que os de gamma-proteobacteria (grupo K) no inverno (e com altas concentrações de nutrientes), já que neste estudo, as amostras de Arraial foram coletadas no verão em época de ressurgência.

O filogrupo G foi o mais cosmopolita de todos, presente em 11 das 12 amostras analisadas, corroborando os resultados obtidos em estudos anteriores (Buchan *et al*., 2005; Yutin *et al., 2007)*.

Entretanto, nas amostras do GOS, este filogrupo está em maior abundância nas amostras do Oceano Índico (menos no GS111), do que nas amostras do Oceano Pacífico (amostras da Ilha de Galápagos) ou da costa dos Estados Unidos. Além disso, é importante ressaltar que a abundância relativa deste grupo é maior nas amostras de membranas de 0,22 µm (amostra P e GS108a) do que nas de 0,8 µm (amostra E e GS108b), sugerindo que o filogrupo G pode preferir o estilo de vida livre. Da mesma maneira, o filogrupo A parece preferir um estilo de vida livre, uma vez que este grupo está ausente em amostras de 0,8 µm. Além disso, este grupo está ausente da amostra anóxica (GS033), sugerindo a importância do oxigênio para este grupo.

Por outro lado, o filogrupo H é mais abundante nas amostras de 0,8 µm e GS033, e o filogrupo E está presente apenas na GS033 e em baixa abundância na GS112.

Surpreendentemente, membros do filogrupo F (clado das *Rhodobacter,* constituído em sua maioria por organismos anaeróbicos) foram encontrados nas duas amostras de Arraial do Cabo (porém com maior abundância na amostra E), mas nas amostras do GOS foram encontrados apenas na amostra anóxica (GS033), como esperado.

A correlação entre a abundância de alguns grupos em membranas de 0,8 µm e na amostra anóxica pode ser talvez explicada pela possível formação de sub-nichos anóxicos em ambientes oxigenados, através do agrupamento de partículas orgânicas, que normalmente ficam retidas em membranas de 0,8 µm.

Os grupos abundantes nestas amostras podem possuir adaptações para ambientes ricos em partículas (e por consequência túrbidos), como por exemplo adaptações na captação de luz ou de utilização de substratos orgânicos presentes nestes ambientes (Cottrell *et al.*, 2010).

6 PERSPECTIVAS

A partir dos resultados deste estudo, será possível o desenho de iniciadores de PCR ou sondas para testes de hibridização em busca de genes de interesse biotecnológico em bibliotecas de fosmídeo construídas com o mesmo DNA utilizado no pirosequenciamento desse estudo. Posteriormente, as bibliotecas poderão ser triadas em busca de clones com sequencias de PKS e NRPS, para que seja possível a expressão heteróloga dos mesmos.

Como trabalhos futuros, podemos ainda iniciar um estudo de metagenômica ao longo do tempo na região, com novas coletas no mesmo local, e em outros pontos da região, com o objetivo de analisar a variação espaço-temporal da comunidade de AAPs em Arraial do Cabo.

7 CONCLUSÕES

Através da metodologia empregada neste estudo, foi possível estimar a diversidade microbiana do ambiente da Praia dos Anjos – Arraial do Cabo – RJ.

As duas amostras mostraram grande abundância de proteobactérias, indicando que este ambiente é amplamente dominado pelas mesmas. Por outro lado, a abundância de Arquéias foi muito baixa, mostrando que no momento da coleta, este domínio da vida era escasso na amostra.

A alta abundância de bactérias do clado *Roseobacter* corrobora a hipótese de que a amostra estudada estava sendo afetada pelo fenômeno da ressurgência e por consequência por um *bloom* do fitoplanctôn. Esta hipótese é ainda corroborada pelos parâmetros físico-químicos da amostra, como baixa concentração de amônia e alta concentração de nitrato.

Através da triagem de genes do metabolismo secundário com interesse biotecnológico (PKS e NRPS), foi possível demonstrar o potencial do ambiente estudado. Foi possível concluir que é possível encontrar novos genes das duas famílias no genoma dos organismos presentes no local.

Através do pipeline desenvolvido para estimar a diversidade e abundância de AAPs, foi possível mostrar que o ambiente estudado possui abundância mais alta do que todos os metagenomas do GOS, com abundância não usual (cerca de 23%) desses organismos na amostra. Foi possível determinar também os filogrupos de AAPs presentes nas amostras, sendo o filogrupo G (clado *Roseobacter*), o mais abundante. Os resultados obtidos mostram que o ambiente estudado possui uma das maiores abundâncias de bactérias do clado *Roseobacter* já encontradas, mesmo levando em consideração estudos utilizando diversas metodologias para estimar esta abundância.

8 REFERÊNCIAS BIBLIOGRÁFICAS

Afiahayati, Sato K, Sakakibara Y. An extended genovo metagenomic assembler by incorporating paired-end information.PeerJ. 2013 Oct 31;1:e196.

Agnan J, Korch C, Selitrennikoff C. Cloning heterologous genes: problems and approaches. Fungal Genet Biol. 1997 Jun;21(3).

Albuquerque ALS, Belem AL, Zuluaga FJB, Cordeiro LGM, Mendoza U, Knoppers BA, et al. Particle Fluxes and Bulk Geochemical Characterization of the Cabo Frio Upwelling System in Southeastern Brazil: Sediment Trap Experiments between Spring 2010 and Summer 2012. An Acad Bras Cienc. 2014 May 14;0(0).

Allen LZ, Allen EE, Badger JH, McCrow JP, Paulsen IT, Elbourne LD, et al. Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic.The ISME journal. 2012;6(7):1403–14.

Atamna-Ismaeel N, Finkel O, Glaser F, von Mering C, Vorholt JA, Koblížek M, et al. Bacterial anoxygenic photosynthesis on plant leaf surfaces. Environmental Microbiology Reports. 2012;4(2):209–16.

Azam F. Microbial control of oceanic carbon flux: the plot thickens. SCIENCE-NEW YORK THEN WASHINGTON-. 1998;694–5.

Balvanera P, Pfisterer AB, Buchmann N, He J-S, Nakashizuka T, Raffaelli D, et al. Quantifying the evidence for biodiversity effects on ecosystem functioning and services: Biodiversity and ecosystem functioning/services. Ecology Letters. 2006 Oct;9(10):1146–56.

Béjà O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, Hamada T, et al. Unsuspected diversity among marine aerobic anoxygenic phototrophs. Nature. 2002;415(6872):630–3.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank.Nucleic Acids Research. 2012 Nov 27;41(D1):D36–D42.

