Universitat
de les Illes Balears

DOCTORAL THESIS

2017

PROTEOGENOMIC AND PHYSIOLOGIC CHARACTERIZATION OF NEW MEMBERS OF ROSEOBACTER LINEAGE FROM MALLORCA HARBORS

Maria Mas Lladó

Universitat de les Illes Balears

DOCTORAL THESIS

2017

Doctoral Program of "Microbiologia Ambiental i Biotecnologia"

PROTEOGENOMIC AND PHYSIOLOGIC CHARACTERIZATION OF NEW MEMBERS OF ROSEOBACTER LINEAGE FROM MALLORCA HARBORS

Maria Mas Lladó

Thesis Supervisor: Dr. Rafael Bosch

Thesis Supervisor: Dra. Balbina Nogales

Doctor by the Universitat de les Illes Balears

A la meva família.

Especialment, als meus pares.

AGRAÏMENTS

A la fi puc donar aquesta tesis per acabada! Ha arribat el moment de tancar una etapa, un capítol de la meva vida i de seguir i millorar amb el que he començat. Per això, ha arribat també el moment d'agrair a tots aquells que d'una manera o d'una altra han contribuït a que aquesta etapa fos exitosa.

En primer lloc, vull donar les gràcies al Dr. Rafael Bosch, per donar-me l'oportunitat de fer els meus estudis doctorals amb ell. Gràcies per haver-me ensenyat tantes coses del món de la microbiologia i de la bioinformàtica, i gràcies també per tenir sempre el teu bon humor i caràcter que et caracteritza. També agrair a la Dra. Balbina Nogales els seus ensenyaments i consells. Gràcies per ser sempre tan perfeccionista ja que m'ha ajudat a ser més meticulosa. En resum, gràcies a tots dos per ensenyar-me el que és fer recerca i a madurar científicament.

Gràcies també als altres professors de la UIB relacionats amb el món de la microbiologia. Principalment gràcies al Dr. Jordi Lalucat, a la Dra. Elena García-Valdés, al Dr. Toni Bennàsar i al Dr. Sebastià Albertí, que han sabut donar-me consells i fer-me suggeriments quan ha estat oportú i sempre amb l'objectiu de millorar en la meva recerca. Especialment, vull destacar el Dr. Toni Bennàsar i la Dra. Elena García-Valdés, que van ser professors meus quan vaig cursar el màster de microbiologia avançada i varen fer que la meva curiositat envers la microbiologia cresqués.

En referència als amics que han fet feina amb jo al laboratori i amb els que he dinat cada dia durant aquests anys, mil gràcies a tots per l'ajuda que m'heu donat sempre. No hagués pogut tenir mai millors companys de feina que vosaltres. Als del "cuartito", que són amb els que sempre he passat més temps...gràcies Toni Busquets i Arantxa Peña per donar-me consells, solucionar problemes, brindar-me moments de desconnexió amb discussions sobre temes intranscendents, escoltar-me quan ho he necessitat...heu estat un gran suport immillorable durant aquests 4 anys. A les altres dues del "cuartito", heu estat les meves companyes de viatge, sempre fent feina colze amb colze, sempre seguint les passes que l'altre feia. Gràcies Cati Alejandro, la meva companya de congressos, per donar-me una mà quan ho he necessitat. I gràcies Bel Brunet, per ensenyar-me tot el que m'has ensenyat i per compartir penes i alegries. Als del laboratori, gràcies Magdalena Mulet, Claudia Prince, Cristina Ramon, David Sánchez i Dani Jaén per escoltar-me quan em desesperava amb els experiments i per ajudar-me sempre que vos ha estat possible. També vull donar les gràcies a na Lady Suárez, que em va ensenyar a treballar amb els aparells d'HPLC i de la que rebia ajuda quan la necessitava. Agrair també el suport de na Marga Gomila, sempre disposada a escoltar, a respondre dubtes i a aportar solucions. Als tècnics de laboratori (Xisco, Àngel i Guillem) per facilitar la meva feina. Gràcies a na Mati pel seu bon humor matiner. I també a na Trinitat, dels Serveis Científico-Tècnics per la seva bona disponibilitat. Pens que els companys de feina són essencials perquè els projectes puguin sortir endavant, i amb vosaltres tot ha estat més fàcil.

Vull agrair a en Joseph les seves ensenyances en el món de la microbiologia i la proteòmica. Gràcies a ell vaig tenir l'oportunitat de conèixer un nou laboratori i una altre forma una mica diferent de treballar de la que estava acostumada. A banda de la part professional, li vull donar les gràcies a ell, a na Mar i a la seva família per la seva hospitalitat i amabilitat durant la meva estància a Coventry.

Thanks to Sue, Alex and Cleidiane for teaching and helping me in the field of proteomics. I also want to thank my friends at Warwick University, Despoina, Amandeep, Satya and Matthias. With your friendship I could fully enjoy my stay. Además quiero agradecer la amistad y ayuda que me brindaron mis amigos españoles que conocí en Inglaterra, sobre todo a Miriam, Kike y Alfredo. Con vosotros, en cierta parte, era como estar como en casa. I sincerely hope to see you all again.

Durant aquesta etapa (i sempre) han estat molt importants les meves amigues. Encara que moltes vegades no entenien molt bé el que feia o quin tipus d'experiments desenvolupava, m'ha ajudat molt el seu suport per desconnectar de la feina. Gràcies Marga Fullana, Maria Francisca Roca, Carme Ferrari i Catalina Maria Mas per la vostra amistat. Gràcies també a n'en Jaume Llorens i na Marga Barceló, per ser grans amics. També vull agrair als amics llucmajorers de la meva parella, que ara també consider els meus, el seu interès pel que estudiava i pels seus ànims.

Infinites gràcies a mon pare i a ma mare. Tot l'agraïment que sent cap a vosaltres no es pot expressar en paraules. Gràcies per ser sempre al meu costat i per animar-me a tirar cap endavant. Sempre m'heu donat l'empenta que he necessitat a l'hora de prendre decisions, ja ho sabeu. Donar-vos les gràcies per la vostra comprensió i ajuda en tot moment. Gràcies també a la meva germana, sobretot per quedar a Mallorca amb nosaltres; has estat un gran suport sempre. Gràcies als meus padrins, que sempre m'heu donat molts d'ànims per continuar cap endavant. Gràcies a la meva parella, en Joan Miquel, que a pesar d'estar a quilòmetres de distància hem pogut fer-nos costat l'un a l'altre quan ho hem necessitat. Igual que tu, desig que no haguem de tornar a separar-nos mai i puguem emprendre el camí junts d'aprop. A tots vosaltres, gràcies mils, sou els pilars de la meva vida.

Finalment, donar les gràcies a na Sebastiana i en Biel per preocupar-se per jo i donar-me coratge. Agraeixo també l'ajuda i costat que he rebut dels meus estimats cunyats, na Xisca i en Pere Guillem; gràcies per tot.

Per acabar, vull donar les gràcies a les diferents entitats que han aportat els fons necessaris per fer possible la realització d'aquesta tesis. En primer lloc, a la Universitat de les Illes Balears per concedir-me una beca de postgrau l'any 2012. Posteriorment al "Ministerio de Educación, Cultura y Deporte" per concedir-me la beca de Formació de Personal Universitari (2012-2016) (referència FPU12/01449). També al projecte CTM2011-24886 (del "Ministerio de Economía, Indústria y Competitividad"), amb cofinançament FEDER. Finalment, al "European Molecular

Biology Organization" per concedir-me una beca per a poder fer una estada a "The University of Warwick" on vaig fer els anàlisis de proteòmica.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAnP, Aerobic anoxygenic phototrophy

ABC, ATP-binding cassette system

AHL, Acylated homoserine lactone

ANIb, Average nucleotide identity based on BLAST

ANIm, Average nucleotide identity based on MUMer

BLAST, Basic local alignment search tool

CDS, Coding DNA sequence

CHMS, 5-carboxymethyl-2-hydroxymuconic semialdehyde

CoA, Coenzyme A

COG, Cluster of orthologous group

DDH, DNA-DNA Hybridization

DMS, Dimethylsulfide

DMSP, Dimethylsulfoniopropionate

DNA, Deoxyribonucleic acid

DTT, Dithiothreitol

ECE, Extrachromosomal element

GGDC, Genome-to-genome distance comparison

GTA, Gene transfer agent

HGT, Horizontal gene transfer

IAA, Iodoacetamide

IS, Insertion sequence

KAAS, KEGG automatic annotation server

KEGG, Kyoto encyclopedia of genes and genomes

MALDI-TOF, Matrix-assisted laser desorption/ionization time-of-flight

MB, Marine broth

MeSH, Methanethiol

MMM, Mineral marine medium

MMMb, Mineral marine medium supplemented with yeast extract

MMPA, 3-methyl-mercaptopropionate

MPA, 3-mercaptopropionate

MS/MS, Tandem mass spectrometry

NanoLC-ESI-MS/MS, Nano-scale liquid chromatographic tandem mass spectrometry

NCBI, National center for biotechnology information

NRP, Nonribosomal peptides

OD_{600}, Optical density at a wavelength of 600 nm

PAH, Polycyclic aromatic hydrocarbon

PCA, Principal component analysis

PCR, Polymerase chain reaction

PGAAP, Prokaryotic genomes automatic annotation pipeline

PK, Polyketides

PTFE, Polytetrafluoroethylene

PVDF, Polyvinylidene fluoride

QS, *Quorum sensing*

RcGTA, GTA of *Rhodobacter capsulatus*

RNA, Ribonucleic acid

T4SS, Type IV secretion system

TCA, Tricarboxylic acid cycle

TCDB, Transporter classification database

TnpA, Transposase

TRAP-T, Tripartite ATP-independent periplasmic transporter system

ABSTRACT

The Roseobacter lineage is an important component of marine bacterioplankton. One of the characteristics of roseobacters is their ability to degrade aromatic compounds, but this evidence is supported mainly by genome data from isolates mostly obtained from non-polluted environments. A collection of Roseobacter isolates from harbors and marinas around Mallorca Island was obtained previously. This was a unique collection of isolates that were obtained from polluted environments, same sampling time, and same geographic area. We hypothesized that these bacteria would differ from other roseobacters in their ability to degrade aromatic compounds and in the degree of genetic exchange by horizontal gene transfer (HGT). The genomes of nine of these isolates have been analyzed in this thesis. Genomic data was used to affiliate taxonomically the isolates and to establish their phylogeny within the Roseobacter lineage. A core genome phylogenetic analysis based on concatenated sequences of 114 proteins encoded in single copy core genes was done (ninety-six genomes). A subset of fourteen of these protein sequences was shown to reproduce well the phylogeny of the lineage and it is proposed as a simplified method for the affiliation of new Roseobacter isolates. The isolates showed typical characteristics (genomic and metabolic) of cultured roseobacters, also with respect to potential for aromatic compound degradation. Seven of them were predicted to possess alkane monoxygenase genes, and three were able to grow on diesel oil as sole carbon and energy source. Proteomic analysis was used to study the response of the isolates to diesel oil exposure. Alkane monooxygenase was detected in three isolates, including two of the isolates able to grow on diesel oil. Proteins involved in maintaining membrane stability, response to oxidative stress, chaperones and membrane transport were detected in cultures exposed to diesel oil, although there were differences in the response of the particular isolates. These results indicated that roseobacters had the basic mechanisms described to respond to hydrocarbons. Finally, three mechanisms of HGT (gene transfer agents, plasmid replicons and transposable elements) were analyzed in a dataset of 96 genomes. At least one of the HGT mechanisms was detected in all genomes except in isolate HTCC2255, considered an example of streamlined genome similar to uncultured representatives of the lineage. Transposases and plasmid replicons were the most widespread HGT mechanisms. There were differences in the number, size and type of plasmid replicons in genomes from the different phylogenomic groups. Phylogeny of RepC replicase of RepABC plasmids evidenced that this type of plasmids have been exchanged promiscuously among roseobacters. Finally, the analysis of shared transposases showed a high degree of connection in the global network but no clear patterns in relation to phylogenomic relatedness, habitat or geographic place of isolation. These findings support the high genetic diversity and genomic plasticity of cultured representatives of the lineage, and a possible difference with uncultured ones.

RESUMEN

El linaje Roseobacter es un componente importante del bacterioplancton marino. Una de sus características es su capacidad para la degradación de compuestos aromáticos, pero esta evidencia se sustenta principalmente en datos genómicos de aislamientos obtenidos generalmente de ambientes no contaminados. En estudios anteriores se obtuvo una colección de aislados de este linaje de puertos deportivos de la isla de Mallorca. Esta colección de aislados es única porque se obtuvieron de ambientes contaminados, en un mismo período de tiempo y en una misma área geográfica. Hipotetizamos que estas bacterias se diferenciarían de los demás roseobacters en su capacidad para la degradación de compuestos aromáticos y en el grado de intercambio genético por transferencia genética horizontal (HGT). En esta tesis se han analizado los genomas de nueve de estos aislados. La información genómica se utilizó para afiliarlos taxonómicamente y para establecer su filogenia dentro del linaje Roseobacter. Se hizo el análisis filogenético del core genoma basado en la concatenación de 114 secuencias de proteínas codificadas en copia única (noventa y seis genomas). Además se mostró que un subconjunto de catorce de las secuencias de estas proteínas reproducía de forma fiel la filogenia del linaje y se propuso como un método simplificado para afiliar nuevos aislados del linaje. Los aislados mostraron las características (genómicas y metabólicas) típicas de los miembros cultivables, también en lo que respecta a la degradación de compuestos aromáticos. En siete de ellos se predijo la presencia de genes para alcano monooxigenasas en los genomas, y tres de ellos fueron capaces de crecer a expensas de diésel como única fuente de carbono y energía. Para estudiar la respuesta a diésel de los aislados se utilizó un análisis proteómico. En tres de los ailados se detectaron alcano monooxigenasas, incluyendo dos de los aislados capaces de crecer a expensas de diésel. Además, en los cultivos expuestos a diésel se detectaron proteínas implicadas en el mantenimiento de la estabilidad de la membrana, la respuesta al estrés oxidativo, chaperonas y el transporte de membrana, aunque hubo diferencias en la respuesta particular de cada aislado. Estos resultados indicaron que los rosobacters tenían los mecanismos básicos descritos para responder a los hidrocarburos. Finalmente, se analizaron tres mecanismos de HGT (agentes de transferencia de genes, replicones de plásmidos y elementos transponibles) en 96 genomas. En todos los genomas excepto el del aislado HTCC2255, considerado un ejemplo de genoma simplificado similar al de los representantes no cultivables del linaje, se detectó al menos uno de los mecanismos de HGT. Se observaron diferencias en el número, tamaño y tipo de replicones de plásmidos en los genomas de los diferentes grupos filogenómicos. La filogenia de la replicasa RepC de los plásmidos RepABC evidenció que este tipo de plásmidos habían sido intercambiados promiscuamente entre roseobacters. Finalmente, el análisis de transposasas compartidas mostró un alto grado de conexión en la red global, pero no se detectaron patrones claros en relación con la proximidad filogenómica, el hábitat o el lugar geográfico de aislamiento. Estos resultados respaldan la gran diversidad genética y plasticidad genómica de los representantes cultivables del linaje y una posible diferencia con los no cultivables.

RESUM

El llinatge Roseobacter és un component important del bacterioplàncton marí. Una de les característiques dels roseobacters és la seva capacitat per a la degradació de compostos aromàtics, però aquesta evidència es recolza principalment en dades genòmiques d'aïllaments obtinguts principalment d'ambients no contaminats. En estudis anteriors es va obtenir una col·lecció d'aïllats d'aquest llinatge de ports esportius de l'illa de Mallorcas. Aquesta col·lecció d'aïllats és única perquè s'havien obtingut d'ambients contaminats, en un mateix període de temps i en una mateixa àrea geogràfica. Hipotetizàrem que aquests bacteris es diferenciarien dels altres roseobacters en la seva habilitat per a la degradació de compostos aromàtics i en el seu grau d'intercanvi genètic per transferència genètica horitzontal (HGT). En aquesta tesi s'han analitzat els genomes de nou d'aquests aïllats. La informació genòmica es va utilitzar per afiliar taxonòmicament els aïllats i establir la seva filogènia dins el llinatge Roseobacter. Es va fer l'anàlisi filogenètic del core genoma basat en la concatenació de 114 seqüències de proteïnes codificades en còpia única (noranta-sis genomes). A més, es va mostrar que un subconjunt de seqüències de catorze d'aquestes proteïnes reproduïa de forma fidel la filogènia del llinatge i es va proposar com un mètode simplificat per afiliar nous aïllats del llinatge. Els aïllats tenien les característiques típiques (genòmiques i metabòliques) dels aïllats cultivables del llinatge, també pel que fa a la degradació de compostos aromàtics. A set d'ells es va predir genòmicament la codificació per alcà monooxigenases, i tres d'ells varen ser capaços de créixer a partir de dièsel com a única font de carboni i energia. Per estudiar la resposta dels aïllats a l'exposició de dièsel es va fer una anàlisi proteòmica. A tres dels aïllats es varen detectar alcà monooxigenases, incloent dos dels aïllats capaços de créixer a expenses de dièsel. A més, en els cultius exposats a dièsel es varen detectar proteïnes implicades en el manteniment de l'estabilitat de la membrana, la resposta a l'estrès oxidatiu, xaperones i el transport de membranal, encara que hi va haver diferències en la resposta particular de cada aïllat. Aquests resultats varen indicar que els roseobacters tenien els mecanismes bàsics descrits per respondre als hidrocarburs. Finalment, es varen analitzar tres mecanismes de HGT (agents de transferència de gens, replicons de plasmidis i elements transposables) en 96 genomes. A tots els genomes, excepte el de l'aïllat HTCC2255, considerat un exemple de genoma simplificat similar als representants no cultivables del llinatge, es va detectar almenys un dels mecanismes de HGT. Es van observar diferències en el nombre, mida i tipus de replicons de plasmidis als genomes dels diferents grups filogenòmics. La filogènia de la replicasa RepC dels plasmidis RepABC va evidenciar que aquest tipus de plasmidis havien estat intercanviats promíscuament entre roseobacters. Finalment, l'anàlisi de transposases compartides mostrà un alt grau de connexió a la xarxa global, però no es varen detectar patrons clars en relació amb la proximitat filogenòmica, l'hàbitat o el lloc geogràfic d'aïllament. Aquests resultats recolzen la gran diversitat genètica i plasticitat genòmica dels representants cultivables del llinatge i mostren una possible diferència amb els no cultivables.

INTRODUCTION

1. The Roseobacter lineage

1.1. Phylogeny, habitat and lifestyle

The Roseobacter lineage is a diverse group within the family *Rhodobacteraceae* in the Class *Alphaproteobacteria*. The 16S rDNA phylogeny shows that it is a coherent group which shares a minimum of 89 % of identity of 16S rRNA gene sequences (Brinkhoff *et al.*, 2008; Buchan & Moran, 2005). While the Roseobacter lineage itself is strongly supported, the within-group phylogeny based on the 16S rDNA is not (Buchan & Moran, 2005). One consequence of this is that the current taxonomy of the Roseobacter group is, in some cases, in disagreement with the phylogeny (Luo *et al.*, 2012, 2013, 2014; Luo & Moran, 2014; Newton *et al.*, 2010). This non-coherent assignment of genus or species names can cause confusion when interpreting the ecology and evolutionary biology of the lineage. The 16S rDNA phylogeny has been recently validated by applying phylogenomic approaches using concatenated protein data sets consisting on conserved single-copy orthologous sequences (Luo *et al.*, 2012; Luo & Moran, 2014; Newton *et al.*, 2010). These studies have revealed that there are five phylogenomic groups within the lineage (see Figure 1I). At the moment of writing this thesis more than 150 genomes, that belong to 47 different genera, within the Roseobacter lineage have been sequenced and published in databases. The information from the sequenced genomes reflects the physiological and genetic diversity found in these organisms. So, there is a need to establish a new robust phylogeny that includes the new sequenced genomes that would allow us to know the evolutionary history of the lineage and would be a helpful tool to affiliate new isolates.

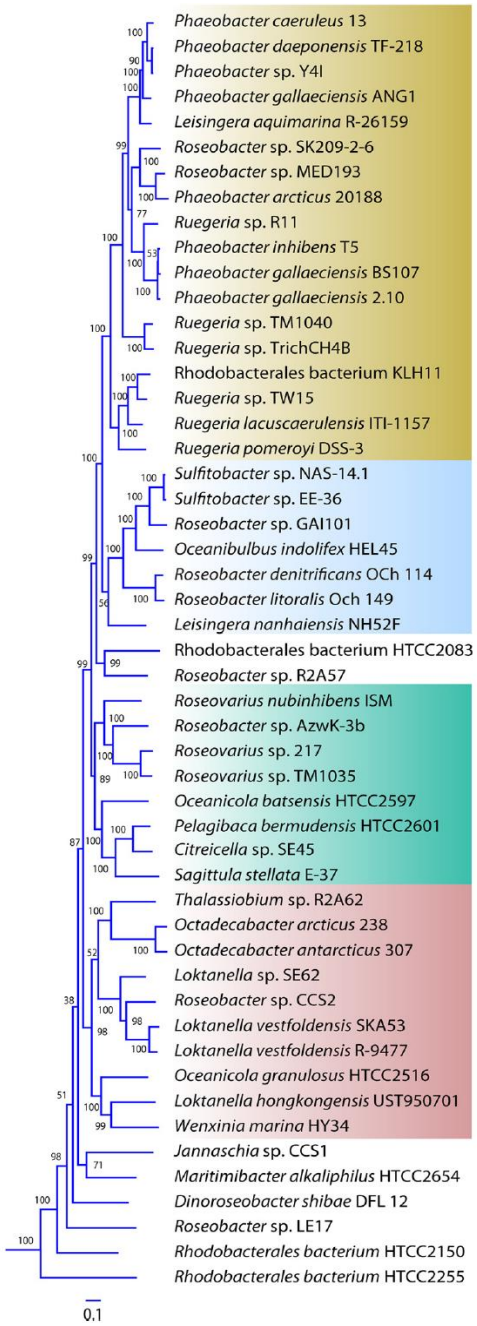

Figure 1I. Phylogenomic tree of fifty-two members of Roseobacter lineage. Colors indicate the four major phylogenomic groups of isolate genomes. The phylogeny was based on the concatenation of ~ 50 single-copy conserved protein sequences. Modified image from Luo & Moran, 2014.

Regarding their lifestyle, roseobacters are characterized for having different types: i) free living in seawater or attached to marine aggregates (González & Moran, 1997; Suzuki *et al*., 2001), and ii) associated with eukaryotic organisms such as phytoplankton (Alavi *et al.*, 2001; Amin *et al.*, 2012; Buchan *et al.*, 2014; González *et al*., 2000; Grossart *et al.*, 2005; Jasti *et al.*, 2005; Zubkov *et al.,* 2001), marine red and green macroalgae (Ashen & Goff, 2000; Rao *et al*., 2007), and various marine animals such as corals, squids and urchins (Apprill *et al.*, 2009; Barbieri *et al*., 2001; Becker *et al.*, 2009; Collins *et al.,* 2012). The interactions with eukaryotic hosts could be mutualistic or pathogenic. One example of mutualistic interactions is *Dinoroseobacter shibae* DFL12^{T}, a symbiont of cosmopolitan marine microalgae including toxic dinoflagellates (it provides vitamins B1 and B12 to the host) (Wagner-Döbler *et al*., 2010). In addition, different Roseobacter lineage members are dominant components associated with the reproductive accessory nidamental glands in cephalopods *Loligo pealei* and *Euprymna scolopes* (squids) (Barbieri *et al*., 2001; Collins & Nyholm, 2011), and *Sepia officinialis* (cuttlefish) (Grigioni *et al*., 2000). Some examples of pathogenic interactions are the cases of *Nautella* sp. R11 that was characterized for causing bleaching disease in the marine red alga *Delisea pulchra* (Case *et al*., 2011) or *Roseovarius crassostreae* that causes a juvenile oyster disease that results in high mortality in *Crassostrea virginica* (Boettcher *et al*., 2005).

Members of the Roseobacter lineage can represent up to 20 % of bacterioplankton communities in coastal ecosystems and from 3 to 5 % in surface waters of the open oceans (Buchan & Moran, 2005; Moran *et al*., 2007). They were isolated from a variety of places such as coastal and deep-sea sediments (Lenk *et al*., 2012; Wang *et al.*, 2009), deep pelagic ocean (Eloe *et al*., 2011), and polar sea ice (Vollmers *et al*., 2013), in which Roseobacter lineage is a major bacterial group (Brinkmeyer *et al*., 2003; Junge *et al.*, 2002). They were also characterized for being isolated from chronically-polluted environments (Harwati *et al*., 2007, 2008, 2009; Piña-Villalonga, 2012; Suárez-Suárez, 2013) (see section 2). So, they are considered a ubiquitous group of microorganisms.

The traditional description of the Roseobacter lineage was based on the cultured members. But recent studies analyzing single-cell genomes of uncultured members of the group showed that there were genomic differences among the cultured and uncultured roseobacters (Luo & Moran, 2014). The majority of cultured roseobacters are considered to be generalists, with large genomes (4.2 ± 0.65 Mb), high G+C content mol % (60 ± 4 %), and versatile metabolic capabilities (Luo & Moran, 2014). The exception is isolate HTCC2255 that has an streamlined genome (2.28 Mb), a low G+C content mol % (37 %), and a scarcity of genes for transcriptional regulation, motility and cell-cell interactions (Luo *et al.*, 2013). These features are typical of uncultured roseobacters that have also streamlined genomes (estimated size 2.87 ± 0.15 Mb), low G+C content mol % (39 ± 1 %) and they also harbor fewer genes for signal transduction than cultured ones (Luo *et al*., 2012, 2014). Other differences in comparison to cultured roseobacters are cell surface modifications (depletion of capsular polysaccharide synthesis and cell adhesion proteins) or the harboring of more genes for phosphorus and sulfate uptake (Luo *et al*., 2012). For these features the uncultured roseobacters are thought to be specialists.

Within the framework of uncultured roseobacters, CHAB-I-5 subcluster, a group that comprises approximately 6 % of all bacterioplankton cells and around 20 % of the roseobacters in some surface ocean waters (Buchan & Moran, 2005), showed to be similar to cultured roseobacters (Zhang *et al*., 2016). The analysis of partial genomes of three single-cell representatives of CHAB-I-5 predicted that these microorganisms have large genomes (between 4.1 and 4.4 Mb) and a high fraction of non-coding DNA (10–12 %) which is different to streamlined genomes (Zhang *et al*., 2016). Thus, new surveys of uncultured and free-living roseobacters expand the previous information of cultured ones in relation to their life-style, genomic traits, gene repertorie, metabolic versatility, etc. (Luo *et al*., 2012, 2014; Luo & Moran, 2014; Zhang *et al*., 2016).

1.2. Relevant features of Roseobacter cultured isolates

Cultured members of Roseobacter lineage are considered mixotrophic bacteria that are able to obtain carbon and energy by chemoorganotrophy, photoheterotrophy and/or chemolitotrophy (Newton *et al*., 2010). Apart from obtaining energy by chemoorganotrophy through the oxidation of organic matter, some of them can also get energy from phototrophy by aerobic anoxigenic phototrophy (AAnP) (Allgaier *et al.*,2003; Newton *et al.*, 2010). Light is used to obtain energy but no carbon fixation is carried out since, in all phototrophic roseobacters, genes for carbon fixation are absent (Ribulose-1,5-bisphosphate carboxylase/oxygenase, RuBisCO). So, this metabolism produce energy but do not directly provide fixed carbon (Béjà *et al*., 2002; Karl, 2002). Phototrophic roseobacters have the genetic information for bacteriochlorophyll biosynthesis (*bch*), carotenoid biosynthesis (*crt*), light harvesting polypeptides (*puc* and *puf*), reaction center proteins (*puhA*, *pufLM*) and their regulators, *ppsR*, *tspO* and *ppaA* (Choudhary & Kaplan, 2000; Wagner-Döbler & Biebl, 2006).

Apart from the possibility of obtaining energy from light of some roseobacters, a common signature of cultured members of this group is a non-obligate chemolithotrophic metabolism. Carbon monoxide (CO) oxidation to CO_2 was experimentally demonstrated in some members of Roseobacter lineage and found to occur at CO concentrations typical of surface seawater (10 nM in coastal and 2 nM in oceanic regions) (Cunliffe, 2011; Moran *et al*., 2004). There are two different types of CO dehydrogenases described (King & Weber, 2007; King, 2003): i) type I or OMP-type (from *Oligotropha*, *Mycobacterium* and *Pseudomonas* type genes) that have high affinity for CO and it is present in the classical carboxidotroph microorganisms such as *Oligotropha carboxidovorans* and *Pseudomonas carboxydohydrogena*, and ii) type II or BMS-type (from *Bradyrhizobium*, *Mesorhizobium* and *Sinorhizobium* type genes), that have a lower affinity for CO than the OMP-type CO dehydrogenase and they are found in non-carboxidotroph microorganisms. Both types of CO oxidation genetic modules were detected in roseobacters (Luo & Moran, 2014; Newton *et al*., 2010). Moreover, *sox* genes that mediate the oxidation of sulfite or thiosulfide to sulfate were also identified in several Roseobacter genomes (Luo & Moran, 2014; Newton *et al*., 2010; Pukall *et al*., 1999; Wagner-Döbler & Biebl, 2006) and, as in the case of CO oxidation, oxidation of sulfur compounds provide an alternative

source of energy. All these mechanisms confirm that these bacteria use a variety of energy sources available to them.

Genomic functional annotation and physiological experiments have revealed that roseobacters have an extraordinary ability to metabolize a wide range of substrates to support their growth, including carbohydrates, sugar alcohols, organic acids, and amino acids (Buchan & Moran, 2005; Moran *et al*., 2004; Newton *et al*., 2010). The large number of proteins related to amino acid transport and metabolism in Roseobacter genomes suggested the importance of amino acids as a carbon source (Moran *et al*., 2004; Newton *et al*., 2010). Recent studies have revealed that choline (essential constituent of eukaryotic cells that occurs in marine microalgae and a variety of coastal plants) can be oxidized by members of Roseobacter lineage to maintain cell viability during periods of carbon starvation (Lidbury *et al*., 2015). Additionally, Roseobacter strains have been among the first aerobic bacteria isolated that grow on dimethylsulfoniopropionate (DMSP) (González *et al*., 2000, 2003; Reisch *et al*., 2011a, 2013). The use of DMSP as a carbon source is relevant because this compound is produced in high concentrations by marine micro-, macroalgae and halophytic plants as an osmoprotectant (Yoch, 2002). It is released into the ocean water and it is probably one of the most important sulfur and carbon sources for marine bacteria. There are two principal degradation pathways (Moran *et al*., 2003; Reisch *et al*., 2011a, 2011b, 2013): i) the cleavage pathway catalized by a lyase that leads to the formation of dimethylsulfide (DMS) and acrylate, which is used by both eukaryotic algae and few bacteria, and ii) the demethylation/dethiolation pathway, which is found only in bacteria and results in the formation of 3-methyl-mercaptopropionate (MMPA) and 5-methyl-tetrahydrofolate. MMPA can be degraded further either by the double demethylation pathway to form 3-mercaptopropionate (MPA), or by dethiolation, resulting in the volatile sulfur compound methanethiol (MeSH) and acrylate. Members of Roseobacter lineage are the only bacteria known to have both, the degradation and the demethylation/dethiolation pathways, sometimes in the same organism (Miller & Belas, 2004; Moran *et al*., 2003). Thus, Roseobacter organisms are adapted to algal blooms (González *et al.*, 2000), where DMSP is released locally in large, varying quantities. Other types of compounds for which there are genomic and physiological evidences that roseobacters are capable of using as a sole source of carbon and energy are alkanes and aromatic compounds (Alejandro-Marín *et al*., 2014; Buchan & Moran, 2005; Harwati *et al*., 2008, 2009; Newton *et al*., 2010) (see section 3).

Most cultured roseobacters harbor genes that encode for chemotaxis and motility functions (Newton *et al*., 2010; Slightom & Buchan, 2009). These abilities could suppose growth advantages to exploit resource-poor ocean waters. It has shown for example, that *Ruegeria* sp. TM1040 was strongly attracted to aminoacids and DMSP while it was only moderately responsive to sugars and tricarboxylic acid intermediates (Miller *et al*., 2004). Furthermore, chemical products of dinoflagellates (Miller *et al*., 2004) and the toxic phytoplankton *Heterosigma akashiwo* (Seymour *et al*., 2009) acted as chemoattractants for this strain. In reference to the capability for motility, flagellar motility has been experimentally

demonstrated for many characterized Roseobacter isolates, and motile ones typically exhibit between one and five polar or subpolar flagella (Biebl *et al*., 2005; Cilia *et al.*, 1998; González *et al.*, 2003; Miller *et al.*, 2004; Shiba, 1991). Polar complex flagella have been documented in most detail for *Ruegeria pomeroyi* DSS-3^T (González *et al*., 2003). Flagellar synthesis and/or motility have been found to be critical for surface recognition, attachment, and biofilm development in many *Proteobacteria*. Roseobacters are described as rapid colonizers of surfaces in coastal environments and its abundance is often high near phytoplankton blooms or macroalgae, or in association with organic particles, suggesting that cell-surface interactions are a defining feature of lineage members (Dang & Lovell, 2000; Mayali *et al*., 2008). For the attachment to surfaces and to other *Proteobacteria*, roseobacters harbor genes encoding components of appendages that allow the formation of biofilms. A genomic island (*tad* genes) encodes the machinery required for the assembly of adhesive fimbrial low-molecular-weight protein (Flp) pili that is important in surface colonization, biofilm formation, and pathogenesis in different bacteria (Tomich *et al.*, 2007). There is strong synteny of seven genes across Roseobacter genomes (*cpaBC-ompA-cpaEF-tadBC*) that is similar to the *cpa* (*Caulobacter* pilus assembly) genes of *Caulobacter crescentus* (Skerker & Shapiro, 2000) and appears to be well-maintained across most *Alphaproteobacteria*. However, in most other *Alphaproteobacteria* instead of *ompA* of roseobacters, its homolog, the *cpaD* gene was found. These genomic island that codifies for these pili was identified in all 28 Roseobacter genomes analyzed by Slightom & Buchan (2009).

Roseobacter lineage is also characterized for the production of secondary metabolites e.g. production of acylated homoserine lactones (AHLs) or antibiotics (Brinkhoff *et al*., 2004; Bruhn *et al*., 2007; Martens *et al*., 2007; Wagner-Döbler *et al*., 2004; Zan *et al*., 2014). The production of AHLs is related to *quorum sensing* (QS) systems that allow bacteria sense and perceive their population density through the use of diffusible signals (Fuqua & Greenberg, 2002; Gram *et al*., 2002; Miller & Bassler, 2001; Zan *et al*., 2014). Production of AHL has been detected in many roseobacters such as *Ruegeria pomeroyi* DSS-3^T, *Phaeobacter inhibens* DSM 14395^T, *Ruegeria* sp. KLH11, *Dinoroseobacter shibae* DLF 12^T and many others (Martens *et al*., 2007; Moran *et al*., 2004; Zan *et al*., 2014). Traits known to be regulated by QS are gene expression in relation to cell density, host-symbiont and host-pathogen interactions, the production of polysaccharides, gene transfer, motility, biofilm formation and the production of antibiotics, toxins and other metabolites (Fuqua & Greenberg, 2002; Zan *et al*., 2014). Regarding the production of antibiotics, three different antibiotics were isolated from roseobacters: i) tryptanthrin from *Oceanibulbus indolifex* HEL-45^T (Wagner-Döbler *et al*., 2004), ii) thiotropocin from strain 27-4 (Bruhn *et al*., 2005), and iii) tropodithietic acid from *Phaeobacter gallaeciensis* (Berger *et al*., 2011, 2012). The production of antibiotics is effective against other marine bacteria and algae that could compete for the same ecological niche, so it supposes an advantage. Other secondary metabolites produced by members of Roseobacter lineage are polyketides (PK) and nonribosomal peptides (nonribosomally produced peptides) (NRP) that are a large family of natural products with antibiotic, immunosuppressive, cytostatic or toxic, activities which are interesting as pharmaceuticals and/or antibacterial agents (Schwarzer *et*

al., 2003; Staunton & Weissman, 2001). Other roseobacters produce other secondary metabolites such as *Roseobacter algicola* that synthetizes the toxin okadaic acid (Lafay *et al*., 1995) or *Oceanibulbus indolifex* HEL-45^T that produces indole, indole derivates and cyclic dipeptides (Wagner-Döbler *et al*., 2004).

Another feature of the lineage is the ability to synthesize vitamins. This is the case of *Dinoroseobacter shibae* DFL12^T that was isolated from *Prorocentrum lima* (a dinoflagellate with whom it has a symbiotic relationship) (Biebl *et al*., 2005). *Dinoroseobacter shibae* DFL12^T is able to synthesize vitamins B1 (thiamin) and B12 (cobalamin) for which its host is auxotrophic (Wagner-Döbler *et al*., 2010). Due to the relevance of vitamin synthesis for symbiotic relationships genes for thiamin, cobalamin and biotin (vitamin B7 or H) synthases were searched in 52 genomes of Roseobacter, being detected in 33, 51 and 21 genomes, respectively (Luo & Moran, 2014).

As a summary, for its abundance and worldwide distributed populations in marine environments and all these abilities to obtain energy, to use a multitude of organic compounds, and to its capability for genetic exchange (see below), members of Roseobacter lineage play a crucial role in the biogeochemical marine cycles, reason why the interest in isolation and description of new members of this lineage has increased in recent years.

1.3. Horizontal gene transfer (HGT) in the Roseobacter lineage

Members of Roseobacter lineage have an evolutionary life history strongly influenced by HGT processes (Luo *et al*., 2013). In fact, one of the two time periods of substantial evolutionary change in Roseobacter genomes is considered to have been due to genome expansion via HGT (Luo *et al*., 2013). The process of HGT, in which DNA is exchanged between different organisms changed the view of the evolutionary relationships among living organisms (Lang *et al.*, 2012). The acquisition of foreign genes occurs extensively among prokaryotes, especially in response to a changing environment, and provides organisms with access to genes increasing its genetic diversity (Jain, 2003; Syvanen, 1994). A study made with 116 prokaryotic complete genomes, estimated that the average proportion of horizontally transferred genes per genome was nearly 12 %, ranging from 0.5 % to 25 % depending on the prokaryotic lineage (Nakamura *et al.*, 2004). The smallest proportion was observed in an endocellular symbiont and the largest in an *Euryarchaea* (Nakamura *et al*., 2004). As a representative of the Class *Alphaproteobacteria*, two members of *Rickettsia* genus were analyzed and the estimated average proportion of horizontally transferred genes they had was 4.2 % and 4.9 % (Nakamura *et al*., 2004). It has been observed that the genetic exchange among organisms is not random, preferently occurring related to a common ancestry or sharing of habitat (Beiko *et al.*, 2005; Jain, 2003; Kloesges *et al.*, 2011).

Because prokaryotic species evolve not only through vertical inheritance but also by DNA acquisition via HGT (Ochman *et al.*, 2000), the study of HGT processes is a way to understand the versatility and genomic diversity that makes them more adaptable to environmental

changes. Some mechanisms of HGT such as gene transfer agents (GTAs) and extrachromosomal elements (ECEs), in particular *repABC* family plasmids, have been studied in the Roseobacter lineage (Biers *et al*., 2008; Petersen *et al*., 2009). On the contrary, there are no previous studies about transposase fluxes among roseobacters that could help in elucidating further possibilities of genetic exchanges among members of Roseobacter lineage.

1.3.1. Gene transfer agents (GTAs)

A GTA is a phage-like entity that was discovered more than four decades ago in *Rhodobacter capsulatus* (from now the GTA of *R. capsulatus* is named as RcGTA). This appeared to be a novel model of HGT similar to virus-mediated generalized transduction where the transducing agent was not a typical bacteriophage (Marrs, 1974). GTAs are not exclusive of *R. capsulatus*. In fact, they have been found in other genomes of the Class *Alphaproteobacteria*, which suggested that this kind of HGT was present in the last common ancestor of this Class and it was subsequently lost in some lineages (Lang *et al*., 2012). GTA particles transfer random genome fragments of a producing cell to recipient cells (Hynes *et al.*, 2012; Lang & Beatty, 2006; Lang *et al*., 2012; Stanton, 2007).

The RcGTA structural genes are encoded by a gene cluster of approximately 15 kb constituted by 17 genes (designated gene 1 to gene 15) (Lang & Beatty, 2000, 2001; Lang *et al.*, 2002). This cluster includes homologs of phage structural genes (capsid, tail, portal, and DNA packaging) but not of self-replication or host lysis genes (Biers *et al*., 2008). In Figure 2I, the structural genes of RcGTA (Genbank reference: AF181080.3) are shown.

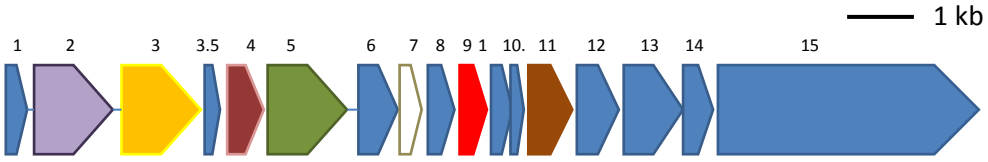

Figure 2I. Genetic cluster of RcGTA. The numbers over the genes indicate genic order in the GTA. Genes with defined functions: 2 (large terminase), 3 (portal protein), 4 (prohead protease), 5 (capsid protein), 7 (head-tail adaptor), 9 (major tail protein) and gene 11 (tail tape measure protein). The rest of genes (blue color) have not assigned function and are defined as a putative GTA.

In general, GTAs have tailed-phage structures and are released to the environment by lysis of the producing cell. Released GTA particles may transfer DNA from the producing cell to a recipient cell (Lang *et al*., 2012). Their production is not the result of a phage infection. In *Rhodobacter capsulatus*, the expression of RcGTA is controlled by cellular regulatory systems that include histidyl-aspartyl signaling (Lang & Beatty, 2000) and *quorum sensing* genes (Schaefer *et al.*, 2002). It has been seen that its production is strongly dependent on the phase of host cell growth, with increased production in the stationary phase (Solioz *et al.*, 1975). On the other hand, it seems that the cells in early-stationary phase are active GTA recipients (Solioz *et al*., 1975). The cell receptor for the adsorption of RcGTA was identified as a capsular polysaccharide receptor (Brimacombe *et al*., 2013).

In reference to the DNA of the host genome that GTAs can transfer, the size of the host DNA in the particle (4.3 kb approximately) is smaller than the DNA encoding the GTA structural proteins (15 kb approximately). Thus a single particle is unable to transfer all the genes required for its own synthesis to a new host and therefore to transfer the ability to produce a GTA to another cell. For that reason, it is though that the GTA common ancestor became defective and lost its capability to transfer to other genomes (Lang *et al*., 2012). This is in contrast from a generalized transducing phage, for which usually only an occasional particle contains host genes, and the fragments of packaged DNA are the size of the phage genome (Ikeda & Tomizawa, 1965).

Regarding to Roseobacter lineage, the production of GTAs was demonstrated in *Ruegeria pomeroyi* DSS-3^T. The produced particles were very similar to RcGTAs particles, however, in *R. pomeroyi* DSS-3^T particles do not appeared to have tails (Biers *et al*., 2008). Two other Roseobacter members have been experimentally demonstrated to produce RcGTA-like particles: *Roseovarius nubinhibens* ISM^T and *Ruegeria mobilis* 45A6 (McDaniel *et al*., 2010). In contrast, more studies about the genomic prediction of the GTA structure have been done. These surveys showed that the presence of GTA in Roseobacter cultured members is a common feature of the lineage (Biers *et al*., 2008; Lang *et al*., 2012; Luo & Moran, 2014).

In order to assess the origin of GTA sequences in the Roseobacter lineage, different studies were made. Lang and Beatty (2006) reported a phylogeny of the capsid sequence of 54 members of the Class *Alphaproteobacteria*. Later, Biers and co-workers (2008) published a phylogeny of the concatenated amino acid sequences of portal protein (gene 3), capsin protein (gene 5) and a protein with no function assigned (gene 12) of 65 bacteria in the Class *Alphaproteobacteria*. In both cases, members of Roseobacter lineage formed a distinct and well-supported clade. In 2014, Luo and Moran detected the GTA genic products in 46 out of 52 Roseobacter genomes. All three surveys pointed out the widespread presence of RcGTA-like structures in *Alphaproteobacteria* Class and the first and second reports suggested the possibility of a prokaryotic unique ancestor for all the Class.

1.3.2. Extrachromosomal elements (ECEs)

The presence of ECEs was previously studied in six members of Roseobacter lineage (Biers *et al*., 2008; Petersen *et al*., 2009, 2011, 2013). Results revealed the presence of, at least, 4 types of plasmids depending on the non-homologous replicases that they harbored (replication types RepABC, RepA, RepB and DnaA-like). The analysis showed that *Dinoroseobacter shibae* DFL-12^T had plasmids with all 4 replication systems, while the other 5 roseobacters analyzed had, at least, one kind of replication system (Petersen *et al*., 2013). DnaA-like type plasmids are only found in *Rhodobacterales*, the order that includes the Roseobacter lineage. RepA and RepB type plasmids have a broader distribution and are found in *Alpha-*, *Beta-*, *Gamma-* and *Deltaproteobacteria* as well as in *Firmicutes* (Petersen, 2011).

Until now, the plasmids that have been most studied in roseobacters are RepABC-family plasmids (with RepC replicase) (Cervantes-Rivera *et al*., 2011; Cevallos *et al*., 2008; Pérez-Oseguera *et al*., 2013; Petersen *et al*., 2009; Venkova-Canova *et al*., 2004). The RepABC-type replicons are widely distributed in the Class *Alphaproteobacteria* as low-copy number plasmids (they are exclusive for this class). They are specially common in *Rhizobiales* (Cevallos *et al*., 2008) and also in the Roseobacter lineage (Petersen *et al.,* 2009, 2012). They were named according to the genetic arrangement of *repA*, *repB* and *repC* (*repABC*). These components are in general physically clustered and the gene order is conserved (Cevallos *et al*., 2008; Thomas, 2000). The *repAB* genes encode for the proteins responsible of the partioning and segregation of the plasmid, and RepC is the replication initiator protein (Bartosik *et al*., 1998; Ramírez-Romero *et al*., 2001). RepA is an ATPase able to polymerize into filaments and RepB binds the ATPase to a centromere-like sequence (*parS*) acting as an adaptor between the plasmid and the filaments that are responsible for the segregation process (Ebersbach *et al*., 2006). The intergenic sequence between *repB* and *repC* genes contains a gene encoding a small antisense RNA (*incα*) (Cervantes-Rivera *et al*., 2010; Ramírez-Romero *et al.*, 2000). It plays an essential role in plasmid replication because it negatively regulates *repC* expression (Chai & Winans, 2005a; MacLellan *et al.*, 2004; Venkova-Canova *et al.*, 2004). RepC is the limiting factor for replication. Mutation, deletion or insertion in this gene, abolish the capacity of the plasmid to replicate (Ramírez-Romero *et al*., 2000; Tabata *et al.*, 1989). In general, one *repABC* module is detected in each plasmid. However, few of these plasmids contain two *repABC* modules, a fact that is common in the order *Rhizobiales* (Cevallos *et al*., 2008). An analysis of 30 completely sequenced Roseobacter genomes that were available in 2009 showed that two-thirds of them contained plausible RepABC-type plasmids (Petersen *et al*., 2009), demonstrating the prevalence of this type of genetic exchange elements in this lineage.

Plasmids are classified in compatibility groups, which are defined on the observation that some plasmids cannot coexist in stable manner within the same bacterium because of the existence of incompatibility regions. That was because of the sharing of the same elements involved in plasmid partioning, replication and control (Novick, 1987). Incompatibility restricts the ability of a bacterial cell to acquire similar plasmids but favors the presence of a wider selection of plasmids within a cell (Cevallos *et al*., 2008). The presence of several RepABC replicons in the same bacteria implies that this plasmid family includes several incompatibility groups. In the case of Roseobacter lineage, *Sulfitobacter* sp. NAS-14.1 harbors five distantly related *repABC* modules that corresponded to 5 different compatibility groups (Petersen *et al*., 2009). It has been shown that RepABC plasmids contain different elements involved in plasmid incompatibility: the RepABC proteins, the small antisense RNA and the *parS* sequences (Cervantes-Rivera *et al*., 2011; Chai & Winans, 2005b; MacLellan *et al*., 2004; Ramírez-Romero *et al*., 2000; Venkova-Canova *et al*., 2004). Nine different compatibility groups of RepABC replicons in *Rhodobacterales* order based on phylogenetic analysis of RepC were identified (Petersen *et al*., 2009).

1.3.3. Transposase fluxes

Insertion Sequences (ISs) are widespread and can occur in high numbers in prokaryotic genomes (Siguier *et al*., 2014). A recent study concluded that transposases were the most abundant genes in both, completely sequenced genomes and environmental metagenomes (Aziz *et al.*, 2010). The rapid dynamics of IS gain, expansion, and loss in genomes of prokaryotes strongly depends on its incorporation by HGT (Touchon & Rocha, 2007) reason why they are considered well suited to provide recent historical information of HGT events between microbes (Hooper *et al.*, 2009). On the other hand, transposases are potentially transferred horizontally more frequently than other genes (Hooper *et al*., 2009) due to its low levels of divergence (Wagner, 2006) compared to other genes. Hooper and co-workers (2009) presented a large scale study of the distribution of transposases across almost 800 bacteria related to its phylogenetic proximity and habitat. They found that the majority of the connections were between closely related organisms with the same or overlapping habitat. However, numerous examples of cross-habitat and cross-phylum connections were also found (Hooper *et al*., 2009).

It has been suggested that transposition of catabolic clusters, followed by DNA rearrangements, might be one of the mechanism whereby degradation genes or operons were recruited (Williams & Sayers, 1994). In fact, the presence of transposase-like genes in close proximity to the upper naphthalene-degradation pathway of *Pseudomonas stutzeri* AN10 can be taken as a supporting evidence that such transposition events have occurred (Bosch *et al.*, 1999). In the case of Roseobacter lineage, the genomes of isolates described as degraders harbor transposase-encoding genes. That was the case, for example, of *Citreicella aestuarii* 357, isolated from a petroleum-polluted beach and able to degrade naphthalene, salicylate and dibenzothiophene. This strain had at least 76 plausible transposases in its genome (Suárez-Suárez *et al*., 2012). Another example is *Celeribacter indicus* P73^{T}, which might have acquired genes for degradation of polyaromatic hydrocarbons by HGT (Cao *et al*., 2015). Because of the relation of transposases with the recruitment of degradation genes or operons, one of the aims of this study has been to analyze the transposases fluxes among members of Roseobacter lineage, and relate these fluxes with phylogenetic relationship and habitat sharing.

2. Importance of the Roseobacter lineage in chronically-polluted marine environments

Large quantities of aromatic compounds are realeased to the environment from both, natural and anthropogenic sources, every year (Head *et al*., 2006). The effect of hydrocarbon pollution on marine microbial communities has been profusely studied; mainly the cases of acute contamination such as the accidents of the Prestige tanker in Spain (Alonso-Gutiérrez *et al*., 2009; Jiménez *et al*., 2007, 2011) or the Deepwater Horizon oil spill in the Gulf of Mexico (Kostka *et al*., 2011; Lamendella *et al*., 2014). In general, the results obtained in these surveys showed a decrease of diversity at the short term because of the disappearance of certain

groups of microorganisms (i.e. archaea and cyanobacteria) and the appearance of blooms of other indigenous dominant bacteria that are specialist hydrocarbon degraders (i.e. *Alcanivorax*, *Marinobacter*, *Thallassolituus*, *Cycloclasticus, Oleispira*), which become abundant mainly when nutrients are added (Head *et al.*, 2006; Yakimov *et al.*, 2007). On the contrary, microbial diversity tends to be high in marine environments chronically polluted by hydrocarbons (Alonso-Gutiérrez *et al*., 2009; Hernandez-Raquet *et al*., 2006; Nogales *et al*., 2007, 2011; Paissé *et al*., 2008) probably due to their adaptation to the presence of these compounds in the environment. In fact, 16S rRNA gene sequences of well-known hydrocarbon degraders such as *Alcanivorax* are seldom detected in chronically-polluted environments (Nogales *et al*., 2011).

