Marine Bacterial Genomics

an ocean of opportunity

Machado, Henrique; Gram, Lone

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Marine bacterial genomics: an ocean of opportunity

Henrique Machado

PhD Thesis

June 2016

DTU Systems Biology
Department of Systems Biology

Pedras no caminho?

Guardo todas, um dia vou construir um castelo…

(Stones in my way? I save every single one, one day I´ll build a castle)

Fernando Pessoa

Preface

This thesis was written as a partial fulfillment of the requirements to obtain a PhD degree from the Technical University of Denmark (DTU). The work here presented was carried out under the supervision of Professor Lone Gram, from May 2013 to June 2016 at the Department of Systems Biology and at the Novo Nordisk Foundation - Center for Biosustainability, both at DTU. Part of the work was also performed at Evolva S/A Copenhagen (September to December 2014).

The PhD student Henrique Machado was supported by a PhD grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7-People-2012-ITN, under grant agreement No. 317058, "BACTORY".

The work resulted in the preparation of ten research articles/manuscripts, six of them included in this thesis, and two communications not included in this thesis.

Acknowledgments

I would like to thank Lone Gram for giving me the opportunity to do a PhD under her supervision and for always believing and trusting in me – Thank you!

Thank you to the whole Bacterial Ecophysiology and Biotechnology group (past and present members) for a great time. Special thank you to Paul, Sonia, Virginie, Eva and Jette for the talks, brainstormings, motivation and help! =)

Thank you to all the BacTory students and supervisors for joining this 3-year journey, and also to Søren Molin and Mette Munk for the time and effort they put into the program. A special thank you to the BacTory Summer School organizing committee (Patri, Sofie and Klara) for all the fun we had doing it! =P

A warm hearted thank you to Mafalda and Isotta for being endless friends and travelling companions! E a ti Mafalda, não só por isso, mas por sempre acreditares em mim, por me ajudares profissionalmente e pessoalmente; é sem dúvida inacreditável estarmos aqui depois de 10 anos, OBRIGADO!

Above all, thank you to my fiancé Klara for the love, patience (most of the times), and positivism – You ROCK!

Thank you to my families (Machado, Ramalho and Bojanovič [to be]), friends and everyone who made these three years in Copenhagen awesome, thank you for being!!!

Abstract

For decades, terrestrial microorganisms have been used as sources of countless enzymes and chemical compounds that have been produced by pharmaceutical and biotech companies and used by mankind. There is a need for new chemical compounds, including antibiotics, new enzymatic activities and new microorganisms to be used as cell factories for production. Therefore exploitation of new microbial niches and use of different strategies is an opportunity to boost discoveries.

Even though scientists have started to explore several habitats other than the terrestrial ones, the marine environment stands out as a hitherto under-explored niche. This thesis work uses high-throughput sequencing technologies on a collection of marine bacteria established during the Galathea 3 expedition, with the purpose of unraveling new biodiversity and new bioactivities. Several tools were used for genomic analysis in order to better understand the potential harbored in marine bacteria.

The work presented makes use of whole genome sequencing of marine bacteria to prove that the genetic repertoire for secondary metabolite production harbored in these bacteria is far larger than anticipated; to identify and develop a new phylogenetic marker for the identification of members of the *Vibrionaceae* family, which led to the identification of two new species using this straightforward pipeline; to discovery of new cytochrome P450 enzymes to be used in

biotechnology; and to a thorough study of the marine genus *Photobacterium,* by means of comparative genomics.

In conclusion, this PhD thesis has contributed to our understanding of the marine microbial environment by studying genomic information of several marine bacteria, expanding the number of marine species taxonomically described, providing identification tools for further marine species documentation and pointing to these organisms as a very promising resource for further bioprospecting.

Dansk resumé

Mikroorganismer isoleret fra jord har i årtier været en kilde til nye enzymer og kemiske forbindelser, der er blevet produceret af farmaceutiske og bioteknologiske virksomheder og anvendt af mennesker. Der er behov for nye kemiske forbindelser, herunder antibiotika, nye enzymatiske aktiviteter og nye mikroorganismer, der skal anvendes som cellefabrikker til produktion. Vi må derfor tænke i udnyttelse af nye mikrobielle nicher og anvendelse af forskellige strategier for at sikre en pipeline af nye opdagelser. Forskere er begyndt at udforske andre nicher end jordbunden, og her skiller havmiljøet sig ud som en hidtil under-udforsket niche. Formålet med dette phd-arbejde er at finde ny biodiversitet og ny bioaktivitet i marine bakterier. I arbejdet er anvendt analyse af genomer og gen-sekvenser på en samling af marine bakterier, der er etableret under Galathea 3 ekspeditionen.

Det konkluderes, ved analyse af hel-genom-sekvenser, at det genetiske repertoire for sekundær metabolit-produktion i disse bakterier er langt større end forventet; dvs som hidtil udforsket med klassisk bio-assay-guided fraktionering. Desuden er genomsekvenserne anvendt til at udvikle en ny fylogenetisk markør (*fur*-genet) til identifikation af medlemmer af *Vibrionaceae* familien, hvilket har ført til identifikation af to nye arter. Genomerne er minet for potentielt bioaktive stoffer, her nye cytokrom P450 enzymer, der kan anvendes i bioteknologi, og de er klonet og udtrykt i både *E. coli* og gær. Endelig er komparativ

analyse af genomer-sekvenserne anvendt til en større fylogenetisk og genetisk analyse af den marine slægten *Photobacterium*.

Samlet har dette ph.d.-arbejde bidraget til vores forståelse af marine mikroorganismer og deres diversitet, til udvikling af nye identifikations-redskaber samt peget på disse organismer som en meget lovende ressource for yderligere bioprospektering.

Resumo em Português

Durante décadas, os microrganismos terrestres foram utilizados como fontes de inúmeras enzimas e compostos químicos que têm sido produzidos por empresas farmacêuticas e de biotecnologia e usadas pela humanidade. Existe uma necessidade de novos compostos químicos, incluindo antibióticos, novas actividades enzimáticas e de novos microrganismos para serem utilizados em produção biotecnológica. Por conseguinte, a exploração de novos nichos microbianos e a utilização de diferentes estratégias é uma oportunidade para intensificar estas descobertas.

Embora os cientistas tenham começado a explorar vários outros nichos para além do terrestre, o ambiente marinho destaca-se como um nicho ainda pouco explorado. Este trabalho utiliza as recentes tecnologias de sequenciação e aplica-as a uma colecção de bactérias marinhas, colecção esta estabelecidos durante a expedição Galathea 3, com o objetivo de desvendar nova biodiversidade e novas bioatividades. Foram utilizadas várias ferramentas de análise de genomas, a fim de melhor compreender o potencial presente em bactérias marinhas.

O trabalho apresentado faz uso da sequenciação do genoma de bactérias marinhas para: provar que o repertório genético para a produção de metabolitos secundários presente nestas bactérias é muito maior do que o previsto; identificar e desenvolver um novo marcador filogenético para a identificação de membros da família

Vibrionaceae, que levou também à identificação de duas novas espécies; a descoberta de novas enzimas do tipo citocromo P450 para ser utilizados em biotecnologia; e para um estudo aprofundado do género *Photobacterium*, por meio de comparação de genomas.

Em conclusão, esta tese de doutoramento contribuiu para compreender melhor o ambiente microbiano marinho, através do estudo da informação genética de várias bactérias marinhas, ampliou o número de espécies marinhas taxonomicamente descritas, proporcionou o desenvolvimento de ferramentas para posterior identificação e classificação de espécies marinhas e comprovou que o estudo deste tipo de microrganismos poderá ser uma mais valia na descoberta de novos compostos, enzimas, etc.

List of Publications

Included in this thesis

Machado H., Sonnenschein E.C., Melchiorsen J. and Gram L. (2015) Genome mining reveals unlocked bioactive potential of marine Gram-negative bacteria, BMC Genomics (2015) 16:158.

Machado H. and Gram L. (2015) The *fur* gene as a new phylogenetic marker in *Vibrionaceae* species identification, Appl Environ Microbiol 81(8):2745-52

Machado H., Cardoso J., Giubergia S., Rapacki K. and Gram L. (2016) Furios: a web-based tool for identification of *Vibrionaceae* species using the *fur* gene, BMC Microbiology (submitted)

Machado H., Giubergia S., Mateiu R.V. and Gram L. (2015) *Photobacterium galatheae* sp. nov., a bioactive bacterium isolated from a mussel in the Solomon Sea, Int J System Evol Microbiol 65(12):4503-7

Cavaleiro A.M.*, **Machado H.***, Casado C., Kol S., Gram L. and Nørholm M.H.H. (2016) Marine bacterial cythochrome P450s and their potential in biotech, PlosOne (submitted)
*equal contribution

Machado H. and Gram L. (2016) Comparative genomics reveals evidence of high genomic exchange driving *Photobacterium* genus evolution (in preparation)

Not included in this thesis

Machado H.*, Cavaleiro A.M.*, D'Arrigo I., Bojanovič K., Nørholm M.H.H. and Gram L. (2016) Exploring marine environments to unravel tolerance mechanisms to relevant compounds and discover new microbial cell factories (in preparation)
*equal contribution

Machado H.*, Cavaleiro A.M.*, Nørholm M.H.H. and Gram L. (2016) Draft genome sequence of *Bacillus* sp. F2HM and *Pseudomonas* sp. D1HM, highly acid tolerant strains (in preparation)
*equal contribution

Machado H.*, Vynne N.G.*, Christiansen G. and Gram L. (2016) Reclassification of *Alteromonas fuliginea* (Romanenko *et al.* 1994) as *Pseudoalteromonas fuliginea* comb. nov. and emended description, Int J System Evol Microbiol (accepted)
*equal contribution

Giubergia S., **Machado H.**, Mateiu R.V. and Gram L. (2016) *Vibrio galatheae* sp. nov., a novel member of the *Vibrionaceae* family isolated from the Solomon Sea, Int J System Evol Microbiol 66(1):347-52

Rasmussen B.B., Nielsen K.F., **Machado H.**, Melchiorsen J., Gram L. and Sonnenschein E.C. (2014) Global and phylogenetic distribution of quorum sensing signals, acyl homoserine lactones in the family of *Vibrionaceae*, Mar Drugs 12(11):5527-46

Machado H., Månsson M. and Gram L. (2014) Draft genome sequence of *Photobacterium halotolerans* S2753, producer of bioactive secondary metabolites, Genome Announc 2(3):e00535-14

Table of Contents

1. Introduction

Terrestrial microorganisms have for decades been successfully used as a source of innumerable enzymes and chemical compounds, which have been put to use by pharmaceutical and biotech companies. Exploitation of new niches and use of different research strategies is required to boost the discovery of new chemical compounds, such as antibiotics, new enzymatic activities and new microorganisms to be used as cell factories for production.

Approximately seventy percent of planet Earth is covered in Oceans. Hence it is no surprise that more than sixty percent of the global prokaryote distribution inhabits the environmental diverse oceanic environment [1]. Nevertheless, the large majority of microbiological studies have dealt with the terrestrial microorganisms, since these environments were more accessible to researchers. In recent years, however, also the marine microbiota has become a topic of intense research [2–9].

Consequently, several research groups and companies have turned to the marine environments with the hope of finding new enzymes, new chemistry, new metabolic pathways, or even new forms of life, that would help addressing the growing demand for efficient and biosustainable products [10–14]. This coupled with the development of technologies for cultivating microorganisms and sequencing genes and genomes, have resulted in an increase of knowledge about the diversity and potential stored in the oceans [15–19].

In the research group of Professor Lone Gram at the Technical University of Denmark, we explore marine bacteria with the aim of finding strains and compounds that can be useful for mankind. In this PhD project we have used second and third generation sequencing technologies to obtain whole genome sequences (WGS) of bioactive marine bacteria from a unique collection of marine bacteria established during the Galathea 3 research expedition [20]. The purpose of this thesis was to understand the genomic potential present in the studied marine bacteria and explore its possible applications. Although based on marine bacteria, the genomic approaches used throughout this PhD project can be applied in any field of Microbiology.

1.1. The Galathea 3 collection

The Galathea 3 global expedition took place in 2006/7 and navigated through the oceans (Figure 1). One of the seventy research projects on board aimed at isolating bacteria that generate bioactive molecules and could potentially be used as sources of novel antibiotics or as probiotic bacteria in marine aquaculture. The Galathea 3 bacterial collection is unique and consists of more than 500 bacterial strains, which have been isolated from all ocean environments due to their antagonistic activity towards the fish pathogen *Vibrio anguillarum* [20].

Figure 1 – Route of the Galathea 3 global expedition (www.galathea3.dk)

Several strains were cultured, isolated and stored with the aim of identifying the compound(s) responsible for the bioactivity. Classical bioassay guided fractionation led to the isolation of several bioactive molecules including new chemical structures. For example, *Photobacterium galatheae* S2753 produces two novel compound families (solonamides and ngercheumicins), which interfere with virulence regulation in *Staphylococcus aureus* [21–23]; and *Vibrio nigripulchritudo* S2604 produces a novel siderophore (nigribactin) [24]. Moreover several known antibiotic compounds were re-discovered, including holomycin, an antibiotic previously only isolated from terrestrial streptomycetes, produced by *P. galatheae* S2753 [6], and andrimid produced by *Vibrio coralliilyticus* S2052 [6]. Additionally, in pigmented *Pseudoalteromonas*, a range of antibiotic compounds (indolmycin, pentabromopseudilin, prodigiosin) were re-identified [25,26].

1.2. Thesis Outline

The thesis comprises an introductory segment (Chapters 1 to 6), where an introduction to the genomic Era and whole genome sequence applications is given, followed by Chapter 7, which includes six research articles based on the work performed during this PhD study.

The PhD project focused on whole genome analyses of several marine bacteria, mostly from the Galathea 3 collection. The initial goal of the project was to resort to genome mining to identify new natural products and their biosynthetic pathways. The main question was if the chemistry so far identified from these marine bacteria matched the genomic potential within these organisms. In Chapter 7, Paper 1 in this thesis, we proved that the genomic potential of several marine species exceeds the chemistry so far identified. We were able to identify several clusters responsible for the production of known molecules, but also clusters that likely encode the biosynthetic machineries necessary to produce potentially new compounds (unpublished data).

However the availability of whole genome sequences far exceeds one single application or opportunity, and soon several other features and details came to our attention. When studying the production of bioactive compounds such as siderophores and the Ferric-uptake regulator (Fur) protein (Paper 1), we noticed good species discrimination within the *Vibrionaceae* family could be found using the Fur protein sequence. Because phylogenetic relationships are

usually hard to determine in *Vibrionaceae*, and several misidentification of species regularly occur, we decided to explore the *fur* gene as a new phylogenetic marker in the *Vibrionaceae* family (Paper 2). This led not only to the proof of its usefulness but also to the identification of two new species within the family, *Photobacterium galatheae* (Paper 3) and *Vibrio galatheae* (paper not included in this thesis). We used this new phylogenetic marker for new species identification (*V. galatheae*) [27], but our main goal is to enable its use among the scientific, clinical and food-surveillance communities, which deal with the identification of isolates from the *Vibrionaceae* family. Therefore we developed a *fur*-database and an online tool, where submission of DNA sequences allows the identification of the isolate at the species level (Paper 4).

We have also used the whole genome sequences in the description of three species, challenging the conservative nature of the microbial taxonomy community (*e.g.* Paper 3). In the reclassification of *Alteromonas fuliginea* (paper not included in this thesis) the power of these new taxonomic tools in resolving misidentification issues previously made using physiological or wet-lab molecular techniques is emphasized.

Challenges associated with natural product discovery relate not only to their synthesis but also to the enzymatic complexity used in the tailoring of the produced compounds. Key enzymes in the production of many bioactive compounds are cytochromes P450, which often play a role in the tailoring of natural products in all domains of life

[28,29]. In Paper 5 we look at the cytochrome P450 diversity in marine bacteria and their potential application in biotechnology. Also, we demonstrate that such key enzymes can be cloned and expressed in model cell factories, such as *Escherichia coli* and *Saccharomyces cerevisiae*.

Finally, in Paper 6 we look into the *Photobacterium* genus (the second largest genus in the *Vibrionaceae* family) and perform a comparative genomics study on this extremely versatile and heterogeneous genus.

2. The Genomic Era

Dr. Thomas Roderick first used the term 'genomics' in 1986 to describe the study and comparison of genomes of different species, including their evolution and relationships [30]. Basically, genomics would require the use of DNA sequencing technology, followed by analyses using *in vitro* experiments and bioinformatic methodology to understand the structure and function of genes.

The establishment of genomics was only possible because of the development of DNA sequencing technologies since the 1970s, which led to the exponential increase of sequencing projects [31–34]. Sequencing became a main feature of biological research, leading to the definition of a genomic Era.

2.1. Developments in Sequencing Technologies

The initial DNA sequencing methods (first-generation sequencing technologies) were based on radioactive phosphorous (^{32}P) and were extremely laborious and low-throughput [35–37]. Developments on the Sanger sequencing technology, with the incorporation of modified and labeled nucleotides, use of higher throughput instruments and better algorithms for result reading, made Sanger sequencing a method of choice for decades.

The first revolution happened with the development of the so-called Next-Generation Sequencing (NGS), now the second-generation [31–33].

The improvements and wide use of the second-generation sequencing technologies generated several opportunities in the field of genomics, but it also emphasized their limitations and the existing room for improvement [34,38].

Although the methods had higher throughput than before, the quality of the sequencing results were actually poorer than the ones performed with older and more expensive techniques [15,16,33,39]. These issues lead to the development of new single-molecule sequencing technologies, the third-generation of sequencing technologies. The read length in these methods is higher, sometimes up to 100 kb, more useful in the analysis of repetitive elements, which translates in a better whole genome reconstruction [15,33]. Third-generation sequencing promises to once again revolutionize the sequencing field [34].

A detailed description of the most used sequencing technologies is provided in Chapter 6 – Appendix 1.

3. Applications of Whole Genome Sequencing

The developments of sequencing technologies associated with the lower cost and higher throughput of DNA sequencing lead microbiology to a genomic Era [16,40]. Statistics on the number of sequences and whole genome sequences deposited at GenBank – NCBI reflect this trend (Figure 2). Bacteria have smaller genome sizes as compared to the ones of higher organisms. Thus whole genome sequencing of bacterial genomes became affordable to most labs around the world; and soon high-throughput sequencing became part of many projects in the bacteriology field.

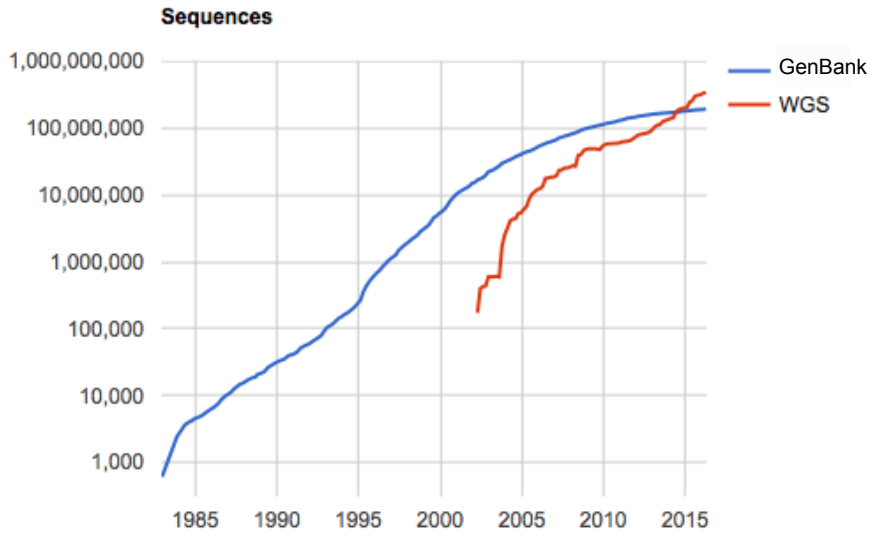


Figure 2 – Number of sequences and whole-genome sequences deposited on NCBI (http://www.ncbi.nlm.nih.gov/genbank/statistics/).

The generation of these whole genome sequences (WGS) provided several new avenues. It broaden our understanding of bacteria, brought WGS to more conservative fields such as taxonomy [41], and opened up new fields such as drug discovery by genome mining and single cell sequencing from environmental samples [12,42,43].

The impact of sequencing technologies in bacteriology has been massive and we here describe some of the fields where this has happened and several of the opportunities it offered.

3.1. Environmental Microbiology

Earth's biosphere is composed of networks of microorganisms, which are important for the ecosystem. Also such networks are important in industrial processes and directly for human wellbeing (*e.g.* human microbiome) [44]. For all these reasons understanding of microbial communities, such as the organisms taking part, their physiology, and the interactions with each other and with other organisms is extremely important.

Major questions in environmental microbiology sciences are: 'who is there?' and 'what are they doing?'. Therefore systematics on the identification and naming of the identified organisms is very important. This systematical classification is named taxonomy.

3.1.1. Taxonomical Identification

Taxonomy is one of the oldest fields in microbiology; the classification of fungi and bacteria goes back hundreds of years. In contrast, genomics is a very recent discipline, a few decades old. In the early years, bacterial classification was based on physiological traits of the different strains, mostly shape, which made it very unstructured. In 1884 after the development of the Gram-staining [45], classification was made by Gram-positive, Gram-negative, Gram-variable or uneven-Gram [46]. Later, more complex phenotypic analyses were performed, such as fatty-acid analyses and growth conditions (temperatures, salinity, carbon-sources, *etc.*); followed by genetic analyses, which included DNA-DNA hybridization (DDH), the gold standard of species definition for many decades [41].

3.1.1.1. Species determination

With the introduction of Sanger sequencing, phylogenies based on the 16S rRNA gene (and other genes) became of common use and started the molecular taxonomy Era [47]. Today, with the use of whole genome sequencing, a new shift in microbial taxonomy is occurring. Several recent studies have been performed in order to validate the use of bacterial WGS in phylogenetic determinations [41,48,49]. Table 1 exemplifies the good correlation obtained between the wet-lab and *in silico* analyses of DNA-DNA relatedness.

Table 1 – Genomic DNA-DNA hybridization (DDH) percentages determined using *in silico* (diagonal up) and wet-lab approaches (diagonal down). Strains: S3431; *P. citrea* LMG 12323^{T}; *A. fuliginea* DSM 15748. ND, not determined (Machado *et al*., 2016).

		***In silico* DDH (% ± SD)**		
		S3431	**LMG 12323**	**DSM 15748**
Wet-lab DDH (%)	**S3431**		19.9 ± 2.30	85.9 ± 2.37
	LMG 12323	20.2		19.6 ± 2.30
	DSM 15748	77.8	ND	

The most accepted *in silico* method in taxonomy is the calculation of the Average Nucleotide Identity (ANI), since several studies have found that it correlates well with DNA-DNA hybridization (DDH) (Figure 3); a similarity value of 95 % ANI corresponds to a 70 % DDH similarity, being these the boundaries for species definition [48,50].

Several online tools have been developed for WGS comparison, to infer the taxonomical relationship of two strains by comparing their WGS. One of the tools is a digital substitute of the wet-lab DNA-DNA hybridization [51], while others perform ANI or Average Amino-acid Identity (AAI) calculations [41,48].

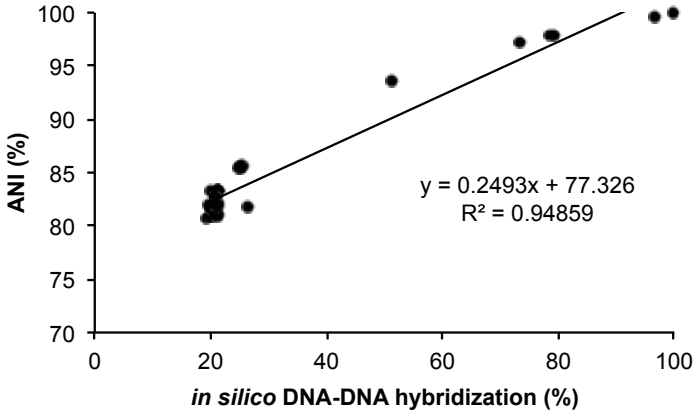


Figure 3 – Correlation of ANI and *in silico* DNA-DNA hybridization values. Correlation performed using a selection of *Photobacterium* genomes representing several species (Machado & Gram, 2016).

Although ANI determinations have been useful in taxonomy inferences, the thresholds used have failed in differentiating several species [41,49]. Most of these cases are already known, since previously used taxonomy approaches have also failed to distinguish between these species. Examples of such cases are *Escherichia coli* vs. *Shigella* species, *Bacillus anthracis* vs. *Bacillus cereus* vs. *Bacillus thuringiensis*, *Yersinia pestis* vs. *Yersinia pseudotuberculosis*, *Mycobacterium tuberculosis* vs. *Mycobacterium bovis*, and *Bordetella pertussis* vs. *Bordetella parapertussis* vs. *Bordetella bronchiseptica* [41].

3.1.1.2. Diversity analysis

It is a fact that our understanding of bacteria was for years limited to the ones that could be grown in regular laboratory conditions. These

are estimated to be only 1% of the whole bacterial diversity present in the environment [19,47,52]. With the implementation of previously described molecular taxonomy methods such as the 16S rRNA gene this changed [53]. Because of 16S rRNA gene conservation, it was possible to PCR amplify these genes without the need for cultivating the organisms present in a given sample, leading to the revelation of an immense microbial diversity previously overlooked [19,53]. With the technological advances of sequencing technologies and the development of different protocols and data analysis pipelines, it became possible to have a more detailed look into the once unseen diversity [16,19,44,52,54].

There are several methods used in environmental sample analyses (Figure 4), but the three main approaches are metagenomics, amplicon sequencing and more recently, single-cell sequencing [44]. All these methods make use of second- or third-generation sequencing technologies in order to give the depth and high throughput needed in these projects and avoid the need for cultivation of all the organisms present in an given environment. These approaches can also be combined with each other or other methods in case there is a specific target organism or gene.

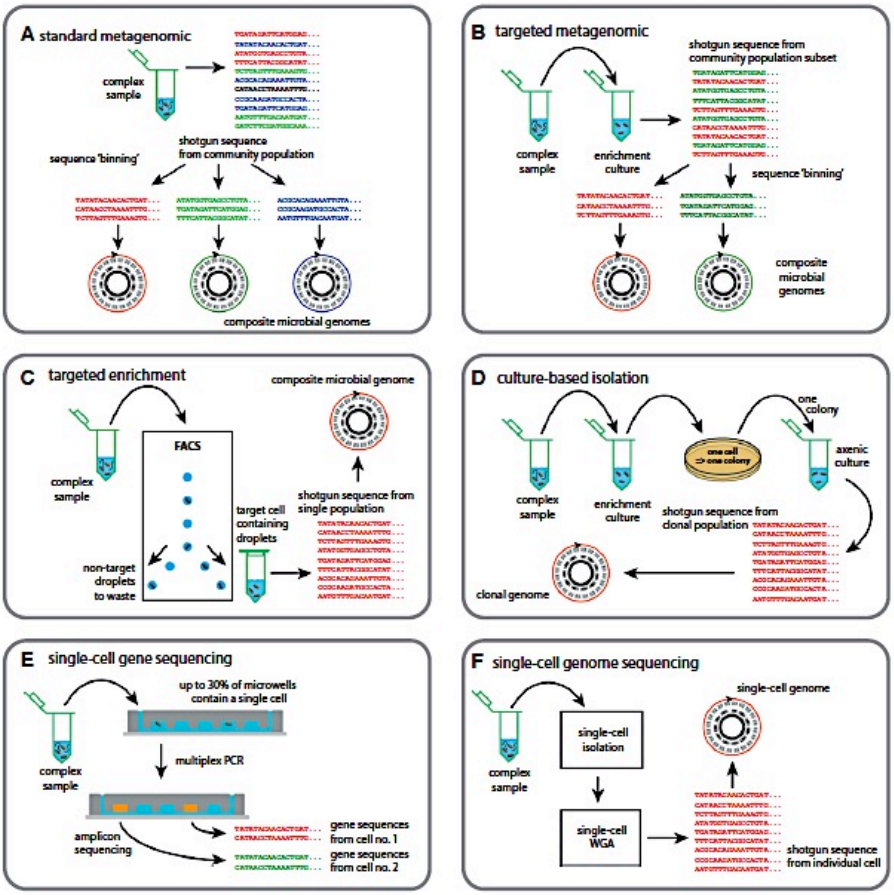


Figure 4 - Methods for microbial genomics. Standard metagenomics and sequence 'binning' to generate microbial genomes (A). Targeted metagenomics by enrichment culture and sequence 'binning' to generate microbial genomes (B). Target enrichment of a given microorganism to produce a single microbial genome (C). Culture based isolation of a microorganism for an axenic culture and generation of a clonal microbial genome (D). Multiplex PCR-based single cell gene sequencing to generate multiple loci sequence information from single cells (E). Single cell whole genome sequencing by single cell isolation and whole genome amplification (WGA) (F). [44]

While metagenomics enables studies of the whole genetic diversity present in a given sample, amplicon sequencing relies on PCR to

amplify a given gene (*e.g.* 16S rRNA gene) and determines the diversity of sequences obtained, which can then be translated into the diversity of the sample [55–57]. A huge drawback of these methodologies is the fact that one can not infer genetic coupling, because it is not clear if the genes identified are present within the same organism or different ones; only gene products encoded by the same organism can freely come in contact with each other and perform multiple enzymatic transformations or regulation of other genes [44]. A recent study comparing metagenomics and amplicon-based methodologies in the study of microbial diversity identified an identification gap with amplicon sequencing, due to mismatches in the primer binding sequences of the 16S rRNA gene 'universal primers' [58].

Another issue that can be encountered in this type of analysis is the variable number and possible low genetic variance of the sequenced gene. Regarding the 16S rRNA gene, although it is provides good differentiation at the genus level, it has several limitations in species distinction [59]. This can be seen for the genera *Pseudoalteromonas* and *Alteromonas*, where species distinction using the 16S rRNA sequence is very poor for several *Pseudoalteromonas* species (Figure 5).

The new single-cell sequencing approaches surpass this issue as it enables the almost full sequencing of a single organism without the need for cultivation [17,44,60]. For example, in 2005, single-cell sequencing of a *Bacillus* sp. strain retrieved two different copies of the 16S rRNA gene. If using metagenomic or amplicon approaches this strain would be considered as two distinct organisms [17].

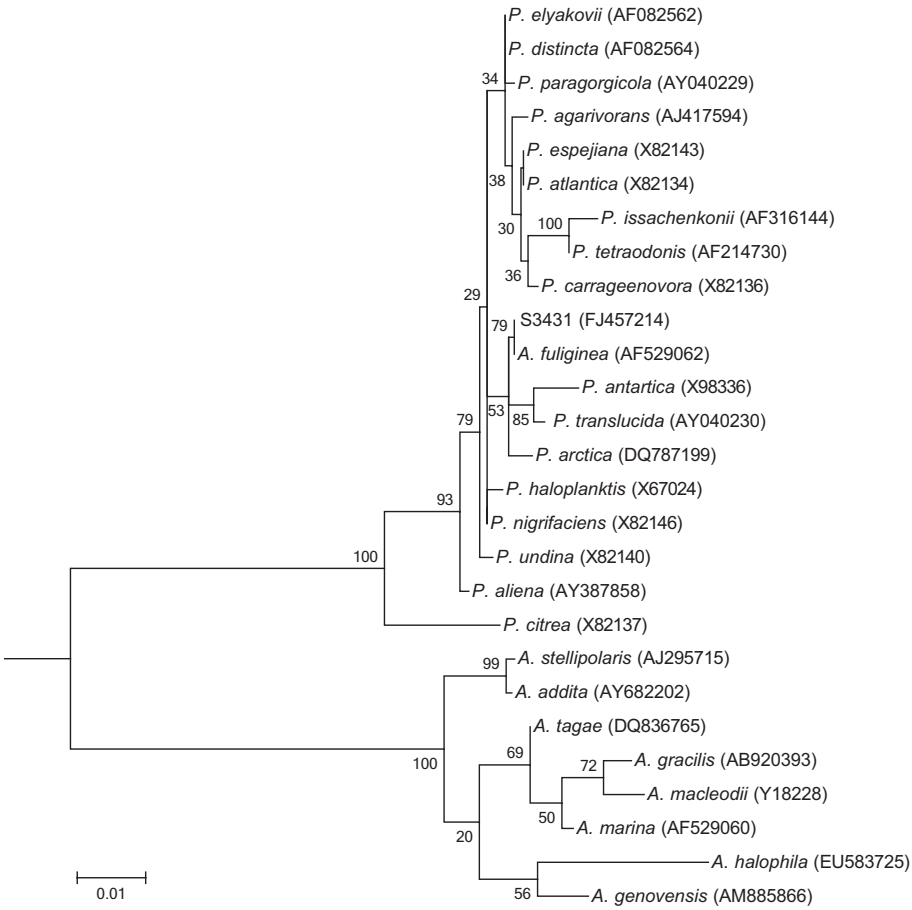


Figure 5 - Maximum likelihood phylogenetic tree based on 16S rRNA gene sequences of *Pseudoalteromonas* and *Alteromonas* type strains. Scale bar: 0,01 substitutions per nucleotide site. Bootstrap support is indicated at nodes (1000 replications). GenBank accession numbers are indicated. *Salinispora arenicola* ATCC BAA-917 was used to root the tree (not included in figure), GenBank accession number NR_042725 (Machado *et al.*, 2016).

3.2. Clinical Microbiology

Whole genome sequencing is also providing information on infectious diseases and epidemiology in clinical settings. With the widespread use of sequencing technologies, their application in clinical settings presents a great opportunity in disease agent surveillance and potentially control of bacterial infections [61,62]. Many infection traits can be now more easily evaluated, including strain transmission within or between hospitals, antibiotic resistance evolution, identification of new bacterial pathogens and evaluation of genomic material exchange between pathogens [63].

Several considerations have to be made when implementing WGS technologies in clinical microbiology: the cost, in-house or outsource sequencing, sequencing capacity, adaptability and data quality. Adding to the sequencing technology considerations, are the bioinformatics analyses performed downstream the process that should also be accounted for their cost, speed, automation, accuracy and detail [62]. All the mentioned considerations present a barrier to the implementation of genomics in clinical settings, but larger hospitals and centers have already started to adopt WGS in diagnostics and epidemiological surveillance [63–65]. The main potential applications of whole genome sequencing (WGS) in the clinical settings are: identification and typing, monitoring of antibiotic resistance and detection of virulence genes.

Owing to its potential, several investments have been made in the development of web tools that use WGS for clinical purposes. An

example of this is the Center for Genomic Epidemiology (www.genomicepidemiology.org), which aims at providing the scientific foundation for web-tools to be used in clinical settings, as well as at creating and validating of the databases used in these tools. It already offers several tools in phenotyping, typing and phylogeny [66–72].

3.2.1. Clinical Typing

When identifying and tracing pathogens in clinical settings, typing is usually performed for epidemiological surveillance of outbreaks. Several methods of typing have been used: serotyping, ribotyping, multilocus variable number tandem repeat analysis (MLVA), pulse-field gel electrophoresis (PFGE) and multilocus sequencing typing (MLST). Different methods have different resolution powers and the use of different methods in the characterization of a strain sometimes hampers strain comparison rather than helping it [62].

The use of WGS may likely contribute to a more unbiased, easier and better comparison of strains from different laboratories, where different typing methods are used. For example, while MLST uses few housekeeping gene sequences in the typing of a specific strain, with WGS a higher number of genes could be used, hence providing a higher discriminatory power. Furthermore, with the more common use of WGS, typing according to single nucleotide polymorphisms (SNPs) has become a common method for distinguishing closely related strains and evaluating their evolutionary history [73]. In the SNP

analyses WGS reads are mapped to a reference genome and SNPs are identified in the sequenced genome compared to the reference one.

The Center for Genomic Epidemiology at DTU has contributed with several typing web-tools that can be used. An example is the MLST web-tool that has been developed using a set of genes from fully sequenced genomes, this allows classification of, for example *Acinetobacter baumannii*, *Clostridium difficile* and *Escherichia coli* [74]. Also tools for serotyping prediction of *E. coli* using whole genome sequences, SerotypeFinder [71] or a fast taxonomy bacterial identification using whole genome sequencing reads, Reads2Type [75], are tools that will revolutionize the clinical typing, once whole genome sequencing reaches the common clinical settings.

As previously described in the 3.1. Environmental Microbiology – 3.1.1. Taxonomical Identification section, WGS offers the possibility of correct, almost flawless identification of strains at the species level without the need to perform time consuming phenotypic or other genetic tests. Therefore, the use of WGS as an identification tool for species where other methods are inconclusive seems very promising. One example is the recent report where WGS was used in the diagnose of neuroleptospirosis [64]. More than 100 different infectious agents can cause encephalitis, making diagnose extremely difficult, and in this case, culturing methods did not work. Therefore, after several unsuccessful hospitalizations and treatments, clinicians used WGS to identify the infectious agent in the cerebrospinal fluid.

With the advances in the third-generation sequencing, it can also be envisioned a time where single-cell sequencing will be performed in diagnoses without the need for bacterial cultivation.

3.2.2. Antibiotic Resistance Detection

Drug-resistant pathogenic microorganisms are becoming a significant threat to public health, due to limited treatment possibilities. In order to avoid the increase of antibiotic resistance infections should be treated with efficient and narrow spectrum antibiotics. This way the use of antibiotics would decrease and pathogens would have a lower evolutionary pressure to acquire resistance.

By sequencing the whole genome of a pathogen it is possible to identify genetic basis of resistance, therefore excluding treatments with antibiotics that the pathogen might be resistant to [76,77]. This *in silico* approach of antibiotic resistance determination avoids the need to culture the pathogens and perform wet-lab antibiotic resistance assays, therefore saving time and expediting treatment adjustments that might be needed.

Several online tools have been lately developed for a fast analysis of the bacterial resistance for clinical purposes, as previously mentioned the Center for Genomic Epidemiology has several tools available including ResFinder, for identification of resistance genes [67]. Although its development dates from 2012, several updates have been made

throughout the years with addition of new resistance genes identified in recent studies.

3.2.3. Virulence Profiling

Another set of important genes to be identified in the clinical setting are the virulence genes encoded by the different pathogens. Virulence genes can be identified from WGS information, allowing a better understanding of the pathogen and help in the prediction of the course of disease, by knowing the capabilities of the pathogen.

Tools to evaluate the pathogenic capabilities of a given bacteria by WGS analysis have been developed, such as PathogenFinder [68]; but also to specifically identify known virulence genes that might have been acquired, for example VirulenceFinder [72]. Using these it is possible to identify genetic basis of for example shiga or cholera toxins or virulence determinants like hemolysins or internalins.

Furthermore, the analysis of WGS can potentially provide an understanding of the evolutionary events that led to the pathogenicity development and in this way expand the knowledge about bacterial pathogens and their interactions with the hosts. It has been previously mentioned that it is difficult to distinguish the species *Bordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica* even using WGS approaches (see section 3.1.1.1. Species determination). All these species colonize the respiratory tract of mammals and *B. pertussis* is a strict human pathogen causing whooping cough [78].

Genomic comparison of these species provided evidence that *B. pertussis* and *B. parapertussis* were independent derivatives of a *B. bronchiseptica*–like ancestor, and that differences in pathogenicity were related with a large-scale gene loss and inactivation by *B. pertussis* [79]. Several other pathogens have been studied and the fact is that different pathogens thrive using different pathogenic approaches, making every case unique [79–81].

3.3. Microbial Biotechnology

Biotechnology consists of any technological application that uses biological systems, living organisms or derivatives thereof, to make or modify products or processes for specific use [82]. It started with the use of microorganisms in the production of beer, cheese, yogurt and bread; and nowadays they are used in much broader applications in the chemical, food, pharmaceutical and biofuel industries, but also in bioremediation and agriculture. Here we focus on the biotechnological applications of bacteria as a source for drug and enzyme discovery, but also as cell factories for biosustainable production of chemicals, drugs and enzymes.

3.3.1. Drug Discovery

As previously mentioned, antibiotic resistance has become a clinical problem and a threat to human life [83]. This is due to fast antibiotic resistance development by the pathogens, but also because the

pharmaceutical industries have failed to keep the antibiotic pipeline running [84–86]. The technical developments on chemical synthesis were expected to supply the drugs of the future, thus natural product discovery was put aside because it was more laborious and expensive. But as chemical synthesis did not deliver, soon research turned back to natural products and a high percentage of new chemicals that are introduced into the markets by pharmaceutical companies are actually derived from natural products [87].

During the last centuries, soil microorganisms have been exploited in the discovery of novel antibiotics and other drugs, supplying more than 80.000 natural products [88]. The most explored genus in terms of secondary metabolites is probably *Streptomyces* [89–93]. Several compounds with antifungal (nystatin, natamycin, amphotericin), antibacterial (chloramphenicol, streptomycin, holomycin) and antiparasitic (ivermectin) activities have been isolated from this genus [94].

A large number of bacterial natural products, including antibitoics, are produced as part of the so-called secondary metabolism and their biosynthesis relies on complex modular enzymes, such as non-ribosomal peptide synthases (NRPSs) and/or polyketide synthases (PKSs), which are modular enzymes [95,96].

The identification and study of the first biosynthetic gene clusters was done using PCR based approaches, which amplified NRPS and PKS domains [97,98]. Later on, DNA libraries like metagenomics were used in the functional screening of biosynthetic clusters followed by

sequencing [99,100]. The links established between natural products and the genetic makeup necessary for their production launched the genomic Era in drug discovery. This led to the great opportunity of using genetic material as a measure of the potential for the production of natural products by a certain organism [12,42,85,86,101,102].

Several bioinformatic tools have been developed to allow the WGS analyses of the genetic potential of a strain for secondary metabolite production [42,85,86,102,103]. These tools have become useful in measuring the potential of organisms to produce natural products, but also in matching identified compounds with their biosynthetic gene clusters. Thalossospiramides and salinosporamides are two examples where identification of the biosynthetic gene clusters allowed the understanding of the structural diversity of the compounds [104,105]. In both cases different chemical structures correlated with differences in the biosynthetic gene clusters of the compounds.

Numerous examples of bioactive molecule discovery using genome-mining tools are available. Several of these are from the proliferous genus *Streptomyces* [106–110]. Examples are the genome mining guided identification of new tris-hydroxamate tetrapeptide in *Streptomyces coelicolor* [107] and new sesquiterpene from *S. avermitilis* [110].

Nowadays, new advances in natural product discovery have been made. Combinations of strategies using mass spectrometry analyses together with peptidogenomics and genome mining to build a molecular network of a given organism are being implemented [111].

This approach has been used in the recent report of the metabolic repertoire of '*Entotheonella*' spp. [43].

3.3.1.1. Bioactive Marine Compounds

Even though scientists have started to explore several habitats other than the terrestrial ones, the marine environment stands out as a hitherto under-explored niche for new bioactive molecules [20,96,102,112,113]. Previous studies have indicated that since the marine environmental conditions are very different from terrestrial habitats, novel compounds and chemical classes are present in this environment, and indeed some marine natural products are characterized by unique marine factors such as halogenation [10,114,115].

During the last years our laboratory has demonstrated the potential of Gram-negative marine bacteria to produce an array of antibiotic and anti-virulence compounds [6,20–22,24–26,116]. In this thesis, using WGS of some of these bacteria we were able to show that the potential for bioactive molecule production is much larger than previously identified by bioassay guided fractionation (Chapter 7, Paper 1) and identify some of the genetic basis for the production of known compounds (Figure 6).

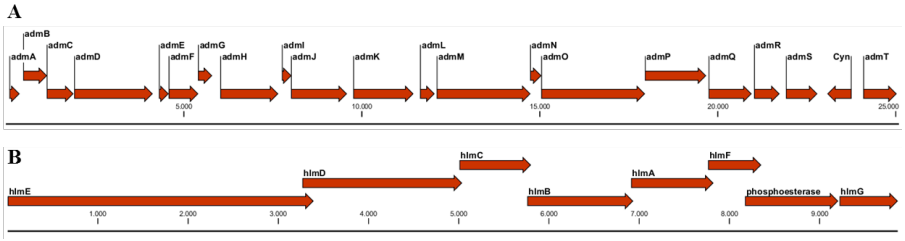


Figure 6 – Previously known biosynthetic clusters identified using genome mining. Andrimid gene cluster from *V. coralliilyticus* S2052 (A) and holomycin gene cluster from *P. galatheae* S2753 (B). (Machado *et al.*, 2015)

The larger genetic repertoire for natural product biosynthesis has also been identified in other marine bacteria, such as cyanobacteria and actinobacteria [95,102,117,118]. Marine actinobacteria from the genus *Salinispora* have been extensively studied regarding natural product biosynthesis [91,102,118]. Several promising compounds have been have been isolated from this genus, with salinosporamides and lomaiviticins representing the most promising leads [119–121].

Sponges have also gained a lot of attention in the natural product discovery field, because different chemical classes of compounds and new scaffolds were identified and are being used nowadays by the pharmaceutical industry [122,123]. Many marine natural products have been isolated and identified from several different sources such as algae, sponges or molluscs; however, several recent studies have attributed the actual production of many of these compounds to microorganisms associated with the eukaryotic producer previously identified. Wilson *et al.*, have in a recent study demonstrated that most of the natural product repertoire isolated from sponges is most

likely produced by previously uncultured bacterial symbionts [43]. The study started by identifying different uncultured bacteria and archaea associated with the marine sponge *Theonella swinhoei* Y. From there whole genome sequencing followed by genome mining was used to identify the genetic potential of the "*Entotheonella*" species and find that "*Entotheonella*" was the symbiont responsible for most of the interesting chemistry identified in the previous studies done on the *T. swinhoei* sponges [43,124].

Although useful in identification of biosynthetic gene clusters for the production of secondary metabolites, second-generation sequencing methodologies are very disadvantageous when analyzing repetitive sequences (see Chapter 6 – Appendix 1). As previously mentioned NRPS and PKS clusters are modular enzymes, with repetitive motifs (Figure 7). This makes it sometimes difficult to assemble such genetic information using solely second-generation sequencing technologies, which provide short readouts. Nowadays, combination of different technologies allows a faster and easier closing of bacterial genomes, therefore avoiding the loss of genetic information. For example, the cluster responsible for the biosynthesis of solonamide B remained unidentifiable until *Photobacterium galatheae* S2753 genome was resequenced using the PacBio third generation sequencing technology (Figure 7) [21,22].

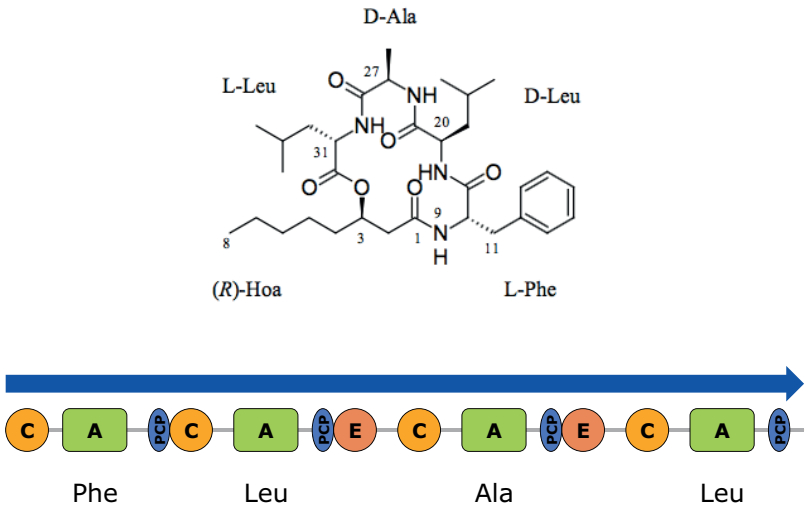


Figure 7 – Putative solonamide B biosynthetic gene cluster. Solonamide B structure and schematic representation of the biosynthetic gene cluster assembled using second- and third-generation sequencing technologies (Machado, unpublished).

3.3.2. Enzyme Discovery

Microorganisms are not only a great source of antimicrobial compounds but they have also supplied many of the enzymes, which have been widely studied and applied in the biotech industry. There is a great interest in proteases, lipases, polysaccharide-degrading enzymes and also in extremozymes (enzymes from extremophile organisms) [7,125].

New sequencing technologies brought great opportunities for enzyme discovery: in uncultivated organisms, using metagenomics or single cell sequencing; but also in the genomes of cultivated organisms. Moreover, the overall process has been expedited by allowing the immediate possibility of cloning and expression of these enzymes in

production organisms, therefore accelerating the path from identification to application.

Genetic information coding for enzymes is much smaller than the one coding for secondary metabolite biosynthetic clusters. Enzymes are usual encoded by few kilo-base pairs, which makes it easier to use metagenomics approaches in enzyme discovery [8,56,126]. When specific enzymatic functions are need, such as biomass degrading enzymes to be used in the production of biofuels, sequencing strategies to identify organisms and enzymes responsible for these functions can be put in place [127,128]. There are several cases where metagenomics and single-cell sequencing of uncultured organisms have been used to enhance the discovery of specific enzymatic functions, such as lignocellulolytic enzymes [127,128].

The most common approach used in enzyme discovery is functional metagenomics, which simultaneously clones and evaluates expression and activity of a certain protein-coding gene. Although straightforward this methodology requires that the heterologous host used for expression (most often *Escherichia coli*) expresses the cloned gene and that the protein is well folded and active [8]. Although it has been predicted that about 40% of the genes from different microorganisms could be expressed in *E. coli* [129], development of alternative heterologous hosts closer related to the species from the studied environments might be a more plausible option [8,126].

With the identification of new marine taxa, certainly new enzymes and pathways will be discovered, enriching the enzyme variety

available to be used in biotechnology. Marine enzymes might present different activity features, when compared to their terrestrial counterparts. The enzyme properties, such as optimal pH, temperature and salinity will most likely reflect the environment where they have been evolving for years.

Identification of a low-temperature-active lipase, two alkane hydroxylases from the deep-sea environment and other hydrolytic enzymes are some examples of enzyme discovery from the marine environment using the metagenomic approach [130–132].

Another type of enzymes that has been gaining much attention in the last decades are cytochromes P450 (CYPs) enzymes. These enzymes can modify several types of compounds, usually by substrate mono-oxygenation, and are of biotechnological interest due to their ability to perform challenging chemistry [28,133]. In comparison with the membrane bound eukaryotic CYPs, the soluble bacterial CYPs are more stable and easier to express and purify [134]. However, their use as biocatalysts in the biotechnological industry has not been fully explored. Marine bacteria are emerging as a reservoir of yet unexplored source of natural products, many of which are dependent on CYP-modifications and we have shown the presence of few CYP coding genes in these bioactive bacteria (Figure 8). Although the number of CYP enzymes in prokaryotes is not as high as in eukaryotes, diverse properties and activities might present new opportunities of usage (Chapter 7 – Paper 5).

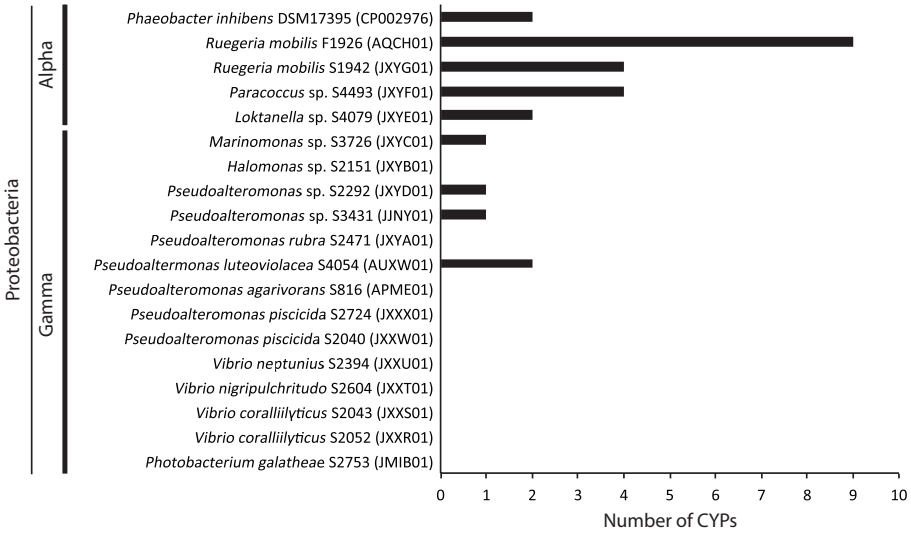


Figure 8 – Overview of the number of identified cytochromes P450 (CYP) in 19 marine bacterial genomes. Both members of alpha and gamma-proteobacteria were mined for putative CYPs. Strain names and numbers are referenced in the vertical axis with the NCBI whole genome sequence accession number in brackets. Black bars represent the number of CYPs identified per genome (Cavaleiro *et al.*, 2016).

3.3.3. Cell Factories

A more recent focus in biotechnology is the use of microorganisms for the production of several products (*e.g.* enzymes, chemicals and biofuels). Previously, different species or strains would be tested to choose the best one, and these would be randomly mutagenized or adaptive laboratory evolved in order to optimize the producing organism. Today the strategies used in the design of cell-factories have been revolutionized by the developments in sequencing technologies, giving rise to the field of metabolic engineering [135–137].

With the availability of WGS it has been possible to get a better understanding of the central metabolism and pathways present in a

given organism and predict which organism would be a better choice. For example, choosing an organism with high tolerance to specific compounds or one that does not encode for degradation pathways of the product to be produced [136]. Furthermore, it made possible a better rational design of the cell-factory by genome engineering, rerouting and optimization of metabolic pathways, etc. [137].

The discovery of new taxa is also an opportunity to discover new organisms to be used as new cell-factories in production. For example, *Vibrio* species have been previously suggested as good candidates due to their genomic and metabolic capabilities [138]. A limitation in using new species in biotechnology is mostly the absence of genetic manipulation tools, which can be used in the design of these new cell factories. Another example is the use of actinomycetes in the production of drugs, where the limited genetic toolbox has been a restrictive factor and has now been targeted for improvement, thanks to the sequencing developments [139].

Nevertheless, understanding of a production strain needs more than just a genomic evaluation. The second-generation of sequencing technologies also allowed the development of methods and protocols for RNA sequencing (RNAseq) [31]. Combined with genome sequencing these RNAseq methods allow the qualitative and quantitative analysis of RNA transcripts in a population at a certain time, the so-called transcriptome. RNAseq studies gave a better understanding of regulatory and transcriptional responses occurring in the microorganisms used in biotechnological applications. Innumerous

studies have used this technique to identify tolerance mechanisms used by cell factories to cope with hazard substrates or products and the industrial production stresses [140]. The analyses of transcriptomic and genomic data also made possible the creation of genome-scale metabolic models that are becoming more and more used in metabolic engineering for the efficient design of cell factories [137].

4. Conclusions and Future Perspectives

The opportunities arising from the developments of sequencing technologies are immense. Some of the fields that can greatly benefit from these developments have been introduced, and are also exemplified in Chapter 7 throughout the several research articles.

The work presented in this thesis made use of whole genome sequencing of marine bacteria for: i) identification of biosynthetic gene cluster for the production of secondary metabolites, ii) taxonomical description of new species, iii) discovery of new cytochrome P450 enzymes, iv) identification and development of new phylogenetic markers to be used in clinical and environmental settings, and v) the better understanding of the *Photobacterium* genus by comparative genomics.

Several examples of the practical applications of sequencing technologies are provided, specifically whole genome sequencing. These were mostly based on the second-generation sequence technologies, which have brought high-throughput to the sequencing field. New third-generation technologies promise to bring sequencing possibilities and applications to an even higher level.

It is expected that third-generation sequence technologies can generate the genetic information from the uncultured bacteria and that this will allow us to develop a metabolic profile of these bacteria and generate laboratory culture conditions. This will allow further study, characterization and description of these new taxa, at the same

time untapping the genomic potential in many habitats and increasing our knowledge about the 99% of the unidentified microbial taxa.

