UiT The Arctic University of Norway

Faculty of Biosciences, Fisheries and Economics, Department of Arctic and Marine Biology

# Hydrocarbon biodegradation potential in environmental bacterial metagenome

Alicia Caro Pascual

Master's thesis in Molecular Environmental Biology – BIO3950 – December 2020

# Table of Contents

# List of Abbreviations

**ABC transporter** ATP-binding cassette transporters

**ABS** Absorbance

**ATP** Adenosine triphosphate

**BLAST** Basic Local Alignment Search Tool

**BLASTn** Nucleotide BLAST

**BLASTx** Translated BLAST

**BP** Base Pair

**BS** Biosurfactant

**CG** Cytosine Guanine

**CO** Crude Oil

**CTAB** Cetyl Trimethyl Ammonium Bromide

**D** Diesel

**DNA** Deoxyribonucleic acid

**DUF** Domain of Unknown Function

**V** Volts

**HDB** Hydrocarbon-degrading bacteria

**KB** Kilobase pair

**LB** Luria Broth

**Mb** Megabase pair

**NCBI** National Center for Biotechnology Information

**NCS** Norwegian Continental Shelf

**ORF** Open Reading Frame

**PAHs** Polycyclic Aromatic Hydrocarbons

**PDB** Phage Dilution Buffer

**PTS transporter** Phosphotransferase system transporter

**RNA** Ribonucleic acid

**RPM** Revolutions Per Minute

**rRNA** Ribosomal RNA

**TCA cycle** TriCarboxylic Acid cycle

# List of Tables

# List of Figures

# Abstract

Hydrocarbon pollution in marine environments present an acute problem which is aggravated by cold temperatures. This is especially relevant in important environmental and economic northern regions such as the Barents Sea. The Barents Sea region has become the focus of oil industries from Russia and Norway, incrementing the risk of hydrocarbon pollution. Bioremediation is a cost-effective and environmentally sound method to remove hydrocarbon pollution. To study the bioremediation potential of native bacteria from a chronically oil polluted region of the Barents Sea a study was done in the Murmansk seaport (Kola Bay, 68°58′00″ N, 33°05′00″ E) analyzing the composition of the bacterial community. This present work aims to do a deeper study of that bacterial community through a metagenomic fosmid library, a selective hydrocarbon-rich medium and bioinformatic analysis. A metagenomic fosmid library was constructed with the environmental DNA, the fosmid clones were transformed in *Escherichia coli* cells and cultivated in minimal media with concentrations 0.05%, 0.5% and 1% of diesel or crude oil. Ten colonies were selected, sequenced, and  subsequently analyzed with the software Geneious 2020.2.2. BLAST searches and other bioinformatic tools were conducted in every Open Reading Frame of the 10 colonies, revealing new links between native bacteria and hydrocarbon degradation as well as promising enzymes for hydrocarbon bioremediation. This work sets the ground for further studies on functional metagenomic analyses with fosmid libraries and further studies with the novel bacterial species and enzymes linked to hydrocarbon biodegradation.

**Keywords:**

Barents Sea, hydrocarbon pollution, hydrocarbon bioremediation, metagenomic fosmid libraries, functional metagenomic analysis.

# Acknowledgements

First and foremost, I would like to thank my supervisors, John Jensen and Anton Liaimer. John for his infinite patience, understanding and encouragement and Anton for his clever insights and guidance. I apologize for the mess I am, thank you for not giving up on me. Also thanks to all the Microorganisms and Plants group for making this thesis possible, especially to Jeanette, Aslak and Mike for their company and friendship.

I would also like to thank my family, gracias a mi familia, especialmente a mis padres, Isabel y Juan, por permitirme cumplir mis sueños y apoyarme siempre, aunque eso signifique no poder vernos tan a menudo como nos gustaría. Esta tesis no hubiera sido posible sin vosotros. Y a Reme por nunca dejar que me olvide de ella, ni siquiera un segundo.

For the person who was there from the very beginning to the end, Bettina, thank you for being the best introduction to Norway and my constant rock and support. Gracias Sandra, por todas las risas, consejos y apoyo, mi pequeña España en Noruega. And thanks to my Stakkevollan people for all the laughs and card games  Ben, Rosa, Greg, Caro, Lisa, Eileen, Niki and Bert.

Gracias Ane por estar tan cerca aunque estés tan lejos y no dejar que nunca me sienta sola, no podría desear una amistad mejor que la nuestra. Y por supuesto Victor y Maikel, gracias por hacerme reír sin parar y perdonarme por estar tan ausente cuando tenía la cabeza en la tesis. Clem thank you for teaching me how to be a better person and all your advice about thesis and life. Y Sergi, gracias.

And lastly, thanks to Anders for all the love, the laughs and taking care of me when I couldn’t. You inspire me to become a better person every day.

# 1 Introduction

## 1.1 Barents Sea and the petroleum industry

Marine environments are a source of biodiversity and economic resources. From fisheries to tourism, many industries depend on a healthy marine ecosystem. This is especially true in countries such as Norway and Russia, where important economic resources, such as petroleum and fisheries, are based in marine environments. One of the most unique marine environments in the Arctic is the Barents Sea. The Barents Sea is part of the Arctic Ocean, located between the Norwegian Sea and the coast of Novaya Zemlya, Russia (Figure 1.a) [1]. The Barents Sea is unique due to its rich and diverse natural resources and being the northernmost ice-free sea during winter [7]. As a result, the southern part of the Barents Sea is open for industrial activities year-round, such as fishing, cargo transportation, and oil exploitation [8]. Furthermore, the Barents Sea is one of the most productive oceanic areas in the world, being the home of thousands of animal species, and a rich environment for plankton and bacteria [1].

The Barents Sea is rich in oil and gas deposits [9]. Because of this, Norway and Russia are both developing petroleum activities in the Barents Sea (Figure 1.b)[10]. Russia is the third largest oil producer in the world, accounting for over 12% of the world's production [9]. Norway occupies the 15th position in oil production worldwide, with the petroleum industry being its largest industry [9, 11]. The Norwegian Barents Sea holds two fields in production, Snøhvit and Goliat [9]. Although there is less petroleum activity in the Barents Sea than in the other areas of the Norwegian coast, it is estimated than half of the undiscovered oil resources are located in the Barents Sea [9]. The Russian Barents Sea has two main active fields, the Prirazlomnoe oil field and the Shtokman gas and condensate field. This last one is located 650 km north-east of Murmansk city and it is one of the largest offshore gas fields in the world [12]. The petroleum industry and the shipping of oil to Europe are the largest contributing factors for the increase growth of human activity in the Barents Sea.

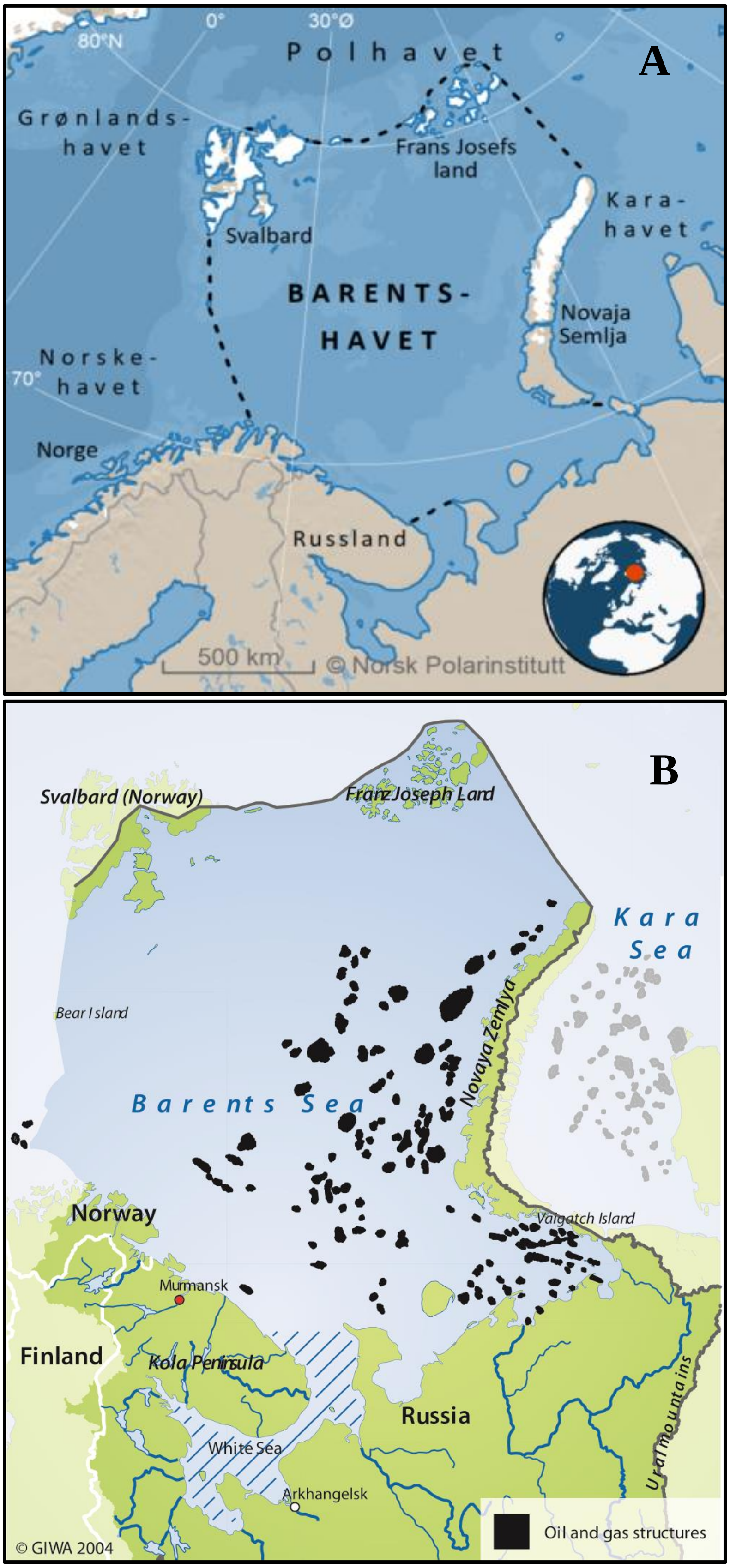


*Figure 1. A) Map of the Barents Sea [1] . B) Map of oil and gas structures in the Barents Sea. The red dot points the place of sampling of this study. Modified from [6]*

## 1.2 Hydrocarbon pollution in cold marine environments

Hydrocarbons are toxic persistent molecules that can have harmful consequences when released in marine environments. The marine environment is especially susceptible to hydrocarbon pollution, from a biological point of view, hydrocarbon pollution can affect photosynthesis processes in the water column, it is lethal or very harmful to marine fauna and flora, and it has long-term carcinogenic potential for humans [13, 14]. Hydrocarbon spills spread easier and are more difficult to remove in marine environments than on land. This is due to sea currents, waves and wind dispersing the oil on the water surface and distributing it through the water column [13, 15]. Moreover, many factors contribute to hydrocarbon pollution at sea. The main factors are natural seeps, involuntary oil spills from oil well blowouts, accidents involving oil vessels and oil transportation, spills from refineries and pipelines, and run-offs from terrestrial sources [2, 16]. In Norway, incidents such as the platform explosion at the EKOFISK oil field, the bulk carrier MV SERVER, and the bulk carrier FULL CITY have released more than 20,400 tons of oil to the sea [15].

### 1.2.1 The fate of hydrocarbons in the marine environment

The fate of hydrocarbons in marine environments depends on the hydrocarbon composition and on the physical and chemical properties of the environment. Petroleum has an extremely complex and diverse composition including thousands of different organic compounds.

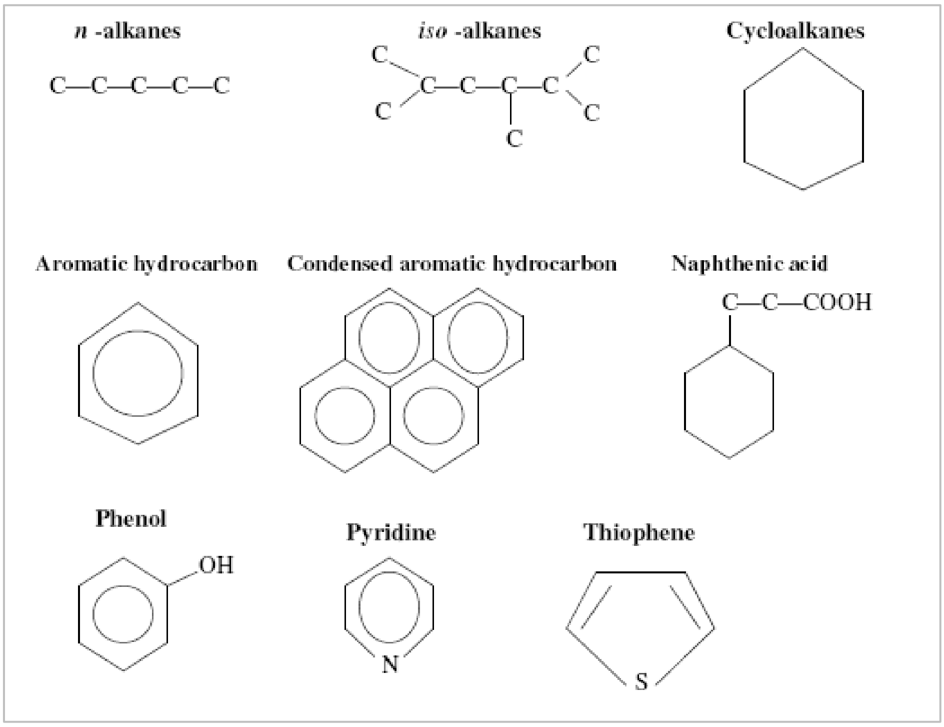


*Figure 2. Structural classification of some crude oil components [2]*

These organic compounds are often grouped into four categories: saturated hydrocarbons, aromatic hydrocarbons, resins, and asphaltenes (Figure 2) [17]. The fractions of these compounds vary within crude oils, creating numerous complex mixtures. Aromatic hydrocarbons, especially PAHs, and the polar fractions (asphaltenes and resins) are the more toxic and persistent in the environment [2, 18], especially at low temperatures, where the physical properties of the oil change and biodegradation processes become slower [18] .

Hydrocarbons interact with biotic and abiotic factors from the marine environment in a process called weathering [2, 4, 13, 18]. The abiotic weathering process comprise evaporation, dissolution, dispersion, emulsification, and sedimentation of the oil, among others. The biotic weathering includes microbial degradation and the ingestion by organisms. These processes act together changing the composition of crude oil and affecting the rate of biodegradation [2, 18]. Significantly, in cold environments the weathering process encounters several obstacles in the removal of hydrocarbons. For instance, the lower temperatures reduce the rates of evaporation and biodegradation of the oil, hence making it more persistent in the environment. The shore-fast ice might encapsulate the oil or trap it underneath it, thus making the oil more difficult to detect and track. Alternatively to these obstacles, the drift ice reduces the wave energy, thus reducing the natural dispersion and emulsification of hydrocarbons and facilitating the physical removal of oil spills [15]. A schematic illustration of the fate of hydrocarbons in seawater and in sea-ice is shown in figure 3. Natural microbial biodegradation is the ultimate fate of most petroleum compounds, especially the most persistent. However, the rate and efficiency of this biodegradation are variable and many factors might influence them [2, 14].

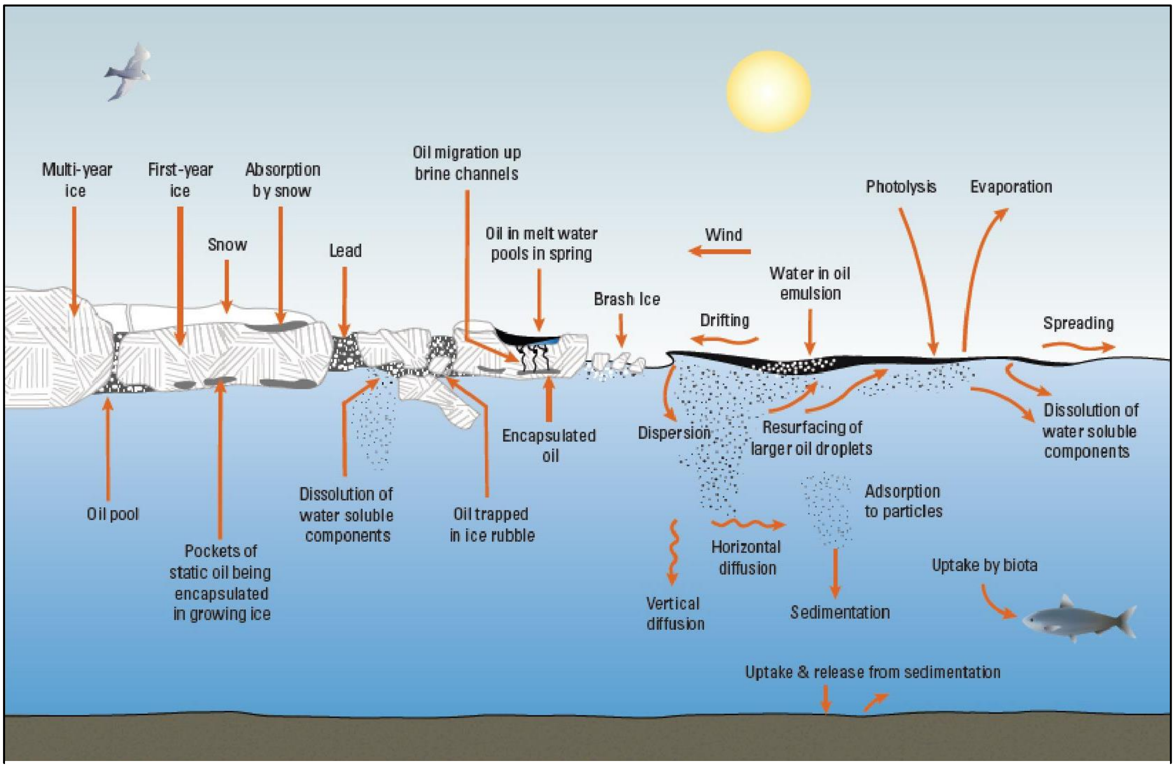

Multi-year ice
First-year ice
Absorption by snow
Snow
Lead
Oil migration up brine channels
Oil in melt water pools in spring
Wind
Photolysis
Evaporation
Water in oil emulsion
Brash Ice
Drifting
Spreading
Oil pool
Pockets of static oil being encapsulated in growing ice
Dissolution of water soluble components
Oil trapped in ice rubble
Encapsulated oil
Dispersion
Resurfacing of larger oil droplets
Dissolution of water soluble components
Adsorption to particles
Horizontal diffusion
Vertical diffusion
Sedimentation
Uptake by biota
Uptake & release from sedimentation

*Figure 3. schematic illustration of the fate of hydrocarbons in sea water and sea ice [4]*

## 1.3 Hydrocarbon degradation by bacteria

### 1.3.1 Metabolic pathways of hydrocarbon degradation in bacteria

The biodegradation of hydrocarbons starts with the destabilization or “activation” of the hydrocarbon molecule. According to the conditions in which this activation occurs, hydrocarbon biodegradation can be divided in two classes: activation in the presence of oxygen (aerobic biodegradation) or anoxic activation (anaerobic biodegradation). Significantly, aerobic biodegradation is the most common of the two and widely spread in the marine environment [2, 17, 19, 20]. Aerobic biodegradation can result in either the degradation of alkanes or degradation of aromatic compounds. Aerobic biodegradation of n-alkanes is typically initiated by monooxygenases, an enzyme that adds an atom of oxygen to the hydrocarbon molecule, activating it. This reaction results in an alcohol which is oxidized into an aldehyde, and finally transformed into a fatty acid. The fatty acids are processed in the β-oxidation pathway resulting in acetyl-CoA, which enters the tricarboxylic acid cycle (TCA cycle) producing biomass and energy (Figure 4) [2, 13, 18].

The degradation of aromatic hydrocarbons typically starts with oxygen activation by a dioxygenase, an enzyme that adds two atoms of oxygen to the aromatic ring. This oxidation results in the formation of a -diol molecule and the cleaving of the aromatic ring. Depending on the starting aromatic molecule, the biodegradation pathway, and the number of rings on the molecule, different key intermediates are formed. For instance, catechol or pyruvate molecules. These intermediates are converted to intermediates of the TCA cycle and used to obtain biomass and energy (Figure 4) [2, 13, 18].

In the anaerobic biodegradation process, the activation of the hydrocarbon molecule is performed without oxygen. Three strategies are used to achieve anoxic activation of hydrocarbons. The first one is the addition of a fumarate molecule to the hydrocarbon. The second strategy uses a water molecule to perform an oxygen-independent hydroxylation. Lastly, the third activation strategy is a carboxylation of the molecule, in other words, the addition of a carbon dioxide molecule [13].

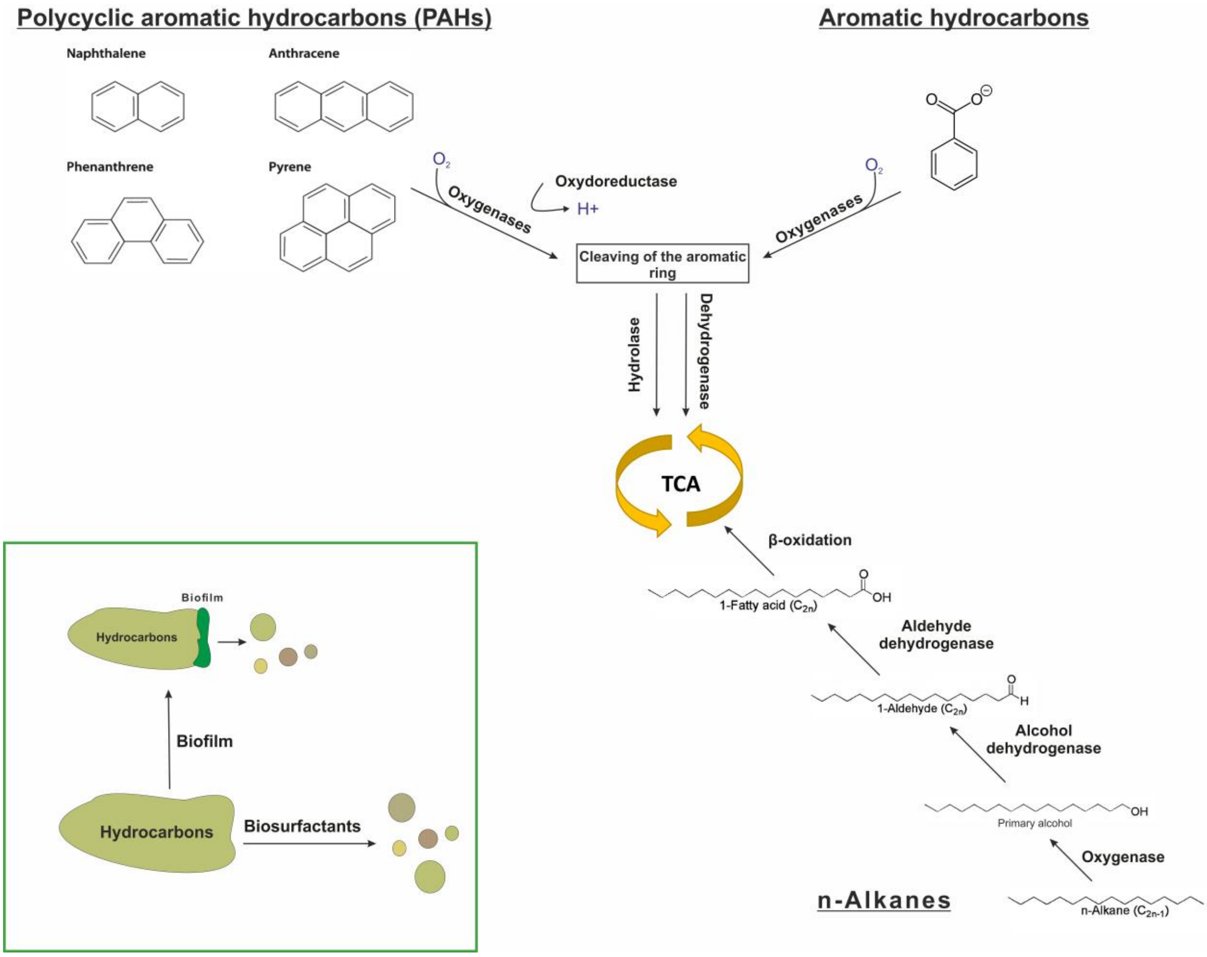


*Figure 4. Graphical representation of some aerobic hydrocarbon degradation pathways: Polycyclic aromatic hydrocarbons, aromatic hydrocarbons, n-alkenes, formation of biofilm, and production of biosurfactants.*

### 1.3.2 Hydrocarbon-degrading bacteria in marine environments

Hydrocarbon-degrading bacteria (HDB) in marine environments are a ubiquitous and diverse group of microorganisms [13, 21]. Normally, HDB constitute around 1% of the total marine bacterial population. However, after an oil spill or in hydrocarbon-rich environments, the dominance of HDB swiftly increases until they represent approximately 90% of the local microbial population [13]. Usually, HDB are specialized in degrading only one or a few fractions of hydrocarbons. Generally, to degrade the entire petroleum fraction a community of HDB has to work together, often forming biofilms [22] . In general, the bacterial biodegradation of hydrocarbon compounds is sequential, depending on the hydrocarbon complexity: linear alkanes > branched alkanes > low molecular weight alkyl aromatics > monoaromatics > cyclic alkanes > polyaromatics > asphaltenes [4, 22]. In cold seawater the same order is expected, although factors such as temperature, oil composition, and the weathering process can alter the order or rate of biodegradation [4].

The HDB in marine environments are a very diverse group. These bacteria can be classified depending on whether they perform an aerobic or an anaerobic degradation of hydrocarbons. The aerobic hydrocarbon degradation is dominant in marine environments [22]. Relevant representatives of aerobic HDB include obligate marine hydrocarbonoclastic bacteria. These bacteria can exclusively use hydrocarbons as their source of energy and carbon. The most significant members of this group belong to the class Gammaproteobacteria, and include the genera *Alcanivorax, Cycloclasticus, Oleispira, Thalassolitus,* and *Oleiphilus* [13]. Besides hydrocarbon obligates, HDB usually are able to use a range of compounds as a source of energy. Some representatives of these bacteria in the marine environment also belong to the class Gammaproteobacteria, such as the genera *Neptumonas*, *Marinobacter* and *Pseudomonas*. Other examples of aerobic HDB are found in the class Alphaproteobacteria, for instance the genera *Sphingomonas*, *Thalassospira*, and *Paracoccus.* In addition, other important genera are *Rhodococcus* and *Gordonia* from the phylum Actinobacteria. In anaerobic hydrocarbon biodegradation the dominant orders are Desulfobacterales and Desulfuromonadales [2, 13].

In cold marine environments, the HDB population is relatively similar to the population in temperate marine environments [4]. The main genus associated with hydrocarbon biodegradation at low temperatures is *Gammaproteobacteria*. Other relevant genera of HDB in Arctic seawater and ice include members of *Alphaproteobacteria, Epsilonproteobacteria,* *Table 1.Taxonomy of Arctic or Antarctic oil-degrading bacteria [18]. a) An Antarctic, Ar Arctic, S sediment, SI sea ice, SW seawater.*

| Class | Family | Genus | Source[a] | References |
|---|---|---|---|---|
| **Alphaproteobacteria** | *Rhodobacteraceae* | *Loktanella* | Ar, SW | [20] |
| | | *Sulfitobacter* | Ar, SW | [23] |
| | *Sphingomonadaceae* | *Sphingopyxis* | Ar, SW | [20] |
| | | *Sphingomonas* | An, SW | [24] |
| **Gammaproteobacteria** | *Alteromonadaceae* | *Alteromonas* | SW | [25] |
| | | *Glaciecola* | Ar, SI | [26] |
| | | *Marinobacter* | An, Ar, SI, SW | [24, 27, 28] |
| | *Colwelliaceae* | *Colwellia* | An, Ar, S, SI, SW | [20, 23-26, 29-31] |
| | | *Thalassomonas* | SW | [25] |
| | *Moritellaceae* | *Moritella* | Ar, S, SI, SW | [23, 29, 30] |
| | *Pseudoalteromonadaceae* | *Algicola* | Ar, SI | [23] |
| | | *Pseudoalteromonas* | An, Ar, S, SI, SW | [20, 26, 27, 29-32] |
| | *Psychromonadaceae* | *Psychromonas* | Ar, SW | [20, 32] |
| | *Shewanellaceae* | *Shewanella* | An, Ar, S, SI, SW | [24, 25, 27, 28, 30, 31] |
| | *Alcanivoracaceae* | *Alcanivorax* | Ar, S, SW | [23, 31] |
| | *Oceanospirillaceae* | *Marinomonas* | An, Ar, S, SI, SW | [24, 26, 31] |
| | | *Oleispira* | Ar, An, SI, SW | [20, 23-26, 32] |
| | *Halomonadaceae* | *Halomonas* | An, Ar, S, SI SW | [24, 28, 31] |
| | *Moraxellaceae* | *Psychrobacter* | Ar, SW | [20, 27] |
| | *Pseudomonadaceae* | *Pseudomonas* | An, Ar, S, SI, SW | [24, 27, 28, 31] |
| | *Piscirickettsiaceae* | *Cycloclasticus* | Ar, S, SW | [25, 31] |
| **Epsilonproteobacteria** | *Campylobacteraceae* | *Arcobacter* | An, Ar, SW | [24, 29, 32] |
| **Bacteroidetes** | *Cytophagales* | *Cytophagia* | An, SW | [24, 32] |
| **Flavobacteriia** | *Flavobacteriaceae* | *Ulvibacter* | Ar, SW | [20] |
| | | *Polaribacter* | Ar, SI, SW | [20, 28, 29] |
| **Actinobacteria** | *Nocardiaceae* | *Rhodococcus* | An, SW | [24] |
| | *Microbacteriaceae* | *Agreia* | Ar, SI, SW | [27, 28] |
| | | *Arthrobacter* | An, SW | 12 |

*Actinobacteria,* and *Bacteroidetes* (Table 1) [4, 18].

### 1.3.3 Bioremediation

Bioremediation is the removal of pollutants from an environment using biological processes [17]. The bioremediation of hydrocarbons transforms hazardous oil fraction into non-toxic compounds. Furthermore, bioremediation is considered the most eco-friendly, cost-effective solution for marine ecological restoration [4, 14, 33]. In marine environments there are two essential bioremediation strategies: biostimulation and bioaugmentation. Biostimulation is the application of treatments to enhance the biodegradation rate of the indigenous bacterial population. Treatments such as application of chemical dispersant to improve hydrocarbon bioavailability or the use of fertilizers for the native HDB. On the other hand, bioaugmentation consists in the inoculation of exogenous HDB, in some cases, genetically modified bacteria. [4, 14, 33].

Bioremediation is a promising tool for the control of hydrocarbon pollution. However, many challenges are yet to be overcome. Some of these challenges include the improvement of the physical contact between bacteria and hydrocarbons, slow biodegradation rates at low temperatures, physical changes of the oil at low temperatures that difficult biodegradation, and the long duration of the process [22].

### 1.3.4 Functional metagenomics and fosmid libraries

A significant limitation in environmental microbial studies is that only a small fraction of all microorganisms on Earth can be cultivated in a standard laboratory [34, 35]. A method to overcome this limitation is the use of metagenomic libraries in functional studies. This approach uses functional genes of a metagenome, these genes are cloned and expressed in culturable microorganisms [34]. Metagenomic fosmid libraries use fosmids, hybrid plasmids based on the bacterial F-plasmid, to clone functional genes and transform them in *Escherichia coli* cells [5, 36]. Fosmid libraries have been previously used in functional studies, for instance to research enzymes in extreme environments [34], to look for novel enzymes such as carboxylesterases and hydrogenases [37, 38] and to study genes involved in degradation of aromatic compounds in sediments [39]. In this study, a metagenomic fosmid library was used to research genes involved in hydrocarbon degradation, a schematic illustration of the process followed is in figure 5.

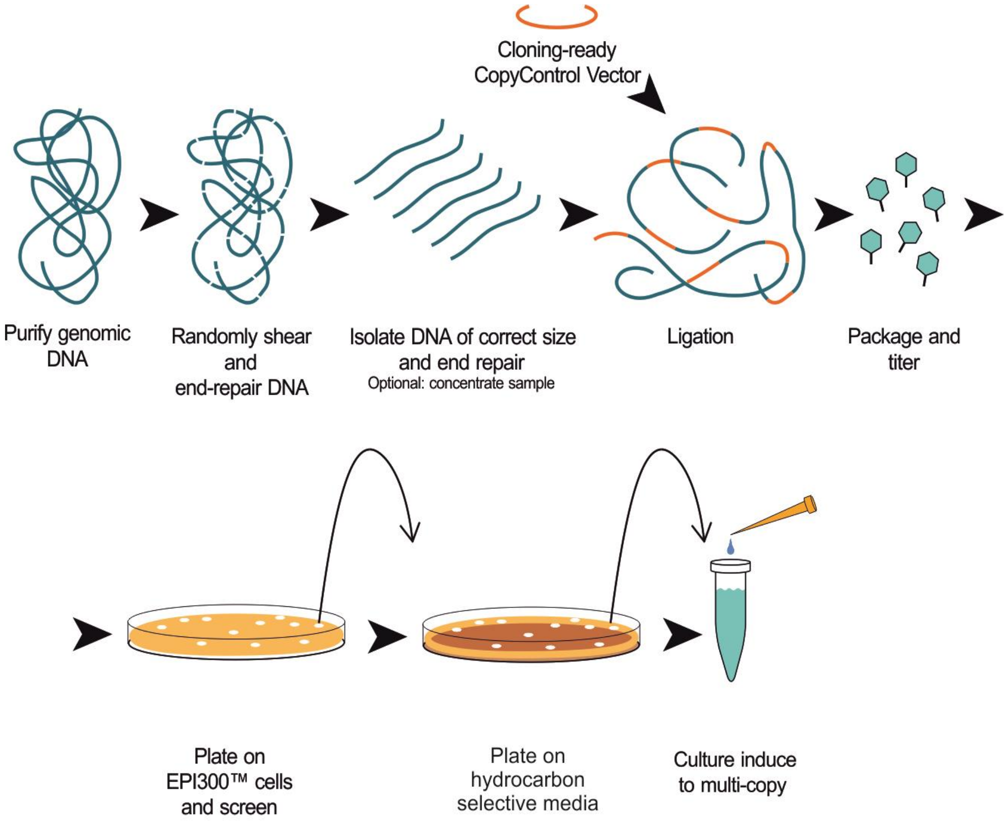

Cloning-ready
CopyControl Vector
Purify genomic DNA
Randomly shear and end-repair DNA
Isolate DNA of correct size and end repair
Optional: concentrate sample
Ligation
Package and titer
Plate on EPI300™ cells and screen
Plate on hydrocarbon selective media
Culture induce to multi-copy

*Figure 5. Schematic illustration of the production of a CopyControl Fosmid library, selection of clones and subsequent induction of clones to high-copy number. Modified from [5].*

## 1.4 Study background

The present work is based in a previous study on the bioremediation potential of bacteria from oil-polluted waters in the Barents Sea [40]. In the previous study, epiphytic bacterial communities in association with the macro-algae *Fucus vesiculosus* were analyzed. The bacterial communities belong to highly different environments: a petroleum-free environment at Dalnie Zelentsy, an environment polluted with eutrophicated urban water at Abramys in Kola Bay, and a chronically oil-polluted environment from Murmansk Sea Port. To study the communities, their 16S rRNA sequences were analyzed using the software MEGAN6 [41]. The results of the analysis showed differences in the composition of the bacterial communities (Figure 6). Notably, the oil-polluted environment presented a distinctive community dominated by Gammaproteobacteria. The difference in community composition was attributed to the presence of hydrocarbons in the environment. In the present study, the metagenome collected at Murmansk Sea Port was further analyzed with the aim of exploring its hydrocarbon bioremediation potential.

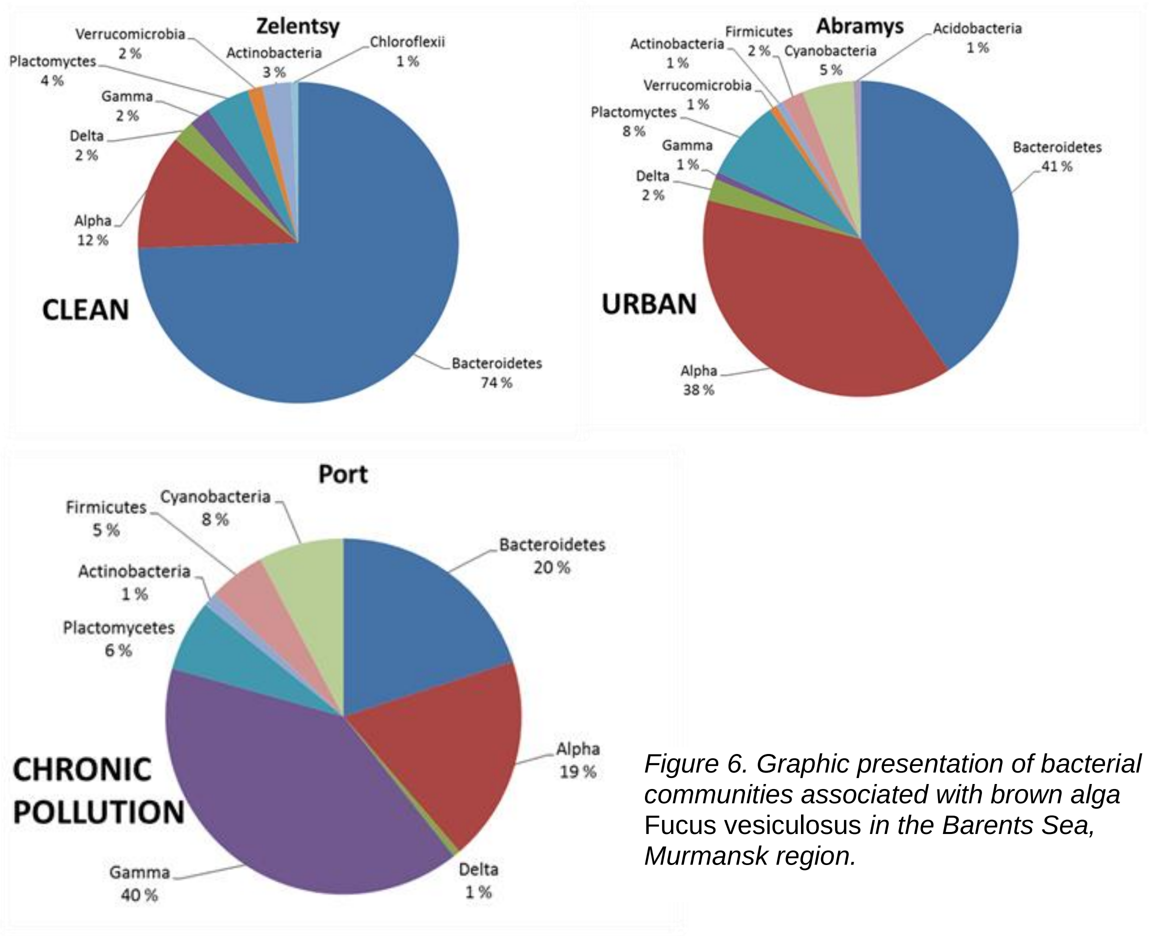


*Figure 6. Graphic presentation of bacterial communities associated with brown alga* Fucus vesiculosus *in the Barents Sea, Murmansk region.*

## 1.5 Objectives

The main goals of this master thesis are:

- First, to study the hydrocarbon bioremediation potential of an indigenous bacterial community. We intended to do so by performing a functional metagenomic analysis and a bioinformatic analysis on a metagenome from an oil-polluted environment.
- Second, to identify native bacterial species that could have a role in bioremediation of hydrocarbons in cold marine environments.
- Third, to identify novel genes and enzymes involved in hydrocarbon degradation in cold environments.

# 2 Material and Methods

## 2.1 Isolation of metagenomic DNA

The metagenomic DNA was sampled in Murmansk seaport (Kola Bay, 68°58′00″ N, 33°05′00″ E). The DNA samples were collected from the surface of the thalli of the macroalgae *F. vesiculosus* [42]. A total of 21 samples were collected and stored at -20 ºC in 1.5 ml tubes until DNA isolation began.

The DNA isolation was performed following the protocol "Bacterial genomic DNA isolation using CTAB" for 1.5 ml samples from the Joint Genome Institute (JGI) [43]. The protocol is attached in appendix 1.

Following the DNA isolation, the quality of the extracted DNA was analyzed on a 0.8% agarose gel electrophoresis ran at 150 V for 15 minutes. In addition, the DNA concentration was measured using a Nanodrop™ 2000 spectrophotometer. After the measurements, the isolated DNA samples were pooled and divided in three 1.5 ml tubes to ensure randomization.

## 2.2 Media preparation

### 2.2.1 Minimal media and Luria broth media

Luria Broth media (LB) and M9 minimal media (M9) were used to produce the metagenomic fosmid library and in testing the transformed *E. coli* cells ability to use of hydrocarbons as sole energy and carbon source. All media were prepared following "Current protocols in molecular biology" [44], attached in appendix 2. To select transformed *E. coli* cells, a concentration of 12 µg/ml of the antibiotic chloramphenicol was added to the media, making M9 Chloramphenicol (M9-C) and LB Chloramphenicol (LB-C) media. The chloramphenicol was diluted in methanol and poured into the media before the solidification of the agar [45].

To test hydrocarbons as sole source of carbon and energy, M9-C media was enriched with different concentrations of Diesel (D) or Crude Oil (CO). The hydrocarbons were mixed with the media in 200 ml sterile glass flasks and the mix was poured in plastic petri dishes. As controls plates 0.2% maltose was added instead of hydrocarbons. A description of the content of the M9-C media used is in table 2.

Prior to the fosmid library production, *E. coli* cells were prepared for their use. "EPI300-T1$^R$ planting strains" *E. coli* cells were plated on LB and incubated at 37ºC overnight. The plate was sealed and stored at 4ºC. A day before the library production, a single colony was inoculated in 50 ml of LB with 10 mM $MgSO_4$, 0.2% Maltose incubated at 37ºC and shaken overnight at 250 rpm. The LB cultures did not contain chloramphenicol.

*Table 2.Content description of M9 minimal media and hydrocarbon quantities added to the media.*

| **Minimal media content per Liter** | **Diesel/ Crude oil per 200 ml** |
|---|---|
| 800 ml autoclaved Milli-Q water | 100 µL D / CO (0.05%) |
| 14 g agar for culture media | 1 ml D / CO (0.5%) |
| 200 ml M9 media 5X | 2 ml D / CO (1%) |
| 1 ml Leucine (100 µg/ml) | 3 ml D / CO (1.5%) |
| 1 ml $FeSO_4 \cdot 7H_2O$ (25 mg/L) | 4 ml D / CO (2%) |
| 3 µl Thiamine (50 µl/L) | 10 ml D / CO (5%) |
| 128.42 µl Chloramphenicol (12µg/ml) | |
| *control: 10 ml Maltose 20% | |

### 2.2.2 Diesel and crude oil

The hydrocarbons used in the study were diesel and crude oil. Diesel is a mixture of paraffins, naphthenes, and aromatic hydrocarbons with carbon numbers between 10 and 22. The diesel used in the study was obtained in Murmansk, Russia and classified as wintertime diesel. Winter diesel contains less wax, shorter carbon chains and a higher content in naphthenes and aromatic hydrocarbons than ordinary diesel [46]. The crude oil used in the study was obtained through the company "Neste Oyj", Finland. The commercial product name is "Crude Oil, Sour (min 0,5 % Sulphur)". The composition is 82–87% carbon, 11–15% hydrogen, with the balance being oxygen, nitrogen, and sulphur [47].