Harbors are one example of chronically hydrocarbon-polluted environments. They are semi closed bodies of water in which there is little exchange with the outside water mass. They have inputs of nutrients from the coast and inputs of hydrocarbon pollution from a variety of sources that derive from anthropogenic activities (e.g. boat traffic, accidental spills, etc.) (Nogales, 2010). Due to nutrient enrichment, the abundance of planktonic prokaryotes and phytoplankton is higher in comparison to locations in the adjacent coast (Nogales *et al*., 2007; Schauer *et al.*, 2000; Zhang *et al.*, 2007). Nogales and co-workers (2007) published a comparative survey of the composition of bacterial communities in a marina (Mallorca Island) and showed that bacterial communities in the harbor had higher diversity and different composition than the bacterial communities from unspoilt areas nearby. Near the harbor, the abundance of typical oligotrophic planktonic bacteria (SAR11) was reduced while typical mesotrophic bacteria (e.g. *Bacteroidetes, Gammaproteobacteria* or members of Roseobacter lineage) increased. In terms of abundance, members of the Roseobacter lineage were detected as the dominating populations, but their abundance was irregular. In her doctoral thesis, Joana Maria Piña-Villalonga studied the diversity of the Roseobacter lineage in seven marinas of the coast of Mallorca (Piña-Villalonga, 2012). Members of the Roseobacter lineage were detected as one of the predominant bacterial groups in all of them. Other diversity studies carried out at Victoria Harbor in Hong Kong (Zhang *et al*., 2009; Zhang *et al*., 2007) and in Xiamen Port in Singapore (Ma *et al*., 2009) also showed that members of Roseobacter lineage were abundant in these environments. Other microbial diversity studies have demonstrated that Roseobacter populations responded favorably to hydrocarbon addition (Buchan & González, 2010). For example, Brakstad and Lødeng (2005) found that the relative contribution of Roseobacter sequences in bacterial 16S rRNA gene amplicon pools increased after the exposure of near shore and offshore water samples to crude oil. On the contrary, in sub-Antarctic waters in the coast of Argentina exposed to crude oil there was a decresase in clones affiliated to Roseobacter (Prabagaran *et al*., 2007). Additional studies showed that some Roseobacter populations may be stimulated by alkanes but inhibited by other components of crude oil (McKew *et al*., 2007). In 2010, Lanfranconi and collaborators performed experiments with microcosms in which they added low amounts of diesel oil without nutrients to seawater collected at different times. After diesel oil treatment there was a reduction in the number of phylotypes related to SAR11, SAR86 and picocyanobacteria while phylotypes of the

Roseobacter lineage increased (Lanfranconi *et al*., 2010). In a recent study in which experiments with mudflat in tidal mesocosms treated with crude oil members of Roseobacter lineage were detected by isolation and by pyrosequencing of 16S rRNA genes (Sanni *et al*., 2015). Another mesocosm experiment, using seawater with the addition of diesel oil, showed an initial bloom of *Vibrio* sp. followed by a substitution of the microbial community by well-known hydrocarbon degraders such as *Oleispira* spp. and *Methylophaga* spp., and opportunistic bacteria such as *Roseobacter* spp. and *Phaeobacter* spp. (Sauret *et al*., 2015).

All these studies highlighted the fact that members of Roseobacter lineage could tolerate the presence of aliphatic and aromatic compounds and/or could proliferate in response to its addition, which might indicate that they have an active role in degradation.

3. Hydrocarbons and aromatic compounds degradation

Members of the Roseobacter lineage have been also characterized for hydrocarbon and aromatic compounds degradation. Their potential for the degradation of aromatic compounds was first reported when Buchan and co-workers (2000) analyzed 6 Roseobacter strains isolated from lignin or aromatic monomer enrichments. These isolates were able to grow with different aromatic compounds (see Table 1I). Subsequently, Moran and co-workers (2007) identified six distinct ring-cleaving pathways for monoaromatic compounds degradation in three Roseobacter genomes. From that moment, the research of genes for aromatic compounds degradation was extended to a higher number of genomes (Buchan & González, 2010; Newton *et al*., 2010). In parallel, new Roseobacter isolates were described as hydrocarbon or aromatic compound degraders (Brito *et al*., 2006; Harwati *et al*., 2007), and some of these were isolated from polluted environments (Giebel *et al*., 2016; Harwati *et al*., 2007, 2008, 2009; Piña-Villalonga, 2012; Suárez-Suárez, 2013) (see Table 1I).

In the following two sections we update the evidences related to the Roseobacter degradation capabilities of aromatic and aliphatic hydrocarbons.

3.1. Degradation of aromatic compounds

Aromatic compounds are aerobically degraded by bacteria by a discrete number of pathways that involve incorporation of molecular oxygen (by ring-hydroxylating dioxygenases and/or ring-cleaving dioxygenases). In addition, Coenzyme A (CoA) activation of benzene ring is also common among aerobic bacteria, including roseobacters (Moran *et al*., 2007; Zaar *et al*., 2004). Regardless of the mechanism of ring cleavage, intermediates from these pathways ultimately feed into the tricarboxylic acid cycle (TCA cycle). Thus, these compounds can serve as primary growth substrates for the bacteria that use them.

Different reports provided evidences for the role of roseobacters in aromatic compound degradation in marine systems. The first evidence came from the isolation of a lignin-transforming bacterium, *Sagittula stellata* E-37^T, from an enrichment culture prepared with

Table 1I. Members of Roseobacter lineage demonstrated to be degraders of hydrocarbons or aromatic compounds or able to grow with these compounds.

Isolate	Source	Compounds[a]	Reference
Sagitulla stellata E-37^T	Coastal marshes (lignin enrichments)	ANT, BEN, COU, FER, HYD, PRO	González *et al*., 1997; Gulvik & Buchan, 2013
Several (17)	Coastal marshes (lignin enrichments)	ANT, BEN, COU, FER, HYD, PRO, SAL, VAN	Buchan *et al*., 2000, 2001, 2004
Ruegeria pomeroyi DSS-3^T	Seawater	BEN, COU, FER, HYD, PRO	Buchan *et al*., 2004; Gulvik & Buchan, 2013
Several (8)	Mangrove sediments	FLU, NAP, PYR	Brito *et al*., 2006
Several (40)	Port seawater	ADB, AFL, ALK, ANA, APH	Harwati *et al*., 2007
Tropicibacter naphthalenivorans C02^T	Port seawater	ANA, NAP, PHE	Harwati *et al*., 2009
Several isolates (8)	Coastal seawater	DPT, PHT	Iwaki *et al*., 2012a
Tropicibacter phthalicus KU27E1^T	Coastal seawater	DPT, PHT	Iwaki *et al*., 2012b
Several isolates (17)	Coastal seawater (harbors)	BEN, GEN, HOM, PHL, PRO, SAL	Piña-Villalonga, 2012
Citreicella aestuarii 357	Crude oil-polluted beach	DIB, GEN, NAP, SAL	Suárez-Suárez, 2013
Celeribacter indicus P73^T	Deep-sea sediment	DIB, DIF, FLO, FLU, NAP, PHE	Cao *et al*., 2015; Lai *et al*., 2014
Phaeobacter gallaeciensis DSM 26640^T	Seawater from larval cultures	BRO	Ichikawa *et al*., 2015
Roseobacter denitrificans OCh 114^T	Marine sediment	BRO	Ichikawa *et al*., 2015
Confluentimicrobium naphthalenivorans NS6^T	Crude oil-contaminated sea-tidal-flat sediment	NAP	Jeong *et al*., 2015
Celeribacter persicus SBU1^T	Mangrove soil	PAH	Jami *et al*., 2016
Strain O3.65	Contaminated deepwater of horizon oil spill	COU, FER, HYD, PRO, VAN	Giebel *et al*., 2016

a: abbreviations: ADB: alkyldibenzothiophenes, AFL: alkylfluorenes, ALK: alkanes, ANA: alkylnaphthalenes, ANT: Anthranilate, APH: alkylphenanthrene, BEN: Benzoate, BRO: bromoform, COU: coumarate, DIB: dibenzothiophene, DIF: dibenzofuran, DPT: dimethylphthalate, FER: ferulate, FLO: fluorene, FLU: fluoranthene, GEN: gentisate, HOM: homogentisate, HYD: 4-hydroxybenzoate, NAP: naphthalene, PAH: polycyclic aromatic hydrocarbons, PHE: phenanthrene, PHL: phenylacetate, PHT: Phthalate, PRO: protocatechuate, PYR: pyrene, SAL: salicylate, VAN: vanillate.

seawater of an effluent of a pulp mill in the coast of Georgia, USA (González *et al*., 1997). Subsequently, Buchan and co-workers (2000) described the first monoaromatic hydrocarbon degradation pathway in roseobacters: the protocatechuate branch of the β-ketoadipate pathway. The *pcaGH* genes encoded for the protocatechuate 3, 4-dioxygenase ring cleavage enzyme, the first enzyme of the pathway that transform protocatechuate to β-carboxymuconate. Gene *pcaH*, was detected in 6 roseobacter isolates (Buchan *et al.*, 2000). All of them could grow on at least three of the eight monoaromatic compounds tested (anthranilate, benzoate, 4-hydroxybenzoate, salicylate, vanillate, ferulate, protocatechuate, and coumarate, see Table 1I). Moreover, protocatechuate 3, 4-dioxygenase activity was detected in cell extracts of four of the isolates grown on 4-hydroxybenzoate (a precursor of protocatechuate). In 2013, Gulvik and Buchan showed the simultaneous catabolism of 4-hydroxybenzoate and benzoate in *Sagittula stellata* E-37^T and *Ruegeria pomeroyi* DSS-3^T.

As the number of Roseobacter genome sequences increased in databases the search for aromatic compound degradation genes started. The first study was done by Moran and collaborators (2007), who analyzed in detail the genomes of the first three sequenced Roseobacter isolates (*Ruegeria* sp. TM1040, *Jannaschia* sp. CCS1 and *Ruegeria pomeroyi* DSS-3^T). They described, based on sequence identity, the presence of six distinct ring-cleaving pathways for monoaromatic compounds (benzoyl-CoA, phenylacetic acid, homoprotocatechuate, homogentisate, gentisate and protocatechuate). Newton and collaborators (2010) extended the analysis of these 6 ring-cleaving pathways to 32 Roseobacter genomes (see Figure 3I). Only in 5 of the 32 analyzed genomes none of these degradation pathways were detected. On the contrary, in 4 of the analyzed genomes genes for all six pathways were detected (*Sagittula stellata* E-37^T, *Jannaschia* sp. CCS1, *Ruegeria pomeroyi* DSS-3^T, and *Roseobacter* sp. GAI101). The most detected pathways corresponded to the protocatechuate branch of the β-ketoadipate, homogentisate and phenylacetic acid pathways (Newton *et al*., 2010).

In 2014, an analysis of detection and organization of genes in 43 members of Roseobacter lineage of the protocatechuate branch of the β-ketoadipate pathway was done (Alejandro-Marín *et al.*, 2014). The eight genes of the protocatechuate pathway, possible regulators, and genes encoding for related functions (such as the catabolism of 4-hydroxybenzoate) were predicted by sequence identity. While most of the roseobacters studied had putatively a complete protocatechuate branch of the β-ketoadipate pathway, in 11 of them not all the genes were detected and incomplete pathways were predicted. Moreover, some of the roseobacters showed to have a potential for transforming i) 4-hydroxybenzoate to protocatechuate, and ii) catechol via *ortho*-cleavage to muconate. Diversity in gene organization was observed among the Roseobacter genomes, with no clear relationship between gene order and taxonomy. The results of this study highlighted the relevance of 4-hydroxybenzoate and β-ketoadipate degradation pathways in the Roseobacter lineage.

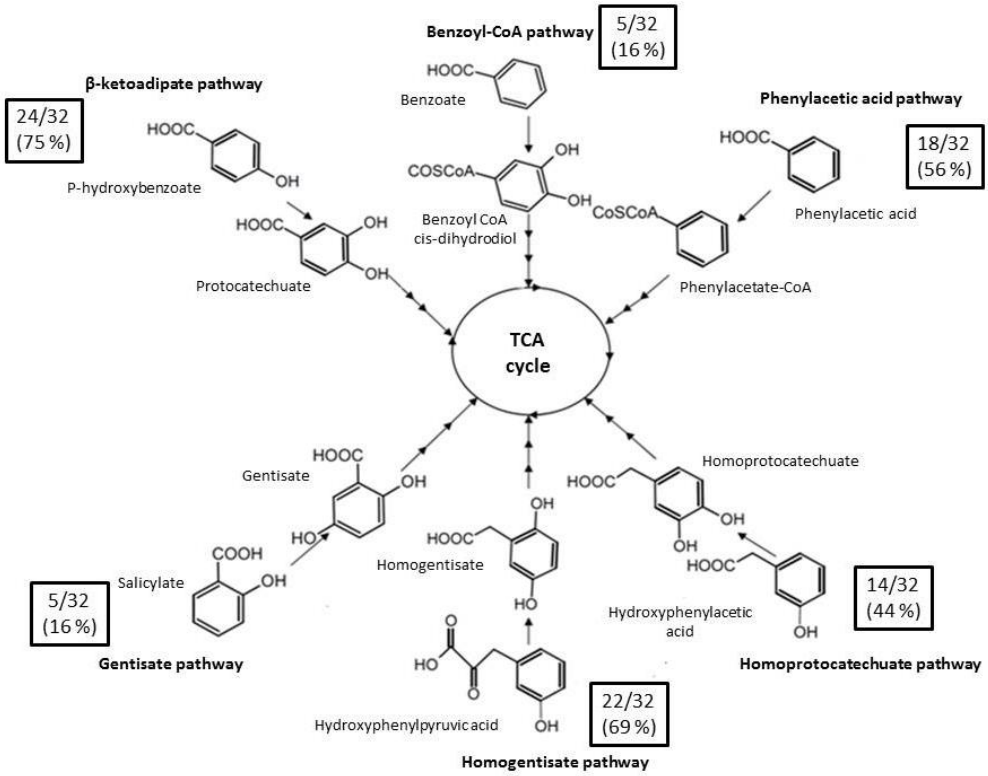

Figure 3I. Aerobic degradation pathways for monoaromatic compounds described in the Roseobacter lineage. Image modified from Moran and co-workers (2007) complemented with the information from Newton and co-workers (2010). The numbers in boxes indicate the number of genomes in which the genes coding for ring-cleaving enzymes of the pathways were detected and the percentage of detection (32 genomes).

Apart from the degradation of monoaromatic compounds, other roseobacters are able to grow using polycyclic aromatic hydrocarbons (PAHs). In 2006, Brito and co-workers isolated a collection of eight Roseobacter bacteria from mangrove sediments in Brazil that were able to grow on pyrene, naphthalene, or fluoranthrene. Other Roseobacter isolates able to degrade polyaromatic hydrocarbons have been described (see Table 1I). Some examples are *Tropicibacter naphthalenivorans* C02^T that was able to degrade naphthalene, alkylnaphthalenes and phenanthrene (Harwati *et al*., 2009); *Celeribacter indicus* P73^T that was capable of degrading a wide range of PAHs including naphthalene, phenantrene, dibenzothiophene and fluoranthene (Cao *et al*., 2015; Lai *et al*., 2014); or *Citreicella aestuarii* 357 that was able to grow using naphthalene and dibenzothiophene as a sole source of carbon and energy (Suárez-Suárez, 2013). In the case of *Celeribacter indicus* P73^T and *Citreicella aestuarii* 357 the genomic prediction of which genes could be involved in the degradation of PAHs was done. The genome of *Celeribacter indicus* P73^T contains 138 candidate genes that may be involved in the metabolism of aromatic compounds, including genes that encode six ring hydroxylating dioxygenases, eight ring cleaving dioxygenases, other catabolic enzymes, transcriptional regulators, and transporters (Cao *et al*., 2015). On the other hand, the genome sequence of *Citreicella aestuarii* 357 revealed the presence of two complete monoaromatic hydrocarbon degradation pathways (homogentisate and gentisate) and a putative naphthalene degradation pathway lacking transcriptional regulators (Suárez-Suárez, 2013).

3.2. Degradation of aliphatic hydrocarbons

Apart from the degradation of aromatic compounds, some members of Roseobacter lineage were also described as aliphatic hydrocarbon degraders (see Table 1I). In the study of Harwati and co-workers (2007) new Roseobacter isolates were identified for its capability for degrading linear hydrocarbons, such as *Tranquillimonas alkanivorans* (Harwati *et al*., 2008). In reference to the genomic characterization, Buchan and González (2010) analyzed 23 Roseobacter genomes in which in the 88 % of cases an alkane hydroxylase was predicted. However, in contrast to monaromatic compounds degradation, there is no biochemical information about how roseobacters metabolize alkanes.

Alkane degradation in highly specialized hydrocarbonoclastic marine bacteria, such as *Alcanivorax borkumensis* has been described (Schneiker *et al*., 2006). As shown in Figure 4I, the aerobic degradation of alkanes can proceed in two ways (Rojo, 2009). The first way is a terminal oxidation which starts by the oxidation of a terminal methyl group to render a primary alcohol, which is further oxidized to the corresponding aldehyde, and finally converted into a fatty acid (Watkinson & Morgan, 1990). The second way to degrade alkanes is by subterminal oxidation (Kotani *et al.*, 2006, 2007). The product that is generated is a secondary alcohol which is converted to a ketone, and subsequently oxidized by a Baeyer-Villiger monooxygenase to render an ester. The ester is hydrolysed by an esterase, generating an alcohol and a fatty acid. Both terminal and subterminal oxidation can coexist in some microorganisms (Rojo, 2009). While the activation of the alkane molecule requires specific en-

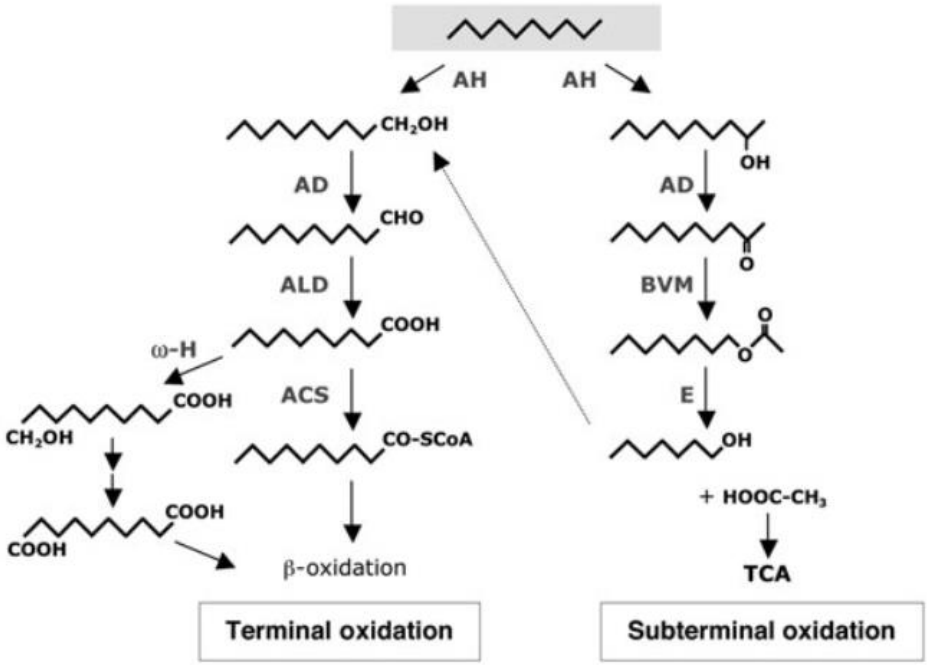

Figure 4I. Aerobic pathways for the degradation of n-alkanes. Initial activation of the alkane molecule requires oxygen as reactant. AH: alkane hydroxylase; AD: alcohol dehydrogenase; ALD: aldehyde dehydrogenase; ACS: acyl-CoA synthetase; ω-H: ω -hydroxylase; BVM: Baeyer-Villiger monooxygenase; E: esterase; TCA: tricarboxylic acids cycle. The image was extracted from Rojo (2009).

zymes that have limited distribution in different bacteria, the oxidation of fatty alcohols and fatty acids is more common among microorganisms (Rojo, 2009).

Since diesel oil is one of the dominant pollutants in harbor environments from where members of Roseobacter lineage were detected in abundance (Piña-Villalonga, 2012), and because no previous studies reported the genomic characterization for degrading diesel oil in roseobacters, in this thesis we have studied the genomic potential and the physiologic capabilities of roseobacters for degrading or tolerating diesel oil. Alkanes are the main component of diesel oil and also of crude oil (up to 50 %). They can be linear (n-alkanes), cyclic (cyclo-alkanes) or branched (iso-alkanes) (Rojo, 2009). As it was previously mentioned, some phylotypes of the Roseobacter lineage increased in the presence of diesel oil in microcosm experiments (Lanfranconi *et al*., 2010; Piña-Villalonga, 2012) and different Roseobacter isolates were able to grow using alkanes (Harwati *et al*., 2007, 2008). Both facts evidenced that roseobacters might have a role in the degradation of alkanes. Moreover, a recent study proved that some members of the Roseobacter lineage (specifically the type species of *Phaeobacter gallaeciensis* and *Roseobacter denitrificans*) were able to decompose brominated hydrocarbons ($CHBr_3$) (Ichikawa *et al*., 2015), as shown in Table 1I.

HYPOTHESIS AND OBJECTIVES

Until 2010, few genomes of isolates of the Roseobacter lineage were sequenced and, although some of the isolates were able to degrade aromatic compounds, most of them, and most of the isolates sequenced later, were not obtained from polluted environments or described as pollutant degraders. Thus, the genomic information available on aromatic (or aliphatic) compound degradation by roseobacters was gathered from isolates that were not linked to pollutants. On the other hand, in previous projects of the research group new Roseobacter isolates were obtained from the Coast of Mallorca Island. Most of the strains were obtained from harbors (i.e. chronically-polluted environments) and were able to grow on different aromatic compounds (Piña-Villalonga, 2012). One of the aims of this thesis was to sequence and analyze the genomes of a selection of nine of these isolates. This provided the unique opportunity of studying a group of roseobacters mostly isolated from chronically-polluted environments (harbors in our case) and from a limited geographic area (Coast of Mallorca).

With this background, the first hypothesis that we proposed was that because these nine Roseobacter isolates were obtained from chronically-polluted environments they would have better (or different) aromatic compound degradation capabilities in comparison to members of Roseobacter lineage isolated from non-polluted environments. The rationale behind this hypothesis was that hydrocarbons, usually present in harbor waters, could be a selection factor favoring the development of populations adapted and/or able to exploit this resource. In relation to this, the second hypothesis was that the Roseobacter harbor isolates would have abilities to cope (tolerate) the presence of diesel oil, commonly present in these environments. Finally, the fact that the nine Roseobacter isolates came from the same geographical area (Mallorca Island), the same type of environment (harbor waters) and they were isolated in a limited period of time (September 2008 and March 2009) offered an unique opportunity for analyzing genetic exchange mechanisms. Therefore, the third hypothesis was that although the nine isolates could belong to different genera and/or species, the sharing of habitat, lifestyle and limited geographical origin would increase the chances for horizontal genetic exchange among them. Therefore, we expected to evidence more signs of horizontal gene transfer between these nine isolates in comparison with roseobacters that came from different locations and habitats.

For testing these hypotheses the following objectives were proposed:

1.- Identify and genomically characterize nine Roseobacter isolates from harbors of Mallorca Island in reference to their metabolic capabilities with particularly interest in the characterization of their ability for aromatic compounds degradation.

2.- Analyze the proteins that could be involved in diesel oil tolerance or degradation of the nine Roseobacter isolates and compare the response of the different isolates to diesel oil.

3.- Determine the relevance of three mechanisms of horizontal gene transfer: gene transfer agents (GTAs); extrachromosomal elements (plasmids), and transposases in genomes of the Roseobacter lineage.

MATERIALS AND METHODS

1. DNA extraction

The nine isolates studied in this thesis were the following: *Phaeobacter* sp. 11ANDIMAR09, *Ruegeria* sp. 6PALISEP08, *Sulfitobacter* sp. 1FIGIMAR09, *Sulfitobacter* sp. 3SOLIMAR09, *Loktanella* sp. 1ANDIMAR09, *Loktanella* sp. 3ANDIMAR09, *Loktanella* sp. 5RATIMAR09, *Thalassobacter* sp. 1CONIMAR09 and *Thalassobacter* sp. 16PALIMAR09. All of them were obtained in March 2009 except 6PALISEP08 (from September 2008) by plating in marine agar (MA). The three letter code in the names indicates the isolation location: AND, Andratx harbor; PAL, Palma harbor; FIG; Cala Figuera harbor; SOL, Sóller harbor; RAT, Cala Ratjada harbor; CON, control unpolluted coastal site. More information on the characteristics of the harbors can be found in Piña-Villonga (2012).

For DNA extractions the isolates were cultured in marine broth (MB) at room temperature and 180 rpm until they reached the exponential growth phase (OD_{600} between 0.4 and 0.8). Cells were collected by centrifugation at 15700 x g for 3 minutes. The supernatant was discarded and the cell pellets were resuspended and washed with Ringer solution (NaCl 0.85 % wt/vol). After this, cells were collected again by centrifugation. DNA extraction was done with the Wizard Genomic DNA Promega® Kit following manufacturer's recommendations. The quality and quantity of extracted DNA was checked with a Nanodrop 2000c spectrophotometer (Thermo Scientific) and by agarose gel electrophoresis. Amplification and sequencing of 16S rDNA was performed as described previously (Lanfranconi *et al.*, 2010; Piña-Villalonga, 2012) in order to discard contaminations.

2. Genome sequencing and assembly

Genome sequencing was done using Illumina technology by the company Lifesequencing (Paterna, Spain, http://www.lifesequencing.com/). The quality of the reads was tested using the program FastQC (Andrews, 2010). We considered that the quality of the reads was acceptable if the average Q score of each nucleotide position was higher than 30 (a Q30 value indicated that the probability of incorrect base call was 1 in 1000, so the base call accuracy was 99.9 %). Between 2 to 4 million reads per genome were assembled with Newbler version 2.9 (454 Life Sciences) using the criterion of a minimum of 60 overlapped nucleotides with at least 98 % identity. The definitive contigs were generated by the comparison of assemblies made with different sets of reads using the Mauve genome alignment software v. 2.3.1 (Darling *et al*., 2004) with default parameters. Only contigs larger than 500 bp were considered for the analysis. The genomes of the nine isolates were annotated using the Prokaryotic Genome Automatic Annotation Pipeline (PGAAP) at National Center for Biotechnology Information (NCBI) (Angiuoli *et al*., 2008). Annotated assemblies were published in the GenBank database (Clark *et al.,* 2016) at the NCBI.

3. Identification and phylogeny of isolates based on genome data

The identification of the isolates based on genome data was done by: i) 16S rDNA sequence identity with the type strains of the EzTaxon database (Kim *et al*., 2012); ii) Average Nucleotide Identity based on BLASTp (ANIb) and based on MUMmer software (ANIm) calculated using the JSpecies program (Richter & Rosselló-Móra, 2009); iii) virtual DNA-DNA hybridization (DDH) formula 2 (recommended for draft genomes) using Genome-to-Genome Distance Comparison (GGDC) (Meier-Kolthoff *et al.*, 2013), and iv) based on a phylogenetic analysis which included other 87 sequenced Roseobacter genomes in the database in August 2014 (96 genomes in total, see details in supplementary Table 1S.1). For this purpose, these genomes were downloaded from the database and, together with the genomes obtained in this study, they were annotated using Prokka program v. 1.7.2 (Seemann, 2014) to have a uniform annotation method. After annotation, genomic features [size, number of genes, number of coding sequences (CDSs) and mol % G+C content] were retrieved using the Ugene software v. 1.12.3 (Okonechnikov *et al.,* 2012).

As a previous step to perform the phylogenomic analysis with the 96 genomes, the annotated proteins were grouped under the criterion of 50 % of identity in at least 50 % of amino acid sequence with CD-Hit program (Huang *et al*., 2010). With this information we defined the pan proteome, the core proteome, the accessory proteome and the singletons. The pan proteome consists of all Cluster of Orthologous Groups of proteins (COGs) and can be divided in: i) core proteome which includes the COGs whose proteins were codified in all genomes, ii) the accessory proteome which includes the COGs whose proteins were codified in two or more genomes (but not all), and iii) singletons which includes COGs whose proteins were exclusively codified in one genome. The functional characterization was done by sequence homology using a reference protein of each COG (the one with longest amino acid sequence) against the Clusters of Orthologous Groups of proteins database (COGs) (Tatusov *et al*., 2000). The criterion for considering that the proteins were homologous was an *E* value below $1x10^{-5}$.

The phylogeny was based on the concatenation of the conserved proteins (114 proteins) of the core proteome whose genes were present in single-copy in the genome. Sequence alignment was done with Clustal Omega (Sievers *et al*., 2011). All poorly aligned and divergent positions were eliminated of the concatenated protein alignment using Gblocks program version 0. 91b with default parameters (Talavera & Castresana, 2007). The phylogeny was done with PROTPARS program from the PHYLIP package (Felsenstein, 1989) using a bootstrap of 100. Trees were drawn using TreeView (Page, 1996) and the Interactive Tree of Life tool (iTOL) (Letunic & Bork, 2007). We also calculated distance matrices using Kimura's model and the PROTDIST program from PHYLIP package (Felsenstein, 1989).

4. Survey of genes for relevant biogeochemical metabolisms

We characterized the genomes of the nine isolates in reference to some of the main ecological features previously described in other Roseobacter members (Luo & Moran, 2014; Newton *et*

al., 2010). The genomic characterization was done by sequence identity using BLASTp (Altschul *et al*., 1990) with proteins of other roseobacters that were well-described [chemolitotrophy, phototrophy and dimethylsulphoniopropionate (DMSP) metabolism] (see in Table 1M). If it was not indicated, the criterion for considering that two proteins were homologous was a minimum of 30 % sequence identity and an *E* value below 1 x 10^{-5}. As can be seen in Table 1M we analyzed the presence of genes for oxidation of carbon monoxide and sulfur compounds, carbon fixation by RuBisCO, genes for aerobic anoxygenic phototrophy, and DMSP degradation pathways.

For searching the genic products of flagella we used the automatic annotation server KAAS from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Moriya *et al*., 2007) using a threshold of 30 and a group of microorganisms that have flagella (including members of Roseobacter lineage). The bacteria used for this purpose were: *Escherichia coli* K-12 MG1655, *Pseudomonas aeruginosa* PAO1, *Pseudomonas putida* F1, *Ruegeria pomeroyi*, *Ruegeria* sp. TM1040, *Jannaschia* sp. CCS1, *Roseobacter denitrificans*, *Dinoroseobacter shibae* and *Ketogulonicigenium vulgare* Y25. We considered that when more than 70 % (25 proteins) of the flagellar assembly proteins (36 proteins in total) were detected in the annotated genome, the isolate would harbor the genic potential for assembling the flagella.

For the analysis of the potential for surface association or colonization, we studied the presence of pili genes. For searching the genomic island that codifies for the Flp pilus in the nine Roseobacter isolates, the proteins encoded by the pilus assembly cluster of the well-characterized *Caulobacter crescentus* NA1000 (genbank accession:CP001340.1) were used as reference (Skerker & Shapiro, 2000). In this case, because the reference strain did not belong to the *Rhodobacterales* order, in some cases the proteins that matched with the reference protein had low values of identity (below 30 %) and higher *E* values (above 1 x 10^{-5}). However, if the gene order in the genomic island that codified by the Flp pilus was conserved, we considered the isolates as a positive for harboring the genic products for Flp pilus.

5. Aromatic and aliphatic compounds degradation

5.1. Genomic characterization

For the genomic characterization of aromatic compound degradation genes, the criterion of protein identity and *E* value was the same as defined above. Gene products for the central degradation pathways that were previously described for roseobacters (see in section 3.1 of introduction) were searched through the nine isolates (see Table 2M).

Gene products involved in the pathway, i.e. from those transforming the central aromatic compound to the final intermediates of the TCA cycle, were analyzed. In the case of the benzoate degradation pathway, apart from the *box* genes that were commonly described in roseobacters (Moran *et al*., 2007), the *ben* genes were also analyzed because they were recently detected in *Roseobacter* sp. MED193 (Alejandro-Marín, personal communication). For

Table 1M. Biogeochemical features analyzed in the nine Roseobacter isolates and the microorganisms used as reference.

Feature	Reference genome	Reference
Carbon monoxide oxidation	*Ruegeria pomeroyi* DSS-3^T	Moran *et al.*, 2004
Type I and II of carbon monoxide dehydrogenase (CoxMLS)		
Sulfur oxidation	*Ruegeria pomeroyi* DSS-3^T	Moran *et al.,* 2004
Sulfide oxidation (SoxVWXYZABCDF)		
Regulators (SoxRS)		
Carbon fixation	*Rhodobacter sphaeroides* 2.4.1^T	Gibson *et al*., 1991, Tabita *et al*., 2007
RuBisCO (RbcLS,RbpL)		
Aerobic anoxygenic phototrophy	*Roseobacter denitrificans* OCh 114^T	Swingley *et al.*, 2007*,* Liotenberg *et al*., 2008
Bacteriochlorophyll biosynthesis (BchCXYZ, BchFNBHLM)		
Carotenoid biosynthesis (CrtCDEF)		
Light harvesting (PucBA, PufAB)		
Reaction center (PuhA, PufLM)		
Regulators (PpsR, TspO and PpaA)		
Dimethylsulphoniopropionate metabolism	*Ruegeria pomeroyi* DSS-3^T	Reisch *et al.*, 2013
Demethylation (DmdA)		
Cleavage (DddP, DddQ, DddW)		
Flagella motility	KEGG	Moriya *et al*., 2007
MotAB		
FlhABCD		
FlgABCDEFGHIKLMN		
FliCDEFGHIJKMNOPQRS		
Adhesion (Flp pili)	*Caulobacter crescentus* NA1000	Skerker & Shapiro, 2000
CpaBC-OmpA-CpaEF-TadBC		

Table 2M. Degradation pathways analyzed in the nine Roseobacter isolates and the microorganisms used as reference.

Degradation pathway	Reference genome	Reference
Benzoate (via benzoyl-CoA)	*Ruegeria pomeroyi* DSS-3^T	Moran *et al*., 2007
Benzoate-CoA ligase (BclA)		
Benzoyl-CoA 2,3-epoxidase (BoxAB)		
Dihydrolase (BoxC)		
Benzoate (via catechol)	*Roseobacter* sp. MED193	Alejandro-Marín (unpublished)
Benzoate 1, 2-dioxygenase (BenABC)		
Benzoate diol dehydrogenase (BenD)		
Phenylacetate	*Ruegeria pomeroyi* DSS-3^T	Moran *et al.*, 2007
Phenylacetyl-CoA ligase (PaaK)		
1, 2-phenylacetyl-CoA monooxygenase (PaaABCDE)		
Additional degradation genic products (PaaFGHIJK, PaaYXZ)		
Homoprotocatechuate	*Jannaschia* sp. CCS1	Moran *et al.*, 2007
Homoprotocatechuate 2, 3 dioxygenase (HpaD)		
Additional degradation genic products (HpaFGHI)		
Homogentisate	*Ruegeria pomeroyi* DSS-3^T	Moran *et al.*, 2007
4-hydroxyphenylpyruvate dioxygenase (HppD)		
Homogentisate 1, 2-dioxygenase (HmgA)		
Additional degradation genic products (HmgCB)		
Gentisate	*Ruegeria pomeroyi* DSS-3^T	Moran *et al.*, 2007
Gentisate 1, 2-dioxygenase (NagI)		
β-ketoadipate (via catechol, *ortho* cleavage)	*Pseudomonas putida* ND6	Peng *et al*., 2008
Catechol 1, 2 dioxygenase (CatA)		
β-ketoadipate (via protocatechuate, *ortho* cleavage)	*Ruegeria pomeroyi* DSS-3^T, *Sagittula stellata* E-37^T, *Citreicella* sp. SE45	Alejandro-Marín *et al.*, 2014
4-Hydroxybenzoate hydroxylase (PobA)		
Protocatechuate 3, 4 dioxygenase (PcaGH)		
Additional degradation genic products (PcaCDQBIJFR)		
Catechol (via *meta* cleavage)	*Pseudomonas putida* G7	Peng *et al*., 2008
Catechol 2, 3 dioxygenase (NahH)		
Protocatechuate (*meta* cleavage)	*Maritimibacter alkaliphilus* HTCC2654	Alejandro-Marín *et al.*, 2014
Protocatechuate 4, 5-dioxygenase		
Alkane hydroxylase	*Alcanivorax borkumensis* SK2^T	Misawa *et al*., 2004
Alkane hydroxylases (AlkB1 and AlkB2)		

protocatechuate degradation, in addition to the proteins of the *ortho* cleavage pathway (β-ketoadipate pathway) we searched for the dioxygenase catalyzing the *meta* cleavage of this compound, protocatechuate 4, 5-dioxygenase, and an hypothetical protein (DUF849) that was associated with *pca* genes in members of Roseobacter lineage (Alejandro-Marín *et al*., 2014; Buchan *et al*., 2004). We also searched both ring-cleaving proteins (*meta*- and *ortho*-cleavage) of catechol because i) it can be a product derived from salycilate (Peng *et al*., 2008), and ii) Alejandro-Marín and co-workers (2014) studied the catechol *ortho*- cleavage in their analysis of the β-ketoadipate pathway. Additionally, for aliphatic hydrocarbon degradation two alkane monooxygenases (AlkB1 and AlkB2) were also searched. In this case, percentages of identity below 30 % and *E* values above $1x10^{-5}$ were accepted because the reference genome did not belong to Roseobacter lineage. We also revised the genome annotation as an alkane monooxygenase or hydroxylase.

5.2. Growth on aromatic compounds and diesel oil

All isolates were cultured in 20 ml of a mineral medium with defined composition (MMM) supplemented with yeast extract (0.05 % w/v) (MMMb) at room temperature and at 180 rpm. MMM medium had the following composition: 4 % Sea Salts (Sigma-Aldrich), Tris-HCl 0.1 M pH 7.4, Na_2HPO_4 6.33 µM, NH_4Cl 56,09 µM and ammonium ferric citrate 0.03 µM. Inocula were prepared in MMMb with succinate at 0.5 % (v/v). Cultures were incubated until they reached the exponential phase (OD_{600} between 0.4 and 0.8) and then transferred (1/100 inoculum) to media with the following substrates: i) monoaromatic compounds (benzoate, phenylacetic acid, homoprotocatechuate, homogentisate, gentisate, salicylate and protocatechuate) at 3 mM concentration, and ii) diesel oil at 0.1 % v/v. These cultures were incubated for three weeks. All cultures were done in triplicates.

6. Proteomic analysis

Given the evidences that i) members of Roseobacter lineage were abundant in chronically hydrocarbon-polluted marine environments (Aguiló-Ferretjans *et al*., 2008; Nogales *et al*., 2007), ii) certain Roseobacter populations increased after addition of diesel oil (Lanfranconi *et al*., 2010; Piña-Villalonga, 2012), iv) some of the tested isolates showed growth with diesel oil, and v) there were not previous studies of the proteins involved in the degradation of diesel oil by roseobacters, we decided to perform proteomic analysis in order to know the proteome that was expressed when the isolates were exposed to this compound.

6.1. Incubation conditions for proteomic analysis

Seawater was collected from a pristine site at the South coast of Mallorca Island (39°21'28.4"N 2°54'52.2"E), in February 2015. The water was filtered successively through polyvinylidene fluoride (PVDF) filters (Millipore) with pore sizes of 5 µm and 0.22 µm. Then, water was autoclaved at 121 °C for 20 minutes. Commercial diesel oil was sterilized by filtration using polytetrafluoroethylene (PTFE) filters (Millipore) with a pore size of 0.20 µm.

The nine isolates were cultured separately in MB at room temperature and 180 rpm until they reached the exponential phase (OD_{600} between 0.4 and 0.8). Then, cells were collected by centrifugation at 18,500 x g for 6 minutes using a centrifuge 5810 R (Eppendorf). The cells were washed with sterile seawater and they were resuspended in the appropriate volume of sterile seawater to achieve an OD_{600} of 0.3. After this, the bacterial suspension was divided in 50 ml aliquots in 6 sterile glass flasks. Three of them were supplemented with 0.1 % (v/v) of filtered diesel oil. Flasks were incubated at room temperature at 180 rpm for 24 hours. After that, the cells were collected in 50 ml tubes by centrifugation (see above). The pellets were resuspended in 1 ml of sterile seawater and the cells were collected again by centrifugation at 15,700 x g for 6 minutes using a centrifuge 5415 R (Eppendorf). The supernatant was removed and the pellet (between 20 and 50 mg of wet biomass per sample) was kept at −80 °C and sent under dry ice to the University of Warwick (United Kingdom) for further processing.

6.2. Sample processing for proteomic analysis

Cell extracts were resuspended in two volumes (2 µl per 1 mg of biomass) of NuPAGE® LDS Sample Buffer (Invitrogen). Four cycles of 5 minutes at 95°C, 1 minute vortex and 5 minutes of incubation in an ultrasonic bath (Branson 1210 Ultrasonic Cleaner) were done in order to lyse the cells. The samples were run in a NuPAGE® Bis-Tris polyacrylamide gel (Invitrogen). Short electrophoresis migrations were carried out at 200 V (Mini Protean Tetra cell, Biorad) until the sample entered the gel. The gel was stained for 1 hour with Coomassie SimplyBlue™ SafeStain (Invitrogen). Polyacrylamide gel bands that contained the proteome fraction were cut in small pieces and kept at −20 °C until in-gel proteolysis with trypsin.

Protein gel pieces were washed twice with ammonium bicarbonate 50 mM in ethanol 50 % (v/v) for 20 minutes at 55 °C. They were dehydrated at 55 °C in agitation using absolute ethanol for 5 minutes. Dry gel pieces were rehydrated for 45 min at 56 °C with 50 mM ammonium bicarbonate containing 10 mM dithiothreitol (DTT) to reduce disulfide bonds. Gel pieces were then treated with 50 mM ammonium bicarbonate containing 55 mM iodoacetamide (IAA) (an alkylating agent that reacts with free sulfhydryl groups of cysteine residues to form S-carboxyamidomethyl-cysteine) for 30 minutes at room temperature in the dark. Gel fractions were then washed and dehydrated as described above. For in-gel protein digestion, gel pieces were rehydrated with 40 µl of 50 mM ammonium bicarbonate containing 2.5 ng/µL of trypsin (Promega). After overnight proteolysis at 37 °C, digested peptides were submerged in formic acid 5 % (v/v) in acetonitrile 50 % (v/v) and incubated 3 times for 10 minutes in an ultrasonic bath (Fisherbrand FB15062 ultrasonic bath) in sweep position in order to extract the digested peptides out of the gel. The gel pieces were removed, the peptides were dried completely at 40 °C under vacuum (miVac DUO concentrator, GeneVac) until the solvent was evaporated, and they were stored at −20 °C until further analysis. Peptides were resuspended in formic acid 1 % v/v in acetonitrile 2 % v/v. An aliquot of 20 µL of peptides was analyzed by Nano-scale Liquid Chromatographic tandem mass spectrometry (nanoLC-ESI-

MS/MS) with an Ultimate 3000/Orbitrap Fusion instrumentation (Thermo Scientific) using conditions previously described (de Groot *et al*., 2009).

The identification and relative quantification of proteins in the samples was done with the program Progenesis QI for proteomics (Nonlinear Dynamics). All runs were aligned to a reference run automatically chosen by the software and the peaks intensities were normalized in reference to the total intensity. Then, a master list of features considering *m/z* values and retention times was generated. The mass peaks with charge states from +2 to +5 were selected for the study in order to discard the peptides or proteins that were not well-digested by trypsin. All MS/MS spectra were exported from Progenesis software as a MASCOT generic file (mgf) and used for peptide identification with MASCOT search engine v. 2.5.0 (Matrix Science) against a polypeptide sequence database for each isolate according to the information provided by genome sequencing and annotation. This step allowed us to correlate the theoretical spectra obtained from the virtual trypsin digestion of the proteome database with the experimental MS/MS spectra, so we could assign a peptide identity to the MS/MS spectra. The identified peptides were imported again to Progenesis where they were matched to the respective features. For considering that a protein was detected they had to have a minimum of 2 non-redundant peptide hits. For selecting the differentially detected proteins we chose the following values for considering that the parameters were statistically significant: minimum fold change ≥ 2 and ANOVA p-value ≤ 0.05.

6.3. Proteome data analysis

In order to determine which fraction of the proteome was common to all analyzed isolates and functionally characterize the detected proteins, we studied if these proteins belonged to the core proteome of these isolates, to the accessory proteome or if they were singletons. For this, a clustering of the genic products of the annotated genomes were made using the criteria of 50 % of identity in at least 50 % of amino acid sequence (50_50) using the program CD-Hit (Huang *et al*., 2010). This clustering was used as a reference to classifiy the Clusters of Ortologous Groups of Proteins (COGs) of detected proteome in the different proteomic categories (core proteome, accessory proteome or singletons). The functional assignation of the COGs was made using the reference sequence of each detected COG against the Clusters of Orthologous Groups of proteins (COGs) database (Tatusov *et al*., 2000).

In order to determine the differences between the proteomes in the two tested incubation conditions (with and without diesel) and among the isolates, a multivariate analysis, Principal Component Analysis (PCA), was done using Past program v. 3.08 (Hammer *et al*., 2001). In order to compare the results of the different isolates we expressed the values of normalized abundance of Progenesis of each protein as a percentage of the total. Different datasets were used for PCA: i) all detected proteins, ii) COGs that were obtained under different criteria [50 % of identity in at least 50 % of amino acid sequence (50_50); 75 % of identity in at least 75 % of amino acid sequence (75_75), and 90 % of identity in at least 95 % of amino acid sequence (90_95)], iii) functional categories obtained from the analysis made with the KEGG (Moriya *et al*.,

2007), iv) lipoproteins, cytoplasmic proteins, transmembrane proteins and signal peptides obtained from the analysis with LipoP Server (Juncker & Willenbrock, 2003), v) cytoplasmic proteins, cytoplasmic-membrane proteins, extracellular proteins, outer membrane proteins, periplasmic proteins and unknown or not-classified proteins using PsortB Server (Nancy *et al*., 2010), and, vi) membrane transport proteins families identified and obtained using the Transporter Classification Database (Saier *et al*., 2014). The selection of the previous datasets was justified because there were previous evidences that the exposure to hydrocarbons could cause changes in cell membrane and transport in bacteria (Bernal *et al*., 2007; García *et al*., 2010; Ramos *et al*., 2015). We also did PCA analysis with the proteome data of each isolate separately in the two incubation conditions. For this, we used data of the abundances normalized given by Progenesis of the proteins clustered using the criterion 90_95 (see above). The proteins that were more relevant for the ordination of the samples in the PCA analysis plus the proteins that were relevant by ANOVA test of Progenesis were analyzed in detail. BLASTp searches (Altschul *et al*., 1990) against UniProtKB/Swiss-Prot (Magrane *et al*., 2011) and non-redundant databases of National Center for Biotechnology Information (NCBI) (Wheeler *et al*., 2007) for each of the selected proteins was done. Those proteins observed in the proteome of one isolate that we considered that could be related to the response to the environmental conditions or could be involved in the tolerance or degradation of diesel oil were selected and searched in the proteomes of other isolates by sequence homology using BLASTp (sequence identity higher than 30 % and *E* value below 1×10^{-5}).

On the other hand, a functional analysis of all detected proteins was done by sequence homology using BLASTp against the COGs database (Tatusov *et al*., 2000). In this case, we used the sum of the percentages of the normalized abundances of the proteins of each functional category (for comparing the results of the different isolates). The percentages were calculated in reference to the sum of the abundandes of all proteins per isolate. Average values and standard deviation for the three technical replicates per condition were calculated. T-student test was done for comparing the percentages of abundances, and we considered that the results were significant when p-value was ≤ 0.05.

Apart from these strategies, a targeted search of proteins that could be important in reference to the degradation and/or tolerance response to diesel oil according to the literature (Ramos *et al*., 2015; Rojo, 2009; Segura *et al*., 2005; Wijte *et al*., 2011) was done. For the analysis of the degradation response we focused on the search of monoaromatic compound degradation proteins, alkane hydroxylases (see Table 2M), and oxygenases involved in the the first step of polyaromatic compound degradation. The search was done by sequence homology using the Aromadeg database (Duarte *et al*., 2014) and by the name of the annotated proteins. Apart from this, the detected proteins that had an annotation that could be related to the degradation of alkanes or aromatic hydrocarbons were checked by sequence homology against UniProtKB/Swiss-Prot (Magrane *et al*., 2011) database and NCBI nr database (Wheeler *et al*., 2007). On the other hand, for studying the tolerance response to diesel oil, we based on a survey that Ramos and collaboratorators published recently in which they analyzed the

mechanisms of solvent resistance by toluene-tolerant *Pseudomonas putida* strains (Ramos *et al*., 2015). These tolerance mechanisms included: changes in microbial cell membranes in response to solvents, removal of reactive oxygen species, expression of chaperones, efflux pumps and energy production (Ramos *et al*., 2015). Furthermore, as another mechanism to face the presence of diesel oil, we studied proteins that could have a role in the attachment to hydrocarbon-water interface, such as pili formation proteins. It has been suggested that *Alcanivorax borkumensis* SK2, an specialist alkane degrader, uses pili to have an easy access to the hydrocarbon (Schneiker *et al*., 2006). For the changes in cell membranes we based in the annotation of the proteins that were upregulated in diesel oil. For the removal of reactive oxygen species we analyzed the upregulated proteins that belonged to (O) "post-translational modification, protein turnover and chaperones" and to the (P) "inorganic transport and metabolism" COG functional categories. In the case of chaperones we studied the upregulated proteins in diesel oil of (O) "post-translational modification, protein turnover and chaperones" COG functional category. For searching the upregulated proteins related to energy production, we analyzed the proteins of (C) "energy production and conversion" COG functional category. Finally, for the pilus assembly protein, we analyzed the upregulated proteins of (U) "intracellular trafficking, secretion, and vesicular transport" COG functional category. For the identification of transporters, they were detected and classified by sequence homology using BLASTp (Altschul *et al*., 1990) against the Transporter Classification Database (TCDB) (Saier *et al*., 2014). Percentages of amino acid identity higher than 30 % and *E* values below $1x10^{-5}$ were considered as valid hits. The percentages of abundances of each category superfamily were used to compare the results. Average and standard deviation of the three replicates per condition were calculated. T-student test was done for comaring the percentages of abundances, and we considered that the results were significant when the p-value was ≤ 0.05.

7. Horizontal gene transfer (HGT) processes

In this thesis we have searched for three mechanisms of HGT (GTAs, ECEs and transposases) in 96 Roseobacter genomes by sequence homology using BLASTp (Altschul *et al*., 1990) considering percentages of amino acid identity higher than 30 % and an *E* value below $1x10^{-5}$.

7.1. Gene transfer agents (GTAs)

For GTA identification, the well-characterized model RcGTA of *Rhodobacter capsulatus* (Lang *et al*., 2000, 2002) (Genbank reference: AF181080.3) was used. Organisms meeting the following criteria were classified as positive for GTA genes: (1) possession of homologs of at least 8 out of the 17 genes present in RcGTA and, (2) conservation of GTA gene order and orientation. In some cases, the proteins were not annotated in the published genomes and we annotated them manually using ExPASy program (Artimo *et al*., 2012) and checked this annotation with BLASTp (Altschul *et al*., 1990) against the nr database of NCBI (Wheeler *et al*., 2007). Firstly, a comparison of gene order and size of the GTA cluster in all genomes was done by checking the annotation of the genomes using Ugene software v. 1.12.3 (Okonechnikov *et al.,* 2012).