Furthermore, new taxa will most likely translate in the discovery of new compounds, enzymes and pathways that will bring microbiology to the foremost front of biosustainable production efforts of ultimately any product.

5. References

1. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95,** 6578–6583 (1998).

2. Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348,** 1262073_1–1262073_9 (2015).

3. Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348,** 1261498 (2015).

4. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348,** 1–10 (2015).

5. De Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348,** 1261605 (2015).

6. Mansson, M., Gram, L. & Larsen, T. O. Production of bioactive secondary metabolites by marine *Vibrionaceae*. *Mar. Drugs* **9,** 1440–68 (2011).

7. Zhang, C. & Kim, S.-K. Research and application of marine microbial enzymes: status and prospects. *Mar. Drugs* **8,** 1920–1934 (2010).

8. Kennedy, J., Marchesi, J. R. & Dobson, A. D. Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microb. Cell Fact.* **7,** 27 (2008).

9. Giordano, D. *et al.* Marine Microbial Secondary Metabolites: Pathways, Evolution and Physiological Roles. *Adv. Microb. Physiol.* **66,** 357–428 (2015).

10. Wietz, M., Mansson, M., Vynne, N. G. & Gram, L. *Marine Microbiology*. *Marine Microbiology : Bioactive Compounds and Biotechnological Applications* 127–159 (Wiley-VCH Verlag GmbH & Co. KGaA, 2013). doi:10.1002/9783527665259

11. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14,** 111–129 (2015).

12. Jensen, P. R., Chavarria, K. L., Fenical, W., Moore, B. S. & Ziemert, N. Challenges and triumphs to genomics-based natural product discovery. *J. Ind. Microbiol. Biotechnol.* **41,** 203–9 (2014).

13. Mai-Prochnow, A. Biofilm development and cell death in the marine bacterium *Pseudoalteromonas tunicata*. *Appl. Environ. Microbiol.* **70,** 3232–3238 (2004).

14. Mitra, S. & Mukherjee, J. Biofilms of Marine and Intertidal Microbes : Ecology and Biotechnology. *Adv Biochem Eng Biotechnol* **123,** 127–141 (2013).

15. Mardis, E. R. Next-Generation Sequencing Platforms. *Annu. Rev. Anal. Chem.* **6,** 287–303 (2013).

16. Loman, N. J. & Pallen, M. J. Twenty years of bacterial genome sequencing. *Nat. Rev. Microbiol.* **13,** 787–794 (2015).

17. Lasken, R. S. Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10,** 631–640 (2012).

18. Choi, E. J. *et al.* Previously Uncultured Marine Bacteria Linked to Novel Alkaloid Production. *Chem. Biol.* **22,** 1270–1279 (2015).

19. Zengler, K. *et al.* Cultivating the uncultured. *Proc. Natl. Acad. Sci.* **99,** 15681–15686 (2002).

20. Gram, L., Melchiorsen, J. & Bruhn, J. B. Antibacterial activity of marine culturable bacteria collected from a global sampling of ocean surface waters and surface swabs of marine organisms. *Mar. Biotechnol. (NY).* **12,** 439–51 (2010).

21. Mansson, M. *et al.* Inhibition of virulence gene expression in *Staphylococcus aureus* by novel depsipeptides from a marine *Photobacterium*. *Mar. Drugs* **9,** 2537–2552 (2011).

22. Nielsen, A. *et al.* Solonamide B inhibits quorum sensing and reduces *Staphylococcus aureus* mediated killing of human neutrophils. *PLoS One* **9,** e84992 (2014).

23. Kjaerulff, L. *et al.* Identification of four new agr quorum sensing-interfering cyclodepsipeptides from a marine *Photobacterium*. *Mar. Drugs* **11,** 5051–62 (2013).

24. Nielsen, A. *et al.* Nigribactin, a novel siderophore from *Vibrio nigripulchritudo*, modulates *Staphylococcus aureus* virulence gene expression. *Mar. Drugs* **10,** 2584–95 (2012).

25. Vynne, N. G., Månsson, M., Nielsen, K. F. & Gram, L. Bioactivity, chemical profiling, and 16S rRNA-based phylogeny of *Pseudoalteromonas* strains collected on a global research cruise. *Mar. Biotechnol. (NY).* **13,** 1062–73 (2011).

26. Vynne, N. G., Mansson, M. & Gram, L. Gene sequence based clustering assists in dereplication of *Pseudoalteromonas luteoviolacea* strains with identical inhibitory activity and antibiotic production. *Mar. Drugs* **10,** 1729–40 (2012).

27. Giubergia, S., Machado, H., Matei, R. V. & Gram, L. *Vibrio galatheae* sp. nov., a member of the family *Vibrionaceae* isolated from a mussel. *Int. J. Syst. Evol. Microbiol.* **66,** 347–352 (2016).

28. Podust, L. M. & Sherman, D. H. Diversity of P450 enzymes in the biosynthesis of natural products. *Nat. Prod. Rep.* **29,** 1251–66 (2012).

29. Daum, M., Herrmann, S., Wilkinson, B. & Bechthold, A. Genes and enzymes involved in bacterial isoprenoid biosynthesis. *Curr. Opin. Chem. Biol.* **13,** 180–8 (2009).

30. Kuska, B. Beer, Bethesda, and Biology: How "Genomics" Came Into Being. *JNCI J. Natl. Cancer Inst.* **90,** 93–93 (1998).

31. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52,** 413–435 (2011).

32. Ansorge, W. J. Next-generation DNA sequencing techniques. *N. Biotechnol.* **25,** 195–203 (2009).

33. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30,** (2014).

34. Lee, H. *et al.* Third-generation sequencing and the future of genomics. *bioRxiv* (2016). doi:10.1101/048603

35. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).

36. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–7 (1977).

37. Maxam, a M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 560–564 (1977).

38. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

39. Loman, N. J. *et al.* High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.* **10,** 599–606 (2012).

40. Reuter, J. a., Spacek, D. V. & Snyder, M. P. High-Throughput Sequencing Technologies. *Mol. Cell* **58,** 586–597 (2015).

41. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 19126–19131 (2009).

42. Weber, T. In silico tools for the analysis of antibiotic biosynthetic pathways. *Int. J. Med. Microbiol.* **304,** 230–5 (2014).

43. Wilson, M. C. *et al.* An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506,** 58–62 (2014).

44. Blainey, P. C. The future is now: Single-cell genomics of bacteria and archaea. *FEMS Microbiol. Rev.* **37,** 407–427 (2013).

45. Gram, H. C. Gram staining. *Fortschritte der Med.* **2,** 185–189 (1884).

46. Gibbons, N. E. & Murray, R. G. E. Proposals Concerning the Higher Taxa of Bacteria. *Int. J. Syst. Bacteriol.* **28,** 1–6 (1978).

47. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276,** 734–740 (1997).

48. Goris, J. *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* **57,** 81–91 (2007).

49. Kim, M., Oh, H. S., Park, S. C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **64,** 346–351 (2014).

50. Tindall, B. J., Rosselló-Móra, R., Busse, H. J., Ludwig, W. & Kämpfer, P. Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol.* **60,** 249–266 (2010).

51. Auch, A. F., von Jan, M., Klenk, H.-P. & Göker, M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* **2,** 117–134 (2010).

52. Lozupone, C. a & Knight, R. Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 11436–11440 (2007).

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

54. Konstantinidis, K. T. & Tiedje, J. M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr. Opin. Microbiol.* **10,** 504–509 (2007).

55. McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C. & Brumfield, R. T. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenet. Evol.* **66,** 526–538 (2013).

56. Hugenholtz, P. & Tyson, G. W. Metagenomics. *Nature* **455,** 481–483 (2008).

57. Wayne, D. Encyclopedia of Metagenomics. (2013). doi:10.1007/978-1-4614-6418-1

58. Eloe-Fadrosh, E. a, Ivanova, N. N., Woyke, T. & Kyrpides, N. C. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat. Microbiol.* **1,** 15032 (2016).

59. Sawabe, T., Kita-Tsukamoto, K. & Thompson, F. L. Inferring the evolutionary history of vibrios by means of multilocus sequence analysis. *J. Bacteriol.* **189,** 7932–6 (2007).

60. Blainey, P. C. & Quake, S. R. Dissecting genomic diversity, one cell at a time. *Nat. Methods* **11,** 19–21 (2014).

61. Salipante, S. J. *et al.* Application of Whole-Genome Sequencing for Bacterial Strain Typing in Molecular Epidemiology. *J. Clin. Microbiol.* **53,** 1072–1079 (2015).

62. Kwong, J. C., McCallum, N., Sintchenko, V. & Howden, B. P. Whole genome sequencing in clinical and public health microbiology. *Pathology* **47,** 199–210 (2015).

63. Roach, D. J. *et al.* A Year of Infection in the Intensive Care Unit: Prospective Whole Genome Sequencing of Bacterial Clinical Isolates Reveals Cryptic Transmissions and Novel Microbiota. *PLoS Genet.* **11,** 1–21 (2015).

64. Wilson, M. R. *et al.* Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N. Engl. J. Med.* **370,** 2408–17 (2014).

65. Moran-Gilad, J. *et al.* Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. *BMC Infect. Dis.* **15,** 174 (2015).

66. Hasman, H. *et al.* Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples. *J. Clin. Microbiol.* **52,** 139–146 (2014).

67. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67,** 2640–2644 (2012).

68. Cosentino, S., Voldby Larsen, M., Møller Aarestrup, F. & Lund, O. PathogenFinder--distinguishing friend from foe using

bacterial whole genome sequence data. *PLoS One* **8,** e77302 (2013).

69. Carattoli, A. *et al.* PlasmidFinder and pMLST: *in silico* detection and typing of plasmid. *Antimicrob. Agents Chemother.* **58,** 3895–3903 (2014).

70. Joensen, K. G. *et al.* Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* **52,** 1501–1510 (2014).

71. Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M. & Scheutz, F. Rapid and easy in silico serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J. Clin. Microbiol.* **53,** 2410–2426 (2015).

72. Kleinheinz, K. A., Joensen, K. G. & Larsen, M. V. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage* **4,** e27943 (2014).

73. Oliver, G. R., Hart, S. N. & Klee, E. W. Bioinformatics for clinical next generation sequencing. *Clin. Chem.* **61,** 124–135 (2015).

74. Larsen, M. V. *et al.* Multilocus Sequence Typing of Total Genome Sequenced Bacteria. *J. Clin. Microbiol.* **50,** 1355–1361 (2012).

75. Saputra, D. *et al.* Reads2Type: a web application for rapid microbial taxonomy identification. *BMC Bioinformatics* **16,** 398 (2015).

76. Köser, C. U., Ellington, M. J. & Peacock, S. J. Whole-genome sequencing to control antimicrobial resistance. *Trends Genet.* **30,** 401–407 (2014).

77. Punina, N. V, Makridakis, N. M., Remnev, M. a & Topunov, a F. Whole-genome sequencing targets drug-resistant bacterial infections. *Hum. Genomics* **9,** 19 (2015).

78. Cherry, J. D. Historical review of pertussis and the classical vaccine. *J. Infect. Dis.* **174 Suppl ,** S259–S263 (1996).

79. Parkhill, J. *et al.* Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat. Genet.* **35,** 32–40 (2003).

80. Marvig, R. L., Sommer, L. M., Molin, S. & Johansen, H. K. Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nat. Genet.* **47,** 57–65 (2015).

81. Burstein, D. *et al.* Genomic analysis of 38 *Legionella* species identifies large and diverse effector repertoires. *Nat. Genet.* **48,** 167–175 (2016).

82. United Nations. Convention on biological diversity. *Diversity* 30 (1992). at <http://www.cbd.int/doc/legal/cbd-en.pdf>

83. Organization, W. H. Antimicrobial resistance: global report on surveillance. *Who* 8 (2014). doi:1.4.2014

84. Xu, J. & Hagler, A. Chemoinformatics and Drug Discovery. *Molecules* **7,** 566–600 (2002).

85. Scheffler, R. J., Colmer, S., Tynan, H., Demain, a L. & Gullo, V. P. Antimicrobials, drug discovery, and genome mining. *Appl. Microbiol. Biotechnol.* **97,** 969–78 (2013).

86. Zerikly, M. & Challis, G. L. Strategies for the discovery of new natural products by genome mining. *Chembiochem* **10,** 625–33 (2009).

87. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.* **70,** 461–77 (2007).

88. Bérdy, J. Thoughts and facts about antibiotics: Where we are now and where we are heading. *The Journal of Antibiotics* **65,** 441–441 (2012).

89. Yu, D., Xu, F., Valiente, J., Wang, S. & Zhan, J. An indigoidine biosynthetic gene cluster from *Streptomyces chromofuscus* ATCC 49982 contains an unusual IndB homologue. *J. Ind. Microbiol. Biotechnol.* **40,** 159–68 (2013).

90. Li, B. & Walsh, C. T. Identification of the gene cluster for the dithiolopyrrolone antibiotic holomycin in *Streptomyces clavuligerus*. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 19731–5 (2010).

91. Udwary, D. W. *et al.* Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 10376–81 (2007).

92. Yin, H. *et al.* Induction of holomycin production and complex metabolic changes by the argR mutation in *Streptomyces clavuligerus* NP1. *Appl. Environ. Microbiol.* **78,** 3431–41 (2012).

93. Li, B. & Walsh, C. T. *Streptomyces clavuligerus* HlmI is an intramolecular disulfide-forming dithiol oxidase in holomycin biosynthesis. *Biochemistry* **50,** 4615–22 (2011).

94. Bhattacharya, D., Nagpure, A. & Gupta, R. K. Bacterial chitinases: properties and potential. *Crit. Rev. Biotechnol.* **27,** 21–8 (2007).

95. Moore, B. S. Biosynthesis of marine natural products: microorganisms (Part A). *Nat. Prod. Rep.* **22,** 580–93 (2005).

96. Xiong, Z.-Q., Wang, J.-F., Hao, Y.-Y. & Wang, Y. Recent advances in the discovery and development of marine microbial natural products. *Mar. Drugs* **11,** 700–17 (2013).

97. Amos, G. C. a *et al.* Designing and implementing an assay for the detection of rare and divergent NRPS and PKS clones in European, Antarctic and Cuban soils. *PLoS One* **10,** 1–15 (2015).

98. Gontang, E. A., Gaudêncio, S. P., Fenical, W. & Jensen, P. R. Sequence-based analysis of secondary-metabolite biosynthesis in marine actinobacteria. *Appl. Environ. Microbiol.* **76,** 2487–2499 (2010).

99. Edwards, D. J. *et al.* Structure and Biosynthesis of the Jamaicamides , New Mixed Polyketide-Peptide Neurotoxins from the Marine Cyanobacterium *Lyngbya majuscula*. **11,** 817–833 (2004).

100. Zhang, G. *et al.* Characterization of the amicetin biosynthesis gene cluster from *Streptomyces vinaceusdrappus* NRRL 2363

implicates two alternative strategies for amide bond formation. *Appl. Environ. Microbiol.* **78,** 2393–2401 (2012).

101. Minowa, Y., Araki, M. & Kanehisa, M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.* **368,** 1500–17 (2007).

102. Ziemert, N. *et al.* Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E1130–9 (2014).

103. Fedorova, N. D., Moktali, V. & Medema, M. H. in *Fungal Secondary Metabolism: Methods and Protocols* (eds. Keller, N. P. & Turner, G.) **944,** 23–45 (Humana Press, 2012).

104. Eustáquio, A. S. *et al.* The discovery of salinosporamide K from the marine bacterium "*Salinispora pacifica*" by Genome mining gives insight into pathway evolution. *ChemBioChem* **12,** 61–64 (2011).

105. Ross, A. C. *et al.* Biosynthetic multitasking facilitates thalassospiramide structural diversity in marine bacteria. *J. Am. Chem. Soc.* **135,** 1155–62 (2013).

106. Aigle, B. *et al.* Genome mining of *Streptomyces ambofaciens*. *Journal of Industrial Microbiology and Biotechnology* **41,** 251–263 (2014).

107. Lautru, S., Deeth, R. J., Bailey, L. M. & Challis, G. L. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nature chemical biology* **1,** 265–269 (2005).

108. Gomez-Escribano, J. P. & Bibb, M. J. Heterologous expression of natural product biosynthetic gene clusters in *Streptomyces coelicolor*: From genome mining to manipulation of biosynthetic pathways. *Journal of Industrial Microbiology and Biotechnology* **41,** 425–431 (2014).

109. Hu, Y., Chou, W. K. W., Hopson, R. & Cane, D. E. Genome mining in *Streptomyces clavuligerus*: Expression and biochemical characterization of two new cryptic sesquiterpene synthases. *Chem. Biol.* **18,** 32–37 (2011).

110. Chou, W. K. W. *et al.* Genome mining in *Streptomyces avermitilis*: Cloning and characterization of sav-76, the synthase for a new sesquiterpene, avermitilol. *J. Am. Chem. Soc.* **132,** 8850–8851 (2010).

111. Nguyen, D. D. *et al.* MS/MS networking guided analysis of molecule and gene cluster families. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E2611–20 (2013).

112. Zhao, X.-Q. Genome-based studies of marine microorganisms to maximize the diversity of natural products discovery for medical treatments. *Evid. Based. Complement. Alternat. Med.* **2011,** 384572 (2011).

113. Wietz, M., Duncan, K., Patin, N. V & Jensen, P. R. Antagonistic interactions mediated by marine bacteria: the role of small molecules. *J. Chem. Ecol.* **39,** 879–91 (2013).

114. Fenical, W. Chemical studies of marine bacteria: developing a new resource. *Chem. Rev.* **93,** 1673–1683 (1993).

115. Lane, A. L. & Moore, B. S. A sea of biosynthesis: marine natural products meet the molecular age. *Nat. Prod. Rep.* **28,** 411–28 (2011).

116. Wietz, M., Mansson, M., Gotfredsen, C. H., Larsen, T. O. & Gram, L. Antibacterial compounds from marine *Vibrionaceae* isolated on a global expedition. *Mar. Drugs* **8,** 2946–60 (2010).

117. Wang, H., Fewer, D. P. & Sivonen, K. Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. *PLoS One* **6,** e22384 (2011).

118. Jensen, P. R., Moore, B. S. & Fenical, W. The marine actinomycete genus *Salinispora*: a model organism for secondary metabolite discovery. *Nat. Prod. Rep.* **32,** 738–751 (2015).

119. Kersten, R. D. *et al.* Bioactivity-guided genome mining reveals the lomaiviticin biosynthetic gene cluster in *Salinispora tropica*. *Chembiochem* **14,** 955–62 (2013).

120. Feling, R. H. *et al.* Salinosporamide A: A highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus *Salinospora*. *Angew. Chemie - Int. Ed.* **42,** 355–357 (2003).

121. Fenical, W. & Jensen, P. R. Developing a new resource for drug discovery: marine actinomycete bacteria. *Nat. Chem. Biol.* **2,** 666–673 (2006).

122. Pérez-López, P. *et al.* Environmental solutions for the sustainable production of bioactive natural products from the marine sponge *Crambe crambe*. *Sci. Total Environ.* **475,** 71–82 (2014).

123. Simmons, T. L. *et al.* Biosynthetic origin of natural products isolated from marine microorganism-invertebrate assemblages. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 4587–4594 (2008).

124. Bewley, C. A., Holland, N. D. & Faulkner, D. J. Two classes of metabolites from *Theonella swinhoei* are localized in distinct populations of bacterial symbionts. *Experientia* **52,** 716–722 (1996).

125. Nigam, P. S. Microbial enzymes with special characteristics for biotechnological applications. *Biomolecules* **3,** 597–611 (2013).

126. Lorenz, P. & Eck, J. Metagenomics and industrial applications. *Nature* **3,** 510–516 (2005).

127. Hess, M. *et al.* Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science* **331,** 463–467 (2011).

128. DeAngelis, K. M. *et al.* Strategies for enhancing the effectiveness of metagenomic-based enzyme discovery in lignocellulolytic microbial communities. *Bioenergy Res.* **3,** 146–158 (2010).

129. Gabor, E. M., Alkema, W. B. L. & Janssen, D. B. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ. Microbiol.* **6,** 879–886 (2004).

130. Acevedo, J. P. *et al.* Cloning of complete genes for novel hydrolytic enzymes from Antarctic sea water bacteria by use of an improved genome walking technique. *J. Biotechnol.* **133,** 277–286 (2008).

131. Xu, M., Xiao, X. & Wang, F. Isolation and characterization of alkane hydroxylases from a metagenomic library of Pacific deep-sea sediment. *Extremophiles* **12,** 255–262 (2008).

132. Hårdeman, F. & Sjöling, S. Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. in *FEMS Microbiology Ecology* **59,** 524–534 (2007).

133. Werck-Reichhart, D. & Feyereisen, R. Cytochromes P450: a success story. *Genome Biol.* **1,** REVIEWS3003 (2000).

134. Urlacher, V. & Schmid, R. D. Biotransformations using prokaryotic P450 monooxygenases. *Curr. Opin. Biotechnol.* **13,** 557–564 (2002).

135. Conrad, T. M., Lewis, N. E. & Palsson, B. Ø. Microbial laboratory evolution in the era of genome-scale science. *Mol. Syst. Biol.* **7,** 509 (2011).

136. Sauer, M. & Mattanovich, D. Construction of microbial cell factories for industrial bioprocesses. *J. Chem. Technol. Biotechnol.* **87,** 445–450 (2012).

137. Lee, J. W. *et al.* Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.* **8,** 536–546 (2012).

138. Reen, F. J., Almagro-Moreno, S., Ussery, D. & Boyd, E. F. The genomic code: inferring *Vibrionaceae* niche specialization. *Nat. Rev. Microbiol.* **4,** 697–704 (2006).

139. Weber, T. *et al.* Metabolic engineering of antibiotic factories: New tools for antibiotic production in actinomycetes. *Trends in Biotechnology* **33,** 15–26 (2015).

140. Mukhopadhyay, A. Tolerance engineering in bacteria for the production of advanced biofuels and chemicals. *Trends in Microbiology* **23,** 498–508 (2015).

6. Appendix 1 – Sequencing Technologies

6.1. First Generation Sequencing Technologies

In 1975, Sanger and collaborators published the original Sanger sequencing method [1,2]. This method used a DNA polymerase for extending a DNA sequencing, which would be labeled using radioactive phosphorous (^{32}P) and then divided into two reactions types: one for amplification of the template DNA; and another 3'-end degradation of the double-stranded template DNA. The amplification reaction would be performed four times in the absence of each nucleotide ("the minus system"), allowing correct amplification until the missing nucleotide; while the degradation reaction would be also performed four times in the presence of each nucleotide ("the plus system), and DNA template degradation would occur until positions where the used nucleotide was present (Figure 1A). The samples would than be subjected to gel electrophoresis for result reading [1].

At the same time Allan Maxam and Walter Gilbert were also working on a DNA sequencing method and in 1977 they published a new method for DNA sequencing [3]. In this method a DNA molecule would be labeled with radioactive phosphorous (^{32}P) and then cleaved in specific residues, depending on the reaction (cleavage reactions: A+G, A, C+T or C). The remaining labeled DNA molecule would be separated in a gel and the reading of the results from the four

different reactions would allow the determination of the DNA sequence (Figure 1B).

Initially, Maxam and Gilberts method was widely used, but developments of the termination reactions in the Sanger sequencing, such as incorporation of modified nucleotides (di-deoxynucliotides - ddNTPs) and later labeled ddNTPs, made this technique less laborious and safer. Therefore Sanger sequencing became the method of choice for the next 30 decades. It benefitted with the automation processes in sample preparation but mostly in result reading and analyses, which made it a throughput technique that enabled the complete sequencing of the first human genome in 2004 [4,5].

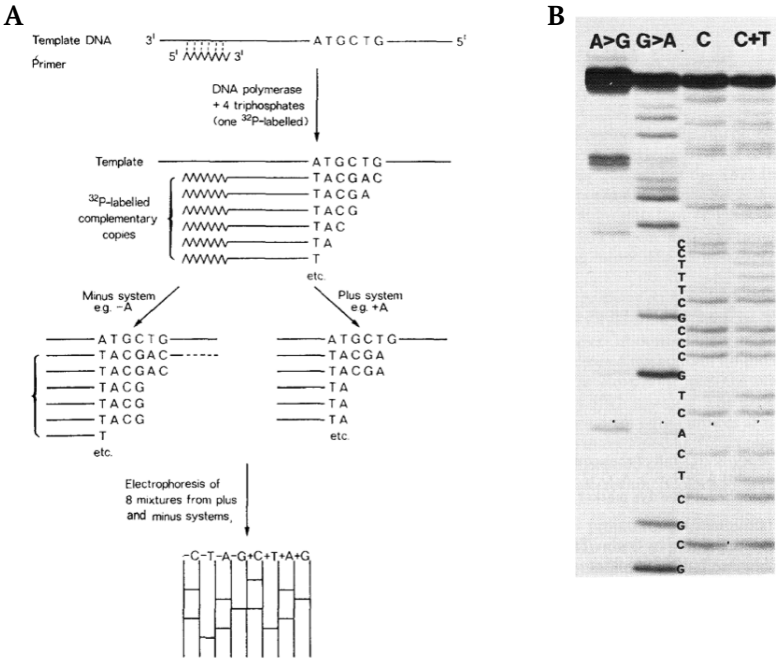


Figure 1 – First Generation Sequencing Methods. The principle of the Sanger sequencing method [1] (A). Example of a autoradiograph of a sequencing gel made using the Maxam and Gilbert sequencing method [3] (B).

6.2. Second Generation Sequencing Technologies - Next Generation Sequencing (NGS)

The Sanger sequencing kept being used and it is still used in smaller-scale sequencing projects. However, with the success of the Human Genome Project, owing to the first complete sequence of the human genome, a need for better sequencing technologies was evident. The full sequence of the human genome had taken 14 years and an estimated cost of $3 billion US dollars [6], so the room for improvement was evident.

Therefore, in 2004 a funding program was initiated by the National Human Genome Research Institute (NHGRI) to reduce the cost of the human genome to $1.000 US dollars in ten years [7]. This was a great driving force behind the development and commercialization of Next Generation Sequencing (NGS) technologies, the second generation of sequencing technologies.

6.2.1. 454 Sequencing

The first NGS technology to reach the market was the 454 sequencing by Life Sciences (now Roche) in 2005 [8]. This method relies on library construction by any method that is able to give a mixture of short adaptor-flanked fragments. These fragments are attached to beads, one fragment per bead, and each fragment amplified individually by emulsion PCR (Figure 2). After denaturation of unattached fragments, the sequencing takes place. During each sequencing amplification cycle (by means of a polymerase) a single species of unlabeled nucleotide is introduced

(Figure 2). If incorporation occurs, pyrophosphate is released and a light signal recorded (by the action of ATP sulfurylase and luciferase). The so-called pyrosequencing is therefore recorded in real time. The sequencing is not simultaneous for all the fragments since some might get ahead or behind others depending on their sequence and the order of base addition.

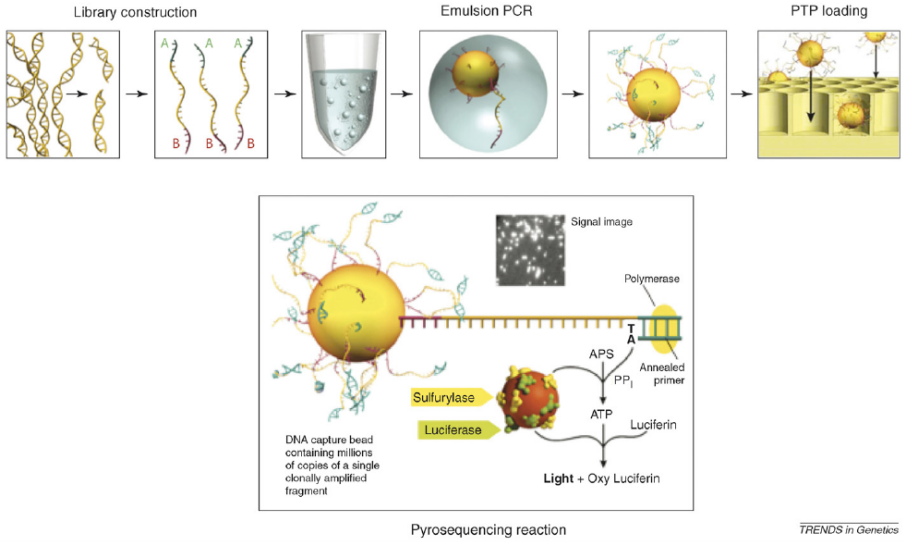


Figure 2 – 454 Sequencing Workflow. The top panel outlines sample preparation. The bottom panel illustrates the pyrosequencing after incorporation of a nucleotide. PTP: picotiter plate. [9]

This technology is able to produce longer reads (maximum 1 kb), which is useful in several applications, and the running times are fairly fast (approx. 23 h). On the other hand, it is quite low throughput, with high reagent cost and presents high error rates when sequencing homopolymer repeats. Homopolymer repeats are consecutive instances of the same base, *e.g.* TTT or CCC. This occurs

because there is no control over multiple consecutive incorporations at a given cycle; the number of incorporations has therefore to be determined by light intensity and it is thus more error prone.

454 sequencing use will most probably be discontinued, as Roche already announced the shut down of 454 and the supporting platform by mid-2016 (http://www.bio-itworld.com/2013/10/16/six-years-after-acquisition-roche-quietly-shutters-454.html).

6.2.2. Solexa/Illumina Sequencing

Later in 2007 a new sequencing technology reached the market. This platform has in its origin the merging of four companies Solexa (Essex, UK), Lynx Therapeutics (Hayward, CA, USA), Manteia Predictive Medicine (Coinsins, Switzerland) and Illumina [10].

As in 454 sequencing, libraries can be constructed by any method that gives rise to several hundred base-pairs adaptor-flanked fragments. These fragments are then bound in one end to a solid surface, which is coated with oligonucleotides complementary to the adaptors (Figure 3B). The single fragments are then amplified using bridge PCR, resulting in clusters of clonal fragments on the solid surface. After removal of unattached fragments by denaturation, the sequencing takes place. The reversible terminator nucleotides used in the sequencing have two modifications: a 3' terminator moiety that allows a single nucleotide insertion and a fluorophore moiety specific for each nucleotide type (Figure 3C). After each cycle and image acquisition, both moieties are cleaved and the cycle is repeated.

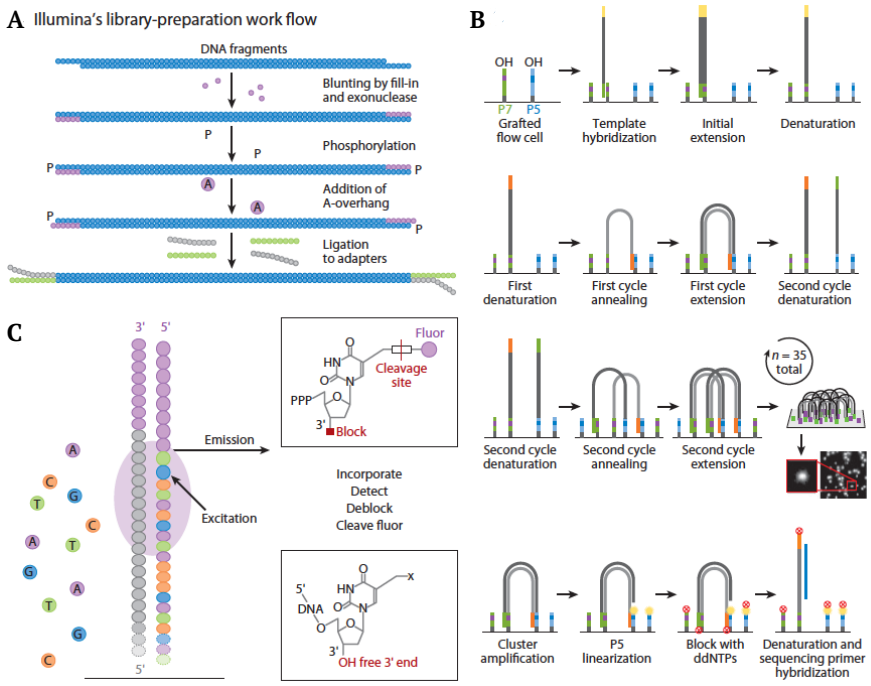


Figure 3 – Solexa/Illumina Workflow. Library preparation (A). Cluster generation on solid surface by bridge PCR (B). Exemplification of sequencing using the fluorescent reversible terminator nucleotides (C). [11]

Currently Illumina has the highest throughput per run and the lowest per-base cost, performing sequencing of up to 300 base-pairs, which makes it compatible with most applications [10,12]. The most common error type is substitution, instead of insertions or deletions (as for 454 sequencing of homopolymers), due to the use of reversible terminators that allow a single base amplification. Drawbacks are the need for tight control of sample loading, as high concentration of fragments might lead to overlapping clonal clusters and consequently low quality sequencing; and the need for sequence

diversity of the fragment library, in order to have effective template generation with appropriate yields [6].

6.2.3. Oligo Ligation Detection (SOLiD) Sequencing

In the same year as Illumina, Sequencing by Oligo Ligation Detection (SOLiD) was released by Applied Biosystems (now Life Technologies) [10].

It also requires a library preparation of a mixture of short adaptor-flanked fragments, that are subsequently bound to a bead for clonal fragment amplification by emulsion PCR, as in the 454 system. Beads containing the clonal fragments are then immobilized in a disordered array manner to a solid planar substrate. Sequencing occurs by the ligation of octamers to the fragment to be sequenced by a DNA ligase that extends the primer used. The readings are made using a flurophore in a determined position of the octamer, which is then cleaved right next to the base read and removal of the fluorophore. For example, if the fluorophore is in the 5th base of the octamer it will enable the reading of every 5th base (*e.g.* bases 5, 10, 15, 20, 25) (Figure 4).

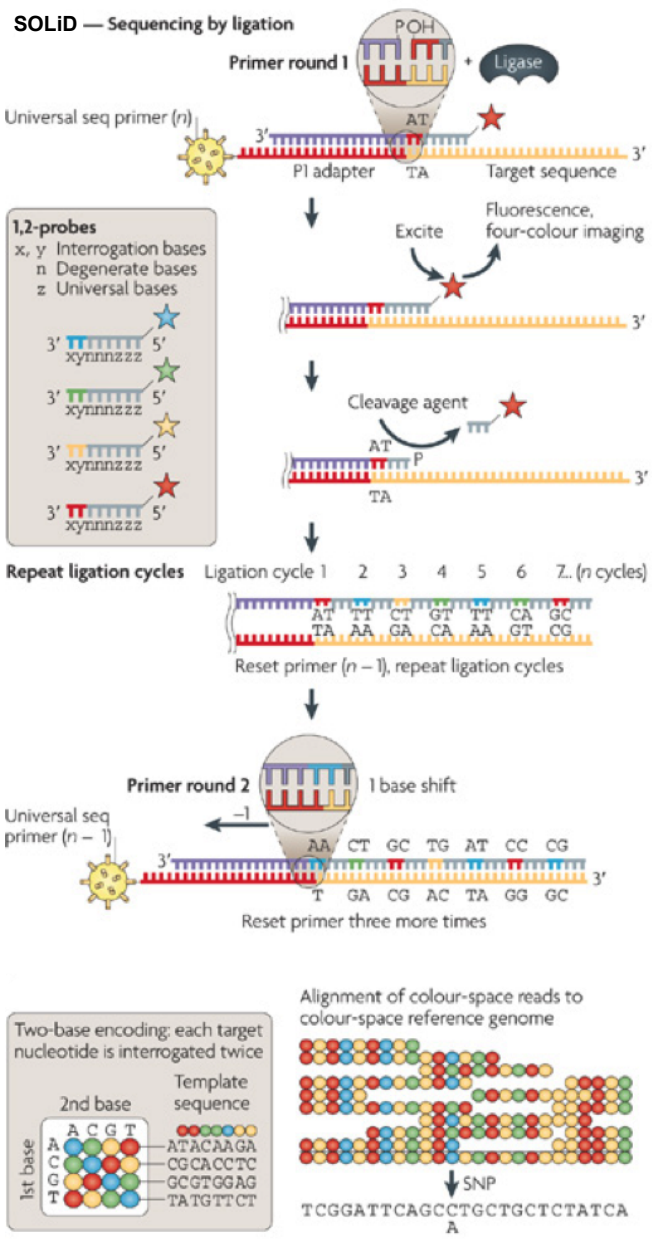


Figure 4 – SOLiD Sequencing Principle. The bottom panel demonstrates the mechanism behind reading of two adjacent bases. [13]

The reads for all the positions of the fragment are then obtained by using different length of primers plus octamers with fluorophores in the same position (*e.g.* the 5th position) or the same primer and octamer with fluorophores in different positions. A feature added to this system was a double reading of each base, by using octamers with fluorophores in two adjacent positions (*e.g.* 4th and 5th bases), improving this way the error correction.

The fact that each base is read twice has made this technology to be claimed the most accurate with 99.94 % accuracy, possibly due to the fact that every base is read twice. Unfortunately this technology produces the shortest reads (maximum 75 bp) with relatively long running times. Because it is less widely use, kits and support services are less developed when compared to the Illumina technologies [6].

6.2.4. Ion Torrent Sequencing

More recently in 2010, Ion Torrent (now Life Technologies) brought to the market the Personal Genome Machine (PGM) [14]. The system is similar to the 454 one, but instead of optical detection it uses semiconductor technology.

A library of mixed short adaptor-flanked fragments is generated by any suitable method and then bound beads. The fragments in the beads are amplified using emulsion PCR to get clonal fragments. The beads are then placed on an Ion Chip, which was designed to detect pH changes in individual wells where each bead is placed. Each different nucleotide is supplied to the system and when incorporation occurs, a hydrogen atom is released creating

acidification of the well environment, which is registered by a detector that translated the level of acidification with the number of nucleotides incorporated.

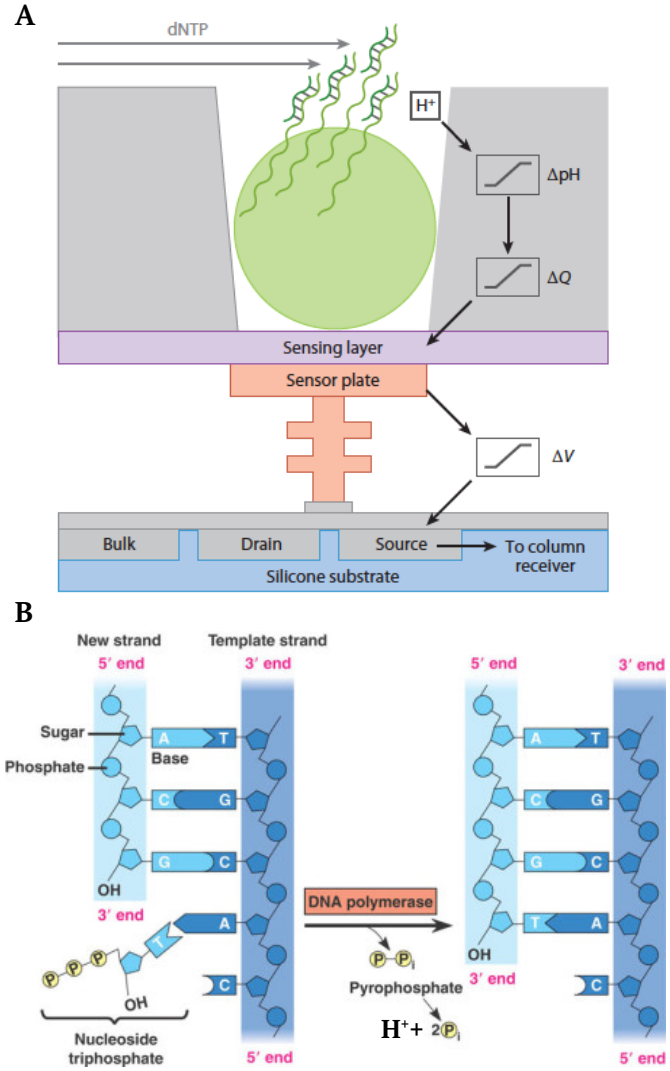


Figure 5 – Ion Torrent Sequencing Scheme. Structure of the Iron Torrent Chip used in pH-bsed sequencing **(A)** [11]. Example of nucleotide incorporation with the release of hydrogen atoms **(B)** (http://www.chromosomechronicles.com/2011/04/12/ion-torrent-the-dark-side-of-dna-sequencing/)

This technology does not need modified nucleotides or an optical reading system, which resulted in a faster, cheaper and smaller instrument and sequencing. The main problem of this technology is the previously mentioned homopolymers and the errors resulting from their occurrence. As for the 454 technology, because there is no control over the number of incorporated nucleotides during each cycle, during homopolymer sequencing the estimation is made by the calculation of a ratio, which might lead to insertion or deletion errors in the sequencing [6,11].

6.3. Third Generation Sequencing Technologies

Three third-generation technologies are nowadays commercially available: Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) sequencing, the Illumina Tru-seq Synthetic Long-Read technology and the Oxford Nanopore Technologies sequencing platform. The commercialization of these technologies open up the possibility of new applications, but these still need further development and improvements in order to compete with the widely used second-generation platforms.

6.3.1. Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) Sequencing

PacBio SMRT technology was introduced in 2010 and is so far the most established from its generation [15]. The SMRT technology does sequencing-by-synthesis, it records the fluorescent nucleotides

incorporated in individual template molecules, therefore not requiring previous amplification of fragments as in the second-generation technologies (Figure 6). The length of the reads produced can be up to 100 kb, although in average these are in the order of a few kilo basepairs (10 – 15 kb) [6,11,16].

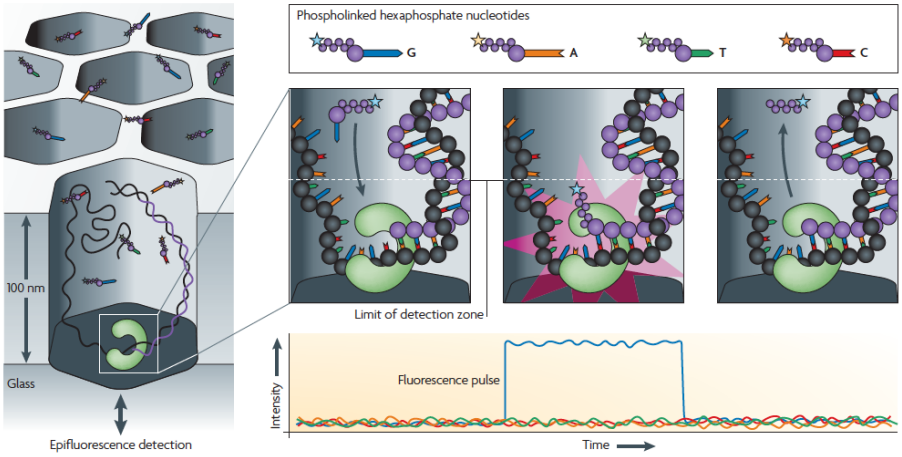


Figure 6 – PacBio SMRT Sequencing. Design of the SMRT cell and the principle behind single molecule real-time sequencing of the PacBio technology. The extremely small cells reduce the number of labeled molecules giving a more precise detection. The four-color nucleotides allow the real-time reading of the sequenced DNA. [13]

The occurrence of long reads is an advantageous in several applications, such as *de novo* sequencing of genomes, or analysis of repetitive sequences; nevertheless, the error rate in these reads is still quite high (approx. 15%). This method is fast, with running times of a few hours, but still low throughput and high cost, when compared to the widely used second-generation systems [6].

6.3.2. Illumina TruSeq Synthetic Long-Read Technology

In 2012, Illumina TruSeq Synthetic Long-Read technology entered the market. As the name suggests, it relies on the amplification of approx. 10 kb barcoded fragments from a DNA template. The barcoding of the fragments allows the synthetic generation of the long reads after sequencing on a short read instrument. The synthetic long reads present a very low error rate (approx. 0.1%), which stands out from the higher error introduced in PacBio sequencing [16].

This technology relies on long fragment amplification and is therefore biased by GC content and repetitive sequences, but also affected by early termination of amplification. Usage of this technology is even more expensive than PacBio, although it has already been used in few studies.

6.3.3. Oxford Nanopore Technologies Sequencing Platform

The latest technology introduced was Oxford Nanopore in 2014. The concept is simple, it makes use of a nanopore and electric current is applied. The current is monitored and as if forces single stranded molecules through the nanopore, different oscillations in the current correspond to the presence of different nucleotides in the nanopore (Figure 7). Several types of nanopores (biological and synthetic) are being tested, as well as improvements in the read measurements and software used.

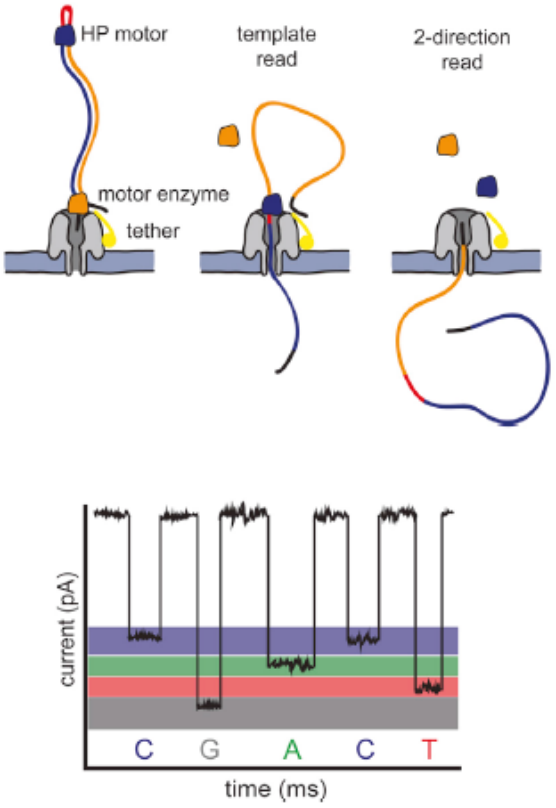


Figure 7 – Oxford Nanopore's Sequencing Strategy. DNA templates are ligated with two adapters. The first adaptor is bound with a motor enzyme and a tether. The second adaptor is a hairpin oligo that is bound by the HP motor protein. Changes in current that are induced as the nucleotides pass through the pore are used to discriminate bases. The library design allows sequencing of both strands of DNA from a single molecule (two-direction reads) [17]

This technology needs further development to reach its full potential, but it might become the cheapest and most portable sequencing method [6,16]. The further developments achieved will determine its usage worldwide. During the writing of this thesis several improvements and opportunities of usage have been reported

for this technology, including the field studying of the Ebola outbreak in Africa [18].

6.4. References

1. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).
2. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–7 (1977).
3. Maxam, a M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 560–564 (1977).
4. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
5. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
6. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30,** (2014).
7. Schloss, J. A. How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.* **26,** 1113–1115 (2008).
8. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–80 (2005).
9. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24,** 133–141 (2008).
10. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).
11. Mardis, E. R. Next-Generation Sequencing Platforms. *Annu. Rev. Anal. Chem.* **6,** 287–303 (2013).

12. Liu, L. *et al.* Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012,** (2012).

13. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

14. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348–352 (2011).

15. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).

16. Lee, H. *et al.* Third-generation sequencing and the future of genomics. *bioRxiv* (2016). doi:10.1101/048603

17. Reuter, J. a., Spacek, D. V. & Snyder, M. P. High-Throughput Sequencing Technologies. *Mol. Cell* **58,** 586–597 (2015).

18. Quick, J. *et al.* Real-time, portable genome sequencing for *Ebola* surveillance. *Nature* **530,** 228–232 (2016).

7. Research Articles

Paper 1 - Machado H., Sonnenschein E.C., Melchiorsen J. and Gram L. (2015) Genome mining reveals unlocked bioactive potential of marine Gram-negative bacteria, BMC Genomics (2015) 16:158.

Paper 2 - Machado H. and Gram L. (2015) The *fur* gene as a new phylogenetic marker in *Vibrionaceae* species identification, Appl. Environ. Microbiol. 81(8):2745-52.

Paper 3 - Machado H., Giubergia S., Mateiu R.V. and Gram L. (2015) *Photobacterium galatheae* sp. nov., a bioactive bacterium isolated from a mussel in the Solomon Sea, Int J System Evol Microbiol. 65(12):4503-7.

Paper 4 - Machado H., Cardoso J., Giubergia S., Rapacki K. and Gram L. (2016) Furios: a web-based tool for identification of *Vibrionaceae* species using the *fur* gene (submitted).

Paper 5 - Cavaleiro A.M.*, Machado H.*, Casado C., Kol S., Gram L. and Nørholm M.H.H. (2016) Marine bacterial cythochrome P450s and their potential in biotech (submitted). *equal contribution

Paper 6 - Machado H. and Gram L. (2016) Comparative genomics reveals evidence of high genomic exchange driving *Photobacterium* genus evolution (in preparation).

Paper 1

Genome mining reveals unlocked bioactive potential of marine Gram-negative bacteria

Machado et al. BMC Genomics (2015) 16:158
DOI 10.1186/s12864-015-1365-z

BMC Genomics

RESEARCH ARTICLE Open Access

Genome mining reveals unlocked bioactive potential of marine Gram-negative bacteria

Henrique Machado[1,2*], Eva C Sonnenschein[2], Jette Melchiorsen[2] and Lone Gram[2]

Abstract

Background: Antibiotic resistance in bacteria spreads quickly, overtaking the pace at which new compounds are discovered and this emphasizes the immediate need to discover new compounds for control of infectious diseases. Terrestrial bacteria have for decades been investigated as a source of bioactive compounds leading to successful applications in pharmaceutical and biotech industries. Marine bacteria have so far not been exploited to the same extent; however, they are believed to harbor a multitude of novel bioactive chemistry. To explore this potential, genomes of 21 marine Alpha- and Gammaproteobacteria collected during the Galathea 3 expedition were sequenced and mined for natural product encoding gene clusters.

Results: Independently of genome size, bacteria of all tested genera carried a large number of clusters encoding different potential bioactivities, especially within the Vibrionaceae and Pseudoalteromonadaceae families. A very high potential was identified in pigmented pseudoalteromonads with up to 20 clusters in a single strain, mostly NRPSs and NRPS-PKS hybrids. Furthermore, regulatory elements in bioactivity-related pathways including chitin metabolism, quorum sensing and iron scavenging systems were investigated both *in silico* and *in vitro*. Genes with siderophore function were identified in 50% of the strains, however, all but one harboured the ferric-uptake-regulator gene. Genes encoding the syntethase of acylated homoserine lactones were found in Roseobacter-clade bacteria, but not in the Vibrionaceae strains and only in one Pseudoalteromonas strains. The understanding and manipulation of these elements can help in the discovery and production of new compounds never identified under regular laboratory cultivation conditions. High chitinolytic potential was demonstrated and verified for *Vibrio* and *Pseudoalteromonas* species that commonly live in close association with eukaryotic organisms in the environment. Chitin regulation by the ChiS histidine-kinase seems to be a general trait of the Vibrionaceae family, however it is absent in the Pseudomonadaceae. Hence, the degree to which chitin influences secondary metabolism in marine bacteria is not known.

Conclusions: Utilizing the rapidly developing sequencing technologies and software tools in combination with phenotypic *in vitro* assays, we demonstrated the high bioactive potential of marine bacteria in an efficient, straightforward manner – an approach that will facilitate natural product discovery in the future.

Keywords: AntiSMASH, Genome mining, *Pseudoalteromonas*, Secondary metabolites, Vibrionaceae

Background

The discovery and development of new molecules for medical treatment is in great need as the 21st century unfolds. Drug-resistant pathogenic microorganisms are becoming a significant threat to public health and the pharmaceutical discovery pipelines have not been delivering the amount of new drugs required for efficient disease treatment [1-3]. Chemical synthesis has developed to be faster and cheaper as compared to biological screenings of organisms and extracts, however, chemical synthetic libraries have not provided the expected novel drugs and a high percentage of new chemicals that are introduced into the markets by pharmaceutical companies are actually derived from natural products [4]. Most of the natural products identified are produced by non-ribosomal peptide synthases (NRPSs) and/or polyketide synthases (PKSs) [5,6]. NRPSs and PKSs are multifunctional modular enzymes that assemble small molecules from monomers like pearls on a string. Both enzyme types

* Correspondence: henma@biosustain.dtu.dk
[1]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle Allè 6, DK-2970 Hørsholm, Denmark
[2]Department of Systems Biology, Technical University of Denmark, Matematiktorvet bldg 301, DK-2800 Kgs Lyngby, Denmark

have core domains responsible for the recognition of the monomer, attachment to the enzyme, condensation and chain-termination. Additionally, domains for tailoring the monomers can be present. In case of PKSs, such as in fatty acid synthesis, the monomers are acyl-CoAs, while NRPSs connect naturally occurring as well as unnatural amino acids to peptide chains. This wide range of possible subunits and the possibilities of their combinations lead to the great diversity of polyketides (PKs) and non-ribosomal peptides (NRPs) [7].

For the last century, soil microorganisms have been isolated and screened intensively to discover novel antibiotics and other drugs, and, in total, microorganisms have supplied more than 80.000 natural products [8]. Today, terrestrial *Streptomyces* is probably the best exploited genus with respect to secondary metabolites [9-13]. *Streptomyces* species produce a great diversity of compounds with antifungal (nystatin, natamycin, amphotericin), antibacterial (chloramphenicol, streptomycin, holomycin) and antiparasitic (ivermectin) activity [14]. Also new cultivation approaches are being used to culture new taxa, which potentially can be a source of novel compounds, as the recently described case of teixobactin [15].

Even though scientists have started to explore several other habitats than the terrestrial, the marine environment stands out as a hitherto under-explored niche for new bioactive molecules [6,16-19]. Previous studies have indicated that since the environmental conditions are very different from terrestrial habitats, novel compounds and chemical classes are present, and indeed some marine natural products are characterized by the unique marine factors such as halogenation [20-22]. Marine natural products have been isolated and identified from several different sources such as algae, sponges or molluscs, however, several recent studies have attributed the production of many of these compounds to microorganisms associated with the eukaryotic producer previously identified [23], bringing marine microorganisms to the spotlight of natural product discovery.

Following the success of terrestrial streptomycetes as producers of natural products, several researchers have focused their search on marine actinobacteria and the discovery of the first truly marine actinobacterium *Salinispora* has provided a number of very interesting bioactive compounds, including the anti-cancer compound salinosporamide [24,25]. Also, subsequent mining of the genome demonstrated an impressive number of potentially bioactive gene clusters [16]. The Gram-negative proteobacteria have generally been thought to have less potential for the production of bioactives than actinobacteria, however, several bioactive compounds have been isolated from the marine genus *Pseudoalteromonas* and more recently also from strains of the *Roseobacter* clade and the Vibrionaceae family [19,26-29].

Hitherto, the vast majority of bioactive compounds have been found using a classical bioassay-guided process, however, this bioprospecting of drugs is expensive and time-consuming, and re-discovery of known compounds is, despite dereplication steps, a major challenge. The process of drug discovery is currently undergoing changes as a result of the rapid developments in sequencing technology and synthetic biology. The number of whole microbial genomes and metagenomic data made publicly available is increasing exponentially and therefore, (meta) genome mining has become an extremely attractive tool for drug discovery [2,3,16,30,31]. It has led to the development of new bioinformatic tools used for screening and identification of the genetic background of the bioactivities including gene clusters responsible for the production of the novel molecules. Many of these clusters are probably silent under most laboratory culture conditions and require induction [32]. Several of the bioinformatic tools have been designed to search specifically for PKS and NRPS clusters, of which the structure is conserved. Several recent reviews provide a comparison between different tools, considering their *modus operandi* [30,31].

AntiSMASH version 2 is a strong comprehensive tool [30] and includes the use of several of the other tools available, such as the CLUSEAN tools [33], NRPSpredictor1/2 [34,35] and a method by Minowa et al. [36]. Even though the occurrence of misidentifications is quite common, it is preferable to "over-identify" rather than missing potential gene clusters [30]. Therefore, complementing antiSMASH analysis with more specific tools aids in the gene cluster identification. In this study, we used three other tools: BAGEL3 for the identification of bacteriocins [37]; NapDos for the identification of keto-synthase (KS-domains) and condensation domains (C-domains) [38]; and NP.search for the identification of whole gene clusters that may be composed of several KS- and/or C-domains [39]. C- and KS-domains catalyze the chain formation of the subunits (peptides or acyl-CoAs), respectively and a high number of these domains reflects the richness of bonds possibly made by an organism and the degree of diversity on non-ribosomal peptide synthesis.