To remove microbial contamination, the diesel and crude oil were filtered through an "Acrodisc syringe filter" with a 0.2 µm supor membrane (Pall life sciences). Diesel and crude oil were kept in sealed glass flasks in the dark at 4ºC.

## 2.3 Metagenomic fosmid library production

The metagenomic fosmid library of the isolated DNA was produced using the kit "CopyControl™ HTP Fosmid Library Production Kit with pCC2FOS™ Vector and Phage T-1 Resistant EPI300™-T1$^R$ *E. coli* Plating Strain" from Epicentre, USA [5] Protocol attached in appendix 3. The steps followed to produce the library were a ligation of the isolated DNA into the plasmid vector, followed by the packaging of the ligated DNA-vector into the lambda phage. Lastly, the lambda phage was transformed into EPI300™-T1$^R$ *E. coli* cells which constituted the fosmid library.

Prior to the fosmid library production, the size of the isolated metagenomic DNA was measured with a gel electrophoresis 1% agar run at 30V overnight. The size control used was a 100 ng of "Fosmid control DNA" size 40kb, supplied by the kit. Subsequently, an "end-repair enzyme mix" was added to the metagenomic DNA. In the present study, isolated metagenomic DNA will be refer to as "insert DNA".

### 2.3.1 Ligation of insert DNA into pCC2FOS plasmid vector

To ligate the insert DNA to the pCC2FOS plasmid vector, a ligation reaction was prepared. by adding 1 µl of 10x Fast-link reaction buffer, 1 µl of 10mM ATP, 1 µl of Copycontrol pCC2FOS vector (0.5 µg/µl), 1 µl of Fast-link DNA ligase, and 1.32 µl of concentrated insert DNA (0.25 µg) to a 1.5 ml tube. The total volume of the ligation reaction was 10 µl, with 10:1 molar ratio of vector and insert DNA. A single ligation reaction produces between $10^3$ to $10^6$ clones, according to producer.

### 2.3.2 Package of fosmid clones into the lambda phage

To make the packaging of the ligated DNA-vector into the lambda phage, one tube of "MaxPlax Lambda Packaging extract" (50 µl) was used per ligation reaction.

First, 25 µl of the MaxPlax packaging extract were mixed with the ligation reaction by pipetting. The mix was incubated at 30 ºC for 2 hours. Subsequently, the other 25 µl of the packaging extract were added, mixed, and incubated for another 2 hours at 30 ºC. At this point, 0.5 ml of Phage Dilution Buffer (PDB) and 25 µl of chloroform were added to preserve the packaged fosmid clones. Samples were kept at 4 ºC until use.

### 2.3.3 Titer and storage of packaged fosmid clones

The objective of the titering was to determine the dilution of packaged fosmid clones necessary to transform sufficient *E. coli* cells to efficiently produce the metagenomic library. The titer was done by making serial dilutions of the fosmid vector and planting them with *E. coli* cells.

To produce the titer, packed fosmid clones were diluted in PDB. Dilutions made were 1:1, 1:10, 1:$10^2$, and 1:$10^3$. Subsequently,10 µl of each dilution were added to separate 1.5 ml tubes with 100 µl of an *E. coli* culture with $A_{600}$ of 0.8. Tubes were incubated at 37 ºC for one hour to induce *E. coli* cells transformation. Following the incubation, cultures were plated on LB-C in duplicates. The plates were incubated overnight at 37ºC and the following day visible colonies were counted. Negative and positive controls consisted of *E. coli* cells without the fosmid vector and *E. coli* cells with the plasmid vector lacking the insert DNA, respectively.

The titer of packaged fosmid clones (CFU/ml) was calculated following the equation:

$$\frac{(number\ of\ colonies)(dilution\ factor)(1{,}000\ \mu\mathrm{l/ml})}{(volume\ of\ phage\ planted[\mu\mathrm{l}])}$$

After the tittering, the fosmid library was prepared for storage. To begin with, 575 µl of packaged fosmid clones were incubated with 5. 75 ml of *E. coli* LB culture for one hour at 37 ºC. Subsequently, the culture was spread on 40 LB-C agar plates and incubated overnight at 37ºC. The following day, colonies on the plates were resuspended with 2 ml of liquid LB-C media and plated on to new LB-C agar plates, incubated at 37ºC overnight. Lastly, *E. coli* colonies were resuspended with 2 ml of LB-C media and pooled in a sterile glass flask. A total of 60 ml of LB-C media containing transformed *E. coli* colonies were retrieved. For storage, the LB-C medium containing the library was mixed with a total concentration of 20% glycerol, aliquoted into sterile eppendorf tubes and stored at -80ºC.

## 2.4 Culture of transformed *E. coli* and selection of clones

Prior to the culturing of the transformed *E. coli*, six concentrations of Diesel (D) were tested on M9-C media (0.05%, 0.5%, 1%, 1.5%, 2% and 5% D). The hydrocarbon concentrations were selected according to prior studies [48-50]. To test bacterial growth in D, a total of 32 agar plates were made including four plates for each D concentration (0.05%, 0.5%, 1%, 1.5%, 2% and 5% D), four plates without D, and four plates with maltose instead of D. Two dilutions of the fosmid library in PDB were tested, dilution 1:1 and dilution 1:10. Plates were incubated at 37ºC and checked for results after one, three, and seven days. As a result of the tests, concentrations 0.05%, 0.5%, and 1% of D /CO and the dilution 1:1 of the fosmid library were used for the rest of the cultures.

To select fosmid clones with hydrocarbon degradation potential, cultures of transformed cells in D and CO media were done. For cultures in M9-C with Diesel (M9-CD), colonies from the fosmid library were placed on five M9-CD plates of each D concentration (0.05%, 0.5%, 1%) and incubated at 37ºC for six days. Additionally, 21 control plates were incubated in the same conditions, for details on the controls refer to figure 11. Subsequently, CO cultures were prepared following the same procedures as the D cultures. Colonies from the fosmid library were plated on five M9-C with CO (M9-CCO) plates of each CO concentration (0.05%, 0.5%, 1%). Plates were incubated at 37ºC for eight days. Controls for M9-CCO plates are shown in figure 11.

To facilitate extraction and analysis of the fosmid clones, single colonies from M9-CD and M9-CCO plates were transferred to LB-C plates, following the T- streaking method to ensure single colonies. The LB-C plates were incubated at 37ºC for two days. To obtain more copies of the fosmid clones, an autoinduction culture was performed. A total of 26 autoinduction cultures were prepared by putting 3 ml of LB-C in 10 ml glass tubes. Subsequently, single colonies from the LB-C plates were introduced in the glass tubes and 6 µl of "500X CopyControl fosmid autoinduction solution" were added to each culture. Autoinduction cultures were incubated for 18 hours at 37ºC, shaking at 230 rpm.

## 2.5 Fosmid clones extraction and enzyme restriction

The extraction of the fosmid clones was performed following the instructions in "Preparation of plasmid DNA by alkaline lysis with SDS: minipreparation.", Cold Spring Harbor Protocols, 2006. [51] attached in appendix 4. Twenty-six extractions were made, 20 fosmids were extracted from D cultures and 6 fosmids were extracted from CO cultures. The concentration and purity of the extracted fosmids were determined by spectrophotometry and gel electrophoresis. The gel electrophoresis was performed using a 1% agarose gel run overnight at 30V.

In addition to the extraction, enzyme restriction analysis of the fosmid clones was performed to check for differences in their sequences. Prior to the enzyme restriction, all fosmid extractions were diluted with Tris-EDTA buffer to obtain the same DNA concentration in every sample, 200 ng/ml.

The enzyme chosen to perform the restriction analysis was *XbaI* (3.000 U) from Thermo Scientific. According to protocol, this restriction enzyme cuts the pCC2FOS vector in two sites: 413 bp and 3234 bp [5]. The restriction sites of *XbaI* are:

**5' T↓C T A G A 3'**

**3' A G A T C↑T 5'**

The enzyme restriction was performed following the Thermo Scientific protocol for *XbaI* (3.000 U) attached in appendix 5. The volumes of reagents used per restriction reactions were: 16 μl of autoclaved Milli-Q water, 2 μl of 10X buffer Tango (Thermo Scientific), 1 μl of DNA (200 ng/ml), and 1 μl of enzyme *XbaI* (10 U/μl). Reactions were incubated at 37ºC for 90 minutes, followed by 20 minutes at 65ºC to inactivate the enzyme. A gel electrophoresis of the restriction fragments was performed using a 1% agarose gel run at 30V overnight. GelRed was used for DNA staining and the DNA ladder was 1kb plus from ThermoScientific. After the visualization of the gel electrophoresis results, 10 fosmids were selected for sequencing, five from D cultures and five from CO cultures.

## 2.6 Sequencing of fosmid clones

Ten fosmid clones were sent for sequencing to IMGM Laboratories GmbH, Martinsried, Germany. Cluster generation and sequencing were performed on the Illumina MiSeq® next generation sequencing system (Illumina Inc.)[52].

Prior to cluster generation, the DNA was fragmented and denature into single stranded DNA and sequencing adapters were added to the DNA fragments. Cluster generation was performed by bridge amplification. The fragments were anchored and immobilize, and by cycles of binding the fragments to the surrounding primers followed by amplifications, approximately 1,000 copies of the original fragment were created, forming a tight cluster.

After cluster generation, sequencing primers were hybridized to the extremes of the DNA fragments. During each sequencing cycle, the cluster was flooded with all four nucleotides

(A, T, G, C). Each nucleotide was labeled with a different fluorophore. In each cycle a base was attached to the growing antisense strand of the DNA fragments, starting from the sequencing primers. A signal corresponding to each nucleotide was emitted and unattached nucleotides were washed away.

In the present study, bidirectional sequencing was performed. Both reads, sense and antisense, had a length of  approximately 150 bases, finally producing 300 bases of sequence information in 2 x 150 bp paired-end reads.[53]

## 2.7 Bioinformatic analysis

### 2.7.1 Treatment of raw data and contig assembly

The method used to sequence the fosmid clones consist of shotgun metagenomic sequencing employing the Illumina MiSeq® next generation sequencing system (Illumina Inc.). The sequencing data of the 10 fosmids was provided in the format **.fastq* as 20 files in separate forward and reverse read lists composed by 2 x 150 bp paired-end reads.

Sequencing data was processed with the software Geneious 2020.2.2 [3]. The 20 **.fastq* files were upload to the software and paired automatically, resulting in 10 files referred to in this work as “Sequence S1 to S10”, respectively.

#### 2.7.1.1 Raw data treatment: Trimming, removing of duplicates, merging of reads and error correction

Prior to the assembly of reads into contigs, a pre-processing of the raw data was performed to prevent assembly errors and to reduce required computational power and time. It consisted of trimming of low quality read ends, removing of read duplicates, merging of overlapping reads and normalization and error correction of read coverage. All processes were performed using programs contained in the software Geneious 2020.2.2 [3] and parametres were set according to Geneious prime [3] recommendations for Illumina sequencing [54].

To begin, a trimming of the read ends was performed with the program “BBDuk: Quality Trimming version 38.84 by Brian Bushnell”. Trimming of low quality read ends such as vectors, primers, and poor-quality bases prevents incorrect assemblies [55]. Next the removing of duplicates followed using the software “Dedupe: Duplicate Read Remover version 38.84 by Brian Bushnell”. This process is for the identification of non-exact duplicates, the identification and removing of exact duplicates, and the removing of duplicates on paired read datasets [55]. Subsequently, an error correction and normalization of the read coverage was performed using “BBNorm: Error correction and read normalization version 38.84 by Brian Bushnell”. To error correct the data or to normalize coverage by discarding reads in regions of high coverage [55]. Lastly, the merge of the reads into contigs was finalized with “BBMerge: Paired Read Merger version 38.84 by Brian Bushnell” [55]. Parameters used in each operation are shown in table 3. A schematic view of the bioformatic tools used is in figure 7.

*Table 3. Parameters used in bioinformatic tools BBDuk, Dedupe, BBNorm, BBMerge and Glimmer contained in the program Geneious 2020.2.2 [10].*

| **Function/ Program** | **Parameters** |
|---|---|
| **Trim**<br>BBDuk | Trim Adapters<br>▪ Adapters: All truseq, Nextera and PhiX adapters<br>▪ Trim: Right end<br>▪ Kmer Length: 27<br>▪ Maximum substitutions: 1<br>Trim Low quality<br>▪ Trim: Both ends<br>▪ Minimum quality: 30 (Q score)<br>Trim adapters based on paired read overhangs<br>▪ Minimum overlap: 24 bp<br>Discard short reads<br>▪ Minimum length: 30 bp |
| **Remove duplicates**<br>Dedupe | Kmer seed Length: 31<br>Maximum edits: 0<br>Maximum substitutions: 0 |
| **Error correction and normalization**<br>BBNorm | Error correction<br>▪ Sensitivity: Default settings<br>▪ Mark: Uncorrectable errors by leaving nucleotide unchanged and assigning low quality |

| | |
|---|---|
| | Normalization<br>▪ Target Coverage Level: 40<br>▪ Minimum Depth: 6 |
| **Merge of reads**<br>BBMerge | Merge Rate: Normal |
| **Gene prediction**<br>Glimmer | Model<br>▪ Compute a new model using long-orfs<br>Genetic code<br>▪ Genetic code: 11 (Bacteria, Archea)<br>▪ Start codons: ATG, GTG, TTG<br>▪ Start codons probabilities: 0.6, 0.35, 0.05<br>▪ Recalculate start codon probabilities for second pass<br>▪ Stop codons: TAG, TGA, TAA<br>Parameters<br>▪ Calculate position weighted matrix<br>▪ Automatic GC% setting<br>▪ Min gene length: 110 bp<br>▪ Max overlap length: 50 bp<br>▪ Threshold score: 30 (Q score) |

#### 2.7.1.2 Contig assembly and gene prediction

To assemble the reads into contigs without a reference genome, *de novo* assemblies were performed on the ten sequences previously treated. The assemblies were executed by "Geneious assembler", a program part of the Geneious 2020.2.2 software [3]. Parameters used are in table 4.

The *de novo* assemblies produced consensus sequences organized into contigs. The longest contig in each sequence, named "contig 1" by the program, was extracted from the assembly file. These contigs were re-circularized (Figures 20 to 26) and set as "Sequences 1 to 10". However, due to their similarity in length, in Sequence 7 "contigs 1, 2 and 3" were grouped together and named “Sequence 7”. Subsequently, a multiple alignment was performed between sequences 1 to 10 using “Clustal omega 1.2.2” [56]. This alignment was performed to search for high similarities or identical sequences among the plasmids. Sequences 1 to 10 are the focus of the subsequent analysis of this thesis.

*Table 4. Parameters for* De novo *assembly and Map to reference alignment tools contained in the program Geneious 2020.2.2 [3].*

| **Assembly/Alignment** | **Parameters** |
|---|---|
| ***De novo* Assembly**<br>(Generation of contigs) | Data:<br>▪ Use: 100% of the data<br>Method:<br>▪ Sensitivity: Medium Sensitivity/Fast<br>Trim:<br>▪ Do not trim before assembly |
| **Map to reference Alignment**<br>(Find Fosmid vector in contigs) | Data:<br>▪ Reference Sequence: Plasmid EU140752.1<br>Method:<br>▪ Mapper: Geneious<br>▪ Sensitivity: Medium Sensitivity/Fast<br>▪ Find structural variants, short insertions, and deletions of any size<br>▪ Find short insertions and large deletions up to 1,000 bp<br>▪ Fine tuning: None (fast/read mapping)<br>Trim:<br>▪ Do no trim before mapping |

To predict genes in the assembled contigs, the software program “Glimmer” [57] was used in Sequences 1 to 10. Parameters used for gene prediction are presented in table 3. The Open Reading Frames (ORFs) found by “Glimmer” were annotated on the sequences files. The ORFs in each sequence can be seen in figures 20 to 26.

To allocate the pCC2FOS vector (named Plasmid EU140752.1 in the NCBI database [58]) in the sequences, a “Map to reference” alignment was carried out. The vector sequence was downloaded from the NCBI gene database [58] and a “Map to reference” alignment was performed using pCC2FOS vector as the reference sequence. The predicted ORFs in Sequences 1 to 10 were aligned to it. Specifications for the assembly are listed on table 4.

### 2.7.2 Assignation of taxonomy and functions to ORFs: MEGABLAST, BLASTn and BLASTx searches

To study the ORFs predicted by the software “Glimmer”, several Basic Local Alignment Search Tool (BLAST)[59] searches were conducted. A summary of the full process is presented in  figure 7. The National Center for Biotechnology Information (NCBI) [58] developed BLAST, a tool which purpose is to find regions of similarity between protein or nucleotide sequences and sequences in NCBI databases. In this study three versions of BLAST were used: MEGABLAST, Nucleotide BLAST (BLASTn), and translated BLAST (BLASTx) [59]. The BLASTn tool searches sequences in nucleotide databases using a nucleotide query. The MEGABLAST tool is a variation on BLASTn designed to perform faster searches. However, it can only find matches if they present long alignments with high similarity. Lastly, BLASTx searches protein databases using a translated nucleotide query [60].

Firstly, MEGABLAST searches were performed for the predicted ORFs to obtain a first classification and facilitate the subsequent BLASTn searches. A division of the ORFs into “hits” and “no hits” bins was made by MEGABLAST. These divisions were based on whether the MEGABLAST tool was able to find any results for the queries (hits) or not (no hits). Secondly, different BLASTn searches were performed on the two types of results. For the “hits”, “query-centric alignment” searches of maximum 3 hits were made and for the “no hits”, "hit table” searches of maximum 10 hits were performed. A “hit table” search returns an alignment for every hit found, providing a table with information for each alignment. Alternatively, a “query-centric alignment” returns only one alignment for every query. This variant presents all the hits aligned against the query sequence, less information is provided than in the “hit table” search. For these reasons, “query-centric alignment” searches were performed for the “hits” results from MEGABLAST and “hit table” searches were performed for the “no hits” results, which presumably required more information. Additionally, BLASTx [59] searches were performed on MEGABLAST results for “hits” and “no hits” in the same manner as stated above for BLASTn, except the maximum number of hits in “query-centric alignment” searches was extended from 3 to 10. Specifications for MEGABLAST, BLASTn and BLASTx searches are shown in table 5 and a schematic representation of the process can be seen in figure 7.

Subsequently, a manual search of the protein functions was completed using the online databases NCBI [58],  Protein Data Bank (PDB) [61], Gene Ontology Annotation (GOA) database [62], and Uniprot [63].

Finally, a BLASTn [59] search was performed for the contigs not included in Sequences 1 to 10 to determine their source. The parameters used were identical to the previous BLASTn searches with “Hit table” search (Table 5).

*Table 5. Parameters for MEGABLAST, BLASTn, BLASTx and custom BLAST searches from the program Geneious 2020.2.2 [3] using NCBI database and tools [59]*

| **Search** | **Parameters** |
|---|---|
| **MEGABLAST**<br>(Nucleotide search) | ▪ Database: Nucleotide collection (nr/nt)<br>▪ Program: Megablast<br>▪ Results: Bin into «has hit» vs. «no hit»<br>▪ Max E-value: 10<br>▪ Word Size: 28<br>▪ Scoring (Match Mismatch): 1-2 |
| **BLASTn**<br>(Nucleotide search)<br>Hit table /Query-centric alignment | ▪ Database: Nucleotide collection (nr/nt)<br>▪ Program: blastn<br>▪ Results: Hit table / Query-centric alignment only<br>▪ Retrieve: Matching region<br>▪ Maximum hits: 10 / 3<br>▪ Max E-value: 10<br>▪ Word Size: 11<br>▪ Scoring (Match Mismatch): 2-3 |
| **BLASTx**<br>(Translated nucleotide to protein search)<br>Hit table /Query-centric alignment | ▪ Database: Non-redundant protein sequences (nr)<br>▪ Program: blastx<br>▪ Results: Hit table / Query-centric alignment only<br>▪ Retrieve: Matching region<br>▪ Maximum hits: 10 / 10<br>▪ Max E-value: 10<br>▪ Word Size: 6<br>▪ Matrix: BLOSUM62<br>▪ Gap cost (Open Extent): 11 1<br>▪ Genetic code: Bacterial (11) |

| | |
|---|---|
| **Custom BLAST**<br>(Protein search) | ▪ Database: AromaDeg database proteins<br>▪ Program: blastp<br>▪ Results: Query-centric alignment only<br>▪ Retrieve: Matching region<br>▪ Maximum hits: 10<br>▪ Max E-value: 5<br>▪ Word Size: 6<br>▪ Matrix: BLOSUM62<br>▪ Gap cost (Open Extent): 11 1 |

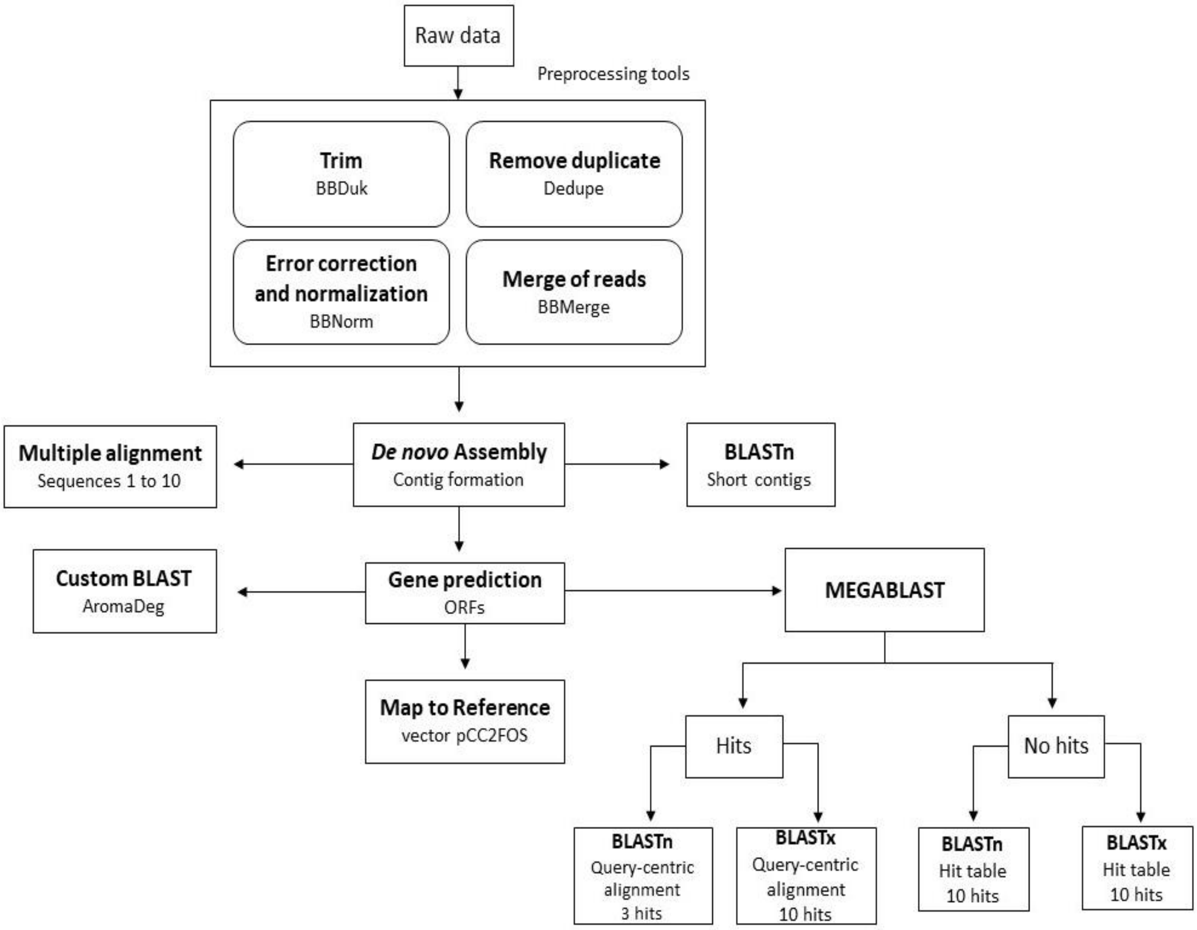


*Figure 7.Schematic representation of the bioinformatics processes and tools used on sequences 1 to 10.*

### 2.7.3 Determination of the phylogeny

Prior to tree building, a multiple alignment of the 491 sequences was performed by “Clustal omega 1.2.2” [56]. The alignment was used to build the phylogenetic tree.

The phylogenetic tree was built by the Geneious Tree Builder [3], parameters used were: Genetic Distance Model: Tamura-Nei, Tree building method: Neighbor-Joining, No Outgroup, Consensus Tree Options: Resample tree, Resampling Method: Bootstrap, Random Seed: 321,056, Number of replicates: 100, Create consensus tree, Support Threshold %: 50. The phylogenetic tree was edited and modified using the online service “iTOL v5.6.3”[64].

### 2.7.4 Databases and Servers

Databases and servers used were the AromaDeg [65] database and online servers MG-RAST version 4.0.3 [66] and AntiSMASH [67].

The AromaDeg [65] database is composed by sequences of proteins involved in PAHs degradation pathways. Amino acids sequences were downloaded and a custom BLAST [59] was performed (Table 5), ORFs from Sequences 1 to 10 were translated to amino acid sequences and blasted against AromaDeg proteins.

The MG-RAST [66] server is a metagenomic analysis server which suggests automatic phylogenetic and functional analysis. Data submitted to the server consisted of preprocessed reads prior to *de novo* assembly.

AntiSMASH [67] is an online tool for the identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genomes. Sequences 1 to 10 were uploaded to the server.

# 3 Results

## 3.1 Isolated metagenomic DNA

Total DNA of six environmental samples was isolated and analyzed by gel electrophoresis (Figure 8) and Nanodrop spectrophotometry (Table 6). In the gel electrophoresis, the DNA bands were visible in every sample and no residual RNA was observed. All DNA bands appeared in the same size range. Bands from samples 4, 5, and 6 appeared more intense, (Figure 8).

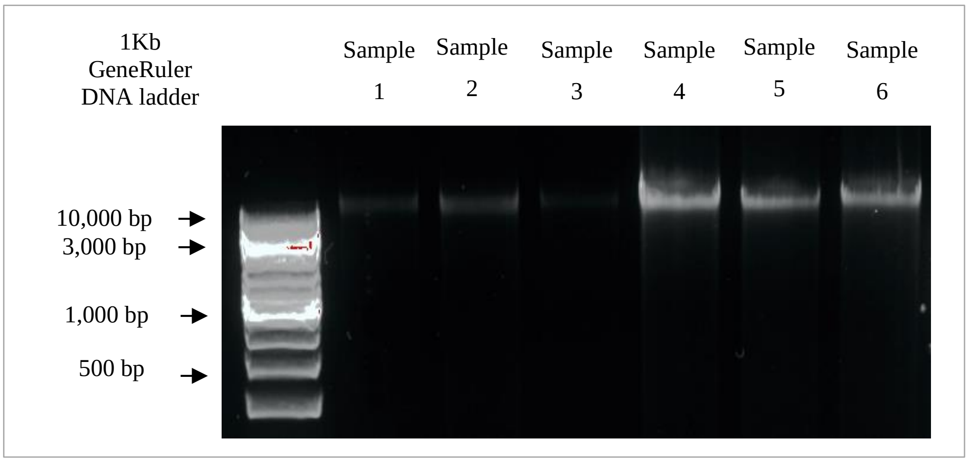


*Figure 8.Gel electrophoresis 0.8% agarose of bacterial isolated DNA from the six environmental samples. The DNA Ladder is 1Kb GeneRuler from Thermofisher.*

The spectrophotometry measurements are shown in table 6. The total quantity of DNA was obtained by multiplying nanograms/microliter times number of microliters in each sample. The absorbance measure at 260 nm (ng/µl) corresponds with the intensity seen in the bands of the electrophoresis in figure 8. Both ratios, A260/280 and A260/230 are lower than expected for pure samples. However, samples in this study were environmental samples with mixed DNA from multiple organisms.

*Table 6. Nanodrop readings of the total DNA samples: DNA concentration (ng/µl), Absorbance ratio 260/280 (nm), Absorbance ratio 260/230 (nm), total quantity of DNA in samples (ng).*

| **Sample number** | **ng/µl** | **Absorbance ratio 260/280 nm** | **Absorbance ratio 260/230 nm** | **Total DNA of the samples (ng)** |
|---|---|---|---|---|
| 1 | 7.4 | 1.20 | 1.76 | 666 |
| 2 | 10.2 | 1.62 | 1.68 | 918 |
| 3 | 5.9 | 0.95 | 1.74 | 531 |
| 4 | 24 | 1.43 | 1.69 | 2160 |
| 5 | 19.9 | 1.54 | 1.72 | 1791 |
| 6 | 22.4 | 1.55 | 2.50 | 2016 |

Prior to the fosmid library production,  gel electrophoresis was performed to measure the size of the isolated DNA. The optimal size, according to producer, is between 30 to 45 kb. The Gel electrophoresis displayed two bands: Copycontrol 40 kb Fosmid control DNA and isolated DNA, a pooled of samples 1 to 6 (Figure 9). The bands presented similar size, thus the isolated DNA was considered as suitable for the ligation.

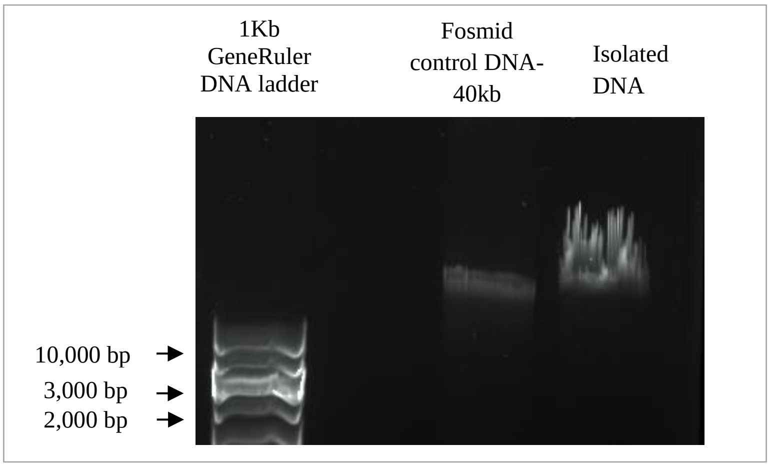


*Figure 9. A 1% agarose gel electrophoresis at 30V, overnight. with control 40 kb Fosmid control DNA and pooled isolated DNA.*

### 3.1.1 Titer of the packaged fosmid clones

The titer cultures were repeated four times with no growth on the plates. The duration of the ligation reaction was changed and new *E. coli* cultures and new media were made, however, the efficiency was still poor. Finally, a new measure of the DNA concentration was taken, it presented lower concentration than expected. The new measure was 131.4 ng/µL instead of the expected 190.7 ng/µL. Due to this result, the ligation reaction was repeated adjusting the DNA concentration to obtain 0.25 µg. The packaging reaction and titering were also repeated, results after the changes are presented in table 7. The titer plates showed visible colonies when using dilutions 1:1 and 1:10 of PDB and packaged fosmid clones (Table 7). No colony growth was observed in dilutions 1:$10^2$ and 1:$10^3$. Dilution 1:1 was used for the subsequent cultures.

*Table 7. Number of colonies grown per plate and CFU/ml in dilutions 1:1 and 1:10 of packaged phage particles.*

| **Dilutions** | **Nº of colonies/plate** | **Titer (CFU/ml)** |
|---|---|---|
| 1:1 | 88 | 8.8*$10^3$ |
| | 46 | 4.6*$10^3$ |
| 1:10 | 6 | 60 |
| | 3 | 30 |

## 3.2 Selection of transformed *E. coli* clones on selective hydrocarbon containing media

The selection for the ability of transformed *E. coli* to use hydrocarbons as sole source of carbon and energy was done by plating the clone library onto M9-C medium with six different concentrations of diesel (0.05%, 0.5%, 1%, 1.5%, 2% and 5% D). Growth was only observed at concentrations 0.05%, 0.5% and 1% D. Due to these results, subsequent cultures were screened for growth only on selection plates containing diesel of 0.05%, 0.5% and 1% D and CO.

After six days of incubation at 37ºC, every 0.05% M9-CD plate had visible bacterial growth. For concentration 0.5% D, growth was visible in every plate except for one plate. However, the growth on those plates was less pronounced than at concentration 0.05% D. Lastly, for 1% M9-CD plates, small colonies were visible on four out of five plates. At this concentration, the petri dishes showed some damage where the plastic became cloudy in appearance which made it difficult to observe bacterial growth.  Examples of these plates and their growth can be seen in figure 10.

Similarly, clone libraries were plated and incubated at concentrations 0.05%, 0.5% and 1% of crude oil (M9-CCO medium). For M9-CCO plates, eight days incubation at 37 ºC were necessary to observe bacterial growth, two more days than in M9-CD plates. Plates with concentration 0.05% CO presented similar growth as for 0.05% D plates, described above. For concentrations 0.5% and 1% CO, brown micelles of oil formed in the plates, which made it difficult to spot the colonies (Figure 10). However, visibility was sufficient to confirm bacterial growth and to select isolated colonies for further analysis.

To obtain more biomass of selected *E. coli* clones, single colonies were picked from M9-CD and M9-CCO plates, transferred to LB-C plates and incubated for two days at 37 ºC. Bacterial growth was observed in every LB-C plate (Figure 10).

In respect to control cultures, for 0.05%, 0.5% and 1% M9-CD plates with an extra carbon source (0.2% maltose) a very similar growth to M9-CD plates was observed. The *E. coli* strain LE392 was used as a control for the *E. coli* EPI300T1$^R$ supplied by the kit. Both strains, *E. coli* LE392 and the *E. coli* EPI300T1$^R$, presented no growth without a fosmid vector in media with chloramphenicol. Control plates LB-C and M9-C without hydrocarbons in the media presented bacterial growth. However, growth was less visible than in 0.05% M9-CD plates. Control parameters used and pictures of control plates are shown in figure 11.

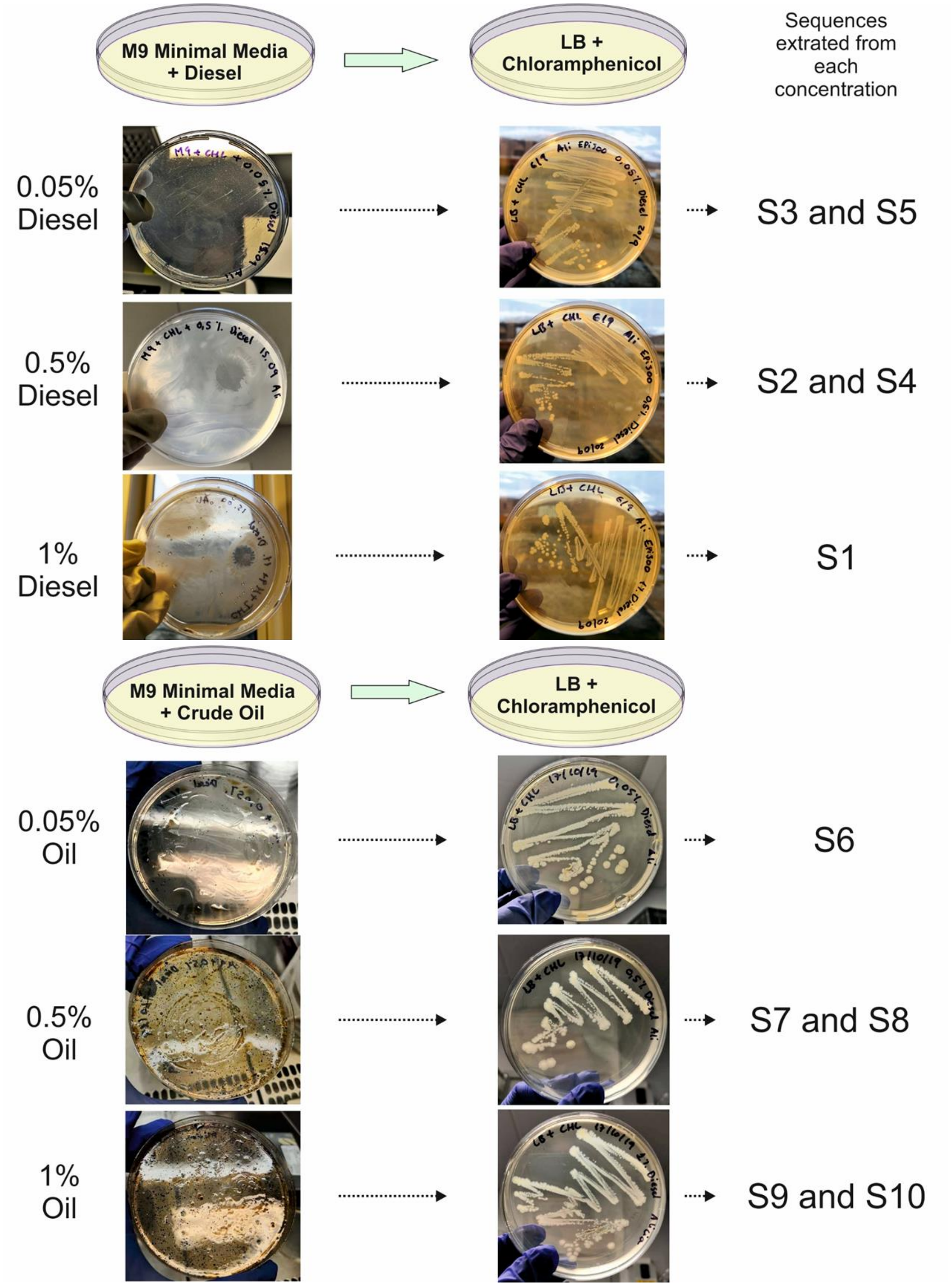


*Figure 10. Cultures of transformed E. coli in 0.05%, 0.5% and 1 % M9-CD (M9 minimal media + Diesel), cultures in 0.05%, 0.5% and 1 % M9-CCO (M9 minimal media + Crude Oil) and in correspondent LB-C media (LB + chloramphenicol). The source of Sequences 1 to 10 (later analyzed with bioinformatics tools) is also shown. Sequence 1 from 1% diesel cultures, Sequences 2 and 4 from 0.5% diesel cultures and Sequences 3 and 5 from 0.05% diesel cultures, Sequence 6 from 0.05% oil cultures, Sequences 7 and 8 from 0.5% oil cultures and Sequences 9 and 10 from 1% oil cultures.*

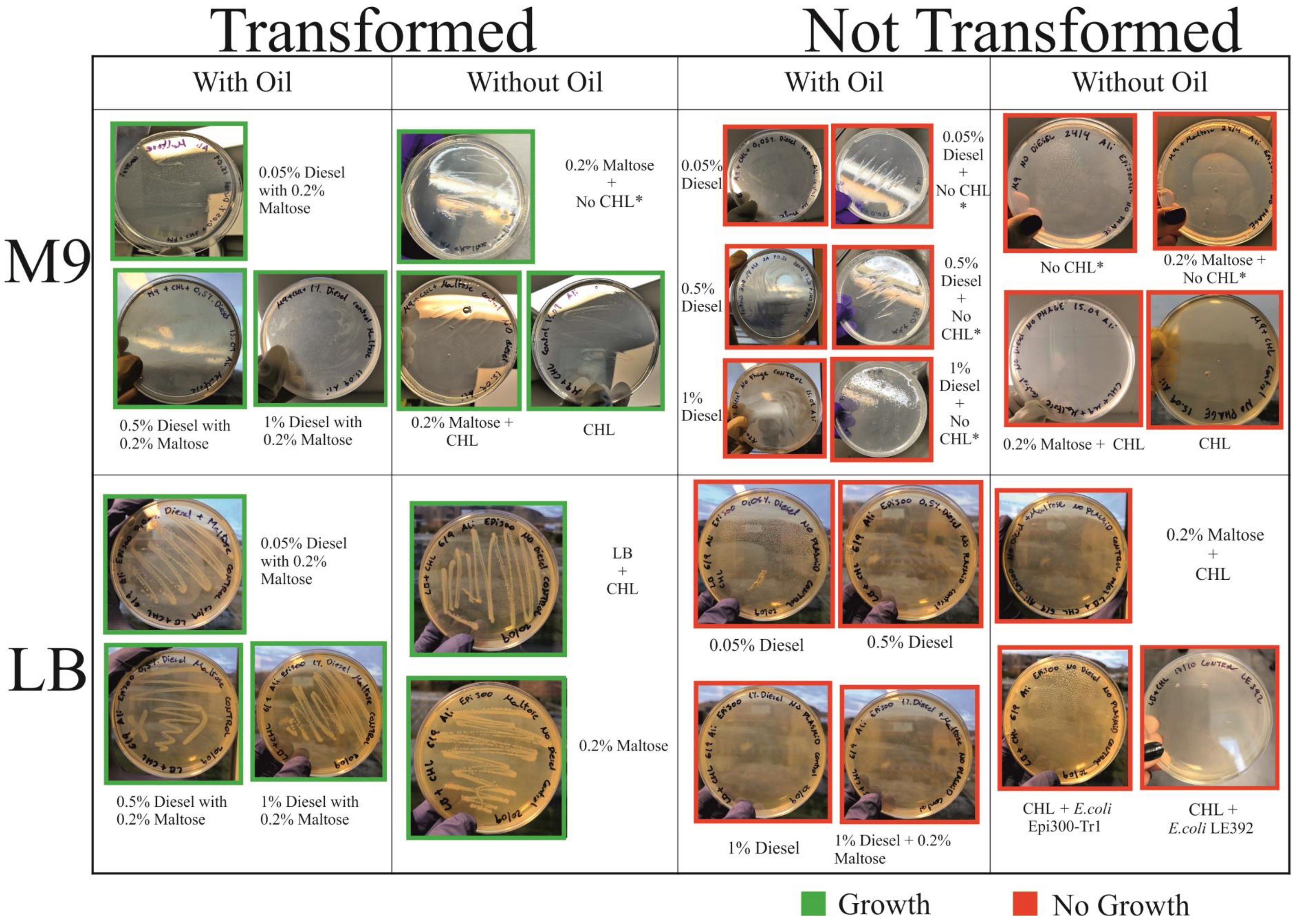


*Figure 11. Control plates for transformed EPI300T1$^R$* E. coli *in M9 media with diesel/oil and LB media with chloramphenicol (CHL). Control* E. coli *transformed or not transformed were plated in M9 minimal media and LB media. Green squares around the photographs indicate bacterial growth, red squares indicate no visible bacterial growth. A 0.2% Maltose dilution was used as substitute of the carbon source for diesel/crude oil.* Escherichia coli *LE392 was used as a substituted for* E. coli *EPI300T1$^R$. *Plates with no chloramphenicol added.*

### 3.2.1 Fosmid clone DNA extraction and restriction enzyme analysis

The fosmid DNA was extracted from the *E. coli* cells grown with hydrocarbons. A total of 20 fosmid extractions were performed on colonies from M9-CD media and six extractions on colonies from M9-CCO media. As an example, a gel electrophoresis of 10 of the fosmid extraction from M9-CD cultures can be seen in figure 12. Subsequently, the fosmids were digested with the restriction enzyme *XbaI* to verify that the clones contained different DNA fragments.