Secondly, we selected the genes that were conserved in all GTA structures for making a concatenation of the genic products. A consensus phylogenetic tree was done by parsimony following the procedure explained in section 3.

7.2. Extrachromosomal elements (ECEs)

Four characteristic replication/partioning modules of extrachromosomal replicons have been described in the Roseobacter lineage (Petersen *et al*., 2011) according to the replicase: RepABC family, RepA, RepB and DnaA-like. For the identification of RepABC family plasmids we used the sequence of the replicase protein RepC using BLASTp (Altschul *et al*., 1990) considering the same criterion of identity and *E* value explained above. Sequences belonging to the 9 different compatibility groups of Roseobacter lineage described (Petersen *et al*., 2009) were used as reference (see Table 3M). Only replicons containing a complete *repABC* operon were considered. The detection of partioning system (*repAB*) was done by searching the annotation of the genes that were located near *repC* and forming a plausible *repABC* operon.

Table 3M. RepC proteins used as reference for the different compatibility groups (Petersen *et al*., 2009).

Compatibility group	Roseobacter genome	GenBank Accession No.
C1	*Sulfitobacter* sp. NAS-14.1	EAP78773.1
	Pseudooceanicola batsensis HTCC2597^T	EAQ01034.1
C2	*Roseovarius* sp. 217	EAQ24825.1
	Dinoroseobacter shibae DFL 12^T	ABV95748.1
C3	*Sagittula stellata* E-37^T	EBA07819.1
	Oceanicola granulosus HTCC2516^T	EAR52081.1
	Rhodobacteraceae bacterium HTCC2150	EBA02778.1
C4	*Oceanibulbus indolifex* HEL-45^T	EDQ03503.1
	Sulfitobacter sp. NAS-14.1	EAP78814.1
C5	*Roseobacter* sp. MED193	EAQ43764.1
C6	*Sulfitobacter* sp. NAS-14.1	EAP78998.1
	Sulfitobacter sp. NAS-14.1	EAP78659.1
C7	*Oceanibulbus indolifex* HEL-45^T	EDQ03745.1
	Roseovarius sp. 217	EAQ22999.1
C8	*Sulfitobacter* sp. EE-36	EAP82486.1
	Oceanibulbus indolifex HEL-45^T	EDQ03325.1
C9	*Roseovarius* sp. TM1035	EDM30090.1
	Roseobacter denitrificans OCh 114^T	ABI93326.1

Petersen and co-workers (2009) defined 9 compatibility groups with 30 genomes of Roseobacter. Now, with 96 analyzed genomes we wanted to reevaluate the compatibily groups. For this, we calculated three different parsimony-based phylogenies using RepA, RepB and RepC (calculated as explained above in section 3). Considering that two plasmids of the same compatibility group cannot coexist in the same bacteria, we defined compatibility (based on the phylogeny of RepC) and sub-compatibility groups of plasmids (based on the phylogenies of RepA and RepB). For determining if two or several genomes shared the same *repABC* module we considered that amino acid sequence of RepA, RepB and RepC had to be at least 95 % identical in 100 % of the sequence.

We also studied which putative RepABC plasmids could be transferred by conjugation. For this, we searched for genes of type IV secretion system (T4SS) in each RepABC putative plasmid.

The VirB4 ATPase, which is highly conserved, was used as marker of the presence of a T4SS (Alvarez-Martinez & Christie, 2009; Guglielmini *et al*., 2013). Other proteins such as pilin protein VirB2, channel proteins VirB6 and VirB8 and the coupling protein VirD4 were also searched to confirm the potential conjugative transferability of these plasmids (Guglielmini *et al*., 2013). We searched these proteins by sequence homology using BLASTp (Altschul *et al*., 1990) and used the proteins of *Agrobacterium tumefaciens* Bo542 Ti plasmid as model: VirB2 (AAZ50519.1), VirB4 (AAZ50521.1), VirB6 (AAZ50523.1), VirB8 (AAZ50525.1) and VirD4 (AAZ50535.1). Once the genic products were identified in a putative plasmid, the genic order and the annotation were checked to confirm that these proteins were T4SS proteins. If we detected at least, four of the five searched proteins in a putative plasmid, we considered that it was conjugative.

Finally, we searched the other replicases (RepA, RepB and DnaA-like) by sequence homology using BLASTp (Altschul *et al*., 1990) considering percentages of amino acid identity higher than 30 % and an *E* value below $1x10^{-5}$. We used as a reference the proteins shown in Table 4M (Petersen *et al*., 2011, 2013).

Table 4M. RepA, RepB and DnaA-like replicases used as reference.

Protein	Roseobacter genome	GenBank Accession No.
RepA	*Sulfitobacter* sp. NAS-14.1	EAP78520.1
	Phaeobacter inhibens DSM 17395	AFO93583.1
RepB	*Octadecabacter arcticus* 238^T	AGI74697.1
	Phaeobacter inhibens DSM 17395	AFO93522.1
	Sulfitobacter guttiformis KCTC32187^T	KIN75153.1
DnaA-like	*Oceanibulbus indolifex* HEL-45^T	EDQ03064.1
	Phaeobacter inhibens DSM 17395	AFO93520.1

7.3. Transposases

For detecting transposases we used two approaches: i) the Prokka annotation of the genomes and ii) using BLASTp searches against a transposase database (percentages of amino acid identity higher than 30 % and *E* value below $1x10^{-5}$) generated in this study with sequences obtained from the ISfinder database (Siguier, 2006). For the transposase database we selected one transposase of each IS family as representative (see Table 5M). The detected transposases were classified into Insertion Sequences (ISs) families (percentages of amino acid identity higher than 30 % and *E* value below $1x10^{-5}$) by sequence homology using the BLAST tool of ISfinder database (Siguier, 2006).

On the other hand, we determined which transposases were shared between the analyzed genomes as a way to study recent HGT processes. For determining the shared transposases, they were clustered under the criteria of 90 % sequence identity in, at least, 95 % of their amino acid sequence with CD-Hit program (Huang *et al*., 2010). The visual representation of the shared Cluster of Orthologous Groups (COGs) of transposases was done using Gephi v. 0.9.1 (Bastian *et al*., 2009).

Table 5M. Transposases included in the database used as reference.

IS family	Bacterium	ISfinder code	GenBank Accession No.
IS1	*Escherichia coli* ECOR-50	U15127	AH003093.2
IS110	*Streptomyces coelicolor* A3	Y00434	Y00434.1
IS1380	*Acetobacter pasteurianus* NC11380	D90424	D10043.1
IS200	*Salmonella typhimurium* LT2	X56834	X56834.1
IS21	*Pseudomonas aeruginosa* PAO25	X14793	X14793.2
IS256	*Staphylococcus aureus* SK18	M18086	GU565967.1
IS3	*Escherichia coli* K12	X02311	X02311.1
IS30	*Escherichia coli* K12	X00792	X00792.1
IS4	*Escherichia coli* K12	J01733	J01733.1
IS481	*Bordetella pertussis* Tohama	M22031	M22031.1
IS5	*Escherichia coli* (bacteriophage λKH100)	J01735	J01735.1
IS6	*Proteus vulgaris* ATCC 29905	X00011	X00011.1
IS605	*Helicobacter pylori* CCUG 17874	HPU60177	U60177.1
IS607	*Helicobacter pylori* CPY0041	AF189015	AF189015.1
IS630	*Shigella sonnei* ATCC 29930	X05955	X05955.1
IS66	*Agrobacterium tumefaciens* Ach5	AF242881	AF242881.1
IS91	*Escherichia coli* ATCC 11775	X17114	X17114.5
IS982	*Lactococcus lactis* SK11	L34754	L34754.1
ISAs1	*Aeromonas salmonicida*	L27156	L27156.1
ISL3	*Lactobacillus delbrueckii* ATCC 11842	X79114	X79114.1
Tn3	*Comamonas testosteroni* BR60	M65135	M65135.1

RESULTS AND DISCUSSION

CHAPTER 1. GENOMIC AND PHYSIOLOGIC CHARACTERIZATION OF NINE ROSEOBACTER ISOLATES FROM MALLORCA COAST

In this chapter, we present the sequenced and annotated genomes of nine Roseobacter isolates from Mallorca coast. A confirmation of their identity and phylogenomic affiliation to the species level was done based on the 16S rDNA similarity, ANI, DDH hybridization and a phylogeny based in the defined core proteome. A new simplified method for establishing a phylogeny of Roseobacter lineage is proposed. The genomes were characterized in reference to biochemical features and aromatic and aliphatic compounds degradation capabilities. In addition, growth of isolates on different aromatic compounds and diesel oil as a sole source of carbon and energy was tested in order to evaluate their abilities for hydrocarbon degradation.

1. Genome assemblies

The genomes of the nine isolates were sequenced and annotated with the annotation pipeline of the NCBI (see the results in Table 1.1). The average genome size was 3.8 ± 0.5 Mb and the average of G+C content mol % was 58.7 ± 1.6 %. These values were in agreement with genomic data from described roseobacters (4.4 ± 0.6 Mb in size and 60 ± 4 G+C content mol %, Luo *et al*., 2014). When we compared the isolates, 6PALISEP08 and 11ANDIMAR09 highlighted for their larger genomic size and 3ANDIMAR09 and 3SOLIMAR09 for their higher G+C content mol %. With respect to the assemblies, isolate 11ANDIMAR09 highlighted for having the largest number of contigs.

Table 1.1. Genomic features of the nine Roseobacter isolates sequenced.

Isolates	Genbank Accession No.	Coverage (fold)	Genome size (Mb)	G+C content mol %	No. contigs[a]	No. of proteins[b]
11ANDIMAR09	LIKT00000000	73	4.7	58.2	101	4,334
6PALISEP08	LGXZ00000000	87	4.5	56.8	42	4,234
1FIGIMAR09	JEMU00000000	88	3.9	58.4	55	3,941
3SOLIMAR09	AXZR00000000	56	3.5	60.3	25	3,247
1ANDIMAR09	LIGP00000000	52	3.7	57.7	18	3,649
3ANDIMAR09	LJAK00000000	105	3.7	62.0	34	3,723
5RATIMAR09	LJAL00000000	105	3.7	57.8	43	3,572
1CONIMAR09	JGVS00000000	57	3.4	58.6	28	3,178
16PALIMAR09	JHAK00000000	48	3.5	58.8	39	3,321

a: longer than 500 bp.

b: Genbank annotation.

2. Taxonomic affiliation of the nine isolates based on genomic data

The isolates analyzed in this study were previously affiliated to the genus level (Piña-Villalonga, 2012). Since their isolation, new type strains of Roseobacter lineage were described. For this reason and in order to see if there were changes in the taxonomic affiliation of the isolates, we compared the similarity of the 16S rDNA gene of the nine isolates with the 16S rDNA gene database of type strains of Ez-Taxon (Kim *et al*., 2012). The results are shown in Table 1.2. In 2012, Piña-Villalonga identified three isolates as *Loktanella*, two as *Thalassobacter*, two as *Sulfitobacter*, one as *Ruegeria* and one as *Phaeobacter* by 16S rDNA sequencing. As expected, the identifications to the genus level were confirmed with genome data with the exception of *Phaeobacter* that now was identified as *Pseudophaeobacter*. At level of species, and according

to 16S rDNA similarity, the isolates were assigned to the closest validated species in Ez-Taxon database. Thus, 8 of these 9 isolates could be adscribed to at least 7 different species (isolate 3ANDIMAR09 could not be identified at species level due to the low sequence similarity to the closest species). Two pair of isolates had identical sequence in the aligned positions: 1ANDIMAR09 and 5RATIMAR09 and the pair 1CONIMAR09 and 16PALIMAR09.

Table 1.2. 16S rDNA-based identification of the isolates.

Isolate	Closest type strain	Identity (%)	Aligned positions (bp)
11ANDIMAR09	*Pseudophaeobacter arcticus* DSM 23566^T	98.9	1,418
6PALISEP08	*Ruegeria atlantica* CECT 4292^T	98.8	1,378
1FIGIMAR09	*Sulfitobacter mediterraneus* KCTC 32188^T	99.8	1,321
3SOLIMAR09	*Sulfitobacter pontiacus* DSM 10014^T	99.9	1,392
1ANDIMAR09	*Loktanella rosea* Fg36^T	99.7	1,392
3ANDIMAR09	*Loktanella salsilacus* LMG 21507^T	97.1	1,401
5RATIMAR09	*Loktanella rosea* Fg36^T	99.7	1,392
16PALIMAR09	*Thalassobacter stenotrophicus* CECT 5294^T	99.8	1,430
1CONIMAR09	*Thalassobacter stenotrophicus* CECT 5294^T	99.8	1,430

For better assignation at the species level we calculated the Average Nucleotide Identity (ANI) (Richter & Rosselló-Móra, 2009) and the virtual DNA-DNA hybridization (DDH) (Meier-Kolthoff *et al.*, 2013) based in genome data. The ANI was calculated with the nucleotide sequence of 96 Roseobacter genomes (see in supplementary Table 1S.1) that were available at the moment to perform the analysis (see in supplementary Tables 1S.2) while the virtual DDH was only calculated with Roseobacter genomes that had the highest ANI values with the isolates. The results are shown in Table 1.3. Values of ANI and DDH higher than 96 % and 70 %, respectively, defined genomically the circumscription to the same species (Richter & Rosselló-Móra, 2009; Rosselló-Móra & Amann, 2015).

Table 1.3. Average Nucleotide Identity (ANI) and virtual DNA-DNA hybridization (DDH) for the Roseobacter isolates and previously sequenced genomes.

Isolate	Closest Roseobacter genome	ANIb (%)	ANIm (%)	DDH (%)
11ANDIMAR09	*Pseudophaeobacter arcticus* DSM 23566^T	81.6	85.4	25.3
6PALISEP08	*Ruegeria atlantica* CECT 4292^T	78.5	86.0	22.3
1FIGIMAR09	*Sulfitobacter mediterraneus* KCTC 32188^T	95.7	96.3	68.7
3SOLIMAR09	*Sulfitobacter donghicola* DSW-25^T	71.8	83.0	18.5
	Sulfitobacter sp. EE-36	97.6	97.7	79.9
	Sulfitobacter sp.NAS-14.1	97.1	97.3	76.9
1ANDIMAR09	*Loktanella vestfoldensis* DSM 16212^T	74.3	83.3	19.2
3ANDIMAR09	*Loktanella vestfoldensis* DSM 16212^T	72.6	83.4	19.1
5RATIMAR09	*Loktanella vestfoldensis* DSM 16212^T	74.3	83.3	19.2
16PALIMAR09	*Thalassobacter stenotrophicus* CECT 5294^T	95.6	96.0	68.5
1CONIMAR09	*Thalassobacter stenotrophicus* CECT 5294^T	96.1	96.8	69.9

According to the results that we obtained, we considered that isolates 1CONIMAR09 and 16PALIMAR09 belonged to *Thalassobacter stenotrophicus* species, although in the case of 16PALIMAR09, the ANIb and virtual DDH were slightly below the cutoff. In fact, the ANIb value (99.4 %) and the virtual DDH value (96.1 %) of the comparison of these two genomes also indicated that they were the same species. A similar situation in which the ANIb and the virtual DDH values were slightly below the cutoff was in the case of *Sulfitobacter* sp. 1FIGIMAR09. Yet,

because the values were near the limits, the ANIm value was higher than the threshold and the homology values of 16S rDNA indicated that they were the same species, we considered that 1FIGIMAR09 isolate was a strain of *Sulfitobacter mediterraneus* species. *Sulfitobacter* sp. 3SOLIMAR09 could not be affiliated to any validly-described species, but the results showed that it belonged to the same species than strains *Sulfitobacter* spp. EE-36 and NAS-14.1 (see Table 1.3). None of the other isolates could be taxonomically affiliated to the species level because of the lack of sequenced genomes of Roseobacter lineage type strains. It is worth mentioning that although *Loktanella* spp. 1ANDIMAR09 and 5RATIMAR09 had identical 16S rDNA sequence, ANI values of the comparison of the two genomes (89.4 % of ANIb and 90 % of ANIm, see supplementary Table 1S.2) showed that in fact, they were not the same species. Therefore, based on genome data we can say that the harbor isolates analyzed in this study corresponded to 8 different species, but only 3 of them were adscribed with confidence to previously characterized species (the two *Thalassobacter stenotrophicus* isolates and *Sulfitobacter mediterraneus* 1FIGIMAR09).

Additionally, a core proteome based phylogeny with all the sequenced genomes that were available in the NCBI database (Wheeler *et al*., 2007) in June of 2014 was done. In supplementary Table 1S.1, a list of the 96 members of Roseobacter lineage used for the phylogeny plus their genomic features and place of isolation are shown. A total of 391,901 proteins from the 96 Roseobacter genomes (see supplementary Tables 1S.3 for the annotation) were clustered in COGs resulting in a panproteome of 88,805 COGs: 134 COGs belonged to the core proteome, 32,711 COGs to accessory proteome and 55,960 COGs were singletons (for more details about the clustering, see supplementary Table 1S.4). Because we were comparing genomes that were very different among them (i.e. different bacterial genera), the number of COGs in the calculated core proteome was low. Only 0.15 % of the COGs (core proteome/panproteome) had representatives of all the genomes. On the contrary, the percentage of singletons was very high (65 % of the COGs were exclusively found in one genome). The changes in the number of COGs that belonged to the core proteome or the panproteome as more genomes are included in the analysis demonstrated the high diversity of the lineage (see Figure 1S.1). As expected, the core proteome had the highest percentage of functional assignation with the majority of functions related to (J) “translation, ribosomal structure and biogenesis” and (O) “post-translational modification, protein turnover and chaperones” COGs categories. The singletons was the fraction with the lowest percentage of functional assignation with functions mainly related to (Q) “secondary metabolites biosynthesis, transport and catabolism” COG category (see supplementary Figures 1S.2 and 1S.3).

We selected the concatenated sequences of the conserved proteins in the core proteome that were present in single-copy in the genomes (114 proteins, listed in Table 1S.5) to perform a phylogenomic analysis of the 96 roseobacter genomes. In total, 21,876 amino acid aligned positions (64 % of the total sequence length) resulting from the exclusion of the highly variable parts of the concatenated sequences were used for the phylogeny. Up to date, this is the

Roseobacter lineage phylogeny made with the largest number of sequenced genomes and the highest number of concatenated proteins. In this new phylogeny (see Figure 1.1), the 5 phylogenomic groups that were defined in previous studies (Newton *et al*., 2010) were maintained. Additionally, a sixth phylogenomic group (G6) that was only constituted by two unclassified isolates (HIMB11 and HTCC2083) could be defined with high bootstrap support.

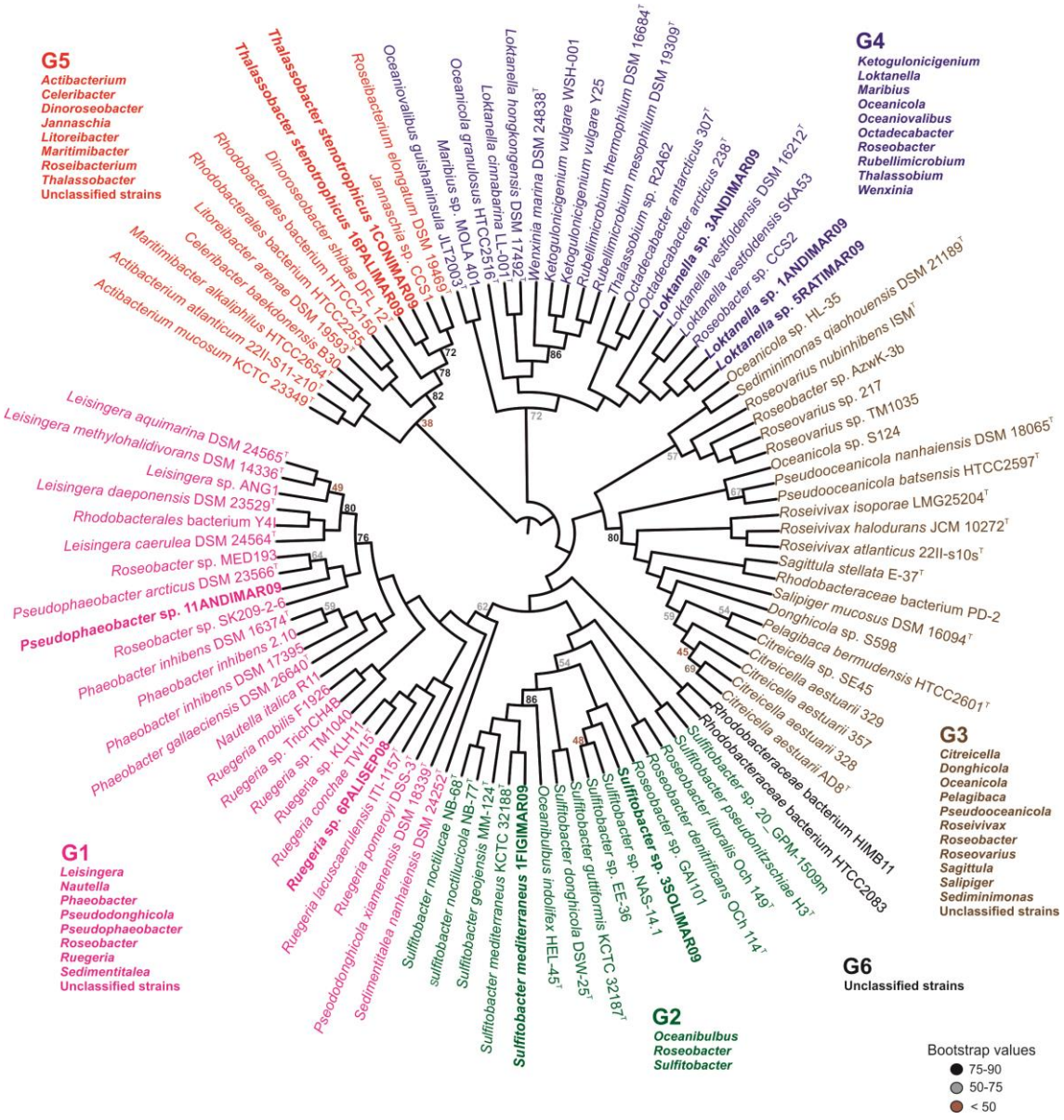

Figure 1.1. Phylogeny of the Roseobacter lineage (96 genomes) based on the concatenation of 114 conserved proteins. The phylogeny was done with PROTPARS package of Phylip program (Felsenstein, 1989) using parsimony and a bootstrap of 100. Only bootstrap values lower than 90 are shown. Phylogenomic group G5 was used as the outgroup. The nine harbor isolates are shown in bold.

In agreement with previous studies (Luo *et al*., 2012, 2014; Newton *et al*., 2010) the phylogeny showed that current Roseobacter taxonomy does not agree with the phylogeny of the lineage. There were species that are not yet classified taxonomically or had an incoherent assignment of scientific names. One example was the unclassified strain Y4I that according to the phylogeny and the ANI results (see supplementary Table 1S.2) should belong to *Leisingera*

daeponensis species. Another one is that the genera *Roseobacter* (with representatives found in G1, G2, G3 and G4 phylogenomic groups) and *Oceanicola* (found in G3 and G4) are clearly paraphyletic, or currently include misnamed strains. The phylogenetic analysis reinforces the evidence of the necessity to revise and improve the current taxonomy of roseobacters. In reference to the 9 harbor isolates analyzed in this study, they were affiliated to 4 different phylogenomic groups: *Ruegeria* sp. 6PALISEP08 and *Pseudophaeobacter* sp. 11ANDIMAR09 were affiliated to phylogenomic group G1; *Sulfitobacter mediterraneus* 1FIGIMAR09 and *Sulfitobacter* sp. 3SOLIMAR09 isolates were circumscribed to the phylogenomic group G2; the three *Loktanella* isolates were affiliated to G4 phylogenomic group, and finally, *Thalassobacter stenotrophicus* 1CONIMAR09 and 16PALIMAR09 were affiliated to phylogenomic group G5.

We calculated the phylogenomic distance values between genomes based on the 114 core protein data (see supplementary Table 1S.6 and for further details Figure 1S.4). The most homogenous phylogenomic group was G1 with an average of phylogenomic distance between genomes of 0.14 ± 0.04, followed by phylogenomic group G2 (0.15 ± 0.04). That was because in phylogenomic group G1, there were many genomes of only three genera (*Ruegeria*, *Leisingera* and *Phaeobacter*) and in phylogenomic group G2, most members belonged to the genus *Sulfitobacter*. In contrast, G3, G4, G5 and G6 phylogenomic groups were more heterogeneous in terms of genera than G1 and G2, and the average phylogenomic distances were also higher (0.21 ± 0.05, 0.26 ± 0.06, 0.28 ± 0.06 and 0.75, respectively).

3. Definition of a new method for establishing the phylogeny of the Roseobacter lineage

In general, there is an accelerated pace of genome sequencing (Land *et al*., 2015) and this is also true for Roseobacter genomes (Luo & Moran, 2014). In fact, from June of 2014 to April of 2016 more than 60 new Roseobacter genomes were available in databases. ANI values provide a relatively rapid affiliation of genomes to different genera. However, as the number of genomes increases the calculation time also increases and computer resources have to be more powerful. On the other hand, in order to define phylogenetic relationships within the Roseobacter lineage as new genomes are published the methodology used in this study and by other authors (Luo & Moran, 2014; Newton *et al*., 2010), i.e. determining conserved proteins and analyze concatenated sequences, is a complex and time-consuming method. Then, as new data is being generated we need simpler and faster methods for defining the phylogeny of the lineage, and for affiliating the new isolates to phylogenomic groups, genera or species with confidence.

In this thesis we propose a new method for determining the phylogeny of members of the Roseobacter lineage. The strategy followed was to start with the 114 conserved core proteins used for the phylogenomic analysis shown in Figure 1.1 and simplify this dataset by generating a concatenate with the least number of proteins possible, while maintaining the tree topology

observed with the whole dataset. As a result we propose a phylogeny based on the concatenation of 14 single-copy proteins conserved in all Roseobacter genomes.

The details of the procedure followed to define this set of protein sequences is the following. Firstly, 114 phylogenomic trees were made using each one of the single-copy proteins of the core proteome of the 96 Roseobacter genomes following the procedure that was done for the core proteome phylogeny (see section 3 of Materials and Methods). After this, the proteins whose phylogenomic tree conserved at least one of the phylogenomic groups of the lineage were selected (32 proteins, see Table 1S.7 for details). Different combinations of concatenated sequences of these proteins were used for phylogenomic tree calculations until we reached a combination that reproduced the topology obtained with the concatenation of the 114 single-copy proteins of the core proteome of the 96 Roseobacter genomes. In Table 1.4 the 14 selected proteins are shown (for more details see supplementary Table 1S.8).

Table 1.4. List of the fourteen proteins selected for phylogenetic analysis of the Roseobacter lineage.

Order	Reference protein	Reference genome	Protein annotation[a]
1	MM48_01386	*Roseivivax atlanticus* 22II-s10s^T	Uridylyltransferase (O)
2	MM78_02585	*Sulfitobacter donghicola* DSW-25^T	RNA polymerase, sigma 70 factor (K)
3	MM83_00446	*Thalassobium* sp. R2A62	Glutamate synthase, small subunit (E, R)
4	MM25_01310	*Pseudooceanicola nanhaiensis* DSM 18065^T	Alpha-ketoglutarate reductase (H, E)
5	MM20_01240	*Oceanibulbus indolifex* HEL-45^T	FADH(2)-oxidizing methylenetetrahydrofolate-tRNA-(uracil(54)-C(5)-methyltransferase (J)
6	MM93_01221	*Maribius* sp. MOLA 401	Preprotein translocase subunit SecY (U)
7	MM34_03560	*Phaeobacter gallaeciensis* DSM 26640^T	Transcription termination factor Rho (K)
8	MM83_03116	*Thalassobium* sp. R2A62	GTPase associated with ribosome 50S subunit (R)
9	MM12_01203	*Leisingera aquimarina* DSM 24565^T	Porphobilinogen synthase (H)
10	MM95_01398	*Citreicella aestuarii* 328	Molecular chaperone DnaJ (O)
11	MM45_00504	*Roseibacterium elongatum* DSM 19469^T	Rod shape-determining membrane protein (D)
12	MM15_02765	*Loktanella cinnabarina* LL-001^T	RNA polymerase, alpha subunit (K)
13	MM06_03051	*Dinoroseobacter shibae* DFL 12^T	ATP phosphoribosyltransferase (E)
14	MM43_02070	*Rhodobacterales* bacterium HTCC2255	Orotate phosphoribosyltransferase (F)

a: functional categories according to COGs database are shown in parenthesis: J (translation, ribosomal structure and biogenesis); K (transcription); D (cell cycle control, cell division, chromosome partitioning); O (post-translational modification, protein turnover, chaperones); U (intracellular trafficking, secretion, and vesicular transport); C (energy production and conversion); E (amino acid transport and metabolism); F (nucleotide transport and metabolism); H (coenzyme transport and metabolism); R (general function prediction only).

The phylogeny obtained with the concatenation of these 14 proteins conserved in the 96 Roseobacter genomes is shown in Figure 1.2. The procedure for calculating the phylogeny was the same as it was done for the core proteome phylogeny (see in section 3 of Materials and Methods). In total, 4,462 amino acid aligned positions were used for the phylogeny (from 6,783 initial positions before Gblocks).

All phylogenomic groups were maintained, although there were two main discrepancies: i) *Maribius* sp. MOLA 401 and *Oceaniovalibus guishaninsula* JLT2003^T were affiliated to G4 phylo-

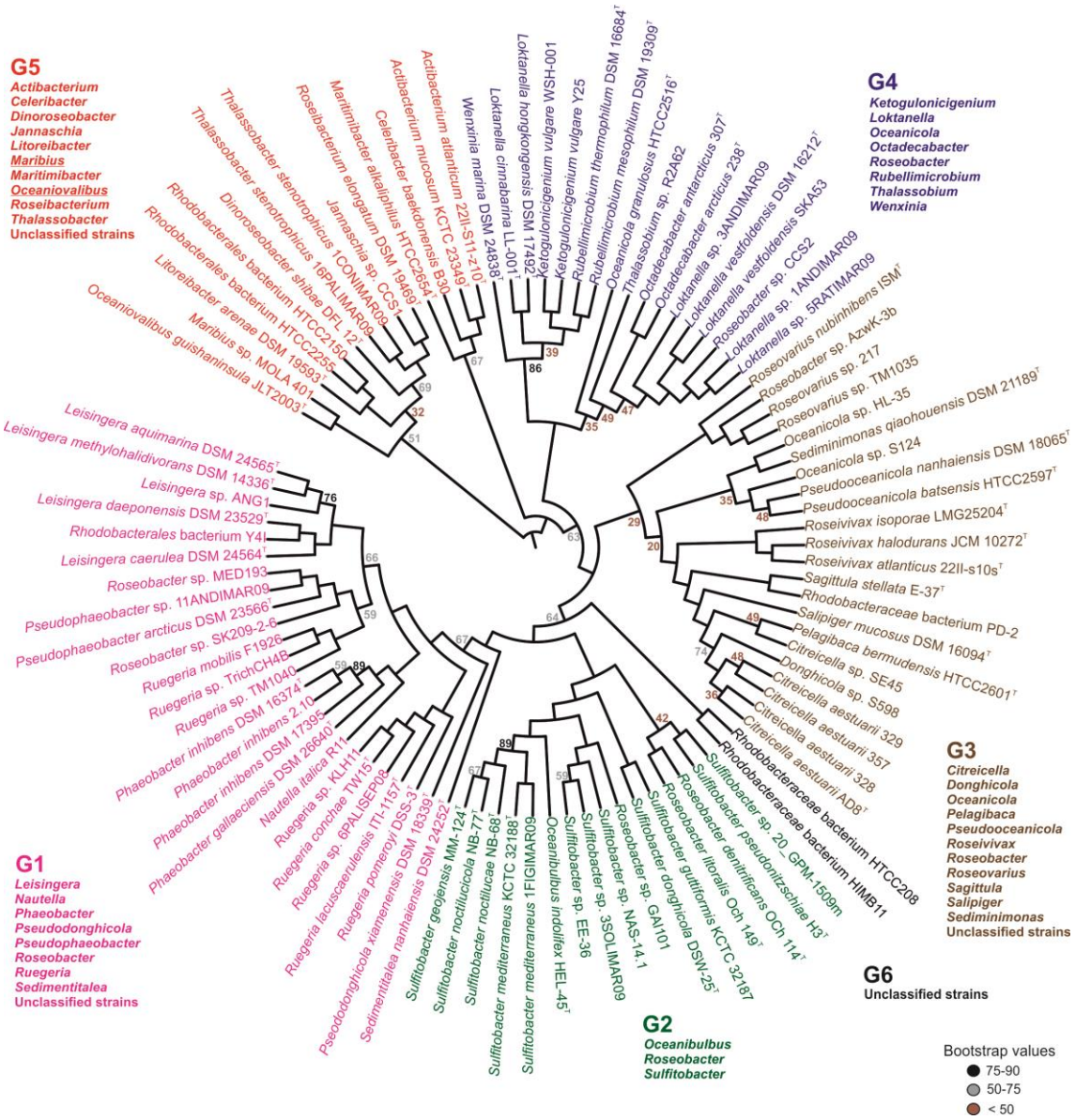

Figure 1.2. Phylogeny of the Roseobacter lineage (96 genomes) based on the concatenation of 14 conserved proteins. The phylogeny was done with PROTPARS package of Phylip program (Felsenstein, 1989) using parsimony and a bootstrap of 100. Only bootstrap values lower than 90 are shown. Phylogenomic group G5 was used as the outgroup. Underlined genera were affiliated to different phylogenomic group than in the phylogeny of the core proteome (see Figure 1.1.).

genomic group in the core proteome phylogeny (although they appeared separately from the other genomes of G4), while in the phylogeny of the concatenated 14 proteins they were affiliated to G5 phylogenomic group; ii) phylogenomic group G5 appeared separated in two branches in the phylogeny of the concatenated 14 proteins. Apart from these cases, the topologies of both phylogenies were in good agreement.

Next, we checked if with the addition of more genomes, the phylogeny obtained with concatenation of these 14 proteins was still useful. For this purpose, a new phylogeny of 165 Roseobacter sequenced genomes (all Roseobacter genomes published until April of 2016) was done (see Figure 1.3). In supplementary Table 1S.9 the genomic features of the 69 new added

roseobacters is shown. The annotation of these new genomes with Prokka program (Seeman, 2014) is shown in supplementary Table 1S.10. Representatives of fourteen new different genera of the Roseobacter lineage were included in the analysis in reference to the previous one (see Figure 1.1.). The 14 selected proteins were identified in the newly annotated 69 genomes by BLASTp using as a reference the proteins shown in Table 1.4 and considering a minimum of identity of 50 % and an *E* value below 1 x 10^{-5}. The 14 genic products identities and sequences of each of the 165 Roseobacter genomes can be checked in supplementary Table 1S.11. A total of 4,412 aligned positions were conserved and selected for calculating the phylogeny. In this case, the phylogeny was done with MEGA7 (Kumar *et al*., 2016) using parsimony and a bootstrap of 100.

In this new phylogeny, the G1, G2, G4 and G5 phylogenomic groups were clearly distinguished and separated with high values of bootstrap. G6 group was resolved with lower confidence (bootstrap value 6). With respect to the G3 phylogenomic group it was not resolved as a monophyletic group. In spite of this, the subgroups (at level of genus or species) within the G3 were well-resolved. In general, all new genomes were placed correctly in the phylogeny, according to their given genera and/or species names (see below). Besides, the inclusion of the new genomes did not alter much the affiliation of the 96 previous genomes. In particular, as it happened with the phylogeny of 14 single-copy proteins of 96 Roseobacter genomes, *Oceaniovalibus guishaninsula* JLT12003^{T} and *Maribius* sp. MOLA 401 were affiliated to phylogenomic group G5 while in the core proteome phylogeny they were affiliated to phylogenomic group G4. Another discrepancy was that with the introduction of the two genomes of the genus *Halocynthiibacter*, strains *Litoreibacter arenae* DSM 19593^{T}, HTCC2150 and HTCC2255 were affiliated to phylogenomic group G4 (with low bootstrap) while in the phylogeny of the core proteome of the 96 roseobacters they were affiliated to phylogenomic group G5.

Using the concatenation of these 14 conserved proteins, from a total of 51 genera, 46 were monophyletic while five were paraphyletic according to the names given to the strains [*Oceanicola* (G3, G4 and G5), *Roseobacter* (G1, G2 and G4), *Phaeobacter* (G1 and G3), *Tropicibacter* (G1 and G3) and *Aestuariivita* (G1 and G3)]. This was also observed in the previous phylogenies (see in Figures 1.1 and 1.2) and in previous studies (Luo & Moran, 2014). The cases of the two genomes of *Tropicibacter* and *Aestuariivita* were particularly stricking because the species included in the analysis are well-described microorganisms (Du *et al*., 2015; Lucena *et al*., 2012). While *Tropicibacter multivorans* and *Aestuariivita boseongensis* type strains were affiliated to phylogenomic group G1, *Tropicibacter naphthalenivorans* and *Aestuariivita atlantica* type strains were affiliated to phylogenomic group G3. Another case was genus *Phaeobacter* that traditionally was also affiliated to phylogenomic group G1 (see Figure 1.1 and 1.2) (Luo & Moran 2014; Newton *et al*., 2010). A new sequenced genome, *Phaeobacter* sp. CECT 7735 (97.5 % of identity in 1,422 bp aligned positions of 16S rDNA with *Phaeobacter gallaeciensis* DSM 26640^{T}), was affiliated to phylogenomic group G3. We calculated the ANIb values for the 165 genomes and compared the results with the phylogene-

Figure 1.3. Phylogeny of the Roseobacter lineage (165 genomes) based on the concatenation of 14 conserved proteins. The phylogeny was done with MEGA7 (Kumar *et al*., 2016) using parsimony and a bootstrap of 100. Only bootstrap values lower than 90 are shown. Phylogenomic group G5 was used as the outgroup. The new genera that were added to each phylogenomic in comparison to the phylogeny of 96 Roseobacter genomes are shown in bold.

tic affiliation obtained, to see if this parameter could help in clarifying the discrepancies. For *T. multivorans* the highest ANIb values (between 73-74 %) were obtained with genomes of group G1, for *T. naphthalenivorans* the highest values (73-75 %) were obtained with genomes of group G3. In the case of *Phaeobacter* sp. CECT 7735 values around 70.5 % were obtained with genomes of G1 and G3. Thus, the affiliation of these genomes could not be resolved. In any case, taking into account the few discrepancies in phylogeny observed, we consider that the

concatenate of 14 protein sequences proposed in this thesis is a useful and rapid method for affiliating new sequenced Roseobacter genomes.

4. Biochemically-relevant pathways predicted in genomes of the harbor isolates

We also wanted to genomically characterize the nine isolates of the coast of Mallorca in reference to different capabilities described as relevant in members of the Roseobacter lineage (see Table 1M). For more details of each feature, see the results in supplementary Table 1S.12. For the characterization we searched proteins from which we found the genes. Firstly we studied the potential for lithotrophy by oxidizing CO to CO_2. The two types of CO dehydrogenase described, CoxMLS types I and II of *Ruegeria pomeroyi* DSS-3^T (Moran *et al*., 2004), were searched in the genomes of the nine isolates. The genes of the type II CO dehydrogenase were found in all isolates, while type I genes were only present in the genomes of the two *Thalassobacter stenotrophicus* strains (1CONIMAR09 and 16PALIMAR09), and in the genome of *Ruegeria* sp. 6PALISEP08 (see Table 1.5). In all cases, the three identified proteins formed a genic cluster. Thus, all isolates harbored the potential for obtaining energy through oxidation of CO to CO_2. Another mechanism for lithoheterotrophic growth by members of Roseobacter lineage is oxidation of sulfur compounds. The complete genic cluster (*sox* genes) was detected in 8 of the isolates but not in 6PALISEP08 (see Table 1.5). These results agreed with previous studies in which the CO genes and *sox* genes were searched in Roseobacter genomes (Luo & Moran, 2014; Newton *et al*., 2010). In 2014, Luo and Moran analyzed 52 Roseobacter genomes; the type I *coxL* gene was detected in all of them while the *soxB* gene (thiosulfohydrolase) was detected in 67 % (35 genomes) of genomes.

In addition to the oxidation of organic matter, genome analysis indicates that some members of the Roseobacter lineage are also capable of phototrophy (Allgaier *et al*., 2003). We searched in the nine Roseobacter genomes the genes that encode the complete gene cluster for AAnP by sequence homology using as reference proteins of the well-characterized photosynthesis gene cluster of *Roseobacter denitrificans* OCh 114^T (Swingley *et al*., 2007). The photosynthesis gene cluster contains many genes involved in: bacteriochlorophyll biosynthesis (*bch*), carotenoid biosynthesis (*crt*), light harvesting complexes (*puc* and *puf*), reaction center proteins (*puhA*, *pufLM*) and their regulators, *ppsR*, *tspO* and *ppaA* (Choudhary & Kaplan, 2000). In addition, we searched the two forms of RuBisCO using as a reference the RbcL and RbpL of *Rhodobacter sphaeroides* 2.4.1^T (Gibson *et al*., 1991; Tabita *et al*., 2007). As shown in Table 1.5, the two *Thalassobacter stenotrophicus* strains harbored all genes for the synthesis of bacteriochlorophyll and carotenoids, for both light harvesting complexes (LH1 and LH2) and two reaction centers. The *puc* genes (for LH2) were located separately to the previous photosynthesis sub-clusters. That was also a common feature found in other AAnP genomes (Choudhary & Kaplan, 2000). In the three *Loktanella* isolates (1ANDIMAR09, 3ANDIMAR09 and 5RATIMAR09) we did not detect genes that codified for the light harvesting complex 2 (see in Table 1.5). It was reported that in other AAnP microorganisms such as the case of *Roseateles depolymerans* or other Roseobacter-like isolates, the LH2 complex was absent (Koblížek *et al*.,

2006; Suyama *et al*., 1999, 2002). In the five isolates we identified two conserved sub-clusters, *crtCDEF–bchCXYZ–pufQBALM* (about 10 kb) and *bchFNBHLM–lhaA–puhAB* (about 12–15 kb). These two sub-clusters have been also detected in a variety of phototropic *Proteobacteria* (Liotenberg *et al*., 2008; Zheng *et al*., 2011). None of two forms of RuBisCO were detected in the isolated genomes.

Table 1.5. Detection of selected genes or pathways with environmental relevance in the nine Roseobacter harbor isolates.

Metabolism	**Proteins**	**11ANDIMAR09**	**6PALISEP08**	**1FIGIMAR09**	**3SOLIMAR09**	**1ANDIMAR09**	**3ANDIMAR09**	**5RATIMAR09**	**16PALIMAR09**	**1CONIMAR09**
CO utilization	Cox proteins, type I	–[a]	+	–	–	–	–	–	+	+
	Cox proteins, type II	+	+	+	+	+	+	+	+	+
Sulfur oxidation	SoxABCDRSVWXYZ proteins	+	–	+	+	+	+	+	+	+
AAnP	Bacteriochlorophyll (Bchl) *a* biosynthesis	–	–	–	–	+	+	+	+	+
	Carotenoid (Crt) biosynthesis	–	–	–	–	+	+	+	+	+
	Light harvesting I PufAB	–	–	–	–	+	+	+	+	+
	Light harvesting II PucAB	–	–	–	–	–	–	–	+	+
	Reaction center PuhA	–	–	–	–	+	+	+	+	+
	Reaction center PufML	–	–	–	–	+	+	+	+	+
Carbon fixation	RuBisCO	–	–	–	–	–	–	–	–	–
DMSP degradation	DMSP demethylase (DmdA)	+	+	+	+	+	–	+	+	+
	DMSP lyase (DddW)	+	–	–	–	–	+	–	–	–
	DMSP lyase (DddQ)	+	–	–	–	–	–	–	–	–
	DMSP lyase (DddP)	+	+	+	–	–	–	–	+	+
Motility	Flagellum	+	–	+	+	–	+	–	+	+
Adhesion	Flp pili	+	+	+	+	+	+	+	+	+

a: +: detected. –: not detected.

We were also interested in analyzing the potential of the harbor isolates for utilizing dimethylsulphoniopropionate (DMSP). So, the well-characterized DmdA and DddP, DddQ and DddW of *Ruegeria pomeroyi* DSS-3^T (Reisch *et al*., 2013) were used as the key enzymes for the search of DMSP demethylation and cleavage pathways, respectively. The gene for the demethylase DmdA was identified in all isolates with the exception of *Loktanella* sp. 3ANDIMAR09. In reference to the cleavage pathway, the *dddP* gene was the most detected lyase being found in 5 isolates, followed by the *dddW* gene detected in 2 and the *dddQ* detected in 1 (see Table 1.5). In 5 isolates genes that codified for DmdA and DddP were detected. In *Pseudophaeobacter* sp. 11ANDIMAR09 genes for DmdA and all DMSP lyases were detected. These results were in agreement with previous surveys that showed that demethyase and lyase genes were common in the lineage (Newton *et al*., 2010).

In addition, the potential for motility and adhesion was also analyzed. We searched through the annotated genomes of the nine isolates the flagellar structural genic products using the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Moriya *et al*., 2007) (see in section 4 of Materials and Methods). According to our results, 6 of the isolates harbored the proteins for the flagellar assembly, while 3 of them not (see Table 1.5). A similar proportion was detected

by Newton and co-workers (2010) that detected the flagellar structural genes in 68.8 % (22 of 32 genomes analyzed). Genes for adhesion pili Flp were detected in all the harbor isolates (see in Table 1.5). The gene order of the genic cluster *cpaBC-ompA-cpaEF-tadBC* was conserved in all isolates. As expected, CpaD was not detected in the Roseobacter isolates and instead we found a colicin A import system protein (OmpA) in the expected position (Slightom & Buchan, 2009). The prevalence of this genic cluster in Roseobacter lineage was demonstrated by Slightom & Buchan (2009) who detected the genes for Flp pili synthesis in all 28 studied Roseobacter genomes. The results suggested that all isolates would be able to synthetize Flp pili.

5. Genomic characterization of the monoaromatic compound and alkane degradation capabilities of harbor isolates

The genomic analysis made up to date revealed that members of Roseobacter lineage could harbor up to 6 different metabolic pathways for the degradation of monoaromatic compounds: benzoyl-CoA, phenylacetic acid, homoprotocatechuate, homogentisate, gentisate and β-ketoadipate pathway (Moran *et al*., 2007; Newton *et al*., 2010). Twenty-seven out of 32 Roseobacter genomes analyzed by Newton and co-workers in 2010 harbored at least one of these six degradation pathways. Four roseobacters harbored genes for all six degradation pathways (*Ruegeria pomeroyi* DSS-3^T, *Roseobacter* sp. GAI101, *Sagittula stellata* E-37^T and *Jannaschia* sp. CCS1) (Newton *et al*., 2010). Therefore, we searched for genes for all these degradation pathways in the isolates obtained from Mallorca harbors. In this study we have searched all the genes involved in the different degradation pathways, and not only the ring-cleaving oxygenases. In addition, alkane hydroxylases, enzymes for the degradation of alkanes, were also searched in genomes of the nine harbor isolates. Alkane hydroxylases were previously detected in Roseobacter genomes (88 % of the 23 analyzed genomes) (Buchan & González, 2010).

The results obtained in this study are summarized in Table 1.6. For more details, see the results in supplementary Table 1S.13. The most detected pathways in genomes of the isolates were the homogentisate and the protocatechuate branch of the β-ketoadipate pathways (detected in 6 isolates). The gentisate pathway was not detected in any isolate and the benzoate pathway (*box* genes) was only detected in 1 isolate. These results were in agreement with the survey published by Newton and co-workers (2010) in which the homogentisate and the β-ketoadipate pathways (protocatechuate branch) were detected in more than the 68 % of the analyzed Roseobacter genomes and the gentisate and benzoate pathways were only detected in the 15 % of the analyzed Roseobacter genomes. We also searched if the isolates harbored alkane hydroxylases using as reference the two alkanes hydroxylases of *Alcanivorax borkumensis* SK2^T (Misawa *et al*., 2004). In all isolates, with the exception of the two *Thalassobacter stenotrophicus* strains (1CONIMAR09 and 16PALIMAR9), genes for alkane hydroxylases were detected with the criteria used in this study (*E* value below 1 x 10^{-5}) (see Table 1.6).

Table 1.6. Detection of monoaromatic compound and alkane degradation pathways in genomes of the Roseobacter isolates from Mallorca harbors.

Aromatic degradation pathway	11ANDIMAR09	6PALISEP08	1FIGIMAR09	3SOLIMAR09	1ANDIMAR09	3ANDIMAR09	5RATIMAR09	16PALIMAR09	1CONIMAR09
Benzoate (via benzoyl-CoA)	–[a]	–	–	+	–	–	–	–	–
Benzoate (via catechol)	–	–	–	–	–	–	–	–	–
Phenylacetate	+	+	+	+	–	–	–	–	–
Homoprotocatechuate	+	+	+	–	–	–	–	–	–
Homogentisate	+	+	+	+	+	–	+	–	–
Gentisate	–	–	–	–	–	–	–	–	–
β-ketoadipate (via catechol)	–	–	–	–	–	–	–	–	–
β-ketoadipate (via protocatechuate)	+	–	+	+	+	+	+	–	–
Catechol (via *meta*)	–	–	–	–	–	–	–	–	–
Protocatechuate (via *meta*)	–	–	–	–	–	–	–	–	–
Alkane hydroxylase	+	+	+	+	+	+	+	–	–

a: +: detected. -: not detected.

The isolate in which more pathways were detected was *Sulfitobacter mediterraneus* 1FIGIMAR09 followed by *Pseudophaeobacter* sp. 11ANDIMAR09 suggesting that these isolates could have a higher potential for the degradation of monoaromatic compounds than the other isolates. On the contrary, none of these degradation pathways were detected in the two strains of *Thalassobacter stenotrophicus* 1CONIMAR09 and 16PALIMAR09.

Apart from detecting the genes involved in the different aromatic compound degradation pathways we analyzed gene location and order, and compared gene structures in the harbor isolates with the model bacteria used for the analysis. The results for the different pathways are explained in the following paragraphs.

Genes for the degradation of homogentisate were detected in 6 of the isolates (see Table 1.6). However, we could not detect 4-hydroxyphenylpyruvate dioxygenase gene (*hppD*) in *Pseudophaeobacter* sp. 11ANDIMAR09 and maleylacetoacetate isomerase gene (*hmgC*) in *Sulfitobacter* sp. 3SOLIMAR09. In *Loktanella* sp. 5RATIMAR09 two different homogentisate 1, 2-dioxygenase (*hmgA*) and fumarylacetoacetate hydrolase (*hmgB*) genes were detected. In Figure 1.4, gene order for the homogentisate degradation pathway of *Ruegeria pomeroyi* DSS-3^T and the 6 isolates is shown. The *hppD* gene that codifies for 4-hydroxyphenylpyruvate dioxygenase (for the conversion of 4-hydroxyphenylpyruvate to homogentisate) was not located near *hmg* genes for degradation to homogentisate. In reference to *hmg* gene order none of the isolates had the same structure but there were some common features. For example, genes *hmgA* and *hmgB* were found together in 4 isolates. Also generally, gene *hmgC* was adjacent to *hmgA-hmgB* or in close proximity. In *Sulfitobacter* sp. 3SOLIMAR09 genes were not clustered. In three isolates, 11ANDIMAR09, 6PALISEP08 and 1FIGIMAR09, as well as in *Ruegeria pomeroyi* DSS-3^T we found genes not related to homogentisate degradation in the region between *hmgA-B* and *hmgC*. Shared proteins not related to the homogentisate degradation pathway were detected in *Ruegeria pomeroyi* DSS-3^T, *Pseudophaeobacter* sp.