The strains investigated in this genome mining study were isolated during the Galathea 3 global expedition in 2006/7. Antagonistic activity towards the human pathogen *Staphylococcus aureus* and the fish pathogen *Vibrio anguillarum* were the main selection criteria [19]. The Galathea 3 bacterial collection has been used in previous studies where identification of new bioactive compounds has been successful. For instance, *Photobacterium halotolerans* strain S2753 produces novel compound families, the solonamides and ngercheumicins, which interfere with virulence regulation in *S. aureus* [40-42]. *Vibrio nigripulchritudo* strain S2604 produces a novel siderophore:

nigribactin [43]. However, also several known antibiotic compounds were re-discovered, for instance, S2753 produces holomycin [28], an antibiotic previously only isolated from terrestrial streptomycetes, and *Vibrio coralliilyticus* S2052 produces andrimid [28]. Also, in pigmented *Pseudoalteromonas*, we have re-identified a range of antibiotic compounds (indolmycin, pentabromopseudilin, prodigiosin) [44,45].

During the last five years, we have demonstrated that marine Gram-negative bacteria produce an array of antibiotic and anti-virulence compounds [19,28,29,40,41,43-45] and here, we ask the question if the classical bioprospecting approach had fully revealed the potential of these bacteria. We present an *in silico* study of different marine bacterial genomes, which were analyzed using several of the prediction tools developed for the identification of secondary metabolism pathways, namely antiSMASH, NapDos, Np.search, and BAGEL3 [37,46,47]. We combined the genome mining with phenotypic evaluation of molecules potentially involved in production or regulation of bioactive compounds; namely, quorum sensing signals, siderophores and chitinases.

Results and discussion

Marine bacterial genomes – genome size

The genomes were assembled using CLC Genomics Workbench 7 (CLC bio, Aarhus, Denmark) to obtain contig-based draft genomes of the strains. These draft genomes were then annotated using the Rapid Annotation using Subsystem Technology (RAST) [48,49]. The subsequent analysis of the genomes was performed using CLC Main Workbench 7 (CLC bio, Aarhus, Denmark).

The genome size varied between 3.6 and 6.2 Mb in the 21 sequenced strains (Table 1). In the six Vibrionaceae, the genomes varied between 4 and 6.2 Mb, and the genomes of the eight *Pseudoalteromonas* spp. ranged from 4.1 to 6.1 Mb. The genomes of the five strains from the Rhodobactereaceae family were slightly smaller; from 3.6

Table 1 Potential for production of bioactive secondary metabolites from 21 marine bacterial strains

Strain	Species	Genome size (Mb)	Antibacterial activity	AntiSMASH (total hits)	BAGEL3	NapDos		NP.search			
						KS-domains	C-domains	NRPS	PKS	Mix	Trans PKS
S2753	*Photobacterium halotolerans*	5.5	yes	12	0	3	19	1	0	1	0
S2052	*Vibrio coralliilyticus*	5.4	yes	7	2	7	13	2	0	2	0
S2043	*Vibrio coralliilyticus.*	5.4	yes	7	2	7	13	2	0	2	0
S2604	*Vibrio nigripulchritudo*	6.2	yes	9	0	6	17	1	0	0	0
S2394	*Vibrio neptunius*	5.2	yes	6	1	4	12	1	0	1	0
S2757	*Vibrio* sp.	4.0	no	2	0	5	0	0	0	0	0
S2040	*Pseudoalteromonas piscicida*	5.3	yes	14	1	8	58	7	0	1	0
S2724	*Pseudoalteromonas piscicida*	5.2	yes	10	1	7	30	2	0	2	0
S816	*Pseudoalteromonas agarivorans*	4.4	no	2	0	5	0	0	0	0	0
S3258	*Pseudoalteromonas ruthenica*	4.1	yes	3	0	5	0	0	0	0	0
S3137	*Pseudoalteromonas ruthenica*	4.1	yes	3	0	5	0	0	0	0	0
S4054	*Pseudoaltermonas luteoviolacea*	6.1	yes	20	1	14	48	3	0	4	1
S2471	*Pseudoalteromonas rubra*	5.8	yes	17	2	12	56	3	0	2	1
S2151	*Halomonas* sp.	5.2	no	5	0	7	0	0	0	0	0
S3726	*Marinomonas* sp.	5.4	yes	5	0	6	17	2	0	0	0
S2292	*Spongiobacter* sp.	4.7	yes	5	1	3	3	0	0	0	0
S4079	*Loktanella* sp.	3.6	no	5	1	3	3	0	0	0	0
S4493	*Paracoccus* sp.	4.0	yes	11	1	3	2	0	0	0	0
S1942	*Ruegeria mobilis*	4.8	yes	8	1	4	1	0	0	0	0
F1926	*Ruegeria mobilis*	4.6	yes	9	0	5	1	0	0	0	0
DSM17395	*Phaeobacter inhibens*	3.8	yes	9 + 1	0	4	1	1	0	0	0

to 4.8 Mb. The *in vitro* bioactivity (antibacterial activity measured as zone size) [19] did not correlate to the genome size (Table 1).

It has been suggested that the potential for production of secondary metabolites would be related to genome size [11,50,51], with a larger genome allowing more genes to be allocated to secondary metabolism. This notion was to some extend developed by studies of the genus *Streptomyces* which is a prolific producer of secondary metabolites and has relatively large genomes of approx. 8 Mb in size as compared to other bacteria. This understanding is changing, as the marine actinomycete *Salinispora* sp. has a genome size of approx. 5 Mb, of which approx. 10% is dedicated to secondary metabolism, whereas only approx. 8% of the genome of *Streptomyces coelicolor* has been reported as dedicated to secondary metabolism [11,16].

Identification of gene clusters potentially encoding secondary metabolites

The genomes were mined using bioinformatic tools for the identification of clusters involved in secondary metabolism, namely antiSMASH, NapDos, Np.search, and BAGEL3 [37-39,47]. We found a high genetic potential for secondary metabolite production also in Gram-negative marine bacteria with genome sizes ranging from 4 to 6 Mb, with some strains reaching the considerable number of eight distinct PKS/NRPS clusters (Table 1 – NP.search). However, some strains with similar genome size harbored none or very few potential bioactive clusters and thus, there was no clear correlation between genome size and number of secondary metabolism gene clusters. Some strains, such as *V. nigripulchritudo* S2604 or *Halomonas* sp. S2151, with larger genomes had a low number of hits; and also contrarily, strains with smaller genomes had a greater number of hits e.g. *P. piscicida* strains S2040 and S2724 (Table 1).

Bioactivity potential - NRPS/PKS

The presence of gene clusters likely encoding bioactive compounds is spread among the different families of Alpha- and Gammaproteobacteria. Although our collection is limited in number, it appears that the Gammaproteobacteria class is richer in NRPS and PKS clusters than the Alphaproteobacteria. The analysis using NapDos and NP.search, in general, identified the same number of potential bioactive gene clusters. A higher frequency of KS- and C-domains was identified in pigmented *Pseudoalteromonas* strains (S2040; S2724; S4054; S2471) followed by Vibrionaceae, with the exception of S2757 (no hits), and *Marinomonas* sp. S3726 (high number of hits).

Some species in the Rhodobactereaceae family (*Ruegeria mobilis* and *Phaeobacter inhibens*) are capable of inhibiting a wide range of other bacteria [52-54]; however, in general, few secondary metabolites have been identified in these strains [54-57]. Here, we show that using bioinformatics tools a few clusters could be identified, but still the bioactive potential harbored in the genome of these genera appears much lower than that observed in Gammaproteobacteria.

A number of strains that were not antagonistic in agar-based assays were included in the analysis and these contained only few gene clusters potentially coding for secondary metabolites (Table 1). This was the case for *Vibrio* sp. S2757 and *P. agarivorans* S816, for which antiSMASH identified only two potential clusters (Table 1).

Another interesting group of strains included those that received five hits in total in the antiSMASH analysis. This includes the bioactives *Marinomonas* sp. S3726 and *Spongiobacter* sp. S2292 and the non-bioactives *Halomonas* sp. S2151 and *Loktanella* sp. S4079. Although, all of them had a considerably lower number of hits in the antiSMASH analysis than the pigmented pseudoalteromonads and the vibrios, the results of the other mining tools (NaPDoS) demonstrate that *Marinomonas* sp. S3726 has a great potential with 6 KS- and 17 C-domains identified (Table 1). Thus, the sole number obtained by one given analysis tool may not reflect the whole potential of the organism, and complementary analysis should be performed to ensure discovery of the full bioactive potential. This should also be done to avoid further work on clusters that may not be true secondary metabolite clusters, it appears from the analysis that NapDos and NP.search tools seemed to identify only a subset of the NRPS/PKS clusters identified by antiSMASH.

The potential for secondary metabolite production in the strains is clearly much larger than so far identified by bioassay-guided fractionation. For instance, the PK/NRP hybrid andrimid has been identified as the bioactive compound in *V. coralliilyticus* S2052 [28,29,58,59]. The genome mining identified the gene cluster likely encoding for andrimid production genes (Figure 1(A)). Also, we found at least three more NRPS clusters using antiSMASH, NapDos and NP.search (Table 1). Similarly, in *P. halotolerans* S2753, the dithiolopyrrolone holomycin was identified in extracts [28] and the corresponding gene cluster was found by the bioinformatic tools used (Figure 1(B)); again, four more NRPS/PKS clusters were found using antiSMASH, although only one more was discovered when using NP.search (Table 1). As indicated, we and others have identified several bioactive compounds from pigmented pseudoalteromonads and here we also identified the respective gene clusters for indolmycin [44], violacein [60] and pentabromopseudilin [5,61,62]. However, the pigmented pseudoalteromonads contained a large number of potential bioactive clusters, including a very high number of C-domains as compared to the other studied strains (NapDos – Table 1).

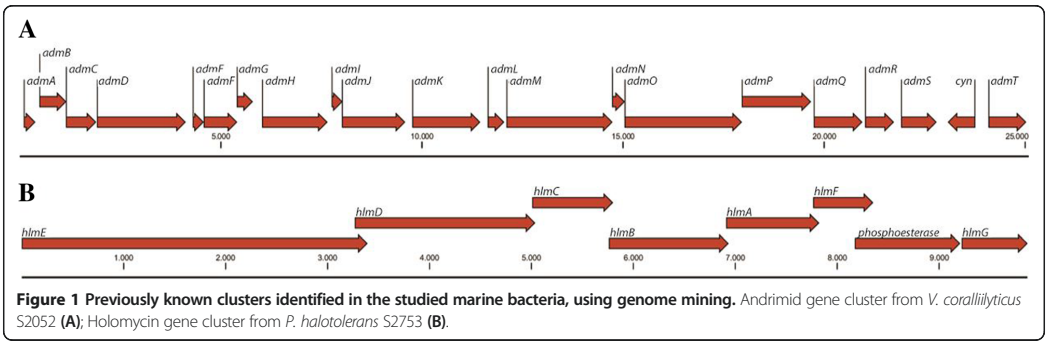


Figure 1 Previously known clusters identified in the studied marine bacteria, using genome mining. Andrimid gene cluster from *V. coralliilyticus* S2052 **(A)**; Holomycin gene cluster from *P. halotolerans* S2753 **(B)**.

The case of Pseudoalteromonas ruthenica

In *in vitro* assays, *Pseudoalteromonas ruthenica* is highly antagonistic against *S. aureus* and *V. anguillarum* causing large clearing zones in agar-based screening assays [19]. However, we have not been able to identify the compound(s) responsible for this inhibition by bioassay-guided fractionation and anticipated that genome mining would reveal potential bioactive gene clusters. AntiSMASH identified three gene clusters (one for siderophore and two for bacteriocin biosynthesis), but only the siderophore cluster was correctly identified, whereas the bacteriocin-related clusters were misidentified and encoded the flagella operon and a cluster encoding for hypothetical proteins, a muramoyltetrapeptide carboxypeptidase and a 2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase, involved in the biosynthesis of peptidoglycan and lysine, respectively.

A second analysis of the *P. ruthenica* strains with antiSMASH based on PFAM domain probabilities increased the number of potential gene clusters from three to thirteen. Mainly clusters encoding for acyl carrier proteins were identified, but we also identified some biosynthetic clusters such as lipopolysaccharide, capsular polysaccharide, legionaminic acid and fatty acid biosynthesis. From all the clusters, only one matched with the RAST annotation as behaving an open reading frame (ORF) encoding a non-ribosomal peptide synthase. Yet this ORF was only 663 bp, and when we blasted the predicted aminoacid sequence against the NCBI protein database, it presented a high similarity with a methionyl-tRNA formyltransferase and not to an NRPS. In agreement with these were NapDos and NP.search, which did not identify any potential bioactive clusters (Table 1). This reduces the likelihood that the clusters identified by antiSMASH using PFAM domains are actually clusters responsible for the production of bioactives.

In genome mining, the identification of clusters likely involved in secondary metabolism, such as NRPS and PKS, have been used as a measure of the potential for finding novel natural bioactive compounds, including antibiotics [63]. Yet, all the bioinformatic tools used to search for the biosynthetic capabilities and potential of *P. ruthenica* failed. This might be the case because the antagonistic activity is due to other biosynthetic pathways as is for instance the case with the antibiotic tropodithietic acid produced by some *Roseobacter* clade bacteria [64]. Also, it can be attributed to limitations in the prediction algorithms. The prediction algorithms of the bioinformatic tools are to some extend based on identification of known biosynthetic activities and one could speculate that truly novel biosynthetic pathways would not be identified. To identify the core genes of a biosynthetic pathway, most of the tools available use profile-HMMs or alignments of conserved domains in biosynthetic enzymes [30]. This is a problem in the identification of non-standard pathways and antiSMASH has therefore implemented an algorithm to identify the distribution of protein domains usually associated with secondary metabolites [30], increasing the probability of identification of clusters responsible for secondary metabolites production. This not only increases the number of hits, but also the time needed for evaluation of the clusters, raising the question of the feasibility of using genome mining in groundbreaking discoveries.

Bacteriocins

The number of clusters identified by antiSMASH as bacteriocins varied between one and five in each strain, with an average of two clusters per strain. However, when the genomes were analyzed using the prediction tool BAGEL3 [37], the presence of bacteriocin-related genes was only confirmed in a few strains. The distribution of bacteriocin clusters did not follow a particular pattern with respect to genera or species.

It seems evident that the specific prediction tools are more accurate in identifying their defined target; therefore, BAGEL 3 being most probably a better indicator of the number of bacteriocin-related genes than antiSMASH

itself. This becomes more evident from the *P. ruthenica* case, where random genes were classified as bacteriocins (see above).

Acyl homoserine lactones

Four of the 21 strains induced a clear response in the AHL (acyl homoserine lactone) biomonitors (Table 2). Three strains, *Vibrio* sp. S2757, *Paracoccus* sp. S4493 and *P. luteoviolaceae* 4054 induced both monitors whereas *P. inhibens* DSM17395 induced only *A. tumefaciens*. This is in agreement with previous studies where also *Phaeobacter* sp. strain S27-4 induced *A. tumefaciens* and chemical analysis identified 3-hydroxy-decanoyl-homoserine lactone [64]. Interestingly, antiSMASH detected homoserine lactone synthases in three of these four strains but not in *Vibrio* S2757. The response in the monitor strains could be caused by other compounds, such as diketopiperazines that have been demonstrated to induce the AHL monitors [65]. The same could be true for the extracts of the *V. coralliilyticus* strains S2052 and S2043, which resulted in a weak reaction in *C. violaceum*, and the genomes did not contain an AHL synthase gene. AntiSMASH detected AHL synthase genes in three strains (*Loktanella* sp. and two *Ruegeria mobilis*) where no AHLs were detected by the monitors (Table 2). These genes could potentially encode novel AHLs not being in the detection range of the used biological monitors [66]. On the other hand, the bacteria may not have been cultured under conditions allowing the expression of the presumed AHL synthase genes or the AHL concentration produced was below the detection limit. We considered if the potential QS systems could be involved in production of secondary metabolites. In *P. luteoviolaceae*, the AHL synthase gene is adjacent to the gene cluster potentially involved in indolmycin production [67], but in the other five strains the HSL synthase genes detected by antiSMASH were not in proximity to identified natural product gene clusters. However, some were close to genes encoding acyl synthases, alcohol dehydrogenases or proteins containing AMP-binding domains, which may potentially be involved in secondary metabolism. Due to draft genomes with multiple contigs, the association with natural product gene clusters could have been lost in the analyzed sequences.

Table 2 Iron system in the studied strains, comprising *in silico* and phenotypical results

Strain	Species	Response in AHL monitor		AntiSMASH	Siderophore (CAS)	AntiSMASH		Fur
		Cv	At	HSL		Siderophore	NRPS	
S2753	*Photobacterium halotolerans*	-	-	0	+	1	4	1
S2052	*Vibrio coralliilyticus*	(+)	-	0	-	1**	4***	1
S2043	*Vibrio coralliilyticus*	(+)	-	0	(+)	1**	4***	1
S2604	*Vibrio nigripulchritudo*	-	-	0	-	0	4***	1
S2394	*Vibrio neptunius*	-	-	0	(+)	1	3***	1
S2757	*Vibrio sp.*	+	+	0	+	1**	0	1
S2040	*Pseudoalteromonas piscicida*	-	-	0	+	0	11***	1
S2724	*Pseudoalteromonas piscicida*	-	-	0	+	0	5***	1
S816	*Pseudoalteromonas agarivorans*	-	-	0	+	1	0	1
S3258	*Pseudoalteromonas ruthenica*	-	-	0	+	1	0	1
S3137	*Pseudoalteromonas ruthenica*	-	-	0	-	1	0	1
S4054	*Pseudoaltermonas luteoviolacea*	+	+	1	(+)	0	11***	1
S2471	*Pseudoalteromonas rubra*	-	-	0	(+)	0	9***	1
S2151	*Halomonas* sp.	-	-	0	(+)	1	0	1
S3726	*Marinomonas* sp.	-	-	0	+	0	3***	1
S2292	*Spongiobacter* sp.	-	-	0	-	0	1	1
S4079	*Loktanella* sp.	-	-	1	(+)	0	1***	1
S4493	*Paracoccus* sp.	+	+	4	-	0	2	0
S1942	*Ruegeria mobilis*	-	-	2	(+)	0	1***	1
F1926	*Ruegeria mobilis*	-	-	2	(+)	0	1***	1
DSM17395	*Phaeobacter inhibens*	-	+	2	+	1*	1	1

Cv: *Chromobacterium violaceum*, At: *Agrobacterium tumefaciens*, HSL: homoserine lactone, CAS: chrome-azurol-S, + : strong bioactivity, (+) : weak bioactivity, – : no bioactivity detected under the tested conditions, NRPS: including single NRPS clusters and NRPS fusion clusters (e.g. NRPS-bacteriocin, NRPS-ectoine). *Located on a plasmid; **Cluster identified as a siderophore – ectoine cluster; ***At least one NRPS is in proximity to siderophore-associated genes (tonB-dependent receptor etc.).

Siderophores and iron regulation

Iron is essential for almost all microorganisms being required for key biological processes [68] and is also one of the most important requirements for successful secondary metabolism. The iron levels in seawater are extremely low, and many marine bacteria are able to sequester iron using siderophores that can also serve as a tool in microbial competition. Hence, siderophores are included as secondary metabolites in the antiSMASH search.

To complement the genetic search, we determined siderophore activity using the CAS assay [69]. Pronounced siderophore activity was detected in eight strains and a weak reaction was observed in eight further strains. Only five strains did not show any activity under the tested conditions (Table 2). The NRPS prediction tools, NapDos and NP.search, do not allow detailed prediction of the type of NRPS coding gene, however, antiSMASH is able to distinguish siderophore synthesis genes. The *in silico* analysis using antiSMASH identified putative siderophore gene clusters in five of the eight strains with a clear CAS reaction, and three of the eight with a weak reaction. In one strain, *P. ruthenica*, antiSMASH detected a siderophore synthesis gene, but the CAS assay was negative. In contrast, the CAS reaction was positive for three strains (two *P. piscicida* and one *Marinomonas*) where a siderophore biosynthesis gene was not detected. AntiSMASH predicts siderophore genes using the currently available sequence information on siderophore-producing NRPSs, which are mainly of terrestrial origin. Terrestrial siderophores differ structurally from marine siderophores that are usually associated with fatty acids [68]. We analysed the identified NRPS gene clusters for siderophore-associated genes such as *tonB*-dependent receptor genes. For twelve strains, these siderophore-associated genes were found close to the NRPS gene leading to the hypothesis that this NRPS gene could likely encode a siderophore-producing NRPS. This would demonstrate that all strains based on their genetic information would be capable of scavenging iron using siderophores. To detect this "hidden" activity for the five non-active strains, the strains might require optimization of culture conditions or certain biological cues from the environment. Iron can also be scavenged by other molecules and non-siderophore iron sequestering systems may be operational in the bacteria where siderophore genes were not detected. Indeed, several heme-related proteins were identified among the studied marine bacteria by an annotation-based search (data not shown).

Even though iron is essential for growth, excess of iron can be toxic to bacteria and thus a tight regulation of uptake is crucial for microbial survival [68]. In Gram-negative bacteria, iron regulation is achieved by a repressor protein named Fur (Ferric-iron uptake regulator) which acts at the transcriptional level [70]. A Fur encoding gene could be identified in all the studied strains and the amino acid sequence predicted, with the sole exception of *Paracoccus* sp. S4493 (Table 2; Figure 2(A)). The verified exception of *Paracoccus* sp. S4493 might be due to sequencing limitations, or the fact that this organism has another regulatory protein involved in iron sensing; in fact other uptake regulators for different metals could be identified (e.g. manganese, potassium, zinc, and nickel).

Within the classes of Proteobacteria, the Fur proteins are relatively conserved at the amino acid level, presenting a higher variation at the C-terminus and the N-terminus.

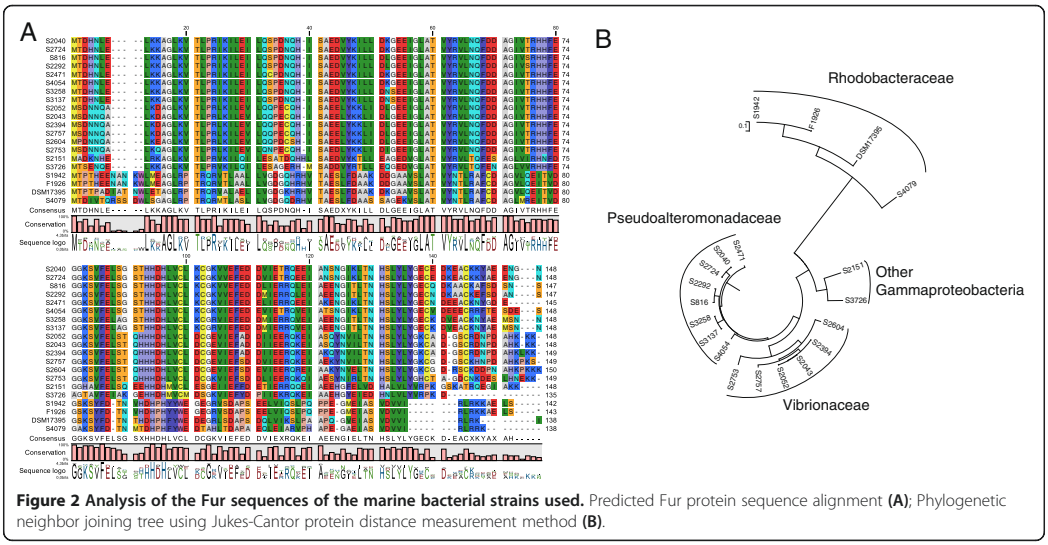


Figure 2 Analysis of the Fur sequences of the marine bacterial strains used. Predicted Fur protein sequence alignment **(A)**; Phylogenetic neighbor joining tree using Jukes-Cantor protein distance measurement method **(B)**.

Nevertheless, the conserved features such as DNA-binding-α-helix and Fe^{2+} and Zn^{2+} binding domains could be identified [71]. A neighbor joining tree using Jukes-Cantor protein distance measurement method (Figure 2(B)) demonstrates the conservation of closely related species, indicating that the *fur* gene is a phylogenetic trait instead of a random species variation or a product of recent horizontal gene transfer. In fact, the clusters based on protein sequences follow the phylogenetic distribution; the analyzed Alphaproteobacteria sequences form a separate, distant group from the families of Gammaproteobacteria, in which the Pseudoalteromonadaceae and the Vibrionaceae families form two distinct clusters from the other Gammaproteobacteria. The only exception here was the *Spongiobacter* sp. S2292, which clustered together with the *Pseudoalteromonas* spp. This is interesting, since the 16S rRNA sequence (GenBank acc. no. FJ457273.1) would place *Spongiobacter* sp. S2292 closer to the *Endozoicomonas* genus and therefore within the order of Oceanopirillales, in which the species *Halomonas* and *Marinomonas* are also included [72]. This fact brings to question the phylogenetic placement of *Spongiobacter* as it remains an unclassified member of Gammaproteobacteria, and this association indicates a closer association with *Pseudoalteromonas* species than with the other Gammaproteobacteria.

Interestingly, this clustering seems to be specific at species level, even for the *Vibrio* strains studied. Identifying *Vibrio* strains to species level typically requires multilocus sequence analysis [73,74]. We recently showed that the *fur* gene is a good phylogenetic marker (Machado & Gram, submitted) to be added to the multilocus sequencing analysis performed nowadays in e.g. *Vibrio* species definition [73-75] and might also be possibly used in other genera for species differentiation.

Chitinases and regulation

Chitin is — after cellulose — the most abundant carbon source on Earth. Enzymes capable of degrading this organic compound are very useful in biotechnological industries. At the same time, chitin is also an important environmental clue influencing regulators of virulence and secondary metabolism [58,76-78]. We have previously shown that an andrimid producing *V. coralliilyticus* S2052 focuses its secondary metabolism exclusively on andrimid when grown on chitin as compared to growth on glucose and casamino acids [58]. This could be coupled with transcriptional changes and we therefore also mined the genomes for chitin catabolic cascade sensor histidine kinase (ChiS) and chitin binding proteins (CBP).

By phenotypic assays, we identified several strongly chitinolytic strains and screened their genomes for chitinase encoding genes. All of the Vibrionaceae and pigmented *Pseudoalteromonas* sp., with the exception of *P. agarivorans* S816, were capable of degrading chitin (Table 3). The genome mining revealed presence of three to nine chitinase encoding genes per strain in the chitinolytic bacteria.

Table 3 Chitinolytic systems in the studied strains, comprising *in silico* and phenotypic results

Strain	Species	Chitinase activity	Chitin		
			Chitinase	ChiS	CBP
S2753	*Photobacterium halotolerans*	+++	3	1	1
S2052	*Vibrio coralliilyticus*	++	9	1	2
S2043	*Vibrio coralliilyticus*	++	9	1	2
S2604	*Vibrio nigripulchritudo*	+	8	2	0
S2394	*Vibrio neptunius*	++	7	1	0
S2757	*Vibrio sp.*	++	3	1	1
S2040	*Pseudoalteromonas piscicida*	++	4	1	0
S2724	*Pseudoalteromonas piscicida*	+++	4	0	0
S816	*Pseudoalteromonas agarivorans*	-	0	0	0
S3258	*Pseudoalteromonas ruthenica*	++	3	0	1
S3137	*Pseudoalteromonas ruthenica*	++	3	0	1
S4054	*Pseudoaltermonas luteoviolacea*	+	10	0	0
S2471	*Pseudoalteromonas rubra*	+	7	0	0
S2151	*Halomonas* sp.	-	0	0	0
S3726	*Marinomonas* sp.	-	0	0	0
S2292	*Spongiobacter* sp.	-	0	0	1
S4079	*Loktanella* sp.	-	0	0	0
S4493	*Paracoccus* sp.	-	0	0	0
S1942	*Ruegeria mobilis*	-	0	1	0
F1926	*Ruegeria mobilis*	-	0	1	0
DSM17395	*Phaeobacter inhibens*	-	0	0	0

ChiS: chitin catabolic cascade sensor histidine kinase, CBP: chitin binding proteins. - : no chitinase activity detected, + : low chitinase activity, ++ : medium chitinase activity, +++ : strong chitinase activity.

Chitin-related genes were present in *Vibrio* species, which is likely related to their ecology and close association with crustaceans [79,80]. The pigmented pseudoalteromonads are also often associated with eukaryotic surfaces [44] including organisms containing no chitin. However, several pseudoalteromonads had genes encoding for chitinases and showed prominent chitinolytic activity.

The chitinolytic cascade has previously been studied in *Vibrio* species where its tight regulation was attributed to the hybrid chitin catabolic sensor/kinase (ChiS) together with a periplasmic chitin oligosaccharide binding protein (CBP) [77]. This regulatory system has been shown to regulate expression of 50 genes, most of which involved in chitin catabolism [77]. Furthermore, it has

been also shown that natural substrates such as chitin influence secondary metabolite production, such as the induction of their production [58]. We searched for the *chiS* gene, which was present in all of the six Vibrionaceae, one *Pseudoalteromonas*, and two *Ruegeria* (Table 3). The Alphaproteobacteria did not degrade chitin, though two *Ruegeria mobilis* strains harbor the chitin sensor genes.

Interestingly, the ChiS regulator was only present in the *Vibrio* strains, suggesting that transcriptional shaping by chitin could be a trait associated with this family. Changes in secondary metabolism by chitin and the presence of the regulator ChiS requires further studies for confirmation.

Conclusions

Here, we presented a straightforward, comprehensive genome mining approach analyzing marine bacterial strains for secondary metabolism and associated features such as quorum-sensing, iron acquisition, chitin use as a carbon source and its regulation. The use of complementary tools for genome mining is of great value in narrowing down the potential gene clusters from a large pool obtained by broad prediction software such as antiSMASH. We demonstrated the great potential of marine bacteria for secondary metabolite production, with special focus on *Vibrio* and pigmented *Pseudoalteromonas* species.

Methods

Bacterial strains and growth conditions

Bacterial strains used in this study were isolated during the Danish Galathea 3 global research expedition (http://www.galathea3.dk/uk) [19] and selected due to their antagonistic activity against a Gram-negative (*Vibrio anguillarum*) and a Gram-positive (*Staphylococcus aureus*) pathogenic bacterium. Pure cultures of each strain were stored in cryoprotectant solution at −80°C from their isolation until the present study. *Phaeobacter inhibens* DSM17395 was obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ, Germany). Some of the strains have previously been used in classical bioassay-guided bioprospecting and produce antibiotics or anti-virulence compounds [19,22,27-29,40-45,58]. Strains were routinely grown on Marine Agar (Difco 2216) and in Marine Broth (Difco 2216).

Genomic DNA isolation and Sequencing

High purity genomic DNA was extracted by successive phenol:chloroform:isoamyl-alcohol purification steps followed by precipitation with isopropanol, treatment with RNase and a final purification and precipitation step [81]. Quantification was done in 1% agarose gel electrophoresis, NanoDrop Spectrometer (Saveen Werner, Sweden) and Qubit 2.0 Analyser (Invitrogen, United Kingdom). Sequencing of the genomes was performed by Beijing Genomic Institute (Shenzhen, China). Libraries of 500 bp were used for 100 bp paired-end sequencing of genomes using the Illumina sequencing technology on a HiSeq2000 with a minimum coverage of 100. Genomic DNA sequences were assembled in contigs using CLC Genomic Workbench (CLC Bio, Aarhus, Denmark). All the genomes had a coverage of 75x or higher. All of them were submitted to the National Center for Biotechnology Information (NCBI) database under the accession numbers AUXW00000000, JMIB00000000, APME00000000, AQCH00000000, CP002972, CP002973, CP002974, CP002975, JXXR00000000, JXXS00000000, JXXT00000000, JXXU00000000, JXXV00000000, JXXW00000000, JXXX00000000, JXXY00000000, JXXZ00000000, JXYA00000000, JXYB00000000, JXYC00000000, JXYD00000000, JXYE00000000, JXYF00000000, JXYG00000000.

Bioinformatic analysis

The draft genomes were annotated using RAST [49] and submitted to secondary metabolite gene cluster analysis using antiSMASH 2.0 [47], NapDos [38], NP.search [39], as well as to the bacteriocin-specific software BAGEL 3 [37]. Following RAST annotation, a homology search was conducted on the *ferric-iron uptake regulator* gene *fur* and an annotation-based search was performed for genes encoding, chitinases and the chitin catabolic cascade sensor gene *chiS*.

Verification of antibacterial activity

The strains were re-tested for their antibacterial activity, as previously described [19]. Briefly, strains to be tested were grown in Marine Agar (Difco 2216) for 24 – 48 h and one colony was spotted in plates of artificial seawater agar with 3% Instant Ocean (IO; Aquarium Systems Inc., Sarrebourg, France) containing *Vibrio anguillarum* strain 90-11-287 serotype O1 [82] or *Staphylococcus aureus* strain 8325 [83] embedded. The plates were incubated and observed for clearing zones in the agar.

Production of acyl homoserine lactones

Production of acyl homoserine lactone (AHL) compounds was analysed using two AHL monitor systems *Agrobacterium tumefaciens* NT1(pZLR4) [84] and *Chromobacterium violaceum* CV026 [85] as described by Ravn et al. [86]. The strains were grown in 10 mL ½YTSS or sea salt medium (1.5% sea salt, 0.3% casamino acids, 0.4% glucose) in 50 mL Falcon tubes for 48 hours at 200 rpm and room temperature and extracted with 10 mL ethyl acetate acidified with 1% formic acid. The extract was dried under nitrogen, resuspended in 0.5 mL ethyl acetate containing 1% formic acid and stored at −20°C. The extracts were tested with the AHL-reporter strains in a plate well assay [87].

Siderophore activity

Siderophore activity was tested using the liquid CAS assay [69]. The marine strains were grown in 10 mL sea salt medium or ½YTSS in 50 mL Falcon tubes at 25°C and 200 rpm for 24 and 48 hours at room temperature. 1 mL of culture was centrifuged for 5 min at 12,100 × g and the supernatant was mixed with CAS solution in a 1:1 ratio. Colour change from blue to orange indicating siderophore activity was observed after 5 min and 24 h.

Chitinase activity

Chitinase activity was tested on chitin containing agar plates. Strains were grown on Marine Agar (Difco 2216) for 24 – 48 h and one colony was spotted on plates containing 20 g/L sea salts, 3 g/L casamino acids, 0.08% hydrolyzed chitin, 20 g/L agar. The plates were incubated for 72 h and chitinase activity monitored at 24, 48 and 72 h. The natural turbidity of the media due to chitin allows the visual evaluation of chitin degradation, which leads to clearance of the media. Chitinase activity was graded qualitatively: low chitinase activity (<1.0 mm) zones were scored with one plus, medium chitinase activity zones (1.0 mm – 3.0 mm) with two pluses, and strong chitinase activity (>3.0 mm) with three plusses.

Abbreviations

NRPS: Non-ribosomal peptide synthase; PKS: Polyketide synthase; KS-domains: Keto-synthase domains; C-domains: Condensation domains; RAST: Rapid Annotation using Subsystem Technology; PFAM: Protein families database; ORF: Open reading frame; profile-HMMs: Profile hidden Markov models; AHL: Acyl homoserine lactone; QS: Quorum sensing; HSL: Homoserine lactone; CAS: Chrome azurol S; Fur: Ferric-iron uptake regulator; ChiS: Chitin catabolic cascade sensor histidine kinase; CBP: Chitin binding protein.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HM and LG designed the study and HM, ESC and JM carried out the experiments. All authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

HM was supported by a PhD grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7-People-2012-ITN, under grant agreement No. 317058, "BACTORY". ECS was supported by the European Union's Seventh Framework Programmes KBBE-2012-6-311975 MaCuMBA (Marine Microorganisms: Cultivation Methods for Improving their Biotechnological Applications) and KBBE-2012-6-312184 PharmaSea (Increasing Value and Flow in the Marine Biodiscovery Pipeline). The present work was carried out as part of the Galathea 3 expedition under the auspices of the Danish Expedition Foundation. This is Galathea 3 contribution no. p110.

Received: 16 January 2015 Accepted: 20 February 2015
Published online: 07 March 2015

References

1. Xu J, Hagler A. Chemoinformatics and drug discovery. Molecules. 2002;7:566–600.
2. Scheffler RJ, Colmer S, Tynan H, Demain AL, Gullo VP. Antimicrobials, drug discovery, and genome mining. Appl Microbiol Biotechnol. 2013;97:969–78.
3. Zerikly M, Challis GL. Strategies for the discovery of new natural products by genome mining. Chembiochem. 2009;10:625–33.
4. Newman DJ, Cragg GM. Natural products as sources of new drugs over the last 25 years. J Nat Prod. 2007;70:461–77.
5. Moore BS. Biosynthesis of marine natural products: microorganisms (Part A). Nat Prod Rep. 2005;22:580–93.
6. Xiong Z-Q, Wang J-F, Hao Y-Y, Wang Y. Recent advances in the discovery and development of marine microbial natural products. Mar Drugs. 2013;11:700–17.
7. Meier JL, Burkart MD. The chemical biology of modular biosynthetic enzymes. Chem Soc Rev. 2009;38:2012–45.
8. Bérdy J. Thoughts and facts about antibiotics: where we are now and where we are heading. J Antibiot. 2012;65:441–41.
9. Yu D, Xu F, Valiente J, Wang S, Zhan J. An indigoidine biosynthetic gene cluster from *Streptomyces chromofuscus* ATCC 49982 contains an unusual IndB homologue. J Ind Microbiol Biotechnol. 2013;40:159–68.
10. Li B, Walsh CT. Identification of the gene cluster for the dithiolopyrrolone antibiotic holomycin in *Streptomyces clavuligerus*. Proc Natl Acad Sci U S A. 2010;107:19731–5.
11. Udwary DW, Zeigler L, Asolkar RN, Singan V, Lapidus A, Fenical W, et al. Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. Proc Natl Acad Sci U S A. 2007;104:10376–81.
12. Yin H, Xiang S, Zheng J, Fan K, Yu T, Yang X, et al. Induction of holomycin production and complex metabolic changes by the *argR* mutation in *Streptomyces clavuligerus* NP1. Appl Environ Microbiol. 2012;78:3431–41.
13. Li B, Walsh CT. *Streptomyces clavuligerus* HlmI is an intramolecular disulfide-forming dithiol oxidase in holomycin biosynthesis. Biochemistry. 2011;50:4615–22.
14. Bhattacharya D, Nagpure A, Gupta RK. Bacterial chitinases: properties and potential. Crit Rev Biotechnol. 2007;27:21–8.
15. Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, et al. A new antibiotic kills pathogens without detectable resistance. Nature. 2015;517:455–9.
16. Ziemert N, Lechner A, Wietz M, Millán-Aguiñaga N, Chavarria KL, Jensen PR. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. Proc Natl Acad Sci U S A. 2014;111:E1130–9.
17. Zhao X-Q. Genome-based studies of marine microorganisms to maximize the diversity of natural products discovery for medical treatments. Evid Based Complement Alternat Med. 2011;2011:384572.
18. Wietz M, Duncan K, Patin NV, Jensen PR. Antagonistic interactions mediated by marine bacteria: the role of small molecules. J Chem Ecol. 2013;39:879–91.
19. Gram L, Melchiorsen J, Bruhn JB. Antibacterial activity of marine culturable bacteria collected from a global sampling of ocean surface waters and surface swabs of marine organisms. Mar Biotechnol (NY). 2010;12:439–51.
20. Fenical W. Chemical studies of marine bacteria: developing a new resource. Chem Rev. 1993;93:1673–83.
21. Lane AL, Moore BS. A sea of biosynthesis: marine natural products meet the molecular age. Nat Prod Rep. 2011;28:411–28.
22. Wietz M, Mansson M, Vynne NG, Gram L. Small molecule antibiotics from marine bacteria and strategies to prevent rediscovery of known compounds. In: Edited by Kim S. Marine microbiology : bioactive compounds and biotechnological applications. Wiley-VCH Verlag GmbH & Co. KGaA; 2013. p. 127-59.
23. Wilson MC, Mori T, Rückert C, Uria AR, Helf MJ, Takada K, et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature. 2014;506:58–62.
24. Beer LL, Moore BS. Biosynthetic convergence of salinosporamides A and B in the marine actinomycete *Salinispora tropica*. Org Lett. 2007;9:845–8.
25. Feling R, Buchanan G. Salinosporamide A: a highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus Salinospora. Angew Chemie. 2003;42:355–7.
26. Still PC, Johnson TA, Theodore CM, Loveridge ST, Crews P. Scrutinizing the scaffolds of marine biosynthetics from different source organisms: Gram-negative cultured bacterial products enter center stage. J Nat Prod. 2014;77:690–702.
27. Månsson M, Phipps RK, Gram L, Munro MHG, Larsen TO, Nielsen KF. Explorative Solid-Phase Extraction (E-SPE) for accelerated microbial natural product discovery, dereplication, and purification. J Nat Prod. 2010;73:1126–32.
28. Mansson M, Gram L, Larsen TO. Production of bioactive secondary metabolites by marine vibrionaceae. Mar Drugs. 2011;9:1440–68.

29. Wietz M, Mansson M, Gotfredsen CH, Larsen TO, Gram L. Antibacterial compounds from marine Vibrionaceae isolated on a global expedition. Mar Drugs. 2010;8:2946–60.
30. Weber T. *In silico* tools for the analysis of antibiotic biosynthetic pathways. Int J Med Microbiol. 2014;304:230–5.
31. Fedorova ND, Moktali V, Medema MH. Bioinformatics approaches and software for detection of secondary metabolic gene clusters. In: Keller NP, Turner G, editors. Fungal secondary metabolism: methods and protocols. 944th ed. Totowa, NJ: Humana Press; 2012. p. 23–45.
32. Seyedsayamdost MR. High-throughput platform for the discovery of elicitors of silent bacterial gene clusters. Proc Natl Acad Sci U S A. 2014;111:7266–71.
33. Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, et al. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol. 2009;140:13–7.
34. Rausch C, Weber T, Kohlbacher O, Wohlleben W, Huson DH. Specificity prediction of adenylation domains in Nonribosomal Peptide Synthetases (NRPS) using Transductive Support Vector Machines (TSVMs). Nucleic Acids Res. 2005;33:5799–808.
35. Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. NRPSpredictor2–a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. 2011;39(Web Server issue):W362–7.
36. Minowa Y, Araki M, Kanehisa M. Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. J Mol Biol. 2007;368:1500–17.
37. Van Heel AJ, de Jong A, Montalbán-López M, Kok J, Kuipers OP. BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Res. 2013;41(Web Server issue):W448–53.
38. Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One. 2012;7:e34064.
39. Li MHT, Ung PMU, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. BMC Bioinformatics. 2009;10:185.
40. Mansson M, Nielsen A, Kjærulff L, Gotfredsen CH, Wietz M, Ingmer H, et al. Inhibition of virulence gene expression in *Staphylococcus aureus* by novel depsipeptides from a marine *Photobacterium*. Mar Drugs. 2011;9:2537–52.
41. Nielsen A, Månsson M, Bojer MS, Gram L, Larsen TO, Novick RP, et al. Solonamide B inhibits quorum sensing and reduces *Staphylococcus aureus* mediated killing of human neutrophils. PLoS One. 2014;9:e84992.
42. Kjaerulff L, Nielsen A, Mansson M, Gram L, Larsen TO, Ingmer H, et al. Identification of four new *agr* quorum sensing-interfering cyclodepsipeptides from a marine *Photobacterium*. Mar Drugs. 2013;11:5051–62.
43. Nielsen A, Mansson M, Wietz M, Varming AN, Phipps RK, Larsen TO, et al. Nigribactin, a novel siderophore from *Vibrio nigripulchritudo*, modulates *Staphylococcus aureus* virulence gene expression. Mar Drugs. 2012;10:2584–95.
44. Vynne NG, Månsson M, Nielsen KF, Gram L. Bioactivity, chemical profiling, and 16S rRNA-based phylogeny of *Pseudoalteromonas* strains collected on a global research cruise. Mar Biotechnol (NY). 2011;13:1062–73.
45. Vynne NG, Mansson M, Gram L. Gene sequence based clustering assists in dereplication of *Pseudoalteromonas luteoviolacea* strains with identical inhibitory activity and antibiotic production. Mar Drugs. 2012;10:1729–40.
46. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39(Web Server issue):W339–46.
47. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, et al. antiSMASH 2.0–a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 2013;41 (Web Server issue):W204–12.
48. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the rapid annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014;42(Database issue):D206–14.
49. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75.
50. Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, et al. Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. Proc Natl Acad Sci U S A. 2001;98:12215–20.
51. Bentley SD, Chater KF, Cerdeño-Tárraga A-M, Challis GL, Thomson NR, James KD, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature. 2002;417:141–7.
52. Buchan A, González J, Moran M. Overview of the marine roseobacter lineage. Appl Environ. 2005;71:5665–77.
53. Wagner-Döbler I, Biebl H. Environmental biology of the marine Roseobacter lineage. Annu Rev Microbiol. 2006;60:255–80.
54. Porsby CH, Nielsen KF, Gram L. *Phaeobacter* and *Ruegeria* species of the Roseobacter clade colonize separate niches in a Danish Turbot (*Scophthalmus maximus*)-rearing farm and antagonize Vibrio anguillarum under different growth conditions. Appl Environ Microbiol. 2008;74:7356–64.
55. Seyedsayamdost MR, Carr G, Kolter R, Clardy J. Roseobacticides: small molecule modulators of an algal-bacterial symbiosis. J Am Chem Soc. 2011;133:18343–9.
56. Brinkhoff T, Bach G, Heidorn T, Liang L, Schlingloff A, Simon M. Antibiotic Production by a Roseobacter Clade-Affiliated Species from the German Wadden Sea and Its Antagonistic Effects on Indigenous Isolates. Appl Environ Microbiol. 2004;70:2560–5.
57. Geng H, Bruhn JB, Nielsen KF, Gram L, Belas R. Genetic dissection of tropodithietic acid biosynthesis by marine roseobacters. Appl Environ Microbiol. 2008;74:1535–45.
58. Wietz M, Månsson M, Gram L. Chitin stimulates production of the antibiotic andrimid in a *Vibrio coralliilyticus* strain. Environ Microbiol Rep. 2011;3:559–64.
59. Jin M, Fischbach MA, Clardy J. A biosynthetic gene cluster for the acetyl-CoA carboxylase inhibitor andrimid. J Am Chem Soc. 2006;128:10660–1.
60. Gauthier MJ. Morphological, physiological, and biochemical characteristics of some violet-pigmented bacteria isolated from seawater. Can J Microbiol. 1976;22:138–49.
61. Laatsch H, Pudleiner H. Marine bakterien, I: synthese von pentabrompseudilin, einem phenylpyrrol aus Alteromonas luteoviolaceus. Liebigs Ann Chem. 1989;9:863–81.
62. Moore BS. Biosynthesis of marine natural products: macroorganisms (Part B). Nat Prod Rep. 2006;23:615–29.
63. Conway KR, Boddy CN. ClusterMine360: a database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res. 2013;41(Database issue):D402–7.
64. Bruhn JB, Nielsen KF, Hjelm M, Hansen M, Bresciani J, Schulz S, et al. Ecology, inhibitory activity, and morphogenesis of a marine antagonistic bacterium belonging to the Roseobacter clade. Appl Environ Microbiol. 2005;71:7263–70.
65. Holden MTG, Chhabra SR, De Nys R, Stead P, Bainton NJ, Hill PJ, et al. Quorum-sensing cross talk: isolation and chemical characterization of cyclic dipeptides from *Pseudomonas aeruginosa* and other gram-negative bacteria. Mol Microbiol. 1999;33:1254–66.
66. Rasmussen BB, Nielsen KF, Machado H, Melchiorsen J, Gram L, Sonnenschein EC. Global and phylogenetic distribution of quorum sensing signals, acyl homoserine lactones, in the family of vibrionaceae. Mar Drugs. 2014;12:5527–46.
67. Vynne NG. Bioactivity and phylogeny of the marine bacterial genus *Pseudoalteromonas*. Kgs. Lyngby: Technical University of Denmark; 2011.
68. Sandy M, Butler A. Microbial iron acquisition: marine and terrestrial siderophores. Chem Rev. 2009;109:4580–95.
69. Schwyn B, Neilands JB. Universal chemical assay for the detection and determination of siderophores. Anal Biochem. 1987;160:47–56.
70. Hider RC, Kong X. Chemistry and biology of siderophores. Nat Prod Rep. 2010;27:637–57.
71. Rudolph G, Hennecke H, Fischer H-M. Beyond the Fur paradigm: iron-controlled gene expression in rhizobia. FEMS Microbiol Rev. 2006;30:631–48.
72. Williams KP, Gillespie JJ, Sobral BWS, Nordberg EK, Snyder EE, Shallom JM, et al. Phylogeny of gammaproteobacteria. J Bacteriol. 2010;192:2305–14.
73. Sawabe T, Ogura Y, Matsumura Y, Feng G, Amin AR, Mino S, et al. Updating the *Vibrio* clades defined by multilocus sequence phylogeny: proposal of eight new clades, and the description of *Vibrio tritonius* sp. nov. Front Microbiol. 2013;4(December):414.
74. Pascual J, Macián MC, Arahal DR, Garay E, Pujalte MJ. Multilocus sequence analysis of the central clade of the genus Vibrio by using the *16S rRNA*, *recA*, *pyrH*, *rpoD*, *gyrB*, *rctB* and *toxR* genes. Int J Syst Evol Microbiol. 2010;60(Pt 1):154–65.

75. Thompson FL, Gevers D, Thompson CC, Dawyndt P, Naser S, Hoste B, et al. Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis. Appl Environ Microbiol. 2005;71:5107–15.
76. Meibom KL, Li XB, Nielsen AT, Wu C-Y, Roseman S, Schoolnik GK. The *Vibrio cholerae* chitin utilization program. Proc Natl Acad Sci U S A. 2004;101:2524–9.
77. Li X, Roseman S. The chitinolytic cascade in Vibrios is regulated by chitin oligosaccharides and a two-component chitin catabolic sensor/kinase. Proc Natl Acad Sci U S A. 2004;101:627–31.
78. Frederiksen RF, Paspaliari DK, Larsen T, Storgaard BG, Larsen MH, Ingmer H, et al. Bacterial chitinases and chitin-binding proteins as virulence factors. Microbiology. 2013;159(Pt 5):833–47.
79. Souza CP, Almeida BC, Colwell RR, Rivera ING. The importance of chitin in the marine environment. Mar Biotechnol (NY). 2011;13:823–30.
80. Stauder M, Huq A, Pezzati E, Grim CJ, Ramoino P, Pane L, et al. Role of GbpA protein, an important virulence-related colonization factor, for *Vibrio cholerae*'s survival in the aquatic environment. Environ Microbiol Rep. 2012;4:439–45.
81. Sambrook J, Russel DW. Molecular cloning: a laboratory manual. Cold Spring Harbour, New York: Cold Spring Harbour Laboratory Press; 2001.
82. Skov MN, Pedersen K, Larsen JL. Comparison of pulsed-field gel electrophoresis, ribotyping, and plasmid profiling for typing of *Vibrio anguillarum* serovar O1. Appl Environ Microbiol. 1995;61:1540–5.
83. Novick R, Morse S. In vivo transmission of drug resistance factors between strains of Staphylococcus aureus. J Exp Med. 1967;125(1):45–59.
84. Cha C, Gao P, Chen YC, Shaw PD, Farrand SK. Production of acyl-homoserine lactone quorum-sensing signals by gram-negative plant-associated bacteria. Mol Plant Microbe Interact. 1998;11:1119–29.
85. Mcclean KH, Winson MK, Fish L, Taylor A, Chhabra SR, Camara M, et al. Quorum sensing and *Chromobacterium violaceum*: exploitation of violacein production and inhibition for the detection of N -acylhomoserine lactones. 1997. p. 3703–11.
86. Ravn L, Christensen AB, Molin S, Givskov M, Gram L. Methods for detecting acylated homoserine lactones produced by Gram-negative bacteria and their application in studies of AHL-production kinetics. J Microbiol Methods. 2001;44:239–51.
87. Gram L, Grossart H. Possible quorum sensing in marine snow bacteria: production of acylated homoserine lactones by Roseobacter strains isolated from marine snow. Appl. 2002;68:4111–6.

Paper 2

The *fur* gene as a new phylogenetic marker for *Vibrionaceae* species identification

AEM
Journals.ASM.org

The *fur* Gene as a New Phylogenetic Marker for *Vibrionaceae* Species Identification

Henrique Machado,[a,b] Lone Gram[a]

Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark[a]; Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Hørsholm, Denmark[b]

Microbial taxonomy is essential in all areas of microbial science. The 16S rRNA gene sequence is one of the main phylogenetic species markers; however, it does not provide discrimination in the family *Vibrionaceae*, where other molecular techniques allow better interspecies resolution. Although multilocus sequence analysis (MLSA) has been used successfully in the identification of *Vibrio* species, the technique has several limitations. They include the fact that several locus amplifications and sequencing have to be performed, which still sometimes lead to doubtful identifications. Using an *in silico* approach based on genomes from 103 *Vibrionaceae* strains, we demonstrate here the high resolution of the *fur* gene in the identification of *Vibrionaceae* species and its usefulness as a phylogenetic marker. The *fur* gene showed within-species similarity higher than 95%, and the relationships inferred from its use were in agreement with those observed for 16S rRNA analysis and MLSA. Furthermore, we developed a *fur* PCR sequencing-based method that allowed identification of *Vibrio* species. The discovery of the phylogenetic power of the *fur* gene and the development of a PCR method that can be used in amplification and sequencing of the gene are of general interest whether for use alone or together with the previously suggested loci in an MLSA.

In microbial science, rapid identification of isolates to the genus or species level is essential in many areas, for instance, when diagnosing an infection or unraveling the microbial diversity in different niches. Therefore, identification and classification must be reliable, reproducible, and informative and at the same time fast and user friendly. Hence, tools that allow identification should be easy and affordable for the average user (1).

Identification of microbial species was originally based on phenotypic assays, due to both the influence of other biology disciplines, such as botany, and the technological limitations at the time (2). While phenotypic identification has been an important tool, the approach was of more limited use for several microbial groups, including the family *Vibrionaceae*, where some species had indistinguishable phenotypes and other species could have divergent phenotypes among strains of the same species (1). More recently, genetically based methods, in particular, have been developed for species identification and phylogeny, with the "molecular clock" approach introduced by Carl Woese proving a strong tool (3). The most common genetic marker has been the 16S rRNA gene, but the sequences of a range of housekeeping genes are also being used. In the future, genome sequences (gene sequences or single-nucleotide polymorphisms [SNP]) will be used in phylogeny and identification.

The family *Vibrionaceae* comprises 159 species in 6 genera, of which the genus *Vibrio* is the largest and most extensively characterized (4, 5). For many genera, the common approach of using 16S rRNA gene similarity as the main phylogenetic species marker has failed, due to its low interspecies resolution (6).

Due to the limitations of the 16S rRNA gene phylogeny in *Vibrio* classification and to the development of molecular techniques, such as fluorescent amplified fragment length polymorphism (FALP) and multilocus sequence analysis (MLSA), the classification of *Vibrionaceae* soon evolved from using a single gene to using several gene sequences for identification and phylogeny. In particular, the introduction of an MLSA scheme using nine gene sequences (*ftsZ*, *gapA*, *gyrB*, *mreB*, *pyrH*, *recA*, *rpoA*, *topA*, and the 16S rRNA gene) has provided much higher resolution for *Vibrionaceae* identification and taxonomy, and this has allowed the identification of new *Vibrio* species previously misidentified (6–11).