Bryant DA, Frigaard N-U. Prokaryotic photosynthesis and phototrophy illuminated. Trends in Microbiology. 2006 Nov;14(11):488–96.

Buchan A, Gonzalez JM, Moran MA. Overview of the Marine Roseobacter Lineage.Applied and Environmental Microbiology. 2005 Oct 4;71(10):5665–77.

Caboche S, Pupin M, Leclere V, Fontaine A, Jacques P, Kucherov G. NORINE: a database of nonribosomal peptides. Nucleic Acids Research. 2007 Dec 23;36(Database):D326–D331.

Campos EJD, Velhote D, Silveira ICA (2000) Shelf break upwelling driven by Brazil Current cyclonic meanders. Geophys Res Lett 27: 751–754.

Cane DE. Harnessing the Biosynthetic Code: Combinations, Permutations, and Mutations.Science. 1998 Oct 2;282(5386):63–8.

Castelao RM, Barth JA (2006) Upwelling around Cabo Frio, Brazil: the importance of wind stress curl. Geophys Res Lett 33: 3602.

Castoe TA, Stephens T, Noonan BP, Calestani C. A novel group of type I polyketide synthases (PKS) in animals and the complex phylogenomics of PKSs. Gene. 2007 May;392(1-2):47–58.

Ceballos-Lascurain H. Tourism, ecotourism, and protected areas: The state of nature-based tourism around the world and guidelines for its development. [Internet]. Iucn; 1996 [cited 2014 Jan 10]. Available from: http://www.cabdirect.org/abstracts/19961808274.html

Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics. 2013 Feb 15;29(4):435–43.

Coates A, Hu Y, Bax R, Page C. The future challenges facing the development of new antimicrobial drugs. Nat Rev Drug Discov. 2002 Nov;1(11):895–910.

Coates AR, Halls G, Hu Y. Novel classes of antibiotics or more of the same?: New antibiotic classes are urgently needed. British Journal of Pharmacology. 2011 May;163(1):184–94.

Coelho-Souza SA, Pereira GC, Coutinho R, Guimarães JR (2013) Yearly variation of bacterial production in the Arraial do Cabo protection area (Cabo Frio upwelling region): an evidence of anthropogenic pressure. Brazilian Journal of Microbiology 44: 1349–1357.

Cottrell MT, Kirchman DL. Photoheterotrophic Microbes in the Arctic Ocean in Summer and Winter. Applied and Environmental Microbiology. 2009 Jun 5;75(15):4958–66.

Cottrell MT, Kirchman DL. Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low-and high-molecular-weight dissolved organic matter. Applied and Environmental Microbiology. 2000;66(4):1692–7.

Cottrell MT, Ras J, Kirchman DL. Bacteriochlorophyll and community structure of aerobic anoxygenic phototrophic bacteria in a particle-rich estuary. The ISME journal. 2010;4(7):945–54.

Courtois S, Cappellano CM, Ball M, Francou F-X, Normand P, Helynck G, et al. Recombinant Environmental Libraries Provide Access to Microbial Diversity for Drug Discovery from Natural Products. Applied and Environmental Microbiology. 2003 Jan 1;69(1):49–55.

Csotonyi JT, Stackebrandt E, Swiderski J, Schumann P, Yurkov V. An alphaproteobacterium capable of both aerobic and anaerobic anoxygenic photosynthesis but incapable of photoautotrophy: Charonomicrobium ambiphototrophicum, gen. nov., sp. nov. Photosynthesis Research. 2011 Feb 10;107(3):257–68.

Csotonyi J, Swiderski J, Stackebrandt E, Yurkov V. A New Extreme Environment for Aerobic Anoxygenic Phototrophs: Biological Soil Crusts. In: Hallenbeck PC, editor. Recent Advances in Phototrophic Prokaryotes [Internet].Springer New York; 2010. p. 3–14.

Cude WN, Mooney J, Tavanaei AA, Hadden MK, Frank AM, et al. (2012) Production of the antimicrobial secondary metabolite indigoidine contributes to competitive surface colonization by the marine roseobacter Phaeobacter sp. strain Y4I. Appl Environ Microbiol 78: 4771–4780. doi:10.1128/AEM.00297-12.

Cury JC, Araujo FV, Coelho-Souza SA, Peixoto RS, Oliveira JAL, Santos HF, et al. Microbial Diversity of a Brazilian Coastal Region Influenced by an Upwelling System and Anthropogenic Activity. Gilbert J, editor. PLoS ONE. 2011 Jan 27;6(1):e16553.

Desriac F, Jégou C, Balnois E, Brillet B, Chevalier P, Fleury Y. Antimicrobial Peptides from Marine Proteobacteria. Marine Drugs. 2013 Sep 30;11(10):3632–60.

Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Computational Biology. 2011 Oct 20;7(10):e1002195.

Ekborg NA, Taylor LE, Longmire AG, Henrissat B, Weiner RM, Hutcheson SW. Genomic and Proteomic Analyses of the Agarolytic System Expressed by Saccharophagus degradans 2-40. Applied and Environmental Microbiology. 2006 May 3;72(5):3396–405.

Ekborg NA, Taylor LE, Longmire AG, Henrissat B, Weiner RM, Hutcheson SW. Genomic and Proteomic Analyses of the Agarolytic System Expressed by Saccharophagus degradans 2-40. Applied and Environmental Microbiology. 2006 May 3;72(5):3396–405.

F P, J B, D V. Differential distribution and ecology of Prochlorococcus and Synechococcus in oceanic waters : a review. Monaco, MONACO: Musée océanographique; 1999.

Fernandes P. Antibacterial discovery and development[mdash]the failure of success? Nat Biotech. 2006 Dec;24(12):1497–503.

Ferrera I, Borrego CM, Salazar G, Gasol JM. 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: Marine AAP dynamics in coastal sea. Environmental Microbiology. 2013 Nov;n/a–n/a.

Fisch KM. Biosynthesis of natural products by microbial iterative hybrid PKS–NRPS.RSC Advances. 2013;3(40):18228.

Foerstner KU, Doerks T, Creevey CJ, Doerks A, Bork P. A Computational Screen for Type I Polyketide Synthases in Metagenomics Shotgun Data. Field D, editor. PLoS ONE. 2008 Oct 27;3(10):e3515.

Gill SR. Metagenomic Analysis of the Human Distal Gut Microbiome. Science. 2006 Jun 2;312(5778):1355–9.

Goericke R. Bacteriochlorophyll a in the ocean: Is anoxygenic bacterial photosynthesis important? Limnology and oceanography. 2002;47(1):290–5.

Gokhale RS, Sankaranarayanan R, Mohanty D. Versatility of polyketide synthases in generating metabolic diversity. Current Opinion in Structural Biology. 2007 Dec;17(6):736–43.