The extraction of fosmids from M9-CD grown *E. coli* clones yielded concentrations of DNA between 48.2 ng/µl and 3314.6 ng/µl .The fosmid extraction from M9-CCO cultures presented concentrations between 739.2 ng/µl and 4558.7 ng/µl (Table 8).

The digestion of the fosmids with the restriction enzyme *XbaI* can be observed in the gel electrophoresis shown in figure 13 for M9-CD cultures and in figure 14 for M9-CCO cultures. The restriction bands of fosmids coming from M9-CD and M9-CCO show different patterns, varying from one to three bands of different sizes, indicating different restriction sites in the sequences. We decided to continue our study with 10 fosmids in total, five from each type of hydrocarbon culture. The fosmids selected to be sequenced were extractions 1, 6, 8, 15, and 20 from M9-CD cultures and extractions 1, 2, 3, 5, and 6 from M9-CCO cultures. .

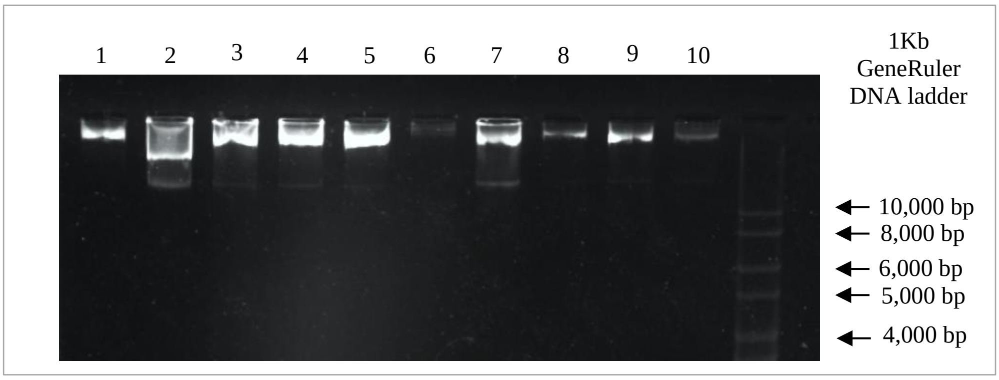


*Figure 12. Gel electrophoresis of fosmid extraction. Samples 1 to 10 from diesel cultures*

*Table 8. Absorbance (230) of fosmid extractions from diesel and crude oil cultures*

| **Diesel** | | | | **Crude oil** | |
|---|---|---|---|---|---|
| **Fosmid extraction** | **Abs 230 (ng/µL)** | **Fosmid extraction** | **Abs 230 (ng/µL)** | **Fosmid extraction** | **Abs 230 (ng/µL)** |
| **1** | 607.5 | **11** | 779.1 | **1** | 2514.2 |
| **2** | 369.9 | **12** | 1941.5 | **2** | 1349.9 |
| **3** | 244.9 | **13** | 1477.7 | **3** | 1385.6 |
| **4** | 3290.7 | **14** | 796.8 | **4** | 1075.8 |
| **5** | 1222.4 | **15** | 1001.4 | **5** | 739.2 |
| **6** | 67.3 | **16** | 3314.6 | **6** | 4558.7 |
| **7** | 546.8 | **17** | 2960.1 | | |
| **8** | 336.8 | **18** | 351.7 | | |
| **9** | 389.0 | **19** | 469.4 | | |
| **10** | 48.2 | **20** | 2560.6 | | |

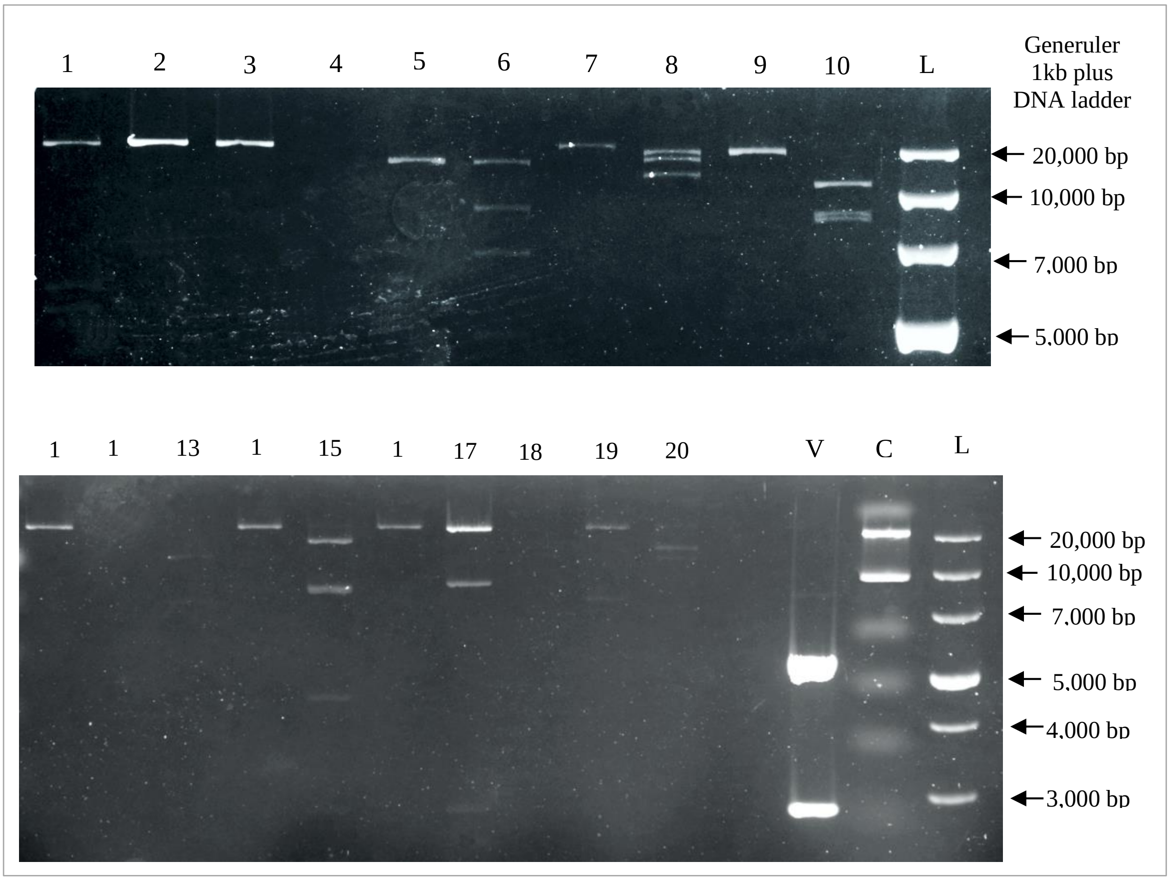


*Figure 13. A 1% agarose gel electrophoresis of 20 fosmid extractions coming from M9-CD cultures digested by the enzyme XbaI. L is the DNA ladder generuler 1Kb plus from Thermofisher. V is the vector without the insert DNA. C is the 40Kb control DNA from the CopyControl kit with the vector. Differences in the band pattern are observed through the plasmid samples.*

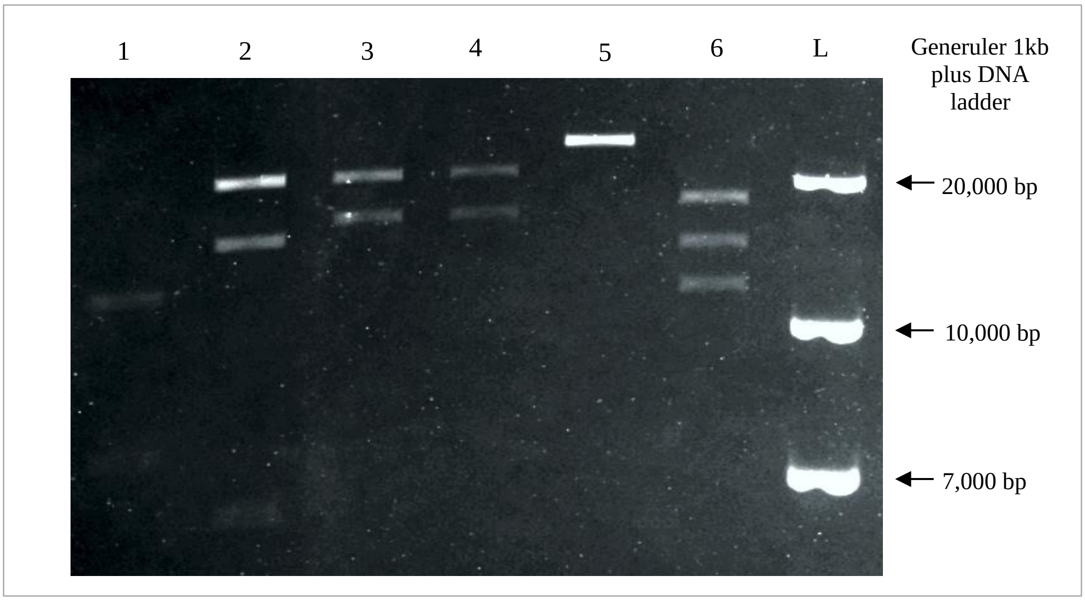


*Figure 14. A 1% agarose gel electrophoresis of six fosmid coming from M9-CCO cultures digested with enzyme XbaI. L is the DNA ladder 1Kb plus from Thermofisher. Differences in the band pattern are observed through the plasmid samples.*

## 3.3 Bioinformatic results

### 3.3.1 Preprocessing of raw data

The fosmid DNA was sequenced and the data received was divided in ten files, one for each extraction. Sequences 1 to 5 correspond to diesel cultures and sequences 6 to 10 correspond to crude oil cultures. The sequencing data was uploaded to the software Geneious 2020.2.2 [3]. Information on raw data is presented in table 9. The information on the data includes Phred quality score (Q) expressed as percentage of bases with a score higher than 30 and percentage of bases with a score higher than 40. The Phred quality score is a measure of the accuracy of the sequencing platform, it does so by indicating the probability of a base being called incorrectly by the sequencer. A Q of 30 is equivalent to the probability of an incorrect base call being 1 in 1,000 which corresponds with a 99.9% of base call accuracy. A Q of 40 corresponds with 1 in 10,000 probability of an incorrect base call, 99.99% of base call accuracy [68]. After the initial treatment of the sequences, it can be observed an improvement on the Phred quality score, from 0 % of bases with $Q \geq 40$ to approximately 80% of bases with $Q \geq 40$ (Table 10).

### 3.3.2 Assembly of contigs

A *De novo* assembly was performed to assemble the reads into contigs. The longest contig assembled in each sequence, or contig 1, had a size between 37 kb to 50 kb (Table 11), which was the expected length of the pCC2FOS vector plus the insert DNA. The size of the pCC2FOS vector is 8.1 kb and the optimal size for insert DNA is between 30 to 45 kb [5]. Sequence 7 was considered an exception since it presented three contigs with a combined length of 42 kb (Table 11). The three contigs were analyzed together. Information on contig length, quantity, CG content, and Phred score is on table 10. These contigs were considered to contain the insert DNA of interest and became the focus of the study. They were named "Sequence + corresponding number". The remaining contigs assembled presented a size ranging from 2kb to 40 bp, a MEGABLAST search [59] showed a high homology, 90% to 100% of pairwise identity, with *E. coli* genes, consequently they were considered contamination from the host *E. coli* and were not studied further.

Following the *de novo* assembly, a multiple alignment was performed by "Clustal omega 1.2.2" [56] to seek similitudes among the sequences. The alignment between Sequences 1 to 10 did not show any significant homology or similarities among the sequences.

*Table 9. Information on raw sequencing data. Number of reads in each sequence, the maximum, minimum, and mean length of the reads, the CG content and Phred score expressed as Q ≥ 30 and Q ≥ 40*

| **Sequence number** | **Reads** | **Max. Length (bp)** | **Min. Length (bp)** | **Mean Length (bp)** | **% CG** | **% bases Q ≥ 30** | **% bases Q ≥ 40** |
|---|---|---|---|---|---|---|---|
| 1 | 3,366,946 | 151 | 35 | 108.0 | 53.9 | 95.5 | 0.0 |
| 2 | 2,562,348 | 151 | 35 | 109.3 | 41.6 | 95.9 | 0.2 |
| 3 | 2,767,748 | 151 | 35 | 108.9 | 52.6 | 94.8 | 0.0 |
| 4 | 3,459,956 | 151 | 35 | 108.6 | 53.3 | 95.5 | 0.0 |
| 5 | 2,226,542 | 151 | 35 | 112.7 | 48.0 | 95.5 | 0.1 |
| 6 | 3,124,356 | 151 | 35 | 118.8 | 52.3 | 95.8 | 0.0 |
| 7 | 2,549,746 | 151 | 35 | 105.5 | 55.0 | 93.7 | 0.0 |
| 8 | 2,991,748 | 151 | 35 | 105.4 | 64.9 | 91.4 | 0.0 |
| 9 | 3,457,984 | 151 | 35 | 109.8 | 52.1 | 95.3 | 0.0 |
| 10 | 2,210,074 | 151 | 35 | 120.0 | 53.4 | 95.5 | 0.0 |

*Table 10. Information on sequencing data after the preprocessing treatment. Number of contigs assembled, maximum, minimum, and mean length of the contigs, CG content and Phred score expressed as Q ≥ 30 and Q ≥ 40.*

| **Sequence** | **Contigs** | **Max. Length (bp)** | **Min. Length (bp)** | **Mean Length (bp)** | **% CG** | **% bases Q ≥ 30** | **% bases Q ≥ 40** |
|---|---|---|---|---|---|---|---|
| 1 | 508 | 42,826 | 47 | 343.0 | 53.3 | 100.0 | 83.1 |
| 2 | 5,698 | 37,621 | 41 | 354.4 | 52.4 | 99.9 | 78.9 |
| 3 | 97 | 50,594 | 69 | 839.2 | 52.7 | 100.0 | 91.4 |
| 4 | 2,542 | 43,581 | 44 | 293.7 | 52.7 | 99.9 | 79.2 |
| 5 | 4,025 | 42,123 | 62 | 1189.2 | 51.1 | 100.0 | 93.9 |
| 6 | 1,279 | 39,380 | 54 | 342.3 | 50.3 | 99.9 | 76.3 |
| 7 | 49 | 19,920 | 51 | 1157.4 | 53.9 | 100.0 | 95.1 |
| 8 | 3,723 | 40,759 | 41 | 300.5 | 52.8 | 99.9 | 78.6 |
| 9 | 4,701 | 40,847 | 51 | 326.5 | 52.4 | 99.9 | 78.7 |
| 10 | 56 | 47,316 | 98 | 1221.8 | 52.8 | 100.0 | 92.7 |

*Table 11. Information on contigs assembled by* de novo *assembly. Information includes length of the contigs, CG content, Phred score (Q ≥ 40), rough melting temperature and the number of ORFs predicted in each contig.*

| Sequence number | Contig number | Length (bp) | % CG | % bases Q ≥ 40 | Rough Tm (ºC) | ORFs |
|---|---|---|---|---|---|---|
| 1 | 1 | 42,826 | 53.7 | 99.9 | 91.3 | 79 |
| 2 | 1 | 37,621 | 37.7 | 100.0 | 84.8 | 41 |
| 3 | 1 | 50,594 | 52.2 | 99.9 | 90.7 | 56 |
| 4 | 1 | 43,581 | 53.1 | 100.0 | 91.1 | 51 |
| 5 | 1 | 42,123 | 43.2 | 99.7 | 87.0 | 62 |
| 6 | 1 | 39,380 | 52.7 | 100.0 | 90.9 | 57 |
| 7 | 1/2/3 | 19,920/ 13,170/ 8,136 | 55.7/51.7/55.7 | 99.9/98.9/100 | 92.1/90.4/92.1 | 44 |
| 8 | 1 | 40,759 | 67.2 | 100.0 | 96.9 | 41 |
| 9 | 1 | 40,847 | 51.7 | 99.6 | 90.5 | 40 |
| 10 | 1 | 47,316 | 53.6 | 99.8 | 91.3 | 56 |

### 3.3.3 Gene prediction and analysis

The Glimmer [57] application on "Geneious 2020.2.2” [54]  was used to predict ORFs in Sequences 1 to 10. The number of ORFs predicted in a sequence range between 40 and 79, with an average of 53 ORFs per sequence. The number of predicted ORFs is in table 11 and a graphical representation of their position in the sequences can be seen in figures 20 to 26.

The MEGABLAST, BLASTn and BLASTx [59] searches paired the majority of predicted ORFs with genes and protein from the NCBI databases [58]. The results of these searches can be seen in figures 20 to 26 and detail tables are presented in appendix 6. Notably, the

BLASTn searches permitted the correlation of Sequences 1 to 10 to a specific bacteria fila or genus. Sequences 1, 3, 7 and 10 present high homologies with *Planctomycetes bacterium,* Sequence 2 with *Maribacter sp.*, Sequences 4 and 6 with *Synechococcales cyanobacterium,* Sequence 5 with *Klebsiella sp.*, Sequence 8 with *Pseudomonas sp.*, and Sequence 9 with *Chromatiales bacterium.* The bacterial species and families with a pairwise identity higher than 85% with ORFs from the sequences of the study are shown in figures 16 and 17.

Once the BLAST tools linked the predicted ORFs to known proteins, a manual search in online databases was performed to research the functions of these proteins. This manual search permitted the classification of the ORFs by their predicted function. The predicted functions are annotated in figures 20 to 26, a more detail description can be found in appendix 6. The general categories used to classify the genes and proteins are biosynthesis or catalysis of any substance, involvement in a hydrocarbon degradation pathway, response to stress, toxicity, and repair; mobile genetic elements, transport in or out of the cell, regulation, signaling for other cells or itself, protein modification, flagella and pili, domain of unknown function, and hypothetical protein.

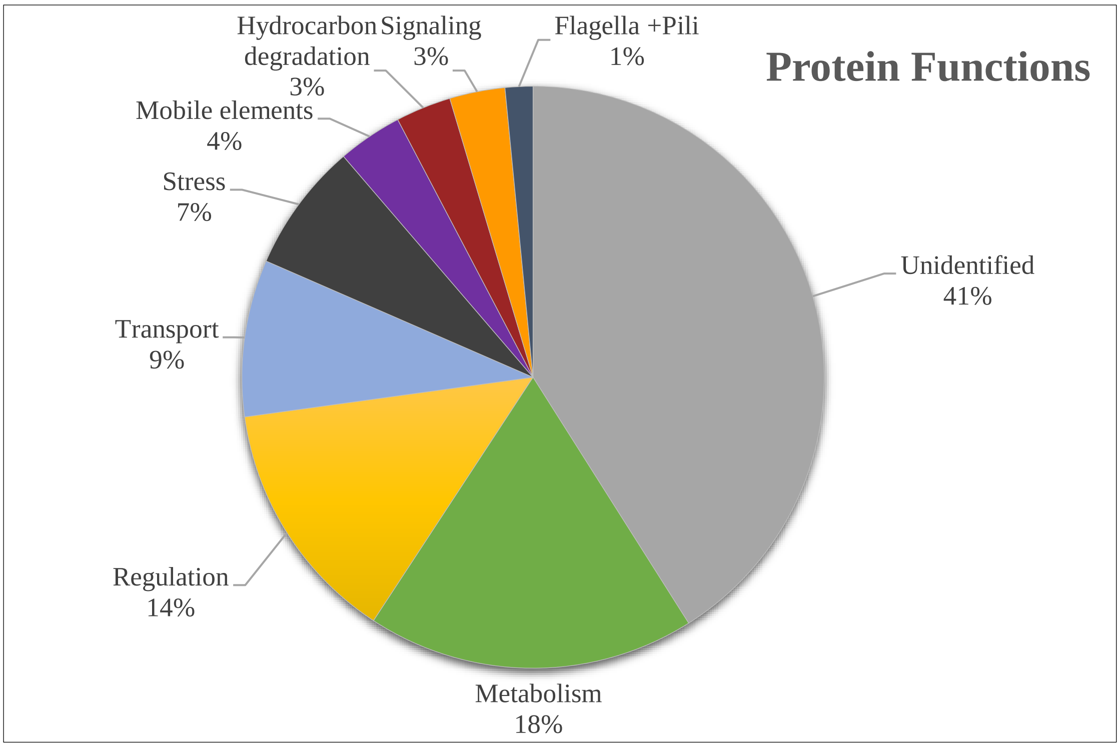


*Figure 15. Percentage of ORFs with certain protein functions in Sequences 1 to 10.*

A summary of the functions found is in figure 15. The majority of ORFs (41%) were assigned as hypothetical proteins, DUFs or  showed no homology with a known protein. Metabolism, regulation, transport, and stress followed with 18%, 14%, 9% and 7% of ORFs, respectively. Significantly, a 3% of ORFs were predicted to be involved in hydrocarbon degradation.

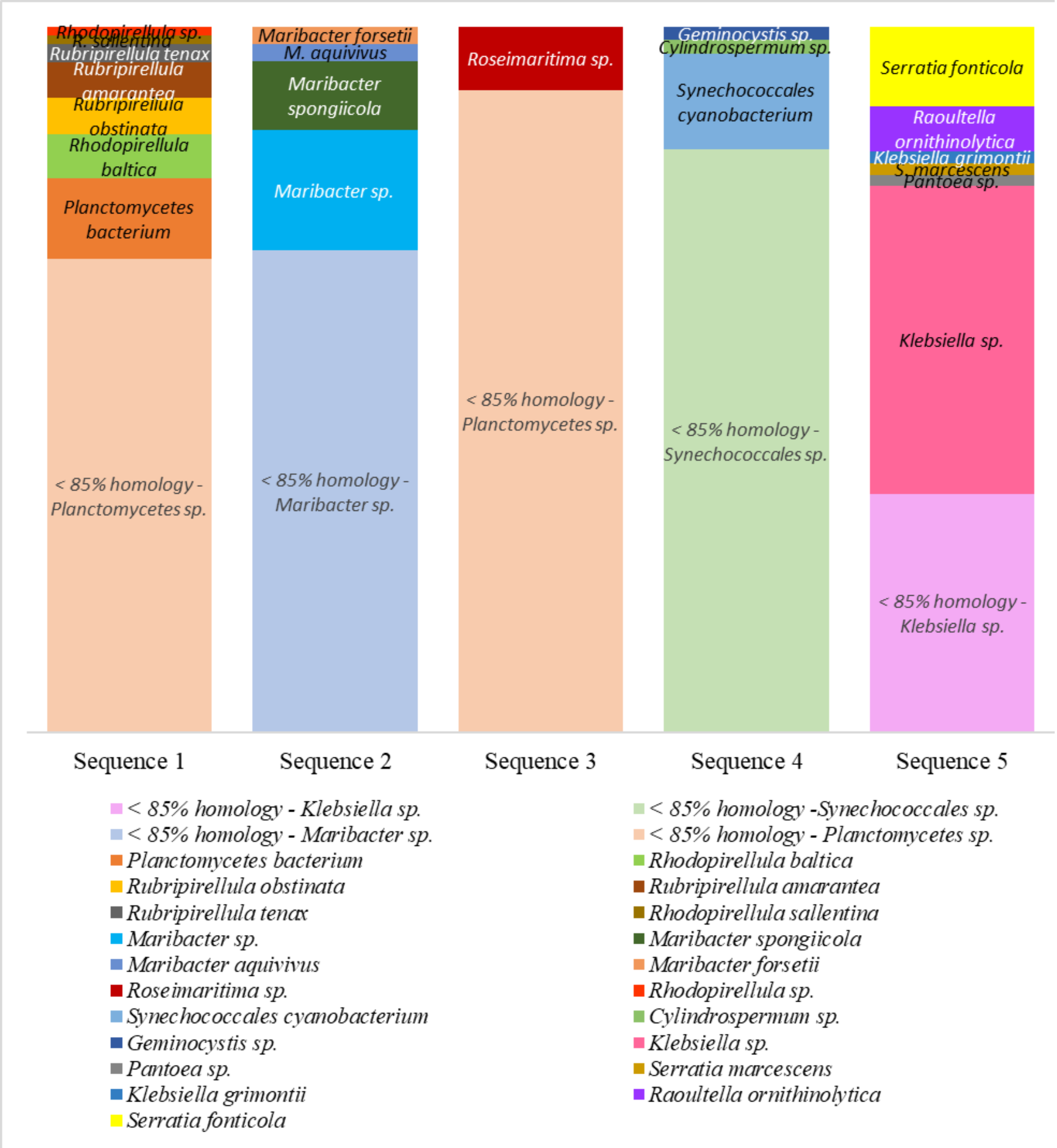


*Figure 16. Representation of the main bacteria phyla and species linked to Sequences 1 to 5. Columns represent the total ORFs in each sequence. At the bottom of the columns are the ORFs with a pairwise identity below 85%, most of them belonged to the same phylum. Below the columns are the bacterial species with a pairwise identity higher than 85% with the ORFs.*

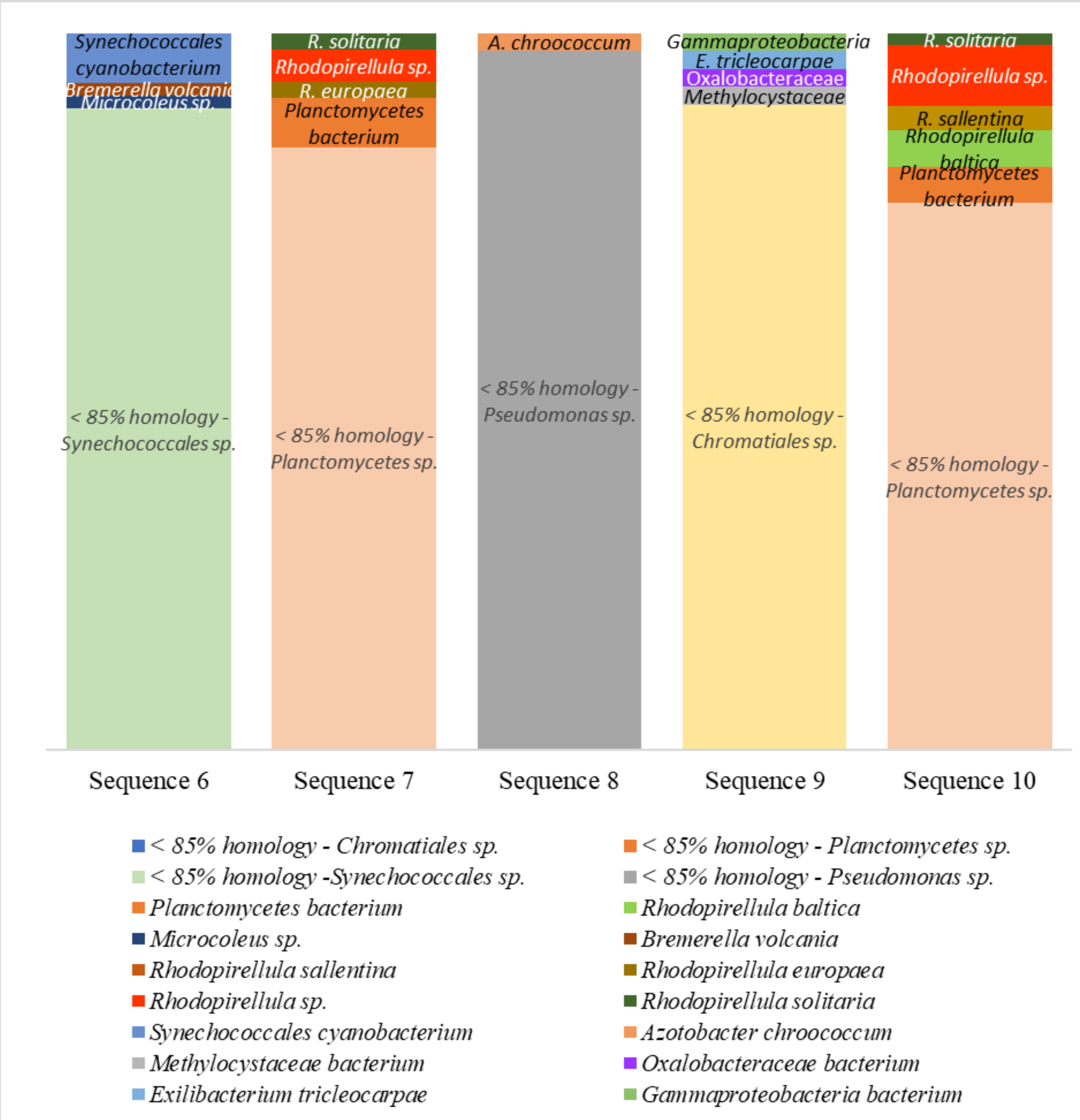


*Figure 17. Representation of the main bacteria phyla and species linked to Sequences 6 to 10. Columns represent the total ORFs in each sequence. At the bottom of the columns are the ORFs with a pairwise identity below 85%, most of them belonged to the same phylum. Below the columns are the bacterial species with a pairwise identity higher than 85% with the ORFs.*

### 3.3.4 Phylogenetic position of the fosmid clones

To determine the phylogenetic relationships between the fosmid library and the metagenomic DNA from the oil-polluted environment, a phylogenetic tree was built. The phylogenetic tree was constructed using 491 sequences of 16S ribosomal RNA (rRNA). The sequences included nine bacterial phyla and genera with high homology to Sequences 1 to 10 and 482 sequences from the bacterial community used to make the metagenomic fosmid library, these sequences were provided by a previous study [42].

The bacterial phyla or species with high homology to the sequences analyzed were: *Rhodopirellula baltica* (S1, S7 and S10), *Planctomycetes sp.* (S1 S3 and S7), *Maribacter sp.* (S2), *Roseimaritima ulvae* (S3), *Synechococcales sp.* (S4 and S6), *Serratia fonticola* (S5), *Klebsiella sp.* (S5), *Gemmatirosa kalamazoonesis* (S8), and *Granulosicoccus antarcticus* (S9). The 16S rRNA sequences of these species were downloaded from the BLAST gene database [59].

A phylogenetic tree was built with the program Geneious 2020.2.2 [3] and edited with the online tool "iTOL v5.6.3"[64]. The tree was made with 16S rRNA sequences from bacteria collected at Murmansk seaport [42] and bacteria species representatives of the sequences of the study. The full tree was composed by 491 branches, branches from environmental samples have the name "SP-number" and the representative bacteria are referred as the name of the species follow by the sequence represented (S-number). The selected bacterial species are distributed among bacteria samples collected at Murmansk in the phylogenetic tree.

The phylogenetic tree was edited by collapsing tree nodes which did not include any of the representatives for the sequences of the study, of the 491 branches only 51 remained shown. Representatives of the sequences were classified by phylum: Gammaproteobacteria including *Serratia fonticola* (S9), *Klebsiella sp.* (S5) and *Granulosicoccus antarcticus* (S9); Gemmatimonadales with *Gemmatirosa kalamazoonesis* (S8); Cyanobacteria having *Synechococcales cyanobacterium* (S4, S6); Planctomycetes including *Planctomycetes bacterium* (S1, S3, S7), *Rhodopirellula baltica* (S1, S, S10) and *Roseimaritima ulvae* (S3) and Flavobacteriales with *Maribacter sp.*(S2). Edited phylogenetic tree can be seen in figure 18 and full tree can be seen in appendix 7.

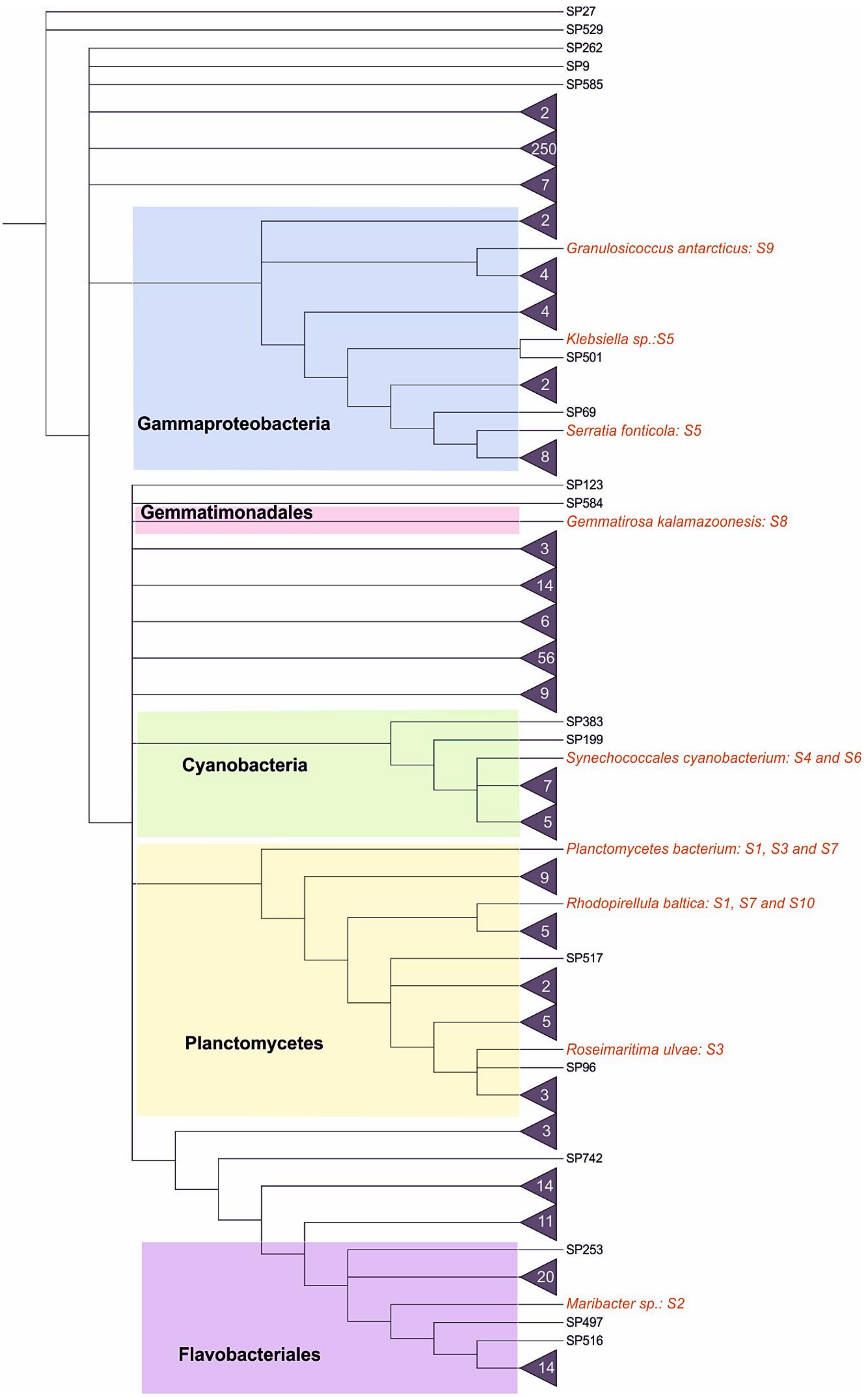


*Figure 18. Phylogenetic tree representing the bacterial metagenome of a hydrocarbon-contaminated site. In red are nine bacterial species added as representatives of the sequences of the study. Collapse tree nodes are represented by triangles with the number of branches collapsed. The phylogenetic tree was built by the Geneious Tree Builder [3], parameters used were: Genetic Distance Model: Tamura-Nei, Tree building method: Neighbor-Joining, No Outgroup, Consensus Tree Options: Resample tree, Resampling Method: Bootstrap, Random Seed: 321,056, Number of replicates: 100, Create consensus tree, Support Threshold %: 50*

### 3.3.5 Analyses by custom BLAST on AromaDeg database and other online databases.

To look for proteins involved in aromatic hydrocarbon degradation, a custom BLAST against the database AromaDeg [65] was performed. Several matches were obtained, however only results with a pairwise identity higher than 70% were annotated (Table 12).

*Table 12. Proteins from database AromaDeg [65] with a pairwise identity 70% or higher with translated amino acids sequences from Sequences 1 to 10. Results include ORFs position in the sequence, name of the protein, bacterial species in which is present, and pairwise identity between proteins.*

| **Sequence** | **ORF** | **Protein** | **Species** | **Pairwise identity (%)** |
|---|---|---|---|---|
| 4 | 61 | VOC family protein | *Pseudomonas fluorescens* | 70.6 |
| 5 | 39 | aromatic 1,2-dioxygenase, alpha subunit | *Ruegeria pomeroyi* | 80.0 |
| 7 | 33 | benzoate 1,2-dioxygenase large subunit | *Alishewanella jeotgali* | 83.3 |
| 7 | 36 | hypothetical protein (record was removed by NCBI) | *Streptomyces prunicolor* | 70.0 |
| 8 | 66 | aromatic-ring-hydroxylating dioxygenase subunit alpha | unclassified *Pseudonocardia* | 83.3 |
| 9 | 45 | catechol 2,3-dioxygenase | *Roseiflexus castenholzii* | 71.4 |
| 10 | 83 | dioxygenase | *Bordetella avium* | 80.0 |

The online server AntiSMASH [67] gave no significant matches to gene clusters involved in biosynthesis of secondary metabolites. The server MG-RAST [66] was used as a guide of which tools to use in the software Geneious [3].

# 4 Discussion

## 4.1 Selection of fosmid clones

After a successful isolation of the metagenomic DNA and the production of the fosmid library, hydrocarbon cultures were analyzed to observe whether they indeed contained genes of interest. Cultures on M9-CD and M9-CCO presented growth of transformed cells in the presence of different concentrations of commercial hydrocarbons. This suggests that transformed *E. coli* cells acquired the capacity to use hydrocarbons as a source of carbon and energy or, at least, acquired the capacity to survive in hydrocarbon-rich environments. This is further supported by control cultures, where *E. coli* cells lacking the fosmid vector were not able to grow in hydrocarbon cultures, even when an additional carbon source was present. These findings suggest that the insert DNA conferred this ability to the cells. It is therefore likely that the insert DNA contains genes involved in hydrocarbon degradation.

It is interesting to note that the control M9 plates with maltose but lacking the selection antibiotic only permitted bacterial growth in transformed cells. A possible explanation might be that the fosmid allows faster cell growth by, for example, providing additional copies of essential genes. Cultures were incubated for the same amount of time, six days. However, *E. coli* presents slow growth in M9 media compared to other minimal media [45], non-transformed *E. coli* colonies might need more time to grow.

## 4.2 Crude oil and diesel cultures

It has been widely reported that microbial biodegradation of Petroleum (Crude Oil) and commercial diesel presents differences in rate, speed, and bacterial communities [13, 69, 70]. Although only 10 samples were analyzed in this study, there are a number of observable differences between D cultures and CO cultures. These differences include the dispersion in M9 media agar plates, the incubation time for visible colonies to appear, the distribution of protein functions, and the results in BLASTs searches.

The differences between M9-CD and M9-CCO media could lead to differences on the selection of fosmid clones. For instance, in CO cultures micelles of dark oil appeared on every plate, whereas in D cultures a thin pellicle of oil covered the surface of the plates. This disparity in oil distribution in the media could lead to differences in oil bioavailability for the cells, leading to the expression of different genes in the hydrocarbon biodegradation pathways. Additionally, hydrocarbon bioavailability could be a possible reason for the difference in incubation time needed to obtain visible colonies, since CO cultures needed two more days than D cultures [22].

To explore the possibility of a variation on the selected genes due to the hydrocarbon source, a comparison was made between protein functions in CO and D cultures (Figure 19). Interestingly, more proteins presented unidentified functions in CO cultures than in D cultures. A possible explanation for this might be the CO has a more complex composition than D [46], presenting a broader range of hydrocarbons as a food source for degraders. Possibly, a wider variety of genes are necessary to use CO as a carbon source, thus making the identification of every protein involved more challenging. Nonetheless, this presents more opportunities to find novel proteins and enzymes involved in hydrocarbon degradation.

Proteins involved in regulation are more prominent in diesel cultures, whether their presence is relevant for regulation of other genes in the clones or they are neighboring sequences for the essential genes remain to be investigated.  Furthermore, proteins involved in transport in and out the cell are more prominent in CO cultures than in D cultures, might be due to the necessity of making CO more available for the cells through biofilm formation or biosurfactant production. As shown by M. Omarova, et al., the formation of biofilms by the

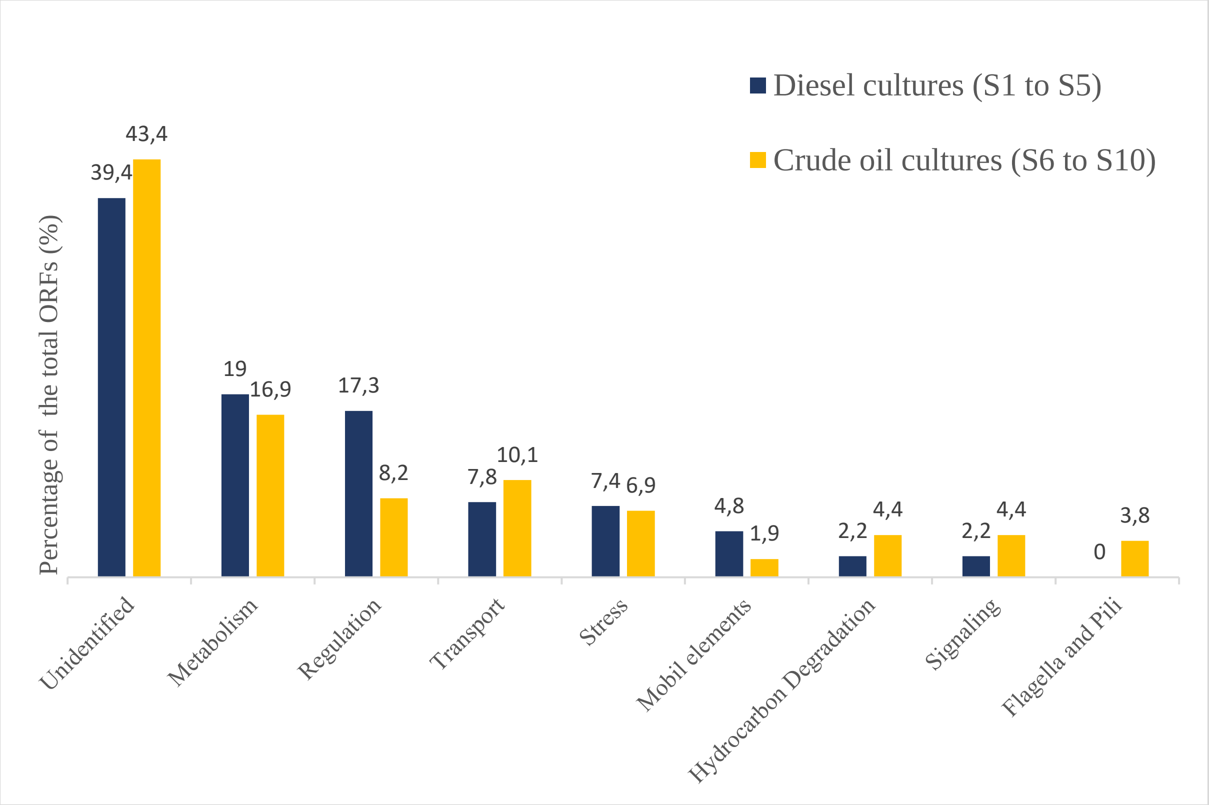


*Figure 19. Comparison of protein functions in diesel and crude oil cultures expressed as percentages of total ORFs found in Sequences 1 to 10.*

hydrocarbon degrader *Alcanivorax borkumensis* promotes the dispersion of an oil slick into stable droplets [71]. Several studies such as the ones by Q. Helmy, et al. with *Azotobacter vinelandii* [72], I. V. Nwaguma, et al. with *Klebsiella pneumoniae* [73] and Y. Huang, et al. with *Serratia marcescens* [74] show that biosurfactants produced by these bacteria increase oil dispersion and enhance bioremediation rates.