While the use of MLSA as described above has improved the resolution of *Vibrionaceae* phylogeny, it has been hampered by requiring several gene sequences. With the explosion of genome sequencing, this will become easier; however, for a great number of laboratories, the use of a single gene sequence for identification and phylogeny will be preferable for years to come. The use of the *fur* gene as a phylogenetic marker in bacteria has been suggested in several studies (12, 13). Also, one study (14) suggested the use of the *fur* gene as a discriminative phylogenetic marker between the species *Alliivibrio salmonicida* and *Alliivibrio logei* (previously *Vibrio salmonicida* and *Vibrio logei*, respectively); however, to our knowledge, no further work has addressed this possible marker. The *fur* gene encodes a ferric uptake regulator (Fur), which in most bacterial species is the major system for maintenance of iron homeostasis. Fur senses excess intracellular Fe^{2+} and binds to the promoter regions of the genes involved in iron acquisition,

Received 7 January 2015 **Accepted** 2 February 2015

Accepted manuscript posted online 6 February 2015

Citation Machado H, Gram L. 2015. The *fur* gene as a new phylogenetic marker for *Vibrionaceae* species identification. Appl Environ Microbiol 81:2745–2752. doi:10.1128/AEM.00058-15.

Editor: C. R. Lovell

Address correspondence to Henrique Machado, henma@biosustain.dtu.dk.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.00058-15.

doi:10.1128/AEM.00058-15

TABLE 1 *Vibrionaceae* strains used in this study

Species	Strain	WGS/*fur* GenBank accession no.
Strains used in the *in silico* study[a]		
Aliivibrio fischeri	ZF-211	AJYI01
Aliivibrio logei	5S-186	AJYJ01
Aliivibrio logei	ATCC 35077	ASAH01
Enterovibrio calviensis	DSM 14347^T	JHZA01
Enterovibrio norvegicus	FF-33	AJYD01
Grimontia hollisae	CIP 101886^T	ADAQ01
Grimontia sp.	AK16	ANFM02
Photobacterium aphoticum	JCM 19237	BBMN01
Photobacterium angustum	S14	AAOJ01
Photobacterium damselae subsp. *damselae*	CIP 102761^T	ADBS01
Photobacterium damselae subsp. *piscicida*	DI21	AKYG01
Photobacterium halotolerans	S2753	JMIB01
Photobacterium halotolerans	DSM 18316^T	AULG01
Photobacterium leiognathi	lrivu.4.1	BANQ01
Photobacterium leiognathi subsp. *mandapamensis*	svers.1.1.	BACE01
Photobacterium phosphoreum	ANT220	CCAR01
Photobacterium profundum	3TCK	AAPH01
Photobacterium sp.	SKA34	AAOU01
Photobacterium sp.	AK15	AMZO01
Salinivibrio costicola subsp. *costicola*	ATCC 33508^T = LMG 11651^T	ASAI01
Shewanella xiamenensis	BC01	JAEC01
Vibrio albensis = *V. cholerae*	VL426	ACHV01
Vibrio anguillarum	96F	AEZA01
Vibrio anguillarum	RV22	AEZB01
Vibrio azureus	NBRC 104587^T	BATL01
Vibrio brasiliensis	LMG 20546^T	AEVS01
Vibrio breoganii	ZF-55	AJYL01
Vibrio campbellii	CAIM 519^T = NBRC 15631^T	AMDG01
Vibrio caribbeanicus	ATCC BAA-2122^T	AEIU01
Vibrio cholerae	MO10	AAKF03
Vibrio cholerae	CP110	AMWF01
Vibrio cholerae	TM 11079-80	ACHW01
Vibrio coralliilyticus	P1	AEQS01
Vibrio coralliilyticus	ATCC BAA-450^T	ACZN01
Vibrio coralliilyticus	OCN008	AVOO01
Vibrio cyclitrophicus	ZF14	AIDH01
Vibrio cyclitrophicus	FF75	ATLT01
Vibrio diazotrophicus	NBRC 103148^T	BBJY01
Vibrio ezurae	NBRC 102218^T	BATM01
Vibrio fortis	Dailan14	JFFR01
Vibrio fluvialis	PG41	ASXS01
Vibrio fluvialis	I21563	ASXT01
Vibrio halioticoli	NBRC 102217^T	BAUJ01
Vibrio harveyi	CAIM 1792	AHHQ01
Vibrio harveyi	NBRC 15634 = ATCC 14126^T	BAOD01
Vibrio harveyi	ZJ0603	AKIH01
Vibrio harveyi	AOD131	AOMR01
Vibrio harveyi	E385	AYKI01
Vibrio harveyi	VHJR4	CAUN01
Vibrio harveyi	VHJR7	CAUO01
Vibrio ichthyoenteri	ATCC 700023^T	AFWF01
Vibrio jasicida	090810c	BAOC01
Vibrio kanaloae	5S-149	AJYX01
Vibrio litoralis	DSM 17657^T	AUFZ01

TABLE 1 (Continued)

Species	Strain	WGS/*fur* GenBank accession no.
Vibrio maritimus	JCM 19240	BBMT01
Vibrio maritimus	JCM 19235	BBMR01
Vibrio metschnikovii	CIP 69.14^T	ACZO01
Vibrio mimicus	VM223	ADAJ01
Vibrio mimicus	MB451	ADAF01
Vibrio mimicus	CAIM 602^T	AOMO01
Vibrio mimicus	SX-4	ADOO01
Vibrio natriegens	NBRC 15636^T = ATCC 14048^T = DSM 759^T	ATFJ01
Vibrio nigripulchritudo	ATCC 27043^T	AFWJ01
Vibrio nigripulchritudo	FTn2	CANW01
Vibrio nigripulchritudo	SOn1	CAOF01
Vibrio ordalii	FS-238	AJYS01
Vibrio ordalii	12B09	AJYV01
Vibrio ordalii	FF-93	AJYT01
Vibrio ordalii	FS-144	AJYU01
Vibrio orientalis	CIP 102891^T = ATCC 33934^T	ACZV01
Vibrio owensii	ATCC 25919	BANZ01
Vibrio owensii	CAIM 1854^T	BAOH01
Vibrio pacinii	DSM 19139^T	JONH01
Vibrio parahaemolyticus	NIHCB0603	AVOM01
Vibrio parahaemolyticus	IDH02189	JAHD01
Vibrio proteolyticus	NBRC 13287^T	BATJ01
Vibrio rhizosphaerae	DSM 18581^T	JONG01
Vibrio rotiferianus	DAT722	AFAJ01
Vibrio rumoiensis	1S-45	AJYK01
Vibrio sagamiensis	NBRC 104589^T	BAOJ01
Vibrio scophthalmi	LMG 19158^T	AFWE01
Vibrio shilonii = *V. mediterranei*	AK1^T	ABCH01
Vibrio sinaloensis	DSM 21326	AEVT01
Vibrio sp.	PPCK-2014	JJMN01
Vibrio splendidus	ZS-139	AJZE01
Vibrio splendidus	FF-6	AJZI01
Vibrio splendidus	ATCC 33789	AFWG01
Vibrio splendidus	12B01	AAMR01
Vibrio splendidus	12E03	AJZD01
Vibrio splendidus	ZF-90	AJZF01
Vibrio splendidus	5S-101	AJZG01
Vibrio splendidus	FF-500	AJZH01
Vibrio splendidus	1F-157	AJZJ01
Vibrio splendidus	1S-124	AJZL01
Vibrio tasmaniensis	1F-187	AJZM01
Vibrio tasmaniensis	1F-155	AJZN01
Vibrio tasmaniensis	5F-79	AJZP01
Vibrio tasmaniensis	ZS-17	AJZQ01
Vibrio tubiashii	ATCC 19109^T	AFWI01
Vibrio tubiashii	NCIMB 1337 = ATCC 19106	AHHF01
Vibrio variabilis	JCM 19239	BBMS01
Vibrio vulnificus	BAA87	JDSE01
Vibrio vulnificus	NBRC 15645^T = ATCC 27562^T	AMQV01
Culture collection strains used in PCR sequencing method development		
Aliivibrio fischeri	DSM 2168	KP721366

(Continued on following page)

TABLE 1 (Continued)

Species	Strain	WGS/*fur* GenBank accession no.
Enterovibrio calviensis	DSM 14347^T	KP721381
Grimontia hollisae	DSM 15132^T	KP721382
Photobacterium damselae	ATCC 33539^T	KP721367
Photobacterium halotolerans	LMG 22194^T	KP721368
Photobacterium angustum	S14	KP721369
Photobacterium rosenbergii	LMG 22223^T	KP721370
Vibrio coralliilyticus	ATCC BAA-450^T = DSM 19607^T	KP721371
Vibrio fluvialis	NCTC 11327^T	KP721372
Vibrio harveyi	DSM 19623^T = ATCC 14126^T	KP721373
Vibrio nigripulchritudo	ATCC 27043^T	KP721374
Vibrio owensii	DY05^T	KP721375
Vibrio parahaemolyticus	ATCC 17802^T	KP721376
Vibrio splendidus	ATCC 33125^T	KP721377
Vibrio vulnificus	ATCC 27562^T = DSM 10143^T	KP721378
Vibrio anguillarum	DSM 21597^T	KP721379
Vibrio ponticus	DSM 16217^T	KP721383
Vibrio chagasii	DSM 17138^T	KP721384
Vibrio brasiliensis	DSM 17184^T	KP721385
Vibrio porteresiae	DSM 19223^T	KP721386
Vibrio pectenicida	DSM 19585^T	KP721387
Salinivibrio costicola subsp. *costicola*	DSM 11403^T	KP721380
Isolates used in the testing of the PCR sequencing method		
Photobacterium halotolerans	S2753	KP721398
Vibrio anguillarum	775	KP721388
Vibrio coralliilyticus	S2043	KP721394
Vibrio coralliilyticus	S2052	KP721395
Vibrio nigripulchritudo	S2604	KP721397
Vibrio neptunius	S2394	KP721396
Vibrio parahaemolyticus	V2	KP721401
Vibrio sp.	S188	KP721389
Vibrio sp.	S203	KP721390
Vibrio sp.	S344	KP721391
Vibrio sp.	S787	KP721392
Vibrio sp.	S1110	KP721393
Vibrio sp.	S2757	KP721399
Vibrio sp.	S4497	KP721400
Vibrio sp.	VibAn	KP721402

[a] Genomes from GenBank.

thereby blocking their transcription. In contrast, when iron availability is limited, derepression of the corresponding genes occurs. In addition to iron transport, Fur controls a range of other processes, such as redox stress resistance, energy metabolism, flagellar chemotaxis, and metabolic pathways (15–17).

The purpose of the present study was to determine if the *fur* gene could be used as a new phylogenetic marker in the identification of *Vibrionaceae* species. The availability of several fully genome-sequenced strains allowed us to address this in an *in silico* analysis. To facilitate broader use of *fur* as a phylogenetic marker, despite highly variable regions outside the *fur* gene, we developed a PCR sequencing-based method for the analysis of the *fur* gene in *Vibrionaceae* species.

MATERIALS AND METHODS

***In silico* analysis of *fur* sequences.** The whole-genome sequences (WGS) from 104 strains were used in this study: 83 *Vibrio* strains representing 44 species, 3 *Aliivibrio* strains representing 2 species, 12 *Photobacterium* strains representing 7 species, 2 *Enterovibrio* strains representing 2 species, 2 *Grimontia* strains representing 1 species, 1 *Salinivibrio costicola* strain, and 1 sequence from *Shewanella xiamenensis* BC01 as an unrelated Gram-negative bacillus (Table 1). The genome sequences were used in the *in silico* analysis of the *fur* gene. The genomes were all analyzed using CLC Main Workbench version 7 (CLC, Aarhus, Denmark). For the genomes annotated by NCBI, an annotation-based search was performed for the *fur* genes. Those not annotated were submitted to a BLAST search against the annotated *fur* genes and manually curated if necessary.

Phylogenetic-data analysis. The *fur* sequences isolated *in silico* or PCR amplified and sequenced in this study were aligned using the alignment tools in CLC Main Workbench version 7 (CLC, Aarhus, Denmark). The Gap cost settings were as follows: gap open cost of 10 points and gap extension cost of 1 point, and end gaps were treated like any other gap. The alignments obtained were used to perform a pairwise comparison of the number of differences and the percent identity using CLC Main Workbench version 7 (CLC, Aarhus, Denmark). Furthermore, maximum-likelihood phylogeny trees were also generated using the CLC Main Workbench version 7 (CLC, Aarhus, Denmark) tools. Neighbor-joining was the tree construction method used, with the Jukes-Cantor nucleotide distance measure. The design of the trees was finalized using MEGA 6 (18).

Bacterial strains and genomic-DNA extraction. The bacterial strains used for development of the PCR method (Table 1) were grown in Marine Broth (Difco; catalog no. 279110) overnight at 25°C, and genomic DNA was isolated using the NucleoSpin Tissue kit (Macherey-Nagel, Düren, Germany). The quality of the genomic DNA was checked by 1% agarose gel electrophoresis and quantified by absorbance using DeNovix DS-11.

Primer design. The degenerate primers fur_AP _fw (5′-CCWCCATAYTGDGWMCGRTTNGCATTCCWCCATAYTGDGWMCGRTTNGCATT-3′) and fur_AP_rv (5′-ACWGTHGGYYTWCGTGATACWTGGG-3′) were designed according to the alignments done using the *fur* regions of several *Vibrio*, *Aliivibrio*, and *Photobacterium* strains. Also, an extra set of primers were designed for amplification in other *Vibrio* species where the AP primers did not work: fur_V_fw (5′-TAACCYYTTGAASTTGAASTTCG-3′), fur_TS_rv (5′-CGWAYDGGHTAYTTCTGTGYDGAT-3′), and fur_OM_rv (5′-GTGGCRGATAAYGTKMGHAAAGG-3′). These primers were then used to amplify the whole *fur* gene. Due to the different sizes of the fragments, internal primers were also designed to confirm the presence of the *fur* gene in the amplified fragment: fur_Sp_internal_fw (5′-CACCAYTTYGAAGGCGGYAAGTC-3′) and fur_Sp_internal_rv (5′-ATYTCTTTYTGKCGYTCTTCRAT-3′).

PCR amplification and sequencing. Amplification reaction mixtures contained 1× *Pfu*X7 buffer [20 mM Tris-HCl, pH 8.8, 10 mM KCl, 6 mM $(NH_4)_2SO_4$, 2 mM $MgSO_4$, 0.1 mg/ml bovine serum albumin (BSA), 0.1% Triton X-100], a 200 μM deoxynucleoside triphosphate (dNTP) mixture, 0.4 μM each primer, 1 μl of a 10-fold dilution of genomic DNA as the template, and 1 μl of *Pfu*X7 polymerase (19) in a final reaction volume of 25 μl. The PCR amplification was carried out in a thermal cycler (Veriti 96-well thermal cycler; Applied Biosystems) as follows: a 2-min initial denaturation step at 98°C, followed by 30 cycles of 98°C for 20 s, 52°C for 20 s, and 72°C for 25 s, with a final extension step of 2 min at 72°C. The amplified products were visualized by agarose gel electrophoresis (1 or 1.5%) and ethidium bromide staining. The PCR products were enzymatically purified by treatment with exonuclease I (ExoI) (Thermo Scientific) and FastAP thermosensitive alkaline phosphatase (Thermo Scientific) before sequencing at GATC Biotech (Cologne, Germany) or Macrogen (Amsterdam, The Netherlands). The sequences were analyzed using CLC Main Workbench version 7 (CLC, Aarhus, Denmark).

Nucleotide sequence accession numbers. The sequence data generated in this study were deposited in GenBank under the accession numbers provided in Table 1.

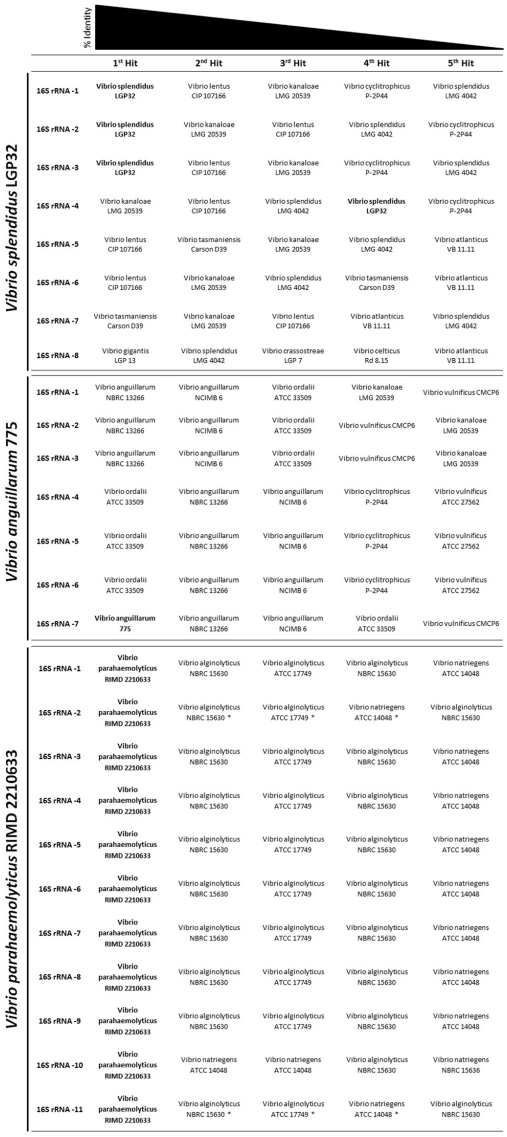


FIG 1 BLAST analyses and closest hits to the different 16S rRNA alleles in three different *Vibrionaceae* strains. Hits from the same strain are in boldface.

RESULTS AND DISCUSSION

16S rRNA gene limitations. 16S rRNA gene sequences have long been used for distinguishing and classifying new strains at the genus and species levels. This approach has been very successful in several bacterial groups, but not in the *Vibrionaceae*, due to low

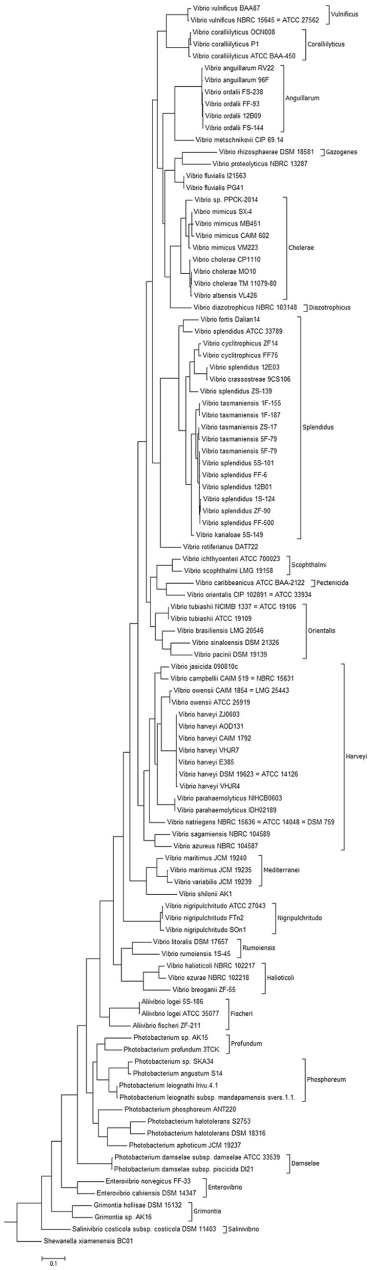


FIG 2 Phylogenetic tree of 103 *Vibrionaceae* strains. The tree is based on *in silico* analyses using the complete *fur* gene sequences and was constructed by the neighbor-joining method. *S. xiamenensis* BC01 was used as the outlier.

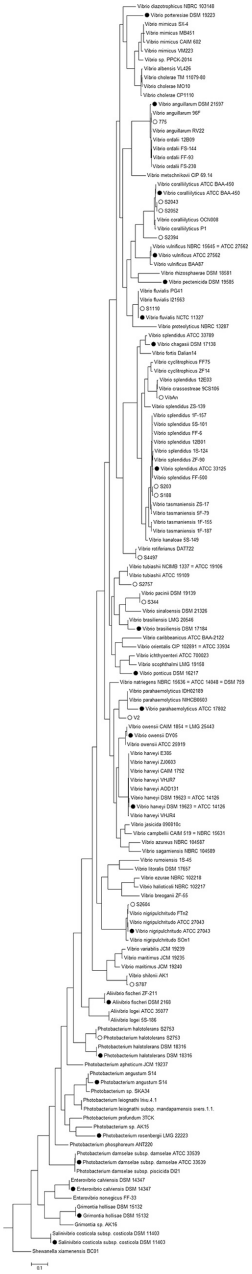


FIG 3 Phylogenetic tree of 140 *Vibrionaceae* isolates (including strains shown in Fig. 1). The tree is based on the PCR-amplified and sequenced complete *fur* gene sequences of representatives of each species and the *fur* sequences of the type strains and of the isolates used in the development and testing of the PCR sequencing-based method and was constructed using the neighbor-joining method. The circles indicate *fur* genes sequenced in the development and testing of the PCR sequencing-based method: the solid circles are the type strains used in the development of the method, and the open circles are the environmental strains used to test the method. *S. xiamenensis* BC01 was used as the outlier.

interspecies resolution achieved using the gene (6). This is in part caused by the many alleles of the 16S rRNA gene that, when cloned and sequenced individually, can identify a strain as belonging to several different species (20). To confirm this observation, we used three closed genomes as examples and used the different 16S rRNA alleles in a BLAST search in order to identify the species (Fig. 1). Not only did the number of alleles seem to be variable among vibrios, but the identification drawn from each allele pointed to a different species, and thus, the identifications made using this approach in vibrios are very questionable. Furthermore, when using next-generation sequencing techniques, such as Illumina, the length of the reads obtained does not allow differentiation of the different alleles, resulting in genome assemblies with only one 16S rRNA allele.

***In silico* analysis of *fur* sequences.** For all 103 strains analyzed, only one copy of the *fur* gene was identified in the whole genome. The *fur* sequences varied in size between 441 and 456 bp, with the exception of *Shewanella xiamenensis* BC01, where the *fur* gene was only 432 bp. The *fur* genes with 441 bp were from *Vibrio halioticoli* and *Vibrio ezurae*, and only *Vibrio nigripulchritudo* SOn1 had a 456-bp *fur* sequence, making the variance of the gene size mostly between 444 and 453 bp (see Fig. S1 in the supplemental material).

The maximum-likelihood phylogeny tree constructed with the alignment of the *fur* genes (Fig. 2) confirmed the clustering of the different species in a manner similar to what has been observed for both 16S rRNA gene- and MLSA-based trees (4, 6, 8, 21–24). The clades recently emended by Sawabe et al. (4) could also be identified (Fig. 2), although a few differences arose. Within the *Vibrio* species, the major differences observed were the clustering of *Vibrio rotiferianus* DAT722 within the *Splendidus* clade rather than within the *Harveyi* clade, as previously described (4). Other species, such as *Vibrio metschnikovii* and *Vibrio fluvialis*, did not cluster in the *Cholerae* clade but were clearly closely related. A similar observation was made for the species *Vibrio orientalis* and *Photobacterium phosphoreum*.

At the species level, some of the strains previously identified as *Vibrio splendidus*, namely, *V. splendidus* 12E03, *V. splendidus* ZS-139, and *V. splendidus* ATCC 33789, clustered separately from the other *V. splendidus* strains. In fact, *V. splendidus* 12E03, *V. splendidus* ZS-139, and *V. splendidus* ATCC 33789 clustered in the *Vibrio cyclitrophicus* and the *Vibrio fortis* branches of the phylogenetic tree (Fig. 2). This could be explained by the previously demonstrated genetic diversity and polyphyletic nature of *V. splendidus* (7, 8, 25), or it could be that these strains were misidentified, which is a recurrent problem within the genus *Vibrio* (22, 26, 27). The second hypothesis seems to be consistent with the genomic data index (ANI) at EzGenome (http://www.ezbiocloud.net/ezgenome/hierarchy?n=Vibrionales&d=2#), where ANI-based trees locate these strains closer to other named *Vibrio* species.

A comparison of the 103 sequences revealed a percent similarity in the same species of more than 97%, with the exception of two species pairs, *V. splendidus* and *Vibrio tasmaniensis*, as well as

FIG 4 Analysis of the variability downstream of the *fur* gene by alignment of the *fur* regions from six whole-genome-sequenced *Vibrio* species.

Vibrio anguillarum and *Vibrio ordalii*, where the above-mentioned species threshold did not apply (see Fig. S2 in the supplemental material). This is similar to what has been shown by others using the genes *rpoA*, *recA*, and *pyrH*, where 98%, 94%, and 94% similarities within the same species were reported (8). Among these genes, *pyrH* has been described as the most discriminatory (28); however, this study was performed in a small selection of strains, focusing mostly on the *Harveyi*, *Campbellii*, and *Rotiferianus* groups, making it difficult to do meaningful comparison between the discriminatory powers of this gene and the *fur* gene described here. Another gene reported as a possible good phylogenetic marker in the family *Vibrionaceae* is *atpA* (22). Compared with this gene, *fur* presents itself as a better phylogeny discriminator. The *atpA* gene phylogeny showed very high homology between species within the main identified groups, which could vary between 94 and 99%, while in the *fur* gene phylogeny, only the two previously described pairs (*V. splendidus*-*V. tasmaniensis* and *V. anguillarum*-*V. ordalii*) show that limitation.

Development of a PCR sequencing-based method. The above-described *in silico* analysis confirmed the value of the *fur* sequence as a phylogenetic marker in the family *Vibrionaceae*, allowing the distinction of most of the strains at species level. Therefore, a PCR-based method that could be widely used by researchers working with *Vibrionaceae* genera would be extremely valuable as a simpler tool in the classification of *Vibrio* isolates or as an extra differentiation marker added to an MLSA.

To develop and validate the PCR method and the designed primers, 22 type strains obtained from several culture collections (Fig. 3) were used in the establishment of a general protocol for the amplification of *fur*. They included 1 *Aliivibrio* type strain, 4 *Photobacterium* type strains, 1 *Enterovibrio* type strain, 1 *Salinivibrio* type strain, 1 *Grimontia* type strain, and 14 *Vibrio* type strains. The degenerate fur_AP primers (see Materials and Methods) amplified the *fur* gene in most *Vibrio* species and in all the *Photobacterium*, *Aliivibrio*, *Enterovibrio*, *Grimontia*, and *Salinivibrio* species tested. The main challenge was the design of a reverse primer suitable for all the *Vibrio* strains, since the variation in sequence downstream of the *fur* gene is significant (Fig. 4). Therefore, we designed an extra set of primers, including one forward and two distinct reverse primers, allowing us to amplify the *fur* genes from the *Vibrio* strains on which the above-mentioned set of primers (AP) did not work. The primers developed here did allow amplification of all the tested strains of *Vibrio* species and the other genera of the family *Vibrionaceae*. The differences between the flanking regions of the *fur* gene have been described previously (13), and the conserved genetic organization of the region upstream from the *fur* gene was not reflected in the downstream region, where there is higher variability between species (13). We also noted in our amplification process that there was a species-dependent fragment size of the amplicon. The relationship between the fragment size amplified and the species needs further investigation, although it could possibly expedite the attribution of a provisional clade or even species at an earlier stage in the classification process.

The limited number of genomes available from *Grimontia*, *Salinivibrio*, *Enterovibrio*, *Photobacterium*, and *Aliivibrio* species hampers the design of more universal primers, and once more WGS from strains belonging to these genera are available, a more in-depth analysis of the *fur* gene flanking regions might provide enough information for the design of primers suitable for all the strains belonging to these species. The so-called "primer problems" have been previously reported in several MLSA analyses, both for strains from the genus *Vibrio* and for strains of the genus

Photobacterium (4, 6). This likely reflects the large genomic variability between members of the family *Vibrionaceae*.

After amplification and sequencing of the amplicons, the sequences were used in two approaches: (i) the *fur* sequences were extracted from the consensus sequence and subjected to BLAST searches against the NCBI database or (ii) the whole consensus sequence was used in the BLAST analysis. Both approaches showed a high level of identification, since the first BLAST hits in both cases corresponded to strains of the same species as the tested strains. Nevertheless, this approach using the NCBI database directly presented some limitations. Because few *fur* sequences are available for the *Vibrio* species, the hits that were obtained corresponded to whole-genome-sequenced strains, limiting the analysis to such strains. In fact, the results were more expressive and clearer when the extracted *fur* sequences were subjected to BLAST searches against a database created with the sequences used in the *in silico* analysis. This shows the importance of creating a database with more *fur* sequences in order to obtain more accurate identifications and to bring this method to its full potential.

Identification of *Vibrio* strains using the *fur* gene sequence. To test the method developed, a collection of previously isolated *Vibrio* species was used. The strains have been categorized in our laboratory by conventional 16S rRNA gene sequence analyses (29), and some also by sequencing of *recA*, *toxR*, and *rpoA* genes (30). PCR, as has been described, worked well for all the tested strains (Fig. 4), so the amplicons were purified and sequenced. The *fur* genes were identified in all the sequenced amplicons.

The isolated *fur* sequences were incorporated in the phylogenetic tree, giving an overview of the distribution of the isolates between the different species and clades (Fig. 3). Of the tested isolates, strains S2757 and S2394 could not be identified with high certainty, although it is obvious that strain S2757 clustered closely with *Vibrio tubiashii* (Fig. 3). This inconclusive identification could indicate that strain S2757 belongs to a new species; further attempts using MLSA were also inconclusive (data not shown). With respect to strain S2394, the fact that no WGS or *fur* sequences from *Vibrio neptunius* are available led to a close association of the strain with *Vibrio coralliilyticus*, which makes sense, given its close phylogenetic proximity (4) (Fig. 3).

It is evident that whole-genome sequencing and bioinformatics will drive identification and taxonomy in the future. Nevertheless, it may be a while until the average laboratory can afford the whole-genome sequencing of their strains and can master the bioinformatics needed to correctly identify them. Therefore, we believe that the discovery of the phylogenetic power of the *fur* gene and the development of a PCR method that can be used in amplification and sequencing of the gene is of general interest, whether for use alone or together with the previously suggested loci in an MLSA.

ACKNOWLEDGMENTS

H.M. was supported by a Ph.D. grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7-People-2012-ITN, under grant agreement no. 317058, "BACTORY."

We thank Paul D'Alvise for helpful discussions and for providing genome sequences and isolates of newly isolated *Vibrio* species.

REFERENCES

1. **Amaral GRS, Dias GM, Wellington-Oguri M, Chimetto L, Campeão ME, Thompson FL, Thompson CC.** 2014. Genotype to phenotype: identification of diagnostic vibrio phenotypes using whole genome sequences. Int J Syst Evol Microbiol **64:**357–365. http://dx.doi.org/10.1099/ijs.0.057927-0.
2. **Drews G.** 2000. The roots of microbiology and the influence of Ferdinand Cohn on microbiology of the 19th century. FEMS Microbiol Rev **24:**225–249. http://dx.doi.org/10.1111/j.1574-6976.2000.tb00540.x.
3. **Wheelis M, Kandler O, Woese C.** 1992. On the nature of global classification. Proc Natl Acad Sci U S A **89:**2930–2934. http://dx.doi.org/10.1073/pnas.89.7.2930.
4. **Sawabe T, Ogura Y, Matsumura Y, Feng G, Amin AR, Mino S, Nakagawa S, Sawabe T, Kumar R, Fukui Y, Satomi M, Matsushima R, Thompson FL, Gomez-Gil B, Christen R, Maruyama F, Kurokawa K, Hayashi T.** 2013. Updating the *Vibrio* clades defined by multilocus sequence phylogeny: proposal of eight new clades, and the description of Vibrio tritonius sp. nov. Front Microbiol **4:**414. http://dx.doi.org/10.3389/fmicb.2013.00414.
5. **Nishiguchi MK.** 2003. Evolution of symbiosis in the Vibrionaceae: a combined approach using molecules and physiology. Int J Syst Evol Microbiol **53:**2019–2026. http://dx.doi.org/10.1099/ijs.0.02792-0.
6. **Sawabe T, Kita-Tsukamoto K, Thompson FL.** 2007. Inferring the evolutionary history of vibrios by means of multilocus sequence analysis. J Bacteriol **189:**7932–7936. http://dx.doi.org/10.1128/JB.00693-07.
7. **Thompson F, Hoste B.** 2001. Genomic diversity amongst *Vibrio* isolates from different sources determined by fluorescent amplified fragment length polymorphism. Syst Appl Microbiol **24:**520–538. http://dx.doi.org/10.1078/0723-2020-00067.
8. **Thompson FL, Gevers D, Thompson CC, Dawyndt P, Naser S, Hoste B, Munn CB, Swings J.** 2005. Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis. Appl Environ Microbiol **71:**5107–5115. http://dx.doi.org/10.1128/AEM.71.9.5107-5115.2005.
9. **Le Roux F, Gay M, Lambert C, Waechter M, Poubalanne S, Chollet B, Nicolas J, Berthe F.** 2002. Comparative analysis of *Vibrio splendidus*-related strains isolated during *Crassostrea gigas* mortality events. Aquat Living Resour **15:**251–258. http://dx.doi.org/10.1016/S0990-7440(02)01176-2.
10. **Gevers D, Cohan F, Lawrence J.** 2005. Re-evaluating prokaryotic species. Nat Rev Microbiol **3:**733–739. http://dx.doi.org/10.1038/nrmicro1236.
11. **Gabriel MW, Matsui GY, Friedman R, Lovell CR.** 2014. Optimization of multilocus sequence analysis for identification of species in the genus *Vibrio*. Appl Environ Microbiol **80:**5359–5365. http://dx.doi.org/10.1128/AEM.01206-14.
12. **Jogler C, Lin W, Meyerdierks A, Kube M, Katzmann E, Flies C, Pan Y, Amann R, Reinhardt R, Schüler D.** 2009. Toward cloning of the magnetotactic metagenome: identification of magnetosome island gene clusters in uncultivated magnetotactic bacteria from different aquatic sediments. Appl Environ Microbiol **75:**3972–3979. http://dx.doi.org/10.1128/AEM.02701-08.
13. **Achenbach LA, Yang W.** 1997. The *fur* gene from *Klebsiella pneumoniae*: characterization, genomic organization and phylogenetic analysis. Gene **185:**201–207. http://dx.doi.org/10.1016/S0378-1119(96)00642-7.
14. **Colquhoun DJ, Sørum H.** 2002. Cloning, characterisation and phylogenetic analysis of the *fur* gene in *Vibrio salmonicida* and *Vibrio logei*. Gene **296:**213–220. http://dx.doi.org/10.1016/S0378-1119(02)00863-6.
15. **Escolar L, Pérez-Martín J, De Lorenzo V.** 1999. Opening the iron box: transcriptional metalloregulation by the Fur protein. J Bacteriol **181:**6223–6229.
16. **McHugh JP, Rodríguez-Quinoñes F, Abdul-Tehrani H, Svistunenko DA, Poole RK, Cooper CE, Andrews SC.** 2003. Global iron-dependent gene regulation in *Escherichia coli*. A new mechanism for iron homeostasis. J Biol Chem **278:**29478–29486. http://dx.doi.org/10.1074/jbc.M303381200.
17. **Vasileva D, Janssen H, Hönicke D, Ehrenreich A, Bahl H.** 2012. Effect of iron limitation and *fur* gene inactivation on the transcriptional profile of the strict anaerobe *Clostridium acetobutylicum*. Microbiology **158:**1918–1929. http://dx.doi.org/10.1099/mic.0.056978-0.
18. **Tamura K, Stecher G, Peterson D, Filipski A, Kumar S.** 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol **30:**2725–2729. http://dx.doi.org/10.1093/molbev/mst197.
19. **Nørholm MHH.** 2010. A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. BMC Biotechnol **10:**21. http://dx.doi.org/10.1186/1472-6750-10-21.
20. **Jensen S, Frost P, Torsvik VL.** 2009. The nonrandom microheterogeneity of 16S rRNA genes in *Vibrio splendidus* may reflect adaptation to versatile lifestyles. FEMS Microbiol Lett **294:**207–215. http://dx.doi.org/10.1111/j.1574-6968.2009.01567.x.
21. **Hoffmann M, Monday SR, Fischer M, Brown EW.** 2012. Genetic and phylogenetic evidence for misidentification of *Vibrio* species within the

Harveyi clade. Lett Appl Microbiol **54:**160–165. http://dx.doi.org/10.1111/j.1472-765X.2011.03183.x.

22. **Thompson CC, Thompson FL, Vicente ACP, Swings J.** 2007. Phylogenetic analysis of vibrios and related species by means of *atpA* gene sequences. Int J Syst Evol Microbiol **57:**2480–2484. http://dx.doi.org/10.1099/ijs.0.65223-0.
23. **Urbanczyk H, Ogura Y, Hayashi T.** 2013. Taxonomic revision of Harveyi clade bacteria (family Vibrionaceae) based on analysis of whole genome sequences. Int J Syst Evol Microbiol **63:**2742–2751. http://dx.doi.org/10.1099/ijs.0.051110-0.
24. **Thompson F, Iida T, Swings J.** 2004. Biodiversity of vibrios. Microbiol Mol Biol Rev **68:**403–431. http://dx.doi.org/10.1128/MMBR.68.3.403-431.2004.
25. **Pascual J, Macián MC, Arahal DR, Garay E, Pujalte MJ.** 2010. Multilocus sequence analysis of the central clade of the genus *Vibrio* by using the *16S rRNA*, *recA*, *pyrH*, *rpoD*, *gyrB*, *rctB* and *toxR* genes. Int J Syst Evol Microbiol **60:**154–165. http://dx.doi.org/10.1099/ijs.0.010702-0.
26. **Lin B, Wang Z, Malanoski AP, O'Grady EA, Wimpee CF, Vuddhakul V, Alves N, Jr, Thompson FL, Gomez-Gil B, Vora GJ.** 2010. Comparative genomic analyses identify the *Vibrio harveyi* genome sequenced strains BAA-1116 and HY01 as *Vibrio campbellii*. Environ Microbiol Rep **2:**81–89. http://dx.doi.org/10.1111/j.1758-2229.2009.00100.x.
27. **Gomez-Gil B.** 2004. *Vibrio hispanicus* sp. nov., isolated from *Artemia* sp. and sea water in Spain. Int J Syst Evol Microbiol **54:**261–265. http://dx.doi.org/10.1099/ijs.0.02775-0.
28. **Chimetto LA, Brocchi M, Gondo M, Thompson CC, Gomez-Gil B, Thompson FL.** 2009. Genomic diversity of vibrios associated with the Brazilian coral *Mussismilia hispida* and its sympatric zoanthids (*Palythoa caribaeorum*, *Palythoa variabilis* and *Zoanthus solanderi*). J Appl Microbiol **106:**1818–1826. http://dx.doi.org/10.1111/j.1365-2672.2009.04149.x.
29. **Gram L, Melchiorsen J, Bruhn JB.** 2010. Antibacterial activity of marine culturable bacteria collected from a global sampling of ocean surface waters and surface swabs of marine organisms. Mar Biotechnol **12:**439–451. http://dx.doi.org/10.1007/s10126-009-9233-y.
30. **Wietz M, Mansson M, Gotfredsen CH, Larsen TO, Gram L.** 2010. Antibacterial compounds from marine Vibrionaceae isolated on a global expedition. Mar Drugs **8:**2946–2960. http://dx.doi.org/10.3390/md8122946.

Paper 3

Photobacterium galatheae sp. nov., a bioactive bacterium isolated from a mussel in the Solomon Sea

International Journal of Systematic and Evolutionary Microbiology (2015), **65**, 4503–4507 **DOI** 10.1099/ijsem.0.000603

Photobacterium galatheae sp. nov., a bioactive bacterium isolated from a mussel in the Solomon Sea

Henrique Machado,[1,2] Sonia Giubergia,[1,2] Ramona Valentina Mateiu[3] and Lone Gram[1]

Correspondence
Lone Gram
gram@bio.dtu.dk

[1]Department of Systems Biology, Technical University of Denmark, Matematiktorvet, bldg 301, DK-2800 Kgs. Lyngby, Denmark

[2]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle Allé 6, DK-2970 Hørsholm, Denmark

[3]Center for Electron Nanoscopy, Danchip, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

A novel, Gram-negative marine bacterium, S2753^T, was isolated from a mussel of the Solomon Sea, Solomon Islands. Analysis of the 16S rRNA gene sequence and whole genome sequence data placed strain S2753^T in the genus *Photobacterium* with the closest relative being *Photobacterium halotolerans* DSM 18316^T (97.7 % 16S rRNA gene similarity). Strain S2753^T was able to grow from 15 to 40 °C and in NaCl concentrations of 0.5 to 9 % (w/v). The predominant fatty acids were 16 : 1ω7c/16 : 1ω6c (27.9 %), 16 : 0 (22.1 %) and 18 : 1ω7c/8 : 1ω6c (21.4 %). The genomic DNA G+C mol content was 49.5 mol%. Based on the phylogenetic, chemotaxonomic and phenotypic differences, strain S2753^T is considered to represent a novel species of the genus *Photobacterium*. Furthermore, whole genome sequence analysis comparing S2753^T and type-strains of closely related species of the genus *Photobacterium* also demonstrated that the strain is genomically distinct enough to be considered a novel species. The name *Photobacterium galatheae* is proposed and the type-strain is S2753^T(=LMG 28894^T=DSM 100496^T).

The genus *Photobacterium* contains 23 species with validly published names (www.bacterio.net) at the time of writing. Beijerinck (1889) first described bacteria in the genus and all of the members have been isolated from marine and aquatic environments: from sediments, water samples or samples associated with marine animals.

In this study we report on the identification and characterization of a novel isolate belonging to the genus *Photobacterium*. In 2006, Gram and colleagues took part in a global research expedition, the so-called Galathea 3 expedition (Gram *et al.*, 2010). During the expedition hundreds of bacteria with antibacterial activity were isolated from seawater and biofilm samples of marine organisms. On 31st December 2006, a swab sample was obtained from a mussel in the Solomon Sea (9 06.480 S 156 51.570 E), near the Solomon Islands. Sterile seawater (3–5 ml) was added to the swab, which was vigorously vortexed, serially diluted and plated onto half-strength Marine Agar (MA; BD Difco 2216). Plates were incubated for 3−7 days until colonies were observed. Colonies were tested for antagonistic activity against other bacteria by replica-plating of the MA plates onto a new agar plate into which *Vibrio anguillarum* had been embedded (Gram *et al.*, 2010). Bacteria capable of inhibiting *V. anguillarum* caused a clear zone in the turbid agar and were isolated and pure cultured from the original MA plates. Strain S2753^T was able to inhibit not just *V. anguillarum* (a Gram-negative fish pathogen) but also *Staphylococcus aureus* (Gram-positive human pathogen) (Gram *et al.*, 2010; Wietz *et al.*, 2010). The strain produced several secondary bioactive metabolites, including the antibiotic holomycin (Mansson *et al.*, 2011a, b).

Strain S2753^T was routinely cultivated on MA or in Marine Broth (MB; BD Difco 2216) for one day at 25 °C, unless otherwise indicated. Cell morphology was investigated by phase-contrast microscopy (Olympus BX51) and scanning electron microscopy (FEI Quanta FEG 200 ESEM) of cells grown for one day at 25 °C in MB. The ability of strain

The GenBank/EMBL/DDBJ accession number for the 16S rRNA gene sequence of strain S2753^T is KR704916.

Three supplementary figures and four supplementary tables are available with the online Supplementary Material.

S2753^T to grow at different temperatures (4, 10, 15, 20, 25, 28, 30, 35, 40, 45 °C) was tested on MA and in different NaCl concentrations [0, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 % (w/v)] in synthetic Zobell medium (5 gl^{-1} Bacto peptone, 1 gl^{-1} yeast extract, 0.1 gl^{-1} ferric citrate) (ZoBell, 1941). Gram testing was performed using the 3 % (w/v) KOH method (Gregersen, 1978) and catalase activity was tested using 3 % (v/v) H_2O_2 (Cowan, 1974). Oxidase activity was assessed using a BBL DrySlide (BD). Antibiotic sensitivity was tested on Iso-sensitest agar (CM0471B, Oxoid) supplemented with 1.5 % (w/v) NaCl using the following antibiotics: 2,4-diamino-6,7-diisopropylpteridine (vibriostatic agent O129 at 10 and 150 µg per disc), chloramphenicol (30 µg/disc), erythromycin (15 µg/disc), tetracycline (30 µg per disc) and penicillin (1.5 IU). API 20 NE (20050, Biomerieux), API ZYM (25200, Biomerieux) and Biolog GN2 (Biolog Inc., USA) tests were performed in accordance with the manufacturers' instructions when comparing metabolic and enzymatic profiles of S2753^T and closely related type-strains of species of the genus *Photobacterium*. Because of the salt requirements for growth of the marine strains, a solution of 1.5 % (w/v) NaCl was used when resuspending colony mass, unless a specific solution was proposed in the manufacturer's protocol. The Identification Service at DSMZ, Braunschweig, Germany, carried out fatty acid analyses using the Sherlock Microbial Identification System (MIDI).

Cells of S2753^T were straight, rod-shaped, with mean length 2.2 ± 0.4 µm and mean width 0.99 ± 0.13 µm. They had a polar flagellum of 4.3 ± 1.14 µm (Fig. S1, available in the online Supplementary Material). The strain was Gram-negative and oxidase- and catalase-positive. S2753^T grew between 15 and 40 °C, and in 0.5 to 9 % (w/v) NaCl. The optimum temperature and salinity were 28 °C and 1–6 % (w/v) NaCl. After 48 h incubation at 25 °C in MA the strain grew in colonies of 3–5 mm diameter, depending on the cell mass and the level of swarming of each colony. S2753^T grew aerobically and anaerobically (using an anaerobic jar and Gas Generaion kit BR0039, Oxoid) and it grew on thiosulfate citrate bile salts sucrose medium (TCBS, CM0333, Oxoid) with small, green colonies (approximately 1 mm). The strain was sensitive to the vibriostatic agent O129, chloramphenicol, erythromycin and tetracycline, but resistant to penicillin (Table 1). Further results from the phenotypical and Biolog GN2 assays performed are available as supplementary data (Tables S1–S4).

The fatty acid profile of strain S2753^T was similar to those of the type strains of closely related species of the genus *Photobacterium* (Table 2) with the dominant cellular fatty acids being 16 : 1ω7*c*/16 : 1ω6*c* (27.89 %), 16 : 0 (22.11 %) and 18 : 1ω7*c*/18 : 1ω6*c* (21.41 %).

The presence of several different 16S rRNA gene sequences in species of the order *Vibrionaceae* has been previously reported (Machado & Gram, 2015) and we identified 12 copies of this gene in the closed genome of S2753^T (unpublished) (Fig. S2). The 16S rRNA gene sequences were not identical, although nine copies shared great similarity. A gene representative of the nine most conserved sequences was chosen to represent the 16S rRNA gene sequence of S2753^T (Fig. S3). This sequence was also the one confirmed by both culture collections where the strain was deposited. The extracted 16S rRNA gene (KR704916) was compared to the 16S rRNA genes from type-strains of species of the genus *Photobacterium* using CLC Main Workbench, version 7 (CLC Bio). The analysis included the generation of a phylogenetic tree using the neighbour-joining method with the Jukes–Cantor nucleotide distance measure. Phylogenetic trees were visualized using MEGA 6 (Tamura *et al.*, 2013). Sequence similarity calculations pointed to the closest relatives of strain S2753^T being *Photobacterium halotolerans* MACL01^T, *Photobacterium aquae* AE6^T, *Photobacterium aphoticum* CECT 7614^T and *Photobacterium ganghwense* FR1311^T, with similarity percentages of 97.67 %, 96.35 %, 96.95 %, and 96.00 %, respectively (Fig. 1). This close relationship was also apparent from the topology of the multi-locus sequence analysis (MLSA) based phylogenetic tree (Fig. 2). The MLSA was performed using seven gene sequences (*gyrB*, *pyrH*, *recA*, *topA*, *gapA*, *ftsZ*, and *rpoA*), which were found and extracted from the whole

Table 1. Differential phenotypic, physiological and biochemical characteristics of strain S2753^T and related strains of species of the genus *Photobacterium*

Strains: 1, S2753^T; 2, *P. halotolerans* DSM 18316^T; 3, *P. ganghwense* DSM 22954^T; 4, *P. aphoticum* DSM 25995^T; 5, *P. aquae* CGMCC 1.12159^T. +, Positive; −, negative; ND, not determined. All data were generated in this study.

Characteristic	1	2	3	4	5
Enzymic profile (API 20 NE and API ZYM):					
Indole Production	−	−	+	−	−
Arginine dihydrolase	−	−	+	+	+
β-Galactosidase	−	+	+	+	−
Assimilation of (API 20 NE):					
L-Arabinose	−	+	−	−	−
D-Mannose	−	+	+	−	−
Potassium gluconate	−	+	+	−	−
Capric acid	−	−	+	−	−
Phenylacetic acid	−	−	+	−	−
Oxidization of (Biolog GN2):					
N-Acetyl-D-galactosamine	+	−	−	+	−
L-Arabinose	−	+	−	−	−
D-Galactose	−	−	+	+	−
D-Mannose	−	+	−	+	+
β-Methyl-D-glucoside	−	+	+	+	+
Sucrose	−	+	−	−	−
cis-Aconitic acid	−	+	+	−	−
D-Gluconic acid	−	+	+	+	−
L-Pyroglutamic acid	−	+	+	−	+
Growth at 10 °C	−	+	ND	ND	ND
Resistance to Penicillin (1.5 IU)	+	−	−	−	−
DNA G+C content (mol%)	49.50	50.90	50.50	49.70	49.10
Genome size (Mb)	4.53	4.69	5.54	5.26	5.08

Table 2. Cellular fatty acid profiles of strain S2753^T and related strains of species of the genus *Photobacterium*

Strains: 1, S2753^T; 2, *P. halotolerans* DSM 18316^T; 3, *P. ganghwense* DSM 22954^T; 4, *P. aphoticum* DSM 25995^T; 5, *P. aquae* CGMCC 1.12159^T. Values given are percentages of total fatty acids. −, Not detected (<0.25 %); TR, trace amounts (0.25–1 %). All data were generated in this study.

Fatty acids	1	2	3	4	5
10 : 0	−	TR	−	−	−
12 : 0	3.92	5.07	3.41	1.96	1.67
11 : 0 3OH	−	−	−	TR	TR
13 : 0 iso	−	−	TR	TR	1.22
13 : 0	−	−	−	TR	TR
12 : 0 3OH	6.11	4.74	2.80	1.86	1.73
14 : 00	1.32	TR	3.30	3.57	2.70
Summed feature 1 (15 : 1 iso H/13 : 0 3OH)	TR	TR	TR	TR	TR
15 : 0 iso	−	−	TR	2.05	2.55
15 : 0 anteiso	−	−	−	TR	2.10
15 : 1ω8*c*	TR	TR	TR	1.53	TR
15 : 1ω6*c*	−	−	−	TR	TR
15 : 0	1.30	1.18	1.15	7.14	2.82
Summed feature 2 (12 : 0 aldehyde?)	4.19	3.85	3.40	2.16	2.16
16 : 0 iso	TR	1.50	TR	TR	1.48
16 : 1ω9*c*	2.19	1.75	−	−	−
Summed feature 3 (16 : 1ω7*c*/16 : 1ω6*c*)	27.89	30.18	31.47	42.96	40.49
16 : 1ω5*c*	TR	−	TR	TR	−
16 : 0	22.11	22.72	21.97	15.09	9.00
17 : 1 isoω9*c*	−	−	TR	1.04	2.07
17 : 0 iso	TR	−	1.01	1.35	2.06
17 : 0 anteiso	−	−	−	1.31	3.52
17 : 1ω6*c*	−	−	TR	TR	TR
17 : 1ω8*c*	2.21	1.43	1.04	1.12	1.01
17 : 0	2.06	1.91	1.61	1.89	1.11
18 : 1ω9*c*	TR	TR	−	−	−
Summed feature 8 (18 : 1ω7*c*/18 : 1ω6*c*)	21.41	22.21	26.17	14.90	18.33
18 : 0	TR	TR	TR	−	−

genome sequences by a similarity-based search to known sequences (Gabriel *et al.*, 2014).

To assess the genomic relatedness of S2753^T and closely related strains of species of the genus *Photobacterium*, whole genome sequencing was performed on S2753^T (Machado *et al.*, 2014), *P. ganghwense* DSM 22954^T (LDOU01), *P. aphoticum* DSM 25995^T (LDOV01) and *P. aquae* CGMCC 1.12159^T (LDOT01). The genome of *P. halotolerans* DSM 18326 was available on GenBank (AULG01). The reference type-strains were obtained from the Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures and from the China General Microbiological Culture Collection, as indicated by the strain designation.

High purity genomic DNA was extracted by successive phenol/chloroform:isoamyl-alcohol purification steps followed by precipitation with 2-propanol, treatment with RNase, and a final purification and precipitation step (Sambrook & Russel, 2001). Quantification of genomic DNA was carried out in 1 % agarose gel electrophoresis, using a NanoDrop Spectrometer (Saveen Werner) and a Qubit 2.0 Analyser (Invitrogen).

Sequencing of the S2753^T genome was carried out by the Beijing Genomic Institute (Shenzhen, China). Libraries of 500 bp were used for 100 bp paired-end sequencing of genomes using the Illumina sequencing technology on a HiSeq2000. Sequencing of the *P. ganghwense*, *P. aphoticum* and *P. aquae* type strains was performed at the Novo Nordisk Foundation Center for Biosustainability (Horshølm, Denmark). Libraries of 300–400 bp were used for 151 bp paired-end sequencing of genomes using the Illumina sequencing technology on a MiSeq. The data were assembled to contigs (>200 bp) using the *De novo* Assembly in CLC Genomic Workbench, version 7 (CLC Bio). The genomes were compared using the Genome-to-Genome Distance Calculator 2.0 (GGDC) tool from DSMZ (http://ggdc.dsmz.de/) (Auch *et al.*, 2010a, b; Meier-Kolthoff *et al.*, 2013) and the Average Nucleotide Identity (ANI) calculator from Kostas Lab (http://enve-omics.ce.gatech.edu/ani/) (Goris *et al.*, 2007). ANI values comparing S2753^T to other species of the genus *Photobacterium* were less than 95 % (the typical threshold for species delimitation) and all DNA–DNA-hybridization values (GGDC) were less than 22 % (Table 3), where 70 % is usually the limit for species identity (Auch *et al.*, 2010a, b; Meier-Kolthoff *et al.*, 2013).

Based on the phenotypic, phylogenetic and genomic data, strain S2753^T should be recognized as a novel species within the genus *Photobacterium* and the name *Photobacterium galatheae* sp. nov. is proposed.

Description of *Photobacterium galatheae* sp. nov.

Photobacterium galatheae (ga.la.the'ae. N.L. gen. n. *galatheae* referring to the research expedition on which the type strain was first isolated).