11ANDIMAR09 and *Sulfitobacter mediterraneus* 1FIGIMAR09 (grey genes in Figure 1.4). Taking as a reference *Ruegeria pomeroyi* DSS-3^T the annotation of these common genes situated between *hmgB* and *hmgC* were i) a metallo-beta-lactamase family protein, ii) a monooxygenase family protein, iii) a hypothetical protein, and iv) a glyoxalase family protein.

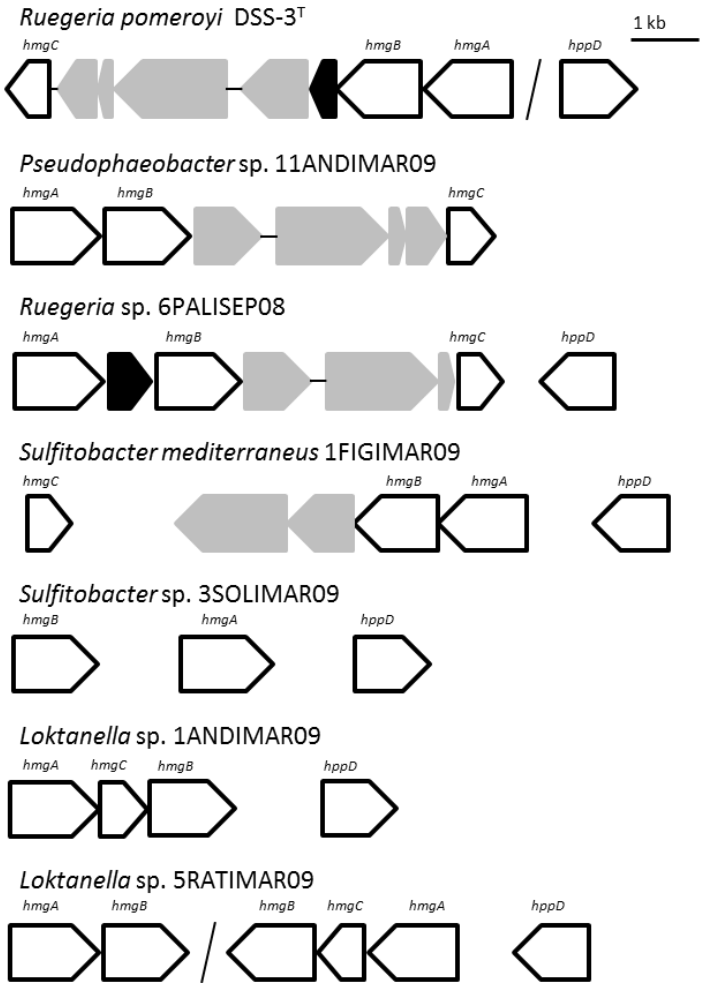

Figure 1.4. Structure of homogentisate degradation genes. Genes shown in black and grey were not related to the homogentisate degradation pathway (grey color indicates genes common to several isolates). Genes that are drawn separately to the other (without bars) indicate that were located in a different contig. The bars separate genes within a chromosome or contig that were not clustered together.

Two different branches of the β-ketoadipate pathway (named the protocatechuate and catechol branches, respectively) convert either protocatechuate or catechol to β-ketoadipate (Harwood & Parales, 1996). We searched the enzymes responsible for the degradation of protocatechuate in the nine isolates. Six isolates harbored genes of the protocatechuate degradation pathway plus the 4-hydroxybenzoate hydroxylase (PobA), a precursor of protocatechuate (see Table 1.6), indicating that they have the potential to degrade 4-hydroxybenzoate and protocatechuate. The regulator PcaQ was detected in *Loktanella* spp. 1ANDIMAR09 and 5RATIMAR09 while genes for the regulators PcaR and PobR were not detected in any genome under the criteria that we used. These two genes, were rarely detected in other Roseobacter genomes that harbored genes of protocatechuate pathway

(Alejandro-Marín *et al*., 2014). In reference to gene organization (see Figure 1.5), *pcaGH* and *pcaIJ* genes were always clustered together as it was reported in other bacteria, including roseobacters (Alejandro-Marín *et al*., 2014; Buchan *et al*., 2004; Harwood & Parales, 1996). In

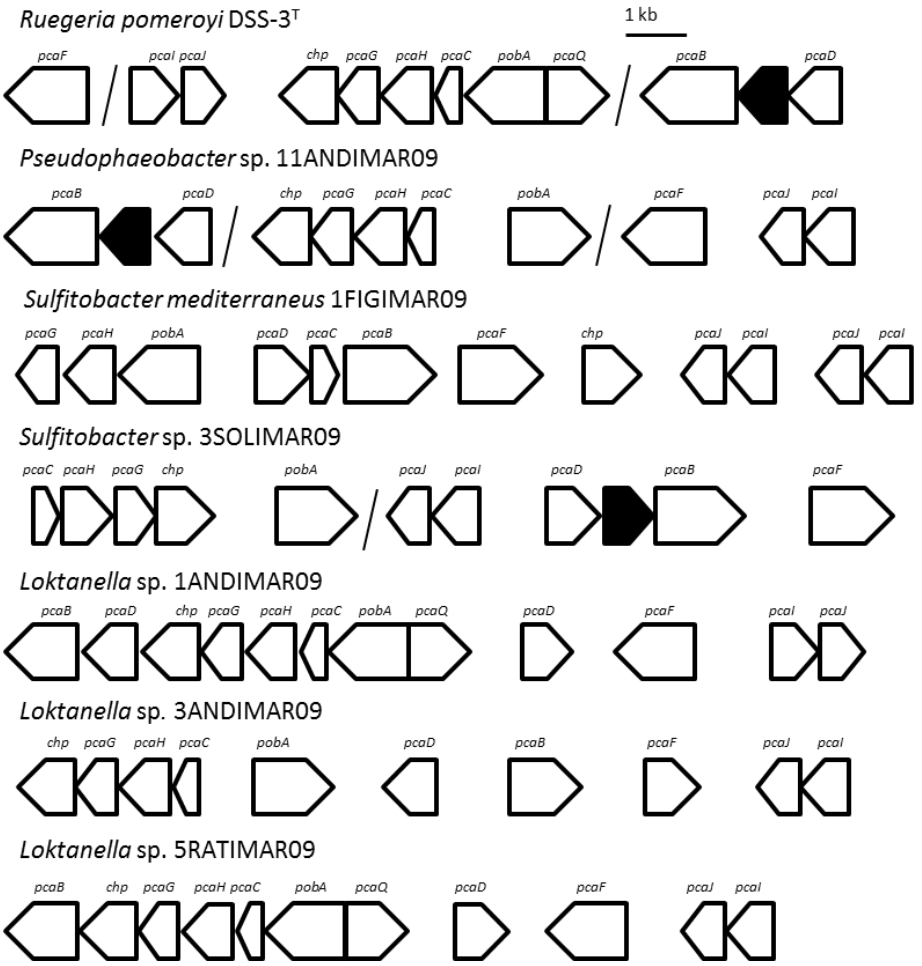

Figure 1.5. Structure of protocatechuate degradation genes. Genes shown in black are predicted as coding for an esterase that has not been related to the protocatechuate degradation pathway. Genes that are drawn separately to the other (without bars) indicate that were located in a different contig. The bars separate the genes within a chromosome or contig that were not clustered together.

the case of *pobA* gene, in three isolates (*Sulfitobacter mediterraneus* 1FIGIMAR09 and *Loktanella* spp. 1ANDIMAR09 and 5RATIMAR09) was next to *pcaGH* genes. Regarding to other genes, they were frequently scattered in the genome in several gene clusters, and gene order was not maintained. In reference to genes that were not related to the degradation of protocatechuate, a gene situated just dowstream *pcaB* that codified for a putative esterase was identified in *R. pomeroyi* DSS-3^T, *Pseudophaeobacter* sp. 11ANDIMAR09, *Sulfitobacter* sp. 3SOLIMAR09, and *Loktanella* sp. 1ANDIMAR09. The protocatechuate 4, 5-dioxygenase (*meta* cleavage), catechol 2, 3-dioxygenase (*meta* cleavage) and catechol 1, 2 dioxygenase (β-ketoadipate, catechol *ortho* cleavage) were not detected in any isolate.

With respect to the phenylacetate degradation pathway, all genes were detected in 4 isolates: *Pseudophaeobacter* sp. 11ANDIMAR09, *Ruegeria* sp. 6PALISEP08, and *Sulfitobacter* isolates

1FIGIMAR09 and 3SOLIMAR09 (see Table 1.6). In the 4 isolates, two different phenylacetyl-CoA ligases (PaaK) were detected with sequence identities below 30 % among them. Both Paak ligases had low values of identity with the reference sequence and were not located near the genic cluster of phenylacetate degradation genes. In the other five isolates genes for phenylacetyl-CoA ligases were also identified but we could not find the multicomponent oxygenase (*paaABCDE*) with the criteria used in this study. In Figure 1.6, the structure of phenylacetate genes of *Ruegeria pomeroyi* DSS-3^T and the 4 harbor isolates is shown. The order of genes that codified for the multicomponent oxygenase (*paaABCDE*) was maintained in all isolates. In all cases, gene *paaJ* encoding for a 3-oxoadipyl-CoA/ 3-oxo-5, 6-dehydrosuberyl-CoA thiolase, an enzyme that belongs to the phenylacetate lower degradation pathway was

Figure 1.6. Structure of phenylacetate degradation genes. Genes that are drawn separately to the other (without bars) indicate that were located in a different contig. The bars separate the genes within a chromosome or contig that were not clustered together.

located upstream of the multicomponent oxygenase *paaABCDE* genes. In *Pseudophaeobacter* sp. 11ANDIMAR09 and *Ruegeria* sp. 6PALISEP08, as in *Ruegeria pomeroyi* DSS-3^T, we found a conserved orf (*orf1*) (annotated as phenylacetic acid degradation protein, putative function) dowstream the *paaABCDE* genes. The role of this putative protein in phenylacetate degradation has not been described yet. The *paaZX* genes (that codified for a oxepin-CoA hydrolase and DNA-binding transcriptional repressor, respectively) appeared clustered in the genomes of all isolates with the exception of *Ruegeria* sp. 6PALISEP08.

The genes of the homoprotecatechuate pathway were identified in three isolates: *Pseudophaeobacter* sp. 11ANDIMAR09, *Ruegeria* sp. 6PALISEP08 and *Sulfitobacter mediterraneus* 1FIGIMAR09 (see Table 1.6). In Figure 1.7, the gene order of the model

organism *Jannaschia* sp. CCS1 and the isolates that harbored the homoprotocatechuate pathway is shown. Gene order for the clusters *hpaFED* and *hpaGH* were conserved among the isolates. Synteny was higher between *Jannashia* sp. CCS1, *Ruegeria* sp. 6PALISEP08 and *Sulfitobacter mediterraneus* 1FIGIMAR09. A gene annotated as a pyridoxal-dependent decarboxylase, and so far not related to the degradation of homoprotocatechuate, was found between *hpaD* and *hpaG* genes in *Jannashia* sp. CCS1, *Sulfitobacter mediterraneus* 1FIGIMAR09 and *Ruegeria* sp. 6PALISEP08.

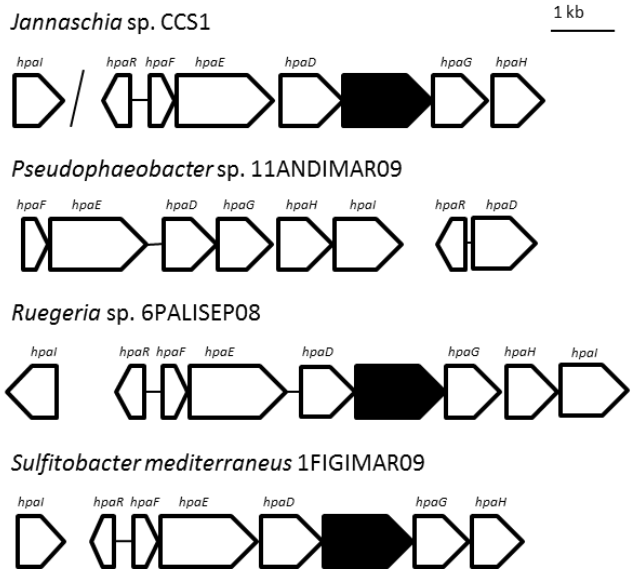

Figure 1.7. Structure of homoprotocatechuate degradation genes. Genes shown in black are predicted as coding for a pyridoxal-dependent decarboxylase. Genes that are drawn separately to the other (without bars) indicate that were located in a different contig. The bars separate the genes within a chromosome or contig that were not clustered together.

Benzoate degradation *box* genes were only detected in *Sulfitobacter mediterraneus* 1FIGIMAR09 (see Table 1.6), and gene structure had a good synteny with benzoate degradation genes of *Ruegeria pomeroyi* DSS-3^T (see Figure 1.8). Between *bclA* and *boxC* genes we detected the same two genes in both genomes. The first one was a putative hydrolase and

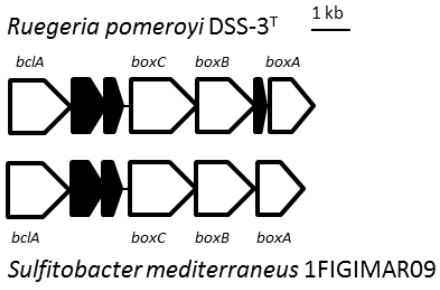

Figure 1.8. Structure of benzoate degradation genes. Genes shown in black are not related to benzoate degradation.

the second was a hypothetical protein. In addition, between *boxB* and *boxA* of *R. pomeroyi* DSS-3^T there was a gene for another hypothetical protein that was not observed in *S.*

mediterraneus 1FIGIMAR09. Genes for degradation of benzoate via catechol pathway (*ben* genes) were not detected in any isolate (see Table 1.6).

6. Growth of harbor isolates on monoaromatic compounds and diesel oil

The nine harbor isolates were grown on monoaromatic compounds and diesel oil as a sole source of carbon and energy (see in section 5.2 of Materials and Methods), and the results were compared with those of pathway prediction based in genome sequence analysis. The results of the growing experiments are shown in Table 1.7.

Table 1.7. Growth of Roseobacter harbor isolates in a variety of monoaromatic compounds and diesel oil.

Compound	11ANDIMAR09	6PALISEP08	1FIGIMAR09	3SOLIMAR09	1ANDIMAR09	3ANDIMAR09	5RATIMAR09	1CONIMAR09	16PALIMAR09
Benzoate (3 mM)	–[a]	–	–	–	–	–	–	–	–
Phenylacetate (3 mM)	+	+	–	+	–	–	–	–	–
Homoprotocatechuate (3 mM)	–	–	–	–	–	–	–	–	–
Homogentisate (3 mM)	–	+	–	–	–	–	–	–	–
Gentisate (3 mM)	–	–	–	–	–	–	–	–	–
Protocatechuate (3 mM)	+	–	–	+	–	–	–	–	–
Diesel oil (0.1 % v/v)	–	+	–	+	–	+	–	–	–

a: +: growth. –: not growth; grey shading indicate the cases in which there were discrepancies between the genomic prediction and the physiologic results.

Only in a few cases, the isolates were able to grow using the tested compounds as a sole source of carbon and energy. *Sulfitobacter* sp. 3SOLIMAR09 and *Pseudophaeobacter* sp. 11ANDIMAR09 grew using phenylacetate and protocatechuate, and *Ruegeria* sp. 6PALISEP08 using phenylacetate and homogentisate. In all cases in which the isolate showed a positive growth, genomic prediction indicated that the isolate harbored the genes for the degradation of these compounds. However, there were many cases in which the isolate putatively harbored the genes for the degradation of a particular aromatic compound but no growth was observed in this compound (see Table 1.7). Alejandro-Marín and co-workers (2014) also detected discrepancies between prediction of genes for protocatechuate branch of β-ketoadipate pathway and the growth of Roseobacter isolates on this compound. In our case, two possible explanations for these discrepancies are proposed: i) the isolates had the genes for degrading the tested compounds but they were not expressed at the conditions tested, and/or ii) the culturing conditions used were not appropriate for their growth in these compounds. We also found some discrepancies with the growth results obtained previously for these isolates (Piña-Villalonga, 2012), but in that case the medium used had a different composition (in the concentration of nitrogen, phosphorus and iron). Since the growth of roseobacters in aromatic compounds is usally weak (see for example Alejandro-Marín *et al.*, 2014) more thorough studies should be performed with the harbor isolates, including chemical analysis of monoaromatic compound degradation, and testing possible inducers of catabolic

pathways. For example, growth of *Sagittula stellata* E-37^T on protocatechuate was observed only if the inoculum was previously grown in the presence of 4-hydroxybenzoate (Alejandro-Marín *et al*., 2014).

In case of growth with diesel oil, despite in 6 of the isolates we detected genes for alkane hydroxylases, only three of the isolates (*Loktanella* sp. 3ANDIMAR09, *Sulfitobacter* sp. 3SOLIMAR09 and *Ruegeria* sp. 6PALISEP08) were able to grow using diesel oil as a sole source of carbon and energy. This growth could be due to the action of the alkane hydroxylases or also because of the action of other enzymes that could metabolize aromatic compounds that are present in diesel oil in lower amounts compared with the alkanes. So, because we wanted to know which genes were expressed and were involved in the tolerance or degradation of diesel oil by these nine isolates, and because no previous studies related to diesel oil tolerance or degradation within the Roseobacter lineage were done, we proposed to do a proteomic analysis (chapter 2).

CHAPTER 2. PROTEOGENOMIC ANALYSIS OF THE ROSEOBACTER HARBOR ISOLATES EXPOSED TO DIESEL OIL

The isolates analyzed in this study were obtained from harbor waters in Mallorca Island and, as shown in the previous chapter, had putative genes for the degradation of alkanes and/or could grow on diesel oil. Because the presence of diesel oil in harbor water is not uncommon we studied the response of the nine isolates when they were in contact with this compound by analyzing their proteomes. Our aim was to decipher which genes were expressed after exposure to diesel oil (either for tolerance or degradation) and which were the differences in the response among the isolates.

1. Proteome identification and functional characterization

The isolates were exposed to seawater (control conditions) or seawater with diesel oil (0.1 % v/v). Digested proteins of the nine isolates were analyzed by nanoLC-ESI-MS/MS (see section 6 of Materials and Methods for details). Two isolates were discarded from the study due to technical reasons: we could not get valid MS spectra for *Loktanella* sp. 5RATIMAR09 and we detected a very low percentage of proteins (5.5 % in reference to the annotated proteome) in the case of *Pseudophaeobacter* sp. 11ANDIMAR09. For the other seven isolates the percentage of detection of the total proteome (considering the two conditions and three replicates for each condition per isolate) was 33.8 % ± 4.6 on average (see Figure 2.1). The list of detected proteins for each isolate analyzed by Progenesis is shown in supplementary Table 2S.1.

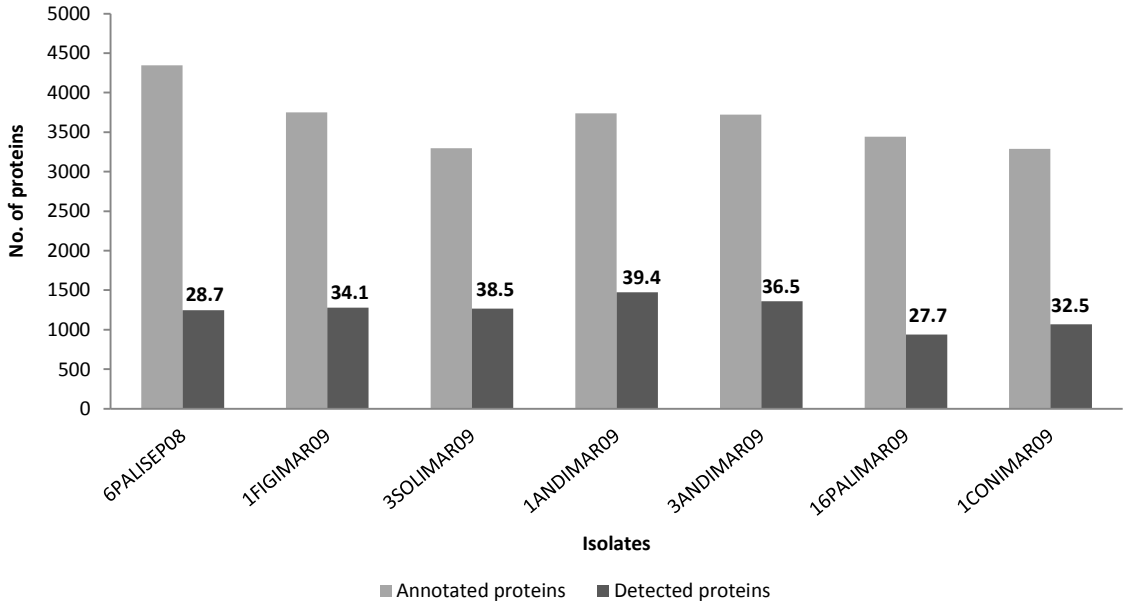

Figure 2.1. Comparison of the annotated and detected proteins. The numbers above the bars indicate the percentage (%) of proteome detection for each isolate. In the percentage of detection we considered the sum of all proteins of the three replicates for the two conditions for each isolate.

For the functional characterization of the detected proteins, first of all we determined the distribution of the detected proteome of the isolates in protein categories. For this, clustering of genic products with 50 % of sequence identity in at least the 50 % of amino acid sequence was done, based on genome sequence data, to determine the different Clusters of Orthologous Groups (COGs) shared by these 7 isolates. By this, we defined the core proteome (core-7), accessory

proteome and singletons of the complete annotated proteome of the 7 isolates. Then, the core proteome, accessory proteome and singletons of the annotated genomes were used to classify the COGs of the detected proteome (we used all the proteins detected in the two conditions and the three replicates for each condition per isolate).

Table 2.1. Assignation of detected proteome in proteomic and functional categories.

	No. of COGs		
Category	**Annotated**	**Detected[a]**	**Functional assignation[b]**
Core proteome	1,206	980 (81.3)	969 (98.9)
Accessory proteome	3,579	1,199 (33.5)	1,050 (87.6)
Singletons	6,454	1,003 (15.5)	737 (73.5)
Panproteome	11,238	3,182 (28.3)	2,756 (86.6)
Total proteins	25,584	8,628 (33.7)	–

a: in parenthesis, percentage of detection in relation to the annotated proteome.

b: in parenthesis, percentage of functional assignation in relation to the detected proteome.

Twenty-eight percent of the panproteome of the isolates was detected by proteomic analysis. This indicates that the remaining 72 % COGs of annotated panproteome seemed to be dedicated to specific functions and will be detected in specific growth conditions as reported previously (Christie-Oleza *et al*., 2012). As we expected (see for example Christie-Oleza *et al*., 2012), the COGs that belonged to the core proteome had the highest value of detection (81.3 %). That is because the proteins of the core proteome are essential and its expression is important for survival (see below and Figure 2S.1). Approximately 46 % of the detected COGs of the core proteome (980) were detected in all isolates (446 COGs).

With regard to functional assignation, in general the percentages were high in all proteomic categories (higher than 70 %), being the core proteome the better characterized fraction (see Table 2.1). The dominating functions of the core proteome were adscribed to categories (E) “amino acid transport and metabolism” and (J) “translation, ribosomal structure and biogenesis”, while in the accessory proteome and singletons we found mainly proteins belonging to (M) “cell wall and envelope biogenesis”, (U) “intracellular trafficking and secretion” and (Q) “secondary metabolites biosynthesis” COG functional categories (see in Figure 2S.1).

In addition, all the detected proteins of each isolate separately (two conditions, three replicates per condition) were also classified according to the COGs database functional categories. For each isolate, the percentages of detected proteins in reference to the total annotated proteins of each COG functional category were calculated (see Table 2.2). In general, the pattern of detection was very similar in all isolates with highest values of detection in the COG functional categories related to basal metabolism [(J) “translation, ribosomal structure and biogenesis” and (F) “nucleotide transport and metabolism”]. For additional information referred to the number of detected and total proteins that constituted each COG functional category check supplementary Table 2S.2.

Table 2.2. Percentages of detected proteins in reference to the annotated proteins per COG functional category of each isolate.

COG functional category	6PALISEP08	1FIGIMAR09	3SOLIMAR09	1ANDIMAR09	3ANDIMAR09	16PALIMAR09	1CONIMAR09
Information storage and processing							
Translation, ribosomal struct(ure and biogenesis (J)	72.1	75.3	81.7	78.4	78.9	64.0	72.8
Transcription (K)	26.1	24.0	28.8	29.7	34.6	19.8	28.6
Replication, recombination and repair (L)	33.3	35.2	31.3	36.0	32.5	22.1	28.7
Cellular processes and signaling							
Cell cycle control, cell division, chromosome partitioning (D)	41.4	48.1	42.3	51.6	57.1	41.4	44.4
Cell wall/membrane/envelope biogenesis (M)	43.8	49.1	44.9	48.9	50.8	39.3	44.7
Cell motility (N)	12.5	15.6	3.2	15.0	8.8	6.8	10.8
Post-translational modification, protein turnover, chaperones (O)	39.8	47.2	50.0	46.5	44.4	33.6	36.2
Signal transduction mechanisms (T)	34.6	26.4	33.3	35.7	26.7	18.0	21.8
Intracellular trafficking, secretion, and vesicular transport (U)	54.5	59.5	42.6	48.4	41.1	23.4	22.6
Defense mechanisms (V)	22.4	39.3	21.2	36.8	37.8	28.1	20.7
Metabolism							
Energy production and conversion (C)	40.7	59.3	55.0	56.8	60.8	48.8	49.3
Amino acid transport and metabolism (E)	35.3	43.0	54.0	47.4	56.9	41.9	43.0
Nucleotide transport and metabolism (F)	64.6	64.1	77.5	73.7	78.9	63.5	68.1
Carbohydrate transport and metabolism (G)	23.6	24.4	35.3	39.8	42.5	31.6	37.5
Coenzyme transport and metabolism (H)	41.2	58.0	56.4	55.6	60.7	47.3	52.7
Lipid transport and metabolism (I)	40.8	51.6	62.8	53.1	59.8	50.0	53.3
Inorganic ion transport and metabolism (P)	12.2	18.2	24.3	20.3	25.0	20.9	18.7
Secondary metabolites biosynthesis, transport and catabolism (Q)	24.1	33.2	31.6	34.8	39.2	30.4	32.5
Poorly characterized							
General function prediction only (R)	25.7	30.7	32.0	34.9	33.6	23.7	27.8
Function unknown (S)	19.2	19.6	21.8	29.5	26.3	14.4	22.3
Not in COGs	9.6	9.1	13.5	11.2	10.2	4.3	8.7

2. Differences in detected protein patterns after exposure to diesel oil

2.1. General functional response

In order to compare the patterns of detected proteins for each isolate and growing condition we used Principal Component Analysis (PCA). Different datasets were used as input for the analysis (see in section 6.3 of Materials and Methods). The PCA that showed the highest level of resolution was the one with COGs obtained from the clustering of detected proteins under the criteria of 90 % of identity in at least the 95 % of amino acid sequence (criterion 90_95). In order to compare the results of the different isolates we expressed the values of normalized abundance of each protein as a percentage of the total. As can be seen in supplementary Figure 2S.2, PCA grouped the isolates according to their phylogenomic origin and not according to the cultured conditions. For this reason, PCA analysis was done separately for each isolate using the normalized abundances given by Progenesis. In Figure 2.2, the result of this analysis for *Ruegeria* sp. 6PALISEP08 is shown as an example.

When each isolate was analyzed individually, the samples grouped according to the culture condition (see Figure 2.2 and Figures 2S.3 to 2S.8). Figure 2.2, shows the separation of the three replicates that were incubated in seawater (control) from those that were exposed to diesel oil which indicated that the detected proteome was different in the two tested conditions. Similar results were obtained for the other isolates (see supplementary Figures from 2S.3 to 2S.8).

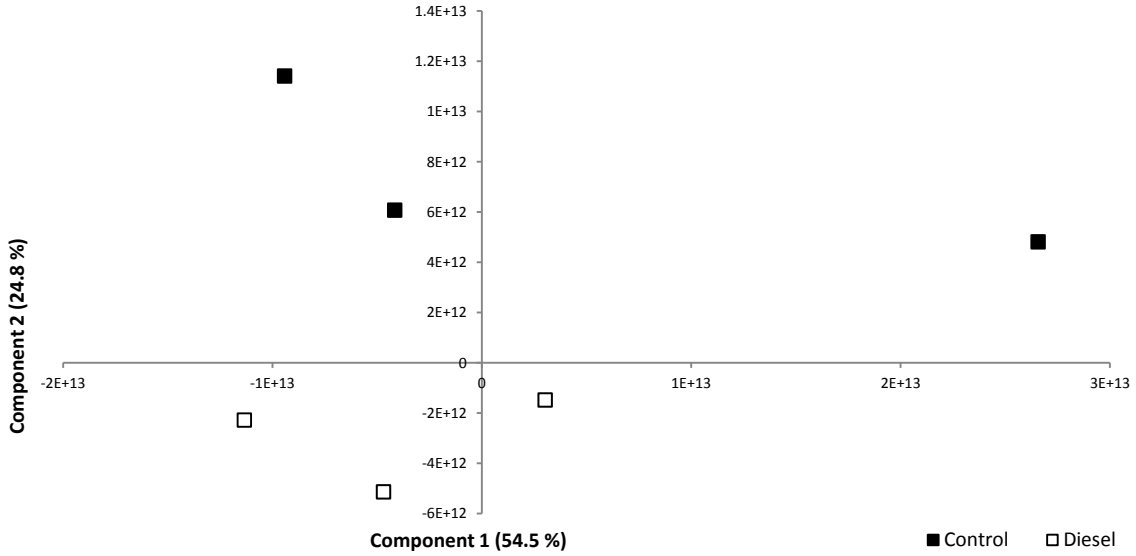

Figure 2.2. PCA of the normalized abundances (Progenesis) of the detected COGs of proteins clustered under the criterion of 90 % of identity in, at least, the 95 % of the amino acid sequence of *Ruegeria* sp. 6PALISEP08 (three replicates).

In order to determine which COG functional categories were detected differently in the two incubation conditions we calculated the percentages of normalized abundances for each isolate and condition (see Table 2.3). Although there were differences in the percentages of abundance of each COG functional category in the different isolates (Table 2.3 and Figures 2S.9 and 2S.10), there was a general pattern. In all isolates, three COG functional categories [(J) "translation, ribosomal structure and biogenesis", (C) "energy production and conversion", and (E) "amino acid transport and metabolism"] had percentages of abundance higher than 12 % while another three [(N) "cell motility", (V) "defense mechanisms" and (D) "cell cycle control, cell division and chromosome partitioning"] had low abundances (less than 1.5 %). On the other hand, there were several COG categories that were particularly interesting because there were significant differences (T-student test, p-value ≤ 0.05) in the percentages of abundance between the control and treated cultures (see bold values in Table 2.3). In three functional categories significant differences were detected in more than one isolate. They belonged to "cellular processes and signaling" COG global category and they were: (D) "cell cycle control, cell division and chromosome partitioning", (U) "intracellular trafficking, secretion, and vesicular transport" (includes proteins such as synthesis pili proteins or outer membrane proteins), and (V) "defense mechanisms" (includes efflux pumps and restriction enzymes). Both (U) and (V) COG functional categories were significantly more abundant in control condition than in the treated one in *Ruegeria* sp. 6PALISEP08 (see in Table 2.3). The isolates in which more significant differences were

Table 2.3. Percentages of abundance of the detected proteins classified in COG functional categories (averages ± sd of three replicates). Values in bold indicate that there was a significant difference between control and diesel oil condition (T-student test p-value ≤ 0.05).

	6PALISEP08		1FIGIMAR09		3SOLIMAR09		1ANDIMAR09		3ANDIMAR09		16PALIMAR09		1CONIMAR09	
	Control	**Diesel**	**Control**	**Diesel**	**Control**	**Diesel**	**Control**	**Diesel**	**Control**	**Diesel**	**Control**	**Diesel**	**Control**	**Diesel**
(J)	21.28±0.92	21.31±0.3	26.64±1.16	26.46±0.43	29.13±0.05	29.92±1.69	15.46±0.77	15.14±0.25	25.90±0.93	25.08±0.23	14.45±0.36	15.28±0.44	18.35±0.33	19.25±0.51
(K)	5.79±0.22	5.54±0.12	3.76±0.08	3.57±0.18	4.07±0.26	3.93±0.16	3.03±0.06	3.00±0.12	4.32±0.02	4.86±0.37	3.26±0.14	3.39±0.11	**3.95±0.02**	**4.26±0.04**
(L)	2.24±0.07	2.12±0.16	1.68±0.06	1.63±0.03	1.40±0.05	1.34±0.08	2.03±0.03	1.90±0.13	1.15±0.06	1.19±0.03	**1.42±0.04**	**1.57±0.07**	1.51±0.02	1.50±0.03
(D)	1±0.08	0.98±0.06	0.23±0.01	0.21±0.02	0.20±0.01	0.20±0.01	**0.24±0.02**	**0.27±0.01**	0.30±0.01	0.31±0.01	**0.26±0.01**	**0.23±0.01**	0.20±0.01	0.21±0.01
(M)	5.51±0.29	5.08±0.1	2.47±0.07	2.44±0.14	2.59±0.08	2.62±0.1	4.07±0.27	3.96±0.18	3.83±0.22	3.80±0.05	2.79±0.11	2.79±0.01	4.22±0.06	4.11±0.08
(N)	0.003±0.001	0.003±0.001	**0.10±0.004**	**0.09±0.003**	0.005±0.0003	0.005±0.0006	0.11±0.02	0.10±0.004	0.01±0.002	0.01±0.003	0.11±0.003	0.10±0.01	0.07±0.005	0.06±0.004
(O)	4.97±0.41	5.65±0.16	8.21±0.45	8.96±0.28	7.14±0.34	7.24±0.37	7.02±0.43	7.74±0.42	8.26±0.69	8.43±0.29	6.48±0.51	6.46±0.03	11.46±0.49	11.28±0.25
(T)	2.89±0.29	3.24±0.23	3.17±0.56	3.37±0.19	1.73±0.09	1.68±0.1	5.01±0.21	5.21±0.17	1.42±0.08	1.23±0.08	3.13±0.51	2.51±0.37	3.28±0.09	3.27±0.21
(U)	**2.27±0.09**	**1.88±0.02**	**0.96±0.06**	**0.80±0.03**	0.94±0.02	0.97±0.03	2.83±0.48	2.14±0.15	1.50±0.03	1.45±0.04	0.76±0.02	0.79±0.02	1.11±0.04	1.01±0.07
(V)	**0.10±0.007**	**0.09±0.004**	0.24±0.02	0.23±0.01	0.05±0.005	0.05±0.01	**0.11±0.002**	**0.10±0.004**	0.12±0.01	0.13±0.01	0.05±0.002	0.05±0.01	0.017±0.001	0.017±0.0003
(C)	19.77±0.34	19.72±0.37	14.93±0.45	15.21±0.18	19.97±0.18	19.55±043	16.03±0.33	16.47±0.27	22.34±1.2	21.86±0.14	22.62±0.67	23.32±0.29	18.56±0.13	18.91±0.37
(E)	13.30±0.27	13.5±0.22	16.58±0.62	16.13±0.5	12.74±0.32	12.38±0.66	21.60±0.72	21.25±0.17	13.40±0.86	13.26±0.25	**22.60±0.55**	**21±0.8**	16.56±0.1	16.59±0.33
(F)	4.61±0.14	4.69±0.14	5.22±0.17	4.96±0.06	3.79±0.05	3.65±0.23	2.65±0.27	2.84±0.09	3.68±0.39	3.94±0.05	5.39±0.23	5.46±0.22	3.93±0.22	4.10±0.21
(G)	6.05±0.28	6.41±0.1	3.73±0.09	3.57±0.11	4.60±0.17	4.30±0.35	4.83±0.18	4.70±0.01	5.45±0.18	5.54±0.13	**3.53±0.07**	**3.26±0.07**	4.47±0.08	4.40±0.06
(H)	3.28±0.15	3.27±0.26	3.21±0.08	3.21±0.08	2.46±0.05	2.42±0.12	1.83±0.12	1.94±0.04	2.60±0.25	2.81±0.05	2.20±0.11	2.34±0.06	**2.06±0.04**	**2.19±0.05**
(I)	3.03±0.22	3.18±0.11	3.38±0.04	3.17±0.2	4.21±0.13	4.28±0.15	3.35±0.23	3.40±0.1	2.54±0.05	2.52±0.05	4.29±0.09	4.39±0.11	3.38±0.11	3.51±0.08
(P)	1.99±0.17	1.93±0.14	1.33±0.04	1.23±0.06	1.68±0.02	1.63±0.18	5.50±0.16	5.65±0.12	1.36±0.12	1.32±0.09	0.95±0.09	1.04±0.004	1.01±0.03	0.95±0.01
(Q)	2.21±0.2	2.23±0.09	3.92±0.44	3.19±0.18	1.84±0.02	1.74±0.07	**1.92±0.05**	**1.73±0.05**	1.37±0.06	1.28±0.04	4.01±0.08	4.12±0.02	**4.71±0.02**	**4.19±0.06**
(R)	4.52±0.06	4.48±0.15	**3.94±0.05**	**3.68±0.08**	3.62±0.02	3.52±0.25	**4.28±0.04**	**4.57±0.06**	4.38±0.29	4.31±0.23	5.21±0.27	5.29±0.24	5.05±0.08	4.85±0.14
(S)	1.49±0.06	1.50±0.05	1.15±0.1	1.28±0.1	1.41±0.05	1.36±0.15	2.00±0.12	1.98±0.06	1.78±0.08	1.76±0.13	1.11±0.04	1.04±0.08	1.44±0.03	1.31±0.08
NC	5.31±0.4	4.81±0.09	**3.59±0.41**	**4.91±0.3**	3.64±0.27	4.40±0.64	3.93±0.26	3.85±0.15	3.87±0.31	3.89±0.07	3.14±0.19	2.97±0.24	**3.28±0.08**	**2.71±0.09**

a: functional categories according to COGs database: J (translation, ribosomal structure and biogenesis); K (transcription); L (replication, recombination and repair); D (cell cycle control, cell division, chromosome partitioning); M (cell wall, membrane, envelope biogenesis); N (cell motility); O (post-translational modification, protein turnover, chaperones); T (signal transduction mechanisms); U (intracellular trafficking, secretion, and vesicular transport); V (defense mechanisms); C (energy production and conversion); E (amino acid transport and metabolism); F (nucleotide transport and metabolism); G (carbohydrate transport and metabolism); H (coenzyme transport and metabolism); I (lipid transport and metabolism); P (inorganic ion transport and metabolism); Q (secondary metabolites biosynthesis, transport and catabolism); R (general function prediction only); S (function unknown) and NC (not in COGs).

detected (at least 4 COG functional categories) among the tested conditions were *Sulfitobacter mediterraneus* 1FIGIMAR09, *Loktanella* sp. 1ANDIMAR09, and the two *Thalassobacter stenotrophicus* strains, 16PALIMAR09 and 1CONIMAR09. On the contrary, in *Loktanella* sp. 3ANDIMAR09 no significant difference in any of the COG functional categories between the conditions was detected.

Because few significant differences in protein abundances were detected in the two conditions tested when their global functionality was analyzed, we studied the differences at higher level of resolution. For this, we selected the proteins that showed statistical differences in the ANOVA analysis done with Progenesis with the criteria of a fold change ≥ 2 and p-value ≤ 0.05 (Table 2S.1). We also included proteins that were important for the ordination obtained with PCA (Figures 2.2 and from 2S.3 to 2S.8): factor loadings higher than 0.1 and lower than −0.1 for each principal component. Based on this criterion, an average of 239 ± 65 proteins per isolate were analyzed in detail (check the list of proteins in Table 2S.3). The results are explained below in section 2.3.

Apart from the study of the statistically significant proteins, we focused on the analysis of proteins that could be involved in degradation and tolerance to diesel oil. For the study of the degradation response we focused on the analysis of monoaromatic compound degradation proteins and alkane hydroxylases (see in section 6.3 of Materials and Methods). For studying tolerance response to diesel oil, we based on a survey that Ramos and co-workers (2015) published recently in which they analyzed the mechanisms of solvent resistance by *Pseudomonas putida* to toluene. These mechanisms of tolerance included: changes in microbial cell membranes in response to solvents, removal of active oxygen species, expression of chaperones and efflux pumps and changes in energy production (Ramos *et al*., 2015). Furthermore, as another mechanism to face the presence of diesel oil, we studied proteins that could have a role in the attachment to hydrocarbon-water interface, such as pili formation proteins. It was suggested that *Alcanivorax borkumensis* SK2 that is an specialist degrader of alkanes, uses pili to have an easy access to the hydrocarbon (Schneiker *et al*., 2006).

2.2. Proteins involved in degradation of diesel oil

Diesel oil is a very complex mixture of thousands of individual compounds, most with carbon numbers between 10 and 22. The majority of them are aliphatic or aromatic compounds (Bacha *et al*., 2007). We searched for proteins that could be involved in the degradation of both types of hydrocarbons. The genomic prediction of monoaromatic compound degradation pathways as well as the identification of putative alkanes monooxygenases in the seven studied isolates was shown in section 5 of chapter 1 (Table 1.6). This information together with the annotation of the detected proteins served us for identifying proteins that could be involved in degradation of diesel oil by these isolates (see Table 2.4 and supplementary Table 2S.4 for detail).

The key enzyme for the degradation of alkanes, alkane 1-monooxygenase was detected in three isolates (see Table 2.4). Two of them (3ANDIMAR09 and 3SOLIMAR09 isolates) were able to grow

using diesel oil as a sole source of carbon and energy in marine mineral medium (see Table 1.7). In 3ANDIMAR09, this enzyme was significantly more abundant in diesel oil condition suggesting an active role in the degradation of alkanes. In 1FIGIMAR09 and 3SOLIMAR09 this enzyme was detected with similar abundances in both conditions suggesting that it was equally induced in both cases. Finally, isolate 6PALISEP08 was able to grow using diesel oil as a sole source of carbon and energy, but the annotated alkane 1-monooxygenase was not detected in the incubation conditions used in this study. This result might indicate that this enzyme would be only expressed under certain culture conditions. Finally, aldehyde dehydrogenases were detected in all isolates and alcohol dehydrogenases in all of them except the two *Loktanella* species (1ANDIMAR09 and 3ANDIMAR09). In all cases there were not significant differences between the tested conditions. These two types of enzymes might be involved in cellular processes other than alkane degradation and therefore this result is not surprising.

In relation to the degradation of aromatic compounds, key enzymes of phenylacetate, homogentisate, β-ketoadipate (via protocatechuate) and benzoate pathways were detected. Phenylacetate-CoA ligase (PaaK) and different proteins (PaaA, PaaC and PaaD) of the multicomponent phenylacetate oxygenase (PaaABCDE) were detected in three of the isolates (6PALISEP08, 1FIGIMAR09 and 3SOLIMAR09) (see Table 2.4). These results agreed with the genomic prediction (see Table 1.6). The fold changes of 1,2-phenylacetyl-CoA monooxygenase (PaaA) normalized abundances in the two *Sulfitobacter* strains were striking. In the case of *Sulfitobacter mediterraneus* 1FIGIMAR09 (fold change 4.8 x and p-value= 0.07) one of the replicates of the control condition was different from the other two replicates showing similar abundances to the diesel oil condition (see in supplementary Table 2S.1.2 for details). This variation in one of the control replicates lowered the statistical significance of the results below the considered cut-off value. In *Sulfitobacter* sp. 3SOLIMAR09 (fold change 7.9 x and p-value= 0.77) the three replicates of the same condition showed variability in the values of abundance (see supplementary Table 2S.1.3 for details). Homogentisate 1, 2-dioxygenase was detected in two isolates (1FIGIMAR09 and 1ANDIMAR09 strains). Both of them were predicted to harbor the complete pathway for the degradation of this compound by genome analysis (see Table 1.6). In reference to the β-ketoadipate pathway (via protocatechuate), 4-hydroxybenzoate 3-monooxygenase (PobA) and the protocatechuate 3, 4-dioxygenase β subunit (PcaH) were detected in two isolates (*Sulfitobacter mediterraneus* 1FIGIMAR09 and *Loktanella* sp. 1ANDIMAR09) and one isolate (*Loktanella* sp. 3ANDIMAR09) respectively. The protocatechuate 3, 4-dioxygenase β subunit had a surprisingly high fold change in control condition without being statistically significant. That was because one of the three replicates of the control condition had a very high normalized abundance in comparison to the other two replicates (see in supplementary Table 2S.1.5 for details). The β-ketoadipate pathway (via protocatechuate) was predicted by genome analysis in 1FIGIMAR09, 1ANDIMAR09 and 3ANDIMAR09 isolates (see Table 1.6). Finally, benzoyl-CoA oxygenase (BoxB) was detected in *Sulfitobacter mediterraneus* 1FIGIMAR09. The degradation pathway of benzoate was predicted in this isolate by sequence homology (see Table 1.6). The fact that virtually, the majority of studied enzymes were detected with similar abundances in both conditions (values near 1 in Table 2.4) suggested that both conditions equally

Table 2.4. Detection of proteins putatively involved in degradation of aliphatic and aromatic compounds. Table shows the values of fold change and statistical probability (in parenthesis) of normalized abundances between the control and diesel oil condition.

Protein	6PALISEP08	1FIGIMAR09	3SOLIMAR09	1ANDIMAR09	3ANDIMAR09	16PALIMAR09	1CONIMAR09
Degradation of aliphatic compounds							
Alkane 1-monooxygenase	n.d.[a]	1.0 (0.91)	1.0 (0.66)	n.d.	3.4 (0.01) D	n.d.	n.d.
Aldehyde dehydrogenases	1.4 (0.26) D[b]	1.4 (0.20) C[b]	1.4 (0.44) D	1.1 (0.43) C	1.5 (0.19) D	1.1 (0.07) C	1.0 (0.08)
		1.1 (0.61) C	1.2 (0.46) C	1.0 (0.94)	2.7 (0.23) C	1.2 (0.22) C	1.0 (0.67)
					1.3 (0.48) C	1.0 (0.96)	
					1.1 (0.95) C		
Alcohol dehydrogenases	1.1 (0.68) D	1.0 (0.82)	1.0 (0.95)	n.d.	n.d.	1.0 (0.78)	1.2 (0.08) D
			1.0 (0.98)				
Degradation of aromatic compounds							
Phenylacetate-CoA ligase PaaK	1.2 (0.21) C	1.1 (0.19) D	1.0 (0.83)	n.d.	n.d.	n.d.	n.d.
	1.2 (0.58) C	1.0 (0.88)	1.0 (0.85)				
1,2-Phenylacetyl-CoA monooxygenase PaaA	1.2 (0.48) C	4.8 (0.07) D	7.9 (0.77) D	n.d.	n.d.	n.d.	n.d.
1,2-Phenylacetyl-CoA structural PaaC	n.d.	n.d.	1.2 (0.63) D	n.d.	n.d.	n.d.	n.d.
1,2-Phenylacetyl-CoA reductase PaaD	n.d.	1.2 (0.08) C	1.1 (0.4)	n.d.	n.d.	n.d.	n.d.
Homogentisate 1,2-dioxygenase HmgA	n.d.	1.1 (0.71) C	n.d	1 (0.74) D	n.d.	n.d.	n.d.
4-Hydroxybenzoate 3-monooxygenase PobA	n.d.	1.3 (0.22) C	n.d.	1.2 (0.15) C	n.d.	n.d.	n.d.
Protocatechuate 3,4-dioxygenase subunit β PcaH	n.d.	n.d.	n.d.	n.d.	15.3 (0.19) C	n.d.	n.d.
Benzoyl-CoA oxygenase, B subunit BoxB	n.d.	1.0 (0.95)	n.d.	n.d.	n.d.	n.d.	n.d.

a: n.d.: not detected.

b: D: overrepresented in proteomes of diesel oil condition, C: overrepresented in proteomes of control condition.

induced the expression of these enzymes. Additionally, it could be that the monoaromatic compounds that we analyzed were not so abundant in diesel oil reason why we did not find a diferential protein expression pattern between the two tested conditions. However, what we expected to find was higher significant differences in the detection of alkane 1-monooxygenase after addition of diesel oil. Since the experiment was done with seawater from a coastal area we could not ensure that there were no hydrocarbons present in the water, even if the samples were taken in a theoretically clean area. Maybe if the experiment would have been done using a defined medium we would have found more differences, at least for the degradation of alkanes.

2.3. Proteins involved in tolerance to hydrocarbons

We searched through the proteomes for proteins that could be involved in tolerance to diesel oil. In particular we studied proteins involved in changes in microbial cell membranes in response to solvents, removal of reactive oxygen species, expression of chaperones, efflux pumps and energy production (Ramos *et al*., 2015). The results are shown in the paragraphs below.

Firstly, we analyzed proteins related to changes in membrane fluidity (Bernal *et al*., 2007; Ramos *et al*., 1997, 2015; Sikkema *et al*., 1995). In the harbor isolates the abundance of proteins involved in *cis* to *trans* isomerization of unsaturated fatty acids was not significantly different in control and diesel-treatment conditions (data not shown). However, in the presence of diesel oil we detected proteins that could play a role in changing the permeability and/or maintaining the integrity of bacterial membrane. For example, the protein lipid A export ATP-binding/permease protein MsbA, an essential component of the outer membrane (Raetz *et al*., 2009), was detected in all isolates (see results in Table 2S.5). In *Loktanella* sp. 3ANDIMAR09 this transporter was significantly more abundant in diesel oil condition (fold change of 2.3 x, p-value= 0.04). Other membrane protein analyzed was a hypothetical protein (DUF2852, pfam11014). This protein (SPO2567) was detected in a previous proteomic study done with *Ruegeria pomeroyi* DSS-3^T when this bacterium was grown in high temperature (40 °C) and in the presence of naphthalene and diesel oil. It was proposed that SPO2567 was a hypothetical membrane protein with a role in membrane stabilization in stress conditions (Christie-Oleza *et al*., 2012). Homologs of protein SPO2567 were detected in all isolates, and in 3 of them the normalized abundance was significantly higher in diesel oil. These isolates were *Sulfitobacter mediterraneus* 1FIGIMAR09 (3.8 x, p-value= 0.002), *Sulfitobacter* sp. 3SOLIMAR09 (11.6 x, p-value= 0.0009) and *Loktanella* sp. 3ANDIMAR09 (6 x, p-value= 0.0004) (see Table 2S.5 for further information). The detection of homologs of protein SPO2567 in the presence of diesel oil in other Roseobacter isolates indicates that this protein is not only characteristic of *Ruegeria pomeroyi* DSS-3^T and that it might have a general function in stress response by roseobacters.

Next, we analyzed proteins involved in response to oxidative stress. The response against oxidative agents is activated by the presence of solvents such as alcohols or aromatic compounds that could provoke oxidative damage in bacteria. This damage is probably caused by the interference of these solvents with electron transport systems, fact that provokes an increase in the production of hydrogen peroxide and other reactive oxygen species (Brynildsen & Liao, 2009;

Domínguez-Cuevas *et al*., 2005; Ramos *et al*., 2015). Several proteins related with the response to oxidative stress were detected in our experiment. A comparison between the number of proteins participating in the removal of reactive oxygen species in the detected proteome and the annotated genome was made for each isolate. The two *Sulfitobacter* isolates (1FIGIMAR09 and 3SOLIMAR09) plus *Loktanella* sp. 3ANDIMAR09 were the isolates in which a higher number of proteins related to the response against oxidative agents were detected in reference to the annotated genome (71.4 %, 55 % and 56 % respectively). In reference to the percentage of abundances of these proteins, there were not significant differences between the control and the diesel oil condition for any of the isolates (see in Figure 2S.11). Apart from the general view, when we analyzed the proteins of this group, some of them were significantly overrepresented in diesel oil condition (Table 2.5).

Table 2.5. Proteins involved in response to oxidative stress significantly overrepresented in diesel oil condition (fold change ≥ 2, p-value ≤ 0.05).