González JM, Kiene RP, Moran MA. Transformation of Sulfur Compounds by an Abundant Lineage of Marine Bacteria in the α-Subclass of the ClassProteobacteria.Applied and environmental microbiology. 1999;65(9):3810–9.

Graça AP, Bondoso J, Gaspar H, Xavier JR, Monteiro MC, et al. (2013) Antimicrobial Activity of Heterotrophic Bacterial Communities from the Marine Sponge Erylus discophorus (Astrophorida, Geodiidae). PLoS ONE 8: e78992.

Grossart H-P, Schlingloff A, Bernhard M, Simon M, Brinkhoff T (2004) Antagonistic activity of bacteria isolated from organic aggregates of the German Wadden Sea. FEMS Microbiol Ecol 47: 387–396.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998 Oct;5(10):R245–249.

Hauruseu D, Koblizek M. Influence of Light on Carbon Utilization in Aerobic Anoxygenic Phototrophs. Applied and Environmental Microbiology. 2012 Aug 10;78(20):7414–9.

Hohmann-Marriott MF, Blankenship RE. Evolution of photosynthesis. Annu Rev Plant Biol. 2011;62:515–48.

Hunter, C. Neil; Daldal, Fevzi; Thurnauer, Marion C.; Beatty JT. [Advances in Photosynthesis and Respiration] The Purple Phototrophic Bacteria Volume 28 || New Light on Aerobic Anoxygenic Phototrophs.2009. Available from: http://libgen.org/scimag/index.php?doi=10.1007/978-1-4020-8815-5_3

Igarashi N, Harada J, Nagashima S, Matsuura K, Shimada K, Nagashima KV. Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria.Journal of molecular evolution. 2001;52(4):333–41.

Ivanova EP. Phylogenetic relationships among marine Alteromonas-like proteobacteria: emended description of the family Alteromonadaceae and proposal of Pseudoalteromonadaceae fam. nov., Colwelliaceae fam. nov., Shewanellaceae fam. nov., Moritellaceae fam. nov., Ferrimonadaceae fam. nov., Idiomarinaceae fam. nov. and Psychromonadaceae fam. nov. INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY. 2004 Sep 1;54(5):1773–88.

Ivanova EP, Mikhailov VV. A new family, Alteromonadaceae fam. nov., including marine proteobacteria of the genera Alteromonas, Pseudoalteromonas, Idiomarina, and Colwellia. Microbiology. 2001;70(1):10–7.

Jamieson RE, Rogers AD, Billett DSM, Smale DA, Pearce DA.Patterns of marine bacterioplankton biodiversity in the surface waters of the Scotia Arc, Southern Ocean.FEMS Microbiology Ecology. 2012 Apr;80(2):452–68.

Jenke-Kodama H. Evolutionary Implications of Bacterial Polyketide Synthases.Molecular Biology and Evolution. 2005 Jun 8;22(10):2027–39.

Jennings, S., Kaiser, M.J., Reynolds, J.D. (2001) "Marine Fisheries Ecology." Oxford: Blackwell Science Ltd.

Joint I, Mühling M, Querellou J. Culturing marine bacteria - an essential prerequisite for biodiscovery: Culturing marine bacteria. Microbial Biotechnology. 2010 Sep;3(5):564–75.

Kennedy J, Codling CE, Jones BV, Dobson ADW, Marchesi JR. Diversity of microbes associated with the marine sponge, Haliclona simulans, isolated from Irish waters and identification of polyketide synthase genes from the sponge metagenome. Environ Microbiol. 2008 Jul;10(7):1888–902.

Kennedy J, Marchesi JR, Dobson AD. Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. Microbial Cell Factories. 2008;7(1):27.

King GM, Smith CB, Tolar B, Hollibaugh JT. Analysis of Composition and Structure of Coastal to Mesopelagic Bacterioplankton Communities in the Northern Gulf of Mexico. Frontiers in Microbiology [Internet]. 2013 [cited 2014 Feb 24];3. Available from: http://www.frontiersin.org/Journal/10.3389/fmicb.2012.00438/full

Kolber ZS. Contribution of Aerobic Photoheterotrophic Bacteria to the Carbon Cycle in the Ocean.Science. 2001 Jun 29;292(5526):2492–5.

Kolber ZS, Van Dover CL, Niederman RA, Falkowski PG. Bacterial photosynthesis in surface waters of the open ocean.Nature. 2000 Sep 14;407(6801):177–9.

Kubo T, Ohtani E, Kondo T, Kato T, Toma M, Hosoya T, et al. Metastable garnet in oceanic crust at the top of the lower mantle. Nature. 2002 Dec 19;420(6917):803–6.

Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, et al. MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. Gilbert JA, editor. PLoS ONE.

2012 Oct 17;7(10):e47656.

Kwak M-J, Song JY, Kim BK, Chi W-J, Kwon S-K, Choi S, et al. Genome Sequence of the Agar-Degrading Marine Bacterium Alteromonadaceae sp. Strain G7. Journal of Bacteriology. 2012 Dec 3;194(24):6961–2.

Lal R, Kumari R, Kaur H, Khanna R, Dhingra N, Tuteja D. Regulation and manipulation of the gene clusters encoding type-I PKSs. Trends in biotechnology. 2000;18(6):264–74.

Lami R, Cottrell MT, Ras J, Ulloa O, Obernosterer I, Claustre H, et al. High Abundances of Aerobic Anoxygenic Photosynthetic Bacteria in the South Pacific Ocean.Applied and Environmental Microbiology.2007 May 11;73(13):4198–205.

Lamy D, De Carvalho-Maalouf P, Cottrell MT, Lami R, Catala P, Oriol L, et al. Seasonal dynamics of aerobic anoxygenic phototrophs in a Mediterranean coastal lagoon. Aquat Microb Ecol. 2011;62(2):153–63.

Lau SCK, Zhang R, Brodie EL, Piceno YM, Andersen G, Liu W-T. Biogeography of bacterioplankton in the tropical seawaters of Singapore.FEMS Microbiology Ecology. 2013 May;84(2):259–69.

Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Pillay M, et al. IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Research. 2013 Oct 16;42(D1):D568–D573.

Meyer F, Paarmann D, D’Souza M, Olson R, Glass E, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes.BMC Bioinformatics. 2008;9(1):386.

Masín M, Zdun A, Sto´n-Egiert J, Nausch M, Labrenz M, Moulisová V, et al. Seasonal changes and diversity of aerobicanoxygenic phototrophs in the Baltic Sea. Aquat Microb Ecol. 2006 Dec;45: 247–254.

Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, et al. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC Genomics. 2009;10(1):624.

Milne PJ, Hunt AL, Rostoll K, Van Der Walt JJ, Graz CJ (1998) The biological activity of selected cyclic dipeptides. J Pharm Pharmacol 50: 1331–1337.

Morel A, Ahn Y-H, Partensky F, Vaulot D, Claustre H. Prochlorococcus and Synechococcus: A comparative study of their optical properties in relation to their size and pigmentation. Journal of Marine Research. 1993;51(3):617–49.

Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al. SAR11 clade dominates ocean surface bacterioplankton communities. Nature. 2002 Dec 19;420(6917):806–10.

Nagashima KV, Hiraishi A, Shimada K, Matsuura K. Horizontal transfer of genes coding for the photosynthetic reaction centers of purple bacteria.Journal of molecular evolution. 1997;45(2):131–6.

Nagle DG, Gerwick WH. Nakienones AC and nakitriol, new cytotoxic cyclic C< sub> 11</sub> metabolites from an okinawan cyanobacterial (Synechocystis sp.) overgrowth of coral. Tetrahedron letters. 1995;36(6):849–52.

Nedashkovskaya OI, Kwon KK, Yang S-H, Lee H-S, Chung KH, Kim S-J. Lacinutrix algicola sp. nov. and Lacinutrix mariniflava sp. nov., two novel marine alga-associated bacteria and emended description of the genus Lacinutrix. INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY. 2008 Dec 1;58(12):2694–8.

Nelson DM, Tréguer P, Brzezinski MA, Leynaert A, Quéguiner B. Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem Cycles. 1995 Sep 1;9(3):359–72.

Oz A, Sabehi G, Koblizek M, Massana R, Beja O. Roseobacter-Like Bacteria in Red and Mediterranean Sea Aerobic Anoxygenic Photosynthetic Populations. Applied and Environmental Microbiology. 2005 Jan 6;71(1):344–53.

Pace NR.A Molecular View of Microbial Diversity and the Biosphere.Science. 1997 May 2;276(5313):734–40.

Palenik B, Ren Q, Dupont CL, Myers GS, Heidelberg JF, Badger JH, et al. Genome sequence of Synechococcus CC9311: insights into adaptation to a coastal environment. Proceedings of the National Academy of Sciences. 2006;103(36):13555–9.

Parsley LC, Linneman J, Goode AM, Becklund K, George I, Goodman RM, et al. Polyketide synthase pathways identified from a metagenomic library are derived from soil Acidobacteria: Polyketide synthase pathways from soil Acidobacteria. FEMS Microbiology Ecology. 2011 Oct;78(1):176–87.

Pereira GC, Coutinho R, Ebecken NFF (2008) Data mining for environmental analysis and diagnostic: a case of upwelling ecosystem of Arraial do Cabo. Braz J Oceanogr 56: 1–12.

Piganeau G, Moreau H. Screening the Sargasso Sea metagenome for data to investigate genome evolution in Ostreococcus (Prasinophyceae, Chlorophyta). Gene. 2007 Dec;406(1-2):184–90.

Preston C, Harris A, Ryan JP, Roman B, Marin R, Jensen S et al. (2011). Application of quantitative PCR on a coastal mooring. PLOS One 6: e22522.

Rappé, Michael S.; Connon, Stephanie A.; Vergin, Kevin L.; Giovannoni SJ.Cultivation of the ubiquitous SAR11 marine bacterioplankton clade.Nature [Internet]. 2002;418(6898).

Raymond J, Blankenship RE. The evolutionary development of the protein complement of Photosystem 2. Biochimica et Biophysica Acta (BBA) - Bioenergetics. 2004 Apr;1655:133–9.

Reisch CR, Crabb WM, Gifford SM, Teng Q, Stoudemayer MJ, Moran MA, et al. Metabolism of dimethylsulphoniopropionate by Ruegeria pomeroyi DSS-3. Mol Microbiol. 2013 Aug;89(4):774–91.

Riedlinger J, Reicke A, Zahner H, Krismer B, Bull AT, Maldonado LA, et al. Abyssomicins, inhibitors of the para-aminobenzoic acid pathway produced by the marine Verrucosispora strain AB-18-032. J Antibiot (Tokyo). 2004 Apr;57(4):271–9.

Ritchie AE, Johnson ZI. Abundance and Genetic Diversity of Aerobic Anoxygenic Phototrophic Bacteria of Coastal Regions of the Pacific Ocean.Applied and Environmental Microbiology. 2012 Feb 3;78(8):2858–66.

Rodrigues RR, Lorenzzetti JA (2001) A numerical study of the effects of bottom topography and coastline geometry on the Southeast Brazilian coastal upwelling. Cont Shelf Res 21: 371–394.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biology. 2007;5(3):e77.

Rye R, Holland HD. Paleosols and the evolution of atmospheric oxygen: a critical review. Am J Sci. 1998 Oct;298(8):621–72.

Schirmer A, Gadkari R, Reeves CD, Ibrahim F, DeLong EF, Hutchinson CR. Metagenomic Analysis Reveals Diverse Polyketide Synthase Gene Clusters in Microorganisms Associated with the Marine Sponge Discodermia dissoluta. Applied and Environmental Microbiology. 2005 Aug 5;71(8):4840–9.

Schneemann I, Nagel K, Kajahn I, Labes A, Wiese J, et al. (2010) Comprehensive Investigation of Marine Actinobacteria Associated with the Sponge Halichondria panicea. Applied and Environmental Microbiology 76: 3702–3714. doi:10.1128/AEM.00780-10.

Schwalbach MS, Fuhrman JA.Wide-ranging abundances of aerobic anoxygenic phototrophic bacteria in the world ocean revealed by epifluorescence microscopy and quantitative PCR.Limnology and oceanography. 2005;50(2):620–8.

Schwarzer D, Marahiel MA. Multimodular biocatalysts for natural product assembly.Naturwissenschaften. 2001 Apr 27;88(3):93–101.

Shen B. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms.Current Opinion in Chemical Biology. 2003 Apr;7(2):285–95.

Shiba T, Simidu U, Taga N. Distribution of aerobic bacteria which contain bacteriochlorophyll a. Applied and environmental microbiology. 1979;38(1):43–5.

Shulse CN, Allen EE. Widespread occurrence of secondary lipid biosynthesis potential in microbial lineages.PLoS One. 2011;6(5):e20146.

Singh SB, Barrett JF. Empirical antibacterial drug discovery—Foundation in natural products.Biochemical Pharmacology. 2006 Mar 30;71(7):1006–15.