## 4.3 Phylogeny

The phylogenetic tree presented the bacterial species representatives of Sequences 1 to 10 distributed among the metagenomic DNA collected at the hydrocarbon-contaminated site. The bacterial representatives are classified in the class Gammaproteobacteria, the order Gemmatimonadales, the phylum Cyanobacteria, the phylum Planctomycetes and the order Flavobacteriales (Figure 18). As shown in the phylogenetic tree, sequences isolated in the background study also belong to these groups. Moreover, when observing the “chronic pollution” chart in figure 6, these classes, orders and phyla are among the most represented in the chronically hydrocarbon polluted environment. These suggest that the insert DNA of the fosmids was isolated correctly and belongs to native bacteria from a chronically oil-polluted environment. Although some of the expected genera did not appeared in the analyzed samples, such as important hydrocarbon degrader *Actinobacteria* [13], the bacterial species present in this study could be considered a good representation of a bacterial community in a low temperature, oil-polluted marine environment.

## 4.4 Analysis of ORFs and Bacteria species

### 4.4.1 Sequences 1, 3, 7 and 10: Phylum *Planctomycetes*

The majority of the predicted ORFs in Sequences 1, 3, 7 and 10 presented high homologies with the phylum Planctomycetes. This Phylum possesses several interesting characteristics, such as a large genome, cryptic morphology and cell compartmentalization, a unique feature for prokaryotes. Furthermore, Planctomycetes are present in a large variety of environments, including important roles in marine snow, degradation and in biofilms formation on diatoms and macroalgae. Macroalgae biofilms present a high number and diversity of Planctomycetes. Therefore, a broad representation of sequences originated from Planctomycetes is not coming as a surprise [75-77].

Starting the ORF analysis with Sequence 1, the genus *Rubripirellula* was notably represented. This genus presented high homology with a total of 16 ORFs in Sequence 1. *Rubripirellula* *sp.* is a relatively new genus, described for the first time in 2015 [78]. In particular, two species represented in Sequence 1 were *Rubripirellula amarantea* and *Rubripirellula tenax,* described as recently as 2019 [79]. This genus has been found forming epiphytic biofilms on macroalgae with other members of the same phylum, such as *Roseimaritima ulvae* and *Mariniblastus fucicola,* both present in this study [80]. Interestingly, no literature has been found linking *Rubripirellula sp.* and hydrocarbon biodegradation or biosurfactant production. However, certain proteins linked to ORFs of Sequence 1 might suggest *Rubripirellula sp.* could have a role in hydrocarbon biodegradation. For instance, the ORF-100 which presents 97.7% of pairwise identity with a protein belonging to the vicinal oxygen chelate (VOC) family protein from *R. amarantea*. A protein superfamily including dioxygenases and bleomycin resistance proteins which have been associated with biodegradation of PAHs as a result of cleaving the polyaromatic ring with oxygen molecules (Figure 27) [81-83].

Sequence 3 is characterized for the genus *Roseimaritima.* Isolated for the first time with *Rubripirellula sp.,* forming part of epiphytic biofilms in macroalgae [78, 80]. Similarly to the genus *Rubripirellula,* not much literature can be found on *Roseimaritima.* Nevertheless, one study in 2020 observed *Roseimaritima sp.* was part of a bacterial community associated with the benthic diatom *Nitzschia sp.* The association increased the diatom degradation efficiency of benzo(a)pyrene and fluoranthene, a PAH [84].

An interesting set of four ORFs in Sequence 3 (ORFs-055, 56, 59 and 60) are predicted to participate in fatty acids biosynthesis, the proteins are beta-ketoacyl-ACP synthase II, Acyl carrier protein (ACP), 3-oxoacyl-[ACP] reductase and ACP S-malonyltransferase. These proteins might be involved in the last step of aerobic degradation of n-alkanes. In this metabolic pathway, the first step would be the oxidation or “activation” of a terminal methyl group to form a primary alcohol, which then would be oxidized to an aldehyde by an alcohol dehydrogenase (most of them contain PQQ as a prosthetic group) and the last steps are the conversion to fatty acids and the entry into the TriCarboxylic Acid (TCA) cycle [85]. Additionally, the ORF-069 could be also linked to hydrocarbon degradation. It presented high homology with a PQQ-binding-like beta-propeller repeat protein found in Planctomycetes. A type of proteins which can be involved in metabolic pathways of hydrocarbon degradation, such as butanoate metabolism, propanoate metabolism, benzoate degradation via hydroxylation and benzoate degradation via CoA ligation [86]. For a graphical representation of hydrocarbon degradation and the ORFs involved refer to figure 27.

Sequences 7 and 10, both from CO cultures, present a majority of ORFs linked to the genus *Rhodopirellula.* To date, there are no studies linking hydrocarbon degradation to this genus [39]. However, in the genome of *Rhodopirellula baltica* (Sequences 1 and 10) genes for cytochrome P450 mono-oxygenase and epoxide hydrolase were observed [87] . These genes possess an important role in oxygenation or “activation” of alkenes, long-chain non-methane alkanes and aromatic hydrocarbons. Oxygenation constitutes the first step in hydrocarbon biodegradation [19, 85, 88-90]. It can thus be suggested that *R. baltica* could be a potential new bacterial hydrocarbon degrader.

In Sequence 7, ORF-033 which did not present any matches in BLAST searches was found to have 83.3% of pairwise identity with the protein benzoate 1,2-dioxygenase large subunit when a search in the AromaDeg database was conducted. This protein performs PAHs oxygenation [19]. Furthermore, ORF-019 and ORF-021 presents homology with a methanol dehydrogenase regulatory protein and a SDR family oxidoreductase, respectively. These proteins are involved in the oxygenation or activation of hydrocarbons [19, 85, 88-90].

In Sequence 10, ORF-083 and ORF-069 have attracted our attention in particular. The ORF-083 presents homology with a hydrocarbon dioxygenase protein, which can activate the first step of hydrocarbon degradation as explain above [19, 85, 88-90]. In ORF-069 the predicted protein is a Gfo/Idh/MocA family oxidoreductase, a structural family which contains enzymes that catalyze oxidation of trans-dihydrodiols, PAHs [58].

Unexpectedly, proteins for pili and flagella were found at the extremes of Sequence 10. The species *R. baltica* displays two life forms, one sessile forming part of macroalgae biofilms and one motile with pili and flagella. Under stress conditions, such as high salinity, the pili and flagella genes are expressed, an example is the study done by P. Wecker, et al. [91] where *R. baltica* SH1T expressed genes of the pili apparatus when exposed to a high salinity environment. In this case, it could be a response for high hydrocarbon concentration in the environment. Another possible explanation might be that the bacteria uses the flagella and pili to adhere to the oil molecules, a way to increase physical contact with the hydrocarbons, an explanation suggested in the review by X. Xu, et al., [22].

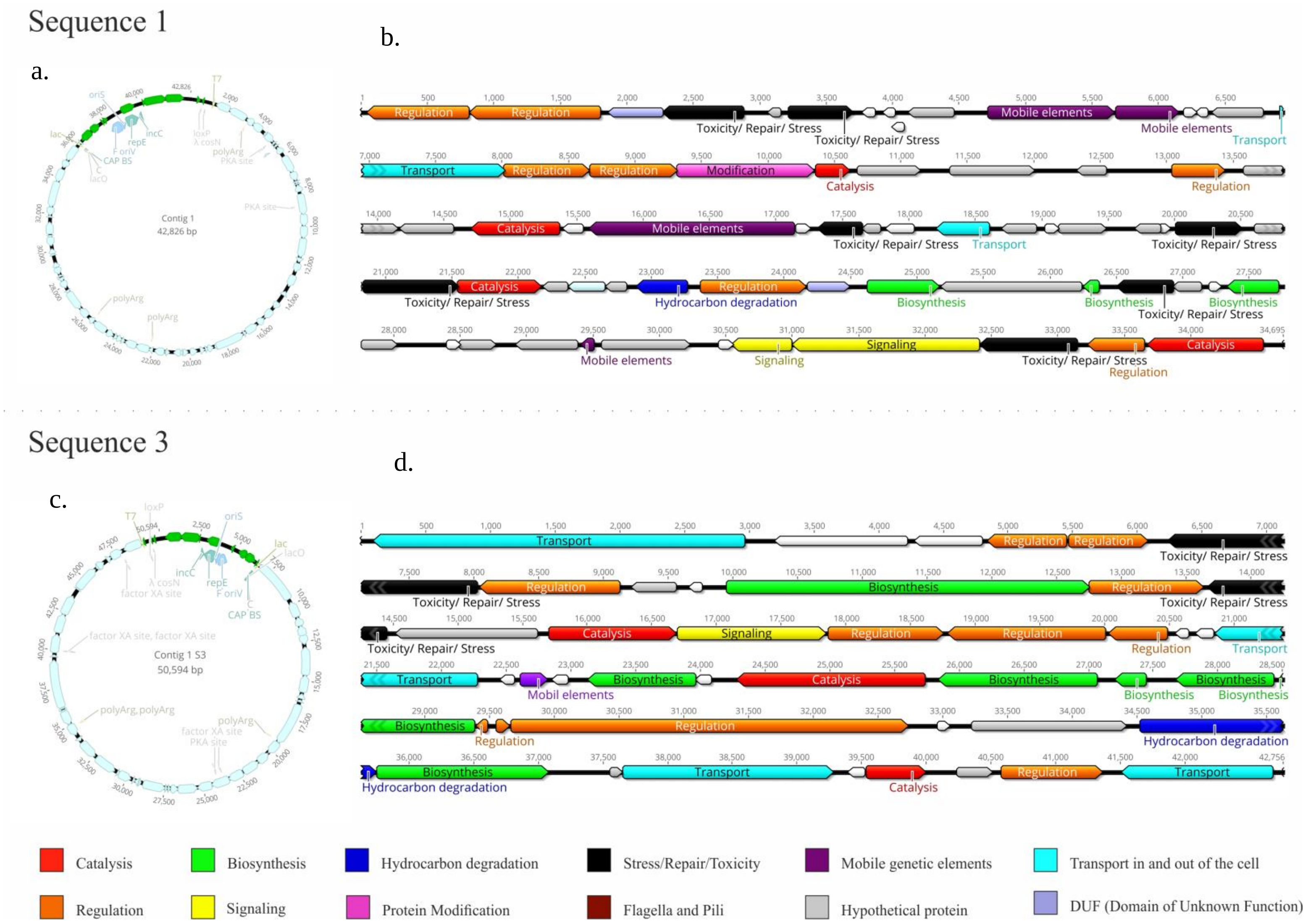


*Figure 20. Graphical representation of Sequences 1 and 3. Figures a and c are a representation of the re-circularized contigs with the pCC2FOS vector in green and ORFs in white. Figures b and d represent the linearized sequences with the predicted functions of the ORFs. Functions are color coded, a legend for the colors can be found below the figures.*

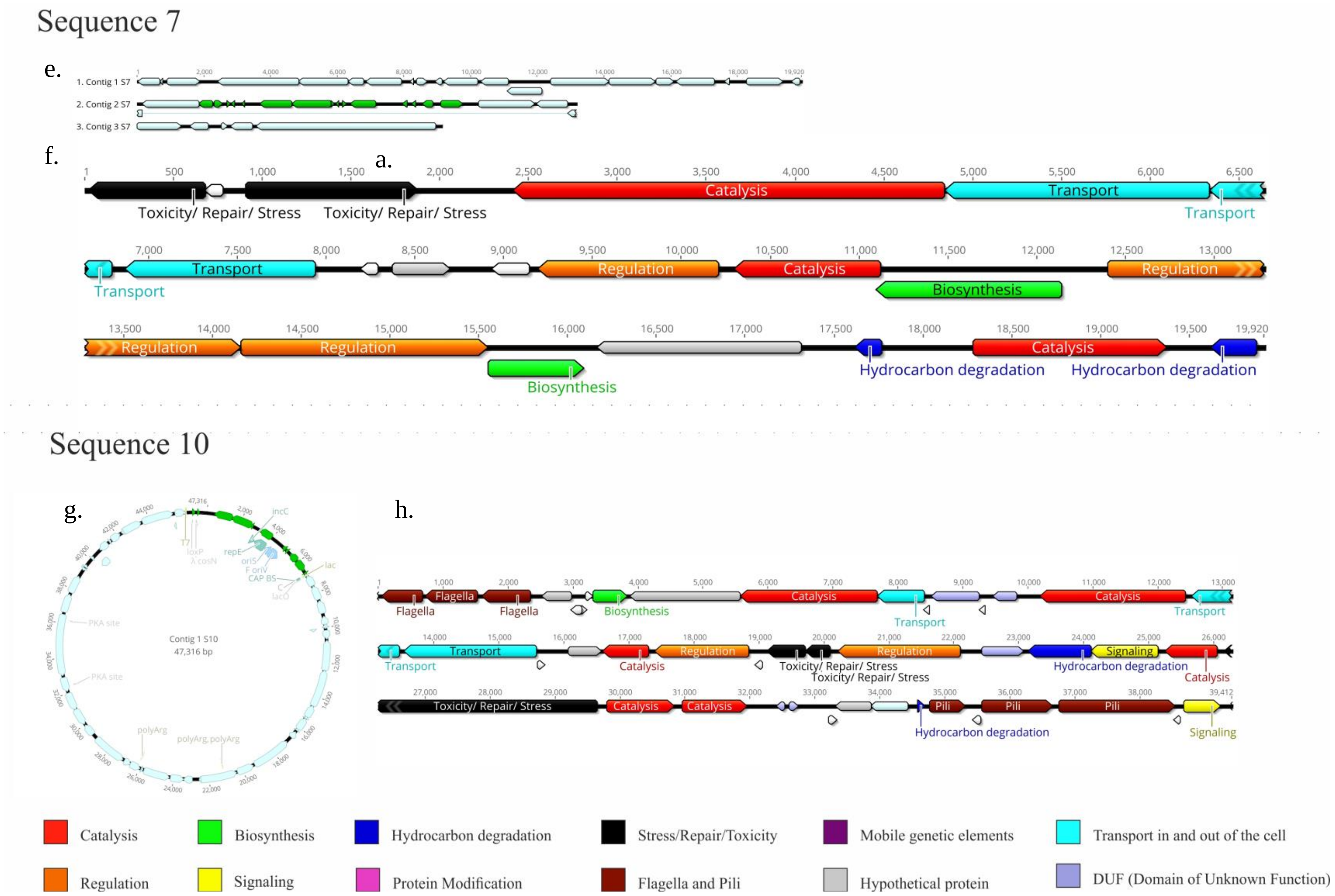


*Figure 21. Graphical representation of Sequences7 and 10. Figure e represents the three contigs of sequence 7, figure g is a representation of the re-circularized contig of sequence 10 with the pCC2FOS vector in green and ORFs in white. Figures f and h represent the linearized sequences with the predicted functions of the ORFs. Functions are color coded, a legend for the colors can be found below*

### 4.4.2 Sequence 2: genus *Maribacter*

Sequence 2 was exclusively matched with species from the genus *Maribacter,* a marine genus of the family *Flavobacteriaceae* isolated for the first time in 2004 [21]. Literature about this genus is scarce. *Maribacter sp.* should not be confused with *Marinobacter sp.* a well-studied hydrocarbon degrader from the family *Alteromonadaceae* [92].

Interestingly, *Maribacter sp.* has been found in cold [93] and template climates [94], can grow between 4 and 32 ºC, produces many types of fatty acids [21], and forms epiphytic biofilms on macroalgae and marine sponges [95, 96]. Additionally, one study found *Maribacter sp.* was able to grow with a concentration of 2% diesel oil and produced biosurfactants. These biosurfactants could reduce the surface tension of hydrocarbons, making them more miscible in water and increasing their bioavailability. Unfortunately, the biosurfactants composition was not analyzed [97].

In Sequence 2, two ORFs are involved in extracellular transport, ORF- 022 and ORF-034. The ORF-022 presents homology with a long-subunit fatty acid transport protein which carries fatty acids to the inside of the cell and it is positioned between a DUF protein (ORF-021) and a ribosomal protein (ORF-023) [58].

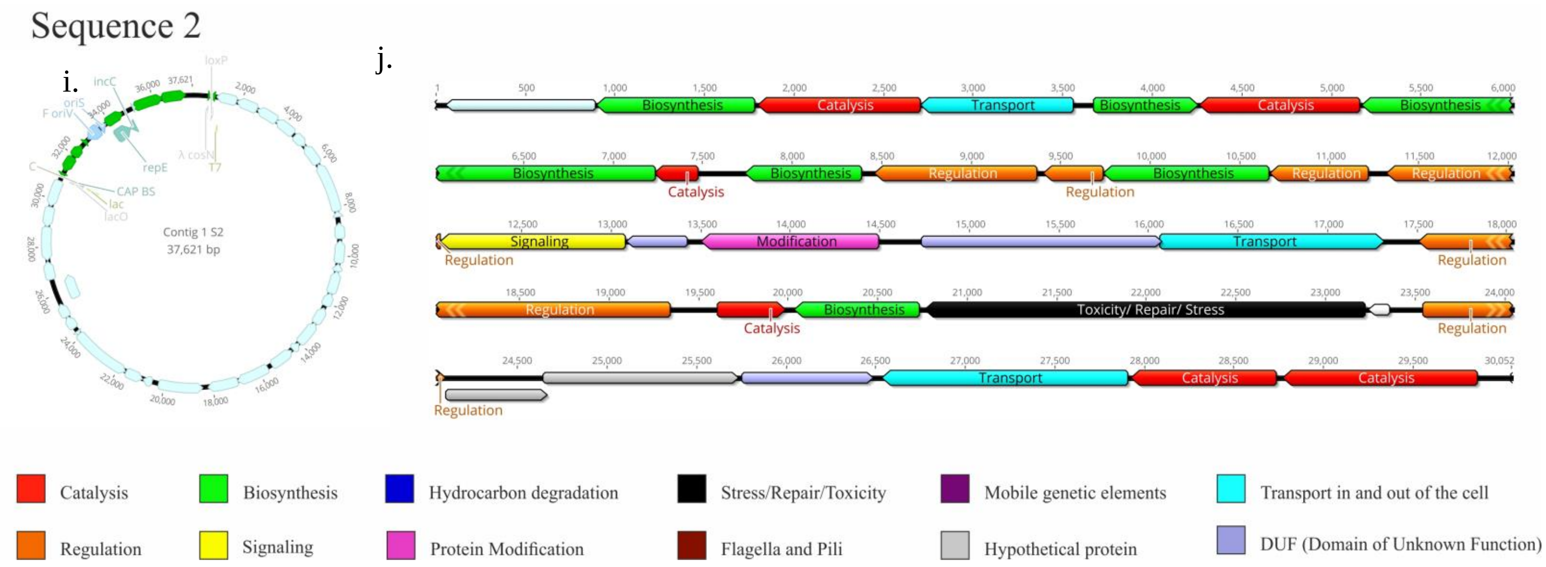


*Figure 22. Graphical representation of Sequence 2. Figure i is a representation of the re-circularized contig with the pCC2FOS vector in green and ORFs in white. Figure j  represents the linearized sequences with the predicted functions of the ORFs. Functions are color coded, a legend for the colors can be found below the figures.*

### 4.4.3 Sequences 4 and 6: *Synechococcales cyanobacterium*

The order Synechococcales belongs to the phylum Cyanobacteria. This order is widely distributed, able to form biofilms, and it is one of the most abundant autotrophs in the marine environment [98-101]. Furthermore, Synechococcales have been found to perform an extracellular degradation of phenols [102] and biodegradation of PAHs [103].

In Sequence 4, the ORF-061 matched with a protein from the AromaDeg database [65]. A VOC family protein, as explained for ORF-100 in Sequence 1, a protein superfamily which includes dioxygenases involved in PAHs degradation [81-83]. In ORF-015 another oxygenase was predicted, in this case an aromatic ring-hydroxylating dioxygenase subunit alpha which can incorporate two atoms of oxygen ($O_2$) into their substrates, starting hydrocarbon degradation [63].

Sequence 6 is one of the most interesting sequences analyzed, since only 8 out of 57 ORFs could have a protein function assigned. The remaining ORFs belonged to hypothetical proteins, DUFs or no matches could be found in any database. Another interesting fact about Sequence 6 is that of the few functions predicted, two were related to hydrocarbon degradation. The ORF-081 showed similarity with a 4a-hydroxytetrahydrobiopterin dehydratase which recycles a co-substrate for several enzymes, including aromatic hydrocarbon hydroxylases. Next to it, in ORF-083, a acetyltransferase was predicted which is involved in the next step of hydrocarbon degradation [63]. A graphical representation of this metabolic pathway and the ORFs involved can be seen in figure 27.

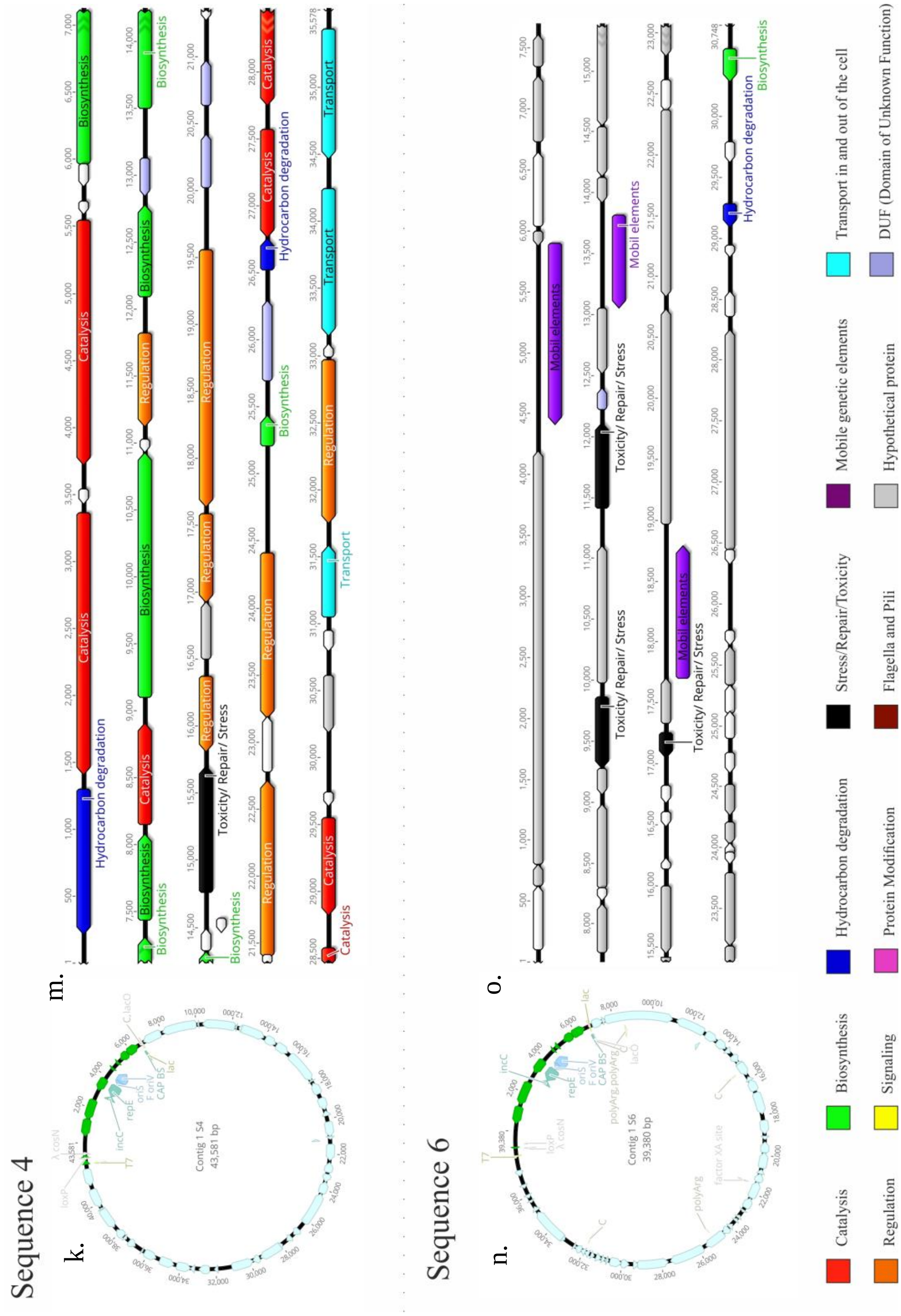


*Figure 23. Graphical representation of Sequences 4 and 6. Figures k and n are a representation of the re-circularized contigs with the pCC2FOS vector in green and ORFs in white. Figures m and o represent the linearized sequences with the predicted functions of the ORFs. Functions are color coded, a legend for the colors can be found below the figures.*

### 4.4.4 Sequence 5: family *Enterobacteraceae*

Sequence 5 was originated from a bacterium of the family *Enterobacteraceae,* most probably of the genus *Klebsiella. Klebsiella* is a genus which possesses many hydrocarbon degrading species and it is often studied for hydrocarbon bioremediation [104-106]. It is considered an important hydrocarbon biosurfactant producer [73, 107, 108]. Additionally, Among the species with high homology to Sequence 5 are *Raoultella ornithinolytica* and *Serratia marcescens.* These bacteria can degrade different types of hydrocarbons, including PAHs, the most persistent in aquatic environments [109, 110]. Notably, the genus *Serratia* is psychrophilic and ubiquitous, appearing in various environment where hydrocarbons are present [111]. Similarly to *Klebsiella sp., S. marcescens* produces biosurfactants in the presence of hydrocarbons to facilitate their bioavailability. Primarily, this species produces lipopeptides and glycolipids [74]. In Sequence 5, six transport related proteins were predicted, among them, a NIPSNAP family protein (ORF-079) which has a role in vesicular transport. An ABC transporter (ORF-084) position between two hypothetical proteins and two proteins related to the PTS sugar transporter (ORF-092 and 94) which could be related to biosurfactant transport across the membrane [63]. The only protein predicted in Sequence 5 involved in hydrocarbon degradation was an aromatic 1,2-dioxygenase alpha subunit (ORF-039) predicted with the custom BLAST with an 80% of pairwise identity. A protein involved in the activation of hydrocarbons. This ORF also presents a 100% pairwise identity with a protein of unknown function from *Klebsiella sp.*

An interesting feature studied in *R. ornithinolytica* is the capacity to change their membrane structure and metabolic pathways when exposed to hydrocarbons. These changes vary on whether the exposure is short- or long-term, days or months, respectively. The presence of benzene, an aromatic hydrocarbon, results in changes in cell shape and stability, uptake and transport proteins, lipid and amino acid biosynthesis and many metabolic pathways. It especially affected the protein synthesis process: tRNA aminoacylation and ribosomal proteins [112]. Interestingly, many of these types of proteins appeared in Sequence 5, such as ORF- 033 an outer membrane protein assembly factor *BamE,* which is involved in the insertion of β-barrels in the outer membrane [61-63].

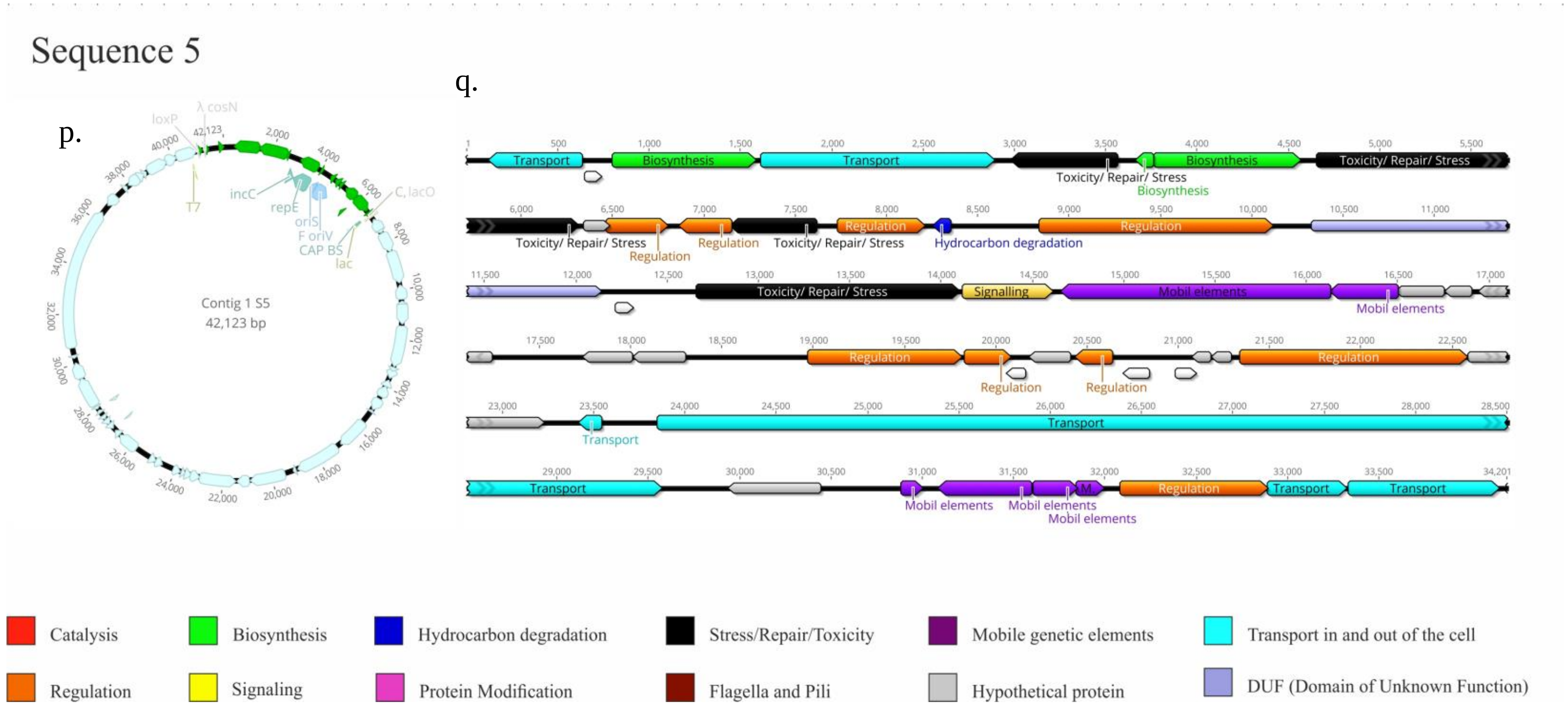


*Figure 24. Graphical representation of Sequence 5. Figures p is a representation of the re-circularized contigs with the pCC2FOS vector in green and ORFs in white. Figure q represents the linearized sequences with the predicted functions of the ORFs. Functions are color coded, a legend for the colors can be found below the figures.*

### 4.4.5 Sequence 8: *Azotobacter chroococcum*

For the ORFs in Sequence 8, the only species that presented high homology was *Azotobacter chroococcum*. This species from the family *Pseudomonadaceae* is a marine bacterium able to fix nitrogen, produce biosurfactants and use hydrocarbons as sole source of energy. For these reasons, *A. chroococcum* is a very valuable species for hydrocarbon degradation. Besides being capable of degrading hydrocarbon, *A. chroococcum* can fix nitrogen which increases the population of hydrocarbon-oxidizing bacteria. Additionally, the production of biosurfactants facilitate hydrocarbon bioavailability for hydrocarbon degraders [72, 113-115].

Surprisingly, the ORF-066 presented matches with two proteins. This ORF has a size of 2973 bp and when searched with BLASTx the best match is a TonB-dependent receptor, which can transport carbohydrates and chelates inside the cell. However, the pairwise identity was of merely 43.5%. Another match was found with the BLAST for the database AromaDeg. The protein found was the aromatic-ring-hydroxylating dioxygenase subunit alpha, a protein involved in PAHs degradation. Despite the pairwise identity being of 83.3%, the dioxygenase has a length of only 991 bp. Possible explanations for this might be the fusion of two proteins or a mismatch from the programs. Unfortunately, the custom BLAST of the database AromaDeg does not provide the information of which part of the ORF was matched to the hydrocarbon degrading protein, further studies on this ORF will be necessary to determine the exact conformation of the proteins.

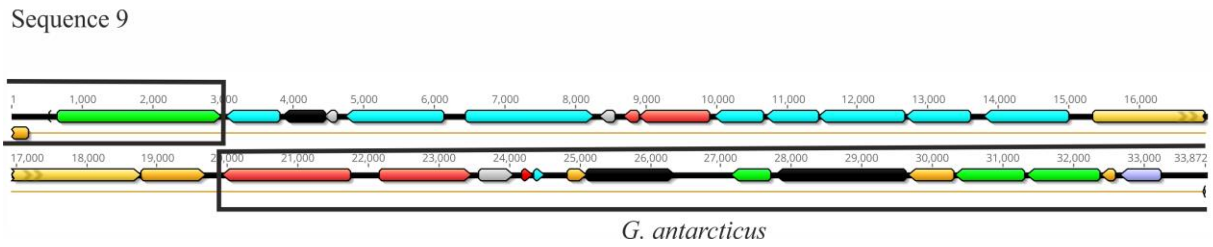
### 4.4.6 Sequence 9: order *Chromatiales*

Sequence 9 presents genes from the order Chromatiales. One of the most represented species is *Granulosicoccus antarcticus* with nine ORFs with approximately a 75% of pairwise identity (Figure 25) (*G. antarcticus* does not appear in figure 17 due to not reaching 85% of pairwise identity). *G. antarcticus* has been found in several marine environments and it is abundant in Antarctic regions. It grows in temperatures ranging between 3 to 25 ºC. Additionally, *G. antarcticus* is one of the main constituents in bacterial epiphytic communities associated with *F. vesiculosus* [116, 117]. Notably, the ORFs linked to *G. antarcticus* in Sequence 9 could indicate a role in hydrocarbon degradation. For instance, ORF-045 did not present any matches in MEGABLAST, BLASTn or BLASTx. However, in the custom BLAST against the database AromaDeg [65] a 71.4% of pairwise identity was found with the protein catechol 2,3-dioxygenase, involved in the cleaving of the aromatic ring in PAHs[118]. The ORF-017 was matched to the N,N-dimethylformamidase large subunit

(68.3% pairwise identity and 48.59% Query coverage) a protein involved in the degradation of a potent organic solvent N-N-dimethylformamide, an important pollutant in aquatic environments [119]. The ORF-049 was linked to the CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase which participates in glycerophospholipid metabolism, a compound that can act as a biosurfactant for hydrocarbons [63].

The ORF-052 and ORF-053 participate in fatty acid and phospholipid metabolism, respectively.  Fatty acid biosynthesis is the last step in hydrocarbon biodegradation and phospholipids are known biosurfactants [19, 85, 88-90]. Furthermore, as seen in figure 25, ORFs with homology to *G. antarcticus* are positioned consecutively. This could indicate that this set of genes represent one or more operons from this bacterium.

*Figure 25. graphical representation of ORFs in Sequence 9 and a selection of ORFs which present high homology with* Granulosicoccus antarcticus.



Besides *G. antarcticus*, several transport related genes are present, including five subunits of the urea ABC transporter (ORFs- 029, 30, 31, 32 and 35) that could be involved in biosurfactant transportation. An interesting protein is predicted in ORF-038, the sensory/regulatory protein *RpfC*, which in the study by P. S. Torres, et al. [120] was strongly linked to the regulation of the formation of a structure biofilm by *Xanthomonas campestris* due to its role in cell-cell signaling. This protein could present the same role in Sequence 9, regulating biofilm formation or facilitating cell-cell communication.

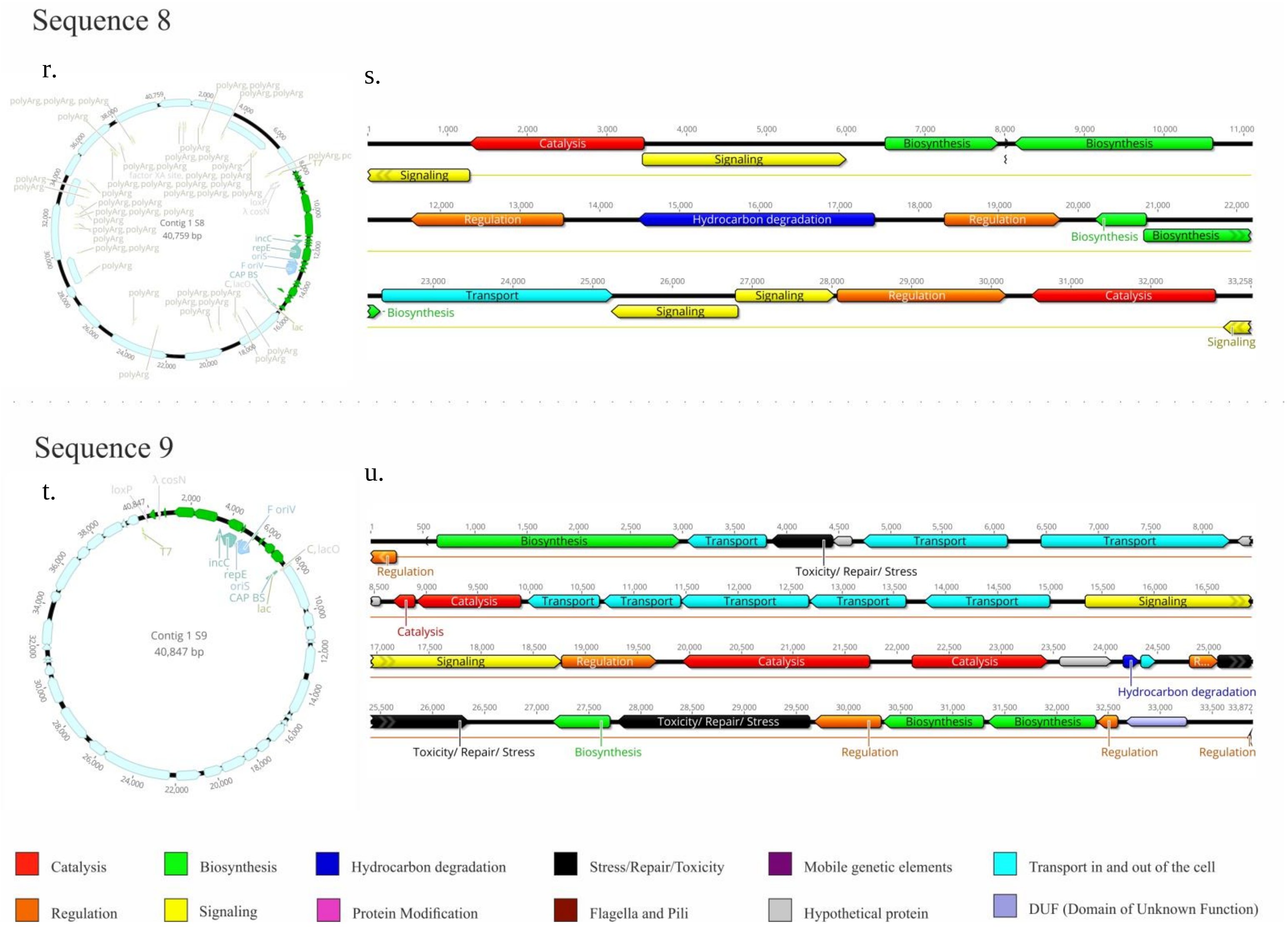


*Figure 26. Graphical representation of Sequences 8 and 9. Figures r and s are a representation of the re-circularized contigs with the pCC2FOS vector in green and ORFs in white. Figures s and u represent the linearized sequences with the predicted functions of the ORFs. Functions are color coded, a legend for the colors can be found below the figures.*

The findings of this study suggest that some marine bacteria not previously linked to hydrocarbon degradation could play a major role in hydrocarbon bioremediation pathways. These bacteria include members of the genus *Rubripirellula, Roseimaritima* and *Rhodopirellula* from the phylum *Planctomycetes.* As well as the genus *Maribacter* and the species *G. antarcticus*. These bacteria are part of native communities present in cold marine environments and symbionts of native macroalgae, these qualities make them very valuable for an environmentally respectful bioremediation.

The results of this investigation show that only 3 % of the predicted functions were related to hydrocarbon degradation. As stated by J. D. Rocca, et al.[121], the abundance of protein-encoding genes does not correlate to the amount of activity of the process they catalyze. However, in further studies a sub-cloning of the sequences, focusing on genes involved in oxidation or activation of hydrocarbons, could help narrowing down the most useful and interesting genes. The reason behind this is that the activation of hydrocarbons is the most selective step in hydrocarbon biodegradation. The enzymes necessary to process this step are not common in bacterial metabolism and oxidation/activation makes hydrocarbons bioavailable for a greater spectrum of bacteria. A further study with more focus on these enzymes is therefore suggested.

Since the aim of the study was not to analyze the entire fosmid library, but to show proof of concept of a method to study an environmental metagenome, the focus was on 10 clone sequences. This proved sufficient to select hydrocarbon degrading genes and to find promising enzymes and bacterial species that could be the focus of future work. The major limitation of the study, taking in consideration the origin of the microbial community,  is the incubation temperature for the transformed cells. The use of lower temperatures would have allowed a better selection of cold-adapted enzymes and bacteria. However, most of the genes analyzed in the study belong to either psychrophilic or very ubiquitous bacteria. These findings suggest that while utilizing the mesophilic metabolic  machinery of the *E. coli* cells to operate at 37ºC, the proteins used to survive in a hydrocarbon-rich medium might work at different temperature ranges. Nevertheless, a study on the same metagenome with different temperatures of incubation, such as 21ºC, would be complementing. In addition, the study could also include subcloning of the interesting sequences and the use of a psychrophilic host for the metagenomic library instead of *E. coli* cells.