Cells are motile rods, 1.6–3.0 μm long and 0.74–1.2 μm wide, with polar flagella 2.5–6.9 μm in length. Cells are Gram-stain-negative, non-spore-forming, facultatively anaerobic and occur mainly as single cells. When grown on MA 2216 at 25 °C, colonies are irregular in shape and size, depending on the biomass and swarming levels. Growth occurs from 15 to 40 °C and in NaCl concentrations of 0.5 to 9 % (w/v). Utilizes dextrin, glycogen, Tween 40, Tween 80, *N*-acetyl-D-galactosamine, *N*-acetyl-D-glucosamine, D-fructose, α-D-glucose, maltose, D-mannitol, D-psicose, trehalose, pyruvic acid methyl ester, acetic acid, citric acid, β-hydroxybutyric acid, DL-lactic acid, succinic acid, bromosuccinic acid, D-alanine, L-alanine, L-alanyl-glycine, L-asparagine, L-aspartic acid, L-glutamic acid, L-histidine,

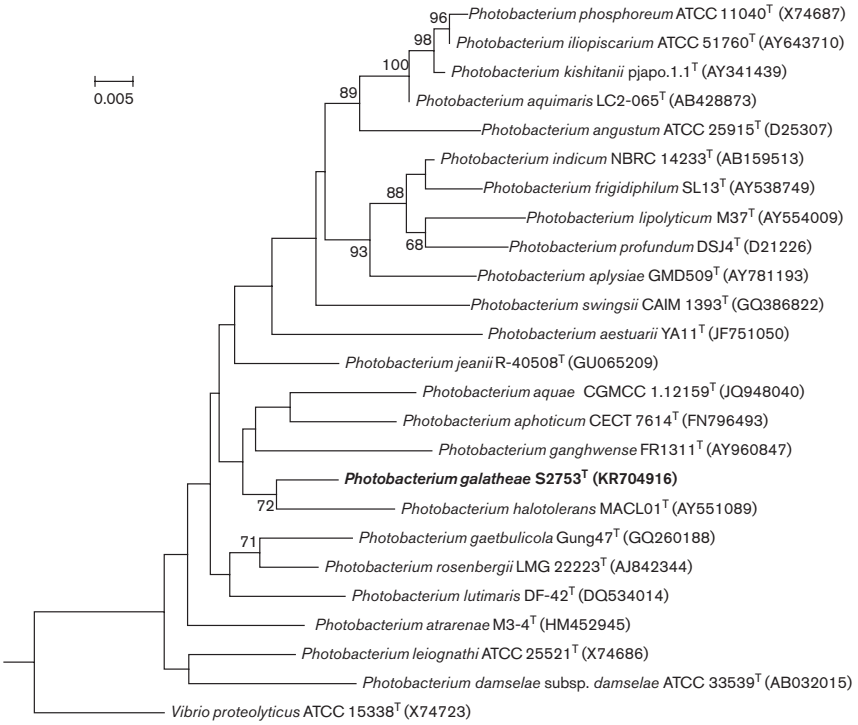


Fig. 1. Phylogenetic tree based on 16S rRNA gene sequences of type strains of species of the genus *Photobacterium*. Strain S2753^T appeared to cluster with *Photobacterium halotolerans* MACL01^T, *Photobacterium aquae* AE6^T, *Photobacterium aphoticum* CECT 7614^T and *Photobacterium ganghwense* FR1311^T, with percentage similarities of 97.67 %, 96.35 %, 96.95 % and 96.00 %, respectively. The nodes with bootstrap support of 70 or more are indicated (1000 replications). GenBank accession numbers are indicated in parentheses. *Vibrio proteolyticus* ATCC 15338^T was used to root the tree. Bar, 0.005 nucleotide substitutions per position.

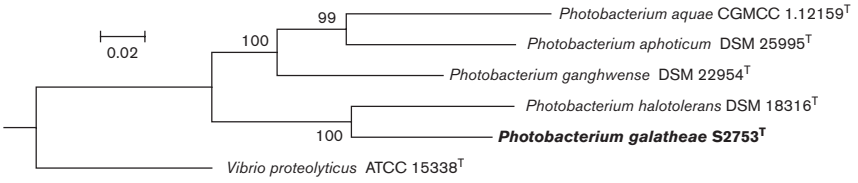


Fig. 2. Phylogenetic tree based on MLSA analysis using seven gene sequences (*gyrB*, *pyrH*, *recA*, *topA*, *gapA*, *ftsZ* and *rpoA*) of closely related type strains of species of the genus *Photobacterium*. Strain S2753^T appeared clustered with *Photobacterium halotolerans* MACL01^T. The nodes with bootstrap support of 70 or more are indicated (1000 replications). GenBank accession numbers are indicated. *Vibrio proteolyticus* ATCC 15338^T was used to root the tree. Bar, 0.02 nucleotide substitutions per position.

Table 3. DDH and ANI percentages of strain S2753^T and type strains of other species of the genus Photobacterium determined using in silico approaches

Strains: 1, S2753^T; 2, *P. halotolerans* DSM 18316^T; 3, *P. ganghwense* DSM 22954^T; 4, *P. aphoticum* DSM 25995^T; 5, *P. aquae* CGMCC 1.12159^T.

		Two-way ANI values (%±SD)				
		1	**2**	**3**	**4**	**5**
DDH – estimate (%±SD)	1		81.02±4.99	79.53±5.39	78.88±4.97	78.87±5.40
	2	22.10±2.36		79.01±4.61	78.57±4.70	78.52±4.97
	3	21.90±2.35	21.00±2.33		80.17±4.98	80.19±5.29
	4	21.00±2.33	20.40±2.32	21.70±2.35		81.04±6.28
	5	20.20±2.31	19.70±2.30	21.10±2.33	21.80±2.35	

L-proline, L-serine, L-threonine, urocanic acid, inosine, uridine, thymidine, putrescine, glycerol, adipic acid, malic acid and trisodium citrate as sole carbon sources. It does not utilize L-arabinose, D-mannose, potassium gluconate, capric acid, D-galactose, beta-methyl-D-glucoside, sucrose, *cis*-aconitic acid, D-gluconic acids, L-pyroglutamic acid and phenylacetic acid. Strain S2753^T reduces nitrate to nitrite and hydrolyses gelatin but it does not hydrolyse aesculin. It is negative for indole production, urease, arginine dihydrolase and β-galactosidase activities. The primary fatty acids are 16 : 1ω7*c*/16 : 1ω6*c* (27.89 %), 16 : 0 (22.11 %) and 18 : 1ω7*c*/18 : 1ω6*c* (21.41 %). Susceptible to the vibriostatic agent O129, chloramphenicol, erythromycin and tetracycline, but resistant to penicillin.

The type strain is S2753^T(=LMG 28894^T=DSM 100496^T), which was isolated from the surface of a mussel in the Solomon Sea, Solomon Islands. The DNA G+C mol content of the type strain is 49.5 %.

Acknowledgements

H. M. and S. G. were supported by a PhD grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7-People-2012-ITN, under grant agreement No. 317058, 'BACTORY'. The present work was carried out as part of the Galathea 3 expedition under the auspices of the Danish Expedition Foundation. This is Galathea 3 contribution no. p112.

References

Auch, A. F., von Jan, M., Klenk, H.-P., Göker, M. & Digital, D. N. A-D. N. A. (2010a). hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci* **2**, 117–134.

Auch, A. F., Klenk, H.-P. & Göker, M. (2010b). Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand Genomic Sci* **2**, 142–148.

Beijerinck, M. W. (1889). Le *Photobacterium luminosum*, Bactérie lumineuse de la Mer du Nord. *Arch Néerlandaises des Sci Exactes Nat*, 401–427.

Cowan, S. T. (1974). *Cowan and Steel's Manual for the Identification of Medical Bacteria*, 2nd edn. Cambridge: Cambridge University Press.

Gabriel, M. W., Matsui, G. Y., Friedman, R. & Lovell, C. R. (2014). Optimization of multilocus sequence analysis for identification of species in the genus *Vibrio*. *Appl Environ Microbiol* **80**, 5359–5365.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P. & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**, 81–91.

Gram, L., Melchiorsen, J. & Bruhn, J. B. (2010). Antibacterial activity of marine culturable bacteria collected from a global sampling of ocean surface waters and surface swabs of marine organisms. *Mar Biotechnol (NY)* **12**, 439–451.

Gregersen, T. (1978). Rapid method for distinction of gram-negative from gram-positive bacteria. *Eur J Appl Microbiol Biotechnol* **5**, 123–127.

Machado, H. & Gram, L. (2015). The fur gene as a new phylogenetic marker for *Vibrionaceae* species identification. *Appl Environ Microbiol* **81**, 2745–2752.

Machado, H., Månsson, M. & Gram, L. (2014). Draft genome sequence of *Photobacterium halotolerans* S2753, producer of bioactive secondary metabolites. *Genome Announc* **2**, e00535–14.

Mansson, M., Gram, L. & Larsen, T. O. (2011a). Production of bioactive secondary metabolites by marine *vibrionaceae*. *Mar Drugs* **9**, 1440–1468.

Mansson, M., Nielsen, A., Kjærulff, L., Gotfredsen, C. H., Wietz, M., Ingmer, H., Gram, L. & Larsen, T. O. (2011b). Inhibition of virulence gene expression in *Staphylococcus aureus* by novel depsipeptides from a marine *photobacterium*. *Mar Drugs* **9**, 2537–2552.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P. & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**, 60.

Sambrook, J. & Russel, D. W. (2001). *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbour, New York: Cold Spring Harbour Laboratory Press.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**, 2725–2729.

Wietz, M., Mansson, M., Gotfredsen, C. H., Larsen, T. O. & Gram, L. (2010). Antibacterial compounds from marine *Vibrionaceae* isolated on a global expedition. *Mar Drugs* **8**, 2946–2960.

ZoBell, C. E. (1941). Studies on Marine Bacteria. I. The cultural requirements of heterotrophic aerobes. *J Mar Res* **4**, 42–75.

Paper 4

Furios: a web-based tool for identification of *Vibrionaceae* species using the *fur* gene

FurIOS: a web-based tool for identification of *Vibrionaceae* species using the *fur* gene

Henrique Machado[1,2], João Cardoso[2], Sonia Giubergia[1,2], Kristoffer Rapacki[3], Lone Gram[1,*]

* corresponding author:

e-mail: gram@bio.dtu.dk

[1] Department of Systems Biology, Technical University of Denmark, Matematiktorvet, bldg. 301, DK-2800 Kgs. Lyngby, Denmark

[2] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle Allè 6, DK-2970 Hørsholm, Denmark

[3] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, bldg. 208, DK-2800 Kgs. Lyngby, Denmark

Abstract

Background

Gene based methods for identification of species from the *Vibrionaceae* family has developed during the last decades to address the limitations of the commonly used 16S rRNA gene phylogeny. Recently, we found that the *ferric-uptake regulator* gene (*fur*) can be used as a single phylogenetic marker providing species discrimination, consistent with Multi-Locus Sequencing Analyses and whole genome phylogenies.

Results

To allow for broader and easy use of this marker, we have developed an online prediction service that allows the identification of *Vibrionaceae* species based on the *fur*-sequence. The input is a DNA sequence that can be uploaded on the web service; the output is a table containing the strain identifier, e-value, and percentage of identity for each of the matches with rows coloured in green for hits with high probability of being the same species. The service is available on the web at: http://www.cbs.dtu.dk/services/furIOS-1.0/. The *fur*-sequences can be derived either from genome sequences or from PCR-amplification of the genomic region encoding the *fur* gene. We have used 191 strains identified as *Vibrionaceae* based on 16S rRNA gene sequence to test the PCR method and the web service on a large dataset. We were able to classify 171 of 191 strains at the species level and 20 strains remained unclassified. Furthermore, the *fur*

phylogenetics and sub-sequent *in silico* DNA-DNA hybridization demonstrated that two strains (ATCC 33789 and ZS-139) that have previously been misidentified as *V. splendidus* and are more closely related to *V. tasmaniensis* and *V. cyclitrophycus*, respectively.

Conclusion

FurIOS is an easy-to-use online service that allows the identification of bacteria from the *Vibrionaceae* family at the species level using the *fur* gene as a single phylogenetic marker. Its simplistic design and straightforward pipeline makes it suitable for any research environment, from academia to industry.

Keywords: *Vibrionaceae*, *Vibrio*, *Photobacterium*, *fur* gene, phylogeny, identification

Background

A key aspect in microbial taxonomy is the identification of microorganisms at the species or genus level. This is important to distinguish pathogenic species in health and environmental sciences, to identify beneficial or symbiotic species and also to study microbial diversity in environmental niches in Nature. Due to the widespread use and importance of bacterial identification, methods and processes should preferably be affordable, fast and easy to use [1].

The *Vibrionaceae* is a large family of marine Gram-negative gamma-proteobacteria, which includes organisms of different environmental importances, *e.g.* symbiotic (*e.g. Vibrio fischeri*), bioactive (*e.g. V. coralliilyticus* and *Photobacterium galatheae*) and pathogenic organisms (*e.g. V. cholerae* and *Aliivibrio salmonicida*). Several of the human pathogenic *Vibrio* species, such as *V. cholera* and *V. parahaemolyticus,* are mesophilic organisms and changes in sea water temperature and salinity influence their distribution [2–5]. The increase in sea water temperature has been linked to a number of human illness outbreaks caused by *Vibrio* species (*e.g. V. cholera*, *V. parahaemolyticus*, *V. vulnificus*) [2] and to several epidemics in marine animals, such as oysters, salmon, sea bass, eel, trout, shrimps and corals, caused by *V. salmonicida*, *V. anguillarum*, *P. damselae*, *V. vulnificus* and *V. coralliilyticus* [2,4,6]. The rising sea water temperatures have also been linked to the increase in the numbers of *V. cholera* associated with plankton in the North Sea [7]. The increased spread of *Vibrio* pathogens and their importance as infectious and/or food

poisoning agents has a direct impact on health of mankind and requires changes in microbiological food control processes and clinical settings [8]. Any survey or intervention to reduce risk requires that the organisms can be rapidly and correctly identified.

Identification of species from the *Vibrionaceae* family has primarily been based on Multi-Locus Sequencing Analysis (MLSA), which relies on the amplification and sequencing of up to nine genes (*ftsZ*, *gapA*, *gyrB*, *mreB*, *pyrH*, *recA*, *rpoA*, *topA*, and the 16S rRNA gene) [9–12]. This has been necessary due to the limitations of 16S rRNA gene phylogeny in this family, where several (7-15) different 16S rRNA genes are encoded in a single genome [13–15]. We have recently identified the *ferric-uptake regulator* gene (*fur*) as a new phylogenetic marker in the *Vibrionaceae* family, and developed a PCR based method for the amplification of the genomic region encoding the *fur* gene [14]. Its use as a single phylogenetic marker in the classification of *Vibrionaceae* at the species level could reduce the timing and cost of strain identification. To enable the use of this potential worldwide, we here describe the design of an online platform, FurIOS 1.0, which allows an easy and fast identification of *Vibrionaceae* species using only their *fur* DNA sequence. We also use a collection of 191 *Vibrionaceae* strains to demonstrate its potential.

Implementation

Database design

A *ferric-uptake regulator* gene (*fur*) sequences database was created extracting the sequences from available whole genome sequences and using the data collected during the design of the gene amplification methodology [14]. The BLAST compatible version was generated using *makeblastdb* from NCBI BLAST+ command line tools. This database includes 134 sequences representing 78 species of the *Vibrionaceae* family, covering the 6 genera (*Vibrio*, *Photobacterium*, *Aliivibrio*, *Grimontia*, *Enterovibrio* and *Salinivibrio*). The use of the *fur* gene as a phylogenetic marker and the publication of more *fur* sequences will allow the improvement of this database by increasing the number of species represented.

Implementation of FurIOS 1.0

FurIOS is implemented in Python, compatible with version 2. The BLAST is performed using NCBI BLAST+ version (2.2.28+) [16], *blastn* calls and output parsing is handled using BioPython [17]. The script verifies the format of the input, parameterizes the *blastn* command call and formats the output. BLAST runs with the following parameters: e-value of 10; gap open penalty of 5; gap extension penalty of 2; mismatch penalty of -3; match reward of 2; word size of 11; maximum number of returned alignments of 50. These parameters are the default CLC Main Workbench (CLC Aarhus, Denmark version 7)

used in the design of the method and evaluation of the potential of the *fur* as a phylogenetic marker [14]. The web interface is provided by the Center for Biological Sequence Analysis (CBS) at the Department of Systems Biology, Technical University of Denmark and the webpages are designed according to their standards. Documentation can be also found on the website.

Testing of the identification pipeline

Bacterial strains and genomic DNA extraction

The bacterial strains used in the implementation of the identification pipeline were 191 *Vibrionaceae* strains from the Galathea 3 culture collection [18]. These strains have been identified as *Vibrionaceae* by analysis of the 16S rRNA gene sequence [18]. Here, we aimed at a more specific classification, at the species level. Strains were grown overnight at 25 °C and 200 r.p.m. in Marine Broth (Difco 279110), before genomic DNA was extracted using the NucleoSpin® Tissue Kit (Macherey-Nagel, Düren, Germany). Genomic DNA quality was checked by 1% agarose gel electrophoresis and quantified by absorbance using DeNovix DS-11 (DeNovix Inc., Wilmington, USA).

PCR amplification and sequencing

Amplification of the *fur* gene was performed as previously described [14], with few changes. Briefly, amplifications were performed in a total volume of 25 µL using 5 ng final concentration of genomic DNA as

template, 0.2 µM of each primer (fur_AP_fw and fur_AP_rv [14]) and TEMPase Hot Start Master Mix Blue (Ampliqon A/S, Odense, Denmark), following the producers instructions. The PCR amplification was carried out in a thermal cycler (Applied Biosystems® Veriti® 96-Well Thermal Cycler) as follows: 15 min initial denaturation step at 95 °C, followed by 30 cycles of 95 °C for 25 s, 52 °C for 25 s, and 72 °C for 1 min, with a final extension step of 5 min at 72 °C. The amplified products were visualised after agarose gel electrophoresis (1%) and ethidium bromide staining. The PCR products were enzymatically purified by treatment with Exonuclease I (ExoI) (Thermo Scientific) and FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific) before being sequenced by Macrogen (Macrogen Europe, Amsterdam, The Netherlands). The sequencing reads were assembled and analysed using the CLC Main Workbench (CLC Aarhus, Denmark version 7).

Species identification

The *fur* gene sequences were merged in a single FASTA file. This file contained 191 *fur* sequences, each with a headline identifying the strain number (*e.g.* ">S2757"). This file was uploaded to the web interface of the FurIOS 1.0 (http://www.cbs.dtu.dk/services/furIOS-1.0/) and the analyses performed. The highest hit was selected from the output for each of the strains (Table S1) and the strain classified at the species level if the percentage of identity was above 95%. In cases where lower than 95% identity was obtained, the strain was classified

as a possible new species. The *fur* gene sequences have been deposited on GenBank under accession numbers KU756296 – KU756481, KP721394, KP721390, KP721391, KP721399 and KP721400.

Genomic analyses of *Vibrio splendidus* strains

In silico DNA-DNA hybridization was performed for strains of the *V. splendidus* representing the three different *fur* phylogenetic clusters of this species. The genomes were compared using the Genome-to-Genome Distance Calculator 2.1 (GGDC) tool from DSMZ (http://ggdc.dsmz.de/) [19–21]. The used whole genome sequences of *V. splendidus* strains FF-500, 1F-157, FF-6, ZF-90, 1S-124, ZS-139, ATCC 33789 and NCCB 53037^{T} are publically available at NCBI under whole-genome sequenced accession numbers AJZH00000000, AJZJ00000000, AJZI00000000, AJZF00000000, AJZL00000000, AJZE00000000, AFWG00000000, LNQX00000000, respectively.

Results and Discussion

The FurIOS web service

Users can access FurIOS via the web interface: http://www.cbs.dtu.dk/services/furIOS-1.0/. The input can be the whole *fur* open-reading frame of the sequence of the strain to be identified, the whole sequenced fragment, whole genome sequences or metagenomics data (with a maximum of 100 contigs of up to 200,000 nucleotides, making a total of 20 million nucleotides per submission).

Submission of untreated sequences (raw sequencing data) is possible and decreases the need for sequence processing by the user, thereby accelerating the process. The sequences can be uploaded on the web service by "copy and paste" or using a file in FASTA format (Fig. 1A). The output is a table containing the strain identifier, e-value, and percentage of identity for each of the matches with rows coloured in green for hits with percentage of identity higher than 95% (Fig. 1B), representing a high probability of being the same species [14]. The service is provided with pre-established settings used in the design and evaluation of *fur* as a phylogenetic marker [14], however a portable version where parameters can be selected by the user will be provided upon request.

Sequence based identification has in other settings also been developed into online tools, which have been successfully used for several years for MLSA of pathogenic bacteria (*e.g.* http://mlst.warwick.ac.uk/mlst/; http://www.mlst.net/; http://bigsdb.web.pasteur.fr/). A tool based on MLSA has been under development for identification of *Vibrio* species (http://www.lge.ibi.unicamp.br/temp/vibrionaceae/), however is appears not to be functional. FurIOS is therefore the first functional online service developed for the classification of *Vibrionaceae* species and it is easier to implement and use than the MLSA based analyses, because it uses a single gene with great discriminatory power [14,22].

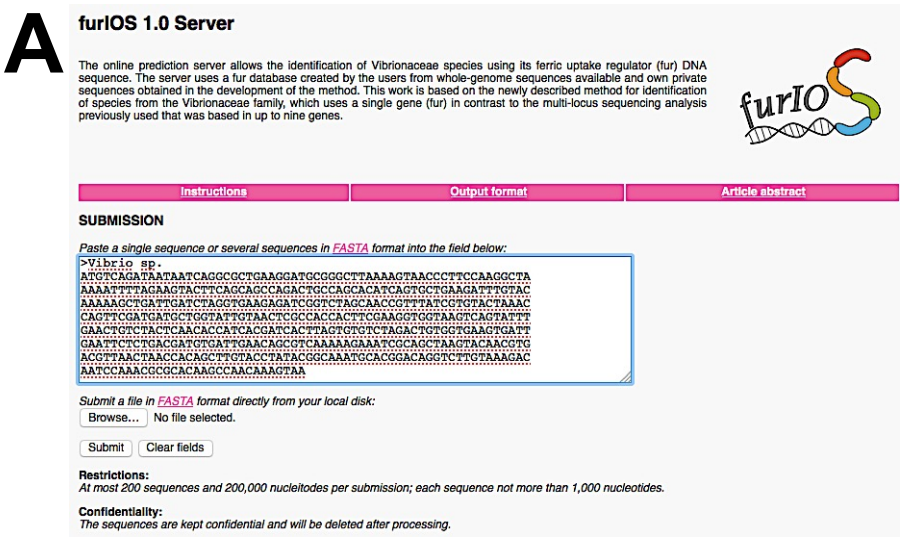


Figure 1 – Workflow and visualization example. (A) Example of sequence submission, this should be in fasta format and can be a multiple sequence submission. (B) Example of a result in table format, which includes the E-value and the percentage of identity; the predicted species is coloured in green.

The corresponding author will act as a curator of the database. Newly published *fur* sequences can be send to the curator, who will use them to update the FurIOS database, in order for the database to be up to date. Publications making use of FurIOS will also be checked for relevant sequences to be added.

Characterization of *Vibrionaceae* isolates

A selection of 191 strains belonging to the Galathea 3 collection [18] was chosen to test the methodology here presented. These strains have been identified as *Vibrionaceae* by 16S rRNA gene sequencing [18], and were used to assess if species affiliation could be done using the FurIOS database. The *fur* sequences were amplified, sequenced and submitted to the FurIOS web service. For each of the strains, the highest identity hit was extracted (Table S1). Species affiliation could be assigned to 171 of the 191 strains (Fig. 2). Based on the *fur* sequence, the remaining 20 isolates had an identity lower than 95% and were therefore classified as "other species" (Fig. 2). These 20 strains could potentially be new *Vibrio* species or species not yet represented in the FurIOS database, due to unavailability of whole genome or *fur* sequences. Of the 20 "other species", 15 were 95 to 90% similar to the *fur* gene sequences of known species, whereas the remaining five strains had similarities between 89 and 80% (Table S1).

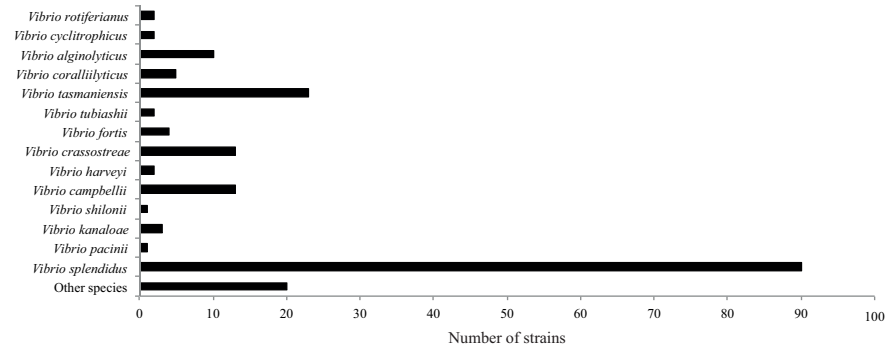


Figure 2 – Classification of 191 strains belonging to the Galathea 3 *Vibrionaceae* culture collection. The identification corresponds to the highest hit obtained from the analyses using FurIOS 1.0. In cases where lower than 95% identity was obtained, the strain was classified as "other species".

The *fur* gene sequence of strain S2757 had an 89% identity to *V. tubiashii* DSM 19142. The isolate was therefore classified as "other species". This strain has been further studied and has been recently described as the type strain of the new species *V. galatheae* [22]. In that study [22], 16S rRNA, MLSA and *fur* gene analyses were performed and the *fur* gene, as a single gene analysis, provided as good discrimination as did the MLSA analysis [22]. This confirms that the *fur* gene analyses can also be used in the classification of new species belonging to the *Vibrionaceae* family.

Ninety of the 171 strains were identified as *V. splendidus*, however, these did not cluster as one tight single cluster but were distributed across the phylogenetic tree (Fig. 3). This species was represented in three clusters, two large and one small, all placed within the *Splendidus* clade, to which *V. fortis*, *V. cyclitrophicus*, *V. crassostreae*,

V. tasmaniensis and *V. kanaloae* species also belong [12]. The three *V. splendidus* clusters correspond to *fur* homology to different *V. splendidus* strains. The first cluster contained strains with *fur* homology to *V. splendidus* strains FF-6, FF-500, 1F-157, 1S-124 and ZF-90 (Table S1). The second smaller cluster and the third cluster contained strains with *fur* homology to *V. splendidus* ZS-139 and ATCC 33789, respectively. The different clustering suggests a different phylogenetic relationship between these strains, previously identified as the same species.

These discrepancies have been previously reported and attributed to the possible misidentification of *Vibrio* strains [23–25] or to the genetic diversity and polyphyletic nature of *V. splendidus* [9,26,27]. To address this, we performed *in silico* DNA-DNA hybridization using the whole genome sequences of the *V. splendidus* strains represented in the three clusters (FF-6, FF-500, 1F-157, 1S-124, ZF-90, ZS-139 and ATCC 33789) and the recently published genome of the *V. splendidus* type strain NCCB 53037^{T} (Table 1). Based on DNA-relatedness, strains ATCC 33789 and ZS-139 were not similar enough to *V. splendidus* type-strain to be considered the same species and possibly represent new species. The *fur* sequences with high homology to the sequence from strain ZS-139 were closely related to *V. cyclitrophycus* (93% identity) while homology to strain ATCC 33789 placed them phylogenetically close to *V. tasmaniensis* (91% identity).

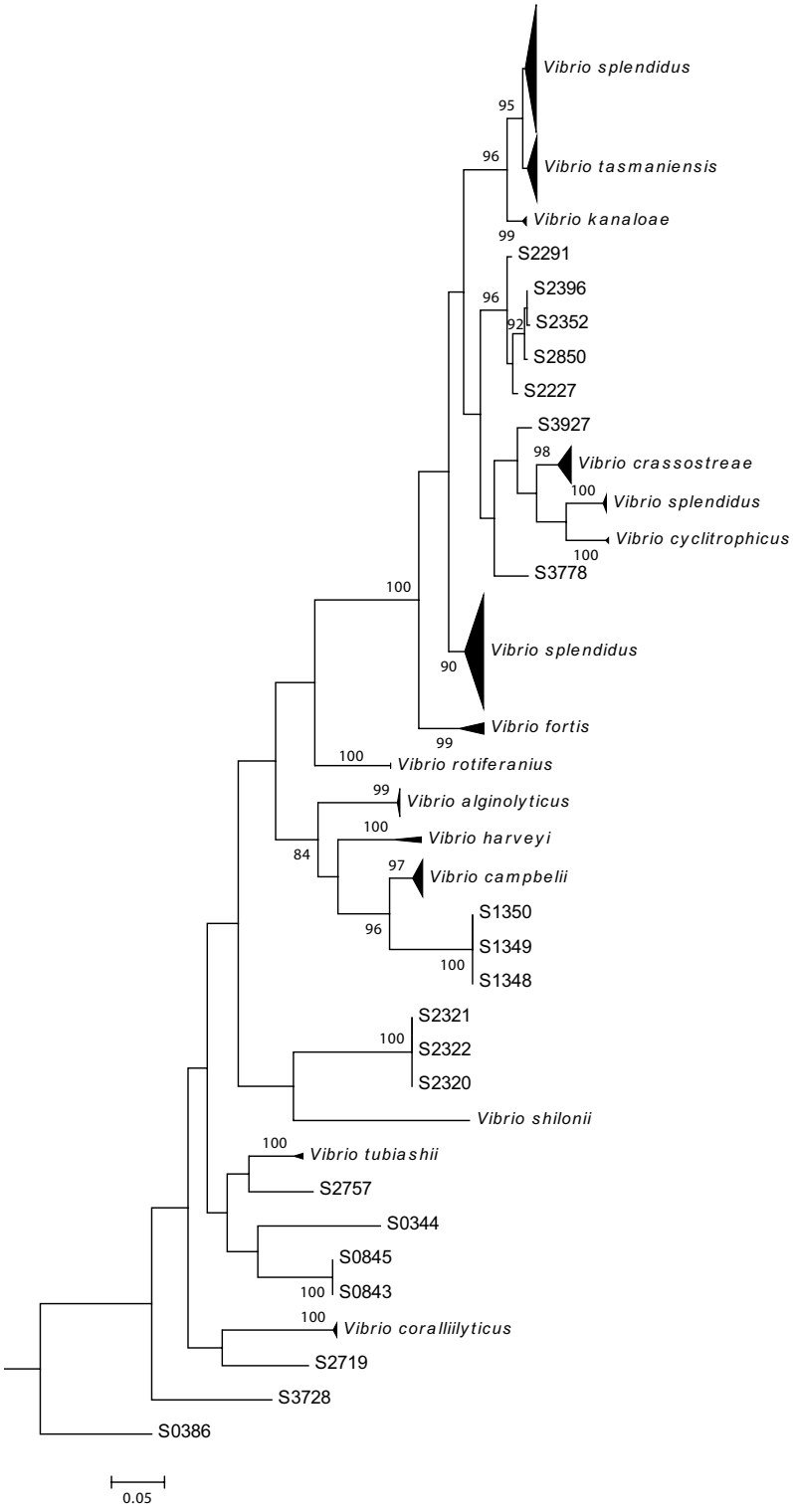

Vibrio splendidus
95
96
Vibrio tasmaniensis
Vibrio kanaloae
99
S2291
96
S2396
92
S2352
S2850
S2227
S3927
98
Vibrio crassostreae
100
Vibrio splendidus
Vibrio cyclitrophicus
100
S3778
100
Vibrio splendidus
90
Vibrio fortis
99
100
Vibrio rotiferanius
99
Vibrio alginolyticus
100
Vibrio harveyi
84
97
Vibrio campbelii
S1350
96
S1349
100
S1348
S2321
100
S2322
S2320
Vibrio shilonii
100
Vibrio tubiashii
S2757
S0344
S0845
100
S0843
100
Vibrio coralliilyticus
S2719
S3728
S0386
0.05

Figure 3 – Phylogenetic tree of 191 Galathea 3 *Vibrionaceae* strains. The tree is based on the complete *fur* gene sequences analysis and was constructed by the neighbor-joining method. S0386 was used as the outlier, since it classified as a *Photobacterium* sp.. The nodes with bootstrap support of 70 or more are indicated (1000 replications).

From the phylogenetic analysis it is also possible to evaluate the relatedness of some of the strains identified as "other species" (Fig. 3). Some of these will most likely be the same species (*e.g.* S1348, S1349 and S1350, or S2320, 2321 and 2322).

Here we identified at the species level 90% of 191 *Vibrionaceae* strains from the Galathea 3 culture collection. The most abundant species was *Vibrio splendidus*, followed by *V. tasmaniensis*, *V. crassostreae* and *V. campbellii* (Fig. 2). The collection was based on culturing from marine samples and subsequent testing of antibacterial activity against the fish pathogen *V. anguillarum* [18], and this may explain the over representation (70%) of the species belonging to the *Splendidus* clade in this collection [12]. Several studies have shown the presence of *V. splendidus* in water samples through all seasons, with higher predominance in summer, and in locations ranging from the artic to the tropics [13,28]. These features reflect the great adaptability of this species [13], which can also explain its over representation in a global culture collection such as the Galathea 3 collection.

Nevertheless, this dataset allowed the evaluation of the potential of this identification method and web tool in the classification of *Vibrionaceae* environmental samples at the species level, by means of

a single PCR reaction and sequencing of its product. The method presented here can be used for identification of *Vibrionaceae* species in less than 24 hours, and it requires minimal equipment and service. The analyses of the 191 sequences using the FurIOS web service took only three minutes, although this time may depend on the server usage at the moment of use.

The recent findings have provided the *Vibrionaceae* taxonomy field with new and more accurate approaches to evaluate the phylogeny and taxonomy relatedness between the different species of this family. This comes as a great opportunity to revaluate the evolutionary theories in this family as well to discover and correctly classify newly identified species.

Conclusions

FurIOS is an easy-to-use online service that allows the identification of bacteria from the *Vibrionaceae* family at the species level using the *fur* gene as a single phylogenetic marker. This online-service available to any user worldwide is an extra effort to implement the use of the recently developed method for the amplification of the *fur* gene sequence, with greater discriminatory power when compared to MLSA or 16S rRNA analyses. Its simplistic design and straightforward pipeline makes it suitable for any research environment, from academia to industry, and especially for quick species identification in clinical and food-safety settings, where *Vibrionaceae* strains are of high risk (*e.g.* microbiological control of sea food products).

Availability and Requirements

Project name: FurIOS

Project home page: http://www.cbs.dtu.dk/services/furIOS-1.0/

Operating system(s): Platform independent

Programming language: Python

License: Apache

Any restrictions to use by non-academics: No

Abbreviations

fur, ferric-uptake regulator; MLSA, Multi-Locus Sequencing Analysis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HM designed the study and the database used, and together with SG performed the testing of the web-based tool. JC carried out the programming and together with KR the implementation of the web-based tool. All authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

HM and SG were supported by a PhD grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7-People-2012-ITN, under grant agreement No. 317058, "BACTORY". JC acknowledges the support by the Novo Nordisk Foundation through The Novo Nordisk Foundation Center for Biosustainability.

References

1. Amaral, G. R. S. *et al.* Genotype to phenotype: identification of diagnostic vibrio phenotypes using whole genome sequences. *Int. J. Syst. Evol. Microbiol.* **64,** 357–65 (2014).

2. Roux, F. Le *et al.* The emergence of *Vibrio* pathogens in Europe: ecology, evolution, and pathogenesis (Paris, 11–12th March 2015). *Front. Microbiol.* **6,** 1–8 (2015).

3. Xu, F. *et al.* Genetic characterization of clinical and environmental *Vibrio parahaemolyticus* from the Northeast USA reveals emerging resident and non-indigenous pathogen lineages. *Front. Microbiol.* **6,** 272 (2015).

4. Huehn, S. *et al.* Pathogenic vibrios in environmental, seafood and clinical sources in Germany. *International Journal of Medical Microbiology* **304,** 843–850 (2014).

5. Reid, P. C., Gorick, G. & Edwards, M. *Climate Change and Marine Ecosystem Research*. 53 (Sir Alister Hardy Foundation for Ocean Science (SAHFOS): Plymouth, 2011).

6. Ottaviani, D. *et al.* Nontoxigenic *Vibrio parahaemolyticus* strains causing acute gastroenteritis. *J. Clin. Microbiol.* **50,** 4141–4143 (2012).

7. Vezzulli, L. *et al.* Long-term effects of ocean warming on the prokaryotic community: evidence from the vibrios. *ISME J.* **6,** 21–30 (2012).

8. Nair, G. B. *et al.* Global dissemination of *Vibrio parahaemolyticus* serotype O3:K6 and its serovariants. *Clin. Microbiol. Rev.* **20,** 39–48 (2007).

9. Thompson, F. L. *et al.* Phylogeny and Molecular Identification of Vibrios on the Basis of Multilocus Sequence Analysis Phylogeny and Molecular Identification of Vibrios on the Basis of Multilocus Sequence Analysis. *Appl. Environ. Microbiol.* **71,** 5107–5115 (2005).

10. Gabriel, M. W., Matsui, G. Y., Friedman, R. & Lovell, C. R. Optimization of multilocus sequence analysis for identification of species in the genus *Vibrio*. *Appl. Environ. Microbiol.* **80,** 5359–5365 (2014).

11. Sawabe, T., Kita-Tsukamoto, K. & Thompson, F. L. Inferring the evolutionary history of vibrios by means of multilocus sequence analysis. *J. Bacteriol.* **189,** 7932–6 (2007).

12. Sawabe, T. *et al.* Updating the *Vibrio* clades defined by multilocus sequence phylogeny: proposal of eight new clades, and the description of Vibrio tritonius sp. nov. *Front. Microbiol.* **4,** 414 (2013).

13. Jensen, S., Frost, P. & Torsvik, V. L. The nonrandom microheterogeneity of 16S rRNA genes in *Vibrio splendidus* may reflect adaptation to versatile lifestyles. *FEMS Microbiol. Lett.* **294,** 207–15 (2009).

14. Machado, H. & Gram, L. The *fur* Gene as a New Phylogenetic Marker for *Vibrionaceae* Species Identification. *Appl. Environ. Microbiol.* **81,** 2745–2752 (2015).

15. Reen, F. J., Almagro-Moreno, S., Ussery, D. & Boyd, E. F. The genomic code: inferring *Vibrionaceae* niche specialization. *Nat. Rev. Microbiol.* **4,** 697–704 (2006).

16. Tatusova, T. a & Madden, T. L. BLAST 2 SEQUENCES, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174,** 247–250 (1999).

17. Cock, P. J. a *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).

18. Gram, L., Melchiorsen, J. & Bruhn, J. B. Antibacterial activity of marine culturable bacteria collected from a global sampling of ocean surface waters and surface swabs of marine organisms. *Mar. Biotechnol. (NY).* **12,** 439–51 (2010).

19. Auch, A. F., von Jan, M., Klenk, H.-P. & Göker, M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* **2,** 117–134 (2010).

20. Auch, A. F., Klenk, H.-P. & Göker, M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand. Genomic Sci.* **2,** 142–148 (2010).

21. Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P. & Göker, M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14,** 60 (2013).

22. Giubergia, S., Machado, H., Matei, R. V. & Gram, L. *Vibrio galatheae* sp. nov., a member of the family *Vibrionaceae*

isolated from a mussel. *Int. J. Syst. Evol. Microbiol.* **66,** 347–352 (2016).

23. Thompson, C. C., Thompson, F. L., Vicente, A. C. P. & Swings, J. Phylogenetic analysis of vibrios and related species by means of *atpA* gene sequences. *Int. J. Syst. Evol. Microbiol.* **57,** 2480–4 (2007).

24. Lin, B. *et al.* Comparative genomic analyses identify the *Vibrio harveyi* genome sequenced strains BAA-1116 and HY01 as *Vibrio campbellii*. *Environ. Microbiol. Rep.* **2,** 81–89 (2010).

25. Gomez-Gil, B. *Vibrio hispanicus* sp. nov., isolated from Artemia sp. and sea water in Spain. *Int. J. Syst. Evol. Microbiol.* **54,** 261–265 (2004).

26. Pascual, J., Macián, M. C., Arahal, D. R., Garay, E. & Pujalte, M. J. Multilocus sequence analysis of the central clade of the genus *Vibrio* by using the 16S rRNA, *recA*, *pyrH*, *rpoD*, *gyrB*, *rctB* and *toxR* genes. *Int. J. Syst. Evol. Microbiol.* **60,** 154–65 (2010).

27. Thompson, F. & Hoste, B. Genomic Diversity Amongst *Vibrio* Isolates from Different Sources Determined by Fluorescent Amplified Fragment Length Polymorphism. *Syst. Appl. Microbiol.* **538,** 520–538 (2001).

28. Thompson, J. R. *et al.* Diversity and dynamics of a north atlantic coastal *Vibrio* community. *Appl Env. Microbiol* **70,** 4103–4110 (2004).

Supplementary Table 1

Table S1 – Results of the identification of Vibrionaceae strains from the Galathea 3 collection using FurIOS1.0.

Strain	Name	E-value	% Identity
S0203	Vibrio splendidus 1F-157	0,00E+00	98
S0278	Vibrio splendidus ZS-139	0,00E+00	99
S0344	Vibrio pacinii DSM 19139	0,00E+00	100
S0387	Vibrio splendidus 1F-157	0,00E+00	100
S0537	Vibrio splendidus ZS-139	0,00E+00	99
S0538	Vibrio splendidus ZS-139	0,00E+00	86
S0539	Vibrio splendidus ZS-139	0,00E+00	100
S0553	Vibrio splendidus ZS-139	0,00E+00	99
S0562	Vibrio splendidus ATCC 33789	0,00E+00	97
S0567	Vibrio splendidus ATCC 33789	0,00E+00	97
S0571	Vibrio splendidus 1F-157	0,00E+00	100
S0573	Vibrio kanaloae 5S-149	0,00E+00	100
S0575	Vibrio splendidus ATCC 33789	0,00E+00	97
S0576	Vibrio splendidus ATCC 33789	0,00E+00	97
S0648	Vibrio splendidus 1F-157	0,00E+00	99
S0649	Vibrio splendidus FF-6	0,00E+00	100
S0679	Vibrio splendidus FF-6	0,00E+00	100
S0681	Vibrio splendidus FF-6	0,00E+00	100
S0682	Vibrio splendidus 1F-157	0,00E+00	100
S0693	Vibrio kanaloae 5S-149	0,00E+00	100
S0703	Vibrio splendidus FF-6	0,00E+00	100
S0787	Vibrio shilonii AK1	0,00E+00	99
S0792	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	99
S0838	Vibrio fortis Dalian14	0,00E+00	97
S0843	Vibrio brasiliensis LMG 20546	0,00E+00	93
S0845	Vibrio brasiliensis LMG 20546	0,00E+00	93
S1063	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	99
S1081	Vibrio splendidus ATCC 33789	0,00E+00	98
S1084	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	98
S1085	Vibrio splendidus ATCC 33789	0,00E+00	99
S1099	Vibrio harveyi VHJR7	0,00E+00	96
S1100	Vibrio crassostreae 9CS106	0,00E+00	97
S1101	Vibrio crassostreae 9CS106	0,00E+00	97

S1104	Vibrio splendidus ATCC 33789	0,00E+00	99
S1108	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	99
S1131	Vibrio splendidus ATCC 33789	0,00E+00	99
S1132	Vibrio splendidus ATCC 33789	0,00E+00	99
S1137	Vibrio splendidus ATCC 33789	0,00E+00	99
S1142	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	99
S1160	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	99
S1171	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	98
S1341	Vibrio splendidus ATCC 33789	0,00E+00	99
S1344	Vibrio splendidus ATCC 33789	0,00E+00	98
S1349	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	91
S1350	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	91
S1362	Vibrio splendidus ATCC 33789	0,00E+00	97
S1364	Vibrio splendidus ATCC 33789	0,00E+00	97
S1365	Vibrio splendidus ATCC 33789	0,00E+00	97
S1371	Vibrio splendidus ATCC 33789	0,00E+00	99
S1383	Vibrio fortis Dalian14	0,00E+00	96
S1394	Vibrio splendidus ATCC 33789	0,00E+00	99
S1399	Vibrio tubiashii DSM 19142	0,00E+00	99
S1604	Vibrio tasmaniensis ZS-17	0,00E+00	99
S1605	Vibrio tasmaniensis ZS-17	0,00E+00	99
S1606	Vibrio splendidus 1F-157	0,00E+00	99
S1651	Vibrio splendidus 1F-157	0,00E+00	98
S1652	Vibrio splendidus 1F-157	0,00E+00	99
S1653	Vibrio splendidus ATCC 33789	0,00E+00	97
S2043	Vibrio coralliilyticus S2052	0,00E+00	100
S2054	Vibrio coralliilyticus S2052	0,00E+00	100
S2148	Vibrio splendidus ATCC 33789	0,00E+00	97
S2149	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	98
S2155	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	98
S2160	Vibrio splendidus ATCC 33789	0,00E+00	97
S2227	Vibrio splendidus ATCC 33789	0,00E+00	93
S2228	Vibrio splendidus ATCC 33789	0,00E+00	97
S2229	Vibrio splendidus ATCC 33789	0,00E+00	97
S2234	Vibrio splendidus ATCC 33789	0,00E+00	98
S2291	Vibrio splendidus ATCC 33789	0,00E+00	93
S2321	Vibrio caribbeanicus ATCC BAA-2122	7,74E-179	90
S2345	Vibrio splendidus 1F-157	0,00E+00	99

S2349	Vibrio splendidus ATCC 33789	0,00E+00	99
S2351	Vibrio splendidus ATCC 33789	0,00E+00	99
S2352	Vibrio splendidus ATCC 33789	0,00E+00	92
S2396	Vibrio splendidus ATCC 33789	0,00E+00	92
S2538	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	99
S2674	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	99
S2681	Vibrio alginolyticus TS13	0,00E+00	97
S2686	Vibrio alginolyticus TS13	0,00E+00	97
S2688	Vibrio alginolyticus TS13	0,00E+00	97
S2689	Vibrio alginolyticus TS13	0,00E+00	97
S2690	Vibrio alginolyticus TS13	0,00E+00	97
S2719	Vibrio brasiliensis LMG 20546	2,09E-154	86
S2723	Vibrio alginolyticus TS13	0,00E+00	97
S2725	Vibrio alginolyticus TS13	0,00E+00	97
S2757	Vibrio tubiashii DSM 19142	3,10E-171	89
S2850	Vibrio splendidus ATCC 33789	0,00E+00	92
S2895	Vibrio tasmaniensis ZS-17	0,00E+00	99
S2946	Vibrio crassostreae 9CS106	0,00E+00	99
S2947	Vibrio splendidus FF-500	0,00E+00	100
S2948	Vibrio splendidus FF-500	0,00E+00	100
S2949	Vibrio tasmaniensis ZS-17	0,00E+00	98
S2951	Vibrio tasmaniensis 1F-187	0,00E+00	99
S3026	Vibrio tasmaniensis ZS-17	0,00E+00	98
S3027	Vibrio crassostreae 9CS106	0,00E+00	99
S3142	Vibrio splendidus FF-500	0,00E+00	100
S3143	Vibrio splendidus FF-500	0,00E+00	100
S3144	Vibrio splendidus FF-500	0,00E+00	100
S3145	Vibrio splendidus ZF-90	0,00E+00	100
S3246	Vibrio tasmaniensis ZS-17	0,00E+00	99
S3247	Vibrio tasmaniensis ZS-17	0,00E+00	99
S3248	Vibrio splendidus FF-500	0,00E+00	100
S3249	Vibrio splendidus FF-500	0,00E+00	100
S3250	Vibrio tasmaniensis ZS-17	0,00E+00	100
S3251	Vibrio tasmaniensis ZS-17	0,00E+00	98
S3252	Vibrio tasmaniensis ZS-17	0,00E+00	98
S3253	Vibrio tasmaniensis ZS-17	0,00E+00	98
S3254	Vibrio splendidus FF-500	0,00E+00	100
S3256	Vibrio tasmaniensis ZS-17	0,00E+00	100

S3259	Vibrio tasmaniensis ZS-17	0,00E+00	100
S3261	Vibrio tasmaniensis ZS-17	0,00E+00	98
S3262	Vibrio tasmaniensis ZS-17	0,00E+00	99
S3301	Vibrio tasmaniensis ZS-17	0,00E+00	98
S3302	Vibrio splendidus ZS-139	0,00E+00	99
S3303	Vibrio splendidus FF-500	0,00E+00	99
S3305	Vibrio splendidus FF-500	0,00E+00	99
S3307	Vibrio splendidus FF-500	0,00E+00	100
S3308	Vibrio splendidus ZF-90	0,00E+00	99
S3309	Vibrio splendidus ZF-90	0,00E+00	99
S3310	Vibrio splendidus FF-500	0,00E+00	100
S3311	Vibrio splendidus 1F-157	0,00E+00	99
S3312	Vibrio splendidus 1F-157	0,00E+00	100
S3314	Vibrio splendidus FF-500	0,00E+00	99
S3315	Vibrio splendidus 1F-157	0,00E+00	99
S3316	Vibrio splendidus 1F-157	0,00E+00	99
S3320	Vibrio splendidus 1F-157	0,00E+00	99
S3434	Vibrio splendidus FF-500	0,00E+00	99
S3435	Vibrio splendidus 1F-157	0,00E+00	99
S3436	Vibrio splendidus FF-500	0,00E+00	99
S3494	Vibrio tasmaniensis ZS-17	0,00E+00	99
S3495	Vibrio tasmaniensis ZS-17	0,00E+00	99
S3658	Vibrio splendidus ATCC 33789	0,00E+00	97
S3659	Vibrio tubiashii DSM 19142	0,00E+00	98
S3660	Vibrio splendidus ATCC 33789	0,00E+00	97
S3667	Vibrio splendidus ATCC 33789	0,00E+00	97
S3669	Vibrio fortis Dalian14	0,00E+00	98
S3706	Vibrio splendidus ATCC 33789	0,00E+00	98
S3707	Vibrio splendidus ATCC 33789	0,00E+00	97
S3728	Vibrio cholerae TM 11079-80	2,53E-134	82
S3729	Vibrio cyclitrophicus FF75	0,00E+00	99
S3778	Vibrio splendidus ATCC 33789	0,00E+00	92
S3781	Vibrio splendidus ATCC 33789	0,00E+00	99
S3783	Vibrio cyclitrophicus FF75	0,00E+00	99
S3789	Vibrio splendidus ATCC 33789	0,00E+00	99
S3791	Vibrio splendidus ATCC 33789	0,00E+00	97
S3792	Vibrio splendidus ATCC 33789	0,00E+00	97
S3793	Vibrio splendidus ATCC 33789	0,00E+00	99

S3794	Vibrio splendidus ATCC 33789	0,00E+00	99
S3854	Vibrio alginolyticus TS13	0,00E+00	97
S3855	Vibrio splendidus ATCC 33789	0,00E+00	97
S3857	Vibrio rotiferianus DAT722	0,00E+00	99
S3858	Vibrio crassostreae 9CS106	0,00E+00	96
S3859	Vibrio fortis Dalian14	0,00E+00	97
S3860	Vibrio alginolyticus TS13	0,00E+00	97
S3861	Vibrio crassostreae 9CS106	0,00E+00	96
S3862	Vibrio crassostreae 9CS106	0,00E+00	97
S3863	Vibrio crassostreae 9CS106	0,00E+00	96
S3864	Vibrio crassostreae 9CS106	0,00E+00	96
S3897	Vibrio crassostreae 9CS106	0,00E+00	96
S3900	Vibrio crassostreae 9CS106	0,00E+00	96
S3925	Vibrio splendidus ATCC 33789	0,00E+00	99
S3927	Vibrio splendidus 12E03	0,00E+00	93
S3941	Vibrio crassostreae 9CS106	0,00E+00	99
S3943	Vibrio crassostreae 9CS106	0,00E+00	96
S4075	Vibrio splendidus ATCC 33789	0,00E+00	98
S4078	Vibrio harveyi VHJR7	0,00E+00	98
S4228	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	99
S4235	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	99
S4309	Vibrio alginolyticus TS13	0,00E+00	97
S4497	Vibrio rotiferianus DAT722	0,00E+00	99
S4635	Vibrio tasmaniensis ZS-17	0,00E+00	99
S4636	Vibrio tasmaniensis ZS-17	0,00E+00	99
S4638	Vibrio splendidus FF-500	0,00E+00	100
S4639	Vibrio kanaloae 5S-149	0,00E+00	99
S4733	Vibrio tasmaniensis 5F-79	0,00E+00	96
S4734	Vibrio tasmaniensis 5F-79	0,00E+00	100
S4735	Vibrio splendidus FF-500	0,00E+00	100
S4736	Vibrio splendidus ZF-90	0,00E+00	100
S4737	Vibrio splendidus FF-500	0,00E+00	100
S4739	Vibrio splendidus 1S-124	0,00E+00	100
S4740	Vibrio splendidus FF-500	0,00E+00	99
S0386	Photobacterium gaetbulicola Gung47	1,31E-131	82
S3662	Vibrio splendidus 1F-157	0,00E+00	99
S0200	Vibrio splendidus 1F-157	0,00E+00	95
S0568	Vibrio splendidus ATCC 33789	0,00E+00	97

S0788	Vibrio coralliilyticus S2052	0,00E+00	99
S1348	Vibrio campbellii CAIM 519 = NBRC 15631	0,00E+00	91
S1396	Vibrio coralliilyticus S2052	0,00E+00	99
S2056	Vibrio coralliilyticus S2052	0,00E+00	100
S2320	Vibrio caribbeanicus ATCC BAA-2122	7,74E-179	90
S2322	Vibrio caribbeanicus ATCC BAA-2122	7,74E-179	90

Paper 5

Marine bacterial cythochromes P450 and their potential in biotechnology

Marine bacterial cytochromes P450 and their potential in biotechnology

Ana Mafalda Cavaleiro[¶,1], Henrique Machado[¶,1,2], Carlos Casado[3], Stefan Kol[1], Morten H. H. Nørholm[1,*], Lone Gram[2]

[1]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle Allè 6, DK-2970 Hørsholm, Denmark

[2]Department of Systems Biology, Technical University of Denmark, Matematiktorvet bldg 301, DK-2800 Kgs Lyngby, Denmark

[3] Evolva Biotech A/S, DK-2200 Copenhagen, Denmark

[¶] **These authors contributed equally to this work**

[*] **Corresponding author**

E-mail: morno@biosustain.dtu.dk.

ABSTRACT

Cytochromes P450 (CYPs) are enzymes that modify several types of compounds, usually by substrate mono-oxygenation, and are of biotechnological interest due to their ability to perform challenging chemistry. In comparison with the membrane bound eukaryotic CYPs, the soluble bacterial CYPs are more stable and easier to express and purify. However, their use as biocatalysts in the biotechnological industry has not been fully explored. Marine bacteria are emerging as a yet unexplored source of natural products, many of which are dependent on CYP-modifications. We mined 19 genomes of bioactive marine bacteria and identified 26 distinct CYP open-reading frames (ORFs). These ORFs were compared to previously studied bacterial CYPs to identify their phylogenetic relationships and putative functions. Five CYPs were analyzed in more detail and expressed in the model cell factories *Escherichia coli* and *Saccharomyces cerevisiae*. Expression in yeast confirmed cytoplasmic localization of the bacterial CYPs. Furthermore, we showed that codon optimized bacterial CYPs were properly folded in yeast. To our knowledge, the present study represents the first successful mining for CYPs from marine bacterial genomes in a merger of white and blue biotechnology.

Keywords Cytochrome P450 (CYP), Marine bacterial CYPs, Biotechnological potential, Yeast expression

INTRODUCTION

Cytochromes P450 (CYPs) are a superfamily of enzymes found in many different taxonomic groups. These heme-thiolate proteins bind to carbon monoxide exhibiting an absorption spectrum at 450 nm, which is the phenomenon responsible for their designation. They use electrons from NAD(P)H to catalyze scission of molecular oxygen and an associated protein, the redox partner, to transfer the reducing equivalents to the heme prosthetic group. CYPs modify several substrates by mono-oxygenation and this is the most common function of these enzymes, although other functions have been described [1,2,3].

Several natural compounds are synthesized in microorganisms, plants and fungi through a plethora of chemical modifications catalyzed by CYPs [3]. Natural product synthesis includes several unique chemical reactions, some of which are performed by the associated CYPs. Therefore, understanding the mechanisms of action of CYP enzymes, associated with natural product synthetic clusters, will allow for catalytic innovation and can be of great use in biotechnological processes [3].

In Nature, eukaryotic CYPs are often part of membrane-associated multi protein complexes and their use as synthetic catalysts at an industrial scale is hampered by this complexity as well as their limited stability and activity when isolated [4,5]. Additionally, the requirements for a constant supply of NAD(P)H and a redox partner are challenges in establishing CYP-based biocatalysis. Despite these obstacles, CYPs have been successfully used in industrial settings. For example, in yeast, the co-expression of a CYP, its redox partner and cytochrome b5 from plants led to high yield production of the antimalarial drug artemisinin [6,7].