Isolate	Protein identity	Annotation	Fold change	p-value
1FIGIMAR09	MM81_00571	Osmotically inducible protein C	3.4	0.005
3SOLIMAR09	MM82_01519	Methionine sulfoxide reductase	7.3	0.004
3ANDIMAR09	MM92_01432	Methionine sulfoxide reductase	12.6	0.007
	MM92_02343	Glutathione S-transferase	6.6	0.010

Only in three isolates (*Sulfitobacter mediterraneus* 1FIGIMAR09, *Sulfitobacter* sp. 3SOLIMAR09 and *Loktanella* sp. 3ANDIMAR09), we observed proteins at significantly higher abundances in diesel oil condition. The proteins detected were: i) the osmotically inducible protein C (OsmC) which has peroxidase activity and is involved in defense to exposure to hyperoxides or elevated osmolarity (Park *et al*., 2008), ii) the methionine sulfoxide reductase (Msr) which catalyzes the reduction of methionine sulfoxide in proteins back to methionine (Cabreiro *et al*., 2006; Vattanaviboon *et al*., 2005; Weissbach *et al*., 2005) and iii) glutathione S-transferase (GST) that constitutes a protein superfamily involved in cellular detoxification against harmful xenobiotics and endobiotics (Allocati *et al*., 2009; Oakley, 2005).

A series of chaperones related to protein re-folding after denaturation by solvents were previously described (Segura *et al*., 2005; Wijte *et al*., 2011). Most of the chaperones that were annotated in the genomes of the isolates were also detected by proteomic analysis. In *Ruegeria* sp. 6PALISEP08 and *Loktanella* sp. 1ANDIMAR09, more than the 90 % of the annotated proteins as chaperones were detected by proteomic analysis. When we compared the percentage of abundances of these proteins with all the detected proteins, we saw that none of the groups of chaperones was significantly overrepresented in diesel oil cultures (see Figure 2S.12). For comparing the abundances of chaperones in the two experimental conditions we considered a fold change ≥ 1.5. This fold change value was lower than in other compared proteins (fold change ≥ 2) because we did not expect relevant changes in abundance of these proteins. Considering this threshold, some of the detected chaperones were significantly more abundant in diesel oil condition than in control condition (see Table 2.6) suggesting a role in the tolerance of diesel oil.

Table 2.6. Chaperones that were overrepresented in diesel oil condition (fold change ≥ 1.5, p-value ≤ 0.05).

Isolate	Protein identity	Annotation	Fold change	p-value
1FIGIMAR09	MM81_02556	DnaK (Hsp70)	1.7	0.004
1ANDIMAR09	MM91_00205	DnaJ	1.6	0.020
3ANDIMAR09	MM92_00005	ClpB	2.1	0.002
	MM92_00236	DnaK (Hsp70)	1.5	0.020
	MM92_00709	DnaJ	1.8	0.020
16PALIMAR09	MM86_02526	DnaK (Hsp70)	1.5	0.003

The results obtained agreed with the described functions of these proteins. DnaK, detected in three isolates in the presence of diesel, is a heat-shock protein essential for stress tolerance that allows organisms to survive conditions that cause protein unfolding (Singh *et al*., 2007). This chaperone together with its co-chaperones (DnaJ and GrpE) play a role in a variety of cellular processes including the folding of newly synthesized proteins (Teter *et al*., 1999) as well as preventing protein aggregation under stress conditions (Gragerov *et al*., 1992; Hesterkamp & Bukau, 1998). The co-chaperone DnaJ was overrepresented in two isolates in diesel oil condition. ClpB protein, that was only significantly detected in 3ANDIMAR09, had the highest fold change. This protein promotes folding and cooperates with DnaK-DnaJ-GrpE chaperone system for suppressing protein aggregation (Barnett *et al*., 2000; Zolkiewski, 1999).

Finally, we concentrated in the analysis of membrane transport proteins. Initial studies showed the importance of efflux pumps in solvent tolerance (Isken & De Bont, 1996; Ramos *et al*., 1997). This efficient mechanism consists in pumping out solvent from the inside of the cell (membrane, periplasm or cytoplasm) to the outer medium, preventing that the chemicals reach lethal concentrations. Because some kind of transporters, such as efflux pumps, were previously described to have a role in the tolerance and/or resistance to toxic compounds (Ramos *et al*., 2015) we analyzed them in detail. Transport proteins were detected and classified in superfamilies by sequence homology using the Transporter Classification Database (TCDB) (Saier *et al*., 2014) (see section 6.3 of Materials and Methods). We classified the detected transporters (both conditions, three replicates) in superfamily categories (see Table 2.7 and for further details check Table 2S.6).

The superfamilies that had the largest number of detected transporters were the ABC superfamily, followed by a group of transporters that could not be classified by the TCDB (unclassified). There were also other abundant superfamilies and they had representative transporters of all isolates: the Mrp Superfamily, the Porin Superfamily I, the CPA Superfamily, the RND Superfamily, the APC Superfamily, the F-ATPase Superfamily, the Mot/Exb Superfamily and the COX Superfamily. Finally, there were superfamilies that had a low number of detected transporters and some of them did not have representatives of all isolates reason why we did not focus on them (i.e. LysE Superfamily, MFS Superfamily, MOP Flippase Superfamily, P-ATPase Superfamily, and UT/RnfD/NqrB Superfamily).

Apart from the number of detected transporters, the percentage of abundance of proteins in each superfamily was analyzed under both tested conditions (see Table 2.8). The transporter superfamilies that had higher percentages of abundance were the ABC superfamily, the porin su-

Table 2.7. Number of detected transporters classified in superfamilies using the TCDB (Saier *et al.*, 2014).

Superfamilies	**6PALISEP08**	**1FIGIMAR09**	**3SOLIMAR09**	**1ANDIMAR09**	**3ANDIMAR09**	**16PALIMAR09**	**1CONIMAR09**
ABC (ATP-binding cassette transporters)	47	62	73	78	66	36	40
ABC importers	38	50	62	64	48	29	31
ABC exporters	9	12	11	14	18	7	9
APC (amino acid-polyamine-organocation)	6	3	7	5	6	2	3
CPA (monovalent cation: proton antiporter)	5	5	6	6	6	5	5
Mrp (Na^+ Transporting Mrp)	12	12	11	13	16	8	11
Porin I	11	11	13	10	10	9	10
RND (resistance-nodulation-cell division)	5	4	4	6	6	6	2
F-ATPase (H^+ or Na^+ translocating F-type, V-type and A-type ATPase)	5	5	3	4	5	3	3
Mot/Exb (outer membrane transport energizer)	2	3	1	2	2	1	1
COX (proton-translocating cytochrome oxidase)	2	2	1	3	1	2	1
VIC (voltage-gated ion channel)	2	1	2	1	1	1	1
RTX-Toxin (repeats in-toxin toxin)	n.d.[a]	2	2	1	2	1	n.d.
P-ATPase (P-type ATPase)	n.d.	n.d.	n.d.	n.d.	1	2	2
UT/RnfD/NqrB (urea transporter/Na^+ exporter)	3	n.d.	n.d.	n.d.	n.d.	2	2
LysE (lysine exporter)	n.d.	n.d.	n.d.	1	n.d.	n.d.	n.d.
MF (major facilitator)	n.d.	n.d.	n.d.	1	n.d.	n.d.	n.d.
MOP Flippase (multidrug/oligosaccharidyl-lipid/polysaccharide flippase)	n.d.	n.d.	1	n.d.	n.d.	n.d.	1
Unclassified	45	39	34	50	59	31	33
Total transporters (% of annotated)	146 (11.7)	151 (11.8)	160 (12.6)	182 (12.4)	184 (13.6)	110 (11.7)	116 (10.9)

a: n.d.: not detected. The identification of these transporters was based on a minimum of 30% of identity in the amino acid sequence and an *E* value below $1 \cdot 10^{-5}$.

perfamily I, the F-ATPase superfamily and the group of unclassified transporters (see in Table 2.8 and 2S.8). There were only few cases in which significant differences were observed between the treatments. Those were the cases of Porin I superfamily, F-ATPase superfamily and in the majority of isolates, the transporters that could not be classified (for more detals check Table 2S.8). Proteomic studies showed that different kind of porins such as OprE, OprH or Opr86 could be upregulated and others such as OprD and OprF could be downregulated in the presence of organic solvents (e.g. phenols or alkanes) indicating a role in the adaptation to solvent exposure (Hemamalini & Khare, 2014; Roma-Rodrigues *et al*., 2010). The transporters of Porin I superfamily were overrepresented i) in control condition in *Ruegeria* sp. 6PALISEP08 and ii) in diesel oil condition in *Sulfitobacter mediterraneus* 1FIGIMAR09. The F-ATPase superfamily transporter catalyzes the synthesis of ATP from ADP and inorganic phosphate when H^+ or Na^+ flow through the ATP synthase. These enzymes also operate in the opposite direction, pumping H^+ or Na^+ from the cytoplasm out of the cell hydrolyzing ATP (Ferguson *et al*., 2006; Nath, 2002). The transporters of F-ATPase superfamily were overrepresented in control condition in *Thalassobacter stenotrophicus* 1CONIMAR09.

Because we detected few significant differences in proteins involved in transport between the treatments when we considered the percentages of abundances of each superfamily category, we increased the resolution analyzing the proteins detected in the main categories one by one

Table 2.8. Percentages of abundance of the detected transporters classified in superfamilies according to the TCDB (Saier *et al.*, 2014) (averages ± sd of three replicates). Values in bold indicate that there was a significant difference between control and diesel oil condition (T-student test p-value ≤ 0.05). Superfamilies that had percentages of abundance lower than 1 % are not shown (see Table 2S.7 for more details).

	6PALISEP08		**1FIGIMAR09**		**3SOLIMAR09**		**1ANDIMAR09**		**3ANDIMAR09**		**16PALIMAR09**		**1CONIMAR09**	
Superfamily	**Control**	**Diesel**	**Control**	**Diesel**	**Control**	**Diesel**	**Control**	**Diesel**	**Control**	**Diesel**	**Control**	**Diesel**	**Control**	**Diesel**
ABC	3.8±0.32	4.3±0.29	5.0±0.76	5.2±0.28	2.5±0.09	2.5±0.14	8.8±0.69	8.6±0.12	2.8±0.11	2.4±0.21	4.6±0.71	3.7±0.44	4.5±0.13	4.3±0.19
Importers	3.5±0.34	4.0±0.31	4.6±0.75	4.9±0.27	2.2±0.08	2.2±0.12	8.6±0.70	8.4±0.12	2.3±0.10	2.0±0.20	4.5±0.70	3.6±0.43	4.3±0.13	4.2±0.20
Exporters	0.3±0.03	0.3±0.03	0.4±0.02	0.4±0.02	0.3±0.02	0.3±0.03	0.3±0.01	0.3±0.02	0.5±0.06	0.5±0.01	0.2±0.01	0.2±0.01	0.2±0.01	0.2±0.003
Porin I	**1.7±0.22**	**1.3±0.04**	**2.5±0.52**	**3.5±0.33**	2.6±0.24	3.4±0.80	2.4±0.16	2.3±0.12	0.7±0.10	0.7±0.04	1.4±0.22	1.2±0.09	0.7±0.06	0.6±0.09
F-ATPase	7.1±0.29	7.0±0.33	4.9±0.27	5.5±0.36	5.3±0.21	5.2±0.19	5.2±0.48	5.4±0.14	6.3±0.30	5.7±0.19	4.1±0.05	4.2±0.16	**4.7±0.08**	**4.3±0.13**
Unclassified	2.9±0.15	2.9±0.17	**3.0±0.04**	**3.5±0.09**	**3.1±0.07**	**3.3±0.07**	2.6±0.12	2.7±0.02	**4.4±0.14**	**4.9±0.09**	**2.8±0.21**	**3.3±0.14**	**3.3±0.11**	**3.5±0.09**

(see Table 2.9). Four of the ABC transporters were significantly more abundant in control condition and all of them were nutrients importers. On the other hand, in 3ANDIMAR09 isolate one ABC exporter was more detected in diesel oil condition. This transporter was the lipid A export ATP-binding/permease protein MsbA that was previously commented when discussing membrane changes. With respect to Porin I superfamily, in isolate 1FIGIMAR09 we detected a protein increased in diesel oil (see Table 2.9). This protein was annotated as an outer membrane protein that belonged to OmpA family. OmpA has both structural and ion-permeable porin roles, with its ionic pore controlled by a salt-influenced electrostatic gating mechanism that allows bacterial survival during osmotic stress (Hong *et al*., 2006). Additionally, in *Escherichia coli*, OmpA was upregulated during phenol exposure to enhance the cell resistance (Zhang *et al*., 2011). Then, this protein could be related to the tolerance to diesel oil in this strain.

Three proteins of *Sulfitobacter mediterraneus* 1FIGIMAR09 and one protein in *Loktanella* sp. 3ANDIMAR09 were overrepresented in one of the tested conditions. In the presence of diesel oil we detected a putative protease in 1FIGIMAR09. This protein seems to be involved in a general secretion pathway for protein export. In 3ANDIMAR09 a protein involved in plasmid replication was significantly more abundant in the presence of diesel oil (see Table 2.9).

Apart from the proteins that had higher percentages of detection (> 1 %, see in Table 2.8), we also focused in the study the RND superfamily, which has been described as the most important for the efflux of organic solvents and solvent tolerance (Ramos *et al*., 2015). In general, we did not detect significant differences between the proteomes of the treatments at a broad level of resolution (see in Table 2S.7). However, when we analyzed the proteins in this superfamily one by one, we found that in *Thalassobacter stenotrophicus* 16PALIMAR09 there was an AcrAD-TolC multidrug efflux transport system–permease subunit (MM86_02170) that was more abundant in diesel oil condition than in the control (2.1 x, p-value= 0.02). AcrAD are believed to form a complex with TolC for multidrug export (Aires *et al*., 2005).

Solvent tolerance is an energy intensive process. When *Pseudomonas putida* DOT-T1E and *Pseudomonas putida* S12 were exposed to toluene, enzymes related to energy production and enzymes of the tricarboxylic acid cycle (TCA) were observed by proteomic analysis (Segura *et a*l., 2005; Wijte *et al*., 2011). That meant that the response to solvent toxicity included a requirement for energy production (Ramos *et al*., 2015). In our experiment, when the percentage of abundance of the detected proteins that belonged to (C) “energy production and conversion” COG functional category were considered all together no significant differences were detected between the two cultured conditions for any isolate (see Table 2.3). However, when the detected proteins were analyzed one by one, few significant differences (p-value ≤ 0.05) were observed. In *Sulfitobacter mediterraneus* 1FIGIMAR09 a carbon monooxide dehydrogenase CoxL (MM81_03049) was more abundant in diesel oil condition (7.7 x, p-value= 0.0004). This enzyme is a member of the *coxSLM* operon which is used by bacteria to oxidize carbon monoxide to carbon dioxide in order to obtain energy (see section 1.2 of Introduction). These results agreed with the genomic prediction in which in all isolates the Cox

Table 2.9. Transporters that were overrepresented in one of the treated conditions (fold change ≥ 2 and p-value ≤ 0.05).

Isolate	Protein identity	Transporter	Fold change	p-value
ABC superfamily				
6PALISEP08	MM88_03273[a]	Glutathione ABC transporter - periplasmic binding protein	2.8 C[c]	0.02
	MM88_02055[a]	Putative ABC transporter, periplasmic substrate-binding protein	2.8 C	0.02
1FIGIMAR09	MM81_01499[a]	Sugar ABC transporter, periplasmic sugar-binding protein	2.1 C	0.02
1ANDIMAR09	MM91_02981[a]	Sugar ABC transporter, periplasmic sugar-binding protein	2.3 C	0.01
3ANDIMAR09	MM92_02989[b]	ATP-binding lipopolysaccharide transport protein	2.3 D[c]	0.04
Porin I superfamily				
6PALISEP08	MM88_03056	Outer membrane porin	4.4 C	0.02
1FIGIMAR09	MM81_01947	Outer membrane protein	11.5 D	1.6×10^{-5}
Unclassified				
1FIGIMAR09	MM81_01974	putative protease, membrane anchored	2.7 D	0.002
	MM81_00694	flagellar motor switch protein FliN	2.4 C	0.03
	MM81_00810	putative secretion protein	2.2 C	0.004
3ANDIMAR09	MM92_03672	protein involved in plasmid replication	3.3 D	0.02

a: ABC importers; b: ABC exporters.

c: D: overrepresented in proteomes of diesel oil condition, C: overrepresented in proteomes of control condition.

proteins type II were predicted (see Table 1.5). In *Loktanella* sp. 1ANDIMAR09 and *Thalassobacter stenotrophicus* 1CONIMAR09 a poly-β-hydroxyalkanoate (PHA) depolymerase (DepA) (MM91_03351 and MM87_01193, respectively) was overrepresented in diesel oil condition (3.9 x, p-value= 0.02 and 1.53 x , p-value= 0.01) although in 1CONIMAR09 isolate the fold change was lower than 2. The degradation of PHAs would be followed by steps of the β-oxidation of fatty acids until the production of acyl-CoAs and acetyl-CoA, which would be metabolized in the TCA as a source of carbon and energy (Ruiz *et al*., 2001). So, the depolymerization of PHA would act as both an input of energy and a source of carbon.

Finally, since it has been suggested that *Alcanivorax borkumensis* SK2^T, a specialist alkane degrader, used pili to have an easy access to the hydrocarbon (Schneiker *et al*., 2006) we also analyzed the presence of pili proteins in the proteomes. None of the detected pili proteins were significantly more abundant in the presence of diesel oil in any of the isolates. Therefore, we could not conclude that the isolates used this strategy for acessing the hydrocarbon. The results of pili protein detection are explained in detail below in section 2.3.

Considering the proteins related to degradation and tolerance to diesel oil that were overrepresented in diesel oil treatments we could highlight some results. For example, isolate *Loktanella* sp. 3ANDIMAR09 had a variety of genes for aromatic and aliphatic compounds degradation (see in Table 1.6 of chapter 1) and it was also able to grow using diesel oil as a sole source of carbon and energy (see in Table 1.7 of chaper 1). In the proteomic analysis, this isolate stood out for the overexpression in diesel oil treatment of i) an alkane 1-monooxygenase (3.4 x, p-value = 0.01); ii) a lipid A export ATP-binding/permease protein MsbA (2.3 x, p-value = 0.04), iii) the hypothetical transmembrane protein (DUF2852) homologous of

SPO2567 (6 x, p-value = 0.0004), iv) proteins involved in response to oxidative stress [methionine sulfoxide reductase (12.6 x, p-value= 0.007) and glutathione S-transferase (6.6 x, p-value= 0.01)], and v) chaperones [ClpB (2.1 x, p-value= 0.02), DnaK (Hsp70) (1.5 x, p-value= 0.02), and DnaJ (1.8 x, p-value= 0.02)]. All these enzymes could probably contribute to the growth and tolerance of to diesel oil of *Loktanella* sp. 3ANDIMAR09.

Other isolates that grew with diesel oil were *Ruegeria* sp. 6PALISEP08 and *Sulfitobacter* sp. 3SOLIMAR09 (see Table 1.7 of chapter 1). However, in these isolates we could not detect any protein related to degradation overrepresented in the presence of diesel oil. We previously mentioned that the fact that we did not know the exact composition of the collected seawater makes that we ignore if there were some aliphatic or aromatic compound that induced the expression of these enzymes. In 6PALISEP08 isolate, alkane 1-monooxygenase was not detected, but there were proteins of the phenylacetate degradation pathway that could contribute to degrade some aromatic compounds that could be present in seawater (see Table 2.4). In *Sulfitobacter* sp. 3SOLIMAR09 the alkane 1-monooxygenase and proteins related to the degradation of phenylacetate were detected without being significant differences in the abundance between the two tested conditions (see Table 2.4). Detected proteins involved in the tolerance of diesel oil were: i) the hypothetical transmembrane protein (DUF2852) homologous of SPO2567 (11.6 x, p-value = 0.0009) and ii) methionine sulfoxide reductase (7.3 x, p-value = 0.004). Few proteins were overrepresented in diesel oil in 6PALISEP08 and 3SOLIMAR09 isolates in comparison with 3ANDIMAR09. It could be that with the cultured and incubation conditions that we used for the proteomic analysis, these two isolates did not induce the expression of the degradation or tolerance genes to face the presence of diesel oil, possibly because of the lack of nutrients (see below in section 2.3).

2.4. Proteins involved in stringent response

Apart form analysing proteins that could be related to degradation and tolerance, the analysis was extended to the detected proteome that were statistically significant (ANOVA, Progenesis analysis) and/or important for the ordination obtained with PCA (see above section 2.1 and Table 2S.3). The aim was to have information about global cell metabolism in the conditions of the experiment. The analysis of these proteins evidenced that the cells were under a situation of stringent response. The stringent response is a broadly conserved bacterial stress response that controls adaptation to nutrient deprivation that allows bacteria to quickly reprogram transcription when there are changes to nutrient availability (Boutte & Crosson, 2013). In the majority of cases there were not significant differences in the percentage of abundance of stringent response related proteins between the two treatments by ANOVA test, which means that this cell response was occurring in both cases. Taking into account how the experiment was done, i.e. transferring cells grown in MB to sterile nutrient-poor oligotrophic seawater (see section 6.1 of Material and Methods), these metabolic changes were coherent. The proteins that evidenced a cellular stringent response are shown in Table 2.10 and presented in the paragraphs below.

Firstly, we detected (p)ppGpp synthetases (RelA/SpoT family) and guanosine pentaphosphate phosphohydrolases (GppA) in all isolates (see Table 2.10). (P)ppGpp is an alarmone that plays a central role in the bacterial stringent response. It is induced by starvation (Srivatsan & Wang, 2008) and, it is synthesized from adenosine triphosphate (ATP) and either guanosine diphosphate (GDP) to make guanosine 5'-diphosphate 3'-diphosphate (ppGpp) or guanosine triphosphate (GTP) to make 5'-triphosphate 3'-diphosphate (pppGpp) (Mechold *et al*., 2013). In *Escherichia coli*, two proteins were involved in stress-induced (p)ppGpp accumulation: RelA and SpoT. RelA is a ribosome-associated (p)ppGpp synthetase responding mainly to uncharged tRNAs that accumulate as a result of amino acid limitation. SpoT is a bifunctional (p)ppGpp synthetase and hydrolase, probably regulating (p)ppGpp levels in response to several conditions other than amino acid limitation (Magnusson *et al*., 2005; Wendrich *et al*., 2002). Guanosine pentaphosphate phosphohydrolase (GppA) catalyse the conversion of pppGpp to ppGpp). Therefore, these two proteins [(p)ppGpp synthetase (RelA/SpoT family) and GppA] are related to the synthesis and conversion of the alarmone (p)ppGpp, suggesting a possible response to starvation in both tested conditions. This conclusion was also supported by the large number of ABC importers detected by proteomics (see Tables 2.7).

Another protein detected in all isolates was the Universal Stress Protein (UspA) (see Table 2.10). The expression of UspA is increased when the cell is exposed to stress agents such as starvation (carbon, nitrogen, phosphate, sulfate, and amino acids), as well as exposure to heat, oxidants, metals, uncouplers of the electron transport chain, polymyxin, cycloserine, ethanol, and antibiotics (Nachin *et al*., 2005). It is induced by the alarmone (p)ppGpp (Kvint *et al*., 2003). Therefore, its detection agreed with the results presented above and indicated a situation of cellular stress response in the conditions of the experiment, independently of the presence of diesel oil.

Other proteins related to stringent response were only detected in some isolates. This was the case of proteins involved in the metabolism of poly-β-hydroxyalkanoates (PHAs). Proteomic data provided evidence of a poly-β-hydroxyalkanoate synthesis repressor (PhaR) in 4 isolates and a poly-β-hydroxyalkanoate depolymerase (DepA) in three of the isolates (see Table 2.10). In isolates 1ANDIMAR09 and 1CONIMAR09, the PHA depolymerase was significantly more detected in diesel oil cultures as described above in section 2.2. These results point to the utilization of intracellular PHA granules by some isolates under the conditions of the experiment, probably to adjust to nutrient and carbon deprivation. This has been described also in a proteomic analysis performed with *Roseobacter litoralis* Och 149^{T} under nutrient and carbon limitation (Zong & Jiao, 2012).

Finally, in some isolates we observed proteins involved in transcription and translation: a bacterial transcription factor (BolA) and a ribosome silencing factor (RsfS). BolA has a relevant role in the adaptation of *Escherichia coli* to general stresses. Under poor growth conditions, BolA is essential for normal cell morphology (Guinote *et al*., 2014). Its overexpression induces biofilm formation (Dressaire *et al*., 2015) and alters the properties of the outer cell membrane (Freire *et al*., 2006), making it less permeable to harmful agents. Its transcription is regulated

Table 2.10. Detected proteins related to stringent response.

Isolate	Protein identity	Fold change	p-value
(p)ppGpp synthetase (RelA/SpoT family)			
6PALISEP08	MM88_03361	1.11 C[a]	0.69
1FIGIMAR09	MM81_03343	1.06 D[a]	0.82
3SOLIMAR09	MM82_00386	1.21 C	0.44
1ANDIMAR09	MM91_02727	1.00	0.88
3ANDIMAR09	MM92_00536	1.34 D	0.28
16PALIMAR09	MM86_02752	1.05 C	0.58
1CONIMAR09	MM87_01462	1.10 C	0.39
Guanosine pentaphosphate phosphohydrolase (GppA)			
6PALISEP08	MM88_00981	1.20 C	0.04
1FIGIMAR09	MM81_00972	1.08 C	0.63
3SOLIMAR09	MM82_03053	1.13 D	0.49
1ANDIMAR09	MM91_01879	1.01 C	0.94
3ANDIMAR09	MM92_02765	1.01 D	0.21
16PALIMAR09	MM86_01352	1.32 D	0.18
1CONIMAR09	MM87_00550	1.11 C	0.52
Universal Stress Protein (UspA)			
6PALISEP08	MM88_01524	1.17 D	0.63
1FIGIMAR09	MM81_01276	1.27 C	0.22
3SOLIMAR09	MM82_03159	1.09 C	0.40
1ANDIMAR09	MM91_01986	1.08 C	0.30
3ANDIMAR09	MM92_00083	1.16 D	0.20
16PALIMAR09	MM86_00860	1.03 D	0.74
1CONIMAR09	MM87_00057	1.11 C	0.26
Poly-β-hydroxyalkanoate synthesis repressor (PhaR)			
6PALISEP08	MM88_00072	1.06 D	0.76
1FIGIMAR09	MM81_01029	22.14 C	0.20
1ANDIMAR09	MM91_03354	1.29 C	0.17
1CONIMAR09	MM87_01190	1.29 C	0.07
Poly-β-hydroxyalkanoate depolymerase (DepA)			
1FIGIMAR09	MM81_01026	1.00	0.90
1ANDIMAR09	MM91_03351	3.91 D	0.02
1CONIMAR09	MM87_01193	1.53 D	0.01
Transcription factor BolA (BolA)			
6PALISEP08	MM88_00901	1.49 D	0.43
1FIGIMAR09	MM81_03712	2.29 C	0.32
3SOLIMAR09	MM82_02344	1.11	0.83
Ribosome silencing factor (RsfS)			
6PALISEP08	MM88_03778	1.36 D	0.45
1FIGIMAR09	MM81_03090	1.25 C	0.27

a: D: overrepresented in proteomes of diesel oil condition, C: overrepresented in proteomes of control condition.

by several factors, including the (p)ppGpp (Guinote *et al*., 2014). This protein was detected in isolates 6PALISEP08, 1FIGIMAR09 and 3SOLIMAR09 (see Table 2.10). In the first two isolates we detected a ribosome silencing factor RsfS as well. This protein plays a role in silencing ribosome activity in the stationary phase and during the transition from rich to poor media; it helps cells to adapt to poor nutrient media and restricted energy conditions by reducing the protein synthesis and thereby saving energy (Häuser *et al*., 2012; Starosta *et al*., 2014).

Two of the proteins detected, BolA and UspA, are involved in the induction of biofilm formation in bacteria (Guinote *et al*., 2014; Zhang *et al*., 2013). Biofilms are also a strategy for cell survival (Dunne, 2002) and, on the other hand, roseobacters are well know for their ability for cell-surface interactions and attachment (Slightom & Buchan, 2009). Therefore, we

analyzed the proteomes for pili proteins. In all isolates proteins related to the synthesis of pili were detected (see Table 2S.9 for further details). In isolates 1FIGIMAR09, 6PALISEP08, 3ANDIMAR09 and 1ANDIMAR09 most of the proteins of the Flp genic cluster *cpaBC-ompA-cpaEF-tadBC* (see section 4 in chapter 1, and Table 1S.12) were detected. In 1FIGIMAR09, the 7 proteins of the Flp cluster were detected; 4 of these proteins were statistically more abundant in control conditions.

CHAPTER 3. ANALYSIS OF HORIZONTAL GENE TRANSFER IN THE ROSEOBACTER LINEAGE

The study of horizontal gene transfer (HGT) processes is a way to understand the versatility and genomic diversity that makes the Roseobacter lineage highly adaptable to environmental conditions. Genes related to adaptation to environmental changes such as those involved in hydrocarbon degradation/tolerance or metal and antibiotic resistance are susceptible to be transferred horizontally (Segura *et al*., 2014). For this reason, we studied the relevance of three mechanisms of HGT in genomes of the lineage: gene transfer agents (GTAs), extrachromosomal elements (ECEs) and transposases in genomes of the Roseobacter lineage. GTAs and RepABC plasmids were previously studied in cultured members of the lineage (21 and 17 analyzed genomes, respectively) showing a widespread distribution among their members (Biers *et al*., 2008; Petersen *et al*., 2009). In reference to ECEs, Petersen and collaborators (2013) showed that roseobacters can have until four non-homologous replication systems (RepABC, RepA, RepB and DnaA-like) and studied the distribution of these replicons in 8 members of *Rhodobacterales* order. What we did was to provide a review that includes an updating of these two mechanisms (GTAs and ECEs) of HGT in 96 roseobacter genomes. On the other hand, transposase presence and fluxes within the lineage have not been evaluated before this study. We mainly focused on the analysis of transposases fluxes based on the study of shared transposases, as an indicative of recent horizontal gene transfer events, and its relation with i) the phylogenomic distance of the genomes, and ii) the sharing of habitat and geographical location of isolation.

1. Gene transfer agents (GTAs)

A GTA is a phage-like entity that constitutes a model of HGT similar to virus-mediated generalized transduction (see section 1.3.1. in the Introduction). GTAs are encoded by a genic cluster formed by 17 structural genes (Lang & Beatty, 2000, 2001; Lang *et al.*, 2002) with an approximate lenght of 15 kb. Some of the GTA genes are homologs of phage structural genes: large terminase (gene 2), portal protein (gene 3), prohead protease (gene 4), major capsid protein (gene 5), head-tail adaptor protein (gene 7), major tail protein (gene 9) and tail tape measure protein (gene 11). Other genes are described as putative GTAs (genes 1, 3.5, 6, 8, 10, 10.5, 12, 13, 14 and 15) (see Figure 2I). This genic cluster was searched in the 96 Roseobacter genomes analyzed in this study (see section 7.1 of Materials and Methods for further details).

The GTA gene cluster was identified in 85 out of 96 genomes. The 11 genomes were we could not predict the GTA cluster belonged to different phylogenomic groups and genera. They were: *Ruegeria mobilis* F1926 (G1), *Ruegeria* sp. 6PALISEP08 (G1), *Ruegeria* sp. KLH11 (G1), *Sulfitobacter mediterraneus* 1FIGIMAR09 (G2), *Roseivivax atlanticus* 22II-s10s^{T} (G3), *Donghicola* sp. S598 (G3), *Thalassobium* sp. R2A62 (G4), *Litoreibacter arenae* DSM 19593^{T} (G5), *Rhodobacterales* bacterium HTCC2255 (G5) and *Rhodobacteraceae* bacterium HIMB11 (G6), *Rhodobacterales* bacterium HTCC2083 (G6). For further details about the complete structure of GTA genic cluster in each genome see supplementary Figure 3S.1. Protein identity and their location can be seen in supplementary Table 3S.1. Despite the GTA module was well-conserved in Roseobacter lineage, four of the genes (1, 2, 3.5 and 10.5) were not detected in

some genomes (see Table 3.1). In the case of gen 1, the genic product was not detected in 5 genomes. It was not found in *Sulfitobacter* sp. 20_GPM-15909m (G2), *Roseovarius* spp. 217 and TM1035 (G3) and, *Celeribacter baekdonensis* B30 and *Dinoroseobacter shibae* DFL 12^T (G5). The genic product of gen 2 was not detected in *Roseovarius* sp. 217 (G3). The genic product of gen 3.5 was not found in the genomes of *Maribius* sp. MOLA 401 and *Oceaniovalibus guishaninsula* JLT2003^T from genomic group G4 and in the genomes of *Celeribacter baekdonensis* B30 and *Maritimibacter alkaliphilus* HTCC2654^T from G5. The genic product of gen 10.5 was not detected in *Sulfitobacter mediterraneus* KCTC 32188^T (G2). The genes that were not detected in the GTA module had small sizes in comparison to other GTA genes: gene 1 (~ 350 bp), gene 3.5 (~ 240 bp) and gene 10.5 (~ 190 bp).

Table 3.1. Percentage of detection of each gene of the GTA module in each phylogenomic group. The genes that are not shown were detected in all genomes.

	Phylogenomic group				
Gene	**G1**	**G2**	**G3**	**G4**	**G5**
1	100	92.8	89.5	100	80
2	100	100	94.7	100	100
3.5	100	100	100	88.8	80
10.5	100	93.3	100	100	100

In all closed genomes, GTA genes were located in the chromosome, in agreement with their proposed vertical transmission (Lang & Beatty, 2006; Lang *et al*., 2002). In the draft genomes in the majority of cases the whole structure was located in the same contig but there were some exceptions. Fragments of GTA module were detected in more than one contig in 5 genomes (see Figure and Table 3S.1): *Leisingera* sp. ANG1 (G1), *Pseudodonghicola xiamenensis* DSM 18339^T (G1), *Rhodobacteraceae* bacterium PD-2 (G3), *Loktanella cinnabarina* LL-001^T (G4) and *Rubellimicrobium mesophilum* DSM 19309^T (G4).

In 13 genomes additional genes not related to GTA were identified in the GTA module (see Table 3.2, and supplementary Table and Figure 3S.1 for details). In 11 of these genomes the non-related GTA genes were located between gene 2 (large terminase) and gene 3 (portal protein). Interestingly, the inserted gene coded for a putative acyltransferase in seven of these eleven roseobacters: the 5 genomes of *Citreicella* spp., *Pelagibaca bermudensis* HTCC2601^T, and *Salipiger mucosus* DSM 16094^T, all of them from phylogenomic group G3. The percentage of aminoacid sequence identity of the putative acyltransferase among the 5 *Citreicella* isolates was 99 % or higher in 100 % of the sequence (158 aligned positions). When we made the comparison with *Pelagibaca bermudensis* HTCC2601^T and *Salipiger mucosus* DSM 16094^T the percentage of identity of this genic product among all 7 genomes was higher than 60 % in 100 % of the sequence (158 or 160 aligned positions respectively). These results suggested that this gene could be inserted in some ancestor of phylogenomic group G3 and as members of G3 evolved, this gene also did it. Other cases of non-related GTA genes located between gene 2 and 3 were possible oxidoreductases but most of them were proteins with unknown function (see Table 3.2). In two cases, the not related GTA genes were located between gene 1 (putative GTA) and gene 2 (large terminase); both of them were uncharacterized proteins. In

Loktanella cinnabarina LL-001^T, gene 2 was truncated, possibly due to the insertion of a gene for a hypothetical protein (see Figure 3S.1 and the annotation of the genome in Table 1S.3.15). In *Octadecabacter arcticus* 238^T two genes (one codified for a transposase and the other one for a putative integrase) interrupted gene 15, the longest of the GTA module. In other members of *Alphaproteobacteria* Class such as *Brucella melitensis* non-GTA genes were reported in its GTA module (Lang & Beatty, 2006).

Table 3.2. Non-related GTA genic products detected in the GTA module of 13 members of Roseobacter lineage.

Position in GTA module	Genome	Protein code	Protein function
Between genes 1 - 2	*Loktanella cinnabarina* LL-001^T	MM15_02207	Hypothetical protein
	Sulfitobacter guttiformis KCTC 32187^T	MM79_01019	Uncharacterized protein
Between genes 2 - 3	*Citreicella aestuarii* 328	MM95_00592	Putative acyltransferase
	Citreicella aestuarii 329	MM96_00959	Putative acyltransferase
	Citreicella aestuarii 357	MM05_00578	Putative acyltransferase
	Citreicella aestuarii AD8^T	MM94_00086	Putative acyltransferase
	Citreicella sp. SE45	MM05_00578	Putative acyltransferase
	Pelagibaca bermudensis HTCC2601^T	MM29_00955	Putative acyltransferase
	Salipiger mucosus DSM 16094^T	MM68_02931	Putative acyltransferase
	Ruegeria lacuscaerulensis ITI-1157	MM64_00841	Oxidoreductase
	Ruegeria conchae TW15^T	MM63_02019	Hypothetical protein
		MM63_02020	Uncharacterized protein
		MM63_02021	Hypothetical protein
		MM63_02022	Cysteine hydrolase
	Sulfitobacter guttiformis KCTC 32187^T	MM79_01019	Uncharacterized protein
	Wenxinia marina DSM 24838^T	MM85_03841	Uncharacterized oxidoreductase
Within gene 15	*Octadecabacter arcticus* 238^T	MM28_01258	IS200/IS605 family TnpA
		MM28_01259	Putative integrase

In Figure 3.1, the genic structures of GTA of *Rhodobacter capsulatus* (RcGTA) (Lang & Beatty, 2000; Lang *et al*., 2002) and the conserved genes of the GTA module of each phylogenomic group are shown. The synteny among the RcGTA and the GTAs of members of Roseobacter lineage was well-conserved. If we compare the RcGTA-like gene cluster organization with members of other orders within the *Alphaproteobacteria* outside the *Rhodobacterales* (e.g. *Rickettsiales*, *Rhizobiales*), this synteny was not maintained (Lang & Beatty, 2006). This result suggests that the GTA structure evolved differently from an original ancestor in the different alphaproteobacterial orders. In fact, phylogenetic analysis using one (protein 5) or the concatenation of three GTA genic products (proteins 3, 5 and 12 hypothetical protein) were done previously and showed a clear difference between the orders of *Alphaproteobacteria* in which the GTA structure was identified (Biers *et al*., 2008; Lang & Beatty, 2006).

In our case, because the synteny of the GTA genes was similar among the members of Roseobacter lineage of different phylogenomic groups, we wanted to study if these genes could be divergent enough to reproduce the core proteome phylogeny. For this purpose, 12 conserved genic products (genes 3–9 and 11–15) were selected for phylogenetic analysis. Figure 3.2 shows the phylogenetic reconstruction obtained with single-copy proteins of core proteome (114 proteins, 96 genomes) in comparison with the phylogeny of the concatenation of the 12 GTA genic products. In total, 2,101 amino acid positions (44 % of the total sequence

length) resulting from the exclusion of the highly variable parts of the concatenated GTA sequences were used for the phylogeny (see section 7.1 of Materials and Methods for details). As can be seen in Figure 3.2 both phylogenies were in good agreement. The defined phylogenomic groups were also maintained in the GTA tree, with the only discrepancy of *Oceaniovalibus guishaninsula* JLT2003^{T} and *Maribius* sp. MOLA 401 which were affiliated to G4 in the core proteome phylogeny and to G5 in the GTA phylogeny. The same discrepancy was also observed in the phylogeny that was made using the concatenation of 14 conserved genic products (see in section 3 of chapter 1). These results indicated that the GTA phylogeny is also a valid method for reproducing the phylogeny of the lineage but has the limitation that not all roseobacters harbor the GTA module.

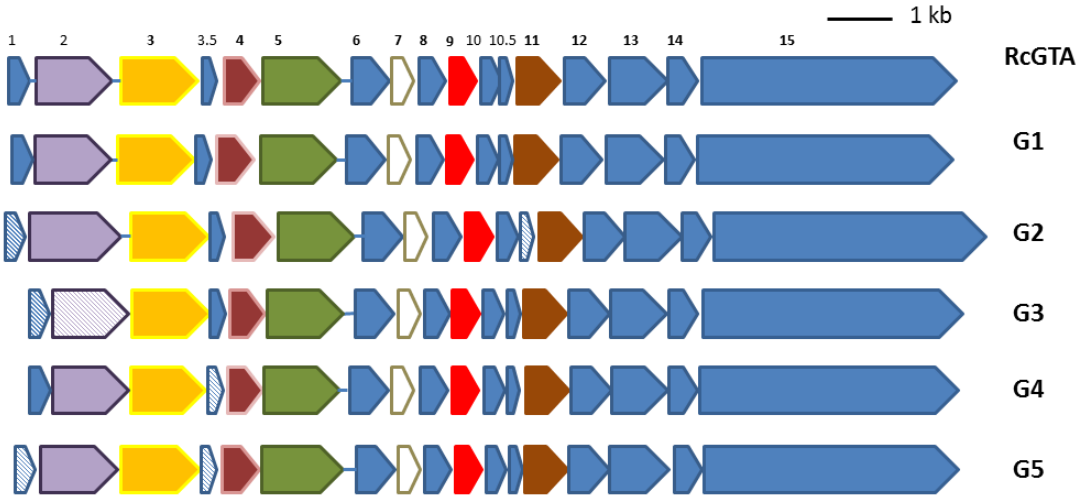

Figure 3.1. Genetic organization of the RcGTA of *Rhodobacter capsulatus* and the representative GTA genes of each phylogenomic group of Roseobacter lineage (G1, G2, G3, G4 and G5). Gene number is indicated in the upper part. Gene with assigned functions: large terminase (2), portal protein (3), prohead protease (4), major capsid protein (5), head-tail adaptor protein (7), major tail protein (9) and tail tape measure protein (11). Genes in blue color have not a defined function. Genes shown in bold were used for phylogenetic analysis. Genes that were not detected in all genomes of each phylogenomic group are shown with line pattern (for more details see in Table 3.1 and supplementary Table and Figure 3.S1).

The presence of almost the complete GTA cluster in the 88.5 % of cultured Roseobacter genomes analyzed made evident that the conservation of GTA cluster could result benefitial for members of the lineage.

2. Extrachromosomal elements (ECEs)

For the study of ECEs elements we analyzed the distribution and phylogeny of RepABC family plasmids and the distribution of other replicons of roseobacters (RepA, RepB, and DnaA-like).

The RepABC family replicons are widely distributed in the Class *Alphaproteobacteria* as low-copy number plasmids. They are named according to the genetic arrangement of *repA*, *repB* (genes for partioning and segregation) and *repC* (replication gene). These components are, in

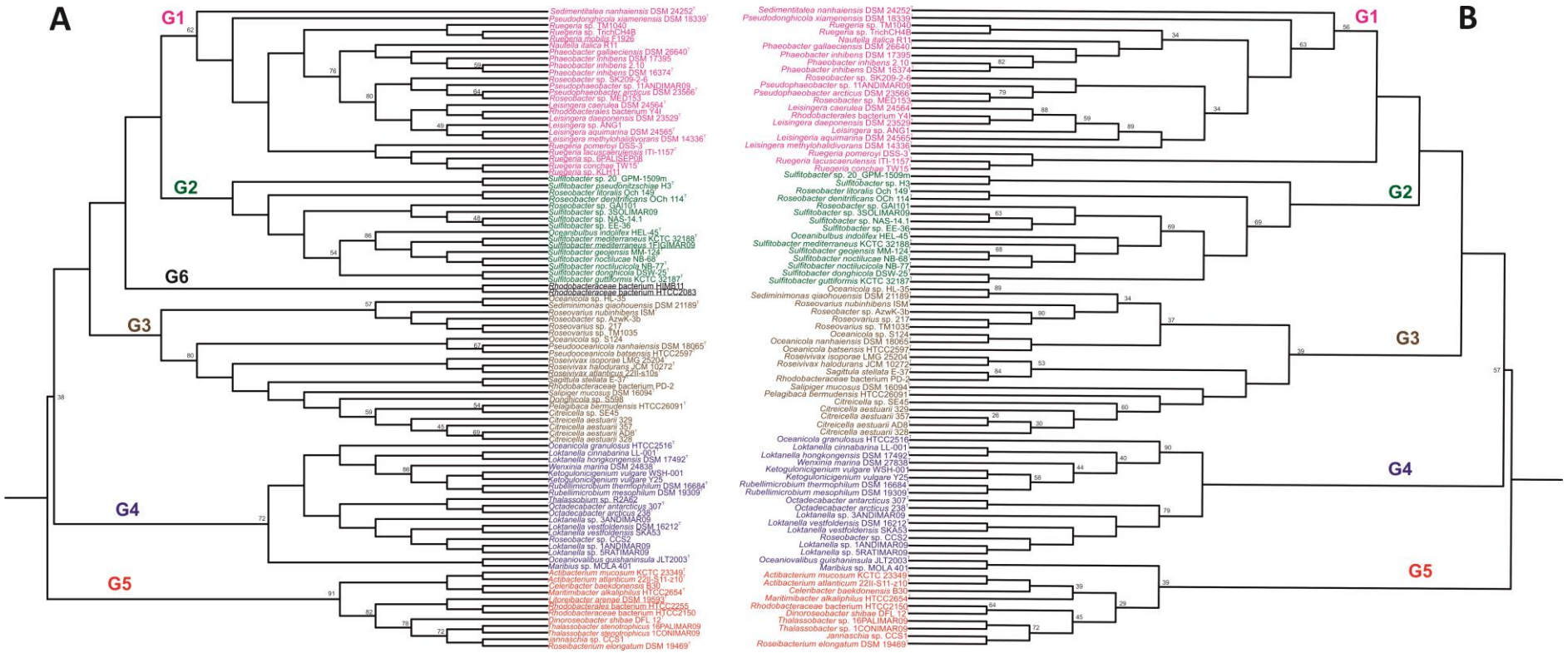

Figure 3.2. Phylogenetic trees of the Roseobacter lineage based on: A) the concatenation of core proteome proteins (114 proteins) of 96 genomes and B) the concatenation of conserved proteins of GTAs (12 proteins) in 85 genomes. Color code indicates phylogenomic groups. GTA genic structure was not detected in the underlined genomes. Bootstrap values higher than 90 are not shown. Both phylogenies were done with PHYLIP (Felsenstein, *et al*., 1989) using parsimony and a bootstrap of 100.

general, physically clustered and the gene order is conserved (Cevallos *et al*., 2008; Thomas, 2000). The RepABC family plasmids are specially common in *Rhizobiales* (Cevallos *et al*., 2008) and are also present in the Roseobacter lineage (Petersen *et al*., 2009, 2012). We searched the *repABC* genic structure in the 96 Roseobacter genomes analyzed in this study (see section 7.2 of Materials and Methods for details). The *repABC* gene structure was detected in 64 Roseobacter genomes. The proportion of detection (two-thirds of the studied genomes) agreed with the previous studies with 30 Roseobacter genomes (Petersen *et al*., 2009). Table 3.3 shows the genomes where *repABC* genes have been found. We analyzed the number of putative plasmids per genome (predicted based on the different *repABC* genes detected), and the number of *repABC* genic clusters present in each plasmid or contig (draft genomes). The table also includes information on the phylogeny of each gene cluster and the compatibility groups defined according to RepC protein phylogeny (see below), and the prediction of presence of type IV secretion systems in the plasmids.

Our results agreed with the previously reported survey of Cevallos and co-workers (2008) in which they analyzed the presence of RepABC modules in 8 members of Roseobacter lineage. The discrepancy was that in our case, several RepABC putative plasmids were found in the same bacterium, such as in i) *Pelagibaca bermudensis* HTCC2601^{T} that in contig AATQ01000004 had two different RepABC modules while in the Cevallos survey only one was found and ii) *Sagittula stellata* E-37^{T} that in contig AAYA01000028 we detected one RepABC module while in the previous survey did not find any. This disagreement was probably due to the use of different search methods, although we could not confirm this because the criteria were not specified in the former study.

The genomes of phylogenomic groups G1, G2 and G3 were the ones in which most RepABC modules were detected (72, 81.3 and 81.8 % of the genomes, respectively). In contrast, only 36.8 % of the genomes of the phylogenomic group G4 contained *repABC* genic structures. In phylogenomic groups G5 and G6, the *repABC* genic structure was detected in approximately half of the genomes (53.3 % and 50 % of genomes, respectively).

The variation in the number of RepABC modules in genomes of the same phylogenomic group was studied (see in Figure 3.3). In genomes of phylogenomic groups G1, G3 and G5, the median number of RepABC modules per genome was 2, while in the phylogenomic group G2 this value was slightly lower (1.5). Phylogenomic groups G1 and G5 were similar in reference to the distribution of the number of RepABC modules per genome having the majority of them between 1 and 3 RepABC modules. Phylogenomic group G3 was the one with more genomes with a higher number of *repABC* modules per genome, mainly due to the *Citreicella* spp. genomes and *Pelagibaca bermudensis* HTCC2601^{T}. Phylogenomic group G4 (not shown in Figure 3.3.) highlighted for the low number of genomes in which RepABC modules was detected and for the low number of RepABC modules detected in each genome (all genomes had only 1 putative plasmid except *Loktanella* sp. 5RATIMAR09 with 2). Two genomes had 5 RepABC modules. They were *Sulfitobacter* sp. NAS-14.1 (G2) and *Citreicella aestuarii* 328 (G3). *Sulfitobacter* sp. NAS-14.1 was previously reported as the Roseobacter genome in which the

Table 3.3. List of Roseobacter genomes in which the *repABC* operon was detected.