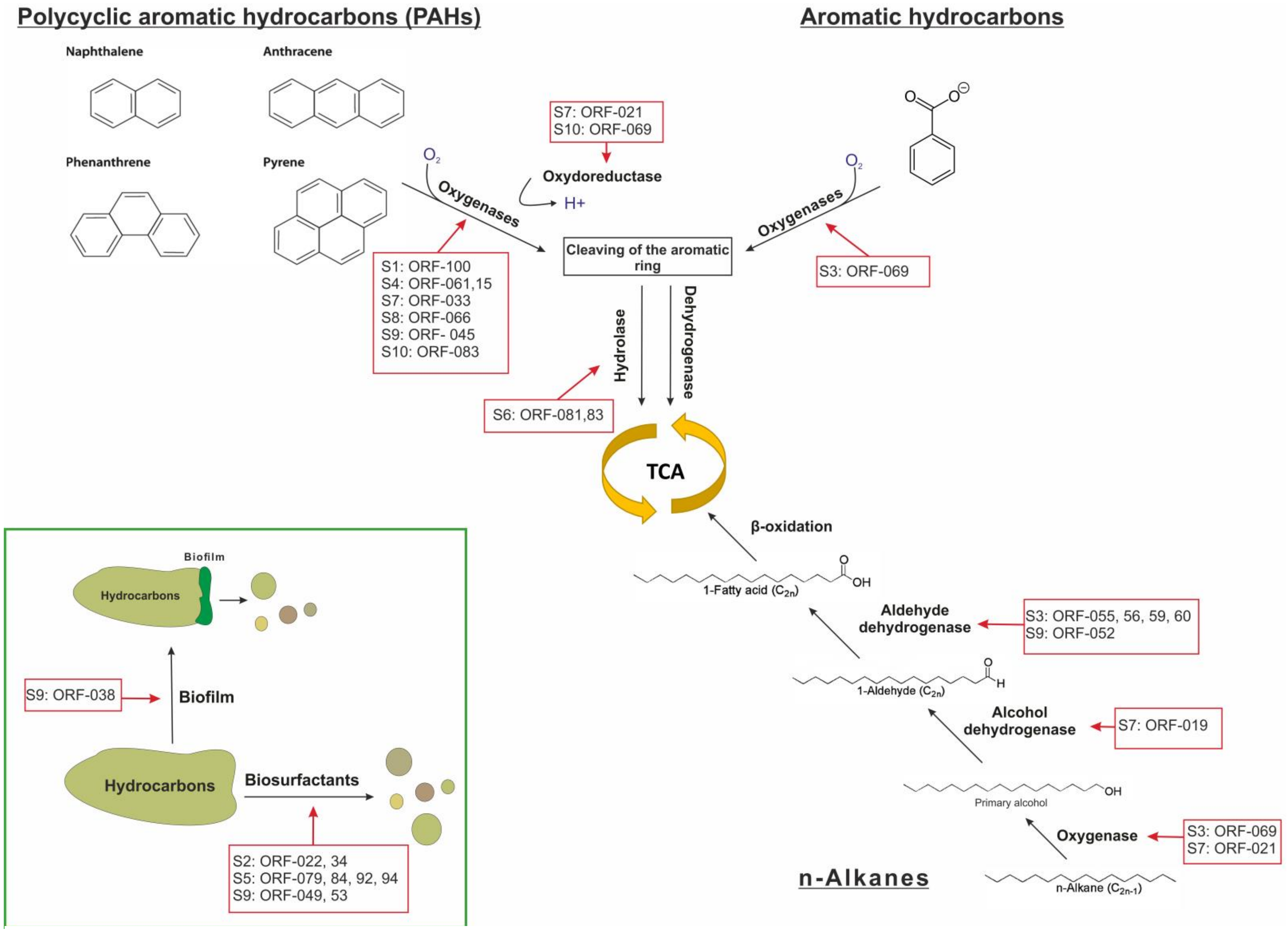


*Figure 27. Graphical representation of some aerobic hydrocarbon degradation pathways: Polycyclic aromatic hydrocarbons, aromatic hydrocarbons, n-alkenes, formation of biofilm, and production of biosurfactants. The ORFs from sequences analyzed in this study are position in the steps in which they are involved.*

## 4.6 Bioremediation in Norway and future studies

Bioremediation is the application of a biological treatment to remove hazardous chemicals from the environment [33]. It includes the use of microorganisms to remove hydrocarbons from a polluted environment. In Norway, bioremediation is regulated by the Act on Protection against pollution and waste (*lov om vern mot forurensing og avfall*) [122] and the Regulation on the limitation of pollution (*forskrift om begrensing av forurensing*)[123]. Bioremediating microorganisms in marine environments are under the regulation umbrella of “Beach purification substances” (*strandrensemidler*). According to these regulations, GMO (Genetically Modified Organisms), pathogenic organisms to the environment and bacteria with antibiotic resistance genes cannot be use for bioremediation [124]. Consequently, the study of native bacteria is needed to improve bioremediation processes and it should be an important issue for future research.

The Norwegian Environment Agency (*Miljødirektoratet*) as well as The Norwegian Scientific Committee for Food Safety (*Vitenskapskomiteen for mattrygghet, VKM*) believe bioremediation is a cost-effective and environmentally friendly solution to clean polluted environments. A study was performed by these agencies to update the regulations on microorganisms used in bioremediation. These agencies recommend the use of native bacteria along with molecular, genetic and metagenomic analysis [33, 125]. Hence, further research should be undertaken in fields such as metagenomic studies of native bacterial communities, biostimulation and bioaugmentation.

# 5 Conclusions

In this study, we were able to isolate, clone and express the DNA of a bacterial community from a chronically oil-polluted environment. The clones were selected in a hydrocarbon selective medium, sequenced, and analyzed through bioinformatic methods. Bacterial species native from a cold marine environment, not previously linked to hydrocarbon degradation, were identified as possible hydrocarbon degraders. Moreover, several proteins with unknown function were identify as possibly being involved in hydrocarbon degradation, setting the ground for further studies.

# Works cited

1. *Norwegian Polar institute* - *[https://www.npolar.no/en/themes/barents-sea/](https://www.npolar.no/en/themes/barents-sea/)*. 2020.
2. Hassanshahian, M. and S. Cappello, *Crude Oil Biodegradation in the Marine Environments, Biodegradation - Engineering and Technology*, ed. R. Chamy and F. Rosenkranz. 2013, IntechOpen.
3. *Geneious 2020.2.2 [https://www.geneious.com](https://www.geneious.com)*. 2020.
4. Helén, J., et al., *Biodegradation of Spilled Fuel Oil in Norwegian Marine Environments.* SINTEF Ocean AS, 2018.
5. Epicentre, *CopyControl HTP Fosmid Library Production Kit with pCC2FOS Vector and Phage T-1 Resistant EPI300-T1R E. coli Plating Strain* 2018.
6. Matishov, G.G., et al., *Global International Waters Assessment :Barents Sea*. 2004, Published by the Univ. of Kalmar on behalf of UNEP: Kalmar, Sweden :.
7. Smedsrud, L.H., et al., *THE ROLE OF THE BARENTS SEA IN THE ARCTIC CLIMATE SYSTEM.* Reviews of Geophysics, 2013. **51**(3): p. 415-449.
8. Halland, E.K., J. Mujezinović, and F. Riis, *CO2 storage atlas at Barents Sea.* Norwegian Petroleum Directorate.
9. *Norwegian Petroleum* - *[https://www.norskpetroleum.no/en/developments-and-operations/activity-per-sea-area/#barents-sea](https://www.norskpetroleum.no/en/developments-and-operations/activity-per-sea-area/#barents-sea)*. 2020.
10. Arild, M., *Russian and Norwegian petroleum strategies in the Barents Sea.* Arctic Review, 2010. **1**(2).
11. *Government.no* - *[https://www.regjeringen.no/en/topics/energy/oil-and-gas/id1003/](https://www.regjeringen.no/en/topics/energy/oil-and-gas/id1003/)*. 2020.
12. Austvik, O.G. and A. Moe, *Oil and Gas Extraction in the Barents Region*. 2016. p. 115-121.
13. Bargiela, R., et al., *Distribution of Hydrocarbon Degradation Pathways in the Sea*, in *Consequences of Microbial Interactions with Hydrocarbons, Oils, and Lipids: Production of Fuels and Chemicals*, S.Y. Lee, Editor. 2017, Springer International Publishing: Cham. p. 629-651.
14. Xue, J., et al., *Marine Oil-Degrading Microorganisms and Biodegradation Process of Petroleum Hydrocarbon in Marine Environments: A Review.* Current Microbiology, 2015. **71**(2): p. 220-228.
15. *ITOPF—International Tanker Owners Pollution Federation [https://www.itopf.org/knowledge-resources/countries-territories-regions/countries/norway/](https://www.itopf.org/knowledge-resources/countries-territories-regions/countries/norway/)*.
16. Chilvers, B.L., K.J. Morgan, and B.J. White, *Sources and reporting of oil spills and impacts on wildlife 1970–2018.* Environmental Science and Pollution Research, 2020.
17. Prince, R.C. and R.M. Atlas, *Bioremediation of Marine Oil Spills*, in *Consequences of Microbial Interactions with Hydrocarbons, Oils, and Lipids: Biodegradation and Bioremediation*, R. Steffan, Editor. 2018, Springer International Publishing: Cham. p. 1-25.
18. Brakstad, O.G., et al., *Biodegradation of Petroleum Oil in Cold Marine Environments*, in *Psychrophiles: From Biodiversity to Biotechnology*, R. Margesin, Editor. 2017, Springer International Publishing: Cham. p. 613-644.
19. Pérez-Pantoja, D., B. González, and D.H. Pieper, *Aerobic Degradation of Aromatic Hydrocarbons*, in *Aerobic Utilization of Hydrocarbons, Oils and Lipids*, F. Rojo, Editor. 2016, Springer International Publishing: Cham. p. 1-44.
20. McFarlin K, L.M., Perkins R, *Biodegradation of oil and dispersed oil by Arctic marine microorganisms.* . American Petroleum Institute, pp 300317. doi:10.7901/2169-3358-2014-1-300317.1, 2014.

21. Nedashkovskaya, O.I., et al., *Maribacter gen. nov., a new member of the family Flavobacteriaceae, isolated from marine habitats, containing the species Maribacter sedimenticola sp. nov., Maribacter aquivivus sp. nov., Maribacter orientalis sp. nov. and Maribacter ulvicola sp. nov.* International Journal of Systematic and Evolutionary Microbiology, 2004. **54**(4): p. 1017-1023.
22. Xu, X., et al., *Petroleum Hydrocarbon-Degrading Bacteria for the Remediation of Oil Pollution Under Aerobic Conditions: A Perspective Analysis.* Front. Microbiol. 9:2885. doi: 10.3389/fmicb.2018.02885, 2018.
23. Garneau M-È, M.C., Meisterhans G, Fortin N, King TL, Greer CW, Lee K *Hydrocarbon biodegradation by Arctic sea-ice and sub-ice microbial communities during microcosm experiments, Northwest Passage (Nunavut, Canada).* FEMS Microbiol Ecol 92(10). doi: 10.1093/femsec/fiw130, 2016.
24. Yakimov MM, G.G., Bruni V, Cappello S, D'Auria G, Golyshin PN, Giuliano L, *Crude oil-induced structural shift of coastal bacterial communities of rod bay (Terra Nova Bay, Ross Sea, Antarctica) and characterization of cultured cold-adapted hydrocarbonoclastic bacteria.* FEMS Microbiol Ecol 49(3):419–432, 2004.
25. Lofthus S, N.R., Lewin A, Brakstad OG *Successions of bacteria adhering to oil surfaces during biodegradation of crude oil in natural seawater at temperatures from 0 to 20°C. In: Abstracts of the 115th General Meeting of the American Society for Microbiology, New Orleans, Louisiana, May 30-June 2 2015.* 2015.
26. Brakstad OG, N.I., Faksness L-G, Brandvik PJ *Responses of microbial communities in Arctic sea ice after contamination by crude petroleum oil.* Microb Ecol 55(3):540–552, 2008.
27. Deppe U, R.H., Michaelis W, Antranikian G, *Degradation of crude oil by an arctic microbial consortium.* Extremophiles 9(6):461–470, 2005.
28. Gerdes B, B.R., Dieckmann G, Helmke E, *Influence of crude oil on changes of bacterial communities in Arctic sea-ice.* FEMS Microbiol Ecol 53(1):129–139, 2005.
29. Bagi A, P.D., Lanzén A, Bilstad T, Kommedal R, *Naphthalene biodegradation in temperate and arctic marine microcosms.* . Biodegradation 25(1):111–125, 2014.
30. Bowman JP, M.R., *Biodiversity, community structural shifts, and biogeography of prokaryotes within Antarctic continental shelf sediment.* . Appl Environ Microbiol 69(5):2463–2483, 2003.
31. Dong C, B.X., Sheng H, Jiao L, Zhou H, Shao Z, *Distribution of PAHs and the PAH-degrading bacteria in the deep-sea sediments of the high-latitude Arctic Ocean.* Biogeosciences 12(7):2163–2177. doi: 10.5194/bg-12-2163-2015, 2015.
32. Brakstad OG, B.K., *Biodegradation of petroleum hydrocarbons in seawater at low temperatures (0–5°C) and bacterial communities associated with degradation.* . Biodegradation 17(1):71–82, 2006.
33. Skaar, I., et al., *Health and environmental risk evaluation of microorganisms used in bioremediation. Scientific Opinion of the Panel on Microbial Ecology of the Norwegian Scientific Committee for Food Safet.* VKM, ISBN: 978-82-8259-232-1, Oslo, Norway, 2016.
34. Mirete, S., V. Morgante, and J.E. González-Pastor, *Functional metagenomics of extreme environments.* Current Opinion in Biotechnology, 2016. **38**: p. 143-149.
35. Amann, R.I., W. Ludwig, and K.-H. Schleifer, *Phylogenetic identification and in situ detection of individual microbial cells without cultivation.* Microbiological reviews, 1995. **59**(1): p. 143-169.
36. Shizuya, H., et al., *Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector.* Proceedings of the

National Academy of Sciences of the United States of America, 1992. **89**(18): p. 8794-8797.

37. Adam, N. and M. Perner, *Activity-Based Screening of Metagenomic Libraries for Hydrogenase Enzymes*, in *Metagenomics: Methods and Protocols*, W.R. Streit and R. Daniel, Editors. 2017, Springer New York: New York, NY. p. 261-270.
38. Popovic, A., et al., *Activity screening of environmental metagenomic libraries reveals novel carboxylesterase families.* Scientific Reports, 2017. **7**(1): p. 44103.
39. Sousa, S.T.P.d., et al., *Exploring the genetic potential of a fosmid metagenomic library from an oil-impacted mangrove sediment for metabolism of aromatic compounds.* Ecotoxicology and Environmental Safety, 2020. **189**: p. 109974.
40. Jensen, A.L.a.J.B., *personal communication.* unpublished, 2018.
41. Huson, D.H., et al., *MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data.* PLOS Computational Biology, 2016. **12**(6): p. e1004957.
42. Pugovkin, D.V., A. Liaimer, and J.B. Jensen, *Epiphytic bacterial communities of the alga Fucus vesiculosus in oil-contaminated water areas of the Barents Sea.* Doklady Biological Sciences, 2016. **471**(1): p. 269-271.
43. William, S., H. Feil, and A. Copeland, *Bacterial genomic DNA isolation using CTAB.* Sigma, 2012. **50**(6876).
44. Elbing, K. and R. Brent, *Media Preparation and Bacteriological Tools.* Current Protocols in Molecular Biology, 2002. **59**(1): p. 1.1.1-1.1.7.
45. Elbing, K.L. and R. Brent, *Recipes and Tools for Culture of Escherichia coli.* Current protocols in molecular biology, 2019. **125**(1): p. e83-e83.
46. John Bacha, J.F., Andy Gibbs, Lew Gibbs, Greg Hemighaus, Kent Hoekman, Jerry Horn,, et al., *Diesel Fuels Technical Review*. Chevron.
47. Schobert, H., *Composition, classification, and properties of petroleum*, in *Chemistry of Fossil Fuels and Biofuels*, H. Schobert, Editor. 2013, Cambridge University Press: Cambridge. p. 174-191.
48. Lee, M., et al., *Enhanced biodegradation of diesel oil by a newly identified Rhodococcus baikonurensis EN3 in the presence of mycolic acid.* Journal of Applied Microbiology, 2006. **100**(2): p. 325-333.
49. Gontikaki, E., et al., *Hydrocarbon-degrading bacteria in deep-water subarctic sediments (Faroe-Shetland Channel).* Journal of Applied Microbiology, 2018. **125**(4): p. 1040-1053.
50. Luo, Q., et al., *Isolation and characterization of marine diesel oil-degrading Acinetobacter sp. strain Y2.* Annals of Microbiology, 2013. **63**(2): p. 633-640.
51. Sambrook, J. and D.W. Russell, *Preparation of plasmid DNA by alkaline lysis with SDS: minipreparation.* Cold Spring Harbor Protocols, 2006. **2006**(1): p. pdb. prot4084.
52. *Illumina Next-Generation Sequencing* [https://www.illumina.com/science/technology/next-generation-sequencing.html](https://www.illumina.com/science/technology/next-generation-sequencing.html). 2020.
53. Alawi, M., *Next Generation Sequencing – Plasmid-Resequencing Analysis.* 2020.
54. Gibbs, M. 2020; Available from: [https://help.geneious.com/](https://help.geneious.com/)
55. *Geneious Prime 2020.2 Manual* 2020.
56. McWilliam H, L.W., Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R *Analysis Tool Web Services from the EMBL-EBI. (2013) Nucleic acids research 2013 Jul;41(Web Server issue):W597-600 doi:10.1093/nar/gkt376* 2013.
57. Delcher, B., Powers and Salzberg, *GLIMMER3 (Bioinformatics 23-6:673-679) Identifying bacterial genes and endosymbiont DNA with Glimmer.* 2007.

58. *National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2020 Oct 14]. Available from: https://www.ncbi.nlm.nih.gov/.*

59. *BLAST [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 10 16]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi.*

60. *Blast Program Selection Guide - https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=Blast*. 2009.

61. H.M. Berman, J.W., Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne., *RCSB PDB - The Protein Data Bank Nucleic Acids Research, 28: 235-242.* 2000.

62. Huntley RP, S.T., Mutowo-Muellenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C., *The GOA database: Gene Ontology annotation updates for 2015. Nucleic Acids Research 2014 doi: 10.1093/nar/gku1113.* 2014.

63. *The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47: D506-515 (2019).*

64. Letunic, I. and P. Bork, *Bioinformatics 23(1):127-8 Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation*. 2006.

65. Duarte, M., et al., *AromaDeg, a novel database for phylogenomics of aerobic bacterial degradation of aromatics*, in *Database*. 2014.

66. Meyer, F., et al., *The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes.* BMC Bioinformatics, 2008. **9**(1): p. 386.

67. Medema, M.H., et al., *antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences.* Nucleic Acids Research, 2011. **39**(suppl_2): p. W339-W346.

68. Illumina, I., *Quality Scores for Next-Generation Sequencing. Assessing sequencing accuracy using Phred quality scoring.*, in *Technical Note: Sequencing*. 2011: illumina.com.

69. Szarlip, P., et al., *Comparison of the dynamics of natural biodegradation of petrol and diesel oil in soil*.* Desalination and Water Treatment, 2014. **52**(19-21): p. 3690-3697.

70. Cruz, J.M., et al., *Biodegradation and Phytotoxicity of Biodiesel, Diesel, and Petroleum in Soil.* Water, Air, & Soil Pollution, 2014. **225**(5): p. 1962.

71. Omarova, M., et al., *Biofilm Formation by Hydrocarbon-Degrading Marine Bacteria and Its Effects on Oil Dispersion.* ACS Sustainable Chemistry & Engineering, 2019. **7**(17): p. 14490-14499.

72. Helmy, Q., et al., *Biosurfactants Production from Azotobacter sp. and its Application in Biodegradation of Petroleum Hydrocarbon.* Journal of Applied and Industrial Biotechnology in Tropical Region, 2008.

73. Nwaguma, I.V., C.B. Chikere, and G.C. Okpokwasili, *Isolation, characterization, and application of biosurfactant by Klebsiella pneumoniae strain IVN51 isolated from hydrocarbon-polluted soil in Ogoniland, Nigeria.* Bioresources and Bioprocessing, 2016. **3**(1): p. 40.

74. Huang, Y., et al., *Isolation and characterization of biosurfactant-producing Serratia marcescens ZCF25 from oil sludge and application to bioremediation.* Environmental Science and Pollution Research, 2020. **27**(22): p. 27762-27772.

75. Bengtsson, M.M. and L. Øvreås, *Planctomycetes dominate biofilms on surfaces of the kelp Laminaria hyperborea.* BMC Microbiology, 2010. **10**(1): p. 261.

76. Lage, O.M. and J. Bondoso, *Planctomycetes and macroalgae, a striking association.* Frontiers in Microbiology, 2014. **5**(267).

77. Lage, O.M. and J. Bondoso, *Planctomycetes diversity associated with macroalgae.* FEMS Microbiology Ecology, 2011. **78**(2): p. 366-375.

78. Bondoso, J., et al., *Roseimaritima ulvae gen. nov., sp. nov. and Rubripirellula obstinata gen. nov., sp. nov. two novel planctomycetes isolated from the epiphytic community of macroalgae.* Systematic and Applied Microbiology, 2015. **38**(1): p. 8-15.

79. Kallscheuer, N., et al., *Three novel Rubripirellula species isolated from plastic particles submerged in the Baltic Sea and the estuary of the river Warnow in northern Germany.* Antonie van Leeuwenhoek, 2019.

80. Faria, M., et al., *Planctomycetes attached to algal surfaces: Insight into their genomes.* Genomics, 2018. **110**(5): p. 231-238.

81. Sharma, V., et al., *Expression, purification, characterization and in silico analysis of newly isolated hydrocarbon degrading bleomycin resistance dioxygenase.* Molecular Biology Reports, 2020. **47**(1): p. 533-544.

82. Duarte, M., et al., *Functional soil metagenomics: elucidation of polycyclic aromatic hydrocarbon degradation potential following 12 years of in situ bioremediation.* Environmental Microbiology, 2017. **19**(8): p. 2992-3011.

83. Vaillancourt, F.H., J.T. Bolin, and L.D. Eltis, *The Ins and Outs of Ring-Cleaving Dioxygenases.* Critical Reviews in Biochemistry and Molecular Biology, 2006. **41**(4): p. 241-267.

84. Kahla, O., et al., *Efficiency of benthic diatom-associated bacteria in the removal of benzo(a)pyrene and fluoranthene.* Science of The Total Environment, 2021. **751**: p. 141399.

85. Moreno, R. and F. Rojo, *Enzymes for Aerobic Degradation of Alkanes in Bacteria*, in *Aerobic Utilization of Hydrocarbons, Oils and Lipids*, F. Rojo, Editor. 2017, Springer International Publishing: Cham. p. 1-25.

86. Letunic, I. and P. Bork, *20 years of the SMART protein domain annotation resource.* Nucleic Acids Res, 2018. **46**(D1): p. D493-d496.

87. Glöckner , F.O., et al., *Complete genome sequence of the marine planctomycete <em>Pirellula</em> sp. strain 1.* Proceedings of the National Academy of Sciences, 2003. **100**(14): p. 8298-8303.

88. Abbasian, F., et al., *A Comprehensive Review of Aliphatic Hydrocarbon Biodegradation by Bacteria.* Applied Biochemistry and Biotechnology, 2015. **176**(3): p. 670-699.

89. Widdel, F. and F. Musat, *Diversity and Common Principles in Enzymatic Activation of Hydrocarbons: An Introduction*, in *Aerobic Utilization of Hydrocarbons, Oils and Lipids*, F. Rojo, Editor. 2016, Springer International Publishing: Cham. p. 1-30.

90. Hyman, M., *Aerobic Degradation of Gasoline Ether Oxygenates*, in *Aerobic Utilization of Hydrocarbons, Oils and Lipids*, F. Rojo, Editor. 2017, Springer International Publishing: Cham. p. 1-31.

91. Wecker, P., et al., *Transcriptional response of the model planctomycete Rhodopirellula baltica SH1T to changing environmental conditions.* BMC Genomics, 2009. **10**(1): p. 410.

92. Gauthier, M.J., et al., *Marinobacter hydrocarbonoclasticus gen. nov., sp. nov., a New, Extremely Halotolerant, Hydrocarbon-Degrading Marine Bacterium.* International Journal of Systematic and Evolutionary Microbiology, 1992. **42**(4): p. 568-576.

93. Cho, K.H., et al., *Maribacter arcticus sp. nov., isolated from Arctic marine sediment.* International Journal of Systematic and Evolutionary Microbiology, 2008. **58**(6): p. 1300-1303.

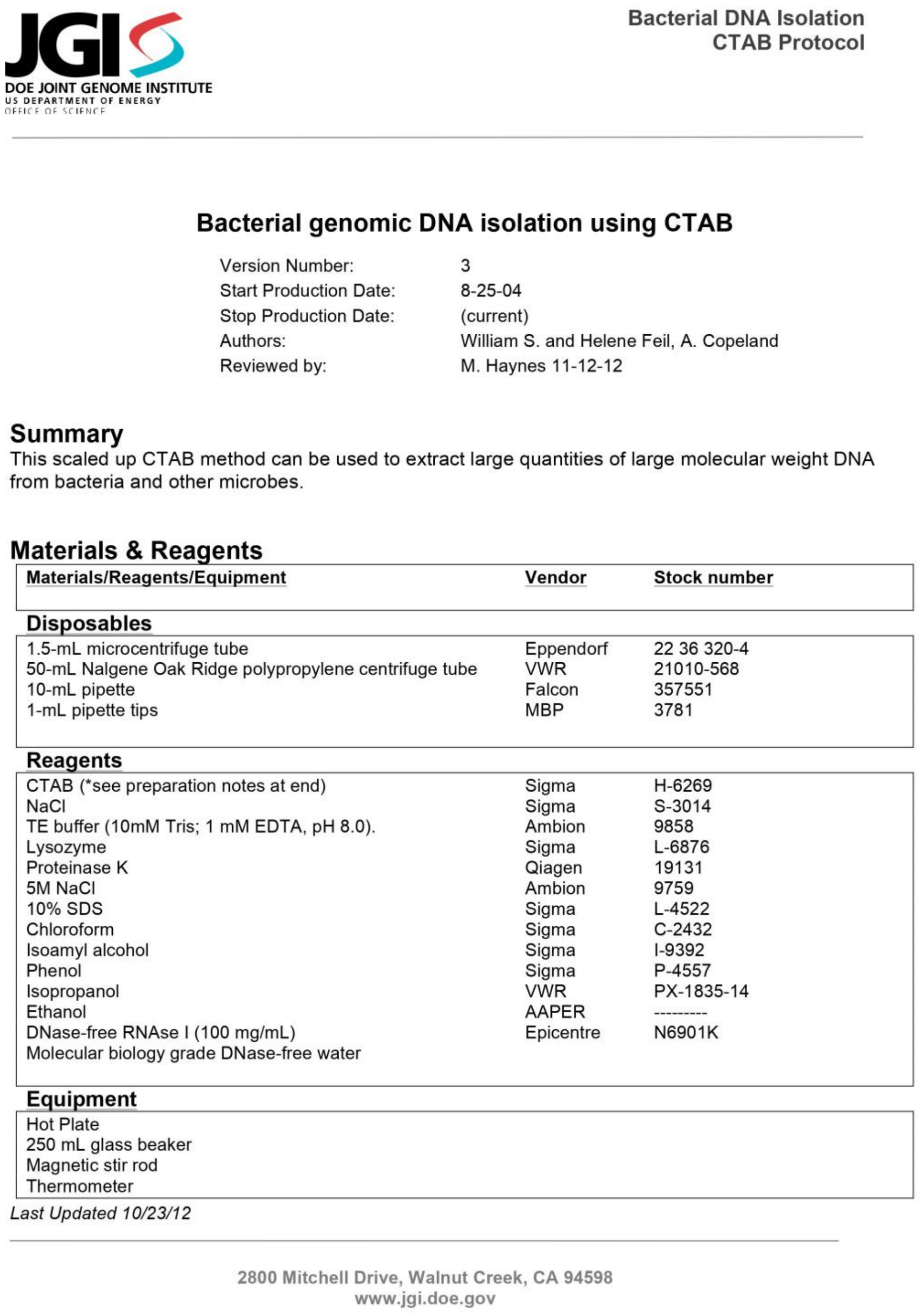
94. Gacesa, R., Baranasic, D., Starcevic, A., Diminic, J., Korlević, M., Najdek, M., ... Zucko, J., *Bioprospecting for Genes Encoding Hydrocarbon-Degrading Enzymes from Metagenomic Samples Isolated from Northern Adriatic Sea Sediments. Food Technology and Biotechnology, 56 (2), 270-277. https://doi.org/10.17113/ftb.56.02.18.5393*. 2018.
95. Kwon, K.K., Y.K. Lee, and H.K. Lee, *Costertonia aggregata gen. nov., sp. nov., a mesophilic marine bacterium of the family Flavobacteriaceae, isolated from a mature biofilm.* International Journal of Systematic and Evolutionary Microbiology, 2006. **56**(6): p. 1349-1353.
96. Jackson, S.A., et al., *Maribacter spongiicola sp. nov. and Maribacter vaceletii sp. nov., isolated from marine sponges, and emended description of the genus Maribacter.* Int J Syst Evol Microbiol, 2015. **65**(7): p. 2097-2103.
97. Rizzo, C., et al., *Influence of salinity and temperature on the activity of biosurfactants by polychaete-associated isolates.* Environmental Science and Pollution Research, 2014. **21**(4): p. 2988-3004.
98. Huang, T.-C., et al., *Organization and expression of nitrogen-fixation genes in the aerobic nitrogen-fixing unicellular cyanobacterium Synechococcus sp. strain RF-1.* Microbiology, 1999. **145**(3): p. 743-753.
99. Parnasa, R., et al., *Small secreted proteins enable biofilm development in the cyanobacterium Synechococcus elongatus.* Scientific Reports, 2016. **6**(1): p. 32209.
100. Becker, S., et al., *Genetic diversity and distribution of periphytic Synechococcus spp. in biofilms and picoplankton of Lake Constance.* FEMS Microbiology Ecology, 2004. **49**(2): p. 181-190.
101. Dvořák, P., et al., *Morphological and molecular studies of Neosynechococcus sphagnicola, gen. et sp. nov.(Cyanobacteria, Synechococcales).* Phytotaxa, 2014. **170**(1): p. 24-34.
102. Wurster, M., et al., *Extracellular degradation of phenol by the cyanobacterium Synechococcus PCC 7002.* Journal of Applied Phycology, 2003. **15**(2): p. 171-176.
103. BUNNAG, S. and H. KANSUNGNOEN. *Biodegradation of used motor oil by Chlorella vulgaris and Synechococcus elongates*. in *emeritus prof. Dr. Faizah binti mohd sharoum chairman umt international annual symposium on sustainability science and management (umtas 2016)*. 2016.
104. Nwangwu, A.M.a.C., *Studies on the bioutilization of some petroleum hydrocarbons by single and mixed cultures of some bacterial species.* Africal Journal of Microbiology Research, 2011.
105. Ali, M.F., M.J. M-Ridha, and A.H. Taly, *Optimization Kerosene Bio-degradation by a Local Soil Bacterium Isolate Klebsiella pneumoniae Sp. pneumonia.* Pure Appl Microbiol, 12(4), 2049-2057, 2018.
106. Li, X., et al., *Anaerobic biodegradation of pyrene by Klebsiella sp. LZ6 and its proposed metabolic pathway.* Environmental Technology, 2020. **41**(16): p. 2130-2139.
107. Lee, S.-C., et al., *Characterization of new biosurfactant produced by Klebsiella sp. Y6-1 isolated from waste soybean oil.* Bioresource Technology, 2008. **99**(7): p. 2288-2292.
108. Jain, R.M., et al., *Production and structural characterization of biosurfactant produced by an alkaliphilic bacterium, Klebsiella sp.: Evaluation of different carbon sources.* Colloids and Surfaces B: Biointerfaces, 2013. **108**: p. 199-204.
109. Alegbeleye, O.O., B.O. Opeolu, and V. Jackson, *Bioremediation of polycyclic aromatic hydrocarbon (PAH) compounds: (acenaphthene and fluorene) in water using indigenous bacterial species isolated from the Diep and Plankenburg rivers,*

*Western Cape, South Africa.* Brazilian Journal of Microbiology, 2017. **48**(2): p. 314-325.

110. Bidja Abena, M.T., et al., *Microbial diversity changes and enrichment of potential petroleum hydrocarbon degraders in crude oil-, diesel-, and gasoline-contaminated soil.* 3 Biotech, 2020. **10**(2): p. 42.
111. Mai, A.-G.M., *Serratia A Novel Source of Secondary Metabolites.* Adv Biotech & Micro, 2018. Volume 11 Issue 3 - September 2018 DOI: 10.19080/AIBM.2018.11.555814.
112. Zdarta, A., et al., *Hydrocarbon-induced changes in proteins and fatty acids profiles of Raoultella ornithinolytica M03.* Journal of Proteomics, 2017. **164**: p. 43-51.
113. Thavasi, R., et al., *Biodegradation of Crude Oil by Nitrogen Fixing Marine Bacteria Azotobacter chroococcum*.* Research Journal of Microbiology, 2006. **1**(5): p. 401-408.
114. Gradova, N., et al., *Use of bacteria of the genus Azotobacter for bioremediation of oil-contaminated soils.* Applied Biochemistry and Microbiology, 2003. **39**(3): p. 279-281.
115. Parnadi, P.S., E. Kardena, and E. Ratnaningsih, *Improving the effectiveness of crude-oil hydrocarbon biodegradation employing Azotobacter chroococcum as co-inoculant.* Microbiology Indonesia, 2007. **1**(1): p. 2-2.
116. Kang, I., Y. Lim, and J.-C. Cho, *Complete genome sequence of Granulosicoccus antarcticus type strain IMCC3135T, a marine gammaproteobacterium with a putative dimethylsulfoniopropionate demethylase gene.* Marine Genomics, 2018. **37**: p. 176-181.
117. Lachnit, T., et al., *Epibacterial community patterns on marine macroalgae are host-specific but temporally variable.* Environmental Microbiology, 2011. **13**(3): p. 655-665.
118. Brusa, T., et al., *Aromatic hydrocarbon degradation patterns and catechol 2,3-dioxygenase genes in microbial cultures from deep anoxic hypersaline lakes in the eastern Mediterranean sea.* Microbiological Research, 2001. **156**(1): p. 49-58.
119. Chetan Kumar Arya, et al., *A 2 ‐ Tyr ‐ 1 ‐ carboxylate Mononuclear Iron Center Forms the Active Site of a Paracoccus Dimethylformamidase.* Wiley ‐ VCH Verlag GmbH & Co. KGaA, Weinheim, 2020.
120. Torres, P.S., et al., *Controlled synthesis of the DSF cell-cell signal is required for biofilm formation and virulence in Xanthomonas campestris.* Environ Microbiol, 2007. **9**(8): p. 2101-9.
121. Rocca, J.D., et al., *Relationships between protein-encoding gene abundance and corresponding process are commonly assumed yet rarely observed.* The ISME Journal, 2014. 9, 1693–1699; doi:10.1038/ismej.2014.252.
122. miljødepartementet, K.-o., *Lov om vern mot forurensninger og om avfall (forurensningsloven).* 1983, ISBN 82-504-1304-0.
123. *Forskrift om begrensning av forurensning (forurensningsforskriften)*, K.-o. miljødepartementet, Editor. 2004: I 2004 hefte 9.
124. *Forskrift om begrensning av forurensning (forurensningsforskriften)*, in *Kapittel 19. Sammensetning og bruk av dispergeringsmidler og strandrensemidler for bekjempelse av oljeforurensing*, K.-o. miljødepartementet, Editor.
125. Øverland, K.R., *Helse- og miljørisikovurdering av mikroorganismer i bioremedieringstiltak.* 2016, Miljødirektoratet

# 6 Appendix

## Appendix 1: Bacterial genomic DNA isolation using CTAB



DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

**Bacterial DNA Isolation**
**CTAB Protocol**

### Bacterial genomic DNA isolation using CTAB

| | |
|---|---|
| Version Number: | 3 |
| Start Production Date: | 8-25-04 |
| Stop Production Date: | (current) |
| Authors: | William S. and Helene Feil, A. Copeland |
| Reviewed by: | M. Haynes 11-12-12 |

### Summary

This scaled up CTAB method can be used to extract large quantities of large molecular weight DNA from bacteria and other microbes.

### Materials & Reagents

| Materials/Reagents/Equipment | Vendor | Stock number |
|---|---|---|
| **Disposables** | | |
| 1.5-mL microcentrifuge tube | Eppendorf | 22 36 320-4 |
| 50-mL Nalgene Oak Ridge polypropylene centrifuge tube | VWR | 21010-568 |
| 10-mL pipette | Falcon | 357551 |
| 1-mL pipette tips | MBP | 3781 |
| **Reagents** | | |
| CTAB (*see preparation notes at end) | Sigma | H-6269 |
| NaCl | Sigma | S-3014 |
| TE buffer (10mM Tris; 1 mM EDTA, pH 8.0). | Ambion | 9858 |
| Lysozyme | Sigma | L-6876 |
| Proteinase K | Qiagen | 19131 |
| 5M NaCl | Ambion | 9759 |
| 10% SDS | Sigma | L-4522 |
| Chloroform | Sigma | C-2432 |
| Isoamyl alcohol | Sigma | I-9392 |
| Phenol | Sigma | P-4557 |
| Isopropanol | VWR | PX-1835-14 |
| Ethanol | AAPER | --------- |
| DNase-free RNAse I (100 mg/mL) | Epicentre | N6901K |
| Molecular biology grade DNase-free water | | |
| **Equipment** | | |
| Hot Plate | | |
| 250 mL glass beaker | | |
| Magnetic stir rod | | |
| Thermometer | | |

*Last Updated 10/23/12*

2800 Mitchell Drive, Walnut Creek, CA 94598
www.jgi.doe.gov

Automatic pipette dispenser
Sorval 500 Plus centrifuge (DuPont, Newtown, CT)
65°C water bath
37°C incubator/heat block
56°C heat block

## Procedure

Cell preparation and extraction techniques.
(Modification of “CTAB method”, in **Current Protocols in Molecular Biology**)

Cell growth:
To minimize gDNA sampling bias (e.g., excess coverage of sequences around the origin of replication) please take precautions NOT to proceed with DNA isolation while most of the cell population is in the stage of active DNA replication. We recommend collaborators to check the cell growth prior to DNA isolation. DNA should be prepared from cell culture that is either in late log phase or early stationary phase. If the cells are in the early log phase, the culture should be placed on ice or 4°C to slow down the growth and allow DNA replication to complete prior to cell lysis and DNA isolation.

If at all possible, please produce more DNA from a single isolation event than is strictly required for library creation and freeze aliquots of the extra DNA. Then, should more DNA be required for finishing it will be available. If extra cells are available instead, please consider storing extra aliquots in 15-40% glycerol at -80°C.

| | **1.5ml** | **30ml** | **60ml** |
|---|---|---|---|
| 1. Grow cells (see above) in broth and pellet at 10,000 rpm for 5 min or scrape from plate. | | | |
| 2. Transfer bacterial suspension to the appropriate centrifuge tube. | | | |
| 3. Spin down cells in microfuge or centrifuge at 10,000 rpm for 5 minute. | | | |
| 4. Discard the supernatant. | | | |
| 5. Resuspend cells in TE. | | | |
| 6. Adjust to $OD^{600}$ ≅ 1.0 with TE buffer (10mM Tris; 1 mM EDTA, pH 8.0) | | | |
| 7. Transfer given amount of cell suspension to a clean centrifuge tube. ------- | 740µl | 14.8ml | 29.6ml |
| 8. Add lysozyme (conc. 100mg/ml). Mix well. ------------------------------------ | 20µl | 400µl | 800µl |
| This step is necessary for hard to lyse gram (+) and some gram (–) bacteria. | | | |
| 9. Incubate for 30 min. at 37°C. | | | |
| 10. Add 10% SDS. Mix well. -------------------------------------------------------- | 40µl | 800µl | 1.6ml |
| 11. Add Proteinase K (10mg/ml). Mix well. --------------------------------------- | 8µl | 160µl | 320µl |
| 12. Incubate for 1-3 hr at 56°C. If cells are not lysed (as seen by cleared solution with increased viscosity) incubation can proceed overnight (16 hrs). | | | |
| 13. Add 5 M NaCl. Mix well. -------------------------------------------------------- | 100µl | 2ml | 4ml |

14. Add CTAB/NaCl (heated to 65°C). Mix well. -------------------------------- 100µl 2ml 4ml

15. Incubate at 65°C for 10 min.

16. Add chloroform:isoamyl alcohol (24:1). Mix well. ---------------------------- 0.5ml 10ml 20ml

17. Spin at max speed for 10 min at room temperature.

18. Transfer aqueous phase to clean microcentrifuge tube (should not be viscous).

19. Add phenol:chloroform:isoamyl alcohol (25:24:1). Mix well. --------------- 0.5ml 10ml 20ml

20. Spin at max speed for 10 min at room temperature.

21. Transfer aqueous phase to clean microcentrifuge tube.

22. Add chloroform:isoamyl alcohol (24:1). Mix well. ---------------------------- 0.5ml 10ml 20ml

23. Spin at max speed for 10 min at room temperature.

24. Transfer aqueous phase and add 0.6 vol isopropanol (-20°C).
(e.g. if 400 µl of aqueous phase is transferred, add 240 µl of isopropanol. ---- Add 0.6 vol. ----

25. Incubate at -20°C for 2 hrs to overnight.

26. Spin at max speed for 15 min at 4°C.

27. Wash pellet with cold 70% ethanol (directly from -20°C freezer), spin at max speed for 5 min.

28. Discard the supernatant and let pellet dry at room temp. This may take some time (20 min. to several hours, depending on humidity).

29. Resuspend in ~170 µl of DNAse-free water. Proceed to RNAse treatment.

1.1 Set up the following reaction in a 1.5ml microcentrifuge tube (multiple reactions can be done in different tubes):

Note: RNase I @ 10U/µl, one unit digests 100 ng of RNA per second

| | |
|---|---|
| DNA (in $H_20$) | 170µl |
| 10X RNase I buffer | 20µl |
| RNase I | 10µl |
| | 200ul |

1.2 Mix & Spin down.

1.3 Incubate tube at 37°C for 1 hr.

*Checkpoint: Check a small aliquot (5ul) on an agarose gel with a no treatment control. Run gel 10-15 min. If there is only a trace of RNA, proceed with next step, heat inactivation. If a large amount of RNA is still present, add another 10µl of RNase I and repeat the incubation.*

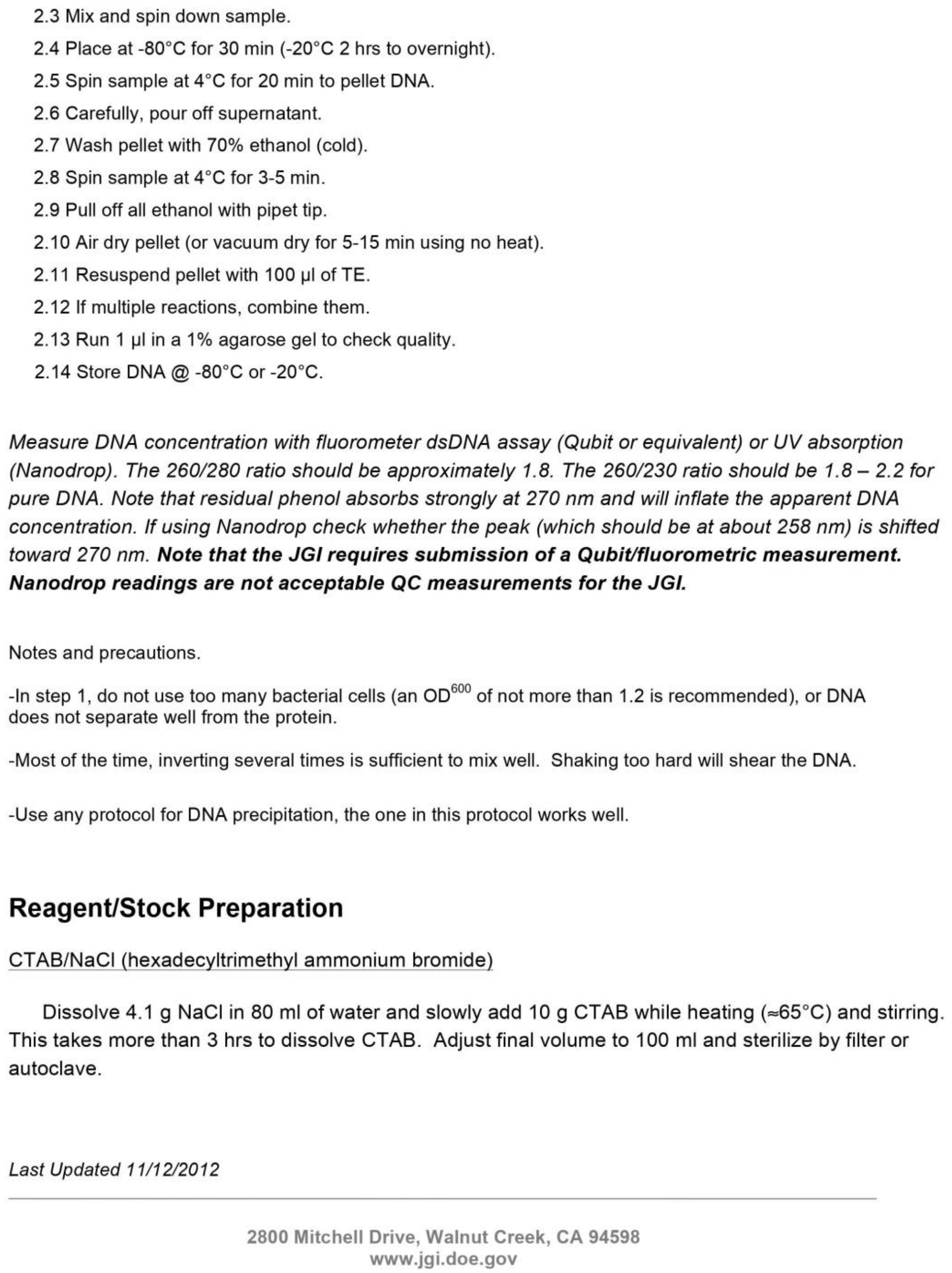
1.4 Heat inactivate enzyme at 70°C for 15 min.