Silva-Stenico ME, Silva CSP, Lorenzi AS, Shishido TK, Etchegaray A, Lira SP, et al. Non-ribosomal peptides produced by Brazilian cyanobacterial isolates with antimicrobial activity. Microbiol Res. 2011 Mar 20;166(3):161–75.

Slightom RN, Buchan A (2009) Surface colonization by marine roseobacters: integrating genotype and phenotype. Appl Environ Microbiol 75: 6027–6037.

Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, et al. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. Nucleic Acids Research. 2010 Nov 2;39(Database):D546–D551.

Swingley WD, Sadekar S, Mastrian SD, Matthies HJ, Hao J, Ramos H, et al. The Complete Genome Sequence of Roseobacter denitrificans Reveals a Mixotrophic Rather than Photosynthetic Metabolism. Journal of Bacteriology. 2007 Feb 1;189(3):683–90.

Swingley W, Blankenship R, Raymond J. Evolutionary Relationships Among Purple Photosynthetic Bacteria and the Origin of Proteobacterial Photosynthetic Systems. In: Hunter CN, Daldal F, Thurnauer M, Beatty JT, editors. The Purple Phototrophic Bacteria [Internet].Springer Netherlands; 2009. p. 17–29. Available from: http://dx.doi.org/10.1007/978-1-4020-8815-5_2

Thomas F, Hehemann J-H, Rebuffet E, Czjzek M, Michel G. Environmental and Gut Bacteroidetes: The Food Connection. Frontiers in Microbiology [Internet]. 2011 [cited 2014 Feb 26];2. Available from: http://www.frontiersin.org/Journal/10.3389/fmicb.2011.00093/full

Tillotson GS, Theriault N.New and alternative approaches to tackling antibiotic resistance. F1000prime reports [Internet]. 2013 [cited 2014 Jan 11];5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3854692/

Trindade-Silva AE, Rua CPJ, Andrade BGN, Vicente ACP, Silva GGZ, Berlinck RGS, et al. Polyketide Synthase Gene Diversity within the Microbiome of the Sponge Arenosclera brasiliensis, Endemic to the Southern Atlantic Ocean. Applied and Environmental Microbiology.2013 Mar 1;79(5):1598–605.

Trindade-Silva AE, Rua C, Silva GGZ, Dutilh BE, Moreira APB, Edwards RA, et al. Taxonomic and Functional Microbial Signatures of the Endemic Marine Sponge Arenosclera brasiliensis. Badger JH, editor.PLoS ONE. 2012 Jul 2;7(7):e39905.

Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nature Reviews Genetics. 2005 Oct 11;6(11):805–14.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, et al. Comparative metagenomics of microbial communities. Science. 2005 Apr 22;308(5721):554–7.

Uchino Y, Hirata A, Yokota A, Sugiyama J. Reclassification of marine *Agrobacterium* species: Proposals of *Stappia stellulata* gen. nov., comb. nov., *Stappia aggregata* sp. nov., nom. rev., *Ruegeria atlantica* gen. nov., comb. nov., *Ruegeria gelatinovora* comb.nov., *Ruegeria algicola* comb. nov., and *Ahrensia kieliense* gen. nov., sp. nov., nom. rev. The Journal of General and Applied Microbiology. 1998;44(3):201–10.

Valentin JL (1984a) Analysis of hydrobiological parameters in yhe Cabo Frio (Brazil) upwelling. Mar Biol 82: 259–276.

Valentin JL, Monteiro-Ribas WM, Mureb MA, Pessotti E (1987) Hydrobiology in the Cabo Frio (Brazil) upwelling two-dimensional structure and variability during a wind cycle. Cont Shelf Res 7: 77–88.

Vila-Costa M, Simo R, Harada H, Gasol JM, Slezak D, Kiene RP. Dimethylsulfoniopropionate uptake by marine phytoplankton.Science. 2006 Oct 27;314(5799):652–4.

Waidner LA, Kirchman DL. Aerobic Anoxygenic Phototrophic Bacteria Attached to Particles in Turbid Waters of the Delaware and Chesapeake Estuaries. Applied and Environmental Microbiology. 2007 Apr 27;73(12):3936–44.

Watanabe A, Ebizuka Y. Unprecedented mechanism of chain length determination in fungal aromatic polyketide synthases. Chemistry & biology. 2004;11(8):1101–6.

Waterbury, J. B. & Rippka, R. (1989) in *Bergey's Manual of SystematicBacteriology*, eds. Staley, J. T., Bryant, M. P., Pfennig, N.&Holt, J. B. (Williams& Wilkins, Baltimore), Vol. 3, pp. 1728–1746

Wawrik B, Kerkhof L, Zylstra GJ, Kukor JJ. Identification of Unique Type II Polyketide Synthase Genes in Soil.Applied and Environmental Microbiology. 2005 May 3;71(5):2232–8.

Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proceedings of the National Academy of Sciences. 1998;95(12):6578–83.

Wommack KE, Bhavsar J, Ravel J. Metagenomics: Read Length Matters. Applied and Environmental Microbiology. 2008 Jan 11;74(5):1453–63.

Worden AZ, Nolan JK, Palenik B. Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. Limnology and Oceanography. 2004;49(1):168–79.

Xiong J, Bauer CE. C OMPLEX E VOLUTION OF P HOTOSYNTHESIS.Annual Review of Plant Biology. 2002 Jun;53(1):503–21.

Yew, SP, Jau MH, Yong KH, Abed RMM, Sudesh K. Morphological Studies of Synechocystis sp. UNIWG under.Polyhydroxyalkanoate Accumulating Conditions.Malaysian Journal of Microbiology. 2005; 1 :48-52.

Yurkov V, Csotonyi J. New Light on Aerobic Anoxygenic Phototrophs. In: Hunter CN, Daldal F, Thurnauer M, Beatty JT, editors. The Purple Phototrophic Bacteria [Internet].Springer Netherlands; 2009. p. 31–55. Available from: http://dx.doi.org/10.1007/978-1-4020-8815-5_3

Yurkov VV, Beatty JT. Aerobic anoxygenic phototrophic bacteria. Microbiology and Molecular Biology Reviews. 1998;62(3):695–724.

Yurkov V, Hughes E. Chapter Eleven - Genes Associated with the Peculiar Phenotypes of the Aerobic Anoxygenic Phototrophs. In: J. Thomas Beatty, editor. Advances in Botanical Research [Internet].Academic Press; 2013. p. 327–58.
Yutin N, Béjà O. Putative novel photosynthetic reaction centre organizations in marine aerobic anoxygenic photosynthetic bacteria: insights from metagenomics and environmental genomics: Novel photosynthetic reaction centres genes in marine AAnPs. Environmental Microbiology. 2005 Jul 15;7(12):2027–33.