Genome	P[a]	Genbank code[b]	RepC protein code	Size (Kb)	Code[c]	Compatibility group
Leisingera aquimarina DSM 24565^T	G1	AXBE01000014	MM12_00050	182	3	C7
		AXBE01000012	MM12_00483	243	1	C2.2
		AXBE01000011	MM12_00688	249	2	C4.1
Leisingera caerulea DSM 24564^T	G1	AXBI01000020	MM31_00552	564	1	C2.1
		AXBI01000019	MM31_00829	246	4	C7
		AXBI01000011	MM31_04473	70	3	C4.2
		AXBI01000001	MM31_05259	271	2	C4.1
Leisingera daeponensis DSM 23529^T	G1	AXBD01000012	MM32_00078	102	2	C1.2
		AXBD01000012	MM32_00093	102	1	C1.1
		AXBD01000007	MM32_00365	102	4	C7
		AXBD01000003	MM32_00627	69	3	C4.2
Leisingera methylohalidivorans DSM 14336^T	G1	CP006774	MM13_04202	221	2	C7
		CP006775	MM13_04341	285	1	C4.1
Leisingera sp. ANG1	G1	AFCF01000001	MM35_04330	25	N	C7
Phaeobacter gallaeciensis DSM 26640^T	G1	**CP006968**	MM34_03938	134	2	C8
		CP006969	MM34_04085	110	1	C7
		CP006973	MM34_04393	40	3	C10
Phaeobacter inhibens DSM 16374^T	G1	**AXBB01000006**	MM37_00128	87	N	C7
Pseudodonghicola xiamenenensis DSM 18339^T	G1	AUBS01000056	MM08_00166	13	2	C4.1
		AUBS01000044	MM08_00380	27	1	C1.2
		AUBS01000018	MM08_01517	87	3	C9
Pseudophaeobacter arcticus DSM 23566^T	G1	**AXBF01000006**	MM30_00127	203	1	C1.1
		AXBF01000005	MM30_00496	229	2	C8
Pseudophaeobacter sp. 11ANDIMAR09	G1	**LIKT01000024**	MM89_03530	75	N	C9
Rhodobacterales bacterium Y4I	G1	ABXF01000062	MM44_00012	10	1	C4.2
		ABXF01000053	MM44_00182	150	2	C7
Roseobacter sp. MED193	G1	**AANB01000013**	MM49_00532	123	2	C8
		AANB01000012	MM49_00705	167	1	C1.1
Ruegeria conchae TW15^T	G1	AEYW01000028	MM63_00019	73	N	C10
Ruegeria mobilis F1926	G1	AQCH01000085	MM65_03002	11	N	C8
Ruegeria sp. 6PALISEP08	G1	LGXZ01000016	MM88_03913	107	N	C10
Ruegeria sp. KLH11	G1	ACCW01000047	MM41_00058	73	2	C7
		ACCW01000009	MM41_03361	65	1	C6
		ACCW01000005	MM41_04059	256	3	C10
Ruegeria sp. TrichCH4B	G1	ACNZ01000047	MM70_03904	46	N	C7
Sedimentitalea nanhaiensis DSM 24252^T	G1	**AXBG01000030**	MM14_00014	227	3	C8
		AXBG01000028	MM14_00241	58	1	C1.1
		AXBG01000023	MM14_01030	123	2	C3.2
Oceanibulbus indolifex HEL-45^T	G2	ABID01000023	MM20_00430	18	2	C7
		ABID01000015	MM20_00611	27	3	C10
		ABID01000008	MM20_00844	56	1	C2.1
Roseobacter denitrificans OCh 114^T	G2	**CP000464**	MM54_03972	106	N	C1.2
Roseobacter litoralis Och 149^T	G2	CP002625	MM55_04475	83	N	C1.2
Roseobacter sp. GAI101	G2	ABXS01000059	MM50_00291	81	N	C4.1
Sulfitobacter geojensis MM-124^T	G2	**JASE01000002**	MM74_03878	194	1	C8
		JASE01000001	MM74_04053	87	2	C9
Sulfitobacter guttiformis KCTC 32187^T	G2	JASG01000001	MM79_03902	53	N	C7
Sulfitobacter mediterraneus 1FIGIMAR09	G2	JEMU01000028	MM81_00107	20	N	C8
Sulfitobacter noctilucicola NB-77^T	G2	**JASD01000006**	MM76_03383	172	1	C1.1
		JASD01000004	MM76_03815	34	2	C7
Sulfitobacter pseudonitschiae H3^T	G2	JAMD01000035	MM77_00061	9	2	C1.2
		JAMD01000020	MM77_00422	57	1	C1.1
		JAMD01000019	MM77_00524	61	3	C2.1
		JAMD01000016	MM77_00667	67	4	C8
Sulfitobacter sp. 20_GPM-1509m	G2	**JIBC01000011**	MM73_00622	166	N	C2.1
Sulfitobacter sp. 3SOLIMAR09	G2	AXZR01000011	MM82_00164	52	N	C7
Sulfitobacter sp. EE-36	G2	AALV01000012	MM71_00054	28	N	C7
Sulfitobacter sp. NAS-14.1	G2	AALZ01000015	MM72_00176	74	1	C2.2
		AALZ01000014	MM72_00284	78	2	C3.2
		AALZ01000013	MM72_00322	84	5	C10
		AALZ01000012	MM72_00447	95	3	C7
		AALZ01000011	MM72_00500	98	4	C8
Citreicella aestuarii 328	G3	LIGK01000017	MM95_01910	69	5	C9
		LIGK01000018	MM95_01956	67	4	C5
		LIGK01000030	MM95_02660	55	2	C1.2
		LIGK01000044	MM95_03239	35	3	C3.1
		LIGK01000082	MM95_04044	16	1	C1.1

Table 3.3. List of Roseobacter genomes in which the *repABC* operon was detected (continued).

Genome	P[a]	Genbank code[b]	RepC protein code	Size (Kb)	Code[c]	Compatibility group
Citreicella aestuarii 329	G3	**LIGL01000015**	MM96_01746	84	2	C3.1
		LIGL01000027	MM96_02523	55	1	C1.2
		LIGL01000031	MM96_02733	50	4	C9
		LIGL01000039	MM96_03032	39	3	C5
Citreicella aestuarii 357	G3	AJKJ01000084	MM05_02279	14	3	C5
		AJKJ01000075	MM05_02567	55	1	C1.2
		AJKJ01000071	MM05_02632	32	4	C9
		AJKJ01000060	MM05_03073	120	2	C3.1
Citreicella aestuarii AD8^T	G3	**LIQE01000008**	MM94_01173	127	2	C1.2
		LIQE01000016	MM94_01904	73	3	C5
		LIQE01000025	MM94_02428	55	1	C1.1
		LIQE01000083	MM94_03989	15	4	C9
Citreicella sp. SE45	G3	ACNW01000121	MM04_00089	109	1	C3.2
		ACNW01000106	MM04_00743	75	2	C9
Oceanicola sp. HL-35	G3	JAFT01000004	MM22_04017	295	N	C3.2
Oceanicola sp. S124	G3	AFPM01000048	MM21_03574	49	N	C4.1
Pelagibaca bermudensis HTCC2601^T	G3	AATQ01000027	MM29_01964	67	3	C6
		AATQ01000009	MM29_03618	136	4	C7
		AATQ01000004	MM29_04297	179	1	C2.1
		AATQ01000004	MM29_04362	179	2	C3.2
Pseudooceanicola batsensis HTCC2597^T	G3	**AAMO01000015**	MM23_00272	87	1	C1.2
		AAMO01000009	MM23_01026	218	3	C6
		AAMO01000007	MM23_01620	285	2	C3.2
Pseudooceanicola nanhaiensis DSM 18065^T	G3	JHZF01000019	MM25_00165	121	1	C4.1
		JHZF01000018	MM25_00248	140	2	C8
Rhodobacteraceae bacterium PD-2	G3	AWRV01000335	MM42_02513	3	1	C1.2
		AWRV01000259	MM42_03020	36	3	C3.2
		AWRV01000134	MM42_03920	60	2	C2.1
Roseivivax atlanticus 22II-s10s^T	G3	AQQW01000012	MM48_00866	124	N	C2.2
Roseivivax halodurans JCM 10272^T	G3	**JALZ01000024**	MM46_01027	39	2	C2.2
		JALZ01000023	MM46_01104	42	1	C2.1
Roseivivax isoporae LMG 25204^T	G3	**JAME01000022**	MM47_01722	92	N	C3.2
Roseovarius nubinhibens ISM^T	G3	AALY01000005	MM57_00088	77	N	C3.2
Roseovarius sp. 217	G3	AAMV01000023	MM56_00189	30	1	C2.1
		AAMV01000021	MM56_00240	35	2	C3.1
		AAMV01000007	MM56_01986	215	3	C3.2
Roseovarius sp. TM1035	G3	**ABCL01000010**	MM58_00338	158	N	C1.2
Sagittula stellata E-37^T	G3	AAYA01000028	MM67_00134	24	2	C6
		AAYA01000007	MM67_02291	239	1	C3.2
Loktanella vestfoldensis DSM 16212^T	G4	ARNL01000015	MM18_03610	18	N	C2.1
Loktanella sp. 1ANDIMAR09	G4	**LIGP01000006**	MM91_03159	203	N	C3.2
Loktanella sp. 3ANDIMAR09	G4	**LJAK01000010**	MM92_02922	129	N	C8
Loktanella sp. 5RATIMAR09	G4	LJAL01000011	MM90_03109	102	2	C3.2
		LJAL01000017	MM90_03532	44	1	C1.1
Oceanicola granulosus HTCC2516^T	G4	AAOT01000006	MM24_02676	147	N	C6
Octadecabacter antarcticus 307^T	G4	CP003741	MM27_05276	63	N	C1.1
Octadecabacter arcticus 238^T	G4	CP003744	MM28_05905	160	N	C1.2
Actibacterium atlanticum 22II-S11-z10^T	G5	**AQQY01000011**	MM01_00375	83	N	C1.1
Celeribacter baekdonensis B30	G5	AMRK01000024	MM03_00054	15	1	C1.2
		AMRK01000021	MM03_00109	27	2	C8
		AMRK01000010	MM03_01047	145	3	C9
Dinoroseobacter shibae DFL 12^T	G5	**CP000831**	MM06_03847	191	1	C1.1
		CP000833	MM06_04125	126	2	C3.2
		CP000834	MM06_04205	86	3	C4.1
Maritimibacter alkaliphilus HTCC2654^T	G5	AAMT01000019	MM19_00707	68	3	C8
		AAMT01000013	MM19_01333	107	2	C3.2
		AAMT01000012	MM19_01427	114	1	C1.2
Rhodobacteraceae bacterium HTCC2150	G5	AAXZ01000006	MM40_00904	116	N	C6
Litoreibacter arenae DSM 19593^T	G5	AONI01000013	MM84_01268	140	N	C1.2
Thalassobacter stenotrophicus 1CONIMAR09	G5	JGVS01000009	MM87_03130	77	1	C3.2
		JGVS01000010	MM87_03171	44	2	C8
Rhodobacteraceae bacterium HTCC2083	G6	ABXE01000002	MM39_04147	17	N	C7

a: phylogenomic group; b: Genbank code that starts with CP are closed genomes; c: enumeration of RepC proteins of each genome; this number was also referred to the phylogeny (see Figure 3.5). Genomes in which only one RepC protein was identified were indicated with letter N.

largest number of different *repABC* genic structures were identified (Cevallos *et al*., 2008; Petersen *et al*., 2009).

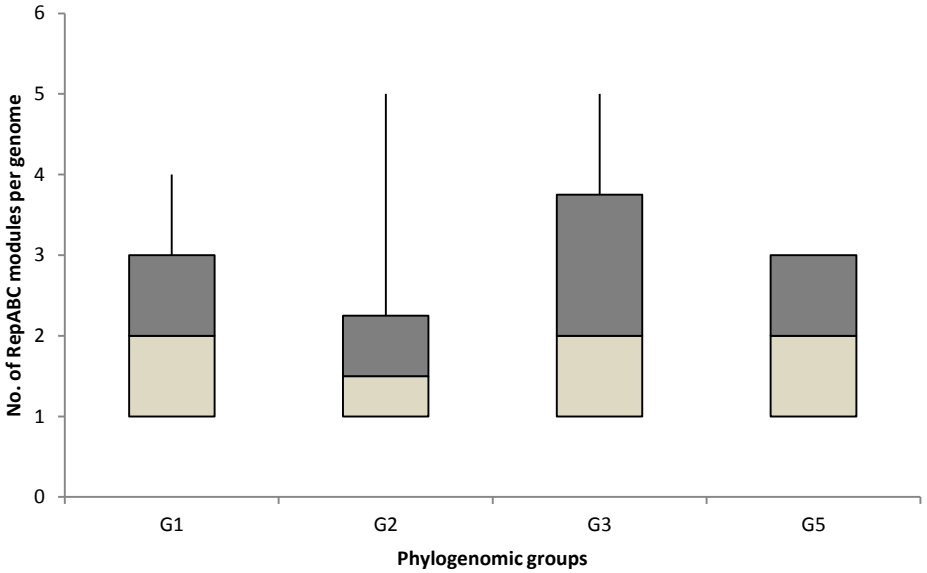

Figure 3.3. Boxplot of the number of RepABC modules per genome in each phylogenomic group (G1:18 genomes, G2:12 genomes, G3: 18 genomes, and G5: 7 genomes). Two phylogenomic groups are not shown in the Figure: G4 had low variability and G6 had only one genome (see Table 3.3). The boxplot represents the percentiles, the median and the extreme values (minimum and maximum). The central box shows the values between the 25 percentile (lower edge of the box) and 75 percentile (upper edge of the box). The middle line of the box represents the median or 50 percentile. The vertical lines show the maximum values.

In complete genomes plasmids are well defined and the size has been determined (see Table 3.3., Genbank codes starting with CP). However, most of the genomes analyzed in this study were in draft status. Therefore we could not determine the size of the putative plasmids detected. In our study, when a *repABC* genic structure was detected in a contig of a draft genome we considered that it was a putative plasmid, and then used contig size as an estimate of minimum plasmid size. According to this, the size of putative plasmids oscillated between 3 Kb and 564 Kb (contig size). In Figure 3.4, the variation of the putative plasmid size per each phylogenomic group is represented. In most cases the sizes of putative plasmids ranged between 40 Kb and 220 Kb. This is within the lower range described in genomes of the Class *Alphaproteobacteria* [47.5 Kb and 2 Mb (Cevallos *et al*., 2008)]. Small putative plasmids (i.e. 3 Kb contigs) might indicate that they were fragments of larger RepABC plasmids. The median of the putative plasmid sizes in genomes of phylogenomic groups G1, G4 and G5 (approximately 100 Kb) was higher than in genomes of phylogenomic groups G2 and G3 (approximately 69 Kb). Phylogenomic group G1 was the one that showed the highest variation in sizes. Moreover, it was the group that had more putative plasmids with high sizes. In contrast, phylogenomic group G2 was the one with the lowest variation in sizes. It was also the group with more putative plasmids with small sizes. Regarding to the minimum and maximum values, phylogenomic groups G1, G2 and G3 had putative plasmids with sizes of 10 Kb or lower and phylogenomics groups G1 and G3 had putative plasmids with sizes of 290 Kb or higher. In fact,

Leisingera caerulea DSM 24564^T (G1) had a putative plasmid size of 564 Kb, the largest of the lineage so far.

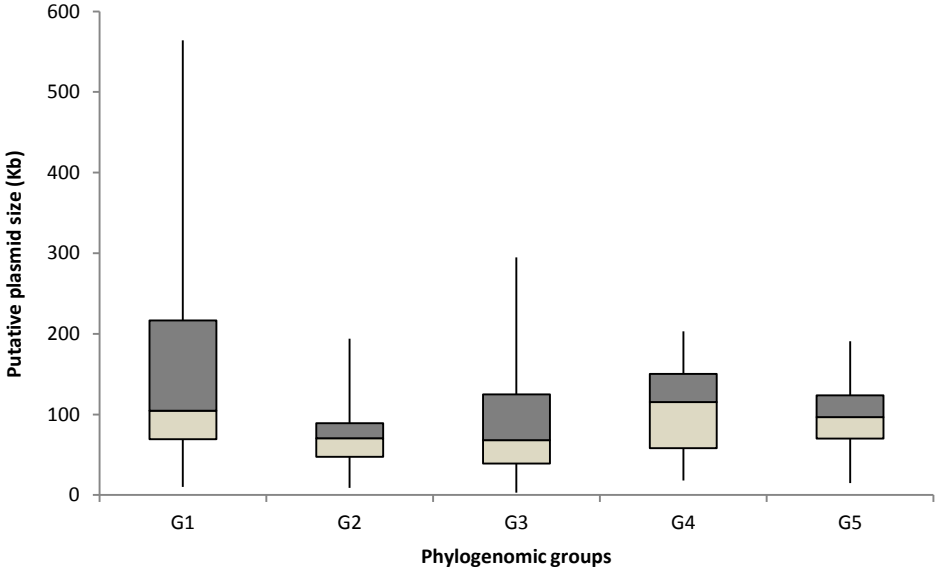

Figure 3.4. Boxplot of sizes of putative RepABC plasmids in phylogenomic groups (G1:18 genomes, G2:12 genomes, G3: 18 genomes, G4: 7 genomes, and G5: 7 genomes). Phylogenomic group G6 was not shown. The boxplot represents the percentiles, the median and the extreme values (minimum and maximum). The central box shows the values between the 25 percentile (lower edge of the box) and 75 percentile (upper edge of the box). The middle line of the box represents the median or 50 percentile. The vertical lines show the minimum and maximum values.

Another interesting aspect was the presence of two different RepABC modules in one putative plasmid (contig). That was the cases of *Leisingera daeponensis* DSM 23529^T and *Pelagibaca bermudensis* HTCC2601^T. This could be an artifact (i.e. genome assembly error) but it could be correct. The fact that one putative plasmid could harbor more than one RepABC module was previously reported in other members of *Alphaproteobacteria*, such as *Rhizobium etli* CFN42, *Sinorhizobium melioti* SM11 (Cevallos *et al*., 2008). This was explained by a co-integration event between an incoming RepABC plasmid and a resident plasmid (Cevallos *et al*., 2008).

We also studied which putative RepABC plasmids could be transferred by conjugation (see Table 3.3). For this, we searched genes of type IV secretion system (T4SS) using as a reference the model of *Agrobacterium tumefaciens* (see section 7.2 of Materials and Methods for details). The VirB4 ATPase is highly conserved in sequence and the only protein with clear-sequence homologs in all known T4SS. For this reason, it is a used as marker for determining the presence of a T4SS (Alvarez-Martinez & Christie, 2009; Guglielmini *et al*., 2013). Other proteins such as pilin protein VirB2, channel proteins VirB6 and VirB8 and the coupling protein VirD4 were also searched to confirm the potential conjugative ability of these plasmids. We observed that between the 23 % and 25 % of the putative RepABC plasmids in genomes of phylogenomic groups G1, G3 and G4 were potentially conjugative, while in the phylogenomic groups G2 and G5 percentages were higher (29 and 35 %, respectively; see Table 3.3 and Table 3S.2 for further details). In the putative plasmid of *Rhodobacteraceae* bacterium HTCC2083

genome (phylogenomic group G6) the type IV secretion system was not found. According to these results the majority of putative RepABC plasmids in roseobacters would not be conjugative plasmids.

The phylogeny of the replicase RepC (Figure 3.5) was established with the objectives of i) determining if there was a common ancestor in agreement with the phylogeny of the lineage (like GTAs) and, ii) defining the compatibility groups of these putative plasmids, since RepC was described as an incompatibility factor (Cervantes-Rivera *et al*., 2011). In total, 129 RepC replicase sequences with 708 amino acids aligned positions were used for the phylogeny. The topology of the RepC tree (see Figure 3.5) disagreed with the phylogeny obtained with the core proteome (see Figure 1.1). The lack of evolutionary congruence in RepC phylogeny agreed with the results obtained in previous studies (Castillo-Ramírez *et al*., 2009; Petersen *et al*., 2009). In addition, we compared the amino acid sequence identity of RepA, RepB and RepC. We considered that the same partioning, segregation and replication system (RepABC module) was shared among roseobacters if the percentage of identity was at least 95 % in 100 % of the amino acid sequence of these three proteins. This is indicated in Figure 3.5 with black boxes. In some cases, RepABC amino acid sequences in genomes that belonged to different phylogenomic groups shared almost exactly the same partioning, segregation and replication structure, such as *Pseudophaeobacter arcticus* DSM 23566^T (G1) and *Sulfitobacter pseudonitzschiae* H3^T (G2) or *Pseudodonghicola xiamenensis* DSM 18339^T (G1) and *Celeribacter baekdonensis* B30 (G5) (see Figure 3.5). These results indicated that bacterial strains (even from different phylogenomic groups) could have freely exchanged this *repABC* genic structure in the past. Thus, RepABC putative plasmids seem to have been transferred promiscuously among roseobacters.

On the other hand, we identified ten different compatibility groups (from C1 to C10) based in the phylogeny made with 129 RepC of 64 Roseobacter genomes (see Figure 3.5). This was one additional compatility group than those described in the analysis of RepC with 30 Roseobacter genomes (Petersen *et al*., 2009). We assumed that coexisting replication genes were compatible if their sequences belonged to different compatibility groups (C1-C10). In general, in genomes with several RepABC putative plasmids, RepC phylogeny showed that each RepC was from a different compatibility group. That was the case of *Sulfitobacter* sp. NAS-14.1 for example. However, there were some exceptions in which a genome had several putative replicons assigned to the same compatibility group using RepC phylogeny (they were indicated with a colored pentagon in the tree): i) *Leisingera daeponensis* DSM 23529^T (C1); ii) *Sulfitobacter pseudonitzschiae* H3^T (C1); iii) *Citreicella aestuarii* 328 (C1); iv) *Roseivivax halodurans* JCM 10272^T (C2); v) *Roseovarius* sp. 217 (C3), and vi) *Leisingera caerulea* DSM 24564^T (C4). For that reason, in these cases we defined sub-compatibility groups based on the phylogenies of RepA and/or RepB, since it has been shown that partition can also have a role in plasmid compatibility (Pérez-Oseguera & Cevallos, 2013; Petersen *et al*., 2009; Soberón *et al*., 2004). With RepA and RepB phylogenies these isolates belonged to different sub-compatibility groups (see supplementary Figures 3S.2 and 3S.3 respectively) with the exception of *Roseivivax*

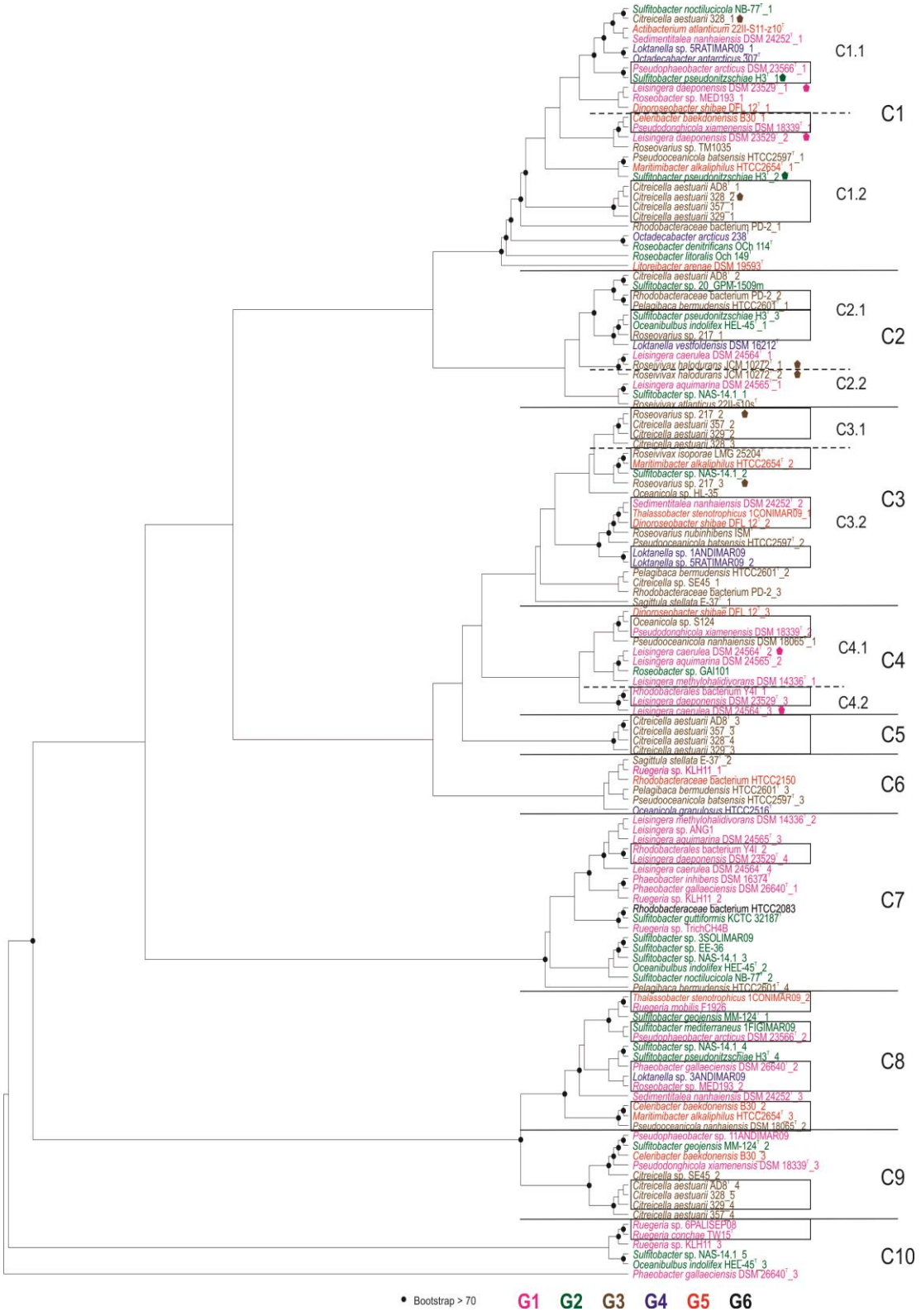

Figure 3.5. Phylogenetic tree of RepC replicase. Color code indicates phylogenomic groups. The number after bacterial names indicates each RepC protein (code column Table 3.3). The boxes indicate the genomes that shared the same amino acid sequence of RepA, RepB and RepC (95 % of identity in 100 % of sequence). Solid lines delimitate the compatibility groups and dashed lines separate the sub-compatibility groups. Pentagon symbols indicate those RepC from the same genome located in different sub-compatibility groups. The phylogeny was done with PHYLIP (Felsenstein, *et al*., 1989) using parsimony and a bootstrap of 100.

halodurans JCM 10272^T. In this case, the incompatibility factor was due to RepA, and not to both proteins (see supplementary Figures 3S.2 and 3S.3).

We also analyzed how the compatibility groups were distributed in each phylogenomic group (see Table 3.4). Genomes of phylogenomic groups G1 and G3 had representatives of 9 different compatibility groups (with the exception of C5 in G1 and C10 in G3). On the other hand, phylogenomic groups G4 and G5 had replicons that belonged to less compatibility groups (they only had representatives of 5 and 6 compatibility groups, respectively). These results suggested that members of phylogenomic groups G1 and G3 had higher genomic versatility and plasticity for recruiting and maintaining a greater diversity of RepABC plasmids than phylogenomic groups G4 and G5. Compatibility group C5 was only constituted by replicons of members of phylogenomic group G3, specifically by *Citreicella aestuarii* strains. In this case, it could be a putative type of RepABC plasmid that was specific of this species.

Table 3.4. Distribution of compatibility groups in the different phylogenomic groups.

	Phylogenomic group					
Compatibility group	**G1**	**G2**	**G6**	**G3**	**G4**	**G5**
C1	6	5	0	9	3	5
C2	2	4	0	6	1	0
C3	1	1	0	13	2	3
C4	7	1	0	2	0	1
C5	0	0	0	4	0	0
C6	1	0	0	3	1	1
C7	10	6	1	1	0	0
C8	5	4	0	1	1	3
C9	2	1	0	5	0	1
C10	4	2	0	0	0	0
Total	38	24	1	44	8	14

Apart from the RepABC type plasmids, other replicons have been described (Petersen *et al*., 2011, 2013). These replication initiator genes were RepA, RepB and DnaA-like. These replication initiator genes were usually located adjacent to partioning genes *parA* and *parB*. With the objective to cover all possible replication systems of ECEs, we searched through the 96 roseobacters if they harbored any of these replicases (see in section 7.2 of Materials and Methods). We considered that the presence of any of the replicases in a contig indicated a plausible ECE. In Table 3.5, we summarized the number of total plausible ECEs that we found in the 96 studied roseobacters (for further details see Table 3S.3). Apart from the four types of replicons described above (Petersen *et al*., 2011, 2013), sometimes we found the replicase RepC, but not associated with the partition proteins RepAB as in RepABC plasmids. All different replication types were detected in 19 roseobacters (underlined in Table 3.5): eight genomes belonged to phylogenomic group G1, three from G2, seven from G3, and one from G4. In 9 genomes we did not find any of the replicases: two from G3 (*Roseobacter* sp. AzwK-3b and *Sediminimonas qiaohouensis* DSM 21189^T), three from G4 (*Loktanella vestfoldensis* DSM 16212^T, *Roseobacter* sp. CCS2 and *Rubellimicrobium thermophilum* DSM 16684^T), three from G5 (*Actibacterium mucosum* KCTC 23349^T, *Rhodobacterales* bacterium HTCC2255 and *Roseibacterium elongatum* DSM 19469^T), and one from G6 (*Rhodobacteraceae* bacterium

Table 3.5. Total number of plausible extrachromosomal elements that were detected in the 96 Roseobacter genomes according to the different replicases.

Genome	P[a]	RepABC	RepA	RepB	RepC	DnaA-like	Total ECEs
Leisingera aquimarina DSM 24565^T	G1	3	2	1	1	1	8
Leisingera caerulea DSM 24564^T	G1	4	1	1	0	1	7
Leisingera daeponensis DSM 23529^T	G1	4	1	1	0	1	6
Leisingera methylohalidivorans DSM 14336^T	G1	2	0	1[b]	0	0	2
Leisingera sp. ANG1	G1	1	1	1	1	1	5
Nautella italica R11	G1	0	0	0	0	1	1
Phaeobacter gallaeciensis DSM 26640^T	G1	3	1	1	1	1	7
Phaeobacter inhibens 2.10	G1	0	1	1	0	1	3
Phaeobacter inhibens DSM 16374^T	G1	1	1	1	0	1	4
Phaeobacter inhibens DSM 17395	G1	0	1	1	0	1	3
Pseudodonghicola xiamenensis DSM 18339^T	G1	3	2	0	0	0	5
Pseudophaeobacter arcticus DSM 23566^T	G1	2	3[c,d]	2[c]	0	0	5
Pseudophaeobacter sp. 11ANDIMAR09	G1	1	1	1	0	1	4
Rhodobacterales bacterium Y4I	G1	2	0	0	0	1	3
Roseobacter sp. MED193	G1	2	0	1	0	0	3
Roseobacter sp. SK209-2-6	G1	0	1	1	0	1	3
Ruegeria conchae TW15^T	G1	1	0	0	0	0	1
Ruegeria lacuscaerulensis ITI-1157^T	G1	0	0	1	0	0	1
Ruegeria mobilis F1926	G1	1	0	1	1	0	3
Ruegeria pomeroyi DSS-3^T	G1	0	0	0	0	1	1
Ruegeria sp. 6PALISEP08	G1	1	1	0	0	1	3
Ruegeria sp. KLH11	G1	3[d]	2[c,d]	1[c]	1[d]	0	3
Ruegeria sp. TM1040	G1	0	1	1	0	0	2
Ruegeria sp. TrichCH4B	G1	1	0	1	1	0	3
Sedimentitalea nanhaiensis DSM 24252^T	G1	3	6[c]	1	0	1	9
Oceanibulbus indolifex HEL-45^T	G2	3	0	0	0	2	5
Roseobacter denitrificans OCh 114^T	G2	1	3	0	0	0	4
Roseobacter litoralis Och 149^T	G2	1	1	0	0	1	3
Roseobacter sp. GAI101	G2	1	0	1	0	0	2
Sulfitobacter donghicola DSW-25^T	G2	0	2	5	0	0	7
Sulfitobacter geojensis MM-124^T	G2	2[d]	1	2[c]	0	1	5
Sulfitobacter guttiformis KCTC 32187^T	G2	1	0	2	0	0	3
Sulfitobacter mediterraneus 1FIGIMAR09	G2	1	1	2	0	0	4
Sulfitobacter mediterraneus KCTC 32188^T	G2	0	0	2	0	0	2
Sulfitobacter noctilucae NB-68^T	G2	0	2[d]	1	1[d]	1	4
Sulfitobacter noctilucicola NB-77^T	G2	2	1	1	1	1	6
Sulfitobacter pseudonitzschiae H3^T	G2	4	1	1	0	2	8
Sulfitobacter sp. 20_GPM-1509m	G2	1	0	2	1	2	6
Sulfitobacter sp. 3SOLIMAR09	G2	1	0	0	0	1	2
Sulfitobacter sp. EE-36	G2	1	0	0	0	1	2
Sulfitobacter sp. NAS-14.1	G2	5	3[d]	0	1[d]	1	9
Citreicella aestuarii 328	G3	5	4	2	1	1	13
Citreicella aestuarii 329	G3	4	3	1	0	1	9
Citreicella aestuarii 357	G3	4	3	2	1	1	11
Citreicella aestuarii AD8^T	G3	4[d]	5	1	1[d]	1	11
Citreicella sp. SE45	G3	2	1	1	1	1	6
Donghicola sp. S598	G3	0	1	0	3	2	6
Oceanicola sp. HL-35	G3	1	4[c,d]	3[c]	0	1	4
Oceanicola sp. S124	G3	1	1	0	0	0	2
Pelagibaca bermudensis HTCC2601^T	G3	4[c]	0	0	1	1	5
Pseudooceanicola batsensis HTCC2597^T	G3	3	1	0	1	1	6
Pseudooceanicola nanhaiensis DSM 18065^T	G3	2	2	1	1	1	7
Rhodobacteraceae bacterium PD-2	G3	3	1	0	0	2	6
Roseivivax atlanticus 22II-s10s^T	G3	1	2	0	1	0	4
Roseivivax halodurans JCM 10272^T	G3	2[d]	1[d]	0	0	0	2
Roseivivax isoporae LMG 25204^T	G3	1	1	0	0	2	4
Roseobacter sp. AzwK-3b	G3	0	0	0	0	0	0
Roseovarius nubinhibens ISM^T	G3	1	0	0	0	0	1
Roseovarius sp. 217	G3	3	0	0	1	0	4
Roseovarius sp. TM1035	G3	1	0	0	0	0	1
Sagittula stellata E-37^T	G3	2	0	1	1	0	4
Salipiger mucosus DSM 16094^T	G3	0	3	2	0	0	5
Sediminimonas qiaohouensis DSM 21189^T	G3	0	0	0	0	0	0
Ketogulonicigenium vulgare WSH-001	G4	0	1[d]	1[d]	0	1	2
Ketogulonicigenium vulgare Y25	G4	0	1[d]	1[d]	0	1	2

Table 3.5. Total number of plausible extrachromosomal elements that were detected in the 96 Roseobacter genomes according to the different replicases (continued).

Genome	P[a]	RepABC	RepA	RepB	RepC	DnaA-like	Total ECEs
Loktanella cinnabarina LL-001^T	G4	0	3[c]	3[c]	0	1	5
Loktanella hongkongensis DSM 17492^T	G4	0	1	1	0	1	3
Loktanella sp. 1ANDIMAR09	G4	1	3[d]	0	1[d]	0	4
Loktanella sp. 3ANDIMAR09	G4	1	2	1	0	0	4
Loktanella sp. 5RATIMAR09	G4	2	0	0	0	0	2
Loktanella vestfoldensis DSM 16212^T	G4	1	0	0	0	0	1
Loktanella vestfoldensis SKA53	G4	0	0	0	0	0	0
Maribius sp. MOLA 401	G4	0	1	0	0	0	1
Oceanicola granulosus HTCC2516^T	G4	1	1	0	0	1	3
Oceaniovalibus guishaninsula JLT2003^T	G4	0	1	1	1	1	4
Octadecabacter antarcticus 307^T	G4	1[c]	0	0	0	0	1
Octadecabacter arcticus 238^T	G4	1[c]	1[d]	3[b,c,d]	0	0	2
Roseobacter sp. CCS2	G4	0	0	0	0	0	0
Rubellimicrobium mesophilum DSM 19309^T	G4	0	0	1	0	0	1
Rubellimicrobium thermophilum DSM 16684^T	G4	0	0	0	0	0	0
Thalassobium sp. R2A62	G4	0	1	0	0	0	1
Wenxinia marina DSM 24838^T	G4	0	1	0	0	0	1
Actibacterium atlanticum 22II-S11-z10^T	G5	1	0	0	0	0	1
Actibacterium mucosum KCTC 23349^T	G5	0	0	0	0	0	0
Celeribacter baekdonensis B30	G5	3	1	0	0	0	4
Dinoroseobacter shibae DFL 12^T	G5	3[c]	1	1	0	0	5
Jannaschia sp. CCS1	G5	0	1	0	0	0	1
Litoreibacter arenae DSM 19593^T	G5	1	0	0	0	0	1
Maritimibacter alkaliphilus HTCC2654^T	G5	3	0	1	0	1	5
Rhodobacteraceae bacterium HTCC2150	G5	1[d]	1	1[d]	0	0	2
Rhodobacterales bacterium HTCC2255	G5	0	0	0	0	0	0
Roseibacterium elongatum DSM 19469^T	G5	0	0	0	0	0	0
Thalassobacter stenotrophicus 16PALIMAR09	G5	0	1	0	1	0	2
Thalassobacter stenotrophicus 1CONIMAR09	G5	2	1	0	0	0	3
Rhodobacteraceae bacterium HIMB11	G6	0	0	0	0	0	0
Rhodobacteraceae bacterium HTCC2083	G6	1	2[c]	0	0	0	2

a: phylogenomic group, b: the replicase was located in the chromosome, c: two replicases of the same type were located in the same contig/plasmid, d: two different kind of replicases were located in the same contig/plasmid. Underlined genomes: 4 different types of replicases were detected.

HIMB11). In phylogenomic groups G1, G2, G3 and G5 the RepC replicase from RepABC family plasmids was the most detected and in phylogenomic groups G2 and G4 the most detected replicase was RepA. The least detected replicases in all phylogenomic groups were RepC (without the clustering with the partioning genes *repAB*) and DnaA-like.

In closed genomes the replicases were always found in plasmids except in two cases where the replicase was found in the chromosome (RepB in *Leisingera methylohalidivorans* DSM 14336^T and *Octadecabacter arcticus* 238^T). Besides, in *Octadecabacter arcticus* 238^T different replicases (one RepA and two RepB) were present in a single plasmid (CP003743) (for further details check Table 3S.3). In reference to the genomes that were in draft we usually found a replicase per contig. However, there were cases in which 2 or more replicases, of the same or different type (indicated with "c" and "d" superscript in Table 3.5, respectively), were detected in the same contig. In those occassions we considered that these contigs were only 1 ECE (see Table 3.5 and 3S.3 for further details). These findings (i.e. replicases in the chromosome and different replicases in the same plasmid or contig could be due to i) mistakes in genome assembly and annotation of genomes or/and ii) genetic rearrangements between the chromosome and the plasmids or between plasmids.

In this study we have considered that the detection of a replicase indicated the presence of an ECE. Therefore, if several replicases from a single plasmid were in different contigs they were counted separately. Moreover, we only looked for the presence of partition genes in the case of RepABC plasmids but not in the analysis of the other replicases. These two limitations of the study might lead to overestimation of the real number of putative plasmids in these genomes. In Figure 3.6 we show the distribution of the number of ECEs per phylogenomic group. Phylogenomic group G3 was the one whose genomes had the higher median of ECEs elements (4.5 ECEs for genome) and the group that had the genomes with the highest number of ECEs). On the contrary, phylogenomic groups G4 and G5 were the groups whose genomes had a lower median of ECEs per genome (2 and 1.5 respectively). All genomes of phylogenomic groups G1 and G2 had, at least, 1 ECEs.

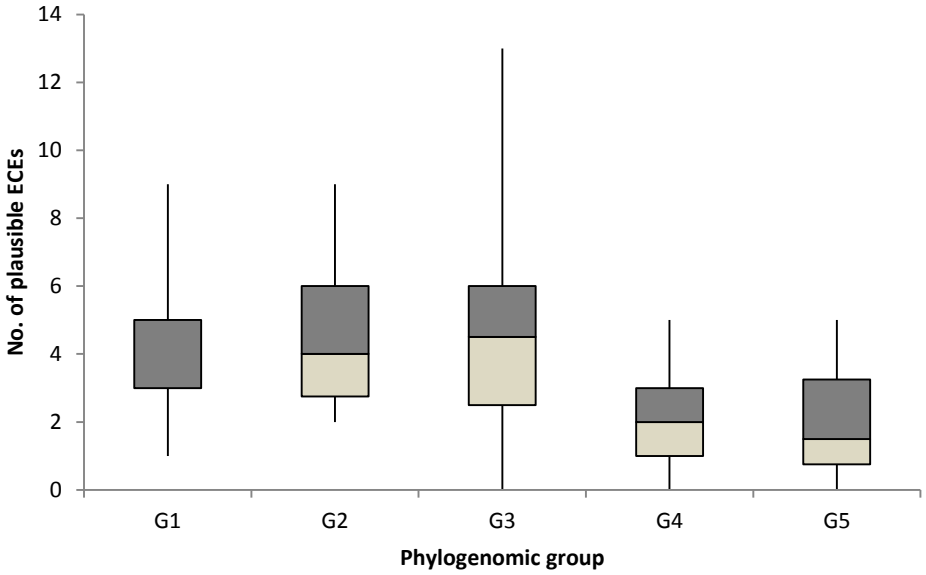

Figure 3.6. Distribution of ECEs in the different phylogenomic groups. The boxplot represents percentiles, the median and the extreme values (minimum and maximum). The central box shows the values between the 25 percentile (lower edge of the box) and 75 percentile (upper edge of the box). The middle line of the box represents the median or 50 percentile. The vertical lines show the minimum and maximum values.

3. Transposase fluxes

Insertion Sequences (ISs) are widespread and can be found in high numbers in prokaryotic genomes (Siguier *et al*., 2014). There were no previous surveys that studied and analyzed the fluxes of transposases among roseobacters as a way to correlate phylogenetic distance or habitat sharing with transposase fluxes. Our interest in analyzing transposases was due to i) previous evidences of correlation between habitat sharing and transposase connections (Hooper *et al*., 2009; Kloesges *et al*., 2011), ii) evidences that transposition of catabolic clusters might be one of the mechanisms whereby degradation genes or operons can be recruited (Bosch *et al*., 1999) and iii) we had a collection of 9 Roseobacter strains isolated at the same time from similar environments (Mallorca Coast) and related to chronically-polluted

environments (8 of them). Then, our hypothesis was that transposase fluxes between these isolates will be higher than those between other roseobacters.

We searched for transposases (TnpAs) in the 96 Roseobacter genomes (see section 7.3 of Materials and Methods for details). Plausible transposases were detected in all genomes except in *Rhodobacterales* bacterium HTCC2255 (see Table 3.6). This is another characteristic that together with its streamlined genome (2.28 Mb) and low G+C content mol % content (37 %) (Luo *et al*., 2013) differenciates strain HTCC2255 from other cultured members of Roseobacter lineage. In total, 6,884 plausible TnpA were detected in the other 95 genomes (see Table 3.6). These TnpAs were classified in IS families and they were clustered in COGs of TnpAs under the criteria of 90 % of identity in, at least, 95 % of the sequence (see Table 3.6 and the text below).

The distribution of the number of TnpAs per phylogenomic group was calculated (see Figure 3.7). Phylogenomic group G6 (only two genomes) and the genomes with more than 300 TnpAs (*Octadecabacter* strains from phylogenomic group G4) were not included in this analysis. In phylogenomic groups G1 and G3 there were more genomes that had high number of TnpAs than in G2, G4 and G5. Fifty percent of the genomes of phylogenomic groups G1 and G3 harbored between 24 and 87 TnpAs, while in G2, G4 and G5, half of genomes harbored between 16 and 67. Phylogenomic group G3 had the highest median of number of TnpAs per genome (71). In addition, it was the phylogenomic group that, together with G1, had the highest variation in number of TnpAs per genome. There were genomes with low number of TnpAs such as *Nautella italica* R11 (G1), *Oceanicola* sp. S124 (G3), *Oceaniovalibus guishaninsula* JLT2003^T (G4), *Roseobacter* sp. CCS2 (G5) and *Actibacterium mucosum* KCTC 23349^T (G5) (see Table 3.6). On the contrary, *Leisingera aquimarina* DSM 24565^T (G1) and *Maribius* sp. MOLA 401 (G4) stood out for having large number of TnpAs in its genomes (248 and 249, respectively, see Table 3.6). Even more stricking was the result of the two *Octadecabacter* genomes of phylogenomic group G4 that had the largest number of TnpAs per genome of the roseobacters included in this survey (see Table 3.6). While the median percentage of CDSs per genome codifying for TnpAs was 0.9 % in roseobacters, in the case of *Octadecabacter antarcticus* 307^T and *O. arcticus* 238^T, these percentages were approximately 13 and 22 %, respectively. This distribution of TnpAs regarding to *Octadecabacter* strains in comparison to other roseobacters was previously mentioned in other studies (Newton *et al*., 2010; Vollmers *et al*., 2013).

We also wanted to determine if there was an association between size of each Roseobacter genome and the number TnpAs that it codified. Touchon & Rocha (2007) suggested that larger genomes are likely to contain more mobile genetic elements than smaller genomes. Our results showed that the number of detected TnpAs in genomes of the Roseobacter lineage was related to genome size (see Figure 3.8). There was a significant positive correlation (Pearson correlation coefficient= 0.35, $p < 0.001$) between the number of TnpAs per genome and genome size. The correlation was higher (0.51, $p < 0.001$) if the genomes with more than 200

Table 3.6. List of predicted transposases in 96 Roseobacter genomes.

Genome	P[a]	Genome size (Mb)	No. TnpAs	No. IS families	No. TnpAs COGs	No. shared TnpA COGs[b]
Leisingera aquimarina DSM 24565^T	G1	5.3	248	16	179	14 (19)
Leisingera caerulea DSM 24564^T	G1	5.3	75	14	59	7 (23)
Leisingera daeponensis DSM 23529^T	G1	4.6	40	7	39	15 (18)
Leisingera methylohalidivorans DSM 14336^T	G1	4.7	166	15	141	14 (31)
Leisingera sp. ANG1	G1	4.6	11	4	11	6 (14)
Nautella italica R11	G1	3.8	6	2	5	1 (1)
Phaeobacter gallaeciensis DSM 26640^T	G1	4.5	29	9	28	8 (32)
Phaeobacter inhibens 2.10	G1	4.2	22	7	14	5 (11)
Phaeobacter inhibens DSM 16374^T	G1	4.1	29	6	17	11 (28)
Phaeobacter inhibens DSM 17395	G1	4.2	19	6	16	5 (6)
Pseudodonghicola xiamenensis DSM 18339^T	G1	4.7	88	12	70	17 (28)
Pseudophaeobacter arcticus DSM 23566^T	G1	5.0	75	15	59	21 (46)
Pseudophaeobacter sp. 11ANDIMAR09	G1	4.7	64	12	64	16 (20)
Rhodobacterales bacterium Y4I	G1	4.3	43	8	19	10 (14)
Roseobacter sp. MED193	G1	4.7	63	14	39	16 (41)
Roseobacter sp. SK209-2-6	G1	4.6	68	9	41	3 (3)
Ruegeria conchae TW15^T	G1	4.5	34	7	32	2 (3)
Ruegeria lacuscaerulensis ITI-1157^T	G1	3.5	33	7	28	2 (3)
Ruegeria mobilis F1926	G1	4.6	24	4	24	6 (4)
Ruegeria pomeroyi DSS-3^T	G1	4.6	35	11	25	3 (3)
Ruegeria sp. 6PALISEP08	G1	4.5	20	8	20	2 (12)
Ruegeria sp. KLH11	G1	4.5	140	13	103	12 (23)
Ruegeria sp. TM1040	G1	4.2	17	3	13	4 (11)
Ruegeria sp. TrichCH4B	G1	4.7	63	9	53	12 (16)
Sedimentitalea nanhaiensis DSM 24252^T	G1	4.9	134	13	91	13 (28)
Oceanibulbus indolifex HEL-45^T	G2	4.1	51	12	49	7 (13)
Roseobacter denitrificans OCh 114^T	G2	4.3	34	9	26	6 (4)
Roseobacter litoralis Och 149^T	G2	4.7	98	12	63	9 (21)
Roseobacter sp. GAI101	G2	4.5	68	11	48	9 (19)
Sulfitobacter donghicola DSW-25^T	G2	3.5	31	11	27	6 (13)
Sulfitobacter geojensis MM-124^T	G2	3.9	25	7	24	6 (16)
Sulfitobacter guttiformis KCTC 32187^T	G2	4.0	24	9	24	3 (5)
Sulfitobacter mediterraneus 1FIGIMAR09	G2	3.9	14	7	13	8 (18)
Sulfitobacter mediterraneus KCTC 32188^T	G2	4.1	32	6	14	5 (7)
Sulfitobacter noctilucae NB-68^T	G2	3.9	28	7	20	3 (14)
Sulfitobacter noctilucicola NB-77^T	G2	4.1	24	11	24	6 (22)
Sulfitobacter pseudonitzschiae H3^T	G2	4.9	33	13	32	13 (19)
Sulfitobacter sp. 20_GPM-1509m	G2	4.7	27	11	25	7 (8)
Sulfitobacter sp. 3SOLIMAR09	G2	3.5	25	5	25	8(11)
Sulfitobacter sp. EE-36	G2	3.5	44	8	36	14 (29)
Sulfitobacter sp. NAS-14.1	G2	4.0	104	13	78	17 (25)
Citreicella aestuarii 328	G3	4.7	74	13	71	61 (31)
Citreicella aestuarii 329	G3	4.5	86	14	86	57 (25)
Citreicella aestuarii 357	G3	4.6	125	11	114	59 (33)
Citreicella aestuarii AD8^T	G3	4.6	82	13	81	62 (28)
Citreicella sp. SE45	G3	5.5	77	9	66	7 (16)
Donghicola sp. S598	G3	3.2	62	14	61	13 (38)
Oceanicola sp. HL-35	G3	4.3	123	10	57	12 (25)
Oceanicola sp. S124	G3	4.7	7	7	7	2 (1)
Pelagibaca bermudensis HTCC2601^T	G3	5.4	88	13	61	14 (27)
Pseudooceanicola batsensis HTCC2597^T	G3	4.4	58	12	44	12 (21)
Pseudooceanicola nanhaiensis DSM 18065^T	G3	4.7	44	9	34	12 (30)
Rhodobacteraceae bacterium PD-2	G3	5.0	43	12	42	3 (7)
Roseivivax atlanticus 22II-s10s^T	G3	4.6	42	10	42	6 (11)
Roseivivax halodurans JCM 10272^T	G3	4.5	69	15	69	8 (11)
Roseivivax isoporae LMG 25204^T	G3	4.9	20	9	20	3 (6)
Roseobacter sp. AzwK-3b	G3	4.2	123	12	66	15 (20)
Roseovarius nubinhibens ISM^T	G3	3.7	7	4	8	2 (7)
Roseovarius sp. 217	G3	4.8	98	16	74	6 (25)
Roseovarius sp. TM1035	G3	4.2	29	9	21	5 (18)
Sagittula stellata E-37^T	G3	5.3	88	9	48	11 (17)
Salipiger mucosus DSM 16094^T	G3	5.8	84	15	62	13 (24)
Sediminimonas qiaohouensis DSM 21189^T	G3	3.6	61	9	37	5 (11)
Ketogulonicigenium vulgare WSH-001	G4	3.3	15	3	9	8 (2)
Ketogulonicigenium vulgare Y25	G4	3.3	15	3	10	8 (2)
Loktanella cinnabarina LL-001^T	G4	3.9	60	10	60	6 (11)
Loktanella hongkongensis DSM 17492^T	G4	3.2	40	7	31	2 (3)

Table 3.6. List of predicted transposases in 96 Roseobacter genomes (continued).

Genome	P[a]	Genome size (Mb)	No. TnpAs	No. IS families	No. TnpAs COGs	No. shared TnpA COGs[b]
Loktanella sp. 1ANDIMAR09	G4	3.7	13	7	12	6 (19)
Loktanella sp. 3ANDIMAR09	G4	3.7	23	7	22	4 (18)
Loktanella sp. 5RATIMAR09	G4	3.7	16	9	15	5 (13)
Loktanella vestfoldensis DSM 16212^T	G4	3.7	30	9	24	6 (9)
Loktanella vestfoldensis SKA53	G4	3.1	17	8	16	3 (6)
Maribius sp. MOLA 401	G4	3.9	249	11	120	9 (28)
Oceanicola granulosus HTCC2516^T	G4	4.0	22	5	21	4 (4)
Oceaniovalibus guishaninsula JLT2003^T	G4	3.0	3	1	3	0 (0)
Octadecabacter antarcticus 307^T	G4	4.9	705	18	366	22 (26)
Octadecabacter arcticus 238^T	G4	5.5	1289	23	577	15 (15)
Roseobacter sp. CCS2	G4	3.5	3	3	3	0 (0)
Rubellimicrobium mesophilum DSM 19309^T	G4	4.9	98	11	80	1 (1)
Rubellimicrobium thermophilum DSM 16684^T	G4	3.2	21	5	20	3 (6)
Thalassobium sp. R2A62	G4	3.5	77	7	56	10 (20)
Wenxinia marina DSM 24838^T	G4	4.2	43	10	30	1 (1)
Actibacterium atlanticum 22II-S11-z10^T	G5	3.2	9	4	9	3 (10)
Actibacterium mucosum KCTC 23349^T	G5	3.7	6	3	6	0 (0)
Celeribacter baekdonensis B30	G5	4.3	39	9	38	5 (17)
Dinoroseobacter shibae DFL 12^T	G5	4.4	87	13	51	11 (30)
Jannaschia sp. CCS1	G5	4.4	19	5	11	0 (0)
Litoreibacter arenae DSM 19593^T	G5	3.7	30	8	30	6 (18)
Maritimibacter alkaliphilus HTCC2654^T	G5	4.5	27	9	22	9 (21)
Rhodobacteraceae bacterium HTCC2150	G5	3.6	46	6	23	8 (16)
Rhodobacterales bacterium HTCC2255	G5	2.3	0	0	0	0 (0)
Roseibacterium elongatum DSM 19469^T	G5	3.6	25	4	12	2 (7)
Thalassobacter stenotrophicus 16PALIMAR09	G5	3.5	20	10	20	12 (28)
Thalassobacter stenotrophicus 1CONIMAR09	G5	3.4	23	9	23	11 (28)
Rhodobacteraceae bacterium HIMB11	G6	3.1	10	4	9	0 (0)
Rhodobacteraceae bacterium HTCC2083	G6	4.0	107	12	45	5 (9)

a: phylogenomic group, b: number between brackets indicates with how many genomes that genome shared TnpA COGs.