1.5 Place tube on ice to cool.

2. Ethanol Precipitation

2.1 Add 1/10 volume of 3M Sodium Acetate to your sample.

2.2 Add 2.5 volumes of 100% ethanol.



2.3 Mix and spin down sample.

2.4 Place at -80°C for 30 min (-20°C 2 hrs to overnight).

2.5 Spin sample at 4°C for 20 min to pellet DNA.

2.6 Carefully, pour off supernatant.

2.7 Wash pellet with 70% ethanol (cold).

2.8 Spin sample at 4°C for 3-5 min.

2.9 Pull off all ethanol with pipet tip.

2.10 Air dry pellet (or vacuum dry for 5-15 min using no heat).

2.11 Resuspend pellet with 100 µl of TE.

2.12 If multiple reactions, combine them.

2.13 Run 1 µl in a 1% agarose gel to check quality.

2.14 Store DNA @ -80°C or -20°C.

*Measure DNA concentration with fluorometer dsDNA assay (Qubit or equivalent) or UV absorption (Nanodrop). The 260/280 ratio should be approximately 1.8. The 260/230 ratio should be 1.8 – 2.2 for pure DNA. Note that residual phenol absorbs strongly at 270 nm and will inflate the apparent DNA concentration. If using Nanodrop check whether the peak (which should be at about 258 nm) is shifted toward 270 nm.* ***Note that the JGI requires submission of a Qubit/fluorometric measurement. Nanodrop readings are not acceptable QC measurements for the JGI.***

Notes and precautions.

-In step 1, do not use too many bacterial cells (an $OD^{600}$ of not more than 1.2 is recommended), or DNA does not separate well from the protein.

-Most of the time, inverting several times is sufficient to mix well. Shaking too hard will shear the DNA.

-Use any protocol for DNA precipitation, the one in this protocol works well.

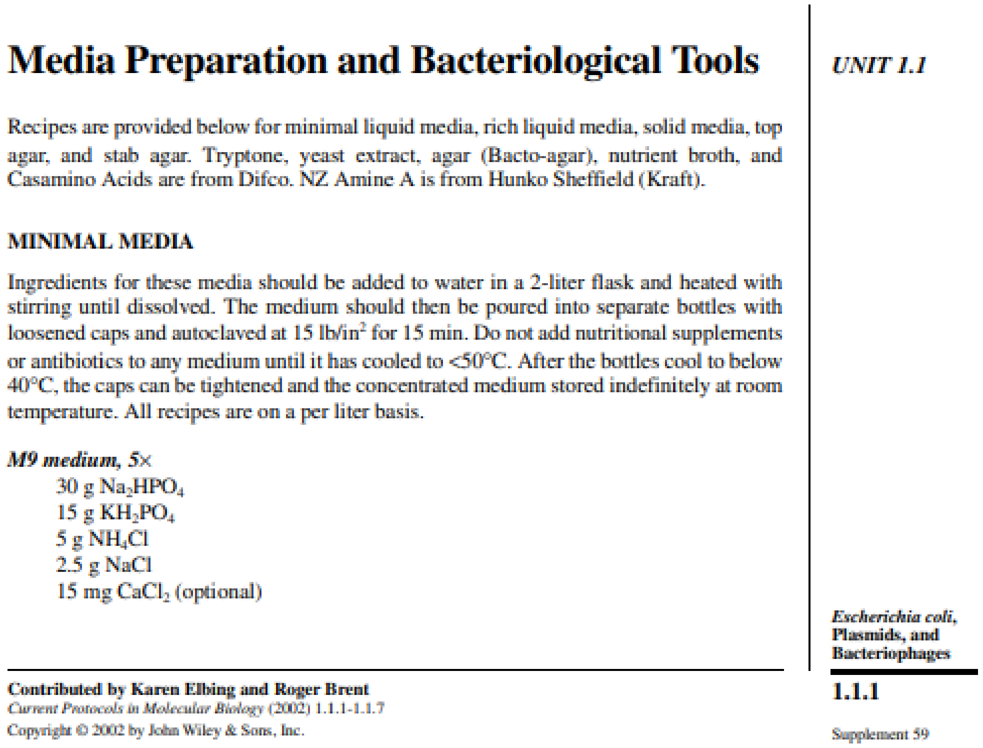
## Reagent/Stock Preparation

CTAB/NaCl (hexadecyltrimethyl ammonium bromide)

Dissolve 4.1 g NaCl in 80 ml of water and slowly add 10 g CTAB while heating (≈65°C) and stirring. This takes more than 3 hrs to dissolve CTAB. Adjust final volume to 100 ml and sterilize by filter or autoclave.

*Last Updated 11/12/2012*

# Appendix 2: Current protocols in molecular biology



## Media Preparation and Bacteriological Tools

UNIT 1.1

Recipes are provided below for minimal liquid media, rich liquid media, solid media, top agar, and stab agar. Tryptone, yeast extract, agar (Bacto-agar), nutrient broth, and Casamino Acids are from Difco. NZ Amine A is from Hunko Sheffield (Kraft).

### MINIMAL MEDIA

Ingredients for these media should be added to water in a 2-liter flask and heated with stirring until dissolved. The medium should then be poured into separate bottles with loosened caps and autoclaved at 15 lb/in$^2$ for 15 min. Do not add nutritional supplements or antibiotics to any medium until it has cooled to <50°C. After the bottles cool to below 40°C, the caps can be tightened and the concentrated medium stored indefinitely at room temperature. All recipes are on a per liter basis.

***M9 medium, 5×***

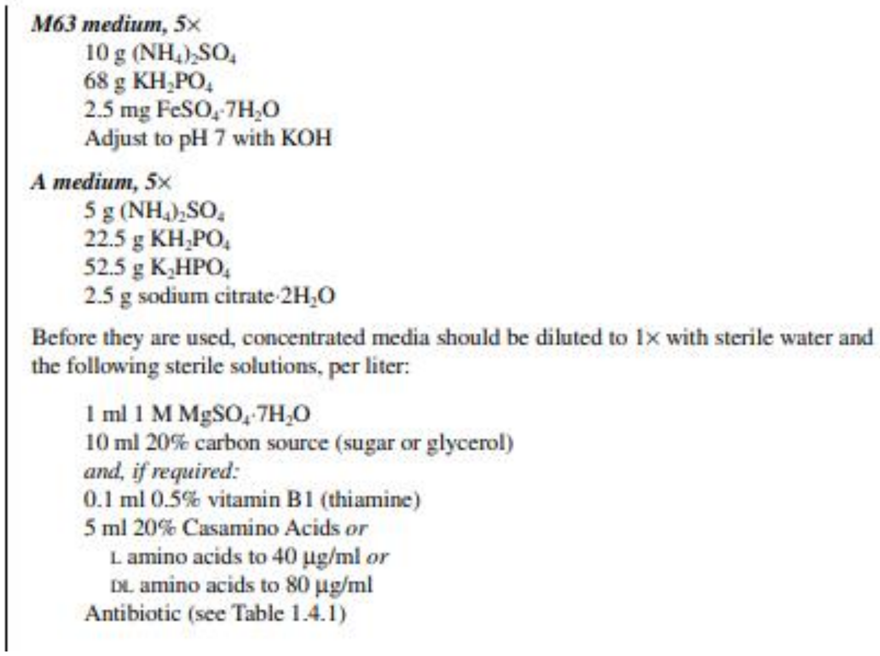
30 g $Na_2HPO_4$
15 g $KH_2PO_4$
5 g $NH_4Cl$
2.5 g NaCl
15 mg $CaCl_2$ (optional)

***Escherichia coli,*** **Plasmids, and Bacteriophages**

**Contributed by Karen Elbing and Roger Brent**
*Current Protocols in Molecular Biology* (2002) 1.1.1-1.1.7
Copyright © 2002 by John Wiley & Sons, Inc.

**1.1.1**

Supplement 59



***M63 medium, 5×***

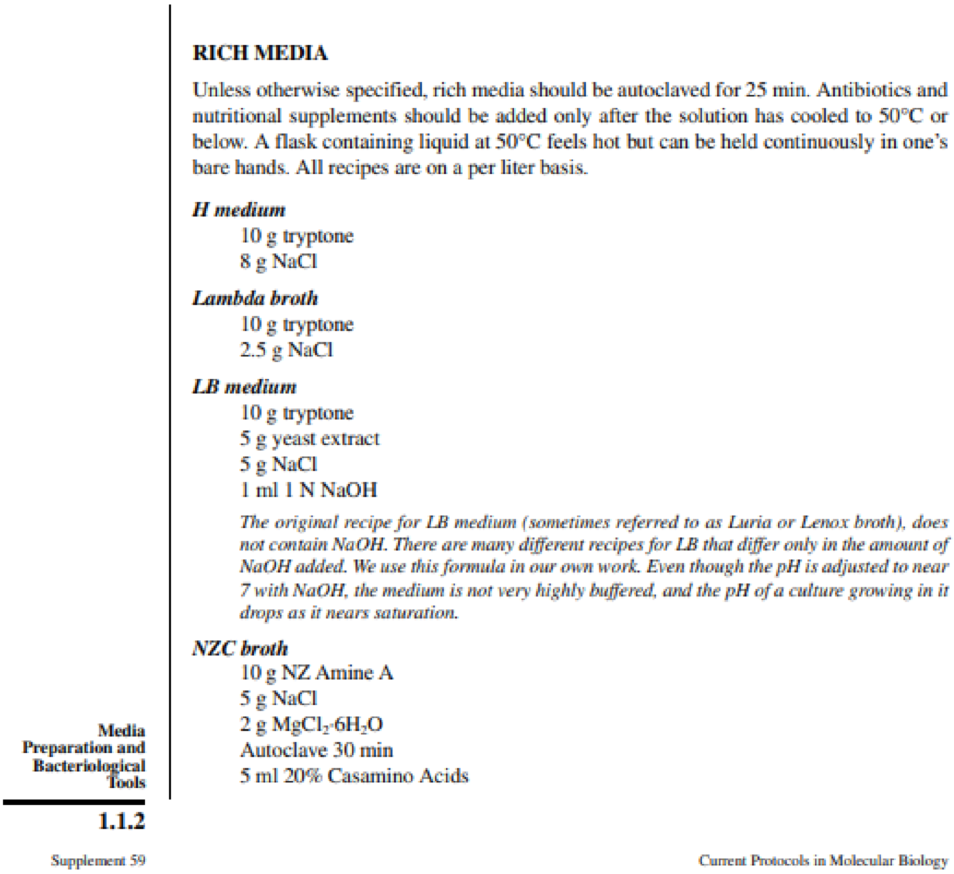
10 g $(NH_4)_2SO_4$
68 g $KH_2PO_4$
2.5 mg $FeSO_4{\cdot}7H_2O$
Adjust to pH 7 with KOH

***A medium, 5×***

5 g $(NH_4)_2SO_4$
22.5 g $KH_2PO_4$
52.5 g $K_2HPO_4$
2.5 g sodium citrate·$2H_2O$

Before they are used, concentrated media should be diluted to 1× with sterile water and the following sterile solutions, per liter:

1 ml 1 M $MgSO_4{\cdot}7H_2O$
10 ml 20% carbon source (sugar or glycerol)
*and, if required:*
0.1 ml 0.5% vitamin B1 (thiamine)
5 ml 20% Casamino Acids *or*
L amino acids to 40 µg/ml *or*
DL amino acids to 80 µg/ml
Antibiotic (see Table 1.4.1)



## RICH MEDIA

Unless otherwise specified, rich media should be autoclaved for 25 min. Antibiotics and nutritional supplements should be added only after the solution has cooled to 50°C or below. A flask containing liquid at 50°C feels hot but can be held continuously in one's bare hands. All recipes are on a per liter basis.

***H medium***

10 g tryptone
8 g NaCl

***Lambda broth***

10 g tryptone
2.5 g NaCl

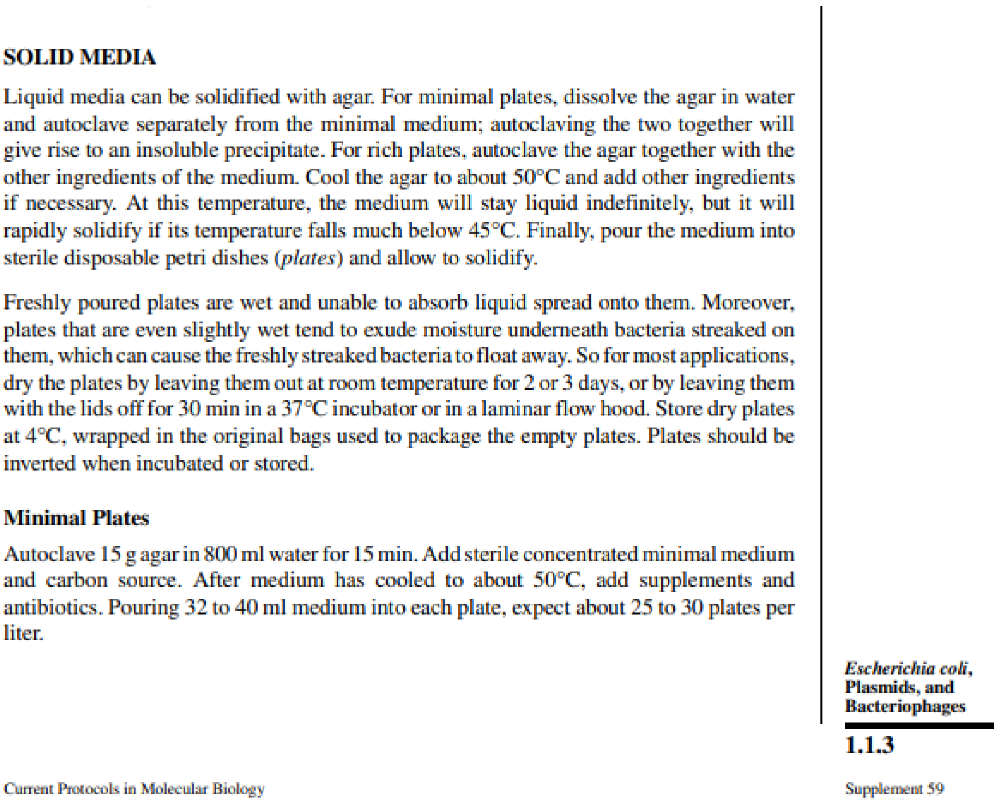
***LB medium***

10 g tryptone
5 g yeast extract
5 g NaCl
1 ml 1 N NaOH

*The original recipe for LB medium (sometimes referred to as Luria or Lenox broth), does not contain NaOH. There are many different recipes for LB that differ only in the amount of NaOH added. We use this formula in our own work. Even though the pH is adjusted to near 7 with NaOH, the medium is not very highly buffered, and the pH of a culture growing in it drops as it nears saturation.*

***NZC broth***

10 g NZ Amine A
5 g NaCl
2 g $MgCl_2 \cdot 6H_2O$
Autoclave 30 min
5 ml 20% Casamino Acids



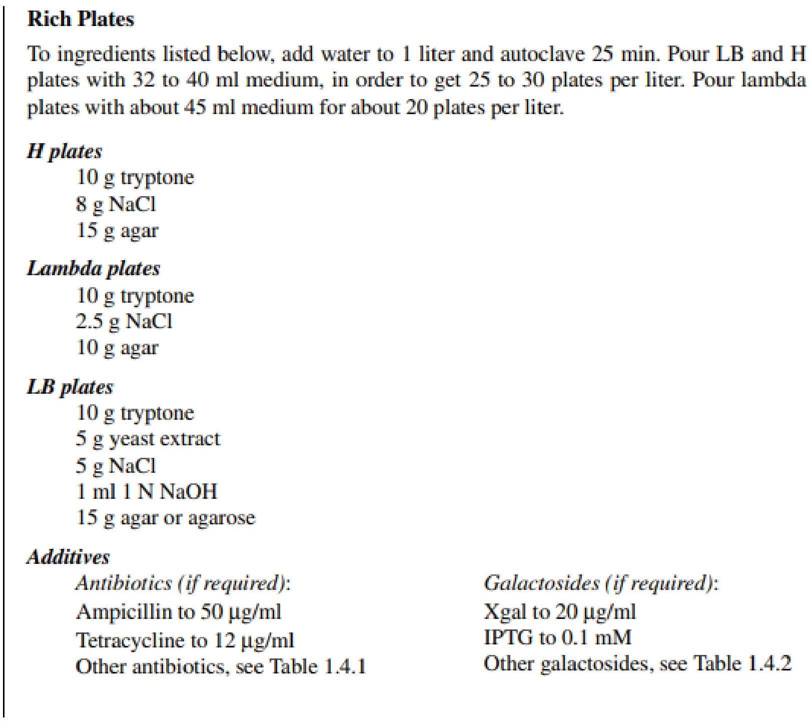
## SOLID MEDIA

Liquid media can be solidified with agar. For minimal plates, dissolve the agar in water and autoclave separately from the minimal medium; autoclaving the two together will give rise to an insoluble precipitate. For rich plates, autoclave the agar together with the other ingredients of the medium. Cool the agar to about 50°C and add other ingredients if necessary. At this temperature, the medium will stay liquid indefinitely, but it will rapidly solidify if its temperature falls much below 45°C. Finally, pour the medium into sterile disposable petri dishes (*plates*) and allow to solidify.

Freshly poured plates are wet and unable to absorb liquid spread onto them. Moreover, plates that are even slightly wet tend to exude moisture underneath bacteria streaked on them, which can cause the freshly streaked bacteria to float away. So for most applications, dry the plates by leaving them out at room temperature for 2 or 3 days, or by leaving them with the lids off for 30 min in a 37°C incubator or in a laminar flow hood. Store dry plates at 4°C, wrapped in the original bags used to package the empty plates. Plates should be inverted when incubated or stored.

### Minimal Plates

Autoclave 15 g agar in 800 ml water for 15 min. Add sterile concentrated minimal medium and carbon source. After medium has cooled to about 50°C, add supplements and antibiotics. Pouring 32 to 40 ml medium into each plate, expect about 25 to 30 plates per liter.



**Rich Plates**

To ingredients listed below, add water to 1 liter and autoclave 25 min. Pour LB and H plates with 32 to 40 ml medium, in order to get 25 to 30 plates per liter. Pour lambda plates with about 45 ml medium for about 20 plates per liter.

***H plates***

10 g tryptone
8 g NaCl
15 g agar

***Lambda plates***

10 g tryptone
2.5 g NaCl
10 g agar

***LB plates***

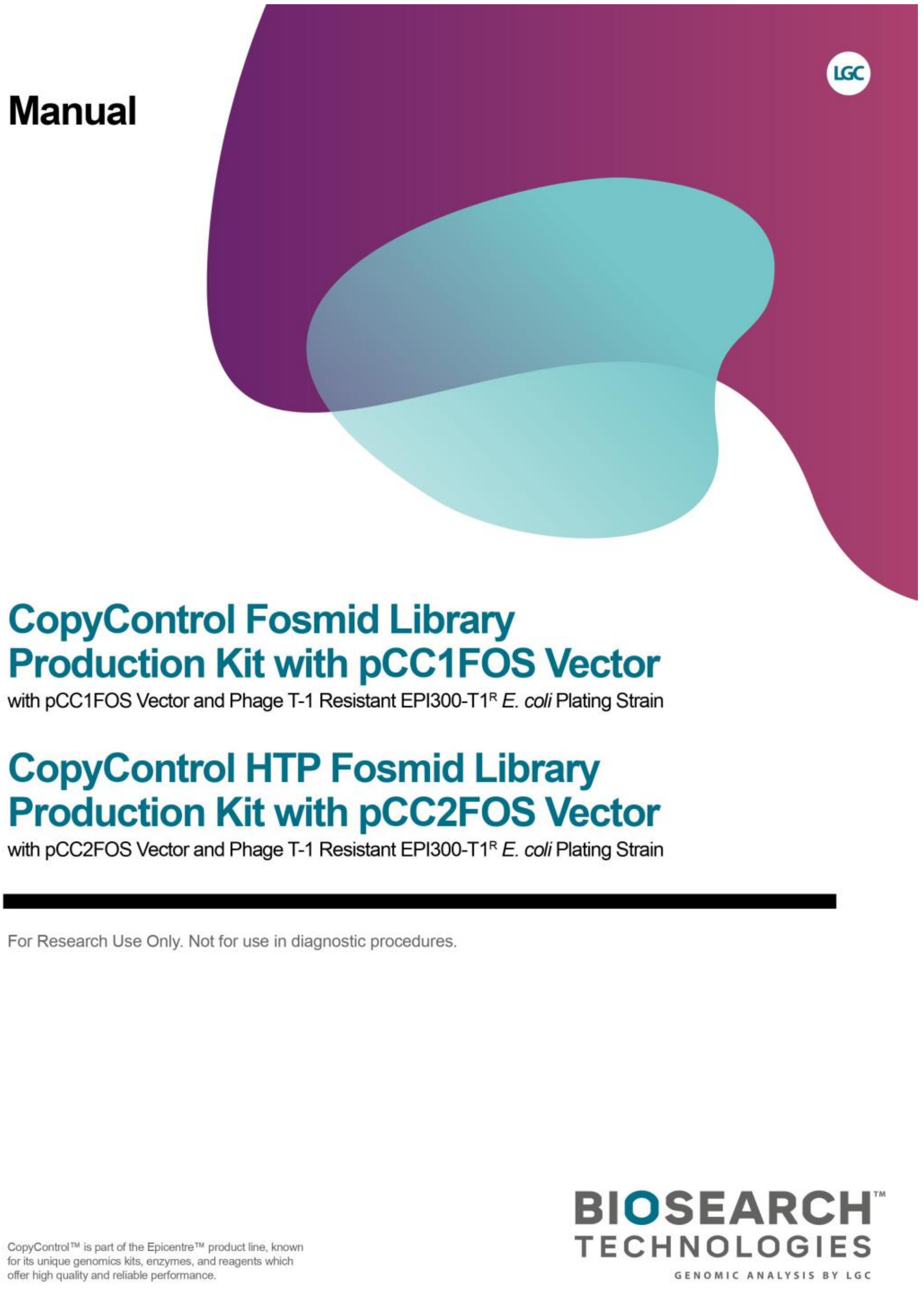
10 g tryptone
5 g yeast extract
5 g NaCl
1 ml 1 N NaOH
15 g agar or agarose

***Additives***

*Antibiotics (if required)*:
Ampicillin to 50 µg/ml
Tetracycline to 12 µg/ml
Other antibiotics, see Table 1.4.1

*Galactosides (if required)*:
Xgal to 20 µg/ml
IPTG to 0.1 mM
Other galactosides, see Table 1.4.2

# Appendix 3: Copycontrol HTP Fosmid library production kit with pCC2FOS vector



Manual

LGC

## CopyControl Fosmid Library Production Kit with pCC1FOS Vector

with pCC1FOS Vector and Phage T-1 Resistant EPI300-T1$^{R}$ *E. coli* Plating Strain

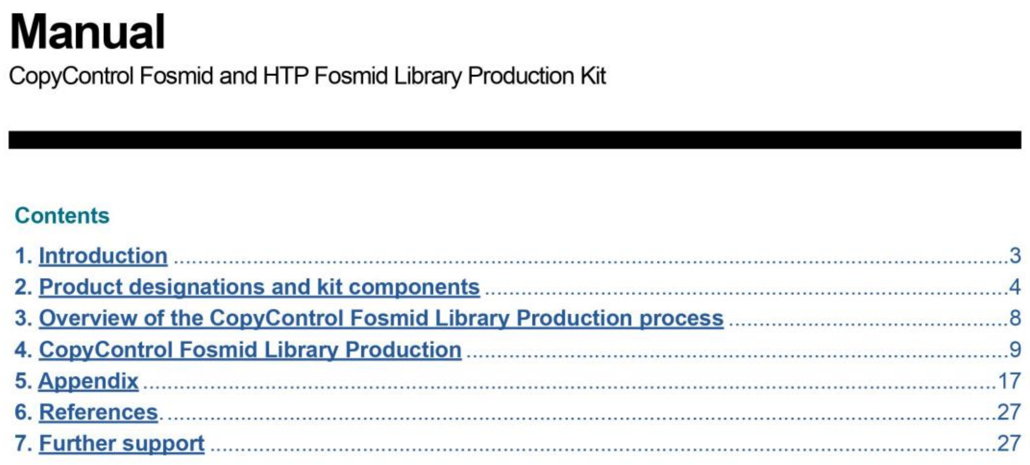
## CopyControl HTP Fosmid Library Production Kit with pCC2FOS Vector

with pCC2FOS Vector and Phage T-1 Resistant EPI300-T1$^{R}$ *E. coli* Plating Strain

For Research Use Only. Not for use in diagnostic procedures.

CopyControl™ is part of the Epicentre™ product line, known for its unique genomics kits, enzymes, and reagents which offer high quality and reliable performance.

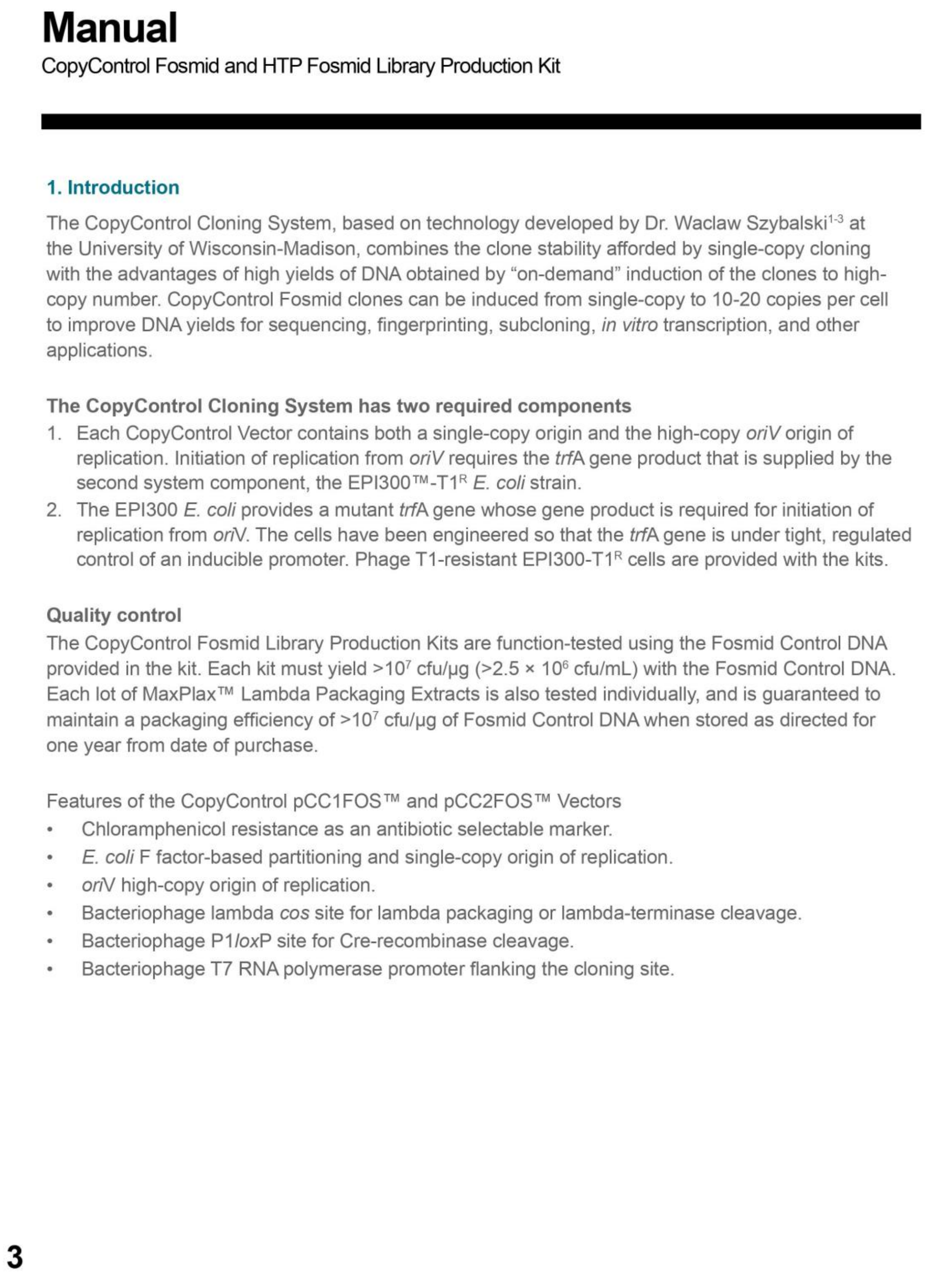
BIOSEARCH™
TECHNOLOGIES
GENOMIC ANALYSIS BY LGC



# Manual

CopyControl Fosmid and HTP Fosmid Library Production Kit

## Contents



## 1. Introduction

The CopyControl Cloning System, based on technology developed by Dr. Waclaw Szybalski[1-3] at the University of Wisconsin-Madison, combines the clone stability afforded by single-copy cloning with the advantages of high yields of DNA obtained by "on-demand" induction of the clones to high-copy number. CopyControl Fosmid clones can be induced from single-copy to 10-20 copies per cell to improve DNA yields for sequencing, fingerprinting, subcloning, *in vitro* transcription, and other applications.

### The CopyControl Cloning System has two required components

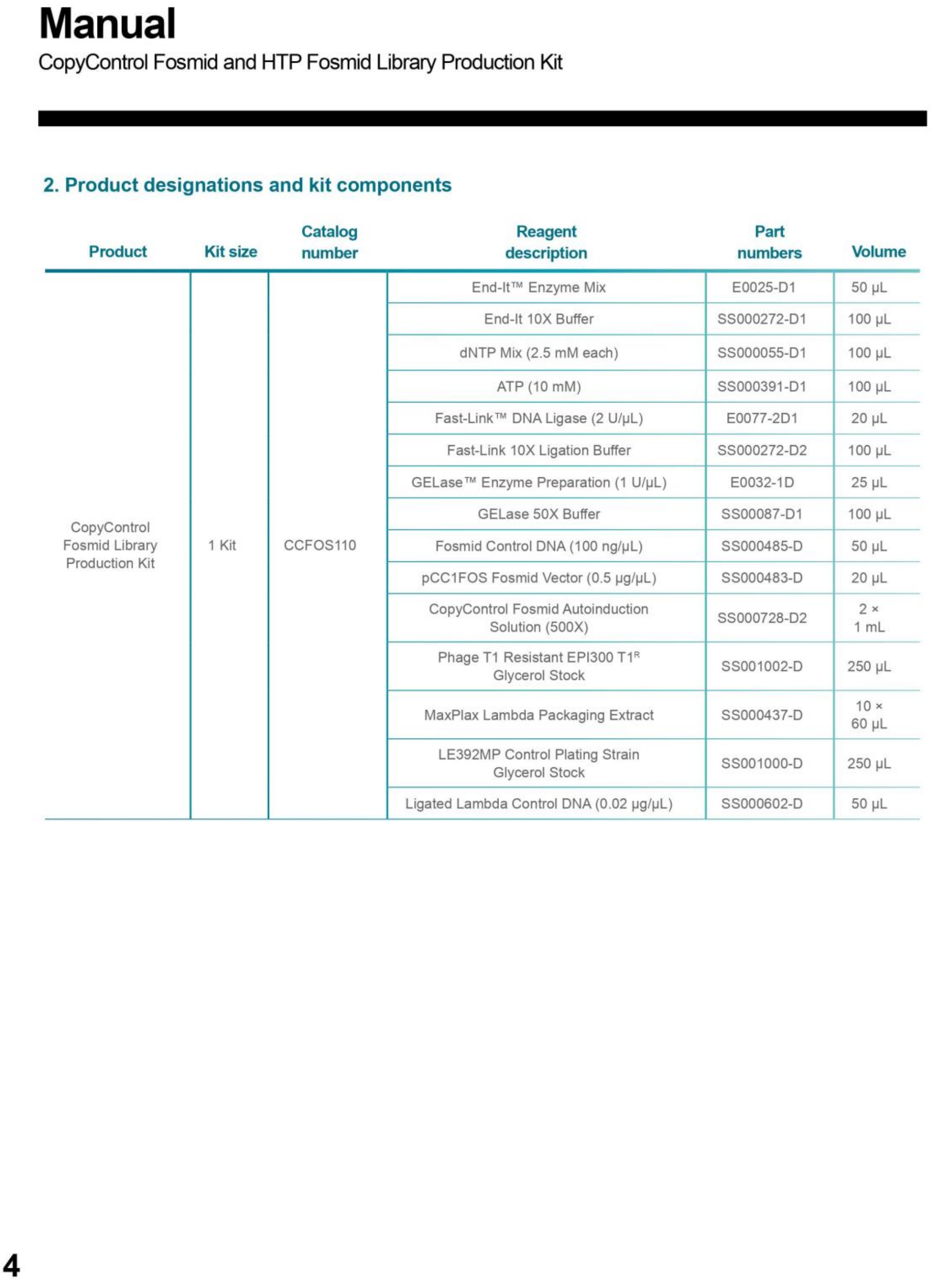
1. Each CopyControl Vector contains both a single-copy origin and the high-copy *oriV* origin of replication. Initiation of replication from *oriV* requires the *trf*A gene product that is supplied by the second system component, the EPI300™-T1$^R$ *E. coli* strain.
2. The EPI300 *E. coli* provides a mutant *trf*A gene whose gene product is required for initiation of replication from *ori*V. The cells have been engineered so that the *trf*A gene is under tight, regulated control of an inducible promoter. Phage T1-resistant EPI300-T1$^R$ cells are provided with the kits.

### Quality control

The CopyControl Fosmid Library Production Kits are function-tested using the Fosmid Control DNA provided in the kit. Each kit must yield >$10^7$ cfu/µg (>2.5 × $10^6$ cfu/mL) with the Fosmid Control DNA. Each lot of MaxPlax™ Lambda Packaging Extracts is also tested individually, and is guaranteed to maintain a packaging efficiency of >$10^7$ cfu/µg of Fosmid Control DNA when stored as directed for one year from date of purchase.

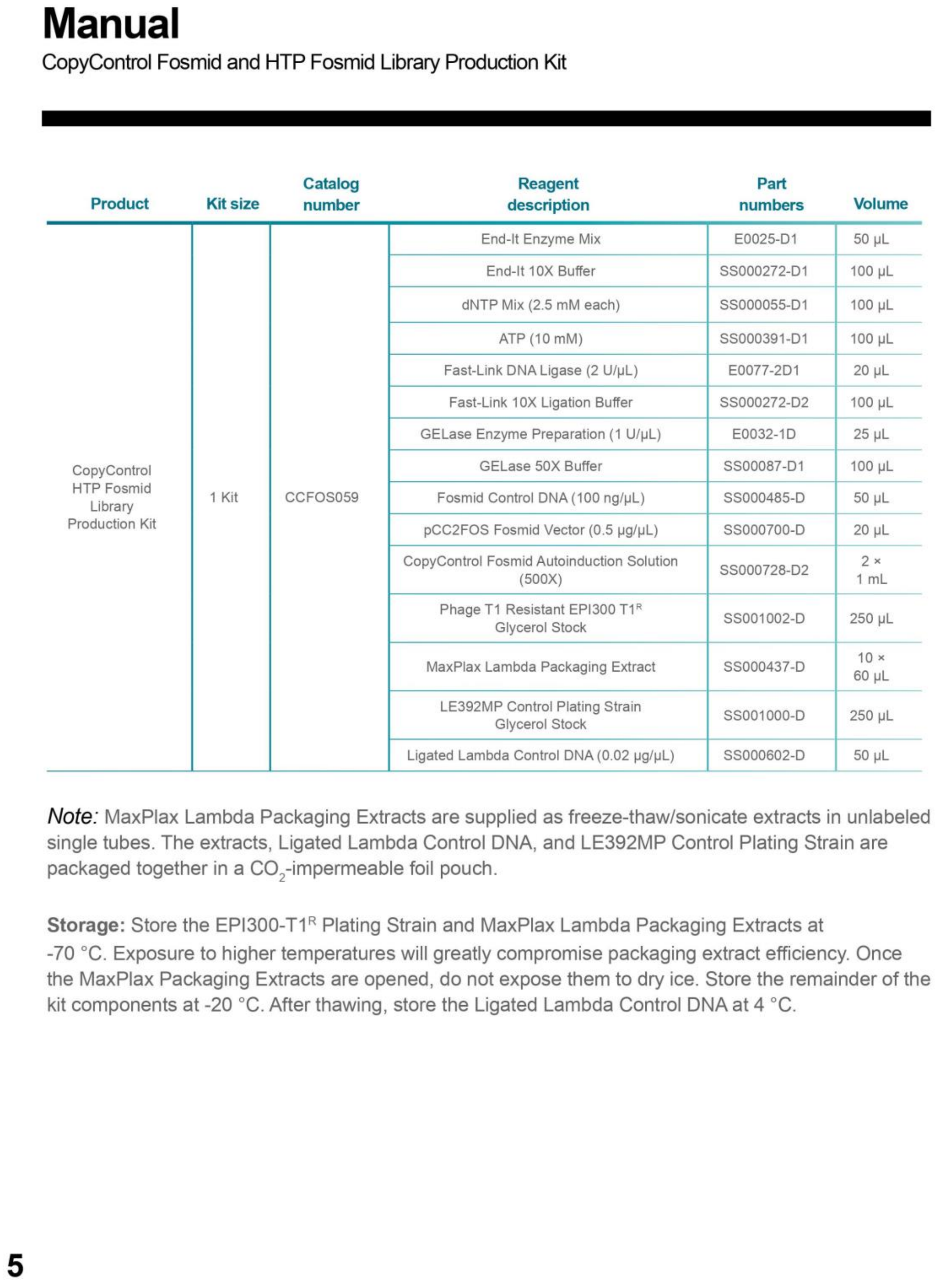
Features of the CopyControl pCC1FOS™ and pCC2FOS™ Vectors

- Chloramphenicol resistance as an antibiotic selectable marker.
- *E. coli* F factor-based partitioning and single-copy origin of replication.
- *ori*V high-copy origin of replication.
- Bacteriophage lambda *cos* site for lambda packaging or lambda-terminase cleavage.
- Bacteriophage P1*lox*P site for Cre-recombinase cleavage.
- Bacteriophage T7 RNA polymerase promoter flanking the cloning site.



## 2. Product designations and kit components

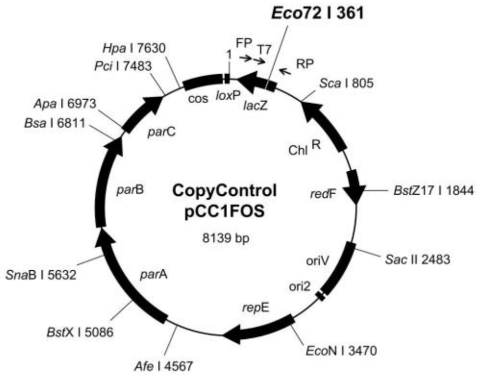
| Product | Kit size | Catalog number | Reagent description | Part numbers | Volume |
|---|---|---|---|---|---|
| CopyControl Fosmid Library Production Kit | 1 Kit | CCFOS110 | End-It™ Enzyme Mix | E0025-D1 | 50 µL |
| | | | End-It 10X Buffer | SS000272-D1 | 100 µL |
| | | | dNTP Mix (2.5 mM each) | SS000055-D1 | 100 µL |
| | | | ATP (10 mM) | SS000391-D1 | 100 µL |
| | | | Fast-Link™ DNA Ligase (2 U/µL) | E0077-2D1 | 20 µL |
| | | | Fast-Link 10X Ligation Buffer | SS000272-D2 | 100 µL |
| | | | GELase™ Enzyme Preparation (1 U/µL) | E0032-1D | 25 µL |
| | | | GELase 50X Buffer | SS00087-D1 | 100 µL |
| | | | Fosmid Control DNA (100 ng/µL) | SS000485-D | 50 µL |
| | | | pCC1FOS Fosmid Vector (0.5 µg/µL) | SS000483-D | 20 µL |
| | | | CopyControl Fosmid Autoinduction Solution (500X) | SS000728-D2 | 2 × 1 mL |
| | | | Phage T1 Resistant EPI300 T1$^R$ Glycerol Stock | SS001002-D | 250 µL |
| | | | MaxPlax Lambda Packaging Extract | SS000437-D | 10 × 60 µL |
| | | | LE392MP Control Plating Strain Glycerol Stock | SS001000-D | 250 µL |
| | | | Ligated Lambda Control DNA (0.02 µg/µL) | SS000602-D | 50 µL |



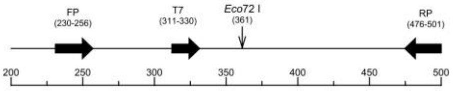
| Product | Kit size | Catalog number | Reagent description | Part numbers | Volume |
|---|---|---|---|---|---|
| CopyControl HTP Fosmid Library Production Kit | 1 Kit | CCFOS059 | End-It Enzyme Mix | E0025-D1 | 50 µL |
| | | | End-It 10X Buffer | SS000272-D1 | 100 µL |
| | | | dNTP Mix (2.5 mM each) | SS000055-D1 | 100 µL |
| | | | ATP (10 mM) | SS000391-D1 | 100 µL |
| | | | Fast-Link DNA Ligase (2 U/µL) | E0077-2D1 | 20 µL |
| | | | Fast-Link 10X Ligation Buffer | SS000272-D2 | 100 µL |
| | | | GELase Enzyme Preparation (1 U/µL) | E0032-1D | 25 µL |
| | | | GELase 50X Buffer | SS00087-D1 | 100 µL |
| | | | Fosmid Control DNA (100 ng/µL) | SS000485-D | 50 µL |
| | | | pCC2FOS Fosmid Vector (0.5 µg/µL) | SS000700-D | 20 µL |
| | | | CopyControl Fosmid Autoinduction Solution (500X) | SS000728-D2 | 2 × 1 mL |
| | | | Phage T1 Resistant EPI300 T1$^R$ Glycerol Stock | SS001002-D | 250 µL |
| | | | MaxPlax Lambda Packaging Extract | SS000437-D | 10 × 60 µL |
| | | | LE392MP Control Plating Strain Glycerol Stock | SS001000-D | 250 µL |
| | | | Ligated Lambda Control DNA (0.02 µg/µL) | SS000602-D | 50 µL |

*Note:* MaxPlax Lambda Packaging Extracts are supplied as freeze-thaw/sonicate extracts in unlabeled single tubes. The extracts, Ligated Lambda Control DNA, and LE392MP Control Plating Strain are packaged together in a $CO_2$-impermeable foil pouch.