During the last decades, CYPs have received increased attention. Of the 905 total identified CYPs in 2009, 30% were of plant origin, but these only represented 11% of total family diversity [8]. Remarkably, bacterial CYPs accounted for only 6% of the total number of

CYPs, but 18% of family diversity [8]. By August 2013, approximately 20,000 CYP enzymes had been identified and catalogued (http://drnelson.uthsc.edu/CytochromeP450.html).

In contrast to eukaryotic CYPs, the bacterial CYPs offer several advantages. So far, all identified bacterial CYPs are soluble (in contrast to membrane-bound) enzymes that are faster and more stable catalysts than their eukaryotic counterparts [4]. The bacterial CYPs are usually associated with ferredoxins and ferredoxin reductases, which are used as redox partners [9]. One of the most extensively studied bacterial CYPs is CYP102A1 from *Bacillus megaterium*, commonly known as BM3 [10]. BM3 is a self-sufficient enzyme, as it is fused with a FAD/FMN reductase (differs from ferredoxin reductases since it has a flavin moiety instead of iron-sulfur core), and it uses C12-C20 saturated or unsaturated fatty acids as natural substrates [11]. For several years, efforts to achieve immobilization, cofactor regeneration and development of assays for CYPs have been focused on BM3 [12]. However, the major achievements with BM3 are mostly related to protein engineering towards, in particular, hydroxylation of BM3-unnatural substrates [5,13,14,15,16,17,18]. Additional efforts have been dedicated to engineering other prokaryotic CYPs into enzymes with the capability to modify a broad range of substrates aiming to exploit their unique ability to perform challenging chemistry in biotechnological processes [19].

Soil microorganisms have been successfully explored as sources of bioactive compounds used by the pharmaceutical and biotech industries [20], but the use of marine bacteria for similar purposes is on the rise [21,22,23]. Marine bacteria are believed to harbor a multitude of novel bioactive compounds due to the unique environmental conditions they have adapted to, such as high pressure, high salinity, high or low temperature or oligotrophic conditions [24,25]. Some of these bioactive compounds are produced by biosynthetic reactions catalyzed by CYPs, which are typically encoded in biosynthetic gene clusters [26]. However, there are very few studies on marine CYPs and bioprospecting marine bacteria would likely provide

novel CYPs. In the present study we have mined the genome sequences of several bioactive marine bacteria for CYP-encoding genes. Selected candidates were cloned and expressed in both *E. coli* and *S. cerevisiae* to explore their cellular localization and biotechnological potential in these model microbial cell factories. This study also explores the heterogeneity of CYP enzymes in marine bacteria and the process of discovering new CYP enzymatic functions.

MATERIALS AND METHODS

Strains and media

All strains used for cloning and expression are listed in Table 1. *Saccharomyces cerevisiae* strain PAP1503 was obtained from Per Amstrup Pedersen (Copenhagen University, Denmark). Yeast strains were grown either in synthetic complete (SC) drop-out media (0.67% Yeast Nitrogen Base with required amino acids from Sigma-Aldrich, St. Louis, MO, USA)) or standard yeast peptone dextrose (YPD) medium (1% yeast extract, 2% peptone, 2% dextrose from Sigma-Aldrich, St. Louis, MO, USA) for both liquid cultures and agar plates. Yeast transformants were selected on SC media without the appropriate selection requirements corresponding to the plasmid-encoded auxotrophic markers. *Escherichia coli* strain NEB5α (New England Biolabs, Ipswich, MA, USA) was used for cloning and propagation of plasmids. *E. coli* strain BL21 (DE3) (Novagen, Madison, WI, USA) was used for radioactive labeling studies. Chemically competent cells of NEB5α and BL21 (DE3) were prepared as described elsewhere [27]. Competency of the cells was 2.6-3.3 × 10^7 CFU/μg DNA. Bacteria were propagated on Luria-Bertani (LB) agar plates (Oxoid, Altrincham, UK) or liquid 2xYT media (1.6% tryptone, 1% yeast extract from Sigma-Aldrich, St. Louis, MO, USA) supplemented with spectinomycin (50 μg/mL) or ampicillin (100 μg/mL) when required and Phosphate, Ammonium and Selenomethionine (PASM) media for ^{35}S-methionine labeling of proteins [28]. ^{35}S-methionine was purchased from Perkin Elmer (Waltham, Massachusetts, USA). Plasmids were isolated using the NucleoSpin® Plasmid QuickPure Kit (Macherey-Nagel, Durën, Germany).

Table 1 – Strains and plasmids used in this study

Strains/Plasmids	Property	Source
Saccharomyces cerevisiae		
PAP1503	*MATα ura3-52 trp1::GAL10-GAL4 lys2-801 leu2Δ1 his3Δ200 pep4::HIS3 prb1Δ1.6R can1 GAL*	[29,30]
PAP1503_P4	PAP1503 transformed with plasmid pEMBLyex4_P4	This study
PAP1503_P5	PAP1503 transformed with plasmid pEMBLyex4_P5	This study
PAP1503_P7	PAP1503 transformed with plasmid pEMBLyex4_P7	This study
PAP1503_P9	PAP1503 transformed with plasmid pEMBLyex4_P9	This study
PAP1503_P29	PAP1503 transformed with plasmid pEMBLyex4_P29	This study
PAP1503_BM3	PAP1503 transformed with plasmid pEMBLyex4_BM3	This study
PAP1503_CYP79A1	PAP1503 transformed with plasmid pEMBLyex4_CYP79A1	This study
Escherichia coli		
NEB5α	*fhuA2* Δ*(argF-lacZ)*U169 *phoA glnV44 Φ80Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*	New England Biolabs
NEB5α_P5	NEB5α transformed with plasmid pCDF_P5	This study
NEB5α_P7	NEB5α transformed with plasmid pCDF_P7	This study
NEB5α_P9	NEB5α transformed with plasmid pCDF_P9	This study
NEB5α_P29	NEB5α transformed with plasmid pCDF_P29	This study
BL21 (DE3)	F^- *ompT gal dcm lon hsdS*$_B$(r_B^- m_B^-) λ(DE3 *[lacI lac*UV5-T7 gene *1 ind1 sam7 nin5])*	Novagen
BL21 (DE3)_P5	BL21 (DE3) transformed with plasmid pCDF_P5	This study
BL21 (DE3)_P7	BL21 (DE3) transformed with plasmid pCDF_P7	This study
BL21 (DE3)_P9	BL21 (DE3) transformed with plasmid pCDF_P9	This study
BL21 (DE3)_P29	BL21 (DE3) transformed with plasmid pCDF_P29	This study
Plasmids		
pEMBLyex4	High copy-number yeast expression vector controlled by the galactose-inducible *GAL1-CYC1* promoter, 2μ origin of replication, *URA3*, *leu2-d*	[31]
pEMBLyex4_P4	*P4* ORF expressed from $P_{GAL1-CYC1}$ promoter in pEMBLyex4	This study
pEMBLyex4_P5	*P5* ORF expressed from $P_{GAL1-CYC1}$ promoter in pEMBLyex4	This study
pEMBLyex4_P7	*P7* ORF expressed from $P_{GAL1-CYC1}$ promoter in pEMBLyex4	This study
pEMBLyex4_P9	*P9* ORF expressed from $P_{GAL1-CYC1}$ promoter in pEMBLyex4	This study
pEMBLyex4_P429	*P29* ORF expressed from $P_{GAL1-CYC1}$ promoter in pEMBLyex4	This study
pEMBLyex4_BM3	*BM3* ORF expressed from $P_{GAL1-CYC1}$ promoter in pEMBLyex4	This study
pEMBLyex4_CYP79A1	*CYP79A1* ORF expressed from $P_{GAL1-CYC1}$ promoter in pEMBLyex4	This study
pCDFDuet-1	Cloning and expression vector, Sp^R	Novagen
pCDF_P5	*P5* ORF expressed from P_{T7} promoter in pCDFDuet-1	This study
pCDF_P7	*P7* ORF expressed from P_{T7} promoter in pCDFDuet-1	This study
pCDF_P9	*P9* ORF expressed from P_{T7} promoter in pCDFDuet-1	This study
pCDF_P29	*P29* ORF expressed from P_{T7} promoter in pCDFDuet-1	This study

Genomic DNA isolation and sequencing

High purity genomic DNA was extracted by successive phenol:chloroform:isoamyl-alcohol purification steps followed by precipitation with isopropanol, treatment with RNase, and a

final purification and precipitation step [32]. Quality assessment and quantification of genomic DNA was done in 1% agarose gel electrophoresis, NanoDrop Spectrometer (Saveen Werner, Sweden), and Qubit 2.0 Analyzer (Invitrogen, United Kingdom). The genomic DNA was used as a template for PCR amplification of the genes of interest.

Bioinformatic analysis

The strains used in the genome mining (Fig. 1) were isolated during the Danish Galathea 3 global research expedition (http://www.galathea3.dk/uk) [33] and whole genome sequenced [23]. Prediction of cytochrome P450 genes was performed using an annotation and homology-based search after RAST genome annotation [34]. Identification of specific biosynthetic gene clusters was performed using antiSMASH 2.0 [35] and annotation-based analyses [34]. Possible identity of all predicted cytochromes P450 to known proteins was done by Basic Local Alignment Search (BLAST) analysis. Membrane protein topology predictions were made using the online tools Topcons http://topcons.cbr.su.se/ [36] and ΔG predictor http://dgpred.cbr.su.se/ [37]. An attempt to identify the families of the new CYPs was performed using the information in the "cytochrome P450 homepage" [8].

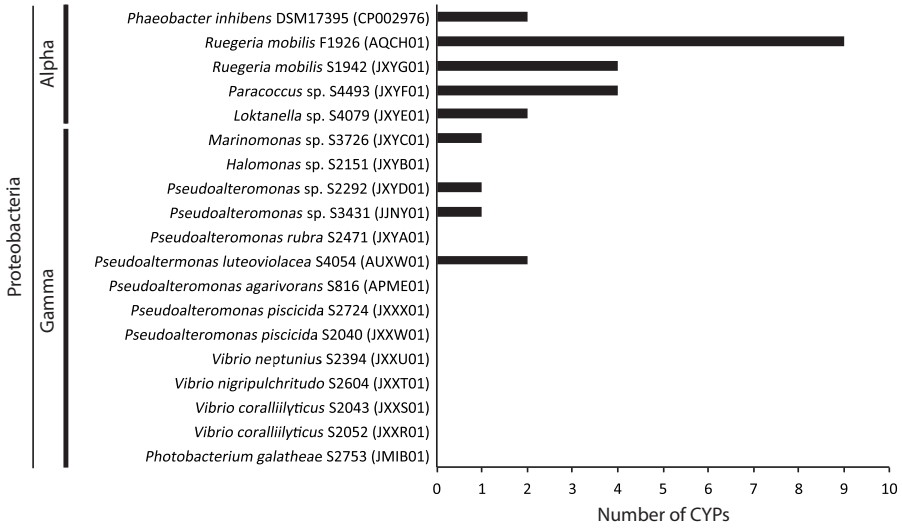


Figure 1. **Overview of the number of identified cytochromes P450 (CYP) in 19 marine bacterial genomes.**

Both members of alpha and gamma-proteobacteria were mined for putative CYPs. Strain names and numbers are referenced in the vertical axis with the NCBI whole genome sequence accession number in brackets. Black bars represent the number of CYPs identified per genome.

PCR, oligonucleotide design and uracil excision cloning

All oligonucleotides used for PCR amplification were purchased from Integrated DNA Technologies (Coralville, IA, USA) and are listed in Table S1. Oligonucleotides were designed with melting temperatures (T_m) of approximately 60°C. The uracil excision-specific oligonucleotides were designed with U-overhangs of T_m varying from 20 to 30°C. Oligonucleotides used for amplification of parts for homologous recombination-based yeast DNA assembly were designed with homology arms of >27 bp. DNA amplification and assembly in *E. coli* with uracil excision cloning were accomplished as described previously [38,39]. PCR products were purified using a PureLink™ Quick Gel Extraction and PCR Purification Combo Kit (Life Technologies, Foster City, USA). PCR products were quantified using a NanoDrop 8000 (Thermo Scientific, Wilmington, USA).

Plasmids and strains construction

All plasmids used for CYP-ORF expression are listed in Table 1. For testing CYP expression in *E. coli*, PCR products containing CYP-predicted ORF and U-overhangs complementary to the entry vector were cloned in a pCDFDuet-1 plasmid via two-fragment uracil excision cloning. BL21 (DE3) was transformed and assayed as described below (see section ^{35}S-methionine labeling of proteins). In order to avoid negative effects of rare codons in *S. cerevisiae*, the genes were codon optimized using the codon optimization tool from Integrated DNA Technologies and then synthetized as gBlocks™ gene fragments (Integrated DNA Technologies, Coralville, IA, USA). The codon optimized sequences of P4, P5, P7, P9 and

P29 have been deposited on GenBank with accession numbers KU710259, KU710260, KU710261, KU710262, and KU710263, respectively. These codon optimized gene fragments were used as PCR templates for assembly in yeast. Three PCR products with long homology arms encoding for protein of interest (CYP), GFP (C-terminally fused to a His8 tag) and pEMBLyex4 plasmid previously digested with *Bam*HI and *Hin*dIII were transformed directly into PAP1503 and assembled in yeast by homologous recombination [40]. The resulting construct encodes for the protein of interest (CYP) C-terminally fused to GFP with a TEV protease cleavage site in between. The newly assembled plasmids were extracted from yeast and transformed into NEB5α for plasmid amplification. All constructs were confirmed by colony PCR and sequencing.

^{35}S-methionine labeling of proteins

Transcription and translation of cytochrome P450 enzymes in *E. coli* was confirmed using the rifampicin blocking technique and ^{35}S- methionine labeling [41] as described previously [42].

Assay for overexpression and production of GFP-fusions

Overexpression of CYP-GFP fusions in *S. cerevisiae* was performed in strain PAP1503 as described elsewhere [40] with minor modifications: synthetic complete (SC) medium was used instead of synthetic minimal (SD) medium. PAP1503 transformants were grown overnight (30°C, 200 rpm) in SC supplemented with glucose (20 g/L) without uracil. Overnight cultures were used to inoculate SC without uracil and leucine (SC-ura-leu) in a 1:50 ratio. Cultures were grown at 30°C for 48 hours before they were used to inoculate 250 mL of SC-ura-leu medium supplemented with glucose (5 g/L) with an initial optical density (OD) of 0.05. This was done in triplicate. Cultures were grown at 30°C, 200 rpm until OD 1.0 before GFP-fusion production was induced by addition of galactose at a final concentration of 2% (m/v). Cultures were then grown for 24 hours at 20°C with shaking at 200 rpm.

Microscopy

S. cerevisiae cells were grown as described above (see previous section). Fluorescence of CYP-GFP fusions was visualized at 400 x magnification with a DM4000B fluorescence microscope from Leica (Wetziar, Germany). CYP79A1 from *Sorghum bicolor* was used as a membrane protein control [43].

Protein purification and quantification

Cell pellets from 0.75 L of CYP-GFP producing cultures were resuspended in 30 mL IMAC buffer A (50 mM TRIS-HCl pH 7.5, 500 mM NaCl, 20 mM imidazole) supplemented with 1 tablet/50 mL complete, EDTA-free Protease Inhibitor Cocktail (Sigma-Aldrich, St. Louis, MO, USA) and cells were lysed by three passes at 20.000 psi through an Emulsiflex-C5 (Avestin, Ottawa, Canada). After centrifugation for 15 min at 15.000*g*, the supernatant was loaded onto a 5 mL HisTrap FF crude column (GE Healthcare, Buckinghamshire, UK) equilibrated with IMAC buffer A. The columns were washed with 10 column volumes of IMAC buffer A and the proteins of interest eluted with a 10 column volume gradient of 0-100% IMAC buffer B (50 mM TRIS-HCl pH 7.5, 500 mM NaCl, 500 mM Imidazole). Fractions containing the proteins of interest were identified by SDS-PAGE, pooled and concentrated using Amicon Ultra-15 Centrifugal Filters (Millipore, Billerica, Massachusetts, USA). The CYP proteins of interest (P4, P5, P7 and P29) were concentrated using a 30 kDa cut-off filter, while P9 and BM3 were concentrated using a 10 kDa and 50 kDa cut-off filter, respectively. Protein concentrations were measured with a NanoDrop ND-1000 (NanoDrop Technologies, Inc., Wilmington, DE, USA). Samples were kept at 4°C until permanent storage at -80°C.

In-gel fluorescence

In-gel fluorescence of purified GFP-fusions was carried out in a G:Box chemi XT4 (Syngene, Cambridge, UK) using Mini-PROTEAN® TGX™ precast gels from Biorad (Hercules, USA) and with 465 nm excitation and 535 nm emission.

Absorption Spectroscopy

Absorption spectra of CYP-GFP fusions at 450 nm were recorded with a SLM Aminco DW-2000 TM spectrophotometer (Spectronic Instruments, Rochester, NY, USA) as described elsewhere [44]. Purified protein-GFP fusions were used in concentrations between 0.07-1.2 mg/mL in Tris-HCl buffer pH 8.0.

RESULTS

Mining the genomes of marine bacteria for cytochromes P450 and redox partners

The genomes of 19 bioactive marine strains, previously identified as secondary metabolite producers [23,45,46], were mined for cytochrome P450 (CYP) encoding genes and a total of 26 putative CYP genes were identified (Table S2). The number of CYPs ranged from zero to nine per bacterial genome (Fig. 1). The identified CYP domain-containing open reading frames (ORFs) had GC contents between 39 and 73% and gene sizes from 300 to 1,400 bp (Table S2). Members of the alpha-proteobacteria had up to nine CYP-encoding genes, while a maximum of two was detected in a single gamma-proteobacterial strain (Fig. 1). CYP genes found in the gamma-proteobacteria were associated with putative specific pathways or biosynthetic clusters, such as a predicted gamma-butyrolactone biosynthetic cluster, a cluster involved in agar degradation or a non-ribosomal peptide/ polyketide synthase (Table 2).

Table 2 - Short-list of putative CYPs to be further studied and their redox partners

P450	Strain	GenBank Accession	Size (bp)	GC (%)	TM domains	Genomic context	CYP family	Ferredoxins	Ferredoxin Reductases
P4	S3431	KDC54395	1173	41	None	Agar degradation	CYP236A1	4	2
P5	S3726	KJZ11492	1209	39	None	Butyrolactone cluster	CYP107E1	3	1
P7	S4079	KJZ21287	1176	54	None	Exopolymer metabolism	CYP104A2	4	3
P9	S4493	KJZ30079	312	56	None	Cobalt-zinc-cadmium resistance	CYP152B1	2	2
P29	S4054	KKE82924	1341	39	None	NRPS-PKS	CYP197A1	3	3

The CYP redox-partners in bacteria are usually ferredoxins or flavoprotein reductases [4] and in line with this putative ferredoxins and ferredoxin reductases were identified in all the nineteen genomes. For CYPs P4 and P7, genes potentially encoding for redox partners were identified immediately downstream of the predicted CYP (Table 2).

Features of CYPs

We used UniProt (Universal Protein Resource), BLAST (Basic Local Alignment Search Tool) and the Cytochrome P450 homepage [8] to compare all predicted CYP sequences with previously studied CYPs. All the 26 identified CYPs contained cytochrome P450 domains, that were also identified by Pfam analyses [47], regardless of their gene size (data not shown). Although sequence conservation is low within this family of enzymes, their general topography and structural fold are highly conserved, allowing CYP domain identification [1]. Next, we compared the putative CYP protein sequences to known bacterial CYPs using the BLAST server on the Cytochrome P450 homepage (http://blast.uthsc.edu/) [8]. Several of the identified CYPs were conserved within the same species or genus. However, some diversity (with respect to CYP family) was found in CYP sequences across the different genera of studied marine bacteria (Fig. 2). The most similar bacterial CYPs, a recently reported marine CYP (AHA34040) and previously studied CYPs (AAC68886 and ADC79647) are also shown (Fig. 2).

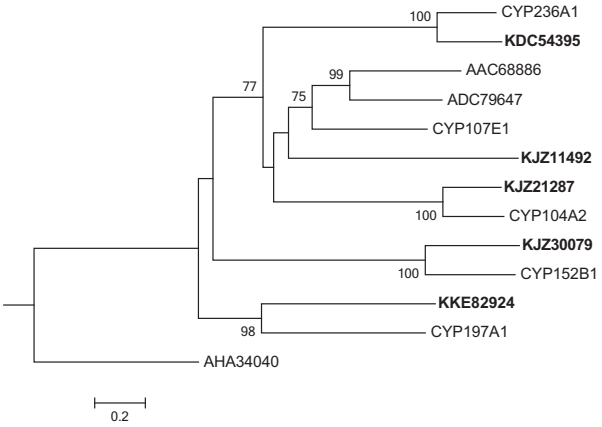


Figure 2. **Phylogenetic tree of the further studied CYPs.**

The five CYPs selected for further studies (highlighted in bold) shown together with the most similar bacterial CYPs, previously studied CYPs (AAC68886 and ADC79647) and a recently studied marine CYP (AHA34040). The nodes with bootstrap support of 70 or more are indicated (1000 replications). GenBank accession numbers are indicated for CYPs, except for the family representatives, where the CYP identification has been used.

Five CYPs were selected for further analyses based on high sequence homology to CYPs belonging to different families with previously identified distinct functions, genome clustering, pathway prediction, and microorganism diversity: one putative CYP from *Pseudoalteromonas* sp., one from *Marinomonas sp*., one from *Loktanella sp*., one CYP from *Paracoccus sp*. and finally, one from *Pseudoalteromonas luteoviolacea* (Table 2). These CYPs not only differed in their primary sequences but were also located in distinct genetic environments, suggesting their participation in different metabolic reactions. The metabolic pathways and specific reactions, including substrate type were predicted using antiSMASH, RAST and BLAST analyses. CYPs P5 and P29 are part of two secondary metabolism clusters, predicted to be involved in gamma-butyrolactone and non-ribosomal peptide - polyketide synthesis, respectively (Table 2). CYPs P4, P7 and P9 are present in genomic

297 regions related to agar degradation, exopolymer metabolism and cobalt-zinc-cadmium
298 resistance, respectively (Table 2).
299
300 **Expression of putative CYPs in model cell factories**
301 To further study the five putative CYPs, all genes were PCR amplified from genomic DNA
302 using the oligonucleotides listed in Table S1 and cloned downstream from a phage T7
303 promoter (P_{T7}) in the pCDFDuet-1 plasmid. Despite several attempts, we were not able to
304 amplify the P4-encoding ORF from the genome of *Pseudoalteromonas* sp. (S3431). Proper
305 transcription and translation in the model bacterium *E. coli* BL21 (DE3) was monitored using
306 ^{35}S-methionine labeling of proteins followed by gel electrophoresis (see materials and
307 methods). BL21 (DE3) with an empty plasmid was used as negative control. P9 was only
308 produced in very low amount, but full-length proteins were observed for all tested CYPs (Fig.
309 S1).
310
311 To elucidate whether these putative CYPs would express in another popular model cell
312 factory, *Saccharomyces cerevisisae*, genes codon optimized for this eukaryotic host were
313 ordered form a commercial source (see materials and methods). Additionally, we included the
314 well studied bacterial CYP, BM3, as a control [10,11]. All five putative marine CYP-
315 encoding codon optimized DNA fragments and the control were fused with DNA encoding
316 GFP and a His8-tag in a pEMBLyex4 plasmid in the *S. cerevisiae* strain PAP1503 by yeast
317 homologous recombination (Fig. 3A). Subsequently the strains were starved for leucine to
318 increase plasmid copy number as described previously [40]. Starved cultures were used as
319 inoculum for higher volume cultures and gene expression was then induced with galactose
320 since CYP-GFP fusion expression is driven by the *CYC1-GAL1* promoter [40]. Expression
321 was confirmed by whole cell fluorescence. All CYP-GFP expressing cells appear green under
322 the microscope suggesting that all fusions are being produced in yeast, despite the different
323 fluorescent intensities observed (Fig. 3B).
324

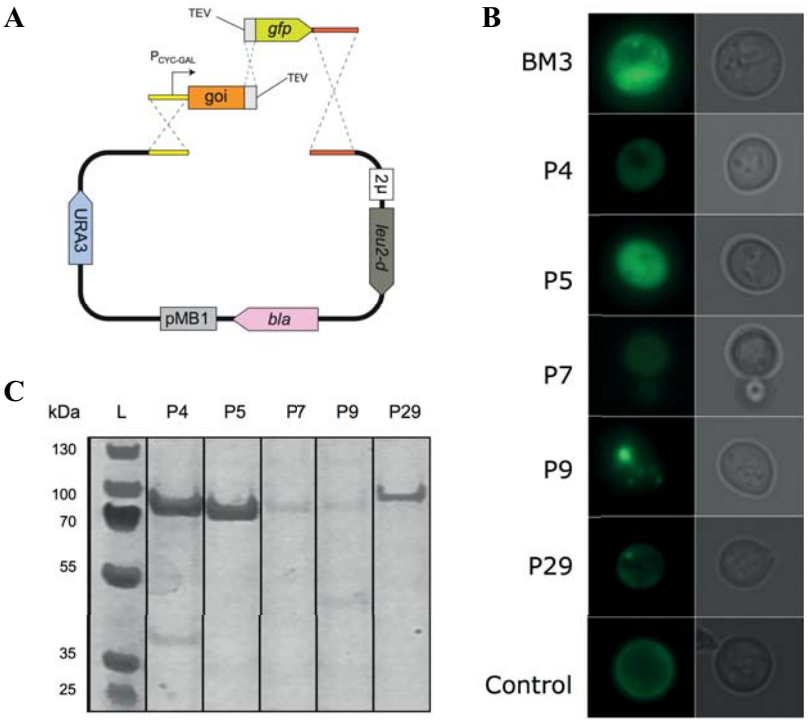


Figure 3. CYPs production in yeast.

(A) Schematic representation of the homologous recombination process that allows the pEMBLyex4 backbone plasmid and the DNA parts encoding for CYPs and GFP to be assembled in yeast. (B) Imaging of live yeast cells, induced with galactose for 24 hours at 20°C, producing the different analyzed CYP-GFP fusions P4, P5, P7, P9 and P29, the BM3 soluble control and the CYP79A1 membrane protein control. (C) SDS-PAGE analysis of purified CYP-GFP fusions P4, P5, P7, P9 and P29. The calculated molecular weights of CYP-GFP fusions are: P4, 71 kDa; P5, 72 kDa; P7, 71 kDa; P9, 38 kDa; P29, 78 kDa.

Cellular localization of bacterial CYPs expressed in *S. cerevisiae*

In contrast to bacterial CYPs, the hydrophobic nature of plant CYPs limits their potential in biotechnological applications. To explore the cellular localization of the marine bacterial CYPs expressed in yeast we monitored their appearance with a fluorescence microscope (Fig. 3B). Topology predictions using the online software Topcons (Topcons

http://topcons.cbr.su.se/ [36] and ΔG predictor http://dgpred.cbr.su.se/ [37] identified no N-terminal transmembrane segments in the CYPs. In line with this, the GFP signal derived from the heterologously expressed bacterial CYPs were homogeneously dispersed in the yeast cytoplasm, as compared to a plant membrane-bound CYP that localized to the periphery of the cells (Fig. 3B). Microscopy also showed that P9 localized in intensely fluorescent spots possibly representing misfolded aggregates (Fig. 3B).

Protein integrity and CO spectra

To evaluate the stability and integrity of the expressed CYP-GFP fusions, all five putative CYPs were purified by affinity using HisTrap columns, and assessed by SDS-PAGE (Fig. 3C) and in-gel fluorescence (Fig. S2). CYP-GFP fusions appeared to be produced as full-length fusions (Fig. 3C) with very little free GFP being detected (Fig. S2). In line with the microscopy data (Fig. 3B), P4 and P5 fusions were produced in the highest yields – the final concentrations being 0.184 and 0.175 mg/mL, respectively. The production levels of P7-GFP and P29-GFP were 0.065 and 0.066 mg/mL, respectively. Surprisingly, the concentration of P9-GFP was high at 0.151 mg/mL but only very little protein was detected when analyzed by SDS-PAGE and in-gel fluorescence (Fig. 3C and Fig. S2).

Proteins belonging to the P450 family exhibit light absorption maxima at 450 nm when correctly incorporated heme is bound to carbon monoxide. Hence, the most widely used method for assaying properly folded and active CYPs involves obtaining CO-binding difference spectra [48]. With this in mind, we used the purified CYP-GFP fusions to perform CO-binding difference spectra using bacterial CYP BM3 as a positive control (Fig. 4A). Again, P4 and P5 were well-behaved, exhibiting the characteristic peak at 450 nm confirming the CYP family predictions and suggesting that the enzymes are expressed in a correctly folded and active form. P7 and P29 were tested, but the CO-binding spectra were inconclusive probably due to low protein concentration (Fig. 4D and 4F). The P9 absorption

spectra (Fig. 4E) showed no peak at all, as expected by the absence of full-length protein and fluorescence (see previous result sections).

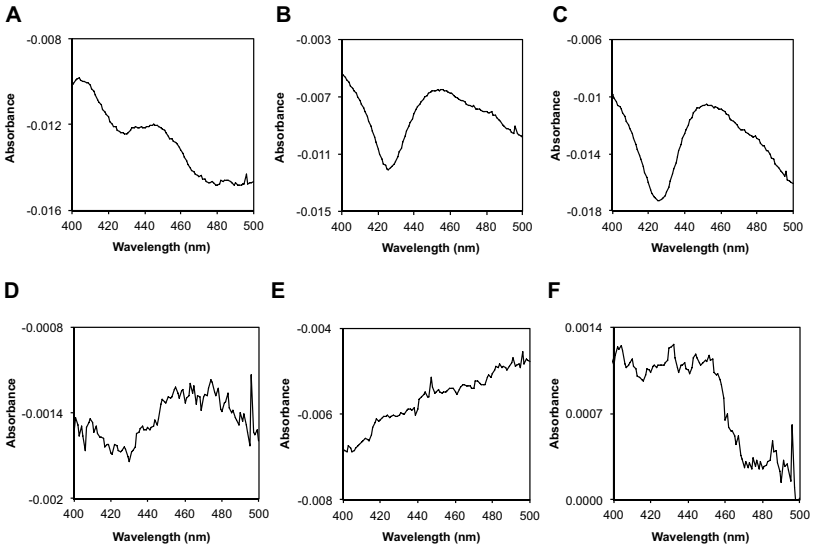


Figure 4. CO spectra different spectra of reduced CYPs.

The analyzed CYP-GFP fusions were kept in 10 mM Tris-HCl buffer pH 7.0 and used in protein concentrations of 0.065-0.184 mg/mL. Absorbance was measured ranging from 400 to 500 nm wavelength and appearance of the 450 nm characteristic peak (or the 420 nm peak for misfolded CYPs) was assessed for (A) BM3; (B) P4; (C) P5; (D) P7; (E) P9; (F) P29.

DISCUSSION

CYPs have received attention due to their association with specific and relevant metabolic pathways and their ability to perform challenging chemistry. Many studies have focused on plants that have a high amount of CYPs per genome [3,49,50,51]. Despite their ubiquity and diversity, the study and engineering of plant CYPs is facing several challenges. In higher organisms, CYPs are membrane associated, representing a major obstacle in expressing and purifying this class of enzymes [52]. Consequently, only two crystal structures of plant-derived CYPs have been published [53,54]. In addition, plant CYPs are less efficient than

their bacterial counterparts [55], meaning that high levels of expression would probably be required for effective and fast conversion of substrates.

Bacterial CYPs, although present in lower number as compared to numbers in eukaryotes, are soluble, more stable and exhibit higher activity [9,55]. Thus, they are easier to engineer, overexpress, purify and crystallize [52]; and known structures of bacterial CYPs have been used to assemble mammalian CYP structures based on homology modeling [56,57,58]. The hydroxylation of camphor by CYP101 (P450cam) in *Pseudomonas putida* [59] was a major breakthrough in the cytochrome P450 field and a few examples of studies on bacterial CYPs have been unraveling important functions in the modification of natural compounds, such as the antibiotic albaflavenone in *S. coelicolor* A3(2) [26] or polybrominated aromatic compounds in *P. luteoviolacea* 2ta16 [60]. However, reports of new bacterial CYPs with new functions/substrates are still scarce.

Marine bacteria are ideal sources of novel chemistry since they are understudied and capable of performing a wide range of biological processes under environmentally challenging conditions [61,62]. Here we present a study on genome mining and expression of bacterial CYPs. Although the number of putative CYPs identified in the 19 analyzed genomes was only 26, the diversity was considerable, with 19 different CYPs with protein identity lower than 40% (Fig. S3). This diversity was also mirrored in their location in distinct genomic environments and similarity to different CYP families, suggesting their participation in different metabolic reactions.

The number of ORFs with CYP domains identified in the studied marine bacteria varied from zero to nine per genome (Fig. 1). The high number identified for *Ruegueria mobilis* F1926 could be due to the poor quality of the genome sequence of this strain, which has more than one thousand contigs. Indeed, the four CYPs found in the other *R. mobilis* strain (S1942) suggest the average number of CYPs probably varies between zero and four in these marine

bacteria. The possible wrong annotation of ORFs with CYP domains can explain the poor results obtained for the 300 bp-CYP P9 (Fig. 3B and Fig. S1). P9 showed high similarity to CYP152B1, which is involved in fatty acid metabolism [63,64]. P9 function is likely different from the other selected CYPs. The P4 and P6 CYPs were present in the same genomic context, sharing 99,74% of amino acid identity (Fig. S3). These CYPs were found within a predicted agar degradation cluster where several agarases are present, pointing to a function in sugar modification. The closest related CYP family identified for these CYPs is CYP236, represented by CYP236A1. The latter was shown to have five agarases placed closely to it in a 101 kb plasmid (pSD15) from the agar degrading *Microscilla* sp. PRE1 [65].

AntiSMASH analysis has shown that P5 CYP is part of a predicted gamma-butyrolactone biosynthesis cluster in *Marinomonas* sp. S3726 [23,35]. Gamma-butyrolactone is a building block compound used in chemical industry and of great interest in bio-based production [66,67,68]. CYP family prediction places P5 in the CYP170 family. CYP170A1 is the family representative and it is responsible for the oxidation of the terpenoid epi-isozizaene to albaflavenone in *Streptomyces coelicolor* A3(2) [61]. CYP170A1 catalyzes a two-step full oxidation of epi-isozizaene to albaflavenone, similar to what could be expected for conversion of tetrahydrofuran to gamma-butyrolactone by oxidation of an aromatic ring (Fig. S4).

The predicted CYP P7 had high similarity to CYP104A2, an *Agrobacterium tumefaciens* CYP responsible for O-demethylation of plant phenolic compounds produced when infected with the plant-pathogen [69,70]. Other CYPs from the same CYP104 family are responsible for the O-demethylation of guaiacol during the degradation of the lignin polysaccharide [71]. The P7-coding gene is located next to predicted fructose transport and capsule polysaccharide export genes, pointing to a possible involvement in O-demethylation of sugars or polysaccharides. Sugar modifications might affect cell fitness and create a bias towards low producers, which could explain the low concentration of P7 in yeast (Fig. 3B and C). Despite

the inconclusive P7 CO-spectrum, the fact that this ORF clusters in the genome with ferredoxins/ferredoxin reductases supports its CYP prediction, as the confirmed P4 CYP (Table 2).

According to antiSMASH predictions, P29 is located right next to a NRPS/PKS gene cluster, suggesting its role as a tailoring enzyme involved in compound modification(s). Homology searches placed it in the CYP197 family, previously identified in *Streptomyces* and *Bacillus* species [72,73]. *Streptomyces* are known for being prominent producers of natural products and contain the highest number of structurally characterized CYPs [3].

All CYPs were successfully expressed in *E. coli* (Fig. S1) as commonly observed for other bacterial CYPs [12]. A growing field aims at introducing chemical modifications, not possible with standard organic chemistry [55]. One such example in *E. coli* is the synthesis of 1α,25-dihydroxyvitamin D3 from vitamin D3, where a bacterial CYP (CYP105A1) from *Streptomyces griseolus* was used [74]. Protein engineering efforts to modify P450cam [75] and BM3 [14] for *in vivo* oxidation of unnatural substrates demonstrate the importance of bacterial CYPs, and support the idea that new bacterial CYPs can be engineered to achieve broader specificity for biotechnological applications. Hence, the marine bacterial CYPs presented here appear as promising candidates for protein engineering.

Most of the successful commercial and industrially relevant CYP-dependent processes have been performed in *S. cerevisiae*. Production of a myriad of different compounds such as hydrocortisone, pregnenolone, amorphadiene (precursor of artemisinic acid), polyketides, isoprenoids, steviol components and opiates has been done in yeast, proving its value as a cell factory for these types of molecules [6,76,77,78]. Two major advantages of using yeast are a well-developed genetic engineering toolbox and a long history of use in the food industry.

Conclusions

To our knowledge, our work is the first demonstration of marine bacterial CYPs being expressed in yeast. This model cell factory offered a convenient assay allowing for cellular localization studies with GFP-fusions that are easily observed under the microscope. We also showed that it is possible, without further optimization besides codon optimization, to successfully express and produce bacterial CYPs in yeast. The stability and integrity of CYP-GFP fusions purified from yeast was evaluated by in-gel fluorescence and CO absorption spectra confirming that P4 and P5 are cytochromes P450 - and that they are properly folded in the yeast system (Fig. 4). Future work should focus on the further optimization of P7 and P29 expression in yeast.

Besides the obvious advantages of unraveling new functions that can be useful in biotechnology and replacing eukaryotic CYPs, this study is a first step towards exploiting the full potential of the bacterial CYPs present in marine environments. Future studies will aim at identifying the specific function of these enzymes.

Acknowledgements

The authors would like to acknowledge Darío Vázquez Albacete for fruitful discussions regarding CYPs, Helle Munck Petersen for technical assistance during protein purification and Per Amstrup Pedersen (Copenhagen University, Denmark) for kindly providing the PAP1503 strain and pEMBLyex4 plasmid. We would also like to thank Swee Hallwyl and Kasia Krzystanek (Evolva Biotech A/S).

Author contributions

Conceived and designed the experiments: AMC, HM, CC. Performed the experiments: AMC, HM, SK. Wrote the paper: AMC and HM with contributions from all the authors.

References

1. Werck-reichhart D, Feyereisen R. (2000) Cytochromes P450: a success story. Genome Biol 1(6): 3003.1–9.

2. Lamb DC, Waterman MR, Kelly SL, Guengerich FP. (2007) Cytochromes P450 and drug discovery. Curr Opin Biotechnol 18(6): 504–512.

3. Podust LM, Sherman DH. (2012) Diversity of P450 enzymes in the biosynthesis of natural products. Nat Prod Rep 29(10): 1251–66.

4. Urlacher V, Schmid RD. (2002) Biotransformations using prokaryotic P450 monooxygenases. Curr Opin Biotechnol 13(6): 557–564.

5. Peters MW, Meinhold P, Glieder A, Arnold FH. (2003) Regio- and Enantioselective Alkane Hydroxylation with Engineered Cytochromes P450 BM-3. J Am Chem Soc 125(44): 13442–13450.

6. Ro D, Paradise EM, Ouellet M, et al. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440: 3–6.

7. Paddon CJ, Westfall PJ, Pitera DJ, et al. (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. Nature 496: 528–532.

8. Nelson DR. (2009) The cytochrome p450 homepage. Hum Genomics 4: 59–65.

9. Kelly SL, Kelly DE. (2013) Microbial cytochromes P450: biodiversity and biotechnology. Where do cytochromes P450 come from, what do they do and what can they do for us? Philos Trans R Soc Lond B Biol Sci. 368(1612): 20120476.

10. Narhi LO, Fulco J. (1986) Characterization of a catalytically self-sufficient 119,000-Dalton Cytochrome P450 Monooxygenase Induced by Barbiturates *Bacillus megaterium*. J Biol Chem 261(16): 7160–7169.

11. Wen L, Fulco AJ. (1987) Cloning of the Gene Encoding a Catalytically Self-sufficient Fatty Acid Monooxygenase Induced by *Bacillus megaterium* and Its Functional

Expression Barbiturates in and Regulation in Heterologous (*Escherichia coli*) and Homologous (*Bacillus megaterium*) Hosts. J Biol Chem 262(14): 6676–6682.

12. Urlacher VB, Lutz-Wahl S, Schmid RD. (2004) Microbial P450 enzymes in biotechnology. Appl Microbiol Biotechnol. 64(3): 317–325.

13. Appel D, Lutz-wahl S, Fischer P, Schwaneberg U, Schmid RD. (2001) A P450 BM-3 mutant hydroxylates alkanes , cycloalkanes , arenes and heteroarenes. J Biotechnol 88:167–171.

14. Carmichael AB, Wong L-L. (2001) Protein engineering of *Bacillus megaterium* CYP102. Eur J Biochem 268(10): 3117–3125.

15. Farinas ET, Schwaneberg U, Glieder A, Arnold FH. (2001) Directed Evolution of a Cytochrome P450 Monooxygenase for Alkane Oxidation. Adv Synth Catal 343: 601–606.

16. Li Q, Ogawa J, Schmid RD, Shimizu S. (2001) Engineering Cytochrome P450 BM-3 for Oxidation of Polycyclic Aromatic Hydrocarbons. Appl Environ Microbiol 67(12): 5735–5739.

17. Glieder A, Farinas ET, Arnold FH. (2002) Laboratory evolution of a soluble, self-sufficient, highly active alkane hydroxylase. Nat Biotechnol 20: 1135–1139.

18. Farwell CC, Zhang RK, Mcintosh JA, Hyster TK, Arnold FH. (2015) Enantioselective Enzyme-Catalyzed Aziridination Enabled by Active-Site Evolution of a Cytochrome P450. ACS Cent Sci 1: 89–93.

19. Cirino PC, Arnold FH. (2002) Protein engineering of oxygenases for biocatalysis. Curr Opin Chem Biol 6(2): 130–135.

20. Bérdy J. (2012) Thoughts and facts about antibiotics: Where we are now and where we are heading. J Antibiot (Tokyo) 65: 441–441.

21. Xiong Z-Q, Wang J-F, Hao Y-Y, Wang Y. (2013) Recent advances in the discovery and development of marine microbial natural products. Mar Drugs 11(3): 700–17.

22. Wilson MC, Mori T, Rückert C, et al. (2014) An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature 506: 58–62.

23. Machado H, Sonnenschein EC, Melchiorsen J, Gram L. (2015) Genome mining reveals unlocked bioactive potential of marine Gram-negative bacteria. BMC Genomics 16(1): 1–12.

24. Fenical W, Jensen PR. (2006) Developing a new resource for drug discovery: marine actinomycete bacteria. Nat Chem Biol 2: 666–673.

25. Lozupone CA, Knight R. (2007) Global patterns in bacterial diversity. Proc Natl Acad Sci U S A 104: 11436–11440.

26. Agarwal V, El Gamal AA, Yamanaka K, et al. (2014) Biosynthesis of polybrominated aromatic organic compounds by marine bacteria. Nat Chem Biol 10(8): 640–647.

27. Inoue H, Nojima H, Okayama H. (1990) High efficiency transformation of *Escherichia coli* with plasmids. Gene 96:23–28.

28. Studier FW. (2005) Protein production by auto-induction in high-density shaking cultures. Protein Expr Purif 41: 207–234.

29. Pedersen PA, Rasmussen JH, Jørgensen PL. (1996) Expression in High Yield of Pig a1b1 Na, K-ATPase and Inactive Mutants D369N and D807N in *Saccharomyces cerevisiae*. J Biol Chem 271(5): 2514–2522.

30. Steffensen L, Pedersen PA. (2006) Heterologous Expression of Membrane and Soluble Proteins Derepresses GCN4 mRNA Translation in the Yeast *Saccharomyces cerevisiae*. Eukaryot Cell 5(2): 248–261.

31. Cesareni G, Murray JAH. (1987) Plasmid vectors carrying the replication origin of filamentous single-stranded phages. Setlow JK, ed. Genet Eng Princ Methods 9: 135–154.

32. Sambrook J, Russel DW. (2001) Molecular Cloning: A Laboratory Manual. Cold Spring Harbour, New York: Cold Spring Harbour Laboratory Press.

574 33. Gram L, Melchiorsen J, Bruhn JB. (2010) Antibacterial activity of marine culturable
575 bacteria collected from a global sampling of ocean surface waters and surface swabs
576 of marine organisms. Mar Biotechnol (NY) 12(4): 439–51.

577 34. Aziz RK, Bartels D, Best AA, et al. (2008) The RAST Server: rapid annotations using
578 subsystems technology. BMC Genomics 9: 75.

579 35. Blin K, Medema MH, Kazempour D, et al. (2013) antiSMASH 2.0 - a versatile
580 platform for genome mining of secondary metabolite producers. Nucleic Acids Res
581 41: W204–12.

582 36. Tsirigos KD, Peters C, Shu N, Kall L, Elofsson A. (2015) The TOPCONS web server
583 for consensus prediction of membrane protein topology and signal peptides. Nucleic
584 Acids Res 43: 401–407.

585 37. Hessa T, Meindl-Beinker NM, Bernsel A, *et al*. (2007) Molecular code for
586 transmembrane-helix recognition by the Sec61 translocon. Nature 450(7172): 1026–
587 1030.

588 38. Cavaleiro AM, Kim SH, Seppa S, Nielsen MT, Nørholm MHH. (2015) Accurate DNA
589 Assembly and Genome Engineering with Optimized Uracil Excision Cloning. ACS
590 Synth Biol 4: 1042–1046.

591 39. Cavaleiro AM, Nielsen MT, Kim SH, Seppälä S, Nørholm MHH. (2015) Uracil
592 excision for assembly of complex pathways. Springer Protoc 1–10.

593 40. Scharff-Poulsen P, Pedersen PA. (2013) *Saccharomyces cerevisiae*-Based Platform for
594 Rapid Production and Evaluation of Eukaryotic Nutrient Transporters and
595 Transceptors for Biochemical Studies and Crystallography. PLoS One 8(10): 1–12.

596 41. Studier FW, Rosenberg AH, Dunn JJ, Debendorff JW. (1990) Use of T7 RNA
597 Polymerase to Direct Expression of Cloned Genes. Methods Enzymol 185(1986): 60–
598 89.

599 42. Nielsen MT, Madsen KM, Seppa S, *et al*. (2014) Assembly of Highly Standardized

Gene Fragments for High-Level Production of Porphyrins in *E. coli*. ACS Synth Biol 4: 274–282.

43. Halkier BA, Sibbesen O, Koch B, Meiler BL. (1995) Characterization of cytochrome P450TYR, a multifunctional haem-thiolate7v-hydroxylase involved in the biosynthesis of the cyanogenic glucoside Dhurrin. Drug Metabol Drug Interact 12(3-4): 285–297.

44. Omura T, Sato R. (1964) The Carbon Monoxide-binding pigment of Liver Microsomes. J Biol Chem 239(7): 2370–2378.

45. Vynne NG. (2011) Bioactivity and phylogeny of the marine bacterial genus *Pseudoalteromonas*.

46. Mansson M, Gram L, Larsen TO. (2011) Production of bioactive secondary metabolites by marine *Vibrionaceae*. Mar Drugs 9(9): 1440–68.

47. Finn RD, Bateman A, Clements J, *et al.* (2014) Pfam: the protein families database. Nucleic Acids Res 42: D222–D230.

48. Guengerich FP, Martin MV, Sohl CD, Cheng Q. (2009) Measurement of cytochrome P450 and NADPH–cytochrome P450 reductase. Nat Protoc 4(9): 1245–1251.

49. Mizutani M, Sato F. (2011) Unusual P450 reactions in plant secondary metabolism. Arch Biochem Biophys 507(1): 194–203.

50. Hamberger B, Bak S. (2013) Plant P450s as versatile drivers for evolution of species-specific chemical diversity. Philos Trans R Soc Lond B Biol Sci 368(1612): 20120426.

51. Nelson D, Werck-Reichhart D. (2011) A P450-centric view of plant evolution. Plant J 66: 194–211.

52. Jung ST, Lauchli R, Arnold FH. (2011) Cytochrome P450: taming a wild type enzyme. Curr Opin Biotechnol 22(6): 809–817.

53. Lee D-S, Nioche P, Hamberg M, Raman CS. (2008) Structural insights into the evolutionary paths of oxylipin biosynthetic enzymes. Nature 455(7211): 363–368.

54. Li L, Chang Z, Pan Z, Fu Z-Q, Wang X. (2008) Modes of heme binding and substrate access for cytochrome P450 CYP74A revealed by crystal structures of allene oxide synthase. Proc Natl Acad Sci U S A 105(37): 13883–13888.

55. Bernhardt R, Urlacher VB. (2014) Cytochromes P450 as promising catalysts for biotechnological application: chances and limitations. Appl Microbiol Biotechnol 98: 6185–6203.

56. Dai R, Pincus MR, Friedman FK. (1998) Molecular Modeling of Cytochrome P450 2B1: Mode of Membrane Insertion and Substrate Specificity. J Protein Chem 17(2): 121–129.

57. Belkina N V, Lisurek M, Ivanov AS, Bernhardt R. (2001) Modelling of three-dimensional structures of cytochromes P450 11B1 and. J Inorg Biochem 87: 197–207.

58. Bathelt C, Schmid RD, Pleiss J. (2002) Regioselectivity of CYP2B6: homology modeling , molecular dynamics simulation , docking. J Mol Model 8: 327–335.

59. Nelson DR, Kamataki T, Waxman DJ, *et al*. (1993) The P450 Superfamily: Update on New Sequences, Gene Mapping, Accession Numbers, Early Trivial Names of Enzymes, and Nomenclature. DNA Cell Biol 12(1): 1–51.

60. Moody SC, Zhao B, Lei L, *et al*. (2012) Investigating conservation of the albaflavenone biosynthetic pathway and CYP170 bifunctionality in streptomycetes. FEBS J 279(9): 1640–9.

61. Fenical W. (1993) Chemical studies of marine bacteria: developing a new resource. Chem Rev 93: 1673–1683.

62. de Carvalho CCCR, Fernandes P. (2010) Production of metabolites as bacterial responses to the marine environment. Mar Drugs 8(3): 705–27.

63. Matsunaga I, Yokotani N, Gotoh O, Kusunose E, Yamada M, Ichihara K. (1997)

Molecular cloning and expression of fatty acid alpha-hydroxylase from *Sphingomonas paucimobilis*. J Biol Chem 272(38): 23592–23596.

64. Fujishiro T, Shoji O, Nagano S, Sugimoto H, Shiro Y, Watanabe Y. (2011) Crystal Structure of H2O2-dependent Cytochrome P450SP{alpha} with Its Bound Fatty Acid Substrate: Insight into the regioselective hydroxylation of fatty acids at the alpha position. J Biol Chem 286(34): 29941–29950.

65. Zhong Z, Toukdarian A, Helinski D, *et al*. (2001) Sequence analysis of a 101-kilobase plasmid required for agar degradation by a *Microscilla isolate*. Appl Environ Microbiol 67(12): 5771–5779.

66. Hwang DW, Kashinathan P, Lee JM, *et al*. (2011) Production of γ-butyrolactone from biomass-derived 1,4-butanediol over novel copper-silica nanocomposite. Green Chem 13(7): 1672.

67. Martin CH, Dhamankar H, Tseng H-C, Sheppard MJ, Reisch CR, Prather KLJ. (2013) A platform pathway for production of 3-hydroxyacids provides a biosynthetic route to 3-hydroxy-γ-butyrolactone. Nat Commun 4: 1414.

68. Dhamankar H, Tarasova Y, Martin CH, Prather KLJ. (2014) Engineering *E. coli* for the biosynthesis of 3-hydroxy-γ-butyrolactone (3HBL) and 3,4-dihydroxybutyric acid (3,4-DHBA) as value-added chemicals from glucose as a sole carbon source. Metab Eng 25: 72–81.

69. Kanemoto RH, Powell AT, Akiyoshi DE, *et al*. (1989) Nucleotide sequence and analysis of the plant-inducible locus pinF from *Agrobacterium tumefaciens*. J Bacteriol 171(5): 2506–2512.

70. Brencic A, Eberhard A, Winans SC. (2004) Signal quenching, detoxification and mineralization of vir gene-inducing phenolics by the VirH2 protein of *Agrobacterium tumefaciens*. Mol Microbiol 51: 1103–1115.

71. Filley TR, Cody GD, Goodell B, Jellison J, Noser C, Ostrofsky A. (2002) Lignin

demethylation and polysaccharide decomposition in spruce sapwood degraded by brown rot fungi. Org Geochem 33: 111–124.

72. Furuya T, Shibata D, Kino K. (2009) Phylogenetic analysis of Bacillus P450 monooxygenases and evaluation of their activity towards steroids. Steroids 74(12): 906–912.

73. Li Z-Z, Li X-F, Yang W, *et al*. (2013) Identification and functional analysis of cytochrome P450 complement in *Streptomyces virginiae* IBL14. BMC Genomics 14(1): 130.

74. Sawada N, Sakaki T, Yoneda S, *et al*. (2004) Conversion of vitamin D3 to 1alpha,25-dihydroxyvitamin D3 by *Streptomyces griseolus* cytochrome P450SU-1. Biochem Biophys Res Commun 320(1): 156–164.

75. Bell SG, Harford-cross CF, Wong L. (2001) Engineering the CYP101 system for *in vivo* oxidation of unnatural substrates. Protein Eng 14(10): 797–802.

76. Kavšček M, Stražar M, Curk T, Natter K, Petrovič U. (2015) Yeast as a cell factory: current state and perspectives. Microb Cell Fact 14(1): 94.

77. Duport C, Spagnoli R, Degryse E, Pompon D. (1998) Self-suffient biosynthesis of pregnenolone and progesterone in engineered yeast. Nat Biotechnol 16: 186–189.

78. Szczebara FM, Chandelier C, Villeret C, *et al*. (2003) Total biosynthesis of hydrocortisone from a simple carbon source in yeast. Nat Biotechnol 21: 143–149.

Supporting information

Table S1. List of oligonucleotides and plasmids.