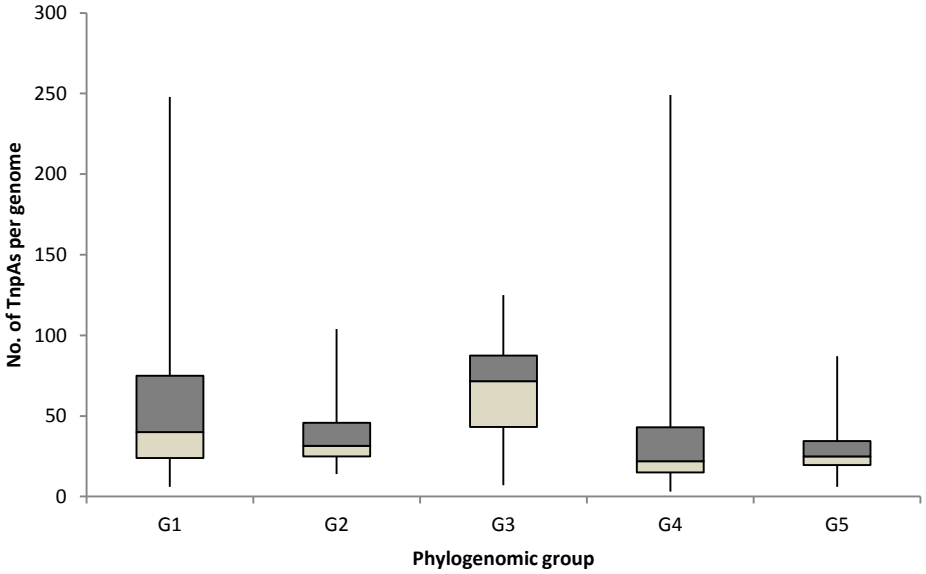

Figure 3.7. Distribution of number of TnpAs per genome for each phylogenomic group. Data from genomes in group G6, and the two *Octadecabacter* genomes (G4) were not included in the analysis. The boxplot represents the percentiles, the median and the extreme values (minimum and maximum). The central box shows the values between the 25 percentile (lower edge of the box) and 75 percentile (upper edge of the box). The middle line of the box represents the median or 50 percentile. The vertical lines show the minimum and maximum values.

TnpAs were excluded from the analysis (*Octadecabacter* spp., *Maribius* sp. MOLA 401 and *Leisingera aquimarina* DSM 24565^T).

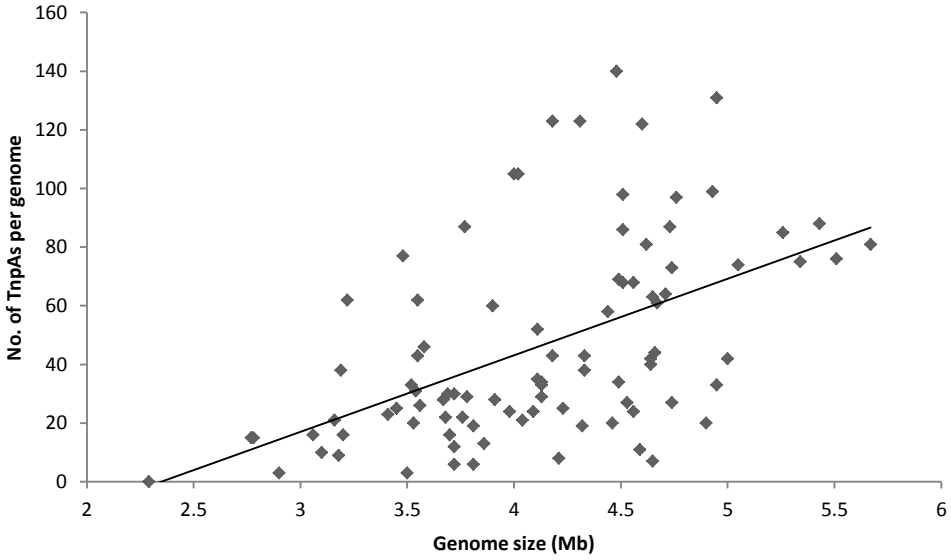

Figure 3.8. Relationship between the number of putative ISs and genome size. Genomes with more than 200 TnpAs were not represented.

With respect to the distribution of TnpAs in IS families, we detected representatives of all the families described in IS database (Siguier, 2006) (see Table 3.6 and Table 3S.4 for further details). This showed the high diversity of putative transposition mechanisms in this lineage. The distribution of the detected TnpAs in IS families is shown in Figure 3.9. The most detected IS families were IS3 with a total of 2,097 TnpAs assigned and IS5 with 1,321 TnpAs (approximately 30 and 18 % of TnpAs, respectively). However, the high percentage of TnpA assignation to these two families could be due to their good representation in the database (Siguier *et al*., 2014). A bias due to database content could be also behind those cases in which the number of TnpAs assigned to an IS family was lower than 10 (that includes 9 IS families that were not represented in Figure 3.9: IS701, ISKra4, IS1634, ISAzo13, IS1, IS607, ISLre2, IS982 and ISH3), which were those with worse representation in the database. Sixty-one TnpAs (0.88 %) could not be assigned to any IS family.

We also determined the distribution of TnpAs in IS families in each phylogenomic group (see Figure 3.10 and Table 3S.4 for further details). Only the IS families that had a minimum of 5 % representation are shown. In genomes of all phylogenomic groups with the exception of G3, IS3 family was the best represented. In G3 the majority of TnpAs belonged to IS5 family. For the rest of IS families there were variations in the percentages depending on the phylogenomic group but, in general the best-represented (apart from IS3 and IS5) were IS66, IS630 and IS481.

Because we wanted to know how many TnpAs were shared among the isolates as a way to make a prediction about possible HGT processes, we studied the TnpAs that were common in the 96 genomes. For that reason, the 6,884 plausible TnpAs were clustered in 4,065 Clusters of Ortologous Groups (COGs) under the strict criteria of 90 % identity in at least 95 % of its amino

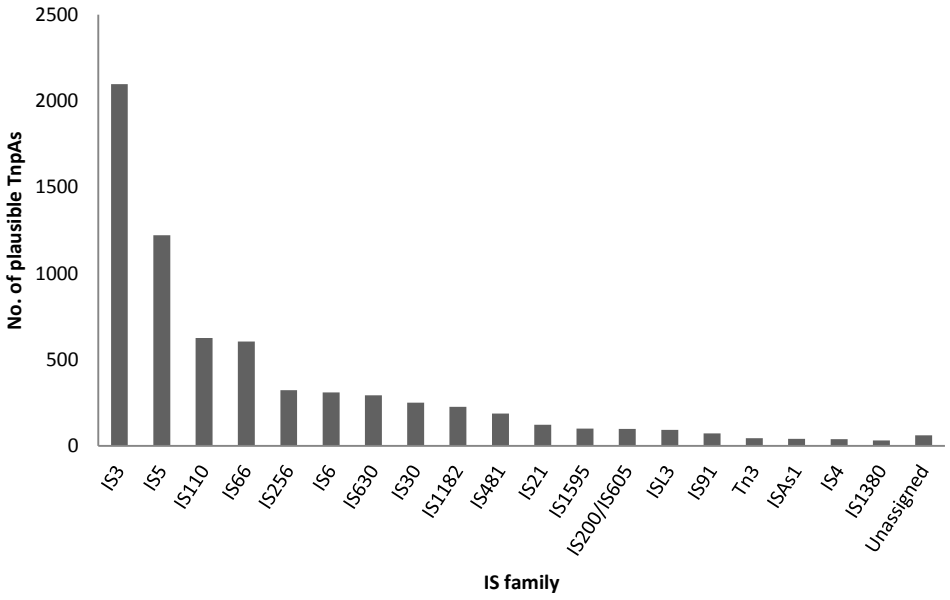

Figure 3.9. Distribution of plausible TnpAs of roseobacters in IS families. The IS families that had a less than 10 plausible TnpAs assigned are not shown in the Figure.

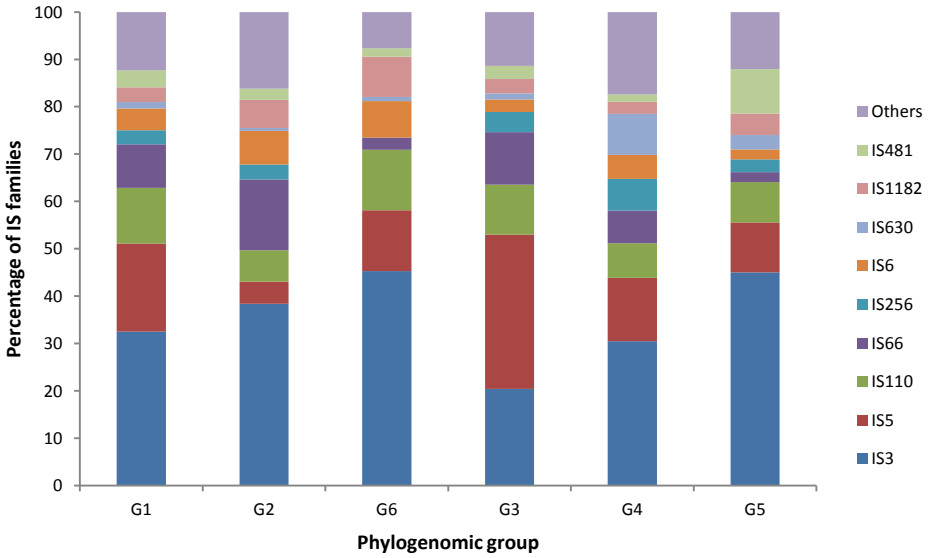

Figure 3.10. Percentage of TnpAs in different IS families per phylogenomic group. IS families that with percentages of occurrence lower than 5 % are not shown.

acid sequence (90_95) (see supplementary Table 3S.5 for detail). This clustering grouped highly similar TnpAs and would allow us to identify recent HGT processes among groups. The number of TnpA COGs of each Roseobacter genome and how many of these COGs were shared with other genomes is shown in Table 3.6. Because each COG could contain TnpAs from different genomes, the study of the distribution of the TnpAs COGs among the genomes was done (for further details check supplementary Table 3S.6). From the 95 genomes that harbored transposases, 5 of them did not share any TnpA COG with other Roseobacter included in this survey [*Oceaniovalibus guishaninsula* JLT2003^{T} (G4), *Roseobacter* sp. CCS2 (G4), *Actibacterium*

mucosum KCTC 23349^T (G5), *Jannaschia* sp. CCS1 (G5), and *Rhodobacteraceae* bacterium HIMB11 (G6)]. From 4,065 Clusters, 324 TnpA COGs (8 %) were shared between 2 or more roseobacters. The number of TnpAs that constituted the 324 shared COGs supposed approximately the 23 % of total TnpAs (1,562 shared TnpAs). That means that few TnpAs were common.

The shared TnpA COGs (90 roseobacters genomes) are shown in Figure 3.11. Phylogenomic groups are differentiated by color code in the Figure. The size of the circles indicates the number of interactions (shared TnpAs) per each genome. Line thickness is related to the number of shared TnpAs COGs. As can be seen in the Figure there was not a pattern that indicated that there were more shared COGs betwen genomes of the same phylogenomic group. The majority of shared TnpA COGs (57 %) were shared between only two genomes. Some genomes highlighted for sharing TnpA COGs with a high number of different genomes [i.e *Pseudophaeobacter arcticus* DSM 23566^T (G1), *Roseobacter* sp. MED193 (G1), *Donghicola* sp. S598 (G3), and *Octadecabacter antarcticus* 307^T (G4)]. Therefore, these genomes would have more possibilities (or have had) for exchanging genetic material with other different genomes by transposition. Five other genomes shared TnpA COGs (i.e. those of *Citreicella aestuarii* strains 328, 329, 357 and AD8^T) with many other genomes (see Table 3.6). However, the most interesting finding was that most of the TnpA COGs in these genomes (35 COGs) were shared between the five *C. aestuarii* strains as shown by the thickness of the connection lines (see Figure 3.11.). This degree of sharing has not been seen between other genomes from the same species. We had additional examples of a higher degree of sharing of TnpA COGs between phylogenetically related species: the two *Thalassobacter stenotrophicus* strains (8 COGs); the two *Ketogulonicigenium vulgare* strains (8 COGs); the two *Octadecabacter* spp. (12 COGs shared). Another interesting case was that 12 COGs were constituted for more than 20 plausible TnpAs that belonged exclusively to one genome (for further details see in Table 3S.7): 10 COGs were exclusive of *Octadecabacter arcticus* 238^T, 1 COG was constituted only for plausible TnpAs that belonged to *Octadecabacter antarcticus* 307^T and the last one was constituted for TnpAs that belonged to *Maribius* sp. MOLA 401 (see Table 3S.7 and for further details see Table 3S.5). These results suggested that the same plausible TnpA was in several copies in the same genome. All three genomes were members of phylogenomic group G4.

Apart from the few examples mentioned above, the analysis shown in Figure 3.11 pointed to a lack of relationship between phylogenomic groups and the degree of intra- and intergroup sharing of TnpAs. In order to analyze the relationship between genome phylogenetic closeness and the sharing of TnpA COGs we did the analysis of: i) shared COGs between genomes of the same phylogenomic group (intragroup), ii) shared COGs between genomes of different phylogenomic groups (intergroup) and iii) relationship between shared COGs and phylogenomic distance. The detailed information on the number of shared COGs between each genome and the rest of genomes in the same phylogenomic group (intragroup) is shown in Table 3S.8 and the graphic representation is shown in Figures 3S.4 to 3S.8. From this data we calculated median values of shared COGs within each phylogenomic group. Median values

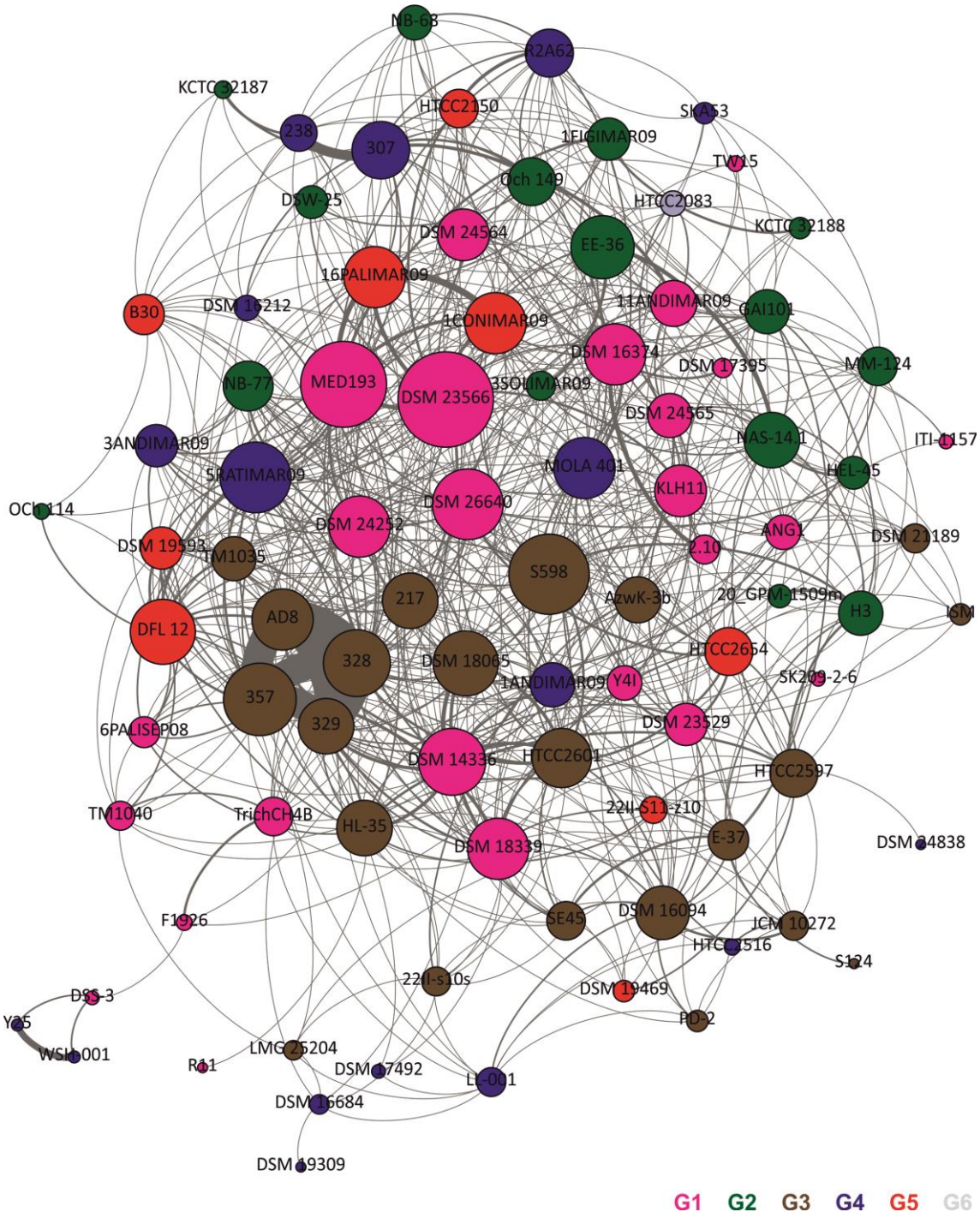

Figure 3.11. Shared COGs of TnpAs among 90 roseobacters differentiated in phylogenomic groups. The criteria used for considering that a plausible TnpA is shared was a minimum of 90 % of identity in, at least, the 95 % of amino acid sequence. The visual representation was done using Gephi v. 0.9.1 (Bastian *et al*., 2009).

of shared COGs were higher for G1, G2 and G3 groups (5, 4.5 and 7 COGs per genome, respectively) than for groups G4, G5 and G6 (3, 1 and 0 per COGs per genome, respectively). That could be because phylogenomic groups G1, G2 and G3 were formed by genomes that were closer phylogenetically (see section 2 of chapter 1). In any case, the degree of intragroup sharing of TnpA COGs can be considered low (as can be also seen in Figure 3.11.).

With respect to the shared COGs among the phylogenomic groups (intergroup) (see Table 3.7) the results show that genomes of all phylogenomic groups shared TnpA COGs with genomes of the other groups, as could be seen in Figure 3.11. However, there were differences in the number of shared COGs between the different groups. Genomes in phylogenomic groups G1 and G3 had the highest number of shared COGs (57). When we referred to percentages, the genomes that shared more TnpA COGs in reference to the total of the COGs predicted in genomes of the group were those of group G5 (20 % of its COGs). All the other shared less than the 15 % of its COGs. In spite of the fact that genomes in the phylogenomic G4 had the largest number of TnpA COGs, it was the group that shared less number of COGs with other phylogenomic groups. That was because the genomes of the two *Octadecabacter* species in G4 had the largest number of TnpAs of the studied roseobacters and most of them were singletons or were shared exclusively between them (see above and in Tables 3S.6 and 3S.8). Two TnpA COGs were shared among all phylogenomic groups G1 to G5, and eight were shared between 4 phylogenomic groups (for further details see Table 3S.5). These ten COGs belonged to families IS3 (COGs 2, 7, 21 and 82), IS5 (COGs 10, 17, 20 and 68) and IS481 (COGs 84 and 94).

As a conclusion of the analysis we can see that although the phylogenetic groups were constituted by a variety of genera and species the genomes shared a certain number of plausible TnpAs COGs (an average of 10 % of the total COGs of TnpAs of each phylogenomic group, see Table 3.7), fact that indicated that there have been a certain level of HGT processes in the recent evolutionary history of these phylogenomic groups.

Table 3.7. Shared TnpA COGs between phylogenomics groups (intra- and intergroup).

	Number and percentages of shared TnpA COGs[a]					
Group[b]	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**
G1	1002 (90.8)	28 (2.5)	57 (5.0)	23 (2.1)	27 (2.4)	3 (0.3)
G2	28 (5.8)	423 (87.8)	22 (4.6)	18 (3.7)	15 (3.1)	1 (0.2)
G3	57 (6.6)	22 (2.3)	867 (90.0)	29 (3.0)	18 (3.7)	1 (0.1)
G4	23 (1.6)	18 (1.3)	29 (2.1)	1380 (96.0)	17 (1.2)	1 (0.1)
G5	27 (11.8)	15 (6.6)	18 (7.9)	17 (7.4)	183 (80.0)	2 (3.7)
G6	3 (5.6)	1 (1.9)	1 (1.9)	1 (1.9)	2 (3.7)	49 (90.7)
Total shared	101 (9.1)	59 (12.2)	93 (9.6)	57 (3.9)	46 (20)	5 (9.2)

a: percentages are shown in parenthesis, b: number of TnpA COGs in each group: G1, 1,103; G2, 482; G3: 963; G4, 1,437; G5, 229; G6, 54.

In addition, we analyzed the relationship between the shared COGs and the phylogenomic distance between genomes. For this, a global study that consisted on the analysis of shared TnpAs COGs between each genome, compared to the phylogenomic distance (see in section 3 of Materials and Methods) was done (Figure 3.12). At phylogenetic distance values higher than 0.35 we did not find shared TnpA COGs. This would mean that there should be a certain level of phylogenomic relatedness between roseobacters for sharing transposases. As the relatedness between genomes increased (lower phylogenetic distance), more shared TnpA COGs were detected. However, there was not a trend that showed a relationship between these two parameters because genomes form a wide range of phylogenomic distances (i.e between 0.05 and 0.3) shared similar number of TnpA COGs (see Figure 3.12). The values near the Y-axis correspond to those pair of genomes which shared many COGs. These were the

genomes from strains of the same species or species of the same genus that have been commented above (i.e. *C. aestuarii* strains or *Octadecabacter* spp.).

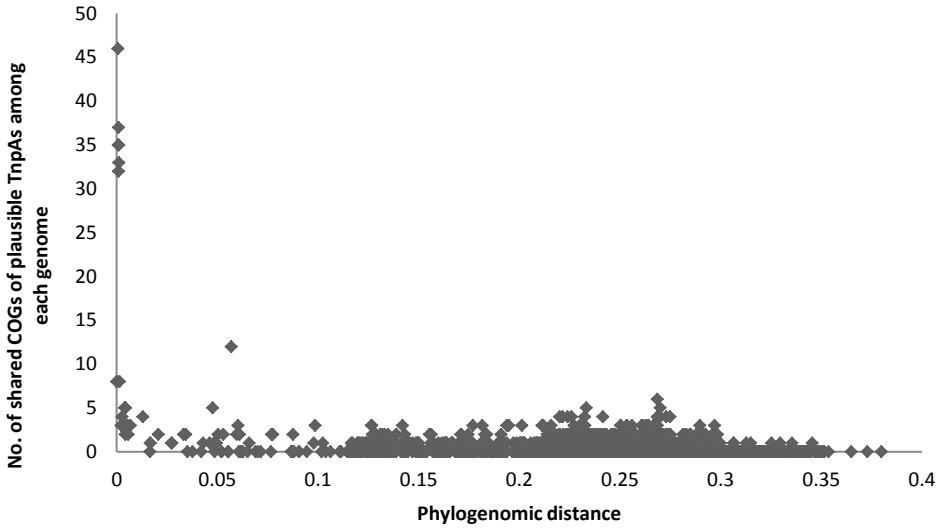

Figure 3.12. Relationship between the number of shared TnpAs COGs between each pair of genomes in comparison to their phylogenomic distance. The shared COGs of plausible TnpAs were defined with the criteria of a minimum of 90 % of identity in, at least, the 95 % of amino acid sequence and the phylogenomic distance was calculated with the concatenation of the single copy proteins of the core proteome using the approximation of Kimura with protein distance (PROTDIST) of the Phylip program (Felsenstein, 1989).

Finally, we analyzed the shared TnpA COGs between genomes from roseobacters that shared the same habitat or that were isolated from the same oceanic regions. Hooper and co-workers (2009) stablished a microbial interaction network through the study of the distribution of transposases using almost 800 sequenced genomes. These authors reached the conclusion that most taxa that shared TnpAs were sharing habitat. Taking into account the information about the origin of the isolates whose genomes have been analyzed in this study we defined 8 habitat categories (see Table 1S.1): i) metazoans (coral, sponge, scallop, etc.), ii) lignin enriched effluents or water, iii) dinoflagellates or other algae, iv) industry processes, v) polar sea ice, vi) seawater, vii) sediments and viii) other sources (soil, biofilms, saline lakes and unknown sources). In the majority of cases (see as an example Figure 3S.9), few genomes of roseobacters from the same habitat category shared TnpA COGs. Moreover, when they were shared they usually were with one or two other genomes (in rare occasions with 3). In some cases, the connections were only with genomes of the same phylogenomic group and therefore could be attributed to both factors, phylogeny and/or habitat. The two habitat categories in which most of the genomes shared COGs with others, and these interactions were usually with more than 3 genomes, were seawater and sediments (Figure 3.13). For the seawater and sediment samples we also defined 7 sub-categories for oceanic regions from which the isolates were obtained (for further details see in Table 1S.1): Mediterranean Sea (Mallorca harbors sub-category), North Atlantic Ocean (Sargasso Sea sub-category), South Atlantic Ocean, North Pacific Ocean, South Pacific Ocean, Central Pacific Ocean and Arctic Ocean. As can be seen in Figure 3.13A many of the genomes from seawater and sediments ha-

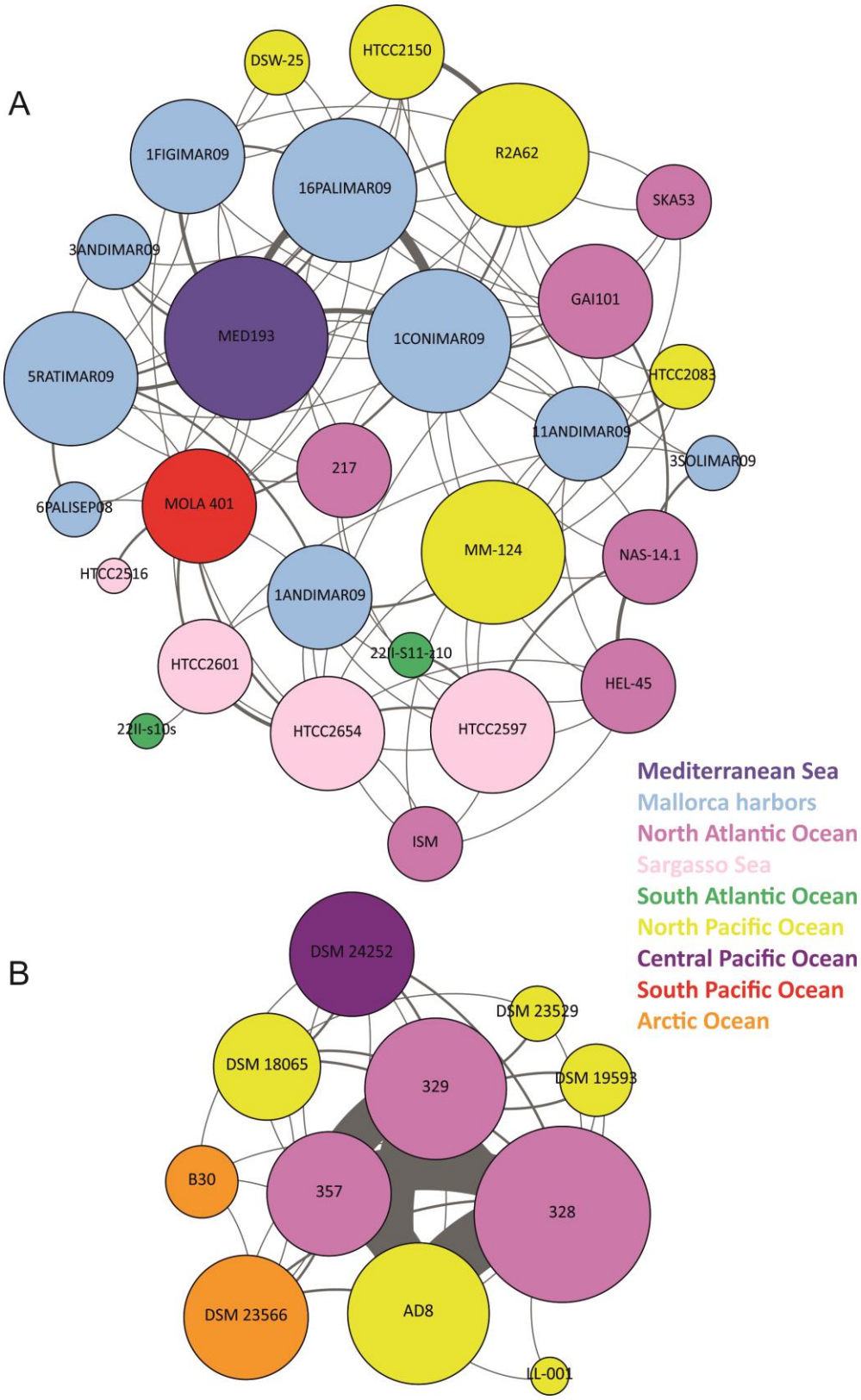

Figure 3.13. Shared TnpA COGs among roseobacters that were isolated from seawater (A) and sediment (B) habitats regarding to oceanic regions from which the isolates were obtained. The criterion used for considering that a plausible TnpA is shared was a minimum of 90 % of identity in, at least, the 95 % of amino acid sequence. The visual representation was done using Gephi v. 0.9.1 (Bastian *et al.*, 2009).

bitats shared TnpA COGs (see the sizes of the circles) but the connections were with multiple genomes, and from different geographic regions. The number of connections between genomes isolated in the same geographic region and same sampling time such as Mallorca harbor isolates or Sargasso Sea isolates was not higher than those between genomes from bacteria from other oceanic samples. The only exceptions were isolates 1CONIMAR09 and 16PALIMAR09, but these two isolates were two strains of *Thalassobacter stenotrophicus* (see section 2 in chapter 1). Interestingly, these genomes shared more TnpA COGs with isolate *Roseobacter* sp. MED193 obtained from a location in the Catalan Coast (NW Mediterranean, Spain), than with other isolates from Mallorca coast. In the case of isolates obtained from sediments we saw similar results, but fewer connections between genomes except for the *Citreicella aestuarii* strains (see Figure 3.13B). *C. aestuarii* isolates 328, 329 and 357 were obtained from the same type of sample in Spain (beach sand polluted with crude oil, same sampling time, and same medium for isolation). Therefore, it was not surprising that they shared a high number of TnpA COGs. But the degree of TnpA COGs sharing with the genome of the type strain AD8^T, isolated from a tidal flat in South Korea three years later, was not expected. Wether this is characteristic of *C. aestuarii* strains or if is related to type of habitat should be evaluated as new genomes of this species are sequenced. The only genome from this genus that we had for comparison was *Citreicella* sp. SE45, isolated from marsh plant detritus, and there were few shared TnpA COGs (see Figure 3.11. and Table 3S.6).

To sum up, we could not find a solid relationship between the sharing of TnpAs and phylogenomic relatedness, habitat or geographical place of isolation in the case of seawater or sediment samples. Our results showed that in some cases, but not always, when the strains belonged to the same genera or species shared more TnpAs. In these cases, the transmission of these shared TnpAs was probably vertical. In relation to habitat or geographical place of isolation we also could not establish patterns. Sometimes isolates from the same region (i.e. some Mallorca isolates and *Roseobacter* sp. MED193) shared TnpA COGs. But we also had the contrary as for example Mallorcan isolate 3SOLIMAR09 did not share transposases with isolates from same geographic area. In any case, the connections shown in this study by shared TnpA COGs indicated that there is a potential flow of genetic material between roseobacters. HGT processes via TnpAs is a factor that should be considered when explaining the diversity and metabolic versatility of cultured members of the Roseobacter lineage.

4. Comparison of the different HGT mechanisms studied

In this survey, we described how GTAs, ECEs and transposases were distributed in 96 members of Roseobacter lineage. In reference to their ocurrence, transposases were the most widespread HGT mechanisms in the lineage (TnpAs were present in 99 % of the genomes), followed by ECEs (in 90.6 % of analyzed genomes) and lastly GTAs (in 88.5 % of analyzed genomes). Eighty-one percent of genomes harbored the three HGT mechanisms studied, 9 % had only ECEs and TnpAs, 7 % had only the GTA genic cluster and TnpAs and a 1 % had only TnpAs (see Table 3S.9). *Rhodobacterales* bacterium HTCC2255 was the only genome in which

none of the studied HGT mechanisms was detected. The lack of the HGT mechanisms that we studied plus its smallest genome size (2.28 Mb), low G+C content mol % and lack of genes for transcriptional regulation, motility and cell-cell interactions (Luo *et al.*, 2013) made it peculiar in reference to the other cultured Roseobacter members.

Our results showed that GTAs were transmitted vertically (GTA phylogeny agreed with the core proteome phylogeny) and were well-conserved in the lineage because they probably evolved from a common ancestor (Biers *et al*., 2008; Lang *et al*., 2012; Lang & Beatty, 2006). On the contrary, RepABC family plasmids and transposases were transmitted promiscuously among members of the lineage. We found genomes that were distantly related but had in common the same partioning, segregation and replication systems or the same transposases. In the case of transposases, we could not determine which factor contributed to HGT fluxes among genomes, although in few cases there were agreements with phylogenomic relatedness (same species or same genus), habitat or oceanic region.

The frequency of the three studied HGT mechanisms in genomes of the cultured members of Roseobacter lineage could have contributed the high genetic diversity and genomic plasticity that is found in the lineage.

CONCLUSIONS

CONCLUSIONS

1. Nine genomes of Roseobacter isolates from harbors of Mallorca Island have been sequenced. Genome sizes were between 3.4 and 4.7 Mb, 56.8 and 62.0 G+C content mol %, and the numbers of predicted proteins were between 3,178 and 4,334. These values are within the ranges described for cultured roseobacters.

2. Isolates have been taxonomically affiliated using Average Nucleotide Identity (ANI) and virtual DNA-DNA hybridization (DDH) to five different genera and eight different genomic species: *Pseudophaeobacter* sp. 11ANDIMAR09, *Ruegeria* sp. 6PALISEP08, *Sulfitobacter mediterraneus* 1FIGIMAR09, *Sulfitobacter* sp. 3SOLIMAR09, *Loktanella* spp. 1ANDIMAR09, 3ANDIMAR09 and 5RATIMAR09, and *Thalassobacter stenotrophicus* 16PALIMAR09 and 1CONIMAR09. Species names could not be given to six isolates because there were no genome sequences available from the closest type strains.

3. A phylogenomic tree of the Roseobacter lineage based in the highest number of core proteins (114) and the highest number of genome sequences (96), compared to previous studies, has been generated. A possible new genomic group (G6), formed by the sequences of unclassified strains HTCC2083 and HIMB11, was evidenced together with the former five genomic groups described previously. A subset of fourteen of these core protein sequences was found to reproduce well the phylogeny of 165 Roseobacter genomes, with minor discrepancies. Therefore, this is proposed as a valid and simpler method for the affiliation of new genome sequences.

4. Predicted metabolic characteristics, as well as catabolic potential of aromatic compounds and alkanes of the nine harbor isolates based on genome sequences did not evidence differences with the other Roseobacter genomes included in this analysis, mostly isolated from non-polluted sources.

5. Proteomic analyses of cultures of the nine isolates exposed to diesel oil has been useful for understanding mechanisms of hydrocarbon tolerance. The general tolerance response described in the literature involving the expression of proteins for maintenance of membrane stability and transport, reduction of oxidative stress and chaperones has been observed after diesel oil treatment, although there were differences in the response of particular isolates. Alkane monooxygenase was detected in three isolates, including two able to grow with diesel oil (*Loktanella* sp. 3ANDIMAR09 and *Sulfitobacter* sp. 3SOLIMAR09).

6. In the conditions of the experiment of exposure to diesel oil, proteomes of all isolates showed that cells had activated a stringent response due to nutrient deprivation (in both, control and diesel treatment). The abundance of these proteins might have limited the detection of proteins induced after diesel oil treatment. This limitation should be taken into account for future experiments.

7. Three mechanisms of horizontal gene transfer (HGT) have been studied in 96 genomes of the Roseobacter lineage: gene transfer agents (GTAs), extrachromosomal elements (ECEs) and transposases (TnpAs). This is the broader study, in terms of HGT mechanisms and number of genomes, done so far. The most widely distributed HGT mechanism among the genomes analyzed were TnpAs followed by ECEs and finally GTAs. *Rhodobacterales* bacterium HTCC2255, an example of streamlined genome similar to uncultured representatives of the lineage, was the only genome in which none of the studied HGT mechanisms was detected.

8. The GTAs-based phylogeny agreed with the core protein phylogeny of the lineage while a phylogeny calculated with RepC replicases from RepABC plasmids did not. This suggests that GTAs transmission in the Roseobacter lineage is vertical, while RepABC plasmids would be exchanged promiscuously among roseobacters, although less than 30 % of these plasmids were potentially conjugative. RepC phylogeny evidenced 10 compatibility groups and 8 compatibility sub-groups based on the phylogenies of RepA and/or RepB in RepABC plasmids.

9. The distribution of the five types of ECEs analyzed (RepABC, RepA, RepB, RepC and DnaA-like) was not uniform in the phylogenomics groups of the lineage. For example, RepABC replicons were more frequent in genomes of phylogenomic groups G1, G2 and G3. There was also variability in the number of RepABC modules per genome and putative replicon size in the different genomic groups.

10. The presence of transposases, their classification in IS families, and possible transposase fluxes (using strict criteria of sequence similarity) between genomes of the Roseobacter lineage have been analyzed for the first time. With the exception of the genome of strain HTCC2255, TnpAs were found in all genomes and they belonged to a variety of IS families (mainly IS3 and IS5). The number of TnpAs per genome was variable but there was a significant correlation between genome size and number of TnpAs (0.35, $p<0.001$).

11. The network of shared transposases was complex, with multiple connections between genomes, but without a clear pattern in relation to phylogenomic relatedness, habitat and geographic place of isolation. This indicates a wide potential for the flow of genetic material between roseobacters.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aguiló-Ferretjans MM, Bosch R, Martín-Cardona C, Lalucat J, Nogales B (2008) Phylogenetic analysis of the composition of bacterial communities in human-exploited coastal environments from Mallorca Island (Spain). Systematic and applied microbiology. 31, 231–240.

Aires JR, Nikaido H (2005) Aminoglycosides are captured from both periplasm and cytoplasm by the AcrD multidrug efflux transporter of *Escherichia coli*. Journal of Bacteriology. 187, 1923–1929.

Alavi M, Miller T, Erlandson K, Schneider R, Belas R (2001) Bacterial community associated with *Pfiesteria*-like dinoflagellate cultures. Environmental Microbiology. 3, 380–396.

Alejandro-Marín CM, Bosch R, Nogales B (2014) Comparative genomics of the protocatechuate branch of the β-ketoadipate pathway in the Roseobacter lineage. Marine Genomics. 17, 25–33.

Allgaier M, Uphoff H, Felske A, Wagner-Döbler I (2003) Aerobic anoxygenic photosynthesis in *Roseobacter* clade bacteria from diverse marine habitats. Applied and Environmental Microbiology. 69, 5051–5059.

Allocati N, Federici L, Masulli M, Di Ilio C (2009) Glutathione transferases in bacteria. FEBS Journal. 276, 58–75.

Alonso-Gutiérrez J, Figueras A, Albaigés J, Jiménez N, Viñas M, Solanas AM, Novoa B (2009) Bacterial communities from shoreline environments (Costa da Morte, Northwestern Spain) affected by the Prestige oil spill. Applied and Environmental Microbiology. 75, 3407–3418.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology. 215, 403–410.

Alvarez-Martinez C, Christie P (2009) Biological diversity of prokaryotic type IV secretion systems. Microbiology and Molecular Biology Reviews. 73, 775–808.

Amin SA, Parker MS, Armbrust EV (2012) Interactions between diatoms and bacteria. Microbiology and Molecular Biology Reviews. 76, 667–684.

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira CD, Kyrpides N, Madupu R, Markowitz V, Tatusova T, Thomson N, White O (2008) Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation. OMICS : a Journal of Integrative Biology. 12, 137-141.

Apprill A, Marlow HQ, Martindale MQ, Rappé MS (2009) The onset of microbial associations in the coral *Pocillopora meandrina*. ISME Journal. 3, 685–699.

Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, De Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, Stockinger H (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Research. 40, 597–603.

Ashen JB, Goff LJ (2000) Molecular and ecological evidence for species specificity and coevolution in a group of marine algal-bacterial symbioses. Applied and Environmental Microbiology. 66, 3024–3030.

Aziz RK, Breitbart M, Edward R (2010) Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Research. 38, 4207–4217.

Bacha J, Freel J, Gibbs A, Gibbs L, Hemighaus G, Hoekman K, Horn J, Gibbs A, Ingham M, Jossens L, Kohler D, Lesnini D, McGeehan J, Nikanjam M, Olsen E, Organ R, Scott B, Sztenderowicz M, Tiedemann A, Walker C, Lind J, Jones J, Scott D, Mills J (2007) Diesel fuels technical review. Chevron Global Marketing.

Barbieri E, Paster BJ, Hughes D, Zurek L, Moser D, Teske A, Sogin ML (2001) Phylogenetic characterization of epibiotic bacteria in the accessory nidamental gland and egg capsules of the squid *Loligo pealei* (Cephalopoda: *Loliginidae*). Environmental Microbiology. 3, 151–167.

Barnett ME, Zolkiewska A, Zolkiewski M (2000) Structure and activity of ClpB from *Escherichia coli*. Role of the amino- and carboxyl-terminal domains. Journal of Biological Chemistry. 275, 37565–37571.

Bartosik D, Baj J, Wlodarczyk M (1998) Molecular and functional analysis of pTAV320, a *repABC*-type replicon of the *Paracoccus versutus* composite plasmid pTAV1. Microbiology. 144, 3149-3147.

Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. 8, 361–362.

Becker PT, Gillan DC, Eeckhaut I (2009) Characterization of the bacterial community associated with body wall lesions of *Tripneustes gratilla* (Echinoidea) using culture-independent methods. Journal of Invertebrate Pathology. 100, 127–130.

Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. Proceedings of the National Academy of Sciences of the United States of America. 102, 14332–14337.

Béjà O, Suzuki MT, Heidelberg JF, Nelson WC, PrestonCM, Hamada T, Eisen JA, Fraser CM, DeLong EF (2002) Unsuspected diversity among marine aerobic anoxygenic phototrophs. Nature. 415, 630–633.

Berger M, Brock NL, Liesegang H, Dogs M, Preuth I, Simon M, Dickschat J, Brinkhoff T (2012) Genetic analysis of the upper phenylacetate catabolic pathway in the production of

tropodithietic acid by *Phaeobacter gallaeciensis*. Applied and Environmental Microbiology. 78, 3539–3551.

Berger M, Neumann A, Schulz S, Simon M, Brinkhoff T (2011) Tropodithietic acid production in *Phaeobacter gallaeciensis* is regulated by *N*-acyl homoserine lactone-mediated quorum sensing. Journal of Bacteriology. 193, 6576–6585.

Bernal P, Segura A, Ramos JL (2007) Compensatory role of the *cis*-trans-isomerase and cardiolipin synthase in the membrane fluidity of *Pseudomonas putida* DOT-T1E. Environmental Microbiology. 9, 1658–1664.

Biebl H, Allgaier M, Tindall BJ, Koblizek M, Lünsdorf H, Pukall R, Wagner-Döbler I (2005) *Dinoroseobacter shibae* gen. nov., sp. nov., a new aerobic phototrophic bacterium isolated from dinoflagellates. International Journal of Systematic and Evolutionary Microbiology. 55, 1089–1096.

Biers EJ, Wang K, Pennington C, Belas R, Chen F, Moran MA (2008) Occurrence and expression of gene transfer agent genes in marine bacterioplankton. Applied and Environmental Microbiology. 74, 2933-2939.

Boettcher KJ, Geaghan KK, Maloy AP, Barber BJ (2005) *Roseovarius crassostreae* sp. nov., a member of the *Roseobacter* clade and the apparent cause of juvenile oyster disease (JOD) in cultured Eastern oysters. International Journal of Systematic and Evolutionary Microbiology. 55, 1531–1537.

Bosch R, García-Valdés E, Moore ERB (1999) Genetic characterization and evolutionary implications of a chromosomally encoded naphthalene-degradation upper pathway from *Pseudomonas stutzeri* AN10. Gene. 236, 149–157.

Boutte CC, Crosson S (2013) Bacterial lifestyle shapes stringent response activation. Trends in Microbiology. 21, 174-180.

Brakstad OG, Lødeng AGG (2005) Microbial diversity during biodegradation of crude oil in seawater from the North Sea. Microbial Ecology. 49, 94–103.

Brimacombe CA, Stevens A, Jun D, Mercer R, Lang AS, Beatty JT (2013). Quorum-sensing regulation of a capsular polysaccharide receptor for the *Rhodobacter capsulatus* gene transfer agent (RcGTA). Molecular Microbiology. 87, 802–817.

Brinkhoff T, Bach G, Heidorn T, Liang L, Schlingloff A, Simon M (2004) Antibiotic production by a *Roseobacter* clade-affiliated species from the German Wadden Sea and its antagonistic effects on indigenous isolates. Applied and Environmental Microbiology. 70, 2560–2565.

Brinkhoff T, Giebel HA, Simon M (2008) Diversity, ecology, and genomics of the *Roseobacter* clade: A short overview. Archives of Microbiology. 189, 531–539.

Brinkmeyer R, Knittel K, Jürgens J, Weyland H, Amann R, Helmke E (2003) Diversity and structure of bacterial communities in Arctic versus Antarctic pack ice. Applied and Environmental Microbiology. 69, 6610–6619.

Brito EMS, Guyoneaud R, Goñi-Urriza M, Ranchou-Peyruse A, Verbaere A, Crapez MAC, Wasserman JCA, Duran R (2006) Characterization of hydrocarbonoclastic bacterial communities from mangrove sediments in Guanabara Bay, Brazil. Research in Microbiology. 157, 752–762.

Bruhn JB, Gram L, Belas R (2007) Production of antibacterial compounds and biofilm formation by *Roseobacter* species are influenced by culture conditions. Applied and Environmental Microbiology. 73, 442–450.

Bruhn JB, Nielsen KF, Hjelm M, Hansen M, Bresciani J, Schulz S, Gram L (2005) Ecology, inhibitory activity, and morphogenesis of a marine antagonistic bacterium belonging to the *Roseobacter* clade. Applied and Environmental Microbiology. 71, 7263–7270.

Brynildsen MP, Liao JC (2009) An integrated network approach identifies the isobutanol response network of *Escherichia coli*. Molecular Systems Biology. 5, 277.

Buchan A, Collier LS, Neidle EL, Buchan A, Collier LS, Neidle EL (2000) Key aromatic-ring-cleaving enzyme, protocatechuate 3, 4-dioxygenase, in the ecologically important marine *Roseobacter* lineage. Applied and Environmental Microbiology. 66, 4662–4672.

Buchan A, González JM (2010) Roseobacter. Handbook of hydrocarbon and lipid microbiology. Timmis, K.N. (ed.). Berlin, Germany: Springer-Verlag. pp. 1335–1343.

Buchan A, LeCleir GR, Gulvik CA, González JM (2014) Master recyclers: features and functions of bacteria associated with phytoplankton blooms. Nature Reviews Microbiology. 12, 686–698.

Buchan A, Moran MA (2005) Overview of the marine *Roseobacter* lineage. Applied and Environmental Microbiology. 71, 5665-5677.

Buchan A, Neidle EL, Moran MA (2001) Diversity of the ring-cleaving dioxygenase gene pcaH in a salt marsh bacterial community. Applied and Environmental Microbiology. 67, 5801-5809.

Buchan A, Neidle EL, Moran MA (2004) Diverse organization of genes of the β-ketoadipate pathway in members of the marine *Roseobacter* lineage. Applied and Environmental Microbiology. 70, 1658–1668.

Cabreiro F, Picot CR, Friguet B, Petropoulos I (2006) Methionine sulfoxide reductases: relevance to aging and protection against oxidative stress. Annals of the New York Academy of Sciences. 1067, 37–44.

Cao J, Lai Q, Yuan J, Shao Z (2015) Genomic and metabolic analysis of fluoranthene degradation pathway in *Celeribacter indicus* P73^{T}. Scientific Reports. 5, 7741.

Case RJ, Longford SR, Campbell AH, Low A, Tujula N, Steinberg PD, Kjelleberg S (2011) Temperature induced bacterial virulence and bleaching disease in a chemically defended marine macroalga. Environmental Microbiology. 13, 529–537.

Castillo-Ramírez S, Vázquez-Castellanos JF, González V, Cevallos MA (2009). Horizontal gene transfer and diverse functional constrains within a common replication-partitioning system in *Alphaproteobacteria*: the *repABC* operon. BMC Genomics. 10, 536.

Cervantes-Rivera R, Pedraza-López F, Pérez-Segura G, Cevallos MA (2011) The replication origin of a *repABC* plasmid. BMC Microbiology. 11, 158.

Cervantes-Rivera R, Romero-López C, Berzal-Herranz A, Cevallos MA (2010) Analysis of the mechanism of action of the antisense RNA that controls the replication of the *repABC* plasmid p42d. Journal of Bacteriology. 192, 3268–3278.

Cevallos MA, Cervantes-Rivera R, Gutiérrez-Ríos RM (2008) The *repABC* plasmid family. Plasmid. 60, 19–37.

Chai Y, Winans SC (2005a). A small antisense RNA downregulates expression of an essential replicase protein of an *Agrobacterium tumefaciens* Ti plasmid. Molecular Microbiology. 56, 1574–1585.

Chai Y, Winans SC (2005b). RepB protein of an *Agrobacterium tumefaciens* Ti plasmid binds to two adjacent sites between *repA* and *repB* for plasmid partitioning and autorepression. Molecular Microbiology. 58, 1114–1129.

Choudhary M, Kaplan S (2000) DNA sequence analysis of the photosynthesis region of *Rhodobacter sphaeroides* 2.4.1. Nucleic Acids Research. 28, 862–867.

Christie-Oleza JA, Fernandez B, Nogales B, Bosch R, Armengaud J (2012) Proteomic insights into the lifestyle of an environmentally relevant marine bacterium. ISME Journal. 6, 124–135.

Cilia V, Lambertl C, Nicolasl JL (1998) *Roseobacter gallaeciensis* sp. nov., a new marine bacterium isolated from rearings and collectors of the scallop *Pecten maximus*. International Journal of Systematic Bacteriology. 2216, 537–542.

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2016) GenBank. Nucleic Acids Research. 44, D67–D72.

Collins AJ, LaBarre BA, Won BSW, Shah MV, Heng S, Choudhury MH, Haydar SA, Santiago J, Nyholm SV (2012) Diversity and partitioning of bacterial populations within the accessory nidamental gland of the squid *Euprymna scolopes*. Applied and Environmental Microbiology. 78, 4200–4208.