**Storage:** Store the EPI300-T1$^R$ Plating Strain and MaxPlax Lambda Packaging Extracts at -70 °C. Exposure to higher temperatures will greatly compromise packaging extract efficiency. Once the MaxPlax Packaging Extracts are opened, do not expose them to dry ice. Store the remainder of the kit components at -20 °C. After thawing, store the Ligated Lambda Control DNA at 4 °C.



Note: Not all restriction enzymes that cut only once are indicated above.
See Appendix E for complete restriction information.
Primers are not drawn to scale.



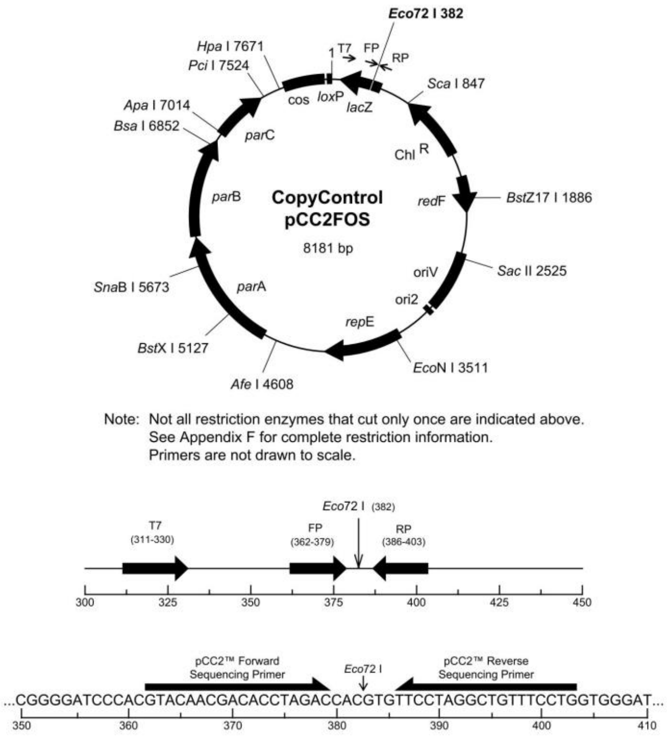
| | | |
|---|---|---|
| FP = pCC1™ | Foward Sequencing Primer | 5' GGATGTGCTGCAAGGCGATTAAGTTGG 3' |
| RP = pCC1™ | Reverse Sequencing Primer | 5' CTCGTATGTTGTGTGGAATTGTGAGC 3' |
| | T7 = T7 Promoter Primer | 5' TAATACGACTCACTATAGGG 3' |

Figure 1. pCC1FOS Vector Map.

### Additional required reagents

In addition to the component supplied, the following reagents are required:

- LB broth + 10 mM $MgSO_4$ + 0.2% Maltose
- Low-melting-point (LMP) agarose
- Ethanol (100% and 70%)
- 3 M Sodium Acetate (pH 7.0)
- Phage Dilution Buffer (10 mM Tris-HCl [pH 8.3], 100 mM NaCl, 10 mM $MgCl_2$)
- TE Buffer (10 mM Tris-HCl [pH 7.5], 1 mM EDTA)



Note: Not all restriction enzymes that cut only once are indicated above.
See Appendix F for complete restriction information.
Primers are not drawn to scale.

| | |
|---|---|
| FP = pCC2™ Foward Sequencing Primer | 5' GTACAACGACACCTAGAC 3' |
| RP = pCC2™ Reverse Sequencing Primer | 5' CAGGAAACAGCCTAGGAA 3' |
| T7 = T7 Promoter Primer | 5' TAATACGACTCACTATAGGG 3' |

Figure 2. pCC2FOS Vector map

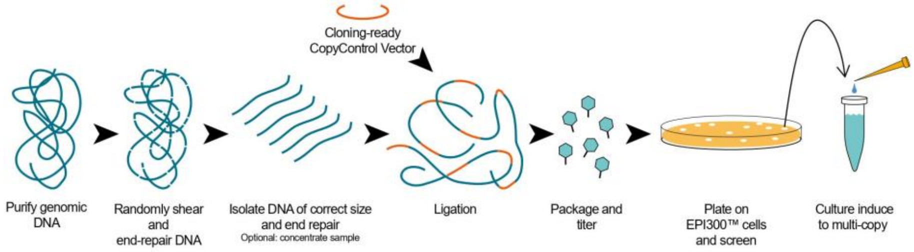

Cloning-ready
CopyControl Vector
Purify genomic DNA
Randomly shear and end-repair DNA
Isolate DNA of correct size and end repair
Optional: concentrate sample
Ligation
Package and titer
Plate on EPI300™ cells and screen
Culture induce to multi-copy

Figure 3. Production of a CopyControl Fosmid library and subsequent induction of clones to high-copy number.

7

**Additional features of the pCC2FOS Vector**

The CopyControl HTP Fosmid Library Production Kit contains the pCC2FOS Vector (Figure 2). The pCC2FOS Vector, a modification of the pCC1FOS (Figure 1) vector, contains a primer cassette that optimises end-sequencing results, especially in a high-throughput setting.[4] The pCC2FOS primer cassette eliminates wasteful vector-derived sequencing reads by having the 3′ terminus of the forward and reverse sequencing primers anneal three nucleotides from the cloning site. In addition, the seven-base sequence at the 3′ end of each primer was specifically designed to minimise mispriming from any contaminating *E. coli* DNA present after template purification.

**How the CopyControl Cloning System works** (Figure 3)

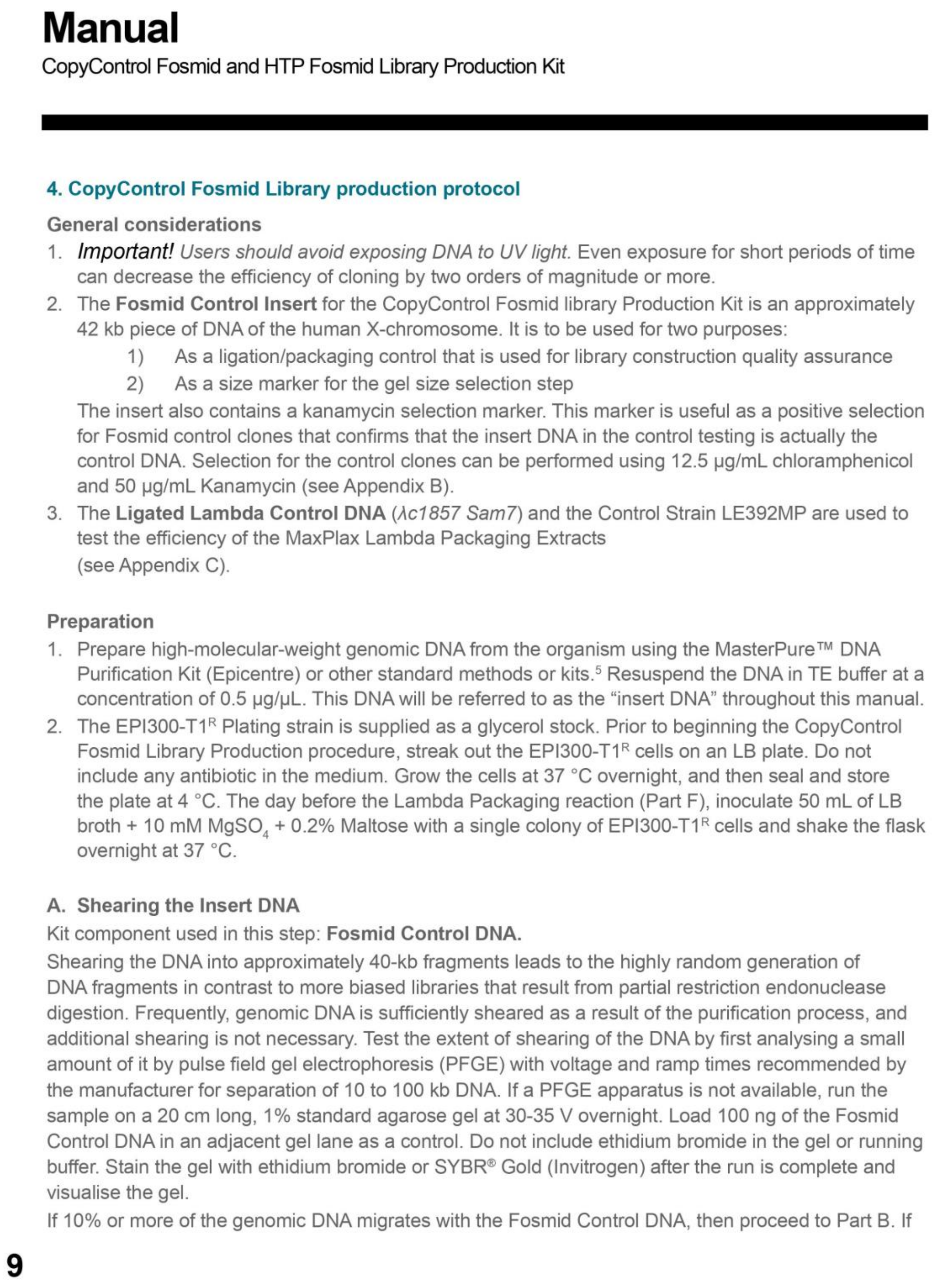
1. Ligate the DNA of interest into the linearised and dephosphorylated CopyControl Cloning-Ready Vector supplied with the respective kit.
2. Package the ligated DNA into the lambda phage and infect EPI300-T1$^R$ E. coli and select on LB-chloramphenicol plates. Under these conditions, the *trfA* gene is not expressed and the clones are maintained at single-copy.
3. Pick individual CopyControl clones from the plate and grow in culture.
4. Add the CopyControl Fosmid Autoinduction Solution (included) or CopyControl Induction Solution (available separately) to induce expression of the *trfA* gene product and subsequent amplification of the clones to high-copy number.
5. Purify plasmid DNA for sequencing, fingerprinting, subcloning, or other applications.

## 3. Overview of the CopyControl Fosmid Library Production process

The CopyControl Fosmid Library Production Kits will produce a complete and unbiased primary fosmid library in about 2 days. The kit utilises a novel strategy of cloning randomly sheared, end-repaired DNA. Shearing the DNA leads to the generation of highly random DNA fragments in contrast to more biased libraries that result from fragmenting the DNA by partial restriction digests.

The steps involved (protocols for steps 2-8 are included in this manual):

1. Purify DNA from the desired source (the kit does not supply materials for this step).
2. Shear the DNA to approximately 40-kb fragments.
3. End-repair the sheared DNA to blunt, 5′-phosphorylated ends.
4. Isolate the desired size range of end-repaired DNA by LMP agarose gel electrophoresis.
5. Purify the blunt-ended DNA from the LMP agarose gel.
6. Ligate the blunt-ended DNA to the Cloning-Ready CopyControl pCC1FOS or pCC2FOS Vector.
7. Package the ligated DNA and plate on EPI300-T1$^R$ plating cells. Grow clones overnight.
8. Pick CopyControl Fosmid clones of interest and induce them to high-copy number using the CopyControl Fosmid Autoinduction Solution.
9. Purify DNA for sequencing, fingerprinting, subcloning, or other applications. The kit does not supply materials for this step.



## 4. CopyControl Fosmid Library production protocol

### General considerations

1. *Important! Users should avoid exposing DNA to UV light.* Even exposure for short periods of time can decrease the efficiency of cloning by two orders of magnitude or more.
2. The **Fosmid Control Insert** for the CopyControl Fosmid library Production Kit is an approximately 42 kb piece of DNA of the human X-chromosome. It is to be used for two purposes:
   1) As a ligation/packaging control that is used for library construction quality assurance
   2) As a size marker for the gel size selection step

   The insert also contains a kanamycin selection marker. This marker is useful as a positive selection for Fosmid control clones that confirms that the insert DNA in the control testing is actually the control DNA. Selection for the control clones can be performed using 12.5 µg/mL chloramphenicol and 50 µg/mL Kanamycin (see Appendix B).
3. The **Ligated Lambda Control DNA** (*λc1857 Sam7*) and the Control Strain LE392MP are used to test the efficiency of the MaxPlax Lambda Packaging Extracts
   (see Appendix C).

### Preparation

1. Prepare high-molecular-weight genomic DNA from the organism using the MasterPure™ DNA Purification Kit (Epicentre) or other standard methods or kits.[5] Resuspend the DNA in TE buffer at a concentration of 0.5 µg/µL. This DNA will be referred to as the "insert DNA" throughout this manual.
2. The EPI300-T1$^R$ Plating strain is supplied as a glycerol stock. Prior to beginning the CopyControl Fosmid Library Production procedure, streak out the EPI300-T1$^R$ cells on an LB plate. Do not include any antibiotic in the medium. Grow the cells at 37 °C overnight, and then seal and store the plate at 4 °C. The day before the Lambda Packaging reaction (Part F), inoculate 50 mL of LB broth + 10 mM $MgSO_4$ + 0.2% Maltose with a single colony of EPI300-T1$^R$ cells and shake the flask overnight at 37 °C.

### A. Shearing the Insert DNA

Kit component used in this step: **Fosmid Control DNA.**

Shearing the DNA into approximately 40-kb fragments leads to the highly random generation of DNA fragments in contrast to more biased libraries that result from partial restriction endonuclease digestion. Frequently, genomic DNA is sufficiently sheared as a result of the purification process, and additional shearing is not necessary. Test the extent of shearing of the DNA by first analysing a small amount of it by pulse field gel electrophoresis (PFGE) with voltage and ramp times recommended by the manufacturer for separation of 10 to 100 kb DNA. If a PFGE apparatus is not available, run the sample on a 20 cm long, 1% standard agarose gel at 30-35 V overnight. Load 100 ng of the Fosmid Control DNA in an adjacent gel lane as a control. Do not include ethidium bromide in the gel or running buffer. Stain the gel with ethidium bromide or SYBR® Gold (Invitrogen) after the run is complete and visualise the gel.

If 10% or more of the genomic DNA migrates with the Fosmid Control DNA, then proceed to Part B. If the genomic DNA migrates slower (higher MW) than the Fosmid Control DNA, then the DNA needs to be sheared further as described below. If the genomic DNA migrates faster than the Fosmid Control DNA (lower MW) then it has been sheared too much and should be reisolated.

If shearing is required, we recommend that at least 2.5 µg (at a concentration of 500 ng/µL) of DNA be used. Randomly shear the DNA by passing it through a 200-µL small-bore pipette tip. Aspirate and expel the DNA from the pipette tip 50-100 times. Examine 1-2 µL of the DNA on a 20-cm agarose gel using the Fosmid Control DNA as a size marker. If 10% or more of the genomic DNA migrates with the Fosmid Control DNA, then proceed to Part B. If >90% of the sheared DNA comigrates with the Fosmid Control DNA and appears as a relatively tight band (as in Fig. 4, lane 3), gel size-selection may not be necessary; you may skip the gel-sizing step and proceed directly with ligation of the DNA to the vector (Part E). If the DNA is still too large, aspirate and expel the DNA from the pipette tip an additional 50 times. Examine 1-2 µL of this DNA by agarose gel electrophoresis as described previously.

### B. End-Repair of the Insert DNA

Kit components used in this step: **End-Repair Enzyme Mix, 10X Buffer, dNTPs, ATP.**

This step generates blunt-ended, 5′-phosphorylated DNA. The end-repair reaction can be scaled as dictated by the amount of DNA available.

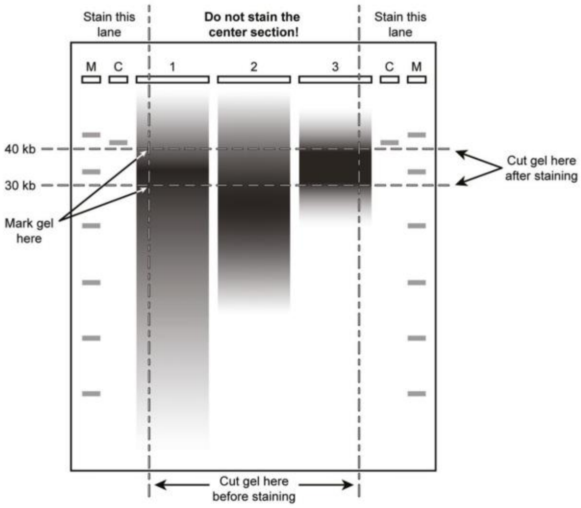

Stain this lane
Do not stain the center section!
Stain this lane
M
C
1
2
3
C
M
40 kb
30 kb
Mark gel here
Cut gel here after staining
Cut gel here before staining

Figure 4. Gel purification of DNA: Keeping ethidium and UV away from your DNA.

1. Thaw and thoroughly mix all of the reagents listed below before dispensing; place on ice. Combine the following on ice:

   x µL sterile water
   8 µL 10X End-Repair Buffer
   8 µL 2.5 mM dNTP Mix
   8 µL 10 mM ATP
   up to 20 µg sheared insert DNA (approximately 0.5 µg/µL)
   4 µL End-Repair Enzyme Mix
   80 µL Total reaction volume

2. Incubate at room temperature for 45 minutes.
3. Add gel loading buffer and incubate at 70 °C for 10 minutes to inactivate the End-Repair Enzyme Mix. Proceed with Size selection of the End-Repaired DNA in Part C.

### C. Size Selection of the End-Repaired DNA

Kit components used in this step: **Fosmid Control DNA.**

If the DNA to be used in the cloning process appears as a long smear (Fig. 4, Lanes 1 and 2), size-select the end-repaired DNA by LMP agarose gel electrophoresis. Ideally, use PFGE with voltage and ramp times recommended by the manufacturer for separation of 10 to 100 kb DNA. If a PFGE apparatus is not available, analyse the sample on a 20 cm long, 1% LMP agarose gel at 30-35 V overnight. **Minigels (e.g. 10 cm) do not provide sufficient resolution of DNA in the 20- to 60-kb size range.**

Fractionate the DNA on an LMP agarose gel. **It is important to perform this electrophoresis in the absence of ethidium bromide (do not add ethidium bromide to the gel). The DNA that will be cloned should not be exposed to UV light under any circumstances.** This can decrease the cloning efficiency by 100-fold or more. A diagram of the recommended method is shown in Fig. 4.

*Note 1:* Even 30 seconds of exposure to 302 nm UV light will cause a 100 to 200 fold drop in ligation and cloning efficiency.

*Note 2:* The protocol below is designed for use with GELase Agarose Gel-Digesting Preparation (kit component), and thus requires LMP agarose. Standard high-melt agarose can also be used and the DNA extracted from the gel slices by other methods.

1. Prepare a 1% **LMP** agarose gel in 1X TAE or 1X TBE buffer. Use a wide comb as needed to be able to load sufficient DNA into the gel (see Figure 4).

   *Note:* Do not include ethidium bromide in the gel solution.
2. Load DNA size markers into each of the outside lanes of the gel. Load 100 ng of Fosmid Control DNA into each of the inner adjacent lanes of the gel. Load the end-repaired insert DNA in the lane(s) between the Fosmid Control DNA lanes.
3. Resolve the samples by gel electrophoresis at room temperature overnight at a constant voltage of 30-35 V. Do not include any DNA stain in the gel or in the gel running buffer during electrophoresis.
4. Following electrophoresis, cut off the outer lanes of the gel containing the DNA size markers, the Fosmid Control DNA, and a small portion of the next lane that contains your random sheared end-repaired genomic DNA (see Figure 4).
5. Stain the cut-off sides of the gel with ethidium bromide or SYBR Gold (Invitrogen), which is more sensitive than ethidium bromide, and visualize the DNA with UV light. Mark the position of the desired size DNA in the gel using a pipet tip or a razor blade.

   *Note:* Do not expose the sample DNA to UV! Even short-duration UV exposure can decrease cloning efficiencies by 100 to 1,000 fold.
6. Reassemble the gel and excise a gel slice that is 2- to 4-mm below the position of the Fosmid Control DNA.

   *Caution:* Be sure to cut the gel slice so that the DNA recovered is ≥25 kb. Cloning DNA smaller than ~25 kb may result in unwanted chimeric clones.

   *Note:* Prior to reassembly, without breaking the gel, carefully rinse the stained gel with distilled water to remove excess stain from the gel pieces. This will prevent the gel pieces containing the sample DNA from being exposed to stain.
7. Transfer the gel slice to a tared, sterile, screw-cap tube for extraction, either by using the GELase method, or other desired method for isolating DNA from agarose gels. The size of the tube to be used will be dictated by the size and number of gel slices being digested with GELase enzyme.
8. Proceed with Recovery of the Size-Fractionated DNA in Part D or store the gel slice at 4 °C to -20 °C for up to 1 year.

### D. Recovery of the Size-Fractionated DNA

Kit components used in this step: **GELase 50X Buffer, GELase Enzyme Preparation.**

Before beginning this step, prepare a 70 °C and a 45 °C water bath or other temperature-regulated apparatus.

1. Weigh the tared tubes to determine the weight of the gel slice(s). Assume 1 mg of solidified agarose will yield 1 µL of molten agarose upon melting.
2. Warm the GELase 50X Buffer to 45 °C. Melt the LMP agarose by incubating the tube at 70 °C for 10-15 minutes. Quickly transfer the tube to 45 °C.

3. Add the appropriate volume of warmed GELase 50X Buffer to 1X final concentration. Carefully add 1 U (1 µL) of GELase Enzyme Preparation to the tube for each 600 µL of melted agarose. Keep the melted agarose solution at 45 °C and gently mix the solution. Incubate the solution at 45 °C for at least 1 hour (overnight incubation is acceptable, if desired).
4. Transfer the reaction to 70 °C for 10 minutes to inactivate the GELase enzyme.
5. Remove 500 µL aliquots of the solution into sterile, 1.5 mL microfuge tube(s).
6. Chill the tube(s) in an ice bath for 5 minutes. Centrifuge the tubes in a microcentrifuge at maximum speed (>10,000 x g) for 20 minutes to pellet any insoluble oligosaccharides. Any "pellet" will be gelatinous, and translucent to opaque. Carefully remove the upper 90%-95% of the supernatant, which contains the DNA, to a sterile 1.5 mL tube. Be careful to avoid the gelatinous pellet.
7. Precipitate the DNA.
   - a) Add 1/10 volume of 3 M sodium acetate (pH 7.0) and mix gently.
   - b) Add 2.5 volumes of ethanol. Cap the tube and mix by gentle inversion.
   - c) Allow precipitation to proceed for 10 minutes at room temperature.
   - d) Centrifuge the precipitated DNA for 20 minutes in a microcentrifuge, at top speed (>10,000 x g).
   - e) Carefully aspirate the supernatant from the pelleted DNA.
   - f) Wash the pellet twice with cold, 70% ethanol, repeating steps d) and e), using care not to disrupt the DNA pellet.
   - g) After the second 70% ethanol wash, carefully invert the tube and allow the pellet to air-dry for 5-10 minutes (longer dry times will make resuspension of the DNA difficult).
   - h) Gently resuspend the DNA pellet in TE Buffer.

   *Note:* A 10 µL ligation reaction volume allows a maximum 6 µL of input DNA.
8. Determine the DNA concentration by fluorimetry. Alternatively, estimate the concentration of the DNA by running an aliquot of the DNA on an agarose gel using dilutions of known amounts of the Fosmid Control DNA as standard.

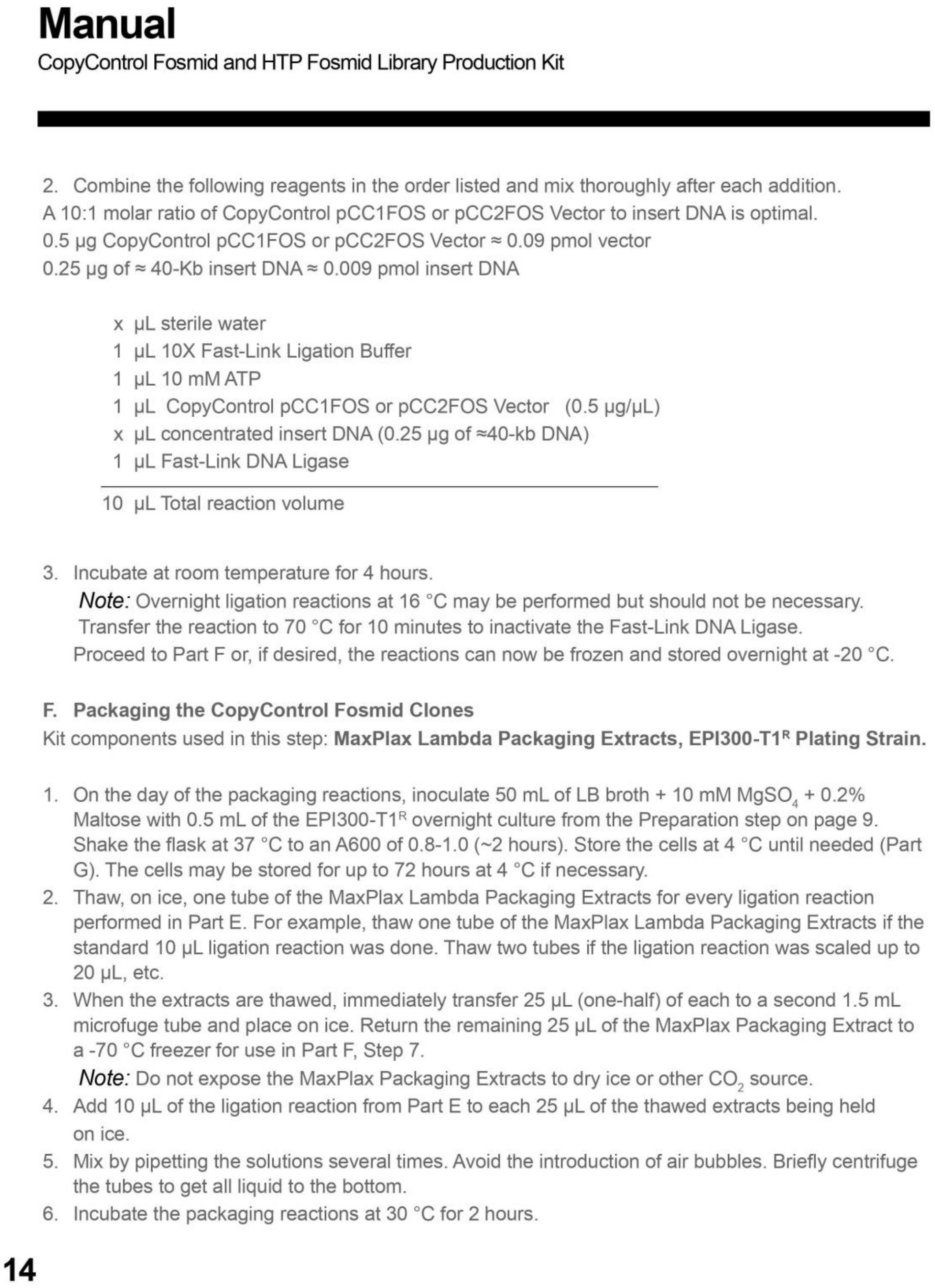
   *Note:* Measuring the DNA concentration by spectrophotometry (A260) is not recommended because the DNA concentration will not be high enough to be measured accurately.

   *Note:* If desired, the reactions can now be frozen and stored overnight at -20 °C.

### E. Ligation Reaction

Kit components used in this step: **Fast-Link 10X Ligation Buffer, Fast-Link DNA Ligase, ATP, CopyControl pCC1FOS or pCC2FOS Cloning-Ready Vector**

1. Please refer to Appendix A to determine the approximate number of CopyControl Fosmid clones that you will need for your library. A single ligation reaction will produce 103-106 clones, depending on the quality of the insert DNA. Based on this information, calculate the number of ligation reactions that you will need to perform. The ligation reaction can be scaled as needed.



2. Combine the following reagents in the order listed and mix thoroughly after each addition.

A 10:1 molar ratio of CopyControl pCC1FOS or pCC2FOS Vector to insert DNA is optimal.
0.5 µg CopyControl pCC1FOS or pCC2FOS Vector ≈ 0.09 pmol vector
0.25 µg of ≈ 40-Kb insert DNA ≈ 0.009 pmol insert DNA

| | |
|---|---|
| x | µL sterile water |
| 1 | µL 10X Fast-Link Ligation Buffer |
| 1 | µL 10 mM ATP |
| 1 | µL CopyControl pCC1FOS or pCC2FOS Vector (0.5 µg/µL) |
| x | µL concentrated insert DNA (0.25 µg of ≈40-kb DNA) |
| 1 | µL Fast-Link DNA Ligase |
| 10 | µL Total reaction volume |

3. Incubate at room temperature for 4 hours.
   *Note:* Overnight ligation reactions at 16 °C may be performed but should not be necessary.
   Transfer the reaction to 70 °C for 10 minutes to inactivate the Fast-Link DNA Ligase.
   Proceed to Part F or, if desired, the reactions can now be frozen and stored overnight at -20 °C.

### F. Packaging the CopyControl Fosmid Clones

Kit components used in this step: **MaxPlax Lambda Packaging Extracts, EPI300-T1$^R$ Plating Strain.**

1. On the day of the packaging reactions, inoculate 50 mL of LB broth + 10 mM $MgSO_4$ + 0.2% Maltose with 0.5 mL of the EPI300-T1$^R$ overnight culture from the Preparation step on page 9. Shake the flask at 37 °C to an A600 of 0.8-1.0 (~2 hours). Store the cells at 4 °C until needed (Part G). The cells may be stored for up to 72 hours at 4 °C if necessary.
2. Thaw, on ice, one tube of the MaxPlax Lambda Packaging Extracts for every ligation reaction performed in Part E. For example, thaw one tube of the MaxPlax Lambda Packaging Extracts if the standard 10 µL ligation reaction was done. Thaw two tubes if the ligation reaction was scaled up to 20 µL, etc.
3. When the extracts are thawed, immediately transfer 25 µL (one-half) of each to a second 1.5 mL microfuge tube and place on ice. Return the remaining 25 µL of the MaxPlax Packaging Extract to a -70 °C freezer for use in Part F, Step 7.
   *Note:* Do not expose the MaxPlax Packaging Extracts to dry ice or other $CO_2$ source.
4. Add 10 µL of the ligation reaction from Part E to each 25 µL of the thawed extracts being held on ice.
5. Mix by pipetting the solutions several times. Avoid the introduction of air bubbles. Briefly centrifuge the tubes to get all liquid to the bottom.
6. Incubate the packaging reactions at 30 °C for 2 hours.

7. After the 2-hour packaging reaction is complete, add the remaining 25 µL of MaxPlax Lambda Packaging Extract from Part F, Step 3 to each tube.
8. Incubate the reactions for an additional 2 hours at 30 °C.
9. At the end of the second incubation, add Phage Dilution Buffer (PDB) to 1 mL final volume in each tube and mix gently. Add 25 µL of chloroform to each. Mix gently and store at 4 °C. A viscous precipitate may form after addition of the chloroform. This precipitate will not interfere with library production. Determine the titer of the phage particles (packaged fosmid clones) in Part G, and then plate the fosmid library in Part H. Or, store the phage particles as described in Appendix D.

*Note:* In the construction of metagenomic fosmid libraries from environmental water or soil microbes, the amount of PDB to be added to the packaged phage may require some adjustment depending on the starting amount of DNA. If the DNA used in ligation is lower than the protocol recommends, then the addition of 0.5 mL of the PDB may be needed.

### G. Titering the packaged CopyControl Fosmid Clones

Kit components used in this step: **EPI300-T1$^R$ Plating Strain from Part F, Step 1.**

Before plating the library, we recommend that you determine the titer of the phage particles (packaged CopyControl Fosmid clones). This will aid in determining the number of plates and dilutions required to obtain a library that meets your needs.

1. Make serial dilutions of the 1 mL of packaged phage particles from Part F, Step 9 into Phage Dilution Buffer (PDB) in sterile microfuge tubes.
   A) 1:101 Dilute 10 µL of packaged phage into 90 µL of PDB.
   B) 1:102 Dilute 100 µL of the 1:101 dilution into 900 µL of PDB.
   C) 1:103 Dilute 100 µL of the 1:102 dilution into 900 µL of PDB.
2. Add 10 µL of each above dilution, and 10 µL of the undiluted phage, individually, to 100 µL of the prepared EPI300-T1$^R$ host cells from Part F, Step 1 above. Incubate each tube for 1 hour at 37 °C.
3. Spread the infected EPI300-T1$^R$ cells on an LB plate + 12.5 µg/mL chloramphenicol and incubate at 37 °C overnight to select for the CopyControl Fosmid clones.
4. Count colonies and calculate the titer of the packaged phage particles from Part F, Step 9.

**Sample Calculation:**

If there were 200 colonies on the plate streaked with the 1:103 dilution, then the titer in cfu/mL, (where cfu represents colony forming units) of this reaction would be:

$$\frac{(\text{\# of colonies})\ (\text{dilution factor})\ (1{,}000\ \mu\text{L/mL})}{(\text{volume of phage plated } [\mu\text{L}])} \quad \textbf{OR} \quad \frac{(200\ \text{cfu})\ (10^3)\ (1{,}000\ \mu\text{L/mL})}{(10\ \mu\text{L})} \ \mathbf{= 2 \times 10^7\ cfu/mL}$$

### H. Plating and selecting the CopyControl Fosmid Library

Based on the titer of the packaged CopyControl Fosmid clones and the estimated number of clones required (see Appendix A), calculate the volume of the packaged fosmid clones that will be needed to prepare the CopyControl Fosmid library.

1. Based on the titer of the phage particles determined in Part G, dilute the phage particles from Part F, Step 9 with Phage Dilution Buffer to obtain the desired number of clones and clone density on the plate. Proceed to the next step or store the diluted phage particles as described in Appendix D.
2. Mix the diluted phage particles from Part H, Step 1 with EPI300-T1$^R$ cells prepared in Part F, Step 1 in the ratio of 100 µL of cells for every 10 µL of diluted phage particles.
3. Incubate the tubes at 37 °C for 1 hour.
4. Spread the infected bacteria on an LB plate + 12.5 µg/mL chloramphenicol and incubate at 37 °C overnight to select for the CopyControl Fosmid clones.
5. We recommend plating as much of the library as possible. Storage of the phage library for more than 72 hours at 4 °C will result in a severe loss of phage viability and the plating efficiency will be severely compromised. We recommend storing the phage as an amplified library (see Appendix D, Method C) for best results.

### Induction of the CopyControl Fosmid Clones to High-Copy Number

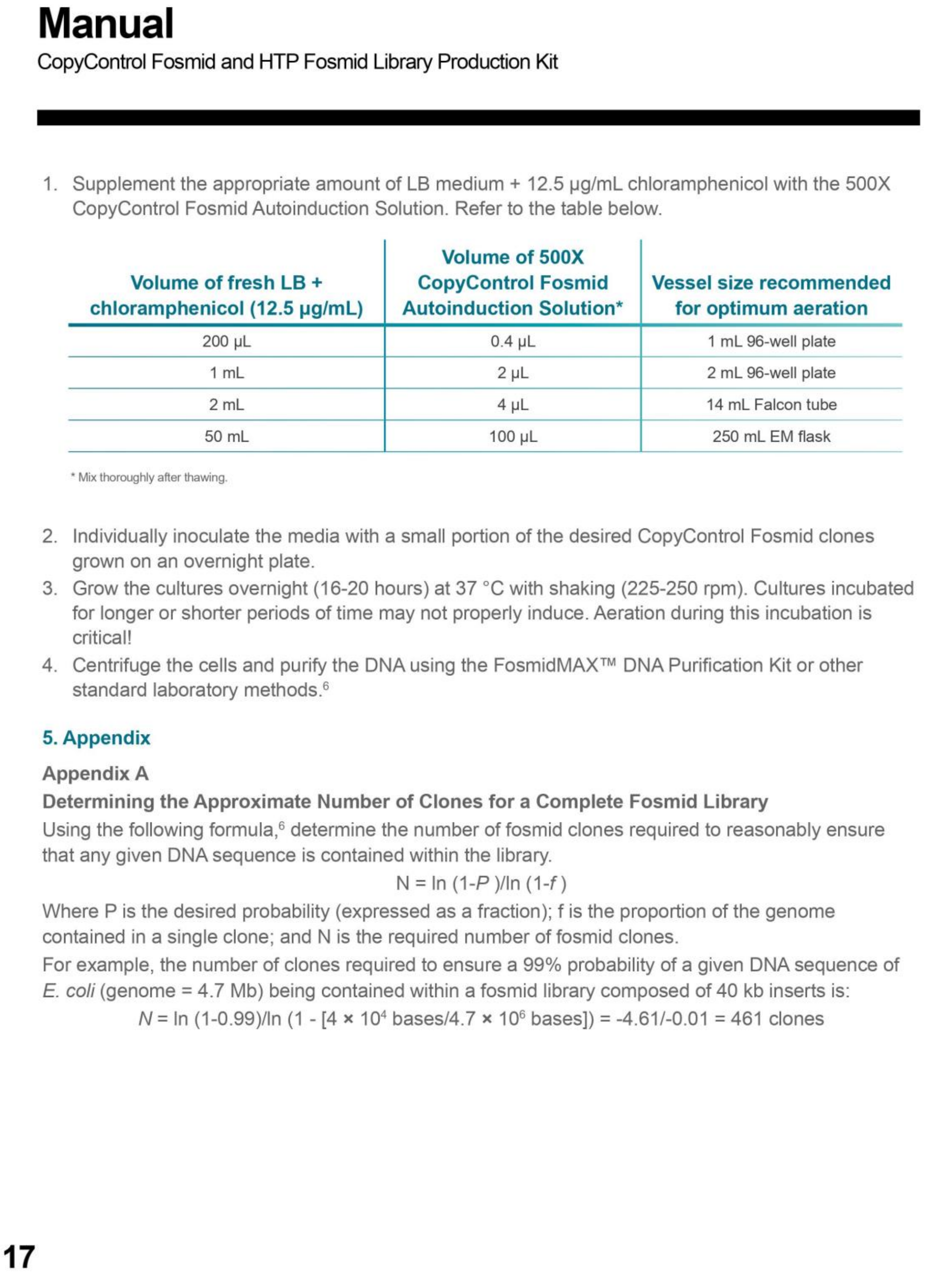
Once the desired CopyControl Fosmid clones are identified, they can be induced to high- copy number for high yields of DNA for sequencing, fingerprinting, or other applications.

The CopyControl Fosmid Autoinduction Solution can be supplemented into the cultures prior to inoculation and requires no subculturing of the bacteria. It is ideal for growing fosmid clones in any culture volume, including 96-well format or other high-throughput applications where subculturing is tedious and time-consuming.

The copy-number induction process can be done in any culture volume desired, depending on your needs. Generally, a 1-mL culture will provide sufficient DNA (typically 1-2 µg) for most applications. Below, we provide the standard autoinduction procedure for amplifying the clones in 200 µL, 1 mL, 2 mL, and 50 mL cultures, and the autoinduction protocol, which is freely scalable.

### Autoinduction using the CopyControl Fosmid Autoinduction Solution

*Note:* If the clones are to be grown in a 96-well plate, we suggest using 1.2 mL of culture in a 2 mL deep-well plate. Incubating the plate at a slight angle can improve culture aeration and provide higher DNA yields. Alternatively, it the clones can be grown in as little as 200 µL of culture in a 1 mL 96-well plate.



1. Supplement the appropriate amount of LB medium + 12.5 µg/mL chloramphenicol with the 500X CopyControl Fosmid Autoinduction Solution. Refer to the table below.

| Volume of fresh LB + chloramphenicol (12.5 µg/mL) | Volume of 500X CopyControl Fosmid Autoinduction Solution* | Vessel size recommended for optimum aeration |
|---|---|---|
| 200 µL | 0.4 µL | 1 mL 96-well plate |
| 1 mL | 2 µL | 2 mL 96-well plate |
| 2 mL | 4 µL | 14 mL Falcon tube |
| 50 mL | 100 µL | 250 mL EM flask |

* Mix thoroughly after thawing.

2. Individually inoculate the media with a small portion of the desired CopyControl Fosmid clones grown on an overnight plate.
3. Grow the cultures overnight (16-20 hours) at 37 °C with shaking (225-250 rpm). Cultures incubated for longer or shorter periods of time may not properly induce. Aeration during this incubation is critical!
4. Centrifuge the cells and purify the DNA using the FosmidMAX™ DNA Purification Kit or other standard laboratory methods.[6]

## 5. Appendix

### Appendix A

### Determining the Approximate Number of Clones for a Complete Fosmid Library

Using the following formula,[6] determine the number of fosmid clones required to reasonably ensure that any given DNA sequence is contained within the library.

$$N = \ln(1-P)/\ln(1-f)$$

Where P is the desired probability (expressed as a fraction); f is the proportion of the genome contained in a single clone; and N is the required number of fosmid clones.

For example, the number of clones required to ensure a 99% probability of a given DNA sequence of *E. coli* (genome = 4.7 Mb) being contained within a fosmid library composed of 40 kb inserts is:

$$N = \ln(1-0.99)/\ln(1 - [4 \times 10^4 \text{ bases}/4.7 \times 10^6 \text{ bases}]) = -4.61/-0.01 = 461 \text{ clones}$$

## Appendix B

### Control Fosmid Library production

The Fosmid Control DNA provided in the kit can be used to familiarise yourself with all the steps involved in producing a CopyControl Fosmid Library. We recommend that new CopyControl Fosmid Kit users perform the control ligation and packaging steps to familiarise themselves with the protocol.

The Fosmid Control DNA, as provided in the kit, is purified, blunt-ended, and ready for ligation to the Cloning-Ready pCC1FOS or pCC2FOS Vector. If desired, the Control DNA can be put through the end-repair and gel purification steps (Parts B, C, D) of the CopyControl Fosmid Library Production procedure.