Yutin N, Suzuki MT, Teeling H, Weber M, Venter JC, Rusch DB, et al. 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. Environmental Microbiology. 2007 Jun;9(6):1464–75.

Zappia G, Menendez P, Delle Monache G, Misiti D, Nevola L, Botta B. The contribution of oxazolidinone frame to the biological activity of pharmaceutical drugs and natural products. Mini reviews in medicinal chemistry. 2007;7(4):389–409.

Zarraonaindia I, Smith DP, Gilbert JA. Beyond the genome: community-level analysis of the microbial world. Biology & Philosophy. 2012 Dec 15;28(2):261–82.

Zhang X-Y, Xie B-B, Qin Q-L, Liu A, Chen X-L, Zhou B-C, et al. Draft Genome Sequence of Strain P7-3-5, a New Flavobacteriaceae Bacterium Isolated from Intertidal Sand. Journal of Bacteriology. 2012 Nov 9;194(23):6632–6632.

Zhang Y, Jiao N. Dynamics of aerobic anoxygenic phototrophic bacteria in the East China Sea: AAPB in the East China Sea. FEMS Microbiology Ecology. 2007 Sep;61(3):459–69.

Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Research. 2010 Jul 1;38(12):e132–e132.

Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. de Crécy-Lagard V, editor. PLoS ONE. 2012 Mar 29;7(3):e34064.

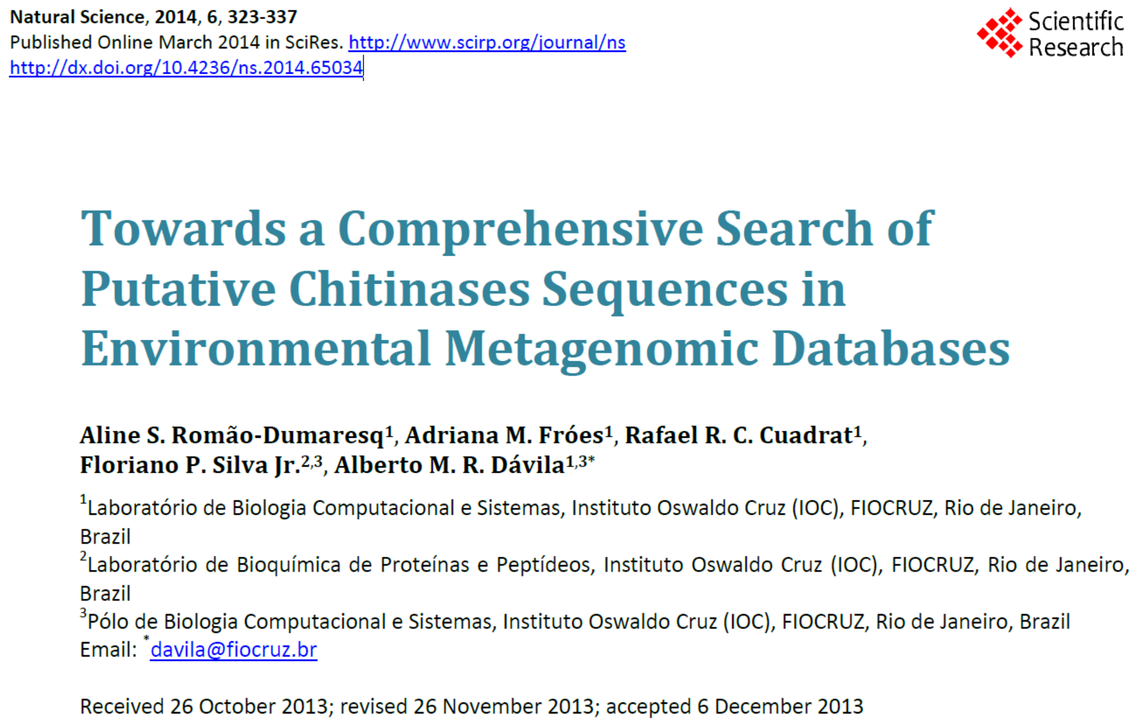
The 10 × '20 Initiative: Pursuing a Global Commitment to Develop 10 New Antibacterial Drugs by 2020. Clinical Infectious Diseases. 2010 Apr 15;50(8):1081–3.

9 ANEXOS

Trabalhos relacionados com a tese

Artigo: Towards a Comprehensive Search of Putative Chitinases Sequences in Environmental Metagenomic Databases

Este trabalho foi publicado em coautoria em março de 2014. O pipeline desenvolvido para esta tese foi aplicado neste trabalho para a detecção de Quitinases em metagenomas públicos.



Natural Science, **2014**, **6**, **323-337**
Published Online March 2014 in SciRes. http://www.scirp.org/journal/ns
http://dx.doi.org/10.4236/ns.2014.65034

Scientific Research

Towards a Comprehensive Search of Putative Chitinases Sequences in Environmental Metagenomic Databases

Aline S. Romão-Dumaresq[1], Adriana M. Fróes[1], Rafael R. C. Cuadrat[1], Floriano P. Silva Jr.[2,3], Alberto M. R. Dávila[1,3*]

[1]Laboratório de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz (IOC), FIOCRUZ, Rio de Janeiro, Brazil
[2]Laboratório de Bioquímica de Proteínas e Peptídeos, Instituto Oswaldo Cruz (IOC), FIOCRUZ, Rio de Janeiro, Brazil
[3]Pólo de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz (IOC), FIOCRUZ, Rio de Janeiro, Brazil
Email: *davila@fiocruz.br

Received 26 October 2013; revised 26 November 2013; accepted 6 December 2013

Resumo estendido: “Exploring the diversity of Polyketide synthases (PKS) and Non-ribosomal peptide synthetases (NRPS) in aquatic environments”

Publicado em “Digital Proceedings of Brazilian Symposium on Bioinformatics (2012:Campo Grande-MS)” ISSN 2316-1248, disponível em http://bsb2012.facom.ufms.br/manager/titan.php?target=openFile&fileId=525.

Exploring the diversity of Polyketide synthases (PKS) and Non-ribosomal peptide synthetases (NRPS) in aquatic environments

Rafael Cuadrat[1,2], Alberto Dávila[1]
rafaelcuadrat@ioc.fiocruz.br,davila@fiocruz.br

[1] Computational and Systems Biology Laboratory, Oswaldo Cruz Institute, Rio de Janeiro, Brasil
[2] Computational and Systems Biology Laboratory - IOC

Abstract. According to the literature marine environments have high microbial diversity and tapproximately 99% of the organisms present in these environ-ments are not cultivable. This rich biodiversity is a major untapped potential of biotechnology, then it is possible to discover new enzymes as PKS and NRPSs in these environments. Therefore, it is necessary to make use of metagenomics approaches, in order to access the genetic material of the organisms in the environment without the need for cultivation. The new high throughput techniques of DNA sequencing allow to obtain data on genes and metabolic pathways pre-sent in these species. This study has aimed to explore the diversity of PKS and NRPS in aquatic environments through the screening of public metagenomes available in IMG/M using hidden markov models (pHMM) and evaluate the potential of aquatic environments for the search of new genes of these families.