Number	Name	Sequence
1	P4_*E.coli*_fw	AGATATACCUATGGCGAAAGTAATATTTATAACCAAAACAAACGA
2	P4_*E.coli*_rv	ATTATGCGUTTATTTTGATTTAAAGCTAACGTGCAGACTGT
3	P5_*E.coli*_fw	AGATATACCUATGAGTATCGAGTTGCATAGTAAGTTAAATAGT
4	P5_*E.coli*_rv	ATTATGCGUTTAGTTAATAACGATAGGTAGGCTATCTGGGC
5	P7_*E.coli*_fw	AGATATACCUATGCTGGATTTGCCAATCAATGAGAC
6	P7_*E.coli*_rv	ATTATGCGUTTAGTCGAGATAGACCGGCAGAT
7	P9_*E.coli*_fw	AGATATACCUATGCTCGACCTCTATGGTACGAAC
8	P9_*E.coli*_rv	ATTATGCGUATGCTCGACCTCTATGGTACGAAC
9	P29_*E.coli*_fw	AGATATACCUATGAAGCAAATACCAAAAGTAACCAACAAAG
10	P29_*E.coli*_rv	ATTATGCGUTTATTGCTCAATTAACTTATTCACTTTCATGACCAAC
11	pCDF_P450_fw	ACGCATAAUGCTTAAGCYGAACAGAAAGTAATCGTATTG
12	pCDF_P450_rv	AGGTATATCUCCTTATTAAAGTTAAACAAAATTATTTCTACAGGGGA ATTG
13	Seq_pCDF_fw	AGGTTTTGCGCCATTCGATGG
14	Seq_pCDF_rv	CGATTATGCGGCCGTGTACAA
15	P4_yGFP_fw	ACACAAATACACACACTAAATTACCGGATCAATTCTAAGATAATT ATGACTTCAAAGATTATAGGCGTTAGTGATTTTCC
16	P4_yGFP_rv	AAATTGACTTTGAAAATACAAATTTTCCTTAGACTTAAAAGAGACGT GCAAGGAA
17	P5_yGFP_fw	ACACAAATACACACACTAAATTACCGGATCAATTCTAAGATAATT ATGAGTATCGAGTTACATTCAAAGTTG
18	P5_yGFP_rv	AAATTGACTTTGAAAATACAAATTTTCGTTAATAACAATTGGCAAAG AATCAGGAC 3’
19	P7_yGFP_fw	ACACAAATACACACACTAAATTACCGGATCAATTCTAAGATAATT ATGCTAGACCTACCAATTAACGAAAC
20	P7_yGFP_rv	AAATTGACTTTGAAAATACAAATTTTCGTCTAAGTATACTGGTAGAT TTAACGG
21	P9_yGFP_fw	ACACAAATACACACACTAAATTACCGGATCAATTCTAAGATAATT ATGTTAGACTTATATGGCACGAACAC
22	P9_yGFP_rv	AAATTGACTTTGAAAATACAAATTTTCATCAAGACAACGTACATTTC TCATGATAAATC
23	P29_yGFP_fw	ACACAAATACACACACTAAATTACCGGATCAATTCTAAGATAATTA TGAAGCAGATCCCGAAG
24	P29_yGFP_rv	AAATTGACTTTGAAAATACAAATTTTCCTGCTCGATCAACTTGTTAA CTTTCATAA
25	BM3_yGFP_fw	ACACAAATACACACACTAAATTACCGGATCAATTCTAAGATAATTA TGACAATTAAAGAAATGCCTCAGCCAAAAACGT
26	BM3_yGFP_rv	AAATTGACTTTGAAAATACAAATTTTCCCCAGCCCACACGTCTTTTG CGT
27	CYC-GAL_long_seq_fw	TTACTATACTTCTATAGACACGCAAACAC
28	GFP_seq_rv	GTAGCATCACCTTCACCTTC
29	GFPup_fw	GAAAATTTGTATTTTCAAAGTCAATTTTCTAAAGGTGAAGAATTAT
30	GFPHISdo_rv	CTTCAATGCTATCATTTCCTTTGATATTGGATCATCTAATGGTGATGG TGATGGTGATGGTGTTTGTACAATTCATCCATACCAT

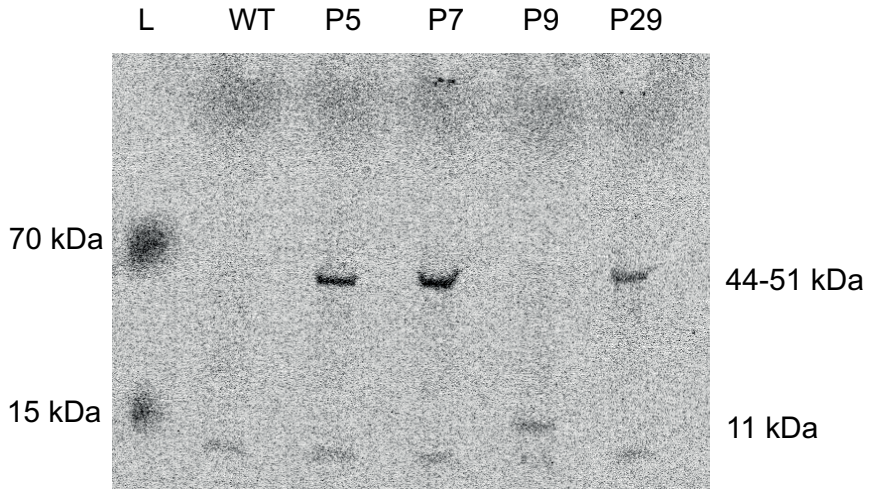


Figure S1. Bacterial CYPs expression in *E. coli* BL21 (DE3) (^{35}S-methionine labeling of proteins).

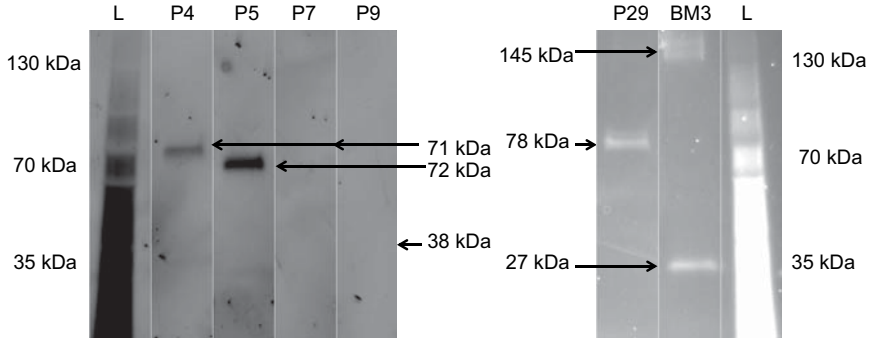


Figure S2. Bacterial CYP-GFP fusions production in yeast *S. cerevisiae* (In-gel fluorescence).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Ruegeria mobilis S1942	1	100,00	71,49	67,40	58,85	35,47	19,02	16,77	12,52	9,68	3,72	12,52	3,62	5,53	3,48	11,92	18,67	8,83	8,83	8,04	10,51	9,96	8,30	8,30	3,84	6,64
Ruegeria mobilis F1926	0	1	71,49	67,40	58,85	35,47	19,02	16,77	12,52	9,68	3,72	12,52	3,62	5,53	3,48	11,92	18,67	8,83	8,83	8,04	10,51	9,96	8,30	8,30	3,84	6,64
Phaeobacter inhibens DSM17395	134	134		64,63	56,38	34,01	19,43	16,67	12,55	9,79	3,37	12,18	2,73	5,33	3,77	12,32	17,76	8,79	8,79	8,80	11,80	9,69	8,09	8,09	3,70	5,35
Loktanella sp. S4079	149	149	168		56,61	32,48	17,96	17,04	11,75	8,45	4,05	13,46	3,06	5,73	4,47	13,27	20,00	10,15	10,15	8,24	10,57	8,33	7,41	7,41	4,19	6,15
Paracoccus sp. S4493	186	186	205	197		34,62	19,71	15,91	12,57	9,00	4,70	11,83	2,65	5,86	3,54	11,82	17,73	7,41	7,41	9,25	10,75	9,68	8,74	8,74	4,13	5,01
Paracoccus sp. S4493	302	302	315	318	306		17,85	15,98	11,71	8,14	4,61	13,84	3,21	6,17	5,15	11,79	18,09	8,21	8,21	9,33	11,63	9,04	8,67	8,67	4,25	5,52
Paracoccus sp. S4493	396	396	394	402	387	401		26,77	15,91	9,19	8,27	8,49	1,49	3,92	3,86	8,85	10,06	12,63	12,63	12,78	13,99	11,48	8,26	8,26	3,81	4,55
Ruegeria mobilis F1926	407	407	410	409	403	397	331		14,31	10,17	5,72	8,17	2,34	3,59	3,12	9,42	10,79	15,16	15,16	11,11	11,96	9,70	9,11	9,11	4,53	3,33
Ruegeria mobilis S1942	447	447	460	443	441	460	407	419		78,26	22,91	10,87	4,88	4,05	2,54	8,30	9,79	14,86	14,86	12,73	12,98	13,75	8,66	8,66	4,04	4,45
Ruegeria mobilis F1926	457	457	470	455	455	474	435	433	91		7,19	7,54	6,19	0,46	3,22	6,40	7,03	9,69	9,69	8,20	8,67	9,46	5,75	5,75	0,23	1,39
Ruegeria mobilis F1926	440	440	459	426	426	435	377	379	323	414		5,53	0,00	12,40	0,00	2,21	3,77	6,70	6,70	5,87	5,51	6,29	4,19	4,19	13,71	12,79
Ruegeria mobilis S1942	468	468	483	456	470	487	465	483	418	429	437		37,50	27,69	25,89	47,80	10,41	7,51	7,51	9,65	9,51	8,13	8,73	8,73	3,74	4,17
Ruegeria mobilis F1926	479	479	498	470	478	483	462	459	409	318	264	280		0,00	0,35	17,44	3,79	3,30	3,30	3,41	3,47	2,62	2,44	2,44	0,37	0,35
Ruegeria mobilis F1926	444	444	462	428	434	441	417	403	428	437	113	324	292		0,00	13,94	3,89	3,92	3,92	3,48	3,80	4,25	5,12	5,12	12,12	13,53
Ruegeria mobilis F1926	444	444	460	427	436	439	410	403	422	331	212	332	282	240		11,04	3,64	1,70	1,70	3,98	3,83	2,52	2,86	2,86	0,46	0,43
Phaeobacter inhibens DSM17395	473	473	484	451	470	479	482	481	431	439	443	237	374	389	403		9,68	6,87	6,87	8,51	9,53	7,95	7,44	7,44	3,05	4,57
Pseudoalteromonas luteoviolacea S4054	427	427	440	408	427	421	474	463	479	489	434	482	457	445	450	485		11,38	11,38	9,84	10,00	12,78	7,29	7,29	2,59	4,70
Pseudoalteromonas sp. S3431	485	485	498	478	487	481	429	396	424	438	376	480	439	392	404	488	448		99,74	10,21	14,59	13,71	9,09	9,09	4,23	5,20
Pseudoalteromonas sp. S2292	485	485	498	478	487	481	429	396	424	438	376	480	439	392	404	488	448	444		10,21	14,59	13,71	9,09	9,09	4,23	5,20
Marinomonas sp. S3726	492	492	499	490	481	476	416	432	432	448	385	468	425	416	410	473	462	422	422		23,08	11,83	6,25	6,25	3,71	3,68
Loktanella sp. S4079	477	477	471	474	473	462	412	427	429	442	377	466	417	405	402	465	459	398	398	320		11,74	6,99	6,99	3,57	3,04
Ruegeria mobilis F1926	479	479	483	473	476	473	432	428	414	421	358	463	409	383	387	463	423	384	384	395	391		6,15	6,15	3,00	4,67
Ruegeria mobilis S1942	530	530	534	537	522	527	489	469	506	509	412	523	480	408	441	535	534	450	450	495	492	488		100,00	21,69	5,91
Ruegeria mobilis F1926	530	530	534	537	522	527	489	469	506	509	412	523	480	408	441	535	534	450	450	495	492	488	0		21,69	5,91
Paracoccus sp. S4493	451	451	468	435	441	451	418	400	428	429	187	438	269	116	217	445	452	385	385	415	405	388	325	325		17,07
Pseudoalteromonas luteoviolacea S4054	436	436	460	427	436	445	420	407	429	426	116	437	282	115	230	439	446	383	383	422	414	388	398	398	102	

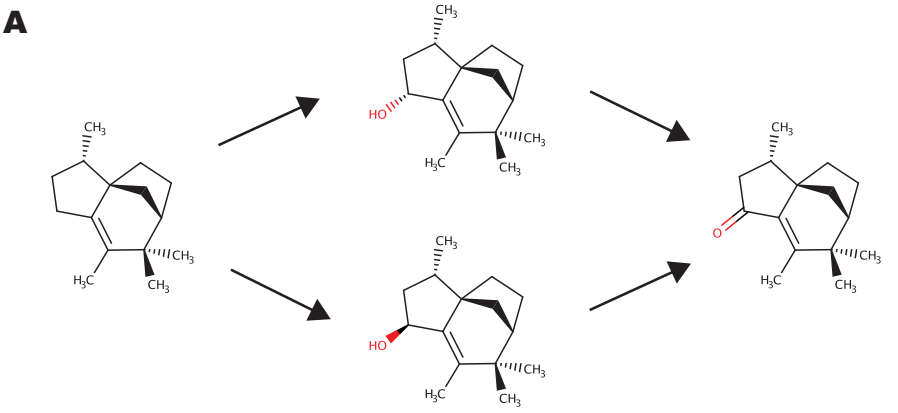
712

713 **Figure S3.** Amino-acid comparison of the 26 marine bacterial CYPs identified in this study.

714 The top quadrant presents the percentage of identity among the CYPs and the lower quadrant

715 the number of differences.

716



717

718 **Figure S4.** The chemical modifications performed by CYPs. (A) epi-isozizaene to

719 albaflavenone and (B) tetrahydrofuran to gamma-butyrolactone.

720

Paper 6

Comparative genomics reveals evidence of high genomic exchange driving *Photobacterium* genus evolution

Comparative genomics reveals evidence of high genomic exchange driving *Photobacterium* genus evolution

Henrique Machado[1,2], Lone Gram[1,*]

* corresponding author:

e-mail: gram@bio.dtu.dk

[1] Department of Systems Biology, Technical University of Denmark, Matematiktorvet, bldg. 301, DK-2800 Kgs. Lyngby, Denmark

[2] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle Allè 6, DK-2970 Hørsholm, Denmark

Abstract

Vibrionaceae family is a large marine bacterial family, which can constitute up to 50% of the prokaryotic population in marine waters. *Photobacterium* is the second largest genus in the family and includes ecological diverse species, with lifestyles ranging from free-living to symbiotic or pathogenic. We used comparative genomics on 35 strains representing 16 of the 28 species described so far, to understand if such diverse lifestyles would be reflected in the genomic information.

We developed a well-supported phylogeny for the genus using the genomic information by standard and novel approaches. This phylogeny correlated well with the occurrence of higher and lower GC content clades. High genomic diversity could be identified and was attributed to the number of foreign DNA acquisition occurrences by means of plasmid uptake, bacteriophage infection and genomic island transposition.

The analyses here presented proves the high exchange of genetic material in the *Photobacterium* genus, therefore increasing the awareness of the variability existent within species and the carefulness needed in generalizing conclusions, as every case seems to be unique.

Introduction

Oceans cover most (70%) of Planet Earth and it has been estimated that they harbor more than 60% of the global prokaryotic diversity [1]. During the last decades, bioprospecting efforts have turned to the marine environments with the hope of finding novel drugs, enzymes, metabolic pathways, or even new forms of life that could address the growing demand for efficient and biosustainable products [2–6]. Also, the rapid development in sequencing technologies and the focus on improved cultivation of hitherto un-cultured microorganisms has dramatically increased our understanding of both diversity and biotechnological potential stored in the marine habitats [7–12].

Vibrionaceae is a prominent marine bacterial family and it represents 0.8% of the bacterial composition as found in the *Tara* oceans metagenomic data [12]. Based on molecular assessment, it can constitute up to 50% of the prokaryotic population in marine waters [13,14]. *Vibrio* is the largest genus within this family (73% in *Tara* oceans data) and it has been extensively studied, in part due to the importance of the human pathogen *V. cholera* [15–19]. The genus *Photobacterium* (16%) is the second largest of the family, followed by *Aliivibrio* (3%).

Species belonging to the *Vibrionaceae* family have been reported to be taxonomically close related [20]. However most of the studies have been based on *Vibrio* species, and few included *P. profundum* strains. Recent metagenomic data shows that although these two genera seem to have similar ecological strategies, *Photobacterium* spp. dominates in

the lower pelagic depths (surface water layer and deep chlorophyll maximum layer) while *Vibrio* spp. takes over in higher depths (mesopelagic zone) (Figure 1) [12]. This suggests different ecological strategies and roles between the two genera.

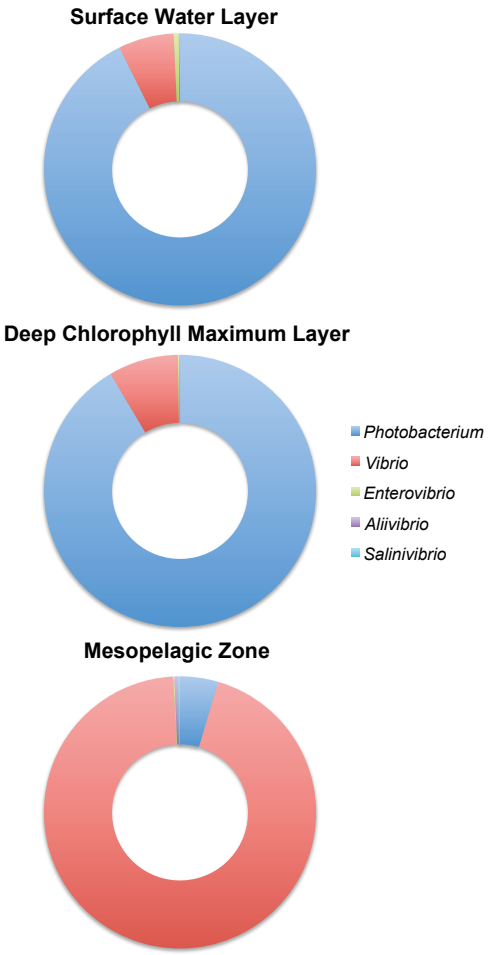


Figure 1 – Taxonomic composition of the *Vibrionaceae* family at different pelagic depths. The data was retrieved from the Tara oceans project, and included stations where information was acquired for the three depths[12]. Surface water layer: down to 5 m depth; deep chlorophyll maximum layer: from 5 to 200 m depth; and mesopelagic zone: from 250 to 1,000 m depth.

Photobacterium includes ecological diverse species, with lifestyles ranging from free-living to symbiotic or pathogenic, as does the *Vibrio* genus [21]. Twenty-eight species belonging to the genus have been described, of which 26 have validly published names (www.bacterio.net) and two are not yet valid (*P. atrarenae* and *P. marinum*) [21–30]. Of these 28, only 27 are available since no type strain has been deposited for *P. aplysiae*, [31,32]. Twenty-two of the 28 species have been described during the last 15 years, making it a relatively new and unstudied genus.

The majority of the *Photobacterium* strains have been isolated from marine environments (both from water samples and marine organisms), although isolation of a *P. halotolerans* from the rhizosphere of a terrestrial weed has recently been described [33,34].

All species of this genus were originally thought to be luminescent, but as more species were characterized, it became evident that a large number did not display this trait [21,24]. This ecological important characteristic is part of the symbiotic life style of some *Photobacterium* species. It is attributed to the expression of the *lux-rib* operon [35], which has been used in taxonomy studies for the division of the genus into clades [21,36].

Recently, we and other scientists have whole genome sequenced several strains of this genus, representing a total of 16 different species. The vast majority of these have been isolated from the marine environment, although with different approaches and goals. Some

were isolated as symbionts of marine animals, others from spoiled fish, and others as commensals in association with oysters, crabs and fish [21,37–39]. Some were isolated as free-living organisms from coastal or pelagic waters [40], and *P. galatheae* S2753, was isolated due to its antagonism of pathogenic bacteria [28,41].

In this study, we use a collection of whole genome sequences from 35 strains representing 16 of the 28 species described so far. As previously mentioned, the genus is widespread in marine environments, and can grow and colonize under different conditions and/or in association with eukaryotic organisms. We hypothesized that these different lifestyles would be reflected in the genomic information. Using comparative genomics we investigated the evolutionary mechanisms responsible for the genomic diversification within the genus, and analyzed the genetic foundation of known physiological traits of *Photobacterium* spp..

The analyses provide better understanding of the genus phylogeny, by elucidating evolutionary relationships using genomic information; it reveals the large genomic diversity within the genus and the role of foreign DNA in the genome evolution.

Material and Methods

General genome statistics

The shotgun whole genome sequences from 35 strains of the *Photobacterium* genus and the sequence of *Vibrio pacinii* DSM 19139 were used in this study (Table 1). Calculation of basic genome statistics such as size, GC content and amino acid usage was done using the CMG biotools [42]. These tools were also used for gene finding using prodigalrunner, generation of BLAST atlases comparing the protein-coding open reading frames and for the pan- and core-genome analyses of the studied strains.

Table 1 – Strains used in this study and accession numbers of its genomes.

Strain	WGS/ChrI	ChrII	Plasmid
***Photobacterium angustum* ATCC 25915**	JZSO01		
***Photobacterium angustum* ATCC 33975**	JZSM01		
***Photobacterium angustum* ATCC 33977**	JZSN01		
***Photobacterium angustum* S14**	AAOJ01		
***Photobacterium aphoticum* DSM 25995**	LDOV01		
***Photobacterium aphoticum* JCM 19237**	BBMN01		
***Photobacterium aquae* CGMCC 1.12159**	LDOT01		
***Photobacterium damselae* subsp. *damselae* ATCC 33539**	JZSI01		
***Photobacterium damselae* subsp. *damselae* CIP 102761**	ADBS01		
***Photobacterium damselae* subsp. *piscicida* DI21**	AKYG01		KC344732
***Photobacterium gaetbulicola* AD005a**	JWLZ01		
***Photobacterium gaetbulicola* Gung47**	CP005973	CP005974	KC687076
***Photobacterium galatheae* S2753**	JMIB01		

***Photobacterium ganghwense* DSM 22954**	LDOU01		
***Photobacterium halotolerans* DSM 18316**	AULG01		
***Photobacterium halotolerans* MELD1**	JWYV01		
***Photobacterium iliopiscarium* ATCC 51760**	JZSQ01		
***Photobacterium iliopiscarium* ATCC 51761**	JZSR01		
***Photobacterium kishitanii* ATCC BAA-1194**	JZSP01		
***Photobacterium kishitanii* GCSL-A1-1**	JZTE01		
***Photobacterium kishitanii* GCSL-A1-2**	JZTD01		
***Photobacterium kishitanii* GCSL-A1-3**	JZTC01		
***Photobacterium kishitanii* GCSL-A1-4**	JZTB01		
***Photobacterium leiognathi* ATCC 25521**	JZSK01		
***Photobacterium leiognathi* ATCC 33979**	JZSL01		
***Photobacterium leiognathi* lrivu.4.1**	BANQ01		
***Photobacterium leiognathi* subsp. *mandapamensis* svers.1.1**	BACE01		
***Photobacterium phosphoreum* ANT-2200**	CCAR01		
***Photobacterium phosphoreum* ATCC 11040**	JZSJ01		
***Photobacterium profundum* 3TCK**	AAPH01		
***Photobacterium profundum* SS9**	CR354531	CR354532	CR377818
***Photobacterium sanctipauli* A-394**	JGVO01		
***Photobacterium* sp. AK15 (*marinum*)**	AMZO01		
***Photobacterium* sp. SKA34**	AAOU01		
***Photobacterium swingsii* CAIM 1393**	LELC01		

Phylogenetic analyses.

The 16S rRNA genes used are the publically available sequences originally published for each strain. The genes used for the MLSA (*ftsZ*, *gapA*, *gyrB*, *mreB*, *pyrH*, *recA*, *rpoA* and *topA*) and the *fur* gene, were obtained from the genomes using the CLC Main Workbench

(CLC Aarhus, Denmark version 7). An annotation-based search was performed for the genes of interest using the NCBI annotation. The genomes not annotated at NCBI were annotated using RAST [43,44]. The accession numbers for 16S rRNA sequences and the gene locus tags are provided in Table S1. Alignments and Maximum Likelihood Phylogeny trees were done using the tools in the CLC Main Workbench (CLC Aarhus, Denmark version 7). Maximum Likelihood Phylogeny trees were constructed using the Neighbor Joining method with the Jukes-Cantor nucleotide distance measure. The topology of the tree was tested with 1000 bootstrap replications. MEGA 6 was used to finalize the tree design [45].

The Genome-to-Genome Distance Calculator by DSMZ was used to determine *in-silico* DNA-DNA reassociation percent values [46], and the Average Nucleotide Identity (ANI) calculator to estimate the ANI values [47]. The nucleotide identity percentages for the fur gene and the MLSA were calculated using the pair-wise comparison tool of CLC Main Workbench (CLC Aarhus, Denmark version 7).

***lux-rib* operon**

The *lux* genes were identified by homology search to previously described genes [36,48,49] using the BLAST tools of CLC Main Workbench (CLC Aarhus, Denmark version 7). The *lux* operons were blasted against whole genome sequences of the studied *Photobacterium* strains using MultiGeneBLAST [50].

Identification of prophages, genomic islands and secondary metabolite clusters

Prophages, genomic islands and secondary metabolite clusters were identified using the online tools PHAST, IslandViewer 3 and antiSMASH 3.0, respectively [51–53]. The whole genome sequences were submitted to the different tools, and the identified clusters compared using MultiGeneBLAST [50].

CRISPR-*Cas*

Genome sequences were analyzed in CRISPRfinder [54] and homology searches for CRISPR associated genes were performed using CLC Main Workbench (CLC Aarhus, Denmark version 7). Using the CLC Main Workbench BLAST tool, direct repeats and protospacers were compared among each other and to previously identified prophages. Confirmation of *cas* genes was performed by BLAST comparison of the identified open-reading frames to the NCBI nucleotide database.

Plasmid comparison

We used previously described *Photobacterium* plasmids to search the genomes for contigs belonging to plasmids using CLC Main Workbench (CLC Aarhus, Denmark version 7). Plasmids used in the search included: pPHDD1 (FN597600.2), pAQU1 (AB571865.1), pP99-018 (AB277723.1), pP91278 (AB277724.1), pPHDP60

(KC344732.1), pPHDP10 (DQ069059.1), pPHDP70 (KP100338.1), pP9014 (AB453229.1), pPH1 (AY789019.1), pPBPR1 (CR377818.1), and the unnamed plasmid from *P. gaetbulicola* Gung47 (KC687076.1).

Identification of virulence and histamine production genes

The genes *hlyA* and *dly* are key virulence genes of *P. damelae* [55–57]. Virulence related genes were identified using the BLAST tools of CLC Main Workbench (CLC Aarhus, Denmark version 7). Also the genetic basis for histamine production was evaluated, by searching for histamine producing genes, previously identified in *Photobacterium* strains [38].

Results

General Features of the sequenced genomes

The genomes ranged in size from 4.2 to 6.4 Mb (Figure 2). The largest genomes were those of *P. profundum* and *P. gaebuticola*, which are the only species with fully sequenced genomes. The GC content of the genomes varied between 38.7 and 50.9%, and the strains clustered in two groups based on GC%; one around 40% and another around 50%. This corresponded to the clustering observed in amino acid usage (Figure S1). The only exception was *P. swingsii*, which had a GC content of 43.4%. The number of genes per genome estimated using prodigalrunner ranged from 4,000 to 7,000.

The only two completely sequenced genomes, from *P. profundum* SS9 and *P. gaetbulicola* Gung47, have a larger and a smaller chromosome of approximately 4 and 2 Mb, plus megaplasmids of 80 and 35 Kb, respectively. The presence of 2 chromosomes is a trend of the *Vibrionaceae* family and it is assumed that the draft genomes used also include two chromosomes and large plasmids, but fully closed genomes would be required to assert this.

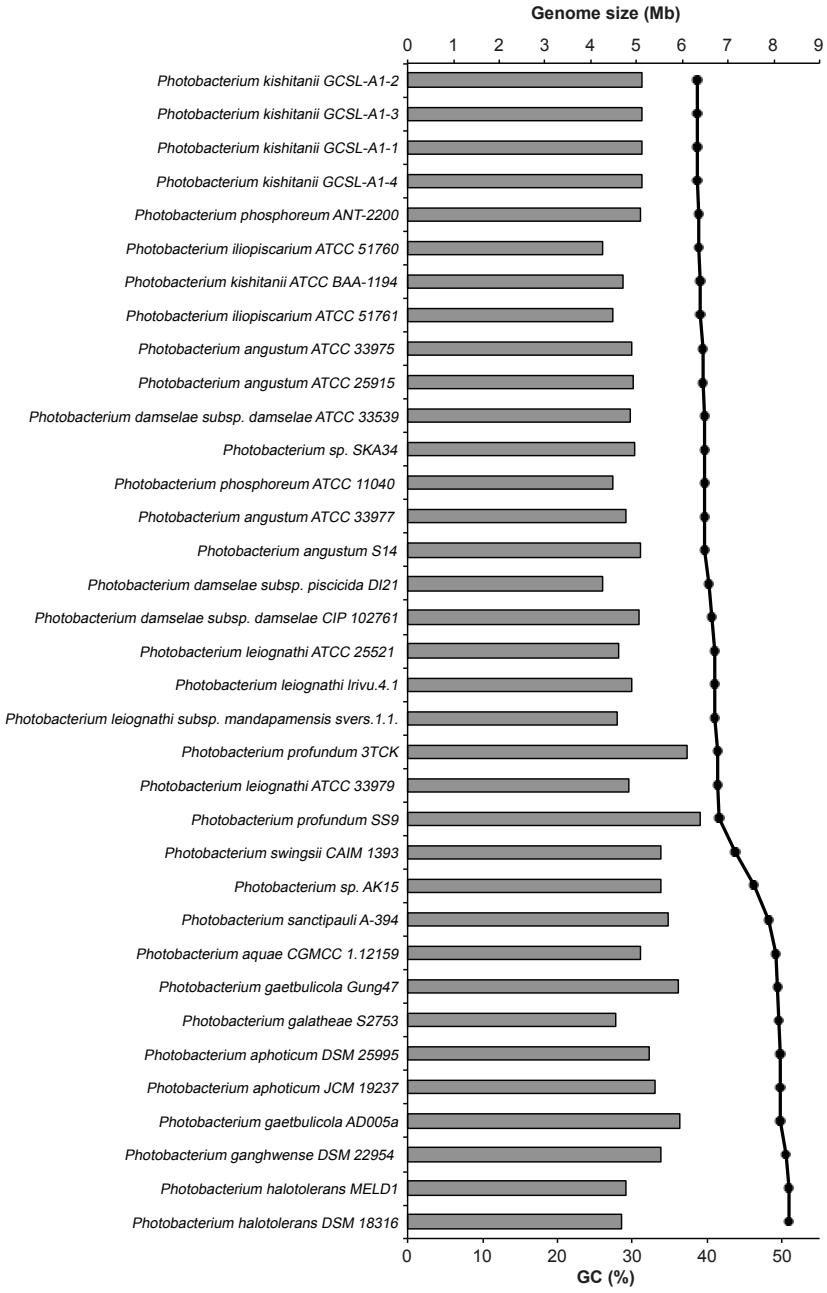


Figure 2 – Genome size and GC content of the studied genomes. The size is represented in mega-bases (Mb) by the horizontal bars and the percentage of GC by the black squares.

Phylogeny

The 16S rRNA phylogeny is widely used in the classification of *Photobacterium* species and it is sometimes the only phylogenetic discrimination provided when new species are described [22,26,58,59]. We constructed a phylogenetic tree with the 16S rRNA gene sequences available from the species (Figure 3A). However, this gene only provides a low discriminatory power in resolving species relatedness for this genus. For instance, *P. angustum* strains appear scattered throughout the phylogenetic tree (Figure 3A), *P. damseale* subsp. *piscicida* DI21 and *Photobacterium marinum* AK15 cluster closely to *V. pacinii* DSM18139, and although *P. profundum* strains cluster together, the phylogenetic distance represented by the length of the branches seems exaggerated.

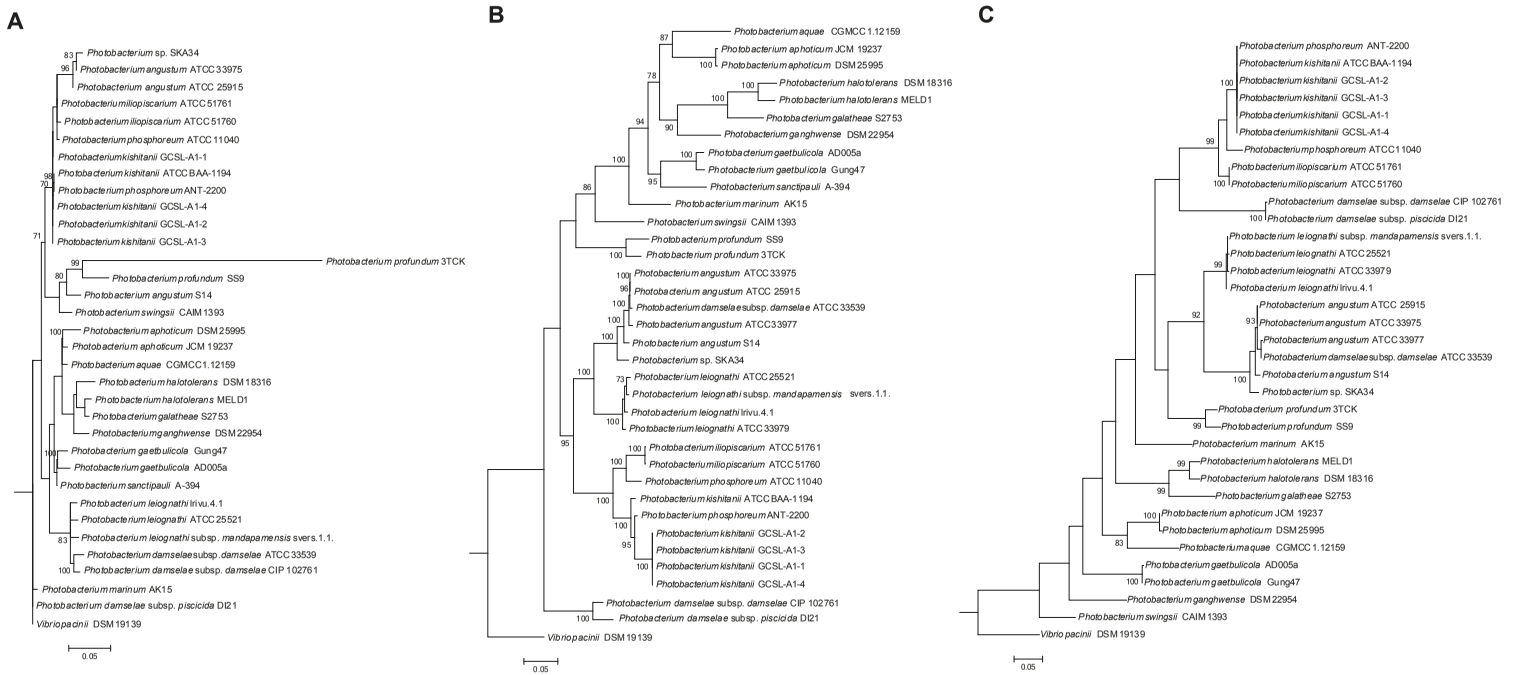
The limitations of the use of the 16S rRNA gene as a phylogenetic marker in *Vibrionaceae* were previously reported [60,61], and the use of MultiLocus Sequence Analysis (MLSA) has been proposed when evaluating *Vibrionaceae* phylogenetic relationships [60,62–65]. We extracted the gene sequences generally used in these analyses (*ftsZ*, *gapA*, *gyrB*, *mreB*, *pyrH*, *recA*, *rpoA*) from the genomes and constructed a phylogenetic tree (Figure 3B). The gene *topA* was not used, since no sequence was found in the genome of *P. galatheae* S2753. The sequences of the seven genes allowed a phylogenetic evaluation using a concatenation length of approximately 7,230 bp. Here, strains belonging to the same species cluster tightly and the phylogenetic relationship between species was clearer. This also

allowed recognizing a mis-identification or mis-naming of two strains: *P. phosphoreum* ANT-2200 and *P. damseale* subsp. *damseale* ATCC 33539, which belong to the species *P. kishitanii* and *P. angustum*, respectively. Also, although the species *P. marinum* has not been recognized, we here show that the type-strain AK15 seems to be indeed the representative of a new species [29].

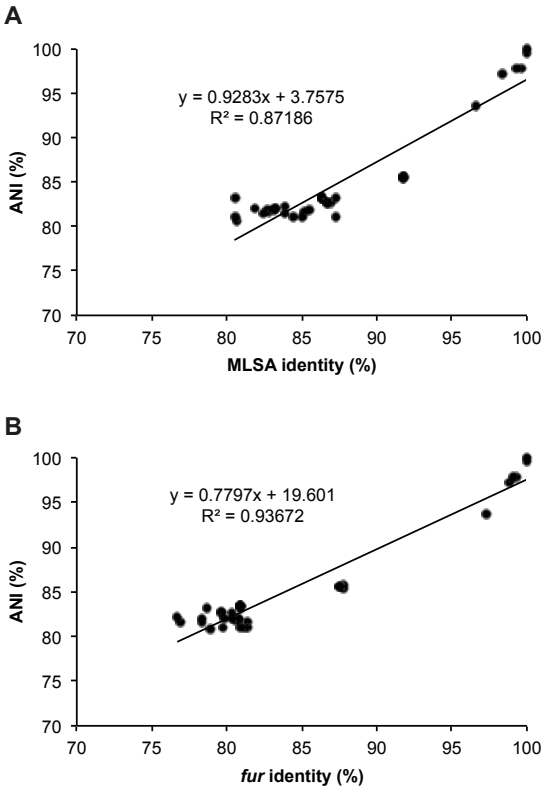
The ferric up-take regulator gene (*fur*) has recently been described as a powerful phylogenetic marker for the *Vibrionaceae* family (Machado & Gram, 2015). In order to evaluate if this was the case for the *Photobacterium* genus, we extracted the *fur* sequences from the sequenced genomes and used them for the construction of a phylogenetic tree (Figure 3C). The disposition of the species in the phylogenetic tree was different from the one in the MLSA analysis, but the core phylogenetic relationships observed were the same. In both cases, species *P. kishitanii*, *P. phosphoreum* and *P. illiopiscarium* formed one cluster, *P. leiognathi* and *P. angustum* another cluster, while *P. damselae* and *P. profundum* had their own cluster. Also two small clusters comprising *P. halotolerans* and *P. galatheae*, and *P. aphoticum* and *P. aquae* could be identified.

>>

Figure 3 – Phylogenetic trees. The trees were constructed using the 16S rRNA gene (A), seven house-keeping genes (*ftsZ*, *gapA*, *gyrB*, *mreB*, *pyrH*, *recA*, *rpoA*) (B) and the *fur* gene (C). The topology of the tree was tested with 1000 bootstrap replications, and Vibrio pacinii DSM 19139 used to root the tree.


A
Photobacterium sp. SKA34
Photobacterium angustum ATCC 33975
Photobacterium angustum ATCC 25915
Photobacterium miliopiscarium ATCC 51761
Photobacterium iliopiscarium ATCC 51760
Photobacterium phosphoreum ATCC 11040
Photobacterium kishitanii GCSL-A1-1
Photobacterium kishitanii ATCC BAA-1194
Photobacterium phosphoreum ANT-2200
Photobacterium kishitanii GCSL-A1-4
Photobacterium kishitanii GCSL-A1-2
Photobacterium kishitanii GCSL-A1-3
Photobacterium profundum 3TCK
Photobacterium profundum SS9
Photobacterium angustum S14
Photobacterium swingsii CAIM 1393
Photobacterium aphoticum DSM 25995
Photobacterium aphoticum JCM 19237
Photobacterium aquae CGMCC 1.12159
Photobacterium halotolerans DSM 18316
Photobacterium halotolerans MELD1
Photobacterium galatheae S2753
Photobacterium ganghwense DSM 22954
Photobacterium gaetbulicola Gung47
Photobacterium gaetbulicola AD005a
Photobacterium sanctipauli A-394
Photobacterium leignathi lrivu.4.1
Photobacterium leignathi ATCC 25521
Photobacterium leignathi subsp. mandapamensis svers.1.1.
Photobacterium damselae subsp. damselae ATCC 33539
Photobacterium damselae subsp. damselae CIP 102761
Photobacterium marinum AK15
Photobacterium damselae subsp. piscicida DI21
Vibrio pacinii DSM 19139
0.05
B
Photobacterium aquae CGMCC 1.12159
Photobacterium aphoticum JCM 19237
Photobacterium aphoticum DSM 25995
Photobacterium halotolerans DSM 18316
Photobacterium halotolerans MELD1
Photobacterium galatheae S2753
Photobacterium ganghwense DSM 22954
Photobacterium gaetbulicola AD005a
Photobacterium gaetbulicola Gung47
Photobacterium sanctipauli A-394
Photobacterium marinum AK15
Photobacterium swingsii CAIM 1393
Photobacterium profundum SS9
Photobacterium profundum 3TCK
Photobacterium angustum ATCC 33975
Photobacterium angustum ATCC 25915
Photobacterium damselae subsp. damselae ATCC 33539
Photobacterium angustum ATCC 33977
Photobacterium angustum S14
Photobacterium sp. SKA34
Photobacterium leignathi ATCC 25521
Photobacterium leignathi subsp. mandapamensis svers.1.1.
Photobacterium leignathi lrivu.4.1
Photobacterium leignathi ATCC 33979
Photobacterium miliopiscarium ATCC 51761
Photobacterium miliopiscarium ATCC 51760
Photobacterium phosphoreum ATCC 11040
Photobacterium kishitanii ATCC BAA-1194
Photobacterium phosphoreum ANT-2200
Photobacterium kishitanii GCSL-A1-2
Photobacterium kishitanii GCSL-A1-3
Photobacterium kishitanii GCSL-A1-1
Photobacterium kishitanii GCSL-A1-4
Photobacterium damselae subsp. damselae CIP 102761
Photobacterium damselae subsp. piscicida DI21
Vibrio pacinii DSM 19139
0.05
C
Photobacterium phosphoreum ANT-2200
Photobacterium kishitanii ATCC BAA-1194
Photobacterium kishitanii GCSL-A1-2
Photobacterium kishitanii GCSL-A1-3
Photobacterium kishitanii GCSL-A1-1
Photobacterium kishitanii GCSL-A1-4
Photobacterium phosphoreum ATCC 11040
Photobacterium iliopiscarium ATCC 51761
Photobacterium miliopiscarium ATCC 51760
Photobacterium damselae subsp. damselae CIP 102761
Photobacterium damselae subsp. piscicida DI21
Photobacterium leignathi subsp. mandapamensis svers.1.1.
Photobacterium leignathi ATCC 25521
Photobacterium leignathi ATCC 33979
Photobacterium leignathi lrivu.4.1
Photobacterium angustum ATCC 25915
Photobacterium angustum ATCC 33975
Photobacterium angustum ATCC 33977
Photobacterium damselae subsp. damselae ATCC 33539
Photobacterium angustum S14
Photobacterium sp. SKA34
Photobacterium profundum 3TCK
Photobacterium profundum SS9
Photobacterium marinum AK15
Photobacterium halotolerans MELD1
Photobacterium halotolerans DSM 18316
Photobacterium galatheae S2753
Photobacterium aphoticum JCM 19237
Photobacterium aphoticum DSM 25995
Photobacterium aquae CGMCC 1.12159
Photobacterium gaetbulicola AD005a
Photobacterium gaetbulicola Gung47
Photobacterium ganghwense DSM 22954
Photobacterium swingsii CAIM 1393
Vibrio pacinii DSM 19139
0.05

In order to provide a quantitatively evaluation of the different phylogenetic approaches, we assessed them by comparing one of the genomes (*P. angustum* ATCC 33975) to all the other 34 *Photobacterium* genomes. *In silico* DNA-DNA hybridization, *fur* and MLSA percentage of identity values were calculated and correlated to the Average Nucleotide Identity (ANI). A correlation coefficient between 0.84 and 0.95 of the phylogenetic information could be observed for all the tested methods (Figure 4). In fact, the *fur* gene identity showed a higher correlation coefficient than the MLSA identity, closer to the coefficient obtained for the methods using the whole genome sequences (DNA-DNA hybridization and ANI).



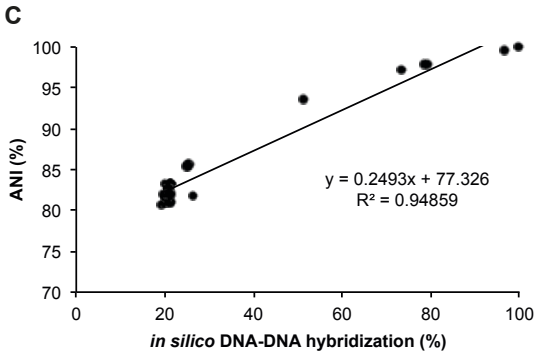


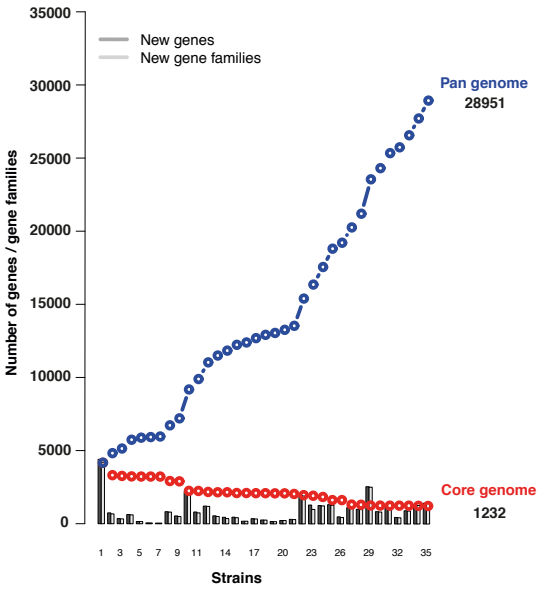
Figure 4 – Evaluation of the discriminatory power of the different methods. Assessment was made using the genomic information of *P. angustum* ATCC 33975 in comparison with the genomic information from all the other *Photobacterium* strains used in the study. The values were plotted and the equation of correlation is presented. The R^2 represents the correlation coefficient.

Pan- and core-genome

The pan-genome refers to the total number of orthologous genes in all the 35 strains, while the core genome represents the number of orthologous genes shared between all the strains. A pan- and core-genome analyses were performed using the 35 strains of the *Photobacterium* genus (Figure 5). The pan-genome consisted of 28,951 gene families, while the core-genome had 1,232 gene families. Taking into consideration the average gene number of 4,750 for the *Photobacterium* strains, 1,232 genes represents 25% of the total genome, which means that 1/4 of the genome is conserved.

Using a power-law regression, it is possible to evaluate the openness of a pan-genome [66]. The *Photobacterium* pan-genome is open, with a γ

parameter of 0.62 in a power-law regression fitting relatively well the data analyzed (R^2 = 0.89) (Figure S4).



Strain	Description
1	*P. phosphoreum* ANT-2200
2	*P. phosphoreum* ATCC 11040
3	*P. kishitanii* ATCC BAA-1194
4	*P. kishitanii* GCSL-A1-1
5	*P. kishitanii* GCSL-A1-2
6	*P. kishitanii* GCSL-A1-3
7	*P. kishitanii* GCSL-A1-4
8	*P. iliopiscarium* ATCC 51760
9	*P. iliopiscarium* ATCC 51761
10	*P. damselae* subsp. *piscicida* DI21
11	*P. damselae* subsp. *damselae* CIP 102761
12	*P. damselae* subsp. *damselae* ATCC 33539
13	*P. leiognathi* ATCC 25521
14	*P. leiognathi* ATCC 33979
15	*P. leiognathi* lrivu.4.1
16	*P. leiognathi* subsp. *mandapamensis* svers.1.1.
17	*P. angustum* ATCC 25915
18	*P. angustum* ATCC 33975
19	*P. angustum* ATCC 33977
20	*P. angustum* S14
21	*Photobacterium* sp. SKA34
22	*P. profundum* 3TCK
23	*P. profundum* SS9
24	*Photobacterium* sp. AK15 (*marinum*)
25	*P. halotolerans* DSM 18316
26	*P. halotolerans* MELD1
27	*P. galatheae* S2753
28	*P. aphoticum* DSM 25995
29	*P. aphoticum* JCM 19237
30	*P. aquae* CGMCC 1.12159
31	*P. gaetbulicola* AD005a
32	*P. gaetbulicola* Gung47
33	*P. ganghwense* DSM 22954
34	*P. swingsii* CAIM 1393
35	*P. sanctipauli* A-394

Figure 5 – Pan- and core-genome plot of the 35 *Photobacterium* strains.

BLAST altases

To visualize the gene content conservation in the *Photobacterium* genus, we constructed a BLAST atlas using 30 *Photobacterium* strains that represented the different species. Because the reference genomes have to be closed genomes, we compared 29 strains to *P. profundum* SS9 and *P. gaetbulicola* Gung47 (Figure 6). The large chromosome is more conserved than the smaller one and the plasmids, which appear to be extremely variable between species and strains. Also, it is

possible to identify genomic regions unique for the reference strain or to its species.

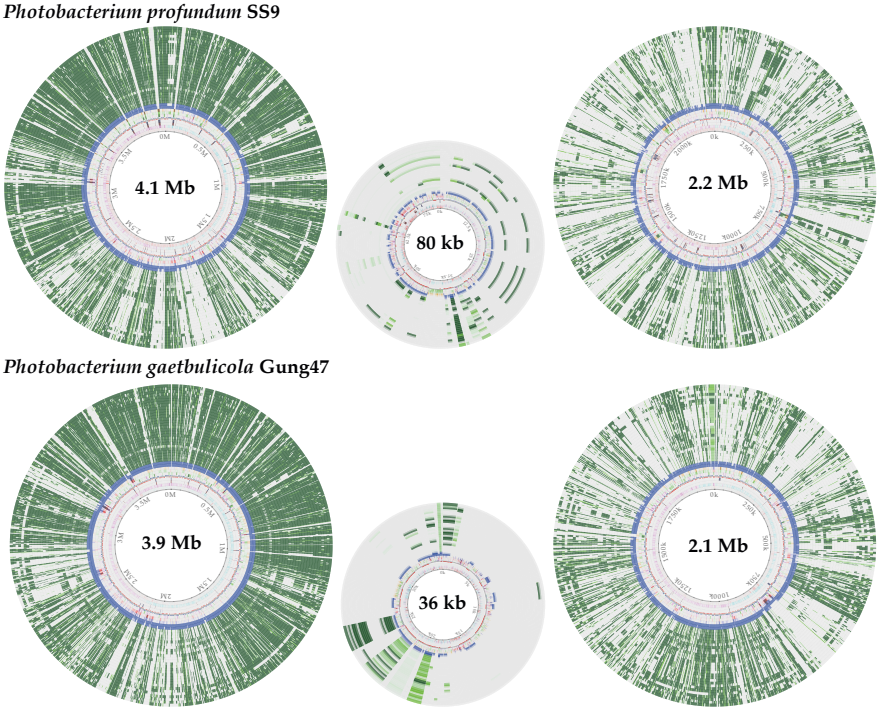


Figure 6 – BLAST atlas using the chromosomes and plasmids of *P. profundum* SS9 and *P. gaetbulicola* Gung47 as reference strains. The first blue lane is the reference genome used, followed by other strains of the same species also colored in blue. The remaining species and strains are colored in green and appear in the outsider rings of the atlas.

Genomic islands and prophages

The genetic exchange of foreign DNA by means of transposable elements, phage infection or conjugative plasmids has been suggested as a driving force in the evolution of members of *Vibrionaceae* [21,67–69].

Using Island Viewer [52], we searched for genomic islands in the fully sequenced genomes of *P. profundum* SS9 and *P. gaetbulicola* Gung47, and compared these to the genomes of other strains from the same species (Figure 7). Some of the major genetic differences between the strains of the same species seem indeed to be related to the presence or absence of specific genomic islands. Genomic regions only present in the reference strain are placed closed to identified genomic islands.

Additionally, using the PHAge Search Tool (PHAST) [51] we identified 32 intact prophage sequences and 59 incomplete ones. From the 32 intact sequences of prophages, 16 were present in multiple strains. Prophage C was present only in *P. profundum* strains, and prophage B seems to be present in all *P. kishitanii* strains. This is an indication of species-specific prophages, although prophage B is also present in other strains and prophage A is spread among several species. Also, the vast majority of prophages identified were a single case.

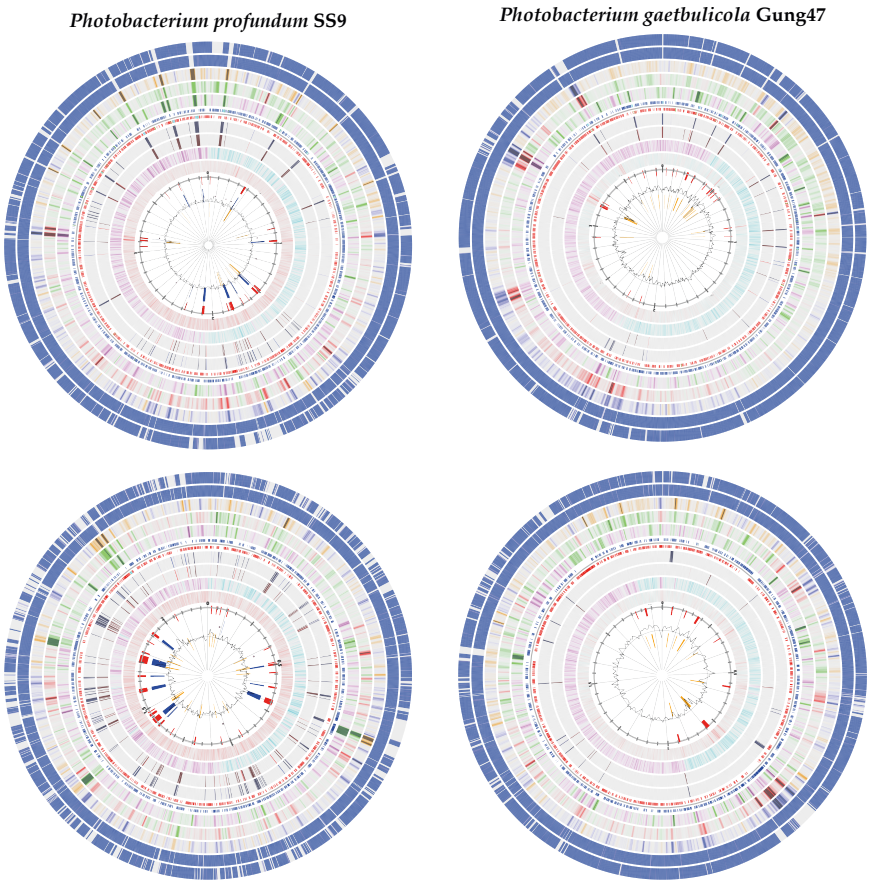
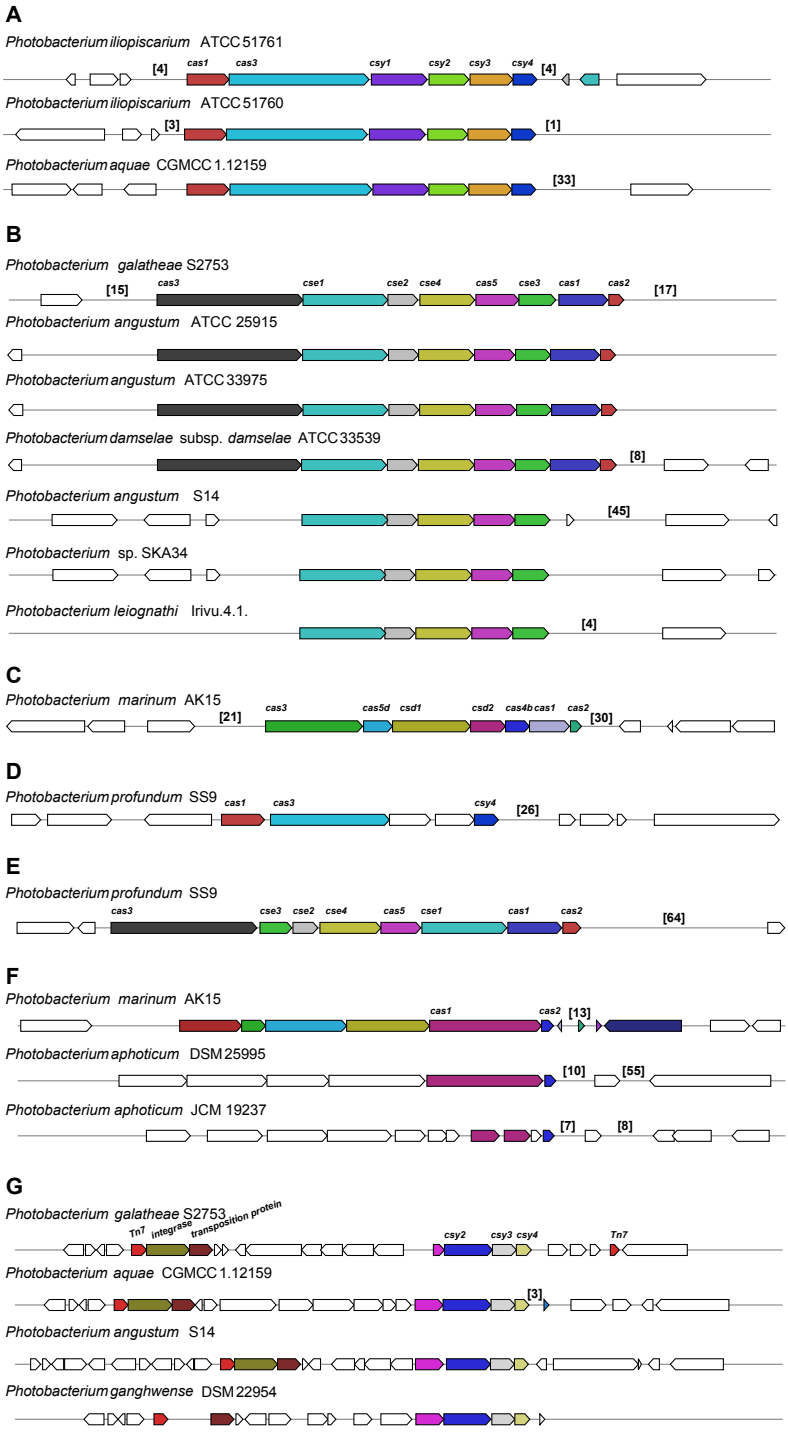


Figure 7 – Identification of genomic islands in the fully sequenced genomes of *P. profundum* SS9 and of *P. gaetbulicola* Gung47. The genomes of the mentioned strains are compared to another from the same species, *P. profundum* 3CTK and *P. gaetbulicola* AD005, respectively. The circle inside the BLAST atlas shows the Island Viewer results (integrated results in red, SIGI-HMM in orange and IslandPath-DIMOB in blue).