1. Prepare EPI300-T1$^R$ host cells as described in Part F, Step 1.
2. Ligate the Fosmid Control DNA to the CopyControl pCC1FOS or pCC2FOS Vector. Combine the following reagents in the order listed and mix after each addition.

| | |
|---|---|
| 3.5 | µL sterile water |
| 1 | µL 10X Fast-Link Ligation Buffer |
| 1 | µL 10 mM ATP |
| 1 | µL CopyControl pCC1FOS or pCC2FOS Vector (0.5 µg/µL) |
| 2.5 | µL Fosmid Control DNA (100 ng/µL) (See general considerations) |
| 1 | µL Fast-Link DNA Ligase |
| 10 | µL Total reaction volume |

3. Incubate at room temperature for 4 hours.
4. Transfer the reaction to 70 °C for 10 minutes to inactivate the Fast-Link DNA Ligase.
5. Package the entire 10 µL ligation reaction as directed in Part F, Steps 2-9.
6. Titer the packaged control clones by making a 1:1000 dilution of the packaged phage extract in Phage Dilution Buffer. Add 10 µL of the diluted packaged phage to 100 µL of EPI300-T1$^R$ host cells. Incubate the tube at 37 °C for 1 hour.
7. Spread the infected EPI300-T1R cells on LB medium + 12.5 µg/mL chloramphenicol. Incubate the plate overnight at 37 °C to select for the control CopyControl Fosmid clones.
8. Count the colonies and determine the titer, cfu/mL of the reaction (refer to Part G, Step 4). You should expect a titer of >1 × $10^7$ cfu/mL; this corresponds to a packaging efficiency of >$10^7$ cfu/µg of the Fosmid Control DNA.
9. The single-copy CopyControl Fosmid clones produced can be induced to high-copy number by following the procedure on page 16.

### Appendix C

### Testing the efficiency of the MaxPlax Packaging Extracts

Kit components used in this step: **Ligated Lambda Control DNA, MaxPlax Lambda Packaging Extracts, LE392MP Plating Strain.**

**Additionally required:**

- **LB Plates without antibiotic**
- **LB Top Agar** (LB broth containing 0.7% [w/v] Bacto-agar supplemented with 10 mM $MgSO_4$)
- **Phage Dilution Buffer** (10 mM Tris-HCl [pH 8.3], 100 mM NaCl, and 10 mM $MgCl_2$) Please see the product literature for the MaxPlax Lambda Packaging Extracts, that was included with the CopyControl Fosmid Library Production Kits, for details on how to test the efficiency of the extracts.

### Appendix D

### Amplification and storage of the Fosmid Library

**Short-Term Storage:** After dilution of the packaging reaction and addition of chloroform, the packaged fosmid library can be stored at 4 °C for several days. For longer-term storage, see recommendations below.

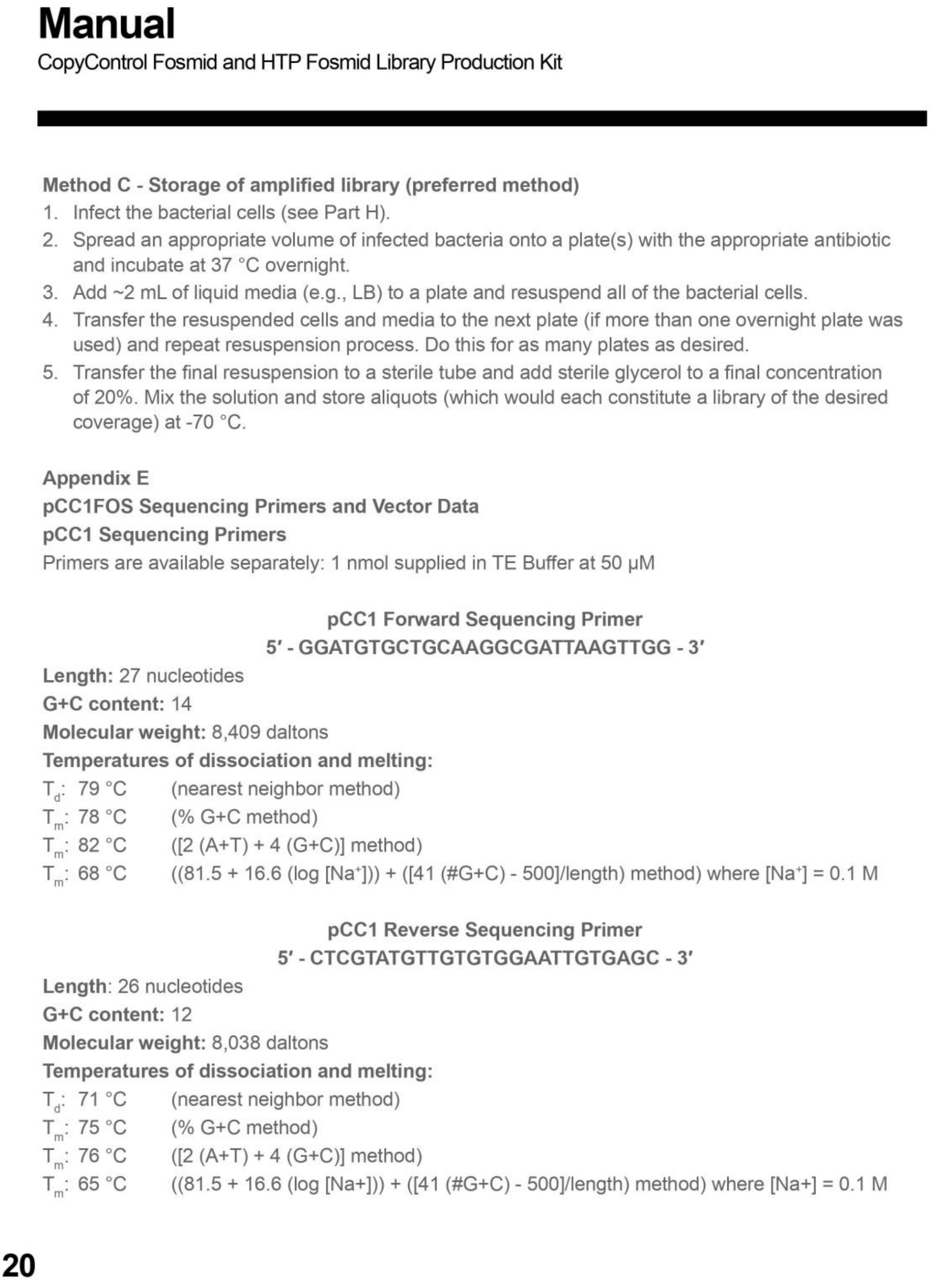
**Long-Term Storage:** For longer-term storage, we recommend storage of the packaged DNA as a primary library, or storage of the library in the EPI300-T1$^R$ Phage T1-resistant *E. coli* plating strain using one of the methods described below.

**Method A - Storage of Packaged DNA**

1. To the packaged fosmid library, add sterile glycerol to a final concentration of 20%, mix, and store at -70 °C.

**Method B - Storage of Infected Cells**

1. Infect the bacterial cells (see Part H).
2. Based on the expected titer, resuspend the cells in an appropriate volume of liquid media.
3. Transfer the final resuspension to a sterile tube and add sterile glycerol to a final concentration of 20%. Mix the solution and store aliquots (which would each constitute a library of the desired coverage) at -70 °C.



**Method C - Storage of amplified library (preferred method)**

1. Infect the bacterial cells (see Part H).
2. Spread an appropriate volume of infected bacteria onto a plate(s) with the appropriate antibiotic and incubate at 37 °C overnight.
3. Add ~2 mL of liquid media (e.g., LB) to a plate and resuspend all of the bacterial cells.
4. Transfer the resuspended cells and media to the next plate (if more than one overnight plate was used) and repeat resuspension process. Do this for as many plates as desired.
5. Transfer the final resuspension to a sterile tube and add sterile glycerol to a final concentration of 20%. Mix the solution and store aliquots (which would each constitute a library of the desired coverage) at -70 °C.

**Appendix E**

**pCC1FOS Sequencing Primers and Vector Data**

**pCC1 Sequencing Primers**

Primers are available separately: 1 nmol supplied in TE Buffer at 50 µM

**pCC1 Forward Sequencing Primer**

**5′ - GGATGTGCTGCAAGGCGATTAAGTTGG - 3′**

**Length:** 27 nucleotides

**G+C content:** 14

**Molecular weight:** 8,409 daltons

**Temperatures of dissociation and melting:**

| | |
|---|---|
| $T_d$: 79 °C | (nearest neighbor method) |
| $T_m$: 78 °C | (% G+C method) |
| $T_m$: 82 °C | ([2 (A+T) + 4 (G+C)] method) |
| $T_m$: 68 °C | ((81.5 + 16.6 (log [Na$^+$])) + ([41 (#G+C) - 500]/length) method) where [Na$^+$] = 0.1 M |

**pCC1 Reverse Sequencing Primer**

**5′ - CTCGTATGTTGTGTGGAATTGTGAGC - 3′**

**Length**: 26 nucleotides

**G+C content:** 12

**Molecular weight:** 8,038 daltons

**Temperatures of dissociation and melting:**

| | |
|---|---|
| $T_d$: 71 °C | (nearest neighbor method) |
| $T_m$: 75 °C | (% G+C method) |
| $T_m$: 76 °C | ([2 (A+T) + 4 (G+C)] method) |
| $T_m$: 65 °C | ((81.5 + 16.6 (log [Na+])) + ([41 (#G+C) - 500]/length) method) where [Na+] = 0.1 M |

*Note:* The sequence of the pCC1 Forward and Reverse Primers do not function well as IRD800-labeled sequencing primers. We recommend using the T7 and pCC1 Primers instead of the pCC1 Forward and Reverse Primers respectively, for this purpose.

**pCC1 RP-2 Reverse Sequencing Primer**
**5′ - TACGCCAAGCTATTTAGGTGAGA - 3′**

### Orientation for Fosmid End-Sequencing

The following is the nucleotide sequence of pCC1FOS (bases 230-501) from the pCC1/ Forward Sequencing Primer (230-256) to the pCC1 Reverse Sequencing Primer (501-476) encompassing the T7 RNA polymerase promoter (311-330) and the Eco72 I site (359-364).

```
230  GGATGTGCTG   CAAGGCGATT   AAGTTGGGTA    ACGCCAGGGT   TTTCCCAGTC
280  ACGACGTTGT   AAAACGACGG   CCAGTGAATT    GTAATACGAC   TCACTATAGG
330  GCGAATTCGA   GCTCGGTACC   CGGGGATCCC    AC - - Cloned insert -
- - - - - - - - - - - - - - - Cloned insert - - - GTGGGATC     CTCTAGAGTC
380  GACCTGCAGG   CATGCAAGCT   TGAGTATTCT    ATAGTCTCAC   CTAAATAGCT
430  TGGCGTAATC   ATGGTCATAG   CTGTTTCCTG    TGTGAAATTG   TTATCCGCTC
480  ACAATTCCAC   ACAACATACG   AG
```

### Appendix F

### pCC2FOS Sequencing Primers and Vector Data

### pCC2 Sequencing Primers

Primers are available separately: 1 nmol supplied in TE Buffer at 50 µM

**pCC2FOS Forward Sequencing Primer**
**5′ - GTACAACGACACCTAGAC - 3′**

**Length**: 18 nucleotides
**G+C content:** 9
**Molecular weight:** 5,462 daltons
**Temperatures of dissociation and melting:**

$T_d$: 48 °C (nearest neighbor method)
$T_m$: 64 °C (% G+C method)
$T_m$: 54 °C ([2 (A+T) + 4 (G+C)] method)
$T_m$: 58 °C ((81.5 + 16.6 (log [Na$^+$])) + ([41 (#G+C) - 500] / length) method) where [Na$^+$] = 0.1 M

### pCC2FOS Reverse Sequencing Primer
**5′ - CAGGAAACAGCCTAGGAA - 3′**

**Length**: 18 nucleotides

**G+C content:** 9

**Molecular weight:** 5,551 daltons

**Temperatures of dissociation and melting:**

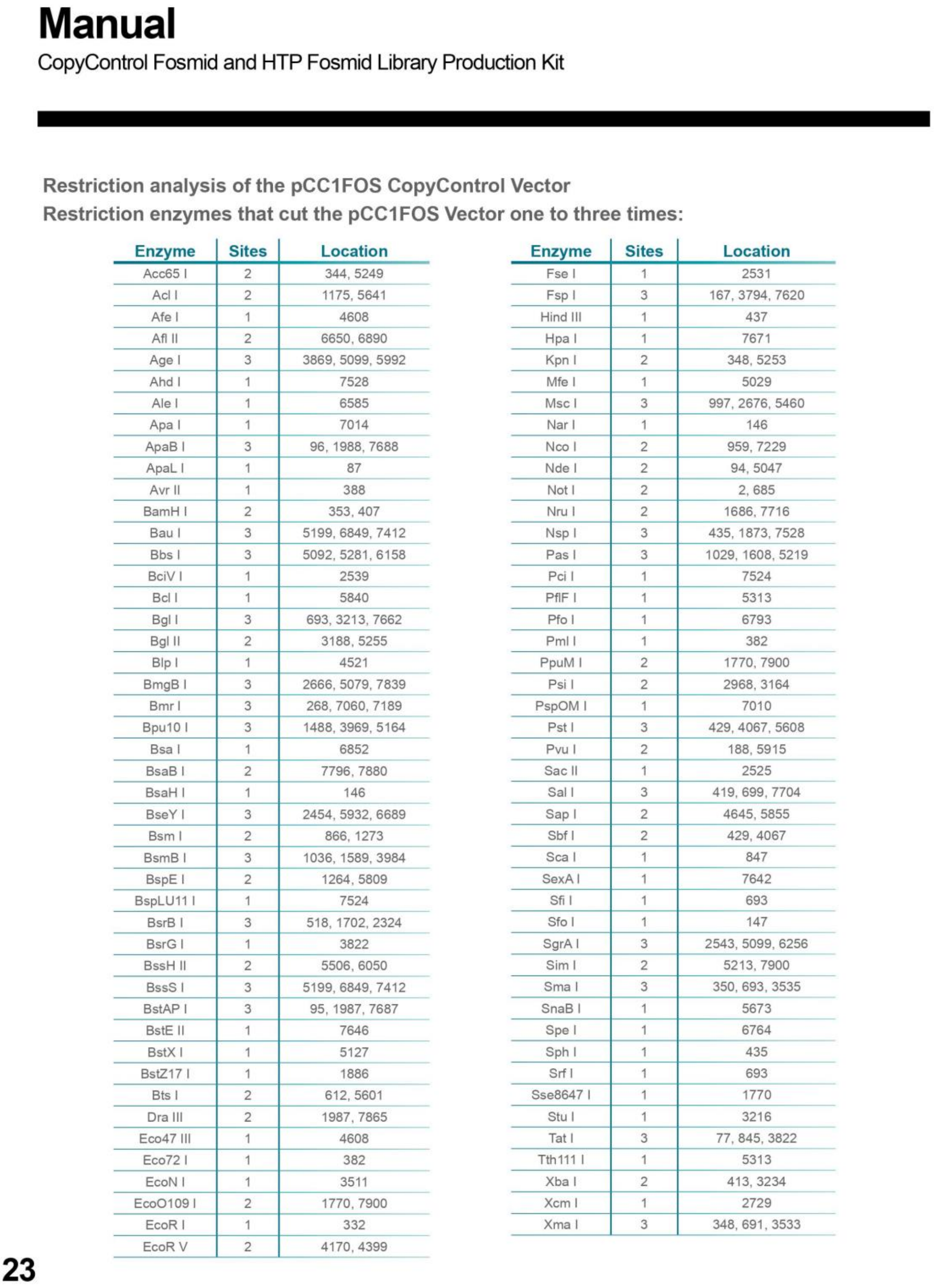
$T_d$: 57 °C (nearest neighbor method)
$T_m$: 64 °C (% G+C method)
$T_m$: 54 °C ([2 (A+T) + 4 (G+C)] method)
$T_m$: 58 °C ((81.5 + 16.6 (log $[Na^+]$)) + ([41 (#G+C) - 500]/length) method) where $[Na^+]$ = 0.1 M

**Orientation for Fosmid End-Sequencing**

The following is the nucleotide sequence of pCC2FOS (bases 360-409) from the pCC2FOS Forward Sequencing Primer (362-379) to the pCC2FOS Reverse Sequencing Primer (403-386) encompassing the *Eco*72 I site (380-385).

```
360 ACGTACAACG    ACACCTAGAC    CAC - - Cloned insert - - GTGTTCC
390 TAGGCTGTTT    CCTGGTGGGA
```

The pCC2FOS sequence is available at lucigen.com/sequences.

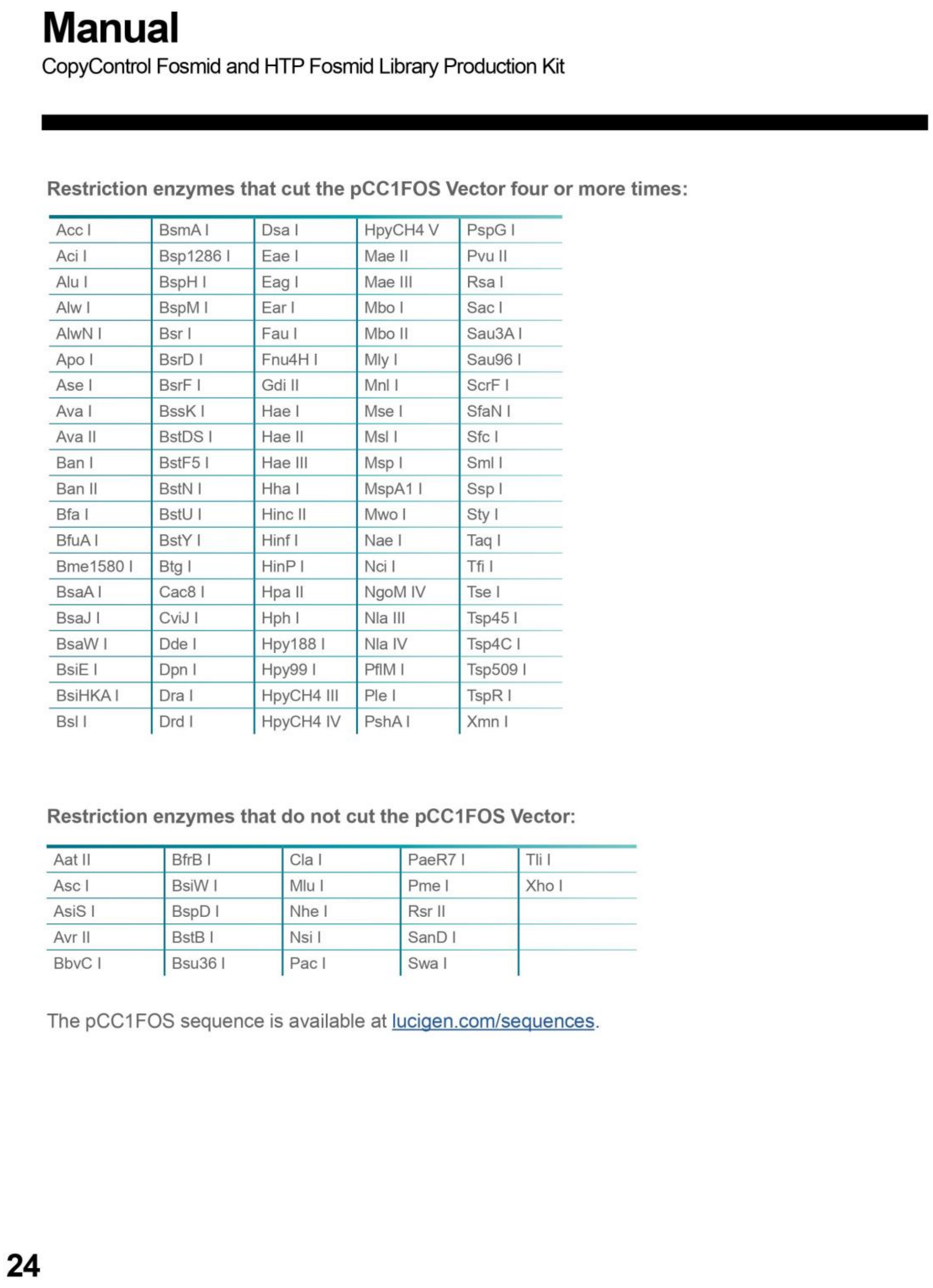


## Restriction analysis of the pCC1FOS CopyControl Vector

## Restriction enzymes that cut the pCC1FOS Vector one to three times:

| Enzyme | Sites | Location |
|---|---|---|
| Acc65 I | 2 | 344, 5249 |
| Acl I | 2 | 1175, 5641 |
| Afe I | 1 | 4608 |
| Afl II | 2 | 6650, 6890 |
| Age I | 3 | 3869, 5099, 5992 |
| Ahd I | 1 | 7528 |
| Ale I | 1 | 6585 |
| Apa I | 1 | 7014 |
| ApaB I | 3 | 96, 1988, 7688 |
| ApaL I | 1 | 87 |
| Avr II | 1 | 388 |
| BamH I | 2 | 353, 407 |
| Bau I | 3 | 5199, 6849, 7412 |
| Bbs I | 3 | 5092, 5281, 6158 |
| BciV I | 1 | 2539 |
| Bcl I | 1 | 5840 |
| Bgl I | 3 | 693, 3213, 7662 |
| Bgl II | 2 | 3188, 5255 |
| Blp I | 1 | 4521 |
| BmgB I | 3 | 2666, 5079, 7839 |
| Bmr I | 3 | 268, 7060, 7189 |
| Bpu10 I | 3 | 1488, 3969, 5164 |
| Bsa I | 1 | 6852 |
| BsaB I | 2 | 7796, 7880 |
| BsaH I | 1 | 146 |
| BseY I | 3 | 2454, 5932, 6689 |
| Bsm I | 2 | 866, 1273 |
| BsmB I | 3 | 1036, 1589, 3984 |
| BspE I | 2 | 1264, 5809 |
| BspLU11 I | 1 | 7524 |
| BsrB I | 3 | 518, 1702, 2324 |
| BsrG I | 1 | 3822 |
| BssH II | 2 | 5506, 6050 |
| BssS I | 3 | 5199, 6849, 7412 |
| BstAP I | 3 | 95, 1987, 7687 |
| BstE II | 1 | 7646 |
| BstX I | 1 | 5127 |
| BstZ17 I | 1 | 1886 |
| Bts I | 2 | 612, 5601 |
| Dra III | 2 | 1987, 7865 |
| Eco47 III | 1 | 4608 |
| Eco72 I | 1 | 382 |
| EcoN I | 1 | 3511 |
| EcoO109 I | 2 | 1770, 7900 |
| EcoR I | 1 | 332 |
| EcoR V | 2 | 4170, 4399 |

| Enzyme | Sites | Location |
|---|---|---|
| Fse I | 1 | 2531 |
| Fsp I | 3 | 167, 3794, 7620 |
| Hind III | 1 | 437 |
| Hpa I | 1 | 7671 |
| Kpn I | 2 | 348, 5253 |
| Mfe I | 1 | 5029 |
| Msc I | 3 | 997, 2676, 5460 |
| Nar I | 1 | 146 |
| Nco I | 2 | 959, 7229 |
| Nde I | 2 | 94, 5047 |
| Not I | 2 | 2, 685 |
| Nru I | 2 | 1686, 7716 |
| Nsp I | 3 | 435, 1873, 7528 |
| Pas I | 3 | 1029, 1608, 5219 |
| Pci I | 1 | 7524 |
| PflF I | 1 | 5313 |
| Pfo I | 1 | 6793 |
| Pml I | 1 | 382 |
| PpuM I | 2 | 1770, 7900 |
| Psi I | 2 | 2968, 3164 |
| PspOM I | 1 | 7010 |
| Pst I | 3 | 429, 4067, 5608 |
| Pvu I | 2 | 188, 5915 |
| Sac II | 1 | 2525 |
| Sal I | 3 | 419, 699, 7704 |
| Sap I | 2 | 4645, 5855 |
| Sbf I | 2 | 429, 4067 |
| Sca I | 1 | 847 |
| SexA I | 1 | 7642 |
| Sfi I | 1 | 693 |
| Sfo I | 1 | 147 |
| SgrA I | 3 | 2543, 5099, 6256 |
| Sim I | 2 | 5213, 7900 |
| Sma I | 3 | 350, 693, 3535 |
| SnaB I | 1 | 5673 |
| Spe I | 1 | 6764 |
| Sph I | 1 | 435 |
| Srf I | 1 | 693 |
| Sse8647 I | 1 | 1770 |
| Stu I | 1 | 3216 |
| Tat I | 3 | 77, 845, 3822 |
| Tth111 I | 1 | 5313 |
| Xba I | 2 | 413, 3234 |
| Xcm I | 1 | 2729 |
| Xma I | 3 | 348, 691, 3533 |



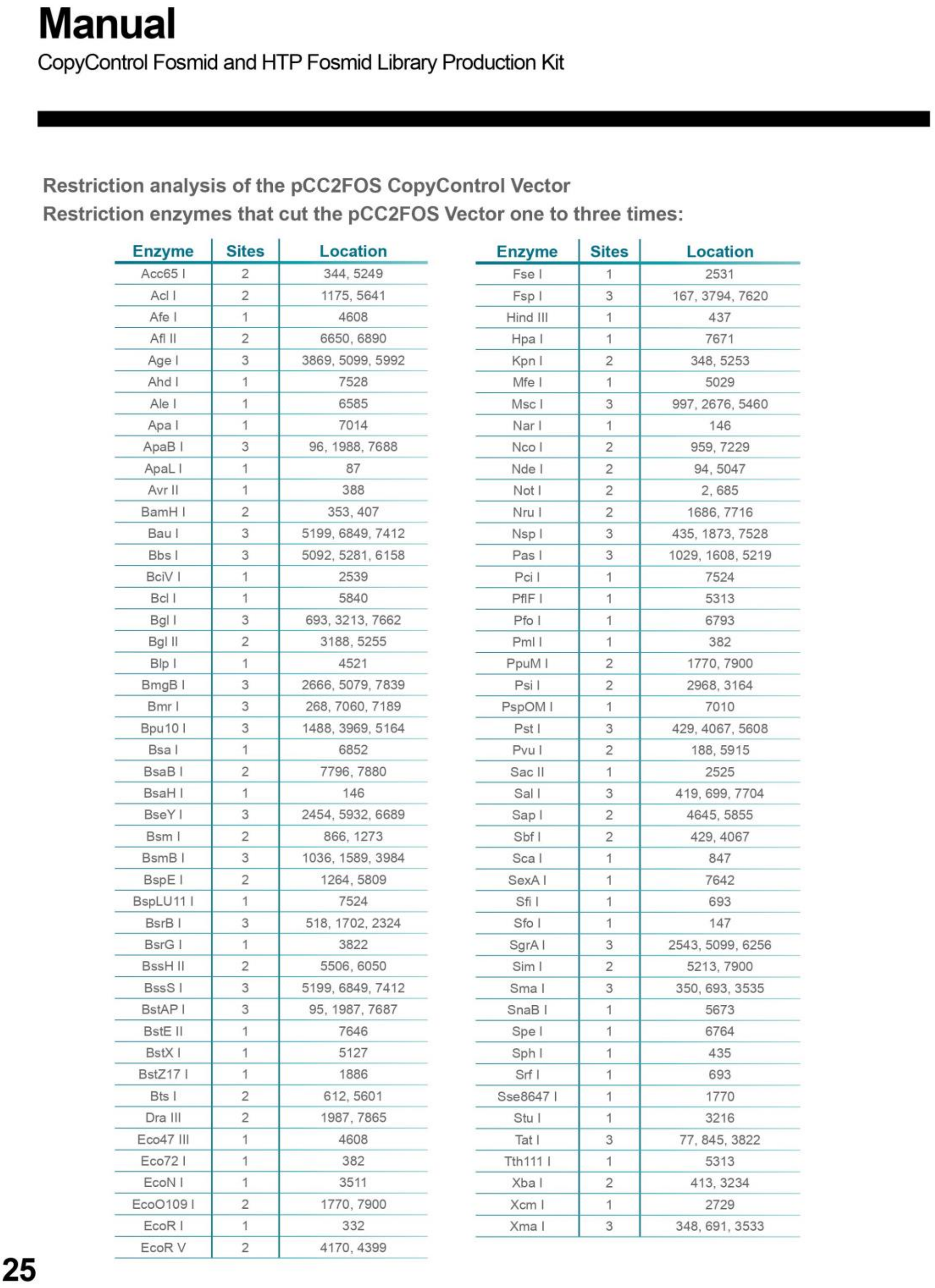
**Restriction enzymes that cut the pCC1FOS Vector four or more times:**

| | | | | |
|---|---|---|---|---|
| Acc I | BsmA I | Dsa I | HpyCH4 V | PspG I |
| Aci I | Bsp1286 I | Eae I | Mae II | Pvu II |
| Alu I | BspH I | Eag I | Mae III | Rsa I |
| Alw I | BspM I | Ear I | Mbo I | Sac I |
| AlwN I | Bsr I | Fau I | Mbo II | Sau3A I |
| Apo I | BsrD I | Fnu4H I | Mly I | Sau96 I |
| Ase I | BsrF I | Gdi II | Mnl I | ScrF I |
| Ava I | BssK I | Hae I | Mse I | SfaN I |
| Ava II | BstDS I | Hae II | Msl I | Sfc I |
| Ban I | BstF5 I | Hae III | Msp I | Sml I |
| Ban II | BstN I | Hha I | MspA1 I | Ssp I |
| Bfa I | BstU I | Hinc II | Mwo I | Sty I |
| BfuA I | BstY I | Hinf I | Nae I | Taq I |
| Bme1580 I | Btg I | HinP I | Nci I | Tfi I |
| BsaA I | Cac8 I | Hpa II | NgoM IV | Tse I |
| BsaJ I | CviJ I | Hph I | Nla III | Tsp45 I |
| BsaW I | Dde I | Hpy188 I | Nla IV | Tsp4C I |
| BsiE I | Dpn I | Hpy99 I | PflM I | Tsp509 I |
| BsiHKA I | Dra I | HpyCH4 III | Ple I | TspR I |
| Bsl I | Drd I | HpyCH4 IV | PshA I | Xmn I |

**Restriction enzymes that do not cut the pCC1FOS Vector:**

| | | | | |
|---|---|---|---|---|
| Aat II | BfrB I | Cla I | PaeR7 I | Tli I |
| Asc I | BsiW I | Mlu I | Pme I | Xho I |
| AsiS I | BspD I | Nhe I | Rsr II | |
| Avr II | BstB I | Nsi I | SanD I | |
| BbvC I | Bsu36 I | Pac I | Swa I | |

The pCC1FOS sequence is available at lucigen.com/sequences.



**Restriction analysis of the pCC2FOS CopyControl Vector**

**Restriction enzymes that cut the pCC2FOS Vector one to three times:**

| Enzyme | Sites | Location |
|---|---|---|
| Acc65 I | 2 | 344, 5249 |
| Acl I | 2 | 1175, 5641 |
| Afe I | 1 | 4608 |
| Afl II | 2 | 6650, 6890 |
| Age I | 3 | 3869, 5099, 5992 |
| Ahd I | 1 | 7528 |
| Ale I | 1 | 6585 |
| Apa I | 1 | 7014 |
| ApaB I | 3 | 96, 1988, 7688 |
| ApaL I | 1 | 87 |
| Avr II | 1 | 388 |
| BamH I | 2 | 353, 407 |
| Bau I | 3 | 5199, 6849, 7412 |
| Bbs I | 3 | 5092, 5281, 6158 |
| BciV I | 1 | 2539 |
| Bcl I | 1 | 5840 |
| Bgl I | 3 | 693, 3213, 7662 |
| Bgl II | 2 | 3188, 5255 |
| Blp I | 1 | 4521 |
| BmgB I | 3 | 2666, 5079, 7839 |
| Bmr I | 3 | 268, 7060, 7189 |
| Bpu10 I | 3 | 1488, 3969, 5164 |
| Bsa I | 1 | 6852 |
| BsaB I | 2 | 7796, 7880 |
| BsaH I | 1 | 146 |
| BseY I | 3 | 2454, 5932, 6689 |
| Bsm I | 2 | 866, 1273 |
| BsmB I | 3 | 1036, 1589, 3984 |
| BspE I | 2 | 1264, 5809 |
| BspLU11 I | 1 | 7524 |
| BsrB I | 3 | 518, 1702, 2324 |
| BsrG I | 1 | 3822 |
| BssH II | 2 | 5506, 6050 |
| BssS I | 3 | 5199, 6849, 7412 |
| BstAP I | 3 | 95, 1987, 7687 |
| BstE II | 1 | 7646 |
| BstX I | 1 | 5127 |
| BstZ17 I | 1 | 1886 |
| Bts I | 2 | 612, 5601 |
| Dra III | 2 | 1987, 7865 |
| Eco47 III | 1 | 4608 |
| Eco72 I | 1 | 382 |
| EcoN I | 1 | 3511 |
| EcoO109 I | 2 | 1770, 7900 |
| EcoR I | 1 | 332 |
| EcoR V | 2 | 4170, 4399 |

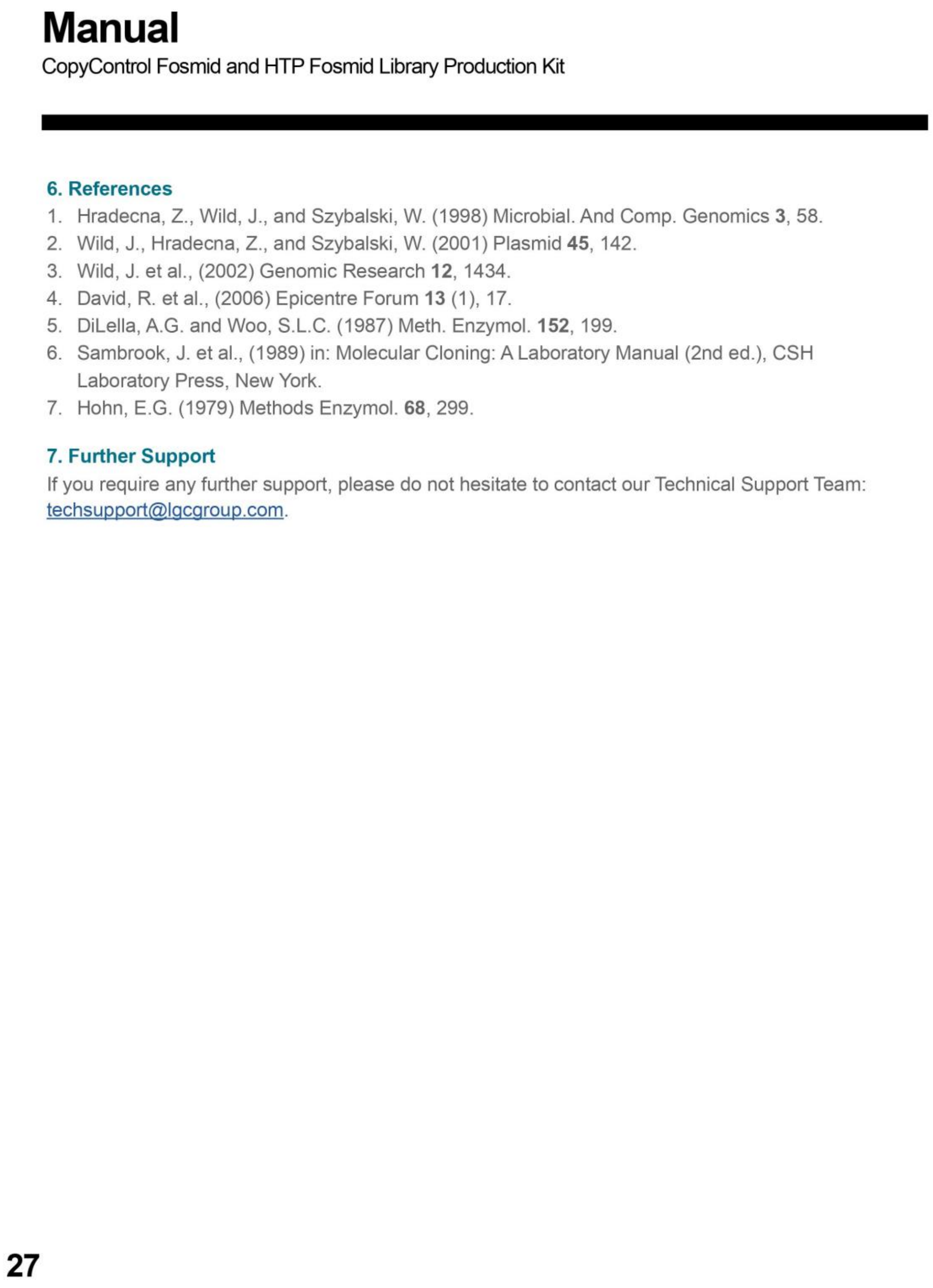
| Enzyme | Sites | Location |
|---|---|---|
| Fse I | 1 | 2531 |
| Fsp I | 3 | 167, 3794, 7620 |
| Hind III | 1 | 437 |
| Hpa I | 1 | 7671 |
| Kpn I | 2 | 348, 5253 |
| Mfe I | 1 | 5029 |
| Msc I | 3 | 997, 2676, 5460 |
| Nar I | 1 | 146 |
| Nco I | 2 | 959, 7229 |
| Nde I | 2 | 94, 5047 |
| Not I | 2 | 2, 685 |
| Nru I | 2 | 1686, 7716 |
| Nsp I | 3 | 435, 1873, 7528 |
| Pas I | 3 | 1029, 1608, 5219 |
| Pci I | 1 | 7524 |
| PflF I | 1 | 5313 |
| Pfo I | 1 | 6793 |
| Pml I | 1 | 382 |
| PpuM I | 2 | 1770, 7900 |
| Psi I | 2 | 2968, 3164 |
| PspOM I | 1 | 7010 |
| Pst I | 3 | 429, 4067, 5608 |
| Pvu I | 2 | 188, 5915 |
| Sac II | 1 | 2525 |
| Sal I | 3 | 419, 699, 7704 |
| Sap I | 2 | 4645, 5855 |
| Sbf I | 2 | 429, 4067 |
| Sca I | 1 | 847 |
| SexA I | 1 | 7642 |
| Sfi I | 1 | 693 |
| Sfo I | 1 | 147 |
| SgrA I | 3 | 2543, 5099, 6256 |
| Sim I | 2 | 5213, 7900 |
| Sma I | 3 | 350, 693, 3535 |
| SnaB I | 1 | 5673 |
| Spe I | 1 | 6764 |
| Sph I | 1 | 435 |
| Srf I | 1 | 693 |
| Sse8647 I | 1 | 1770 |
| Stu I | 1 | 3216 |
| Tat I | 3 | 77, 845, 3822 |
| Tth111 I | 1 | 5313 |
| Xba I | 2 | 413, 3234 |
| Xcm I | 1 | 2729 |
| Xma I | 3 | 348, 691, 3533 |

### Restriction enzymes that cut the pCC2FOS Vector four or more times:

| | | | | |
|---|---|---|---|---|
| Acc I | Bsl I | Drd I | HpyCH4 III | PshA I |
| Aci I | BsmA I | Dsa I | HpyCH4 IV | PspG I |
| Afl III | Bsp1286 I | Eae I | HpyCH4 V | Pvu II |
| Alu I | BspH I | Eag I | Mae II | Rsa I |
| Alw I | BspM I | Ear I | Mae III | Sac I |
| AlwN I | Bsr I | Fat I | Mbo I | Sau3A I |
| Apo I | BsrD I | Fau I | Mbo II | Sau96 I |
| Ase I | BsrF I | Fnu4H I | Mly I | ScrF I |
| Ava I | BssK I | Gdi II | Mnl I | SfaN I |
| Ava II | BstDS I | Hae I | Mse I | Sfc I |
| Ban I | BstF5 I | Hae II | Msl I | Sml I |
| Ban II | BstN I | Hae III | Msp I | Ssp I |
| Bcc I | BstU I | Hha I | MspA1 I | Sty I |
| Bfa I | BstY I | Hinc II | Mwo I | Taq I |
| BfuA I | Btg I | Hinf I | Nae I | Tfi I |
| Bme1580 I | Cac8 I | HinP I | Nci I | Tse I |
| BsaA I | Cvi II | Hpa II | NgoM IV | Tsp45 I |
| BsaJ I | CviJ I | Hph I | Nla III | Tsp4C I |
| BsaW I | Dde I | Hpy188 I | Nla IV | Tsp509 I |
| BsiE I | Dpn I | Hpy188 III | PflM I | TspR I |
| BsiHKA I | Dra I | Hpy99 I | Ple I | Xmn I |

### Restriction enzymes that do not cut the pCC1FOS Vector:

| | | | | |
|---|---|---|---|---|
| Aat II | Bmt I | Cla I | PaeR7 I | Swa I |
| Asc I | BsiW I | Mlu I | Pme I | Tli I |
| AsiS I | BspD I | Nhe I | PspX I | Xho I |
| BbvC I | BstB I | Nsi I | Rsr II | Zra I |
| BfrB I | Bsu36 I | Pac I | SanD I | |



## 6. References

1. Hradecna, Z., Wild, J., and Szybalski, W. (1998) Microbial. And Comp. Genomics **3**, 58.
2. Wild, J., Hradecna, Z., and Szybalski, W. (2001) Plasmid **45**, 142.
3. Wild, J. et al., (2002) Genomic Research **12**, 1434.
4. David, R. et al., (2006) Epicentre Forum **13** (1), 17.
5. DiLella, A.G. and Woo, S.L.C. (1987) Meth. Enzymol. **152**, 199.
6. Sambrook, J. et al., (1989) in: Molecular Cloning: A Laboratory Manual (2nd ed.), CSH Laboratory Press, New York.
7. Hohn, E.G. (1979) Methods Enzymol. **68**, 299.

## 7. Further Support

If you require any further support, please do not hesitate to contact our Technical Support Team: techsupport@lgcgroup.com.

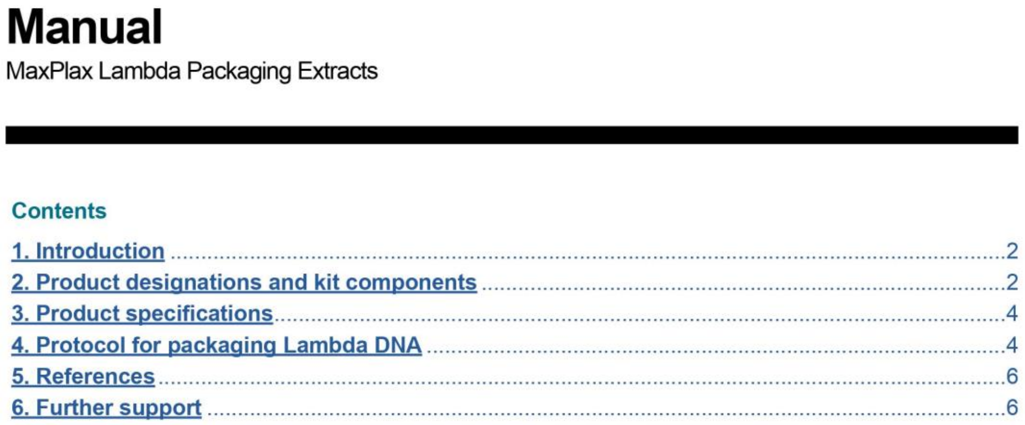
Manual

# MaxPlax Lambda Packaging Extracts

For Research Use Only. Not for use in diagnostic procedures.

MaxPlax™ is part of the Epicentre™ product line, known for its unique genomics kits, enzymes, and reagents, which offer high quality and reliable performance.