Collins AJ, Nyholm SV (2011) Draft genome of *Phaeobacter gallaeciensis* ANG1, a dominant member of the accessory nidamental gland of *Euprymna scolopes*. Journal of Bacteriology. 193, 3397–3398.

Cunliffe M (2011) Correlating carbon monoxide oxidation with *cox* genes in the abundant Marine Roseobacter Clade. ISME Journal. 5, 685–691.

Dang H, Lovell CR (2000) Bacterial primary colonization and early succession on surfaces in marine waters as determined by amplified rRNA gene restriction analysis and sequence analysis of 16S rRNA genes. Applied and Environmental Microbiology. 66, 467–475.

Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Research. 14, 1394–1403.

Domínguez-Cuevas P, Marín P, Ramos JL, Marqués S (2005) RNA polymerase holoenzymes can share a single transcription start site for the Pm promoter critical nucleotides in the -7 to-18 region are needed to select between RNA polymerase with σ38 or σ32. Journal of Biological Chemistry. 280, 41315–41323.

Dressaire C, Moreira RN, Barahon, S, Alves de Matos AP, Arriano CM (2015) BolA is a transcriptional switch that turns off motility and turns on biofilm development. mBio. 6, e02352–14.

Du Y, Li G, Shao Z, Liu X, Lai Q, Sun F (2015) *Aestuariivita atlantica* sp. nov., isolated from deep-sea sediment. International Journal of Systematic and Evolutionary Microbiology. 65, 3281–3285.

Duarte M, Jauregui R, Vilchez-Vargas R, Junca H, Pieper DH (2014) AromaDeg, a novel database for phylogenomics of aerobic bacterial degradation of aromatics. Database. 2014:bau1.

Dunne WM (2002) Bacterial adhesion: seen any good biofilms lately? Clinical Microbiology Reviews. 15, 155-166.

Ebersbach G, Ringgaard S, Moller-Jensen J, Wang Q, Sherratt DJ, Gerdes K (2006) Regular cellular distribution of plasmids by oscillating and filament-forming ParA ATPase of plasmid pB171. Molecular Microbiology. 61, 1428–1442.

Eloe EA, Malfatti F, Gutierrez J, Hardy K, Schmidt WE, Pogliano K, Pogliano J, Azam F, Bartlett, DH (2011) Isolation and characterization of a psychropiezophilic alphaproteobacterium. Applied and Environmental Microbiology. 77, 8145–8153.

Felsenstein J (1989) PHYLIP-Phylogeny Inference Package (Version 3.2). Cladistics. 5, 164–166.

Ferguson SA, Keis S, Cook GM, Al FET, Acteriol JB (2006) Biochemical and molecular characterization of a Na^+-Translocating F_1F_0 -ATPase from the thermoalkaliphilic bacterium *Clostridium paradoxum*. Journal of Bacteriology. 188, 5045–5054.

Freire P, Vieira HL, Furtado AR, de Pedro MA, Arraiano CM (2006) Effect of the morphogene *bolA* on the permeability of the *Escherichia coli* outer membrane. FEMS Microbiology Letters. 260, 106–111.

Fuqua C, Greenberg EP (2002) Listening in on bacteria: acyl-homoserine lactone signalling. Nature Reviews in Molecular Cell Biology 3, 685–695.

García V, Godoy P, Daniels C, Hurtado A, Ramos JL, Segura A (2010) Functional analysis of new transporters involved in stress tolerance in *Pseudomonas putida* DOT-T1E. Environmental Microbiology Reports. 2, 389–395.

Gibson JL, Falcone DL, Tabita FR (1991) Nucleotide sequence, transcriptional analysis, and expression of genes encoded within the form I CO_2 fixation operon of *Rhodobacter sphaeroides*. Journal of Biological Chemistry. 266, 14646–14653.

Giebel HA, Klotz F, Voget S, Poehlein A, Grosser K, Teske A, Brinkhoff T (2016) Draft genome sequence of the marine *Rhodobacteraceae* strain O3. 65, cultivated from oil-polluted seawater of the Deepwater Horizon oil spill. Standards in Genomic Sciences. 11, 81.

González JM, Covert JS, Whitman WB, Henriksen JR, Mayer F, Scharf B, Schmitt R, Buchan A, Fuhrman JA, Kiene RP, Moran MA (2003) *Silicibacter pomeroyi* sp. nov. and *Roseovarius nubinhibens* sp. nov., dimethylsulfoniopropionate-demethylating bacteria from marine environments. International Journal of Systematic and Evolutionary Microbiology. 53, 1261–1269.

González JM, Mayer F, Moran MA, Hodson RE, Whitman WB (1997) *Sagittula stellata* gen. nov., sp. nov., a lignin-transforming bacterium from a coastal environment. International Journal of Systematic and Evolutionary Microbiology. 47, 773–780.

González JM, Moran MA (1997) Numerical dominance of a group of marine bacteria in the α-subclass of the class *Proteobacteria* in coastal seawater. Applied and Environmental Microbiology. 63, 4237–4242.

González JM, Simó R, Massana R, Covert JS, Casamayor EO, Pedrós-Alió C, Moran MA (2000) Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. Applied and Environmental Microbiology. 66, 4237–4246.

Gragerov A, Nudler E, Komissarova N, Gaitanaris GA, Gottesman ME, Nikiforov V (1992) Cooperation of GroEL/GroES and DnaK/DnaJ heat shock proteins in preventing protein misfolding in *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America. 89, 10341-10344.

Gram L, Grossart H, Schlingloff A, Kiørboe T (2002) Possible quorum sensing in marine snow bacteria: production of acylated homoserine lactones by *Roseobacte*r strains isolated from marine snow. Applied and Environmental Microbiology. 68, 4111–4116.

Grigioni S, Boucher-Rodoni R, Demarta A, Tonolla M, Peduzzi R (2000) Phylogenetic characterisation of bacterial symbionts in the accessory nidamental glands of the sepioid *Sepia officinalis* (Cephalopoda: Decapoda). Marine Biology. 136, 217–222.

de Groot A, Dulermo R, Ortet P, Blanchard L, Guérin P, Fernandez B, Vacherie B, Dossat C, Jolivet E, Siguier P, Chandler M, Barakat M, Dedieu A, Barbe V, Heulin T, Sommer S, Achouak W, Armengaud J (2009) PLoS Genetics. 5: e1000434.

Grossart H, Levold F, Allgaier M, Simon M, Brinkhoff T (2005) Marine diatom species harbour distinct bacterial communities. Environmental Microbiology. 7, 860–873.

Guglielmini J, De La Cruz F, Rocha EPC (2013) Evolution of conjugation and type IV secretion systems. Molecular Biology and Evolution. 30, 315–331.

Guinote IB, Moreira RN, Barahona S, Freire P, Vicente M, Arraiano CM (2014) Breaking through the stress barrier: the role of BolA in Gram-negative survival. World Journal of Microbiology and Biotechnology. 30, 2559–2566.

Gulvik CA, Buchan A (2013) Simultaneous catabolism of plant-derived aromatic compounds results in enhanced growth for members of the *Roseobacter* lineage. Applied and Environmental Microbiology. 79, 3716–3723.

Hammer Ø, Harper DAT, Ryan PD (2001) PAST-Palaeontological statistics software package for education and data analysis. Palaeontologia Electronica. 4, 1-9.

Harwati TU, Kasai Y, Kodama Y, Susilaningsih D, Watanabe K (2007) Characterization of diverse hydrocarbon-degrading bacteria isolated from Indonesian seawater. Microbes and Environments. 22, 412–415.

Harwati TU, Kasai Y, Kodama Y, Susilaningsih D, Watanabe K (2008) *Tranquillimonas alkanivorans* gen. nov., sp. nov., an alkane-degrading bacterium isolated from Semarang Port in Indonesia. International Journal of Systematic and Evolutionary Microbiology. 58, 2118–2121.

Harwati TU, Kasai Y, Kodama Y, Susilaningsih D, Watanabe K (2009) *Tropicibacter naphthalenivorans* gen. nov., sp. nov., a polycyclic aromatic hydrocarbon-degrading bacterium isolated from Semarang Port in Indonesia. International Journal of Systematic and Evolutionary Microbiology. 59, 392–396.

Harwood CS, Parales RE (1996) The β-ketoadipate pathway and the biology of self-identity. Annual Reviews in Microbiology. 50, 553–590.

Häuser R, Pech M, Kijek J, Yamamoto H, Titz B, Naeve F, Tovchigrechko A, Yamamoto K, Szaflarski W, Takeuchi N, Stellberger T, Diefenbacher ME, Nierhaus KH, Uetz P (2012) RsFA (YbeB) proteins are conserved ribosomal silencing factors. PLoS Genetics. 8, 1–12.

Head IM, Jones DM, Röling WFM (2006) Marine microorganisms make a meal of oil. Nature Reviews Microbiology. 4, 173–182.

Hemamalini R, Khare S (2014) A proteomic approach to understand the role of the outer membrane porins in the organic solvent-tolerance of *Pseudomonas aeruginosa* PseA. PLoS One. 9, e103788.

Hernandez-Raquet G, Budzinski H, Caumette P, Dabert P, Le Ménach K, Muyzer G, Duran R (2006) Molecular diversity studies of bacterial communities of oil polluted microbial mats from the Etang de Berre (France). FEMS Microbiology Ecology. 58, 550–562.

Hesterkamp T, Bukau B (1998) Role of the DnaK and HscA homologs of Hsp70 chaperones in protein folding in *E. coli*. EMBO Journal. 17, 4818-4828.

Hong H, Szabo G, Tamm LK (2006) Electrostatic couplings in OmpA ion-channel gating suggest a mechanism for pore opening. Nature Chemical Biology. 2, 627-365.

Hooper SD, Mavromatis K, Kyrpides NC (2009) Microbial co-habitation and lateral gene transfer: what transposases can tell us. Genome Biology. 10, R45.

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. BMC Bioinformatics. 26, 680–682.

Hynes AP, Mercer RG, Watton DE, Buckley CB, Lang AS (2012) DNA packaging bias and differential expression of gene transfer agent genes within a population during production and release of the *Rhodobacter capsulatus* gene transfer agent, RcGTA. Molecular Microbiology. 85, 314–325.

Ichikawa K, Kurihara M, Tamegai H, Hashimoto S (2015) Decomposition of brominated organic halogens by cultures of marine proteobacteria: *Phaeobacter*, *Roseobacter*, and *Rhodobacter*. Marine Chemistry. 176, 133–141.

Ikeda H, Tomizawa J (1965) Transducing fragments in generalized transduction by phage P1. Journal of Molecular Biology. 14, 85–109.

Isken S, De Bont JAM (1996) Active efflux of toluene in a solvent-resistant bacterium. Journal of Bacteriology. 178, 6056–6058.

Iwaki H, Nishimura A, Hasegawa Y (2012a) Isolation and characterization of marine bacteria capable of utilizing phthalate. World Journal of Microbiology and Biotechnology. 28, 1321–1325.

Iwaki H, Nishimura A, Hasegawa Y (2012b) *Tropicibacter phthalicus* sp. nov., a phthalate-degrading bacterium from seawater. Current Microbiology. 64, 392–396.

Jain R (2003) Horizontal gene transfer accelerates genome innovation and evolution. Molecular Biology and Evolution. 20, 1598–1602.

Jami M, Lai Q, Ghanbari M, Moghadam MS, Kneifel W, Domig KJ (2016) *Celeribacter persicus* sp. nov., a polycyclic-aromatic-hydrocarbon-degrading bacterium isolated from mangrove soil. International Journal of Systematic and Evolutionary Microbiology. 66, 1875-1880.

Jasti S, Sieracki ME, Poulton NJ, Giewat MW, Rooney-Varga JN (2005) Phylogenetic diversity and specificity of bacteria closely associated with *Alexandrium* spp. and other phytoplankton. Applied and Environmental Microbiology. 71, 3483–3494.

Jeong HI, Jin HM, Jeon CO (2015) *Confluentimicrobium naphthalenivorans* sp. nov., a naphthalene-degrading bacterium isolated from sea-tidal-flat sediment, and emended description of the genus *Confluentimicrobium* Park *et al*. 2015. International Journal of Systematic and Evolutionary Microbiology. 65, 4191–4195.

Jiménez N, Viñas, M, Bayona JM, Albaiges J, Solanas AM (2007) The Prestige oil spill: bacterial community dynamics during a field biostimulation assay. Applied Microbiology and Biotechnology. 77, 935–945.

Jiménez N, Viñas M, Guiu-Aragonés C, Bayona JM, Albaigés J, Solanas AM (2011) Polyphasic approach for assessing changes in an autochthonous marine bacterial community in the presence of Prestige fuel oil and its biodegradation potential. Applied Microbiology and Biotechnology. 91, 823–834.

Juncker A, Willenbrock H (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Science. 12, 1652–1662.

Junge K, Imhoff F, Staley T, Deming W (2002) Phylogenetic diversity of numerically important Arctic sea-ice bacteria cultured at subzero temperature. Microbial Ecology. 43, 315–328.

Karl DM (2002) Microbiological oceanography: hidden in a sea of microbes. Nature. 415, 590–591.

Kim OS, Cho YJ, Lee K, Yoon SH, Kim M, Na H, Park SC, Jeon YS, Lee JH, Yi H, Won S, Chun J (2012) Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. International Journal of Systematic and Evolutionary Microbiology. 62, 716–721.

King GM (2003) Molecular and culture-based analyses of aerobic carbon monoxide oxidizer diversity. Applied and Environmental Microbiology. 69, 7257–7265.

King GM, Weber CF (2007) Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. Nature Reviews Microbiology. 5, 107–118.

Kloesges T, Popa O, Martin W, Dagan T (2011) Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Molecular Biology and Evolution. 28, 1057–1074.

Koblížek M, Falkowski PG, Kolber ZS (2006) Diversity and distribution of photosynthetic bacteria in the Black Sea. Deep-Sea Research Part II-Topical Studies in Oceanography. 53, 1934–1944.

Kostka JE, Prakash O, Overholt WA, Green SJ, Freyer G, Canion A, Delgardio J, Norton N, Hazen TC, Huettel M (2011) Hydrocarbon-degrading bacteria and the bacterial community response in Gulf of Mexico beach sands impacted by the Deepwater Horizon oil spill. Applied and Environmental Microbiology. 77, 7962–7974.

Kotani T, Kawashima Y, Yurimoto H, Kato N, Sakai Y (2006) Gene structure and regulation of alkane monooxygenases in propane-utilizing *Mycobacterium* sp. TY-6 and *Pseudonocardia* sp. TY-7. Journal of Bioscience and Bioengineering. 102, 184–192.

Kotani T, Yurimoto H, Kato N, Sakai Y (2007) Novel acetone metabolism in a propane-utilizing bacterium, *Gordonia* sp. strain TY-5. Journal of Bacteriology. 189, 886–893.

Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular Biology and Evolution. 33, 1870–1874.

Kvint K, Nachin L, Diez A, Nyström T (2003) The bacterial universal stress protein: function and regulation. Current Opinion in Microbiology. 6, 140–145.

Lafay B, Ruimy R, de Traubenberg CR, Breittmayer V, Gauthier MJ, Christen R (1995) *Roseobacter algicola* sp. nov., a new marine bacterium isolated from the phycosphere of the toxin-producing dinoflagellate *Prorocentrum lima*. International Journal of Systematic Bacteriology. 45, 290–296.

Lai Q, Cao J, Yuan J, Li F, Shao Z (2014) *Celeribacter indicus* sp. nov., a polycyclic aromatic hydrocarbon-degrading bacterium from deep-sea sediment and reclassification of *Huaishuia halophila* as *Celeribacter halophilus* comb. nov. International Journal of Systematic and Evolutionary Microbiology. 64, 4160–4167.

Lamendella R, Strutt S, Borglin S, Chakraborty R, Tas N, Mason OU, Hultman J, Prestat E, Hazen TC, Jansson JK (2014) Assessment of the Deepwater Horizon oil spill impact on Gulf coast microbial communities. Frontiers in Microbiology. 5, 130.

Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW (2015) Insights from 20 years of bacterial genome sequencing. Functional and Integrative Genomics. 15, 141–161.

Lanfranconi MP, Bosch R, Nogales B (2010) Short-term changes in the composition of active marine bacterial assemblages in response to diesel oil pollution. Microbial Biotechnology. 3, 607–621.

Lang AS, Beatty JT (2000) Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of *Rhodobacter capsulatus*. Proceedings of the National Academy of Sciences of the United States of America. 97, 859–864.

Lang AS, Beatty JT (2001) The gene transfer agent of *Rhodobacter capsulatus* and "constitutive transduction" in prokaryotes. Archives of Microbiology. 175, 241–249.

Lang AS, Beatty JT (2006) Importance of widespread gene transfer agent genes in α-proteobacteria. Trends in Microbiology. 15, 54–62.

Lang AS, Taylor TA, Beatty TJ (2002) Evolutionary implications of phylogenetic analyses of the Gene Transfer Agent (GTA) of *Rhodobacter capsulatus*. Journal of Molecular Evolution. 55, 534–543.

Lang AS, Zhaxybayeva O, Beatty JT (2012) Gene transfer agents: phage-like elements of genetic exchange. Nature Reviews Microbiology. 10, 472–482.

Lenk S, Moraru C, Hahnke S, Arnds J, Richter M, Kube M, Reinhardt R, Brinkhoff T, Harder J, Amann R, Mußmann M (2012) *Roseobacter* clade bacteria are abundant in coastal sediments and encode a novel combination of sulfur oxidation genes. ISME Journal. 6, 2178–2187.

Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. BMC Bioinformatics. 23, 127–128.

Lidbury I, Kimberley G, Scanlan DJ, Murrell JC, Chen Y (2015) Comparative genomics and mutagenesis analyses of choline metabolism in the marine *Roseobacter* clade. Environmental Microbiology. 17, 5048-5062.

Liotenberg S, Steunou AS, Picaud M, Reiss-Husson F, Astier C, Ouchane S (2008) Organization and expression of photosynthesis genes and operons in anoxygenic photosynthetic proteobacteria. Environmental Microbiology. 10, 2267–2276.

Lucena T, Pujalte MJ, Ruvira MA, Garay E, Macián MC, Arahal DR (2012) *Tropicibacter multivorans* sp. nov., an aerobic alphaproteobacterium isolated from surface seawater. International Journal of Systematic and Evolutionary Microbiology. 62, 844–848.

Luo H, Csuros M, Hughes AL, Moran MA (2013) Evolution of divergent life history strategies in marine alphaproteobacteria. mBio. 4, e00373–13.

Luo H, Löytynoja A, Moran MA (2012) Genome content of uncultivated marine Roseobacters in the surface ocean. Environmental Microbiology. 14, 41–51.

Luo H, Moran MA (2014) Evolutionary ecology of the marine *Roseobacter* clade. Microbiology and Molecular Biology Reviews. 78, 573–587.

Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA (2014) Evolutionary analysis of a streamlined lineage of surface ocean Roseobacters. ISME Journal. 8, 1428–39.

Ma Y, Xiong H, Tang S, Yang Q, Li M (2009) Comparison of the community structure of planktonic bacteria in ballast water from entry ships and local sea water in Xiamen Port. Progress in Natural Science. 19, 946–953.

MacLellan SR, Smallbone LA, Sibley CD, Finan TM (2004) The expression of a novel antisense gene mediates incompatibility within the large *repABC* family of *α*-proteobacterial plasmids. Molecular Microbiology. 55, 611–623.

Magnusson LU, Farewell A, Nyström T (2005) ppGpp: a global regulator in *Escherichia coli*. Trends in Microbiology. 13, 236–242.

Magrane M, and UniProt Consortium (2011) UniProt knowledgebase: A hub of integrated protein data. Database. 2011, bar009.

Marrs B (1974) Genetic recombination in *Rhodopseudomonas capsulata*. Proceedings of the National Academy of Sciences of the United States of America. 71, 971–973.

Martens T, Gram L, Grossart HP, Kessler D, Müller R, Simon M, Wenzel SC, Brinkhoff T (2007) Bacteria of the *Roseobacter* clade show potential for secondary metabolite production. Microbial Ecology. 54, 31–42.

Mayali X, Franks PJS, Azam F (2008) Cultivation and ecosystem role of a marine *Roseobacter* clade-affiliated cluster bacterium. Applied and Environmental Microbiology. 74, 2595–2603.

McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH (2010) High frequency of horizontal gene transfer in the oceans. Science. 330, 50.

McKew BA, Coulon F, Osborn AM, Timmis KN, McGenity TJ (2007) Determining the identity and roles of oil-metabolizing marine bacteria from the Thames estuary, UK. Environmental Microbiology. 9, 165–176.

Mechold U, Potrykus K, Murphy H, Murakami KS, Cashel M (2013) Differential regulation by ppGpp versus pppGpp in *Escherichia coli*. Nucleic Acids Research. 41, 6175-6189.

Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 14, 60.

Miller MB, Bassler BL (2001) Quorum sensing in bacteria. Annual Review of Microbiology. 55, 165-199.

Miller TR, Belas R (2004) Dimethylsulfoniopropionate metabolism by *Pfiesteria*-associated *Roseobacter* spp. Applied and Environmental Microbiology. 70, 3383–3391.

Miller TR, Hnilicka K, Dziedzic A, Desplats P, Belas R (2004) Chemotaxis of *Silicibacter* sp. strain TM1040 toward dinoflagellate products. Applied Environmental Microbiology. 70, 4692–4701.

Misawa N, Smits THM, Beilen JBV, Harayama S (2004) Cloning and functional analysis of *alkB* genes in *Alcanivorax borkumensis* SK2. Environmental Microbiology. 6, 191–197.

Moran MA, Belas R, Schell MA, González JM, Sun F, Sun S, Binder BJ, Edmonds J, Ye W, Orcutt B, Howard EC, Meile C, Palefsky W, Goesmann A, Ren Q, Paulsen I, Ulrich LE, Thompson LS, Saunders E, Buchan A (2007) Ecological genomics of marine Roseobacters. Applied and Environmental Microbiology. 73, 4559–4569.

Moran MA, Buchan A, González JM, Heidelberg JF, Whitman WB, Kiene RP, Henriksen JR, King GM, Belas R, Fuqua C, Brinkac L, Lewis M, Johri S, Weaver B, Pai G, Eisen JA, Rahe E, Sheldon WM, Ye W, Miller TR, Carlton J, Rasko DA, Paulsen IT, Ren Q, Daugherty SC, Deboy RT, Dodson RJ, Durkin AS, Madupu R, Nelson WC, Sullivan SA, Rosovitz MJ, Haft DH, Selengut J, Ward N (2004) Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature. 432, 910–913.

Moran MA, González JM, Kiene RP (2003) Linking a bacterial taxon to sulfur cycling in the sea: studies of the marine *Roseobacter* group. Geomicrobiology Journal. 20, 375–388.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research. 35, W182–W185.

Nachin L, Nannmark U, Nyström T (2005) Differential roles of the universal stress proteins of Escherichia coli in oxidative stress resistance, adhesion, and motility. Journal of Bacteriology. 187, 6265–6272.

Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nature Genetics. 36, 760–766.

Nancy YY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FS (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. BMC Bioinformatics. 26, 1608–1615.

Nath S (2002) The molecular mechanism of ATP synthesis by F_1F_0-ATP synthase: A scrutiny of the major possibilities. Tools and Applications of Biochemical Engineering Science. Springer Berlin Heidelberg. pp. 65–98.

Newton RJ, Griffin LE, Bowles KM, Meile C, Gifford S, Givens CE, Howard EC, King E, Oakley CA, Reisch CR, Rinta-Kanto JM, Sharma S, Sun S, Varaljay V, Vila-Costa M, Westrich JR, Moran MA (2010) Genome characteristics of a generalist marine bacterial lineage. ISME Journal. 4, 784-798.

Nogales B (2010) Harbors and marinas. Handbook of hydrocarbon and lipid microbiology. Timmis, K.N. (ed.). Berlin, Germany: Springer-Verlag. pp. 2361–2367.

Nogales B, Aguiló-Ferretjans MM, Martín-Cardona C, Lalucat J, Bosch R (2007) Bacterial diversity, composition and dynamics in and around recreational coastal areas. Environmental Microbiology. 9, 1913–1929.

Nogales B, Lanfranconi MP, Piña-Villalonga JM, Bosch R (2011) Anthropogenic perturbations in marine microbial communities. FEMS Microbiology Reviews. 35, 275–298.

Novick RP (1987) Plasmid incompatibility. Microbiological Reviews. 51, 381–395.

Oakley AJ (2005) Glutathione transferases: new functions. Current Opinion in Structural Biology. 15, 716–723.

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature. 405, 299–304.

Okonechnikov K, Golosova O, Fursov M (2012) Unipro UGENE: a unified bioinformatics toolkit. BMC Bioinformatics. 28, 1166–1167.

Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences. 12, 357–358.

Paissé S, Coulon F, Goñi-Urriza M, Peperzak L, McGenity TJ, Duran R (2008) Structure of bacterial communities along a hydrocarbon contamination gradient in a coastal sediment. FEMS Microbiology Ecology. 66, 295–305.

Park SC, Pham BP, Van Duyet L, Jia B, Lee S, Yu R, Han SW, Yang JK, Hahm KS, Cheong GW (2008) Structural and functional characterization of osmotically inducible protein C (OsmC) from *Thermococcus kodakaraensis* KOD1. Biochimica et Biophysica Acta. 1784, 783–788.

Peng RH, Xiong AS, Xue Y, Fu XY, Gao F, Zhao W, Tian YS, Yao QH (2008). Microbial biodegradation of polyaromatic hydrocarbons. FEMS Microbiology Reviews. 32, 927–955.

Pérez-Oseguera Á, Cevallos MA (2013) RepA and RepB exert plasmid incompatibility repressing the transcription of the *repABC* operon. Plasmid. 70, 362–376.

Petersen J (2011) Phylogeny and compatibility: plasmid classification in the genomic era. Archives of Microbiology. 193: 313–321.

Petersen J, Brinkmann H, Berger M, Brinkhoff T, Päuker O, Pradella S (2011) Origin and evolution of a novel DnaA-Like plasmid replication type in *Rhodobacterales*. Molecular Biology and Evolution. 28, 1229–1240.

Petersen J, Brinkmann H, Bunk B, Michael V, Päuker O, Pradella S (2012) Think pink: photosynthesis, plasmids and the *Roseobacte*r clade. Environmental Microbiology. 14, 2661–2672.

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

Petersen J, Frank O, Göker M, Pradella S (2013) Extrachromosomal, extraordinary and essential -the plasmids of the *Roseobacter* clade. Applied Microbiology and Biotechnology. 97, 2805–2815.

Piña-Villalonga JM (2012) Diversidad e importancia ecológica del grupo Roseobacter en aguas costeras sometidas a impacto antropogénico. Universitat de les Illes Balears. Doctoral thesis.

Prabagaran SR, Manorama R, Delille D, Shivaji S (2007) Predominance of *Roseobacter*, *Sulfitobacter*, *Glaciecola* and *Psychrobacter* in seawater collected off Ushuaia, Argentina, Sub-Antarctica. FEMS Microbiology Ecology. 59, 342–355.

Pukall R, Buntefuss D, Frühling A, Rohde M, Kroppenstedt RM, Burghardt J, Lebaron P, Bernard L, Stackebrandt E (1999) *Sulfitobacter mediterraneus* sp. nov., a new sulfite-oxidizing member of the α-Proteobacteria. International Journal of Systematic and Evolutionary Microbiology. 49, 513–519.

Raetz CRH, Guan Z, Ingram BO, Six DA, Song F, Wang X, Zhao J (2009) Discovery of new biosynthetic pathways: the lipid A story. Journal of Lipid Research. 50, S103–S108.

Ramírez-Romero MA, Soberón N, Pérez-Oseguera A, Téllez-Sosa J, Cevallos MA (2000). Structural elements required for replication and incompatibility of the *Rhizobium etli* symbiotic plasmid. Journal of Bacteriology. 182, 3117–24.

Ramírez-Romero, MA, Téllez-Sosa J, Barrios H, Pérez-Oseguera A, Rosas V, Cevallos MA (2001) RepA negatively autoregulates the transcription of the *repABC* operon of the *Rhizobium etli* symbiotic plasmid basic replicon. Molecular Microbiology. 42, 195–204.

Ramos JL, Sol Cuenca M, Molina-Santiago C, Segura A, Duque E, Gómez-García MR, Udaondo Z, Roca A (2015) Mechanisms of solvent resistance mediated by interplay of cellular factors in *Pseudomonas putida*. FEMS Microbiology Reviews. 39, 555–566.

Ramos JL, Duque E, Reyes F, Ferna A (1997) Mechanisms for solvent tolerance in bacteria. Journal of Biological Chemistry. 272, 3887–3890.

Rao D, Webb JS, Holmström C, Case R, Low A, Steinberg P, Kjelleberg S (2007) Low densities of epiphytic bacteria from the marine alga *Ulva australis* inhibit settlement of fouling organisms. Applied and Environmental Microbiology. 73, 7844–7852.

Reisch CR, Crabb WM, Gifford SM, Teng Q, Stoudemayer MJ, Moran MA, Whitman WB (2013) Metabolism of dimethylsulphoniopropionate by *Ruegeria pomeroyi* DSS-3^T. Molecular Microbiology. 89, 774–791.

Reisch CR, Moran MA, Whitman WB (2011a) Bacterial catabolism of dimethylsulfoniopropionate (DMSP). Frontiers in Microbiology. 2, 1–12.

Reisch CR, Stoudemayer MJ, Varaljay VA, Amster IJ, Moran MA, Whitman WB (2011b) Novel pathway for assimilation of dimethylsulphoniopropionate widespread in marine bacteria. Nature. 473, 208–211.

Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proceedings of the National Academy of Sciences of the United States of America. 106, 19126–19131.

Rojo F (2009) Degradation of alkanes by bacteria. Environmental Microbiology. 11, 2477–2490.

Roma-Rodrigues C, Santos PM, Benndorf D, Rapp E, Sá-Correia I (2010) Response of *Pseudomonas putida* KT2440 to phenol at the level of membrane proteome. Journal of Proteomics. 73, 1461–1478.

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

Ruiz JA, López NI, Fernández RO, Méndez BS (2001) Polyhydroxyalkanoate degradation is associated with nucleotide accumulation and enhances stress resistance and survival of *Pseudomonas oleovorans* in natural water microcosms. Applied and Environmental Microbiology. 67, 225–230.

Saier MH, Reddy VS, Tamang DG, Västermark Å (2014) The transporter classification database. Nucleic Acids Research. 42, D251-D258.

Sanni GO, Coulon F, McGenity TJ (2015) Dynamics and distribution of bacterial and archaeal communities in oil-contaminated temperate coastal mudflat mesocosms. Environmental Science and Pollution Research. 22, 15230–15247.

Sauret C, Böttjer D, Talarmin A, Guigue C, Conan P, Pujo-Pay M, Ghiglione JF (2015) Top-down control of diesel-degrading prokaryotic communities. Microbial Ecology. 70, 445–58.

Schaefer AL, Taylor TA, Beatty JT, Greenberg EP (2002) Long-chain acyl-homoserine lactone quorum-sensing regulation of *Rhodobacter capsulatus* gene transfer agent production. Journal of Bacteriology. 184, 6515–21.

Schauer M, Massana R, Pedrôs-Alio C (2000) Spatial differences in bacterioplankton composition along the Catalan coast (NW Mediterranean) assessed by molecular fingerprinting. FEMS Microbiology Ecology. 33, 51–59.

Schneiker S, Martins dos Santos VA, Bartels D, Bekel T, Brecht M, Buhrmester J, Chernikova TN, Denaro R, Ferrer M, Gertler C, Goesmann A, Golyshina OV, Kaminski F, Khachane AN, Lang S, Linke B, McHardy AC, Meyer F, Nechitaylo T, Pühler A, Regenhardt D, Rupp O, Sabirova JS,

Selbitschka W, Yakimov MM, Timmis KN, Vorhölter FJ, Weidner S, Kaiser O, Golyshin PN (2006) Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium *Alcanivorax borkumensis*. Nature Biotechnology. 24, 997–1004.

Schwarzer D, Finking R, Marahiel MA (2003) Nonribosomal peptides: from genes to products. Natural Product Reports. 20, 275–287.

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. BMC Bioinformatics. 30, 2068-2069.

Segura A, Molina L, Ramos JL (2014) Plasmid-mediated tolerance toward environmental pollutants. Microbiology Spectrum. 2, 1–23.

Segura A, Godoy P, Van Dillewijn P, Hurtado A, Arroyo N, Santacruz S, Ramos JL (2005) Proteomic analysis reveals the participation of energy- and stress-related proteins in the response of *Pseudomonas putida* DOT-T1E to toluene. Journal of Bacteriology. 187, 5937–5945.

Seymour JR, Ahmed T, Stocker R (2009) Bacterial chemotaxis towards the extracellular products of the toxic phytoplankton *Heterosigma akashiwo*. Journal of Plankton Research. 31, 1557–1561.

Shiba T (1991) *Roseobacter litoralis* gen. nov., sp. nov., and *Roseobacter denitrificans* sp. nov., aerobic pink-pigmented bacteria which contain bacteriochlorophyll *a*. Systematic and Applied Microbiology. 14, 140–145.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 7, 539.

Siguier P (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Research. 34, D32–D36.

Siguier P, Gourbeyre E, Chandler M (2014) Bacterial insertion sequences: Their genomic impact and diversity. FEMS Microbiology Reviews. 38, 865–891.

Sikkema J, de Bont JAM, Poolman B (1995) Mechanisms of membrane toxicity of hydrocarbons. Microbiological Reviews. 59, 201–22.

Singh VK, Utaida S, Jackson LS, Jayaswal RK, Wilkinson BJ, Chamberlain NR (2007) Role for *dnaK* locus in tolerance of multiple stresses in *Staphylococcus aureus*. Microbiology. 153, 3162–3173.

Skerker JM, Shapiro L (2000) Identification and cell cycle control of a novel pilus system in *Caulobacter crescentus*. EMBO Journal. 19, 3223–3234.

Slightom RN, Buchan A (2009) Surface colonization by marine roseobacters: integrating genotype and phenotype. Applied and Environmental Microbiology. 75, 6027–6037.

Soberón N, Venkova-Canova T, Ramírez-Romero MA, Téllez-Sosa J, Cevallos MA (2004) Incompatibility and the partitioning site of the *repABC* basic replicon of the symbiotic plasmid from *Rhizobium etli*. Plasmid. 51, 203–216.

Solioz M, Yen HC, Marris B (1975) Release and uptake of gene transfer agent by *Rhodopseudomonas capsulata*. Journal of Bacteriology. 123, 651–657.

Srivatsan A, Wang JD (2008) Control of bacterial transcription, translation and replication by (p)ppGpp. Current Opinion in Microbiology. 11, 100–105.

Stanton TB (2007) Prophage-like gene transfer agents—novel mechanisms of gene exchange for *Methanococcus*, *Desulfovibrio*, *Brachyspira*, and *Rhodobacter* species. Anaerobe. 13, 43–49.

Starosta AL, Lassak J, Jung K, Wilson DN (2014) The bacterial translation stress response. FEMS Microbiology Reviews. 38, 1172–1201.

Staunton J, Weissman KJ (2001) Polyketide biosynthesis: a millennium review. Natural Product Reports. 18, 380–416.

Suárez-Suárez, LY (2013) Caracterización genómica y fisiológica de la capacidad degradadora de hidrocarburos aromáticos de *Citreicella aestuarii* 357. Universitat de les Illes Balears. Doctoral thesis.

Suárez-Suárez LY, Brunet-Galmes I, Piña-Villalonga JM, Christie-Oleza JA, Peña A, Bennasar A, Armengaud J, Nogales B, Bosch R (2012) Draft genome sequence of *Citreicella aestuarii* strain 357, a member of the *Roseobacter* clade isolated without xenobiotic pressure from a petroleum-polluted beach. Journal of Bacteriology. 194, 5464–5465.

Suyama T, Shigematsu T, Suzuki T, Tokiwa Y, Kanagawa T, Nagashima KV, Hanada S (2002) Photosynthetic apparatus in *Roseateles depolymerans* 61A is transcriptionally induced by carbon limitation. Applied Environmental Microbiology. 68, 1665–1673.

Suyama T, Shigematsu T, Takaichi S, Nodasaka Y, Fujikawa S, Hosoya H, Tokiwa Y, Kanagawa T, Hanada S (1999) *Roseateles depolymerans* gen. nov., sp. nov., a new bacteriochlorophyll *a*-containing obligate aerobe belonging to the β-subclass of the *Proteobacteria*. International Journal of Systematic and Evolutionary Microbiology. 49, 449–457.

Suzuki MT, Preston CM, Chavez FP, DeLong EF (2001) Quantitative mapping of bacterioplankton populations in seawater: field tests across an upwelling plume in Monterey Bay. Aquatic Microbial Ecology. 24, 117–127.

Swingley WD, Sadekar S, Mastrian SD, Matthies HJ, Hao J, Ramos H, Acharya CR, Conrad AL, Taylor HL, Dejesa LC, Shah MK, O'huallachain ME, Lince MT, Blankenship RE, Beatty JT,

Touchman JW (2007) The complete genome sequence of *Roseobacter denitrificans* reveals a mixotrophic rather than photosynthetic metabolism. Journal of Bacteriology. 189, 683–690.

Syvanen M (1994) Horizontal gene transfer: evidence and possible consequences. Annual Review of Genetics. 28, 237–261.

Tabata S, Hooykaas PJ, Oka A (1989) Sequence determination and characterization of the replicator region in the tumor-inducing plasmid pTiB6S3. Journal of Bacteriology. 171, 1665–1672.

Tabita FR, Hanson TE, Li H, Satagopan S, Singh J, Chan S (2007) Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. Microbiology and Molecular Biology Reviews. 71, 576–599.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology. 56, 564–577.

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research. 28, 33–36.

Teter SA, Houry WA, Ang D, Tradler T, Rockabrand D, Fischer G, Blum P, Georgopoulos C, Hartl FU (1999) Polypeptide Flux through Bacterial Hsp70 : DnaK Cooperates with Trigger Factor in Chaperoning Nascent Chains. Cell. 97, 755-765.

Thomas CM (2000) Paradigms of plasmid organization. Molecular Microbiology. 37, 485–491.

Tomich M, Planet PJ, Figurski DH (2007) The *tad* locus: postcards from the widespread colonization island. Nature Reviews Microbiology. 5, 363–375.

Touchon M, Rocha EPC (2007) Causes of insertion sequences abundance in prokaryotic genomes. Molecular Biology and Evolution. 24, 969–981.

Vattanaviboon P, Seeanukun C, Whangsuk W, Utamapongchai S, Mongkolsuk S (2005) Important role for methionine sulfoxide reductase in the oxidative stress response of *Xanthomonas campestris* pv . phaseoli. Journal of Bacteriology. 187, 5831–5836.

Venkova-Canova T, Soberón NE, Ramírez-Romero MA, Cevallos MA (2004) Two discrete elements are required for the replication of a *repABC* plasmid: an antisense RNA and a stem-loop structure. Molecular Microbiology. 54, 1431–1444.

Vollmers J, Voget S, Dietrich S, Gollnow K, Smits M, Meyer K, Brinkhoff T, Simon M, Daniel R (2013) Poles apart: Arctic and Antarctic *Octadecabacter* strains share high genome plasticity and a new type of xanthorhodopsin. PLoS One. 8, e63422.

Wagner A (2006) Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. Molecular Biology and Evolution. 23, 723–733.

Wagner-Döbler I, Ballhausen B, Berger M, Brinkhoff T, Buchholz I, Bunk B, Cypionka H, Daniel R, Drepper T, Gerdts G, Hahnke S, Han C, Jahn D, Kalhoefer D, Kiss H, Klenk HP, Kyrpides N, Liebl W, Liesegang H, Meincke L, Pati A, Petersen J, Piekarski T, Pommerenke C, Pradella S, Pukall R, Rabus R, Stackebrandt E, Thole S, Thompson L, Tielen P, Tomasch J, von Jan M, Wanphrut N, Wichels A, Zech H, Simon M (2010) The complete genome sequence of the algal symbiont *Dinoroseobacter shibae*: a hitchhiker's guide to life in the sea. ISME Journal. 4, 61–77.

Wagner-Döbler I, Biebl H (2006) Environmental biology of the marine *Roseobacter* lineage. Annual Review of Microbiology. 60, 255–280.

Wagner-Döbler I, Rheims H, Felske A, El-Ghezal A, Flade-Schröder D, Laatsch H, Lang S, Pukall R, Tindall BJ (2004). *Oceanibulbus indolifex* gen. nov., sp. nov., a North Sea alphaproteobacterium that produces bioactive metabolites. International Journal of Systematic and Evolutionary Microbiology. 54, 1177–1184.

Wang B, Tan T, Shao Z (2009) *Roseovarius pacificus* sp. nov., isolated from deep-sea sediment. International Journal of Systematic and Evolutionary Microbiology. 59, 1116–1121.

Watkinson RJ, Morgan P (1990) Physiology of aliphatic hydrocarbon-degrading microorganisms. Biodegradation. 1, 79–92.

Weissbach H, Resnick L, Brot N (2005) Methionine sulfoxide reductases: history and cellular role in protecting against oxidative damage. Biochimica et Biophysica Acta. 1703, 203–212.

Wendrich TM, Blaha G, Wilson DN, Marahiel MA, Nierhaus KH (2002) Dissection of the mechanism for the stringent factor RelA. Molecular Cell. 10, 779–788.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler, GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 35, D5–D12.

Wijte D, Van Baar BLM, Heck AJR, Altelaar AFM (2011) Probing the proteome response to toluene exposure in the solvent tolerant *Pseudomonas putida* S12. Journal of Proteome Research. 10, 394–403.

Williams PA, Sayers JR (1994) The evolution of pathways for aromatic hydrocarbon oxidation in *Pseudomonas*. Biodegradation. 5, 195–217.

Yakimov MM, Timmis KN, Golyshin PN (2007) Obligate oil-degrading marine bacteria. Current Opinion in Biotechnology. 18, 257–266.

Yoch DC (2002) Dimethylsulfoniopropionate: its sources, role in the marine food web, and biological degradation to dimethylsulfide. Applied and Environmental Microbiology. 68, 5804–5815.

Zaar A, Gescher J, Eisenreich W, Bacher A, Fuchs G (2004) New enzymes involved in aerobic benzoate metabolism in *Azoarcus evansii*. Molecular Microbiology. 54, 223–238.

Zan J, Liu Y, Fuqua C, Hill RT (2014) Acyl-homoserine lactone quorum sensing in the *Roseobacter* clade. International Journal of Molecular Sciences. 15, 654–669.

Zhang DF, Li H, Lin XM, Wang SY, Peng XX (2011) Characterization of outer membrane proteins of *Escherichia coli* in response to phenol stress. Current Microbiology. 62, 777-783.

Zhang R, Lau SCK, Ki JS, Thiyagarajan V, Qian PY (2009) Response of bacterioplankton community structures to hydrological conditions and anthropogenic pollution in contrasting subtropical environments. FEMS Microbiology Ecology. 69, 449–460.

Zhang R, Liu B, Lau SCK, Ki JS, Qian PY (2007) Particle-attached and free-living bacterial communities in a contrasting marine environment: Victoria Harbor, Hong Kong. FEMS Microbiology Ecology. 61, 496–508.

Zhang Y, Sun Y, Jiao N, Stepanauskas R, Luo H (2016) Ecological genomics of the uncultivated marine *Roseobacter* lineage CHAB-I-5. Applied and Environmental Microbiology. 82, 2100–2111.

Zhang W, Wang Y, Lee OO, Tian R, Cao H, Gao Z, Li Y, Yu L, Xu Y, Qian PY (2013) Adaptation of intertidal biofilm communities is driven by metal ion and oxidative stresses. Scientific Reports. 3, 3180.

Zheng Q, Zhang R, Koblížek M, Boldareva EN, Yurkov V, Yan S, Jiao N (2011) Diverse arrangement of photosynthetic gene clusters in aerobic anoxygenic phototrophic bacteria. PLoS One. 6, 1–7.

Zolkiewski M (1999) ClpB cooperates with DnaK, DnaJ , and GrpE in suppressing protein aggregation. Journal of Biological Chemistry. 274, 28083–28086.

Zong R, Jiao N (2012) Proteomic responses of *Roseobacter litoralis* OCh149 to starvation and light regimen. Microbes and Environments. 27, 430–442.

Zubkov MV, Fuchs BM, Archer SD, Kiene RP, Amann R, Burkill PH (2001) Linking the composition of bacterioplankton to rapid turnover of dissolved dimethylsulphoniopropionate in an algal bloom in the North Sea. Environmental Microbiology. 3, 304–311.

ANNEXES

ANNEX 1. SUPPLEMENTARY FIGURES

LIST OF SUPPLEMENTARY FIGURES

All supplementary Figures are attached together with this document.

Figure 1S.1. Changes in the core proteome and panproteome of 96 members of Roseobacter lineage as new genomes were added.

Figure 1S.2. Distribution of COGs of panproteome, core proteome, accessory proteome and singletons in functional categories.

Figure 1S.3. Percentages of assignation in COGs functional categories of the core proteome, accessory proteome and singletons.

Figure 1S.4. Phylogenomic distances among the genomes of each phylogenomic group.

Figure 2S.1. Functional assignation of the COGs of detected proteins distributed according their proteomic category.

Figure 2S.2. PCA of normalized abundances (to 100) of the detected proteins clustered under the criterion of 90 % of identity in, at least, the 95 % of its amino acid sequence of the seven Roseobacter isolates.

Figure 2S.3. PCA of normalized abundances (Progenesis) of the detected proteins clustered under the criterion of 90 % of identity in, at least, the 95 % of its amino acid sequence of *Sulfitobacter mediterraneus* 1FIGIMAR09.

Figure 2S.4. PCA of normalized abundances (Progenesis) of the detected proteins clustered under the criterion of 90 % of identity in, at least, the 95 % of its amino acid sequence of *Sulfitobacter* sp. 3SOLIMAR09.

Figure 2S.5. PCA of normalized abundances (Progenesis) of the detected proteins clustered under the criterion of 90 % of identity in, at least, the 95 % of its amino acid sequence of *Loktanella* sp. 1ANDIMAR09.

Figure 2S.6. PCA of normalized abundances (Progenesis) of the detected proteins clustered under the criterion of 90 % of identity in, at least, the 95 % of its amino acid sequence of *Loktanella* sp. 3ANDIMAR09.

Figure 2S.7. PCA of normalized abundances (Progenesis) of the detected proteins clustered under the criterion of 90 % of identity in, at least, the 95 % of its amino acid sequence of *Thalassobacter stenotrophicus* 16PALIMAR09.

Figure 2S.8. PCA of normalized abundances (Progenesis) of the detected proteins clustered under the criterion of 90 % of identity in, at least, the 95 % of its amino acid sequence of *Thalassobacter stenotrophicus* 1CONIMAR09.

Figure 2S.9. Percentages of normalized abundances of each COG functional category of the seven isolates in control condition.

Figure 2S.10. Percentages of normalized abundances of each COG functional category of the seven isolates in treated condition (diesel oil).

Figure 2S.11. Comparison of the percentage abundance of detected proteins involved in the response to oxidative stress that belonged to (O) “post-translational modification, protein turnover and chaperone” and (P) “inorganic transport and metabolism” COG functional categories of all isolates.

Figure 2S.12. Comparison of the percentage abundance of detected chaperones that belonged to (O) “post-translational modification, protein turnover and chaperones” COG functional category of all isolates.

Figure 3S.1. Synteny of the GTA module among the 85 Roseobacter genomes in which it was detected.

Figure 3S.2. Phylogenetic tree of RepA.

Figure 3S.3. Phylogenetic tree of RepB.

Figure 3S.4. Shared TnpA COGs among roseobacters of phylogenomic group G1.

Figure 3S.5. Shared TnpA COGs among roseobacters of phylogenomic group G2.

Figure 3S.6. Shared TnpA COGs among roseobacters of phylogenomic group G3.

Figure 3S.7. Shared TnpA COGs among roseobacters of phylogenomic group G4.

Figure 3S.8. Shared TnpA COGs among roseobacters of phylogenomic group G5.

Figure 3S.9. Shared TnpA COGs among roseobacters that were isolated from metazoan (A) and dinoflagellates or other algae (B).

ANNEX 2. SUPPLEMENTARY TABLES

LIST OF SUPPLEMENTARY TABLES

All supplementary Tables are attached together with this document.

Table 1S.1. Genome features of 96 Roseobacter genomes analyzed in this study.

Table 1S.2. Average Nucleotide Identity (ANIb and ANIm) of 96 Roseobacter genomes.

Table 1S.3. Genome annotation with Prokka of the 96 Roseobacter genomes.

Table 1S.4. Clustering of the annotated proteins of 96 Roseobacter genomes under the criteria of 50 % of identity in at least, the 50 % of its amino acid sequence.

Table 1S.5. Reference protein of each cluster of the conserved proteins of the core proteome that were used for making the concatenation for the Roseobacter phylogeny.

Table 1S.6. Distance matrix of 96 genomes of Roseobacter lineage calculated by Kimura distance method using PHYLIP.

Table 1S.7. Clusters of Orthologous Groups of proteins (COGs) whose single phylogeny maintained at least one or a part of a phylogenomic group.

Table 1S.8. Clusters of Orthologous Groups of proteins that were found in single-copy of the core proteome of 96 Roseobacter genomes that were selected for making the phylogeny of the 165 Roseobacter genomes.

Table 1S.9. Genome features of 69 new Roseobacter genomes included later in the study.

Table 1S.10. Genome annotation with Prokka of 69 new Roseobacter genomes.

Table 1S.11. Protein codes and sequences of each of the 14 Clusters of Orthologous Groups of proteins of the 165 Roseobacter genomes annotated with Prokka program used for making the phylogeny of the 165 members of Roseobacter lineage.

Table 1S.12. Detection of biochemically-relevant proteins in the nine Roseobacter isolates.

Table 1S.13. Detection of aliphatic and monoaromatic degradation proteins in the nine Roseobacter isolates.

Table 2S.1. List of detected proteins of the seven Roseobacter isolates considered as significant by the analysis with Progenesis.

Table 2S.2. Classification of annotated genome and detected proteome in COG functional categories.

Table 2S.3. List of detected proteins that had statistical relevance by Progenesis and by PCA of the seven Roseobacter isolates.

Table 2S.4. Detected proteins that could have a role in the degradation of diesel oil in the seven Roseobacter isolates.

Table 2S.5. Lipid A export ATP-binding/permease protein MsbA and hypothetical transmembrane protein (DUF2852) detected in the seven Roseobacter isolates.

Table 2S.6. Detected transporters and its classification according to the Transporter Classification Database (TCDB).

Table 2S.7. Percentages of abundance of the detected transporters classified in superfamilies according to the TCDB.

Table 2S.8. Classification in families of the detected transporters that could not be assigned to any superfamily according to the TCBD.

Table 2S.9. Detected proteins related to the pilus assembly in the seven Roseobacter isolates.

Table 3S.1. GTA module of each genome ordered by phylogenomics groups.

Table 3S.2. Identification of five genic products of type IV secretion system (T4SS) of RepABC putative plasmids.

Table 3S.3. Identity and location of the RepA, RepB, RepC (without RepAB) and Dna-like replicases through the 96 Roseobacter genomes.

Table 3S.4. Total number of detected plausible TnpAs and its distribution in IS families for each Roseobacter genome.

Table 3S.5. Clustering of detected TnpAs of 96 Roseobacter genomes under the criteria of 90 % of identity in at least the 95 % of amino acid sequence.

Table 3S.6. Shared COGs of plausible TnpAs clustered under the criterion of the 90 % of identity in at least the 95 % of amino acid sequence among all 96 Roseobacter.

Table 3S.7. COGs constituted for more than 20 TnpAs that belonged exclusively to one genome.

Table 3S.8. Number of shared COGs of TnpAs between genomes of the same phylogenomic group.

Table 3S.9. Overview of the detection of HGT mechanisms in 96 members of Roseobacter lineage.