CRISPR-Cas System

The bacterial and archaeal adaptive immune systems entail CRISPR-Cas modules [70] and we queried the genomes for the architecture of the CRISPR-Cas systems including the *cas* gene organization, the direct repeats and the protospacers in the CRISPR locus. We divided the architecture of the CRISPR-Cas systems into seven clusters (Figure 8). Most of the architectures identified were similar to the previously described ones of *Yersinia pestis* (Figure 8A), *Escherichia coli* (Figure 8B) and *Desulfivibrio vulgaris* (Figure 8C) [71]. Also, we identified two clusters in *P. profundum* SS9 similar to the ones of *Y. pestis* and *E. coli*, encoded in the chromosome and plasmid, respectively (Figure 8D and 8E, respectively). These clusters had different gene arrangement and/or included genes coding for unknown proteins in the operon. Other two clusters containing CRISPR-associated genes were identified (Figure 8F and 8G).

In most of the cases, both direct repeats and protospacers could be identified both upstream and downstream to the *cas* operon. These were also identified elsewhere in the genome or in distinct contigs, usually consisting of one of the ends of the contig. For example, further upstream from the Tn7 in the *P. galatheae* S2753 (Figure 8G), an array of 16 protospacers could be identified beside a gene coding for a DNA nicking enzyme.


A
Photobacterium iliopiscarium ATCC 51761
[4]
cas1
cas3
csy1
csy2
csy3
csy4
[4]
Photobacterium iliopiscarium ATCC 51760
[3]
[1]
Photobacterium aquae CGMCC 1.12159
[33]
B
Photobacterium galatheae S2753
[15]
cas3
cse1
cse2
cse4
cas5
cse3
cas1
cas2
[17]
Photobacterium angustum ATCC 25915
Photobacterium angustum ATCC 33975
Photobacterium damselae subsp. damselae ATCC 33539
[8]
Photobacterium angustum S14
[45]
Photobacterium sp. SKA34
Photobacterium leiognathi Irivu.4.1.
[4]
C
Photobacterium marinum AK15
[21]
cas3
cas5d
csd1
csd2
cas4b
cas1
cas2
[30]
D
Photobacterium profundum SS9
cas1
cas3
csy4
[26]
E
Photobacterium profundum SS9
cas3
cse3
cse2
cse4
cas5
cse1
cas1
cas2
[64]
F
Photobacterium marinum AK15
cas1
cas2
[13]
Photobacterium aphoticum DSM 25995
[10]
[55]
Photobacterium aphoticum JCM 19237
[7]
[8]
G
Photobacterium galatheae S2753
Tn7
integrase
transposition protein
csy2
csy3
csy4
Tn7
Photobacterium aquae CGMCC 1.12159
[3]
Photobacterium angustum S14
Photobacterium ganghwense DSM 22954

<<

Figure 8 – Different CRISPR/Cas subtypes identified across the *Photobacterium* genomes. CRIPR/Cas similar to Yersinia pestis (A), *Escherichia coli* (B) and *Desulfivibrio vulgaris* (C), and other new gene organizations (D – F).

The Direct Repeats (DR) that flanked the protospacers were extracted and compared. The similarity of the DRs correlated with the type of clusters the strain had. *P. aphoticum* JCM 19237 and *P. aphoticum* DSM 25995 had the same DR, so did *P. iliopiscarium* ATCC 51760, *P. iliopiscarium* ATCC 51761, *P. aquae* CGMCC 1.12159 and *P. profundum* SS9. *P. angustum* ATCC 25915, *P. damselae* subsp. *damselae* ATCC 33539 (actually *P. angustum*) and *P. angustum* ATCC 33975 had also high similarity DRs (> 95% Identity), which were also similar (95% < Identity > 86%) to the ones from *P. profundum* SS9, *P. galatheae* S2753, *P. angustum* S14 and *P. leiognathi* lrivu.4.1.

The protospacers are short sequences that are derived from bacteriophages or other foreign DNA, such as conjugative plasmids [70,72]. Therefore, these sequences may provide a history of encounters of a specific bacterium with phages and/or plasmids. The number of spacers varied considerable between different strains, from 1 to 64 protospacers in the same CRISPR array (Figure 8). The 4 spacers in *P. leiognathi* lrivu.4.1 were 100% identical to the first 4 spacers in *P. angustum* S14, which had an array of 45 spacers. Also high similarity (> 93% Identity) could be identified in the first spacers of *P. angustum* ATCC 25915, *P. damselae* subsp. *damselae* ATCC 33539 (actually *P.*

angustum) and *P. angustum* ATCC 33975, although for the *P. angustum* strains the protospacers have not been identified just downstream of the *cas* operon, but elsewhere. We have also compared the protospacers to the prophage sequences identified in these genomes. Interestingly for *P. angustum* ATCC 33975 an array of 9 sequential spacers were 100% identical to an intact prophage sequence identified within the same genome.

Secondary metabolism

Although *Photobacterium* strains have mostly been described due to their association with marine animals, recently their potential in drug discovery and other applications have been reported [28,33,34,73–76]. Using antiSMASH 3.0 [53], different biosynthetic gene clusters were identified (Table S2) with some clusters being present across the genus and others being species specific (Figure 9). The same terpene biosynthetic cluster could be identified in all the *P. angustum* strains (including the mis-identified *P. damselae* subsp. *damselae* ATCC 33539, actually a *P. angustum*). Another species-specific cluster was the polyunsaturated fatty acid cluster present in both *P. profundum* strains.

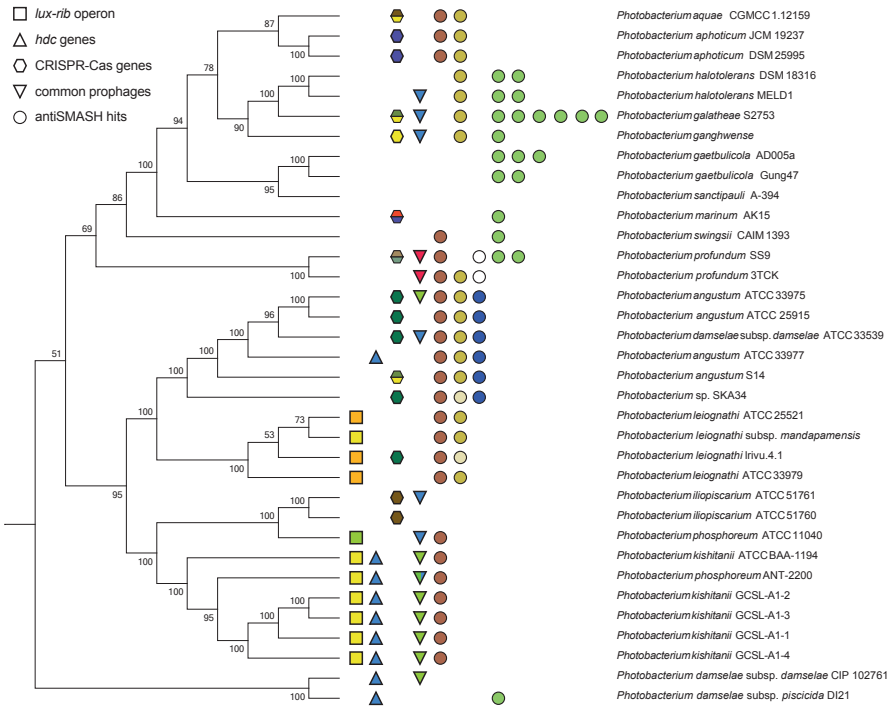


Figure 9 – Summary figure of some of the shared genomic traits between the *Photobacterium* strains. The phylogenetic tree is based on the MLSA phylogenetic distances. Squares represent the lux-rib operon, color-coded for the different gene cluster architectures. Triangles indicate the presence of genes coding for histamine producing proteins. Hexagons indicate the presence of CRISPR-Cas loci, color-coded for the different gene cluster architectures. Inverted triangles indicate the common prophages, each different also color-coded. Circles represent antiSMASH hits: brown for aryl-polyene biosynthetic gene cluster; yellow for ectoine biosynthetic gene cluster; blue for terpene biosynthetic gene cluster; white for polyunsaturated fatty acid biosynthesis gene cluster; and green for NRPS biosynthetic gene clusters.

Other biosynthetic clusters such as for siderophore, aryl-polyene and ectoine were widely distributed across the genus (Figure 9). A siderophore cluster was present in 11 strains, although two distinct

siderophore biosynthetic clusters were identified (Figure S2). The ectoine cluster was present in 18 strains (two of these had two out of three genes needed for its biosynthesis) (Figure 9 – yellow circles). Finally, the most widely distributed cluster was the aryl-polyene biosynthetic cluster, which was identified in 23 out of the 35 strains (Figure 9 – brown circles).

Plasmids and virulence genes

Most *Photobacterium* species have been described as symbiotic or associated with other marine organisms [21,26,35,39,77]. Furthermore, strains of *P. damselae* are pathogens of marine organisms, especially fish [78,79]. The genes *hlyA* and *dly* are key virulence genes of *P. damelae* [55–57], and we therefore conducted an homology search to evaluate the possible virulence of other *Photobacterium* species. The genes could only be identified in *P. damselae* subsp. *damselae* CIP 102761. In this strain, two copies of the *hlyA* gene were identified, one in contig_1, close to an IS4 transposase and a phage integrase and the other in contig_4, next to the *dly* gene.

Often draft whole genome sequences contain plasmid sequences. Plasmids are important mediators of several physiological traits of *Photobacterium*, such as virulence, drug resistance and biosynthetic capabilities [55,80–82]. We compared known *Photobacterium* plasmids to the here studied genomes. Plasmid pPHDD1 showed high similarity to contig_4 of *P. damselae* subsp. *damselae* CIP 102761 genome. Some of the contigs of *P. damselae* subsp. *piscicida* DI21 also had high

similarity to plasmids pPHDP60, pPHDP10 and pPHDP70, which have been previously isolated from this strain [80,83].

Bioluminescence

One of the traits of the *Photobacterium* genus is, as the name suggests, the fact that it used to be composed only of luminous bacteria [21]. However, this is not true for all species of the genus, but this trait and the *lux* genes have been used in the general classification of the different species and strains [21,35,36]. We identified the *lux-rib* operon and compared it between the strains. The *lux-rib* operon was only found in *P. kishitanii*, *P. phosphoreum*, and *P. leiognathi* (Figure 9). In all the strains of *P. kishitanii*, *P. leiognathi* subsp. *mandapamensis* svers.1.1., and *P. phosphoreum* ANT-2200, the genes *luxF* and *ribE* were present (Figure 10A). *P. phospohoreum* ATCC 11040 had no *ribE* encoded in the operon (Figure 10B), while all the other strains of *P. leiognathi* (ATCC 25521, ATCC 33979 and Irivu 4.1) carried neither *ribE* nor *luxF* genes (Figure 10C). Although it had been suggested before that some strains of *P. angunstum* are luminous [21], here none of the six *P. angustum* strains (S14, ATCC 25915, ATCC 33977, ATCC 33975, SKA34 and *P. dameselae* subsp. *damselae* ATCC 33539) encoded the genes responsible for luminescence. An extra set of *rib* genes (*ribDEBH*) could be identified elsewhere in the genome for all the 35 strains (data not shown).

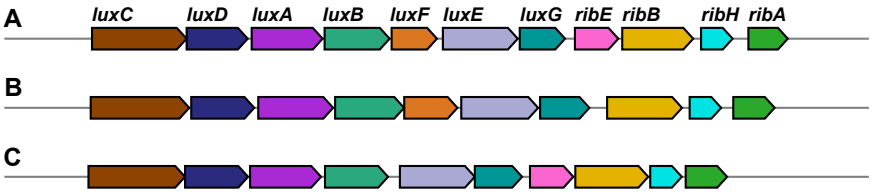


Figure 10 – Different organizations of the *lux-rib* operon in *Photobacterium*. *P. kishitanii*, *P. leiognathi* subsp. *mandapamensis* svers.1.1., and *P. phosphoreum* ANT-2200 (A); *P. phospohoreum* ATCC 11040 (B); *P. leiognathi* ATCC 25521, ATCC 33979 and Irivu 4.1 (C).

Histamine production

Ingestion of fish containing high levels of histamine is the most frequent cause of fish poisoning incidents within the United States [84]. This is denominated scombrotoxin fish poisoning. Recently, high-histamine producing *Photobacterium* strains have been isolated from freshly caught fish [38], raising awareness for the need to revise food safety rules regarding sea food. The histamine production relies on histidine decarboxylase enzyme (Hdc). We searched for histamine producing genes, previously identified in *Photobacterium* strains [38]. Genes could be identified in all *P. kishitanii* strains, *P. angustum* ATCC 33977, *P. phosphoreum* ANT-220 (actually *P. kishitanii*), *P. damselae* subsp. *damselae* ATCC 33539 (actually *P. angustum*) and *P. damselae* subsp. *damselae* CIP 102761 (Figure 9 and Figure S1).

Discussion

This study aimed at understanding the genomic diversity of the *Photobacterium* genus, the drives for its evolution and impact on ecological strategies by analyzing whole genome sequences of 35 strains representing 16 different species.

The genomes revealed a 10% GC variation within the genus, with the formation of two distinct GC% groups. The GC content has been previously correlated to several environmental conditions as well as to the amino acid usage [85]. The strains with higher GC content clustered together in the amino acid usage analysis, creating two different clusters, the high and low GC clusters. The genome size range was in accordance with the ones reported for other genera from the *Vibrionaceae* family [86]. The largest genomes identified belonged to the *P. profundum* SS9, which has been shown to be an environmentally extremely versatile species. It is able to grow at cold temperatures, down to 0 °C, and in pressurized environments from 0.1 to 70 MPa [87–89]. It seems therefore logical that such environmental versatility is associated with a larger genome size [90]. It was previously shown that genome size and GC content are related with the ecological strategies carried by the different marine bacteria, with free-living bacteria having lower GC and smaller genomes, when compared to patch-associated bacteria [91]; and also that symbionts, parasites and commensals would experience genome reduction due to specialization [92,93]. Here we observe that the smallest genomes are indeed among the known symbiotic (*P. iliopiscarium*, *P. damselae*, *P. phosphoreum*), but

also *P. galatheae* and *P. halotolerans* had small genomes. An interesting observation is that symbiotic species presented lower GC content than the free-living or patch associated ones.

Phylogeny

The poor phylogenetic discriminatory power of the 16S rRNA gene in the *Vibrionaceae* family was previously reported [61,86]. Here, we used several clustering methods to investigate the phylogenetic relationships within the *Photobacterium* genus. Different approaches led to distinct phylogenetic results (MLSA, *fur* and amino acid usage), nevertheless the core groups were the same (Figure 3). Independently of the method used, the species with higher GC content clustered together, suggesting an evolutionary correlation of GC content and phylogenetic proximity within the genus.

We suggest the use of MLSA or *fur* phylogenetic analysis in the future characterization of new isolates at the species level. Interestingly, the *fur* gene of approximately 450 bp had a better correlation to the ANI than the MLSA, which used approximately 6,800 bp more in the analysis. This observation strengthens the previously suggestion of the use of the *fur* gene as a single phylogenetic marker in the *Vibrionaceae* family [61].

The phylogenetic analysis showed a misidentification of two isolates, *P. phosphoreum* ANT-2200 and *P. damselae* subsp. *damselae* ATCC 33539. The *P. phosphoreum* ANT-2200 isolate is definitely a *P.*

kishitanii, which was previously suggested by others [21,36,38], and we confirmed this in the phylogenetic assessment, but also throughout the different genomic features studied. The genome identified as belonging to *P. damselae* subsp. *damselae* ATCC 33539 was similar to the ones of *P. angustum*, but the 16S rRNA phylogeny using the original sequence placed this strain elsewhere. Since no other physiological discrepancies have been previously reported, we believe that the case of *P. damselae* subsp. *damselae* ATCC 33539 might be a wrong whole genome sequence submission to NCBI or the sequencing of the wrong strain; this mistake should nevertheless be fixed, since it might affect future studies.

It has been suggested to divide the *Photobacterium* genus into two or three clades [21,24]. The most recent review on this genus suggested two clades, which would divide the species into clade 1 (*P. angustum*, *P. leiognathi*, *P. aquimaris*, *P. kishitanii*, *P. phosphoreum* and *P. iliopiscarium*) and clade 2 (*P. aplysiae*, *P. lipolyticum*, *P. profundum*, *P. fridigiphilum*, *P. indicum*, *P. damselae*, *P. jeanii*, *P. ganghwense*, *P. halotolerans*, *P. gaetbulicola*, *P. lutimaris* and *P. rosenbergii*) [21]. This suggestion was based both on the molecular phylogenetic evaluation, but also on the creation of a luminous/symbiotic and a non-luminous/non-symbiotic cluster. Our results support the phylogenetic relationship previously observed between the species of each cluster, but the identification of genes responsible for luminescence suggest that *P. angustum* and *P. damselae* are not luminescent species as previously reported [21]. *P. damselae* has acquired *lux* genes through

horizontal gene transfer [49], which might explain the absence of these genes in the studied strains. On the other hand, luminescence of *P. angustum* strains may vary [21], although we could not identify *lux* genes in any of the six *P. angustum* strains (ATCC 25915, ATCC 33975, ATCC 33977, S14, SKA34, and the misidentified *P. damselae* ATCC 33539). Furthermore, *P. ganghwense* has been defined as being bioluminescent, however, it has not been described how this feature was evaluated in the original study [59], nor could we identify *lux* genes in the genome of the type-strain. Therefore, the generic division of luminous versus non-luminous clades should be avoided. Analysis of the *lux-rib* operon further showed that strains ATCC 25521, ATCC 33979 and Irivu.4.1 are most likely *P. leiognathi* subsp. *leiognathi*, according to their *lux-rib* gene organization [36].

Genomic diversity

We identified a pan-genome of 28,951 genes and a core-genome of 1,232 genes using 35 genomes (Figure 5). The pan-genome for this genus alone is greater than what has been reported in a study of the *Vibrionaceae* family, where 32 genomes of a total of 13 species from three different genera were used [86]. This suggests a highly genomic diverse genus, matching their ability to colonize diverse environmental niches [90,94] and supports the theory that high gene content variation exists in environmental marine strains [66,94]. The number of core-genome genes is in agreement to the observed in other marine *Gammaproteobacteria* [95], although this is two times

higher than what has been shown for the *Vibrio* genus, where the core-genome comprised approximately 500 genes [86]. This difference can be explained by the high number of *Vibrio* species (more than 120), which reflects the genomic and ecological diversity of that genus; the higher the number of species, the higher possibility of a species missing a core-gene.

The gene content of thirty whole genome sequences was compared to the reference genomes of *P. profundum* SS9 and *P. gaetbulicola* Gung47 (Figure 6). The large chromosome seems to be more conserved between species than the smaller chromosome or plasmids, as previously described for other members of the *Vibrionaceae* family [86]. The smaller chromosomes appear to be very variable and even differences between strains of the same species can be observed (Figure 7).

Genome evolution

It has been several times suggested that horizontal gene transfer was probably the cause and means of genomic evolution in the *Vibrionaceae* family [20,21,67,69], but not many studies have been performed to address this proposal. In the *Photobacterium* genus, studies on horizontal gene transfer are limited to the *lux-rib* operon and a *chitinase A* gene [49,96]. Horizontal gene transfer can occur by

uptake of environmental DNA, conjugative plasmids and bacteriophage infection.

Identification of genomic islands in the closed genomes of the reference strains showed their participation in the genomic variation of strains belonging to the same species (Figure 7). Phage DNA has been suggested as one of the main sources for lateral gene transfer in bacteria, example in the *Vibrionaceae* family is the most important virulence factor in *V. cholera*, encoded in the CTXϕ phage integrated in the large chromosome [18,20]. We identified 32 intact and 59 incomplete prophage sequences, suggesting an intense exchange of genomic material, with an average of 2.6 prophages per genome. These prophages were also distinct; from the 20 different intact prophages identified in the 35 studied genomes, only 3 of them had a reoccurrence in a second strain.

The bacterial immune system (CRISPR-Cas) allows the protection against re-infection by the same bacteriophage [72]. The system uses sequence homology to identify and neutralize possible threats. Several *cas* gene arrays have been described in several bacterial species [71]. Interestingly, the *cas* operons previously associated with *Escherichia coli*, *Desulvibrio vulgaris* and *Yersinia pestis* were identified in strains of the *Photobacterium* genus (Figure 8). From the seven clusters identified only two (Figure 8A and 8F) seem to have a correlation with species, since presence/absence and type of cluster seems randomly spread across the different species (Figure 10).

The estimation of direct repeats (DR) and protospacers in the CRISPR arrays is very difficult in draft genomes due to the short sequencing reads and the repetitive nature of the sequences, which make them difficult to correctly assemble. Nevertheless, several CRISPR arrays have been identified in the studied genomes, some of them containing sequences from prophages encoded in the same genome, suggesting multiple phage infections throughout evolution. Arrays with extensive number of protospacers in some strains indicate innumerous bacteriophage infections.

Another evidence of high genomic exchange is in the number and distribution of secondary metabolism biosynthetic clusters (Figure 9). The ectoine and aryl-polyene clusters seem to be present in two and three branches of the phylogenetic tree, respectively. This suggests various events of gain and loss of these clusters throughout the evolution of *Photobacterium* species. On the other hand, terpene and poly-unsaturated fatty acid clusters are species specific, *P. angustum* and *P. profundum*, respectively. These must have been acquired in the speciation process and maintained due to physiological advantages.

Other functions

Some phenotypes of interest have been observed and further studied in *Photobacterium*. These include bioluminescence, virulence and histamine production [21,35,36,38,78,79]. We confirmed that genes encoding

virulence traits are not commonly present in *Photobacterium* genomes and that these are most likely acquired from the environment by *e.g.* plasmid uptake [80].

On the other hand, bioluminescence was initial thought to be a widespread feature of the genus [21]. We identified the genetic basis for bioluminescence only in three species *P. kishitanii*, *P. phosphoreum*, and *P. leiognathi*. This correlated with phylogenetic analysis, with the exception of *P. iliopiscarium*, that might have lost this trait due to niche adaptation. Although it had been suggested before that some strains of *P. angunstum* are luminous [21], here none of the six *P. angustum* strains (S14, ATCC 25915, ATCC 33977, ATCC 33975, SKA34 and *P. dameselae* subsp. *damselae* ATCC 33539) encoded the genes responsible for luminescence.

Histamine production seems to be specific for *P. kishitanii* and *P. damselae*, however genes responsible for this feature could be identified also in one *P. angustum* strain (ATCC 33977). The same has been observed for the type IV secretion systems in *Pseudovibrio* and attributed to the frequency with which those genes are horizontally acquired [99,100]. There are also several reports describing *P. phosphoreum* as histamine producing species [97,98]. In this study we only had two *P. phosphoreum* genomes and one was a mis-identified *P. kishitanii*. In the P. phosphoreum genome we did not identify the genes responsible for histamine production, but it might be that it is present in some *P. phosphoreum* strains. This once again reveals the

plasticity of the genomic information in this genus, which has an impact on the observed phenotypes.

Conclusion

We have here shown the genetic diversity harbored in the extremely versatile genus *Photobacterium*. We confirmed the non-pathogenic nature of these genus, and increase the awareness to genetic material exchange happening in this genus, which might lead to acquisition of virulence factors, as in the case of *P. damelae*. This genus seems to be extremely versatile in environments colonized, possibly due to the genetic diversity observed. Furthermore, we present evidence for the exchange of genetic material at high levels, reflected in the variability of the smaller chromosome, in the number of protospacers, prophages and genomic islands, and in the number of secondary metabolite clusters. We have also shown the potential of some of the species for the production of bioactive molecules, considering their secondary metabolite cluster richness.

References

1. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95,** 6578–6583 (1998).

2. Wietz, M., Mansson, M., Vynne, N. G. & Gram, L. *Marine Microbiology*. *Marine Microbiology : Bioactive Compounds and Biotechnological Applications* 127–159 (Wiley-VCH Verlag GmbH & Co. KGaA, 2013). doi:10.1002/9783527665259

3. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **14,** 111–129 (2015).

4. Jensen, P. R., Chavarria, K. L., Fenical, W., Moore, B. S. & Ziemert, N. Challenges and triumphs to genomics-based natural product discovery. *J. Ind. Microbiol. Biotechnol.* **41,** 203–9 (2014).

5. Mai-prochnow, A. *et al.* Biofilm Development and Cell Death in the Marine Bacterium Pseudoalteromonas tunicata Biofilm Development and Cell Death in the Marine Bacterium Pseudoalteromonas tunicata. (2004). doi:10.1128/AEM.70.6.3232

6. Mitra, S. & Mukherjee, J. Biofilms of Marine and Intertidal Microbes : Ecology and Biotechnology. *Adv Biochem Eng Biotechnol* **123,** 127–141 (2013).

7. Mardis, E. R. Next-Generation Sequencing Platforms. *Annu. Rev. Anal. Chem.* **6,** 287–303 (2013).

8. Loman, N. J. & Pallen, M. J. Twenty years of bacterial genome sequencing. *Nat. Rev. Microbiol.* **13,** 787–794 (2015).

9. Lasken, R. S. Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10,** 631–640 (2012).

10. Choi, E. J. *et al.* Previously Uncultured Marine Bacteria Linked to Novel Alkaloid Production. *Chem. Biol.* **22,** 1270–1279 (2015).

11. Zengler, K. *et al.* Cultivating the uncultured. *Proc. Natl. Acad. Sci.* **99,** 15681–15686 (2002).

12. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science (80-.).* **348,** 1–10 (2015).

13. Wietz, M., Gram, L., Jørgensen, B. & Schramm, a. Latitudinal patterns in the abundance of major marine bacterioplankton groups. *Aquat. Microb. Ecol.* **61,** 179–189 (2010).

14. Gilbert, J. A. *et al.* Defining seasonal marine microbial community dynamics. *ISME J.* **6,** 298–308 (2012).

15. Meibom, K. L. *et al.* The Vibrio cholerae chitin utilization program. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 2524–9 (2004).

16. Papenfort, K., Förstner, K. U., Cong, J.-P., Sharma, C. M. & Bassler, B. L. Differential RNA-seq of Vibrio cholerae identifies the VqmR small RNA as a regulator of biofilm formation. *Proc. Natl. Acad. Sci.* 201500203 (2015). doi:10.1073/pnas.1500203112

17. Stauder, M. *et al.* Role of GbpA protein, an important virulence-related colonization factor, for Vibrio cholerae's survival in the aquatic environment. *Environ. Microbiol. Rep.* **4,** 439–45 (2012).

18. Kim, E. J., Lee, C. H., Nair, G. B. & Kim, D. W. Whole-genome sequence comparisons reveal the evolution of Vibrio cholerae O1. *Trends Microbiol.* 1–11 (2015). doi:10.1016/j.tim.2015.03.010

19. Rajpara, N. *et al.* Antimicrobials , Resistance and Chemotherapy Role of integrons , plasmids and SXT elements in multidrug resistance of Vibrio cholerae and Providencia vermicola obtained from a clinical isolate of diarrhea Article type : Received on : Accepted on : Frontie. *Front. Microbiol.* (2015). doi:10.3389/fmicb.2015.00057

20. Vitulo, N., Vezzi, A., Romualdi, C., Campanaro, S. & Valle, G. A global gene evolution analysis on Vibrionaceae family using phylogenetic profile. *BMC Bioinformatics* **8 Suppl 1,** S23 (2007).

21. Urbanczyk, H., Ast, J. C. & Dunlap, P. V. Phylogeny, genomics, and symbiosis of Photobacterium. *FEMS Microbiology Reviews* **35,** 324–342 (2011).

22. Liu, Y. *et al.* Photobacterium aquae sp. nov., isolated from a recirculating mariculture system. *Int. J. Syst. Evol. Microbiol.* **64,** 475–480 (2014).

23. Moreira, A. P. B. *et al.* Photobacterium sanctipauli sp. nov. isolated from bleached Madracis decactis (Scleractinia) in the St Peter & St Paul Archipelago, Mid-Atlantic Ridge, Brazil. *PeerJ* **2,** e427 (2014).

24. Lucena, T. *et al.* Photobacterium aphoticum sp. nov., isolated from coastal water. *Int. J. Syst. Evol. Microbiol.* **61,** 1579–1584 (2011).

25. Lo, N., Jin, H. M. & Jeon, C. O. Photobacterium aestuarii sp. nov., a marine bacterium isolated from a tidal flat. *Int. J. Syst. Evol. Microbiol.* **64,** 625–630 (2014).

26. Gomez-Gil, B. *et al.* Photobacterium swingsii sp. nov., isolated from marine organisms. *Int J Syst Evol Microbiol* **61,** 315–319 (2011).

27. Figge, M. J. *et al.* Photobacterium piscicola sp. nov., isolated from marine fish and spoiled packed cod. *Syst. Appl. Microbiol.* **37,** 329–335 (2014).

28. Machado, H., Giubergia, S., Mateiu, R. V. & Gram, L. Photobacterium galatheae sp. nov., a bioactive bacterium isolated from a mussel in the Solomon Sea. *Int. J. Syst. Evol. Microbiol.* **65,** 4503–4507 (2015).

29. Srinivas, T. N. R., Vijaya Bhaskar, Y., Bhumika, V. & Anil Kumar, P. Photobacterium marinum sp. nov., a marine bacterium isolated from a sediment sample from Palk Bay, India. *Syst. Appl. Microbiol.* **36,** 160–5 (2013).

30. Kim, B. C. *et al.* Photobacterium atrarenae sp. nov. a novel bacterium isolated from sea sand. *Curr. Microbiol.* **63,** 433–438 (2011).

31. Holmes, B. & Farmer, J. J. International Committee on Systematics of Prokaryotes; Subcommittee on the taxonomy of Aeromonadaceae, Vibrionaceae and related organisms: Minutes of the meetings, 6 August 2008, Istanbul, Turkey. *Int. J. Syst. Evol. Microbiol.* **59,** 2638–2640 (2008).

32. Seo, H. J., Bae, S. S., Yang, S. H., Lee, J. H. & Kim, S. J. Photobacterium aplysiae sp. nov., a lipolytic marine bacterium isolated from eggs of the sea hare Aplysia kurodai. *Int. J. Syst. Evol. Microbiol.* **55,** 2293–2296 (2005).

33. Mathew, D. C., Mathew, M., Gicana, G. & Huang, C. Genome Sequence of Photobacterium halotolerans MELD1 , with Mercury Reductase (merA), Isolated from Phragmites australis. *Genome Announc.* **3,** 12204 (2015).

34. Mathew, D. C. *et al.* A rhizosphere-associated symbiont, Photobacterium spp. strain MELD1, and its targeted synergistic activity for phytoprotection against mercury. *PLoS One* **10,** 1–18 (2015).

35. Dunlap, P. V. in *Encyclopedia of Microbiology (Third Edition)* 45–61 (2009). doi:http://dx.doi.org/10.1016/B978-012373944-5.00066-3

36. Ast, J. C. & Dunlap, P. V. Phylogenetic analysis of the lux operon distinguishes two evolutionarily distinct clades of Photobacterium leiognathi. *Arch. Microbiol.* **181,** 352–361 (2004).

37. Gomez-Gil, B. *et al.* Photobacterium swingsii sp. nov., isolated from marine organisms. *Int. J. Syst. Evol. Microbiol.* **61,** 315–319 (2011).

38. Bjornsdottir-Butler, K., McCarthy, S., Dunlap, P. V. & Benner, R. a. *Photobacterium angustum* and *Photobacterium kishitanii* : Psychrotrophic High Histamine-Producing Bacteria Indigenous to Tuna. *Appl. Environ. Microbiol.* AEM.02833–15 (2016). doi:10.1128/AEM.02833-15

39. Ast, J. C. *et al.* Photobacterium kishitanii sp. nov., a luminous marine bacterium symbiotic with deep-sea fishes. *Int. J. Syst. Evol. Microbiol.* **57,** 2073–2078 (2007).

40. Nogi, Y., Masui, N. & Kato, C. Photobacterium profundum sp. nov., a new, moderately barophilic bacterial species isolated from a deep-sea sediment. *Extremophiles* **2,** 1–7 (1998).

41. Gram, L., Melchiorsen, J. & Bruhn, J. B. Antibacterial activity of marine culturable bacteria collected from a global sampling of ocean surface waters and surface swabs of marine organisms. *Mar. Biotechnol. (NY).* **12,** 439–51 (2010).

42. Vesth, T., Lagesen, K., Acar, Ö. & Ussery, D. CMG-Biotools, a Free Workbench for Basic Comparative Microbial Genomics. *PLoS One* **8,** (2013).

43. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9,** 75 (2008).

44. Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42,** D206–14 (2014).

45. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–9 (2013).

46. Auch, A. F., von Jan, M., Klenk, H.-P. & Göker, M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* **2,** 117–34 (2010).

47. Goris, J. *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* **57,** 81–91 (2007).

48. Urbanczyk, H., Furukawa, T., Yamamoto, Y. & Dunlap, P. V. Natural replacement of vertically inherited lux-rib genes of Photobacterium aquimaris by horizontally acquired homologues. *Environ. Microbiol. Rep.* **4,** 412–416 (2012).

49. Urbanczyk, H., Ast, J. C., Kaeding, A. J., Oliver, J. D. & Dunlap, P. V. Phylogenetic analysis of the incidence of lux gene horizontal transfer in Vibrionaceae. *J. Bacteriol.* **190,** 3494–3504 (2008).

50. Medema, M. H., Takano, E. & Breitling, R. Detecting sequence homology at the gene cluster level with multigeneblast. *Mol. Biol. Evol.* **30,** 1218–1223 (2013).

51. Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: A Fast Phage Search Tool. *Nucleic Acids Res.* **39,** 347–352 (2011).

52. Dhillon, B. K. *et al.* IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Res.* **43,** W104–8 (2015).

53. Weber, T. *et al.* antiSMASH 3.0--a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* 1–7 (2015). doi:10.1093/nar/gkv437

54. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **35,** 52–57 (2007).

55. Rivas, A. J., Balado, M., Lemos, M. L. & Osorio, C. R. The Photobacterium damselae subsp. damselae hemolysins damselysin and HlyA are encoded within a new virulence plasmid. *Infect. Immun.* **79,** 4617–4627 (2011).

56. Rivas, A. J., Balado, M., Lemos, M. L. & Osorio, C. R. Synergistic and additive effects of chromosomal and plasmid-encoded hemolysins contribute to hemolysis and virulence in Photobacterium damselae subsp. damselae. *Infect. Immun.* **81,** 3287–3299 (2013).

57. Roux, F. Le *et al.* The emergence of Vibrio pathogens in Europe: ecology, evolution, and pathogenesis (Paris, 11–12th March 2015). *Front. Microbiol.* **6,** 1–8 (2015).

58. Rivas, R., García-Fraile, P., Mateos, P. F., Martínez-Molina, E. & Velázquez, E. Photobacterium halotolerans sp. nov., isolated from Lake Martel in Spain. *Int. J. Syst. Evol. Microbiol.* **56,** 1067–71 (2006).

59. Park, Y. D. *et al.* Photobacterium ganghwense sp. nov., a halophilic bacterium isolated from sea water. *Int. J. Syst. Evol. Microbiol.* **56,** 745–749 (2006).

60. Sawabe, T., Kita-Tsukamoto, K. & Thompson, F. L. Inferring the evolutionary history of vibrios by means of multilocus sequence analysis. *J. Bacteriol.* **189,** 7932–6 (2007).

61. Machado, H. & Gram, L. The fur Gene as a New Phylogenetic Marker for Vibrionaceae Species Identification. *Appl. Environ. Microbiol.* **81,** 2745–2752 (2015).

62. Pascual, J., Macián, M. C., Arahal, D. R., Garay, E. & Pujalte, M. J. Multilocus sequence analysis of the central clade of the genus Vibrio by using the 16S rRNA, recA, pyrH, rpoD, gyrB, rctB and toxR genes. *Int. J. Syst. Evol. Microbiol.* **60,** 154–65 (2010).

63. Gabriel, M. W., Matsui, G. Y., Friedman, R. & Lovell, C. R. Optimization of multilocus sequence analysis for identification of species in the genus Vibrio. *Appl. Environ. Microbiol.* **80,** 5359–5365 (2014).

64. Thompson, F. L. *et al.* Phylogeny and Molecular Identification of Vibrios on the Basis of Multilocus Sequence Analysis Phylogeny and Molecular Identification of Vibrios on the Basis of Multilocus Sequence Analysis. *Appl. Environ. Microbiol.* **71,** 5107–5115 (2005).

65. Sawabe, T. *et al.* Updating the Vibrio clades defined by multilocus sequence phylogeny: proposal of eight new clades, and the description of Vibrio tritonius sp. nov. *Front. Microbiol.* **4,** 414 (2013).

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

67. Lilburn, T. G., Gu, J., Cai, H. & Wang, Y. Comparative genomics of the family Vibrionaceae reveals the wide distribution of genes encoding virulence-associated proteins. *BMC Genomics* **11,** 369 (2010).

68. Gu, J. *et al.* Genomic and systems evolution in Vibrionaceae species. *BMC Genomics* **10 Suppl 1,** S11 (2009).

69. Reen, F. J., Almagro-Moreno, S., Ussery, D. & Boyd, E. F. The genomic code: inferring Vibrionaceae niche specialization. *Nat. Rev. Microbiol.* **4,** 697–704 (2006).

70. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* 1–15 (2015). doi:10.1038/nrmicro3569

71. Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* **1,** 0474–0483 (2005).

72. Attar, N. Structural biology: How CRISPR captures spacer invaders. *Nat. Rev. Microbiol.* **13,** 15760 (2015).

73. Machado, H., Månsson, M. & Gram, L. Draft genome sequence of Photobacterium halotolerans S2753, producer of bioactive secondary metabolites. *Genome Announc.* **2,** 9–10 (2014).

74. Nielsen, A. *et al.* Solonamide B inhibits quorum sensing and reduces Staphylococcus aureus mediated killing of human neutrophils. *PLoS One* **9,** e84992 (2014).

75. Wietz, M., Mansson, M., Gotfredsen, C. H., Larsen, T. O. & Gram, L. Antibacterial compounds from marine Vibrionaceae isolated on a global expedition. *Mar. Drugs* **8,** 2946–60 (2010).

76. Machado, H., Sonnenschein, E. C., Melchiorsen, J. & Gram, L. Genome mining reveals unlocked bioactive potential of marine Gram-negative bacteria. *BMC Genomics* **16,** 1365 (2015).

77. Urbanczyk, H., Urbanczyk, Y., Hayashi, T. & Ogura, Y. Diversification of two lineages of symbiotic photobacterium. *PLoS One* **8,** 1–9 (2013).

78. Andreoni, F. & Magnani, M. Photobacteriosis: Prevention and diagnosis. *J. Immunol. Res.* **2014,** (2014).

79. Hundenborn, J., Thurig, S., Kommerell, M., Haag, H. & Nolte, O. Severe Wound Infection with Photobacterium damselae ssp. damselae and Vibrio harveyi, following a Laceration Injury in

Marine Environment: A Case Report and Review of the Literature. *Case Rep. Med.* **2013,** 610632 (2013).

80. Osorio, C. R. *et al.* A transmissible plasmid-borne pathogenicity island confers piscibactin biosynthesis in the fish pathogen Photobacterium damselae subsp. piscicida. *Appl. Environ. Microbiol.* **81,** 5867–5879 (2015).

81. Nonaka, L. *et al.* Novel Conjugative Transferable Multiple Drug Resistance Plasmid pAQU1 from Photobacterium damselae subsp. damselae Isolated from Marine Aquaculture Environment. *Microbes Environ.* **27,** 263–272 (2012).

82. Kim, M. J. *et al.* Complete DNA sequence and analysis of the transferable multiple-drug resistance plasmids (R plasmids) from Photobacterium damselae subsp. piscicida isolates collected in Japan and the United States. *Antimicrob. Agents Chemother.* **52,** 606–611 (2008).

83. Osorio, C. R. *et al.* Genomic and functional analysis of ICEPdaSpa1, a fish-pathogen-derived SXT-related integrating conjugative element that can mobilize a virulence plasmid. *J. Bacteriol.* **190,** 3353–3361 (2008).

84. Pennotti, R., Scallan, E., Backer, L., Thomas, J. & Angulo, F. J. Ciguatera and scombroid fish poisoning in the United States. *Foodborne Pathog. Dis.* **10,** 1059–66 (2013).

85. Lightfield, J., Fram, N. R. & Ely, B. Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One* **6,** (2011).

86. Thompson, C. C. *et al.* Genomic taxonomy of Vibrios. *BMC Evol. Biol.* **9,** 258 (2009).

87. Lauro, F. M. *et al.* Ecotype diversity and conversion in Photobacterium profundum strains. *PLoS One* **9,** 1–10 (2014).

88. Eloe, E. a., Lauro, F. M., Vogel, R. F. & Bartlett, D. H. The deep-sea bacterium Photobacterium profundum SS9 utilizes separate flagellar systems for swimming and swarming under high-pressure conditions. *Appl. Environ. Microbiol.* **74,** 6298–6305 (2008).

89. Lauro, F. M. *et al.* Large-scale transposon mutagenesis of Photobacterium profundum SS9 reveals new genetic loci important for growth at low temperature and high pressure. *J. Bacteriol.* **190,** 1699–1709 (2008).

90. Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 3160–5 (2004).

91. Luo, H. & Moran, M. A. How do divergent ecological strategies emerge among marine bacterioplankton lineages? *Trends Microbiol.* 1–8 (2015). doi:10.1016/j.tim.2015.05.004

92. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J.* **8,** 1–13 (2014).

93. Morris, J. J., Lenski, R. E. & Zinser, E. R. The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *MBio* **3,** (2012).

94. Konstantinidis, K. T., Braff, J., Karl, D. M. & DeLong, E. F. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific Subtropical Gyre. *Appl. Environ. Microbiol.* **75,** 5345–5355 (2009).

95. Qin, Q. L. *et al.* Comparative genomics of the marine bacterial genus Glaciecola reveals the high degree of genomic diversity and genomic characteristic for cold adaptation. *Environmental Microbiology* **16,** 1642–1653 (2013).

96. Hunt, D. E., Gevers, D., Vahora, N. M. & Polz, M. F. Conservation of the chitin utilization pathway in the Vibrionaceae. *Appl. Environ. Microbiol.* **74,** 44–51 (2008).

97. Kanki, M., Yoda, T., Ishibashi, M. & Tsukamoto, T. Photobacterium phosphoreum caused a histamine fish poisoning incident. *Int. J. Food Microbiol.* **92,** 79–87 (2004).

98. Kanki, M., Yoda, T., Tsukamoto, T. & Baba, E. Histidine decarboxylases and their role in accumulation of histamine in tuna and dried saury. *Appl. Environ. Microbiol.* **73,** 1467–1473 (2007).

99. Romano, S. *et al.* Comparative genomic analysis reveals a diverse repertoire of genes involved in prokaryote-eukaryote interactions within the Pseudovibrio Genus. *Front. Microbiol.* **7,** (2016).

100. Cascales, E. & Christie, P. J. The versatile bacterial type IV secretion systems. *Nat. Rev. Microbiol.* **1,** 137–49 (2003).

Supplementary Information

Table S1 – GenBank assession numbers for the 16S rRNA gene and MLSA phylogenies.

Strain	16S rRNA	ftsZ	gapA	gyrB	mreB	pyrH	recA	rpoA	topA
P. damselae subsp. damselae CIP 102761	AB032015	VDA_001180	VDA_001240	VDA_003250	VDA_003504	VDA_001360	VDA_001282	VDA_003464	VDA_002384
P. damselae subsp. piscicida DI21		RAST	RAST	RAST	RAST	RAST	RAST	RAST	RAST
P. damselae subsp. damselae ATCC 33539	NR_040831	UB36_13705	UB36_17335	UB36_18700	UB36_13420	UB36_20750	UB36_17065	UB36_20875	UB36_04120
Photobacterium. sp. SKA34	AF255623	SKA34_22402	SKA34_22052	SKA34_17333	SKA34_22729	SKA34_21349	SKA34_21755	SKA34_22977	SKA34_03419
Photobacterium sp. AK15 (marinum)	NR_133050.1	C942_04822	C942_01118	C942_02909	C942_04764	C942_03492	C942_01172	C942_02953	C942_01398
P. profundum 3TCK	DQ027054.1	P3TCK_24370	P3TCK_24736	P3TCK_22929	P3TCK_24070	P3TCK_25440	P3TCK_25016	P3TCK_23870	P3TCK_27904
P. profundum SS9	AB003191	PBPRA3211	PBPRA3132	PBPRA0011	PBPRA3270	PBPRA2966	PBPRA3068	PBPRA0345	PBPRA2477
P. leiognathi subsp. mandapamensis svers.1.1.	KC456610.1	RAST	RAST	RAST	RAST	RAST	RAST	RAST	RAST
P. leiognathi lrivu.4.1	NR_115541.1	RAST	RAST	RAST	RAST	RAST	RAST	RAST	RAST
P. leiognathi ATCC 25521	D25309.1	UB42_05010	UB42_10705	UB42_10125	UB42_05295	UB42_15885	UB42_10975	UB42_19320	UB42_03440
P. leiognathi ATCC 33979		UB34_03160	UB34_06890	UB34_12385	UB34_02875	UB34_16825	UB34_07160	UB34_19695	UB34_11980
P. angustum S14	AJ630163.2	VAS14_19911	VAS14_20231	VAS14_22357	VAS14_19626	VAS14_20966	VAS14_20496	VAS14_19336	VAS14_18414
P. angustum ATCC 25915	NR_119046.1	UB33_14340	UB33_16680	UB33_17120	UB33_14050	UB33_19385	UB33_16410	UB33_20160	UB33_12670
P. angustum ATCC 33977		UB35_14360	UB35_16015	UB35_16905	UB35_14070	UB35_19055	UB35_15750	UB35_20175	UB35_01635
P. angustum ATCC 33975	AY900628.1	UB39_10920	UB39_15480	UB39_18015	UB39_10630	UB39_17475	UB39_15210	UB39_20870	UB39_10245
P. halotolerans DSM 18316	NR_042975.1	RAST	RAST	RAST	RAST	RAST	RAST	RAST	RAST
P. galatheae S2753	KR704916	EA58_05035	EA58_04580	EA58_08470	EA58_05325	EA58_18525	EA58_04325	EA58_08130	
P. halotolerans MELD1	KC903134.1	KY46_10315	KY46_09915	KY46_18430	KY46_10600	KY46_18145	KY46_19655	KY46_21110	KY46_05665
P. phosphoreum ANT-2200	EU881910.1	PPBDW_I20087	PPBDW_I20168	PPBDW_I10065	PPBDW_I20029	PPBDW_I20304	PPBDW_I20220	PPBDW_I10338	PPBDW_I21735
P. phosphoreum ATCC 11040	NR_119047.1	UB41_05570	UB41_09335	UB41_11210	UB41_12100	UB41_17225	UB41_09585	UB41_19395	UB41_00885
P. gaetbulicola Gung47	NR_117271.1	H744_2c0707	H744_2c0784	H744_2c0322	H744_2c0648	H744_2c0925	H744_2c0841	H744_2c0600	H744_2c2703
P. gaetbulicola AD005a		RJ45_05475	RJ45_25915	RJ45_14550	RJ45_16640	RJ45_03805	RJ45_17415	RJ45_17760	RJ45_22920
P. sanctipauli A-394	NR_126301.1	RAST	RAST	RAST	RAST	RAST	RAST	RAST	RAST
P. swingsii CAIM 1393	NR_117351.1	AB733_13525	AB733_21045	AB733_18230	AB733_13235	AB733_17490	AB733_21315	AB733_22840	AB733_00535
P. ganghwense DSM 22954	NR_043295.1	ABT57_19625	ABT57_18865	ABT57_22530	ABT57_19910	ABT57_00220	ABT57_19135	ABT57_10545	ABT57_01030
P. aquae CGMCC 1.12159	NR_133815.1	ABT56_09255	ABT56_07995	ABT56_20835	ABT56_08970	ABT56_03810	ABT56_07730	ABT56_21155	ABT56_02290
P. kishitanii GCSL-A1-3		UA42_10690	UA42_16515	UA42_17650	UA42_10405	UA42_17110	UA42_16770	UA42_22285	UA42_03665
P. kishitanii GCSL-A1-2		UA40_10935	UA40_16310	UA40_17100	UA40_11220	UA40_20535	UA40_16565	UA40_22275	UA40_05810
P. kishitanii ATCC BAA-1194	NR_042852.1	UB40_07970	UB40_14090	UB40_16315	UB40_08255	UB40_19380	UB40_14340	UB40_20590	UB40_01490
P. kishitanii GCSL-A1-1		UA41_06365	UA41_15730	UA41_16960	UA41_06650	UA41_20485	UA41_15985	UA41_22160	UA41_03300
P. kishitanii GCSL-A1-4		UA38_09245	UA38_14910	UA38_15685	UA38_08960	UA38_19805	UA38_15165	UA38_21845	UA38_12435
P. iliopiscarium ATCC 51761	AY849432.1	UB37_10495	UB37_13050	UB37_09940	UB37_11510	UB37_11190	UB37_17965	UB37_18570	UB37_01590
P. iliopiscarium ATCC 51760	NR_111990.1	UB38_13970	UB38_10860	UB38_13175	UB38_12625	UB38_12205	UB38_10615	UB38_18020	UB38_01660
P. aphoticum JCM 19237 (C119)		JCM19237_2573	JCM19237_4425	JCM19237_6778	JCM19237_2485	JCM19237_880	JCM19237_4500	JCM19237_718	JCM19237_4151
P. aphoticum DSM 25995	FN796493.1	ABT58_02040	ABT58_19985	ABT58_01355	ABT58_02325	ABT58_18090	ABT58_19735	ABT58_21575	ABT58_14605
Vibrio pacinii DSM 19139	AJ316194	DQ907366.1	DQ907300.1	AB298264.1	KP635265.1	AJ842486.1	AJ580850.1	AJ842674.1	DQ907508.1

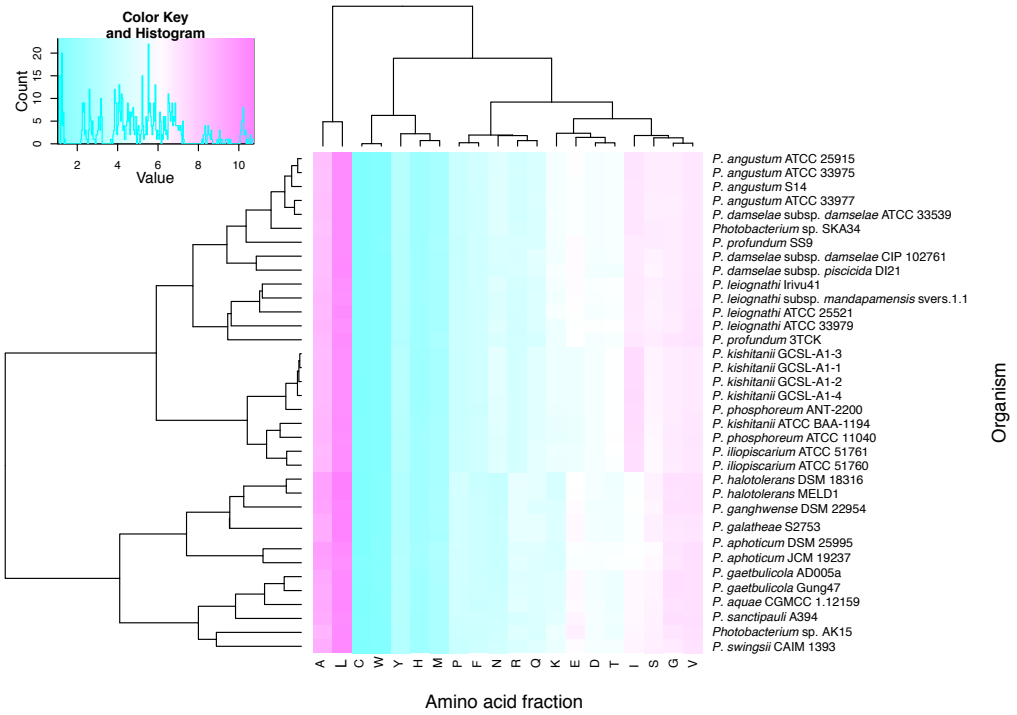


Figure S1 – Amino-acid usage in the *Photobacterium* strains.

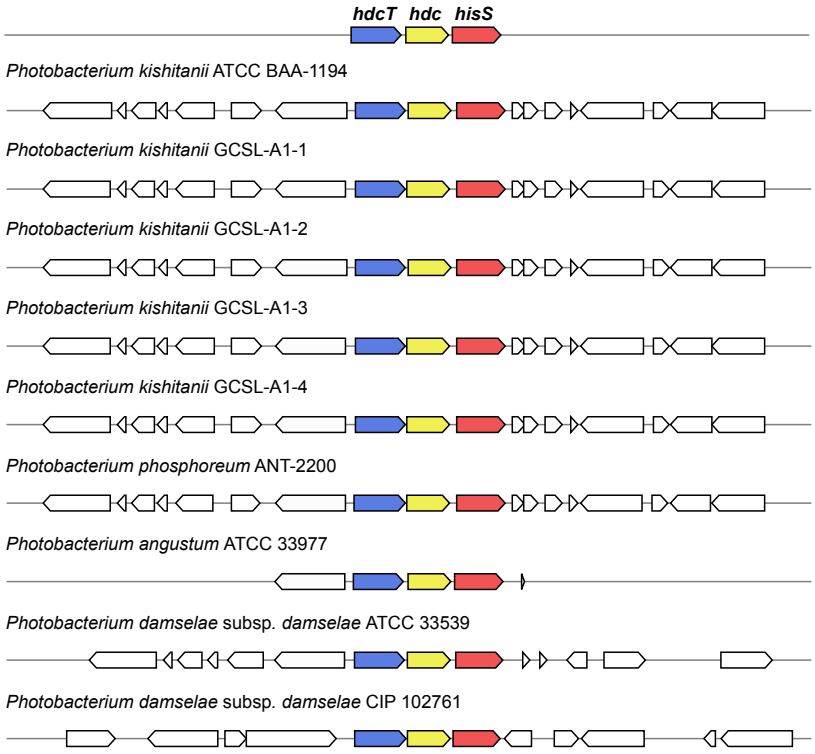


Figure S2 – Cluster encoding the proteins responsible for histamine production in *Photobacterium*.