Keywords: PKS, NRPS, Metagenomic, environmental

1 Introduction

There are approximately 3.67 x 1030 microbial cells [1] in marine environments. It is estimated that the abundance of bacteria is of up to 106 cells per milliliter of water in the sea (pelagic zone), representing most of oceanic biomass [2]. This huge biodiversity has great potential, as its study allows the discovery of new enzymes of interest for biotechnology industry. Several groups of marine micro-organisms are known for their high production of secondary metabolites, such as cyanobacteria. Marine cyanobacteria are a rich source of complex bioactive secondary metabolites which derive from mixed biosynthetic pathways [3]. Another group of marine bacteria known to be a producer of bioactive natural products are the vibrio’s (a total of 93 compounds have been isolated from Vibrionaceae [4]. Moreover, cyanobacteria pre-sent in fresh waters are known to produce toxic secondary metabolites and other types of non-ribosomal peptides [5]. Most of these metabolites are produced by two large families of enzymes (i) Polyketide synthases (PKS) and (ii) non-ribosomal peptide synthetases (NRPS) that account for many clinically important pharmaceutical products [5]. The main objective of the present study is to explore the diversity of PKS and NRPS in aquatic environments.

2 Material and Methods

2.1 Reference Database

All curated type I PKS sequences (iterative and modular) were obtained from MAPSIDB (http://gate.smallsoft.co.kr:8080/pks/mapsidb) in fasta format and the domains KS, AT and ACP were extracted using the fastacmd program (BLAST 2.2.21 package). The orthologs groups K05551 and K05552 (containing sequences of KS II alfa and beta subunit respectively) were downloaded from KEGG (http://www.genome.jp/kegg/). The protein sequences of Adenilation (A) and Con-densation (C) domains from NRPS were obtained from NRPSDB (http://linux1.nii.res.in/~zeeshan/webpages/home.html).

2.2 Metagenomes

The protein sequences (translated ORFS) of metagenomes were obtained from IMG/M (http://img.jgi.doe.gov/cgi-bin/m/main.cgi).

2.3 Screening PKS and NRPS in metagenomes

A pipeline to screen PKS and NRPS in public metagenomes was built in RUBY (http://www.ruby-lang.org/). The first step from the pipeline is the alignment of each domain using MAFFT v6.717b. Then, the multiple alignments are used to generate hidden markov models (pHMM) using hmmbuild from HMMER 3.0 package. These pHMMs are then used to search for enzymes domains in public metagenomes using hmmsearch (from HMMER 3 package). A parser is then used to (i) generate CSV tables, (ii) extract specific informations from the HMMER results, and (iii) to generate fasta files from the hits found by the pHMM used. The number of hits found with HMMER is normalized by calculating the percentage of hits obtained in relation to the total of the metagenomic sequences. The KS domain sequences obtained by

pHMM from metagenomes are filtered by size (> 150 aminacids) and by the presence of the catalytic site. Subsequently, the environmental KS domain sequences are aligned with the reference sequences (protein fasta sequences used to build the multiple alignment) and outgroup (fabB, fabF and fabH), then the alignments are trimmed and converted to Phylip format using Trimal 1.2 (with –automatic1 parameter). The alignments are then submitted to RaxML version 7.2.2 to generate phylogenetic trees whit bootstrap support (100 replicates) using the WAG model. Finally, the newick (nwk) files generated by RaxML are parsed with a BIORUBY link/version script to classify the domains.

3 Results and discussion

3.1 Searching for profiles

A total of 52801 environmental sequences similar to the 4 profiles of KS domains (hmmsearch hits between pHMM and environmental sequences) were obtained. Using the AT and ACP pHMM, a total of 11468 and 4421 hits were obtained, respectively. Using NRPS domains pHMM, 50750 hits were obtained. The table 1 shows the hits obtained by each pHMM used.

Table 1. Number of hits obtained using each pHMM

Profile (pHMM)	*Number of hits*
Modular KS	12403
Iterative KS	12990
Type II KS (alfa subunit)	14596
Type II KS (beta subunit)	12812
Modular AT	5959
Iterative AT	5509
Modular ACP	2286
Iterative ACP	1743
Type II ACP	392
NRPS A	45589
NRPS C	5161

The number of hits of the NRPS A domain is 8.9 times higher than the domain C. As the domains A and C are essential for minimum functionaly of NRPS, it was expected A similar number of hits for the two domains, however the Phmm built with sequences of domain A may be more sensitive than the one from C domain pHMM. Among the number of hits obtained with the pHMM of type II PKS, a discrepancy is also observed, the KS alpha and beta domains are much more abundant than the ACP do-main. The reason for this difference may be the same as discussed for the case of NRPS or by the fact that the KS domains are the most conserved domain in PKS, in fact, because of this conservation, it is the most used region for PCR primers design [6].

The figure 1 shows the distribution of hits between the types of aquatic environment.

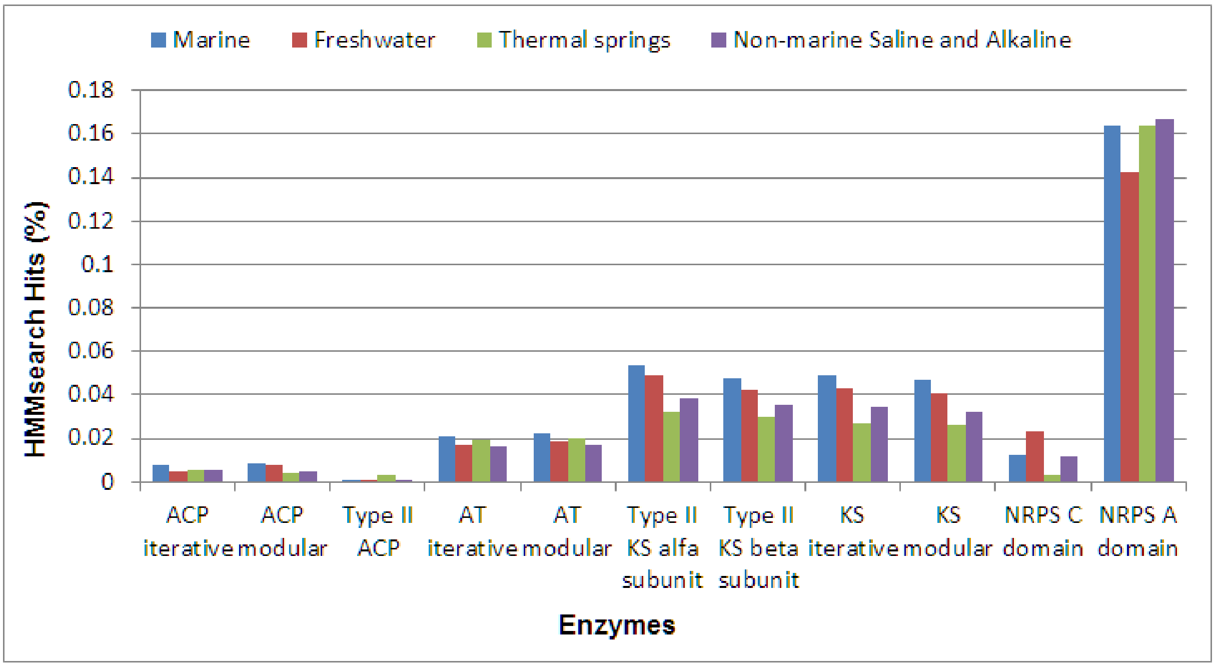


Fig. 1. Distribuition of pHMM hits between the types of aquatic environment.

In most cases, the environment with greater abundance of hits is the ocean, but the difference between the types of aquatic environments is little.

3.2 Classifying environmental KS domains with phylogenetic trees

Due to the similarities between fatty acid synthases (FAS) and PKSs, the pHMM approach is not sufficient to determine whether a sequence is FAS or PKS [7]. To overcome this limitation, one of the ways to classify the sequences is through phylog-eny. Trees were constructed through the ML method for the hits obtained with type I KS domains (modular and iterative) and TYPE II pHMMS in 22 selected aquatic environments. From a total of 796 and 846 hits obtained from aquatic metagenomes using KS modular and iterative pHMM, respectively. From these, only 56 and 52 were selected for phylogenetic analysis because they have more than 150 aminoacids and cysteine active site. The sequences (hits) related to type II KS were filtered only by size (larger than 150 amino acids) and from 851, only 159 were selected.
By this analysis it was possible to determine which environmental sequences are potential true KS, and also classifies them according to their type and modularity. A total of 98 sequences were classified as true KS domain. The figure 2 shows the clas-sification of sequences obtained with each pHMM.

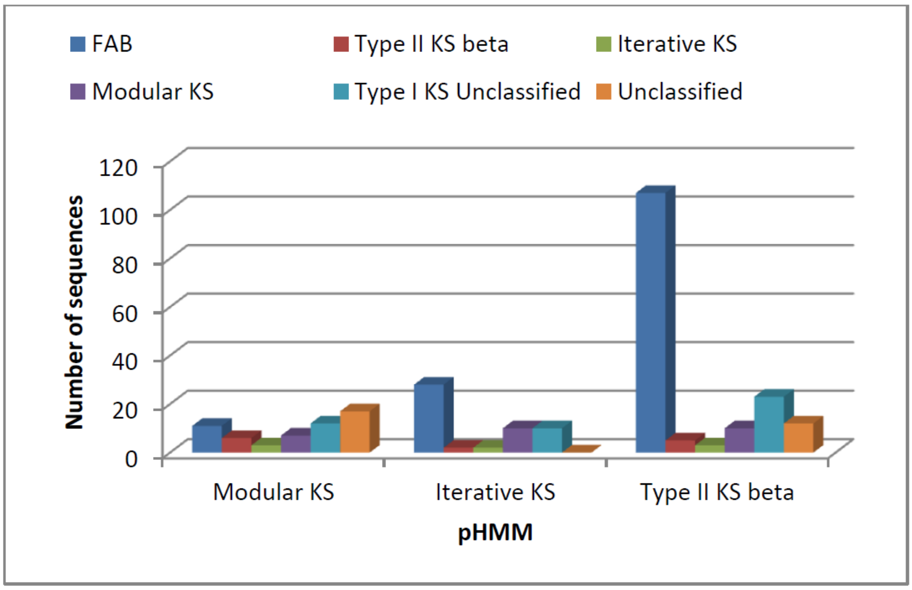


Fig. 2. Classification of hits obtained by each KS pHMM using ML phylogenetic trees with reference sequences. The legend shows the classification obtained by ML as: type I modular KS, type I iterative KS, type I KS unclassified (not defined as modular or iterative), type II KS, FAB domains and unclassified sequences (do not groups with any reference sequence).

4 Conclusion

By using the pipeline developed, it was possible to screen metagenomes for en-zymes of biotechnological interest. The use of pHMMs is a fast and sensitive way to obtain the homologous sequences of interest to the study, but due to sensitivity of the approach, it is necessary to classify the hits obtained using phylogenetic trees. This classification is generally slower and requires manual verification in most cases. The pipeline showed in this study worked well by classifying the sequences of interest automatically from a NWK file. In this study it was possible to confirm the potential of aquatic metagenomes to uncover new PKS and NRPS enzymes

5 References

1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proceedings of the National. Academy of Sciences U S A 1998; 95(12):6578-6583.

2. Azam F. Oceanography: Microbial Control of Oceanic Carbon Flux: The Plot Thickens. Science 1998; 280:694-696

3. Joshawna K. Nunnery, Emily Mevers, William H. Gerwick1 Biologically active secondary metabolites from marine Cyanobacteria Curr Opin Biotechnol. 2010 December ; 21(6): 787–793

4. Ken-ichi HARADA, Production of Secondary Metabolites by Freshwater Cyanobacteria *Chem. Pharm. Bull.* 52(8) 889—899 (2004)

5. Ayuso-Sacido a, Genilloud O. New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: detection and distribution of these biosynthetic gene sequences in major taxonomic groups. Microbial ecology. 2005 Jan;49(1):10–24.

6. Foerstner KU, Doerks T, Creevey CJ, Doerks A, Bork P. A computational screen for type I polyketide synthases in metagenomics shotgun data. PloS one, 2008 Jan;3(10):e3515.

Trabalhos não relacionados com a tese

Artigo: "An Orthology-Based Analysis of Pathogenic Protozoa Impacting Global Health: An Improved Comparative Genomics Approach with Prokaryotes and Model Eukaryote Orthologs."

Este artigo foi publicado na revista OMICS: A Journal of Integrative Biology. DOI: 10.1089/omi.2013.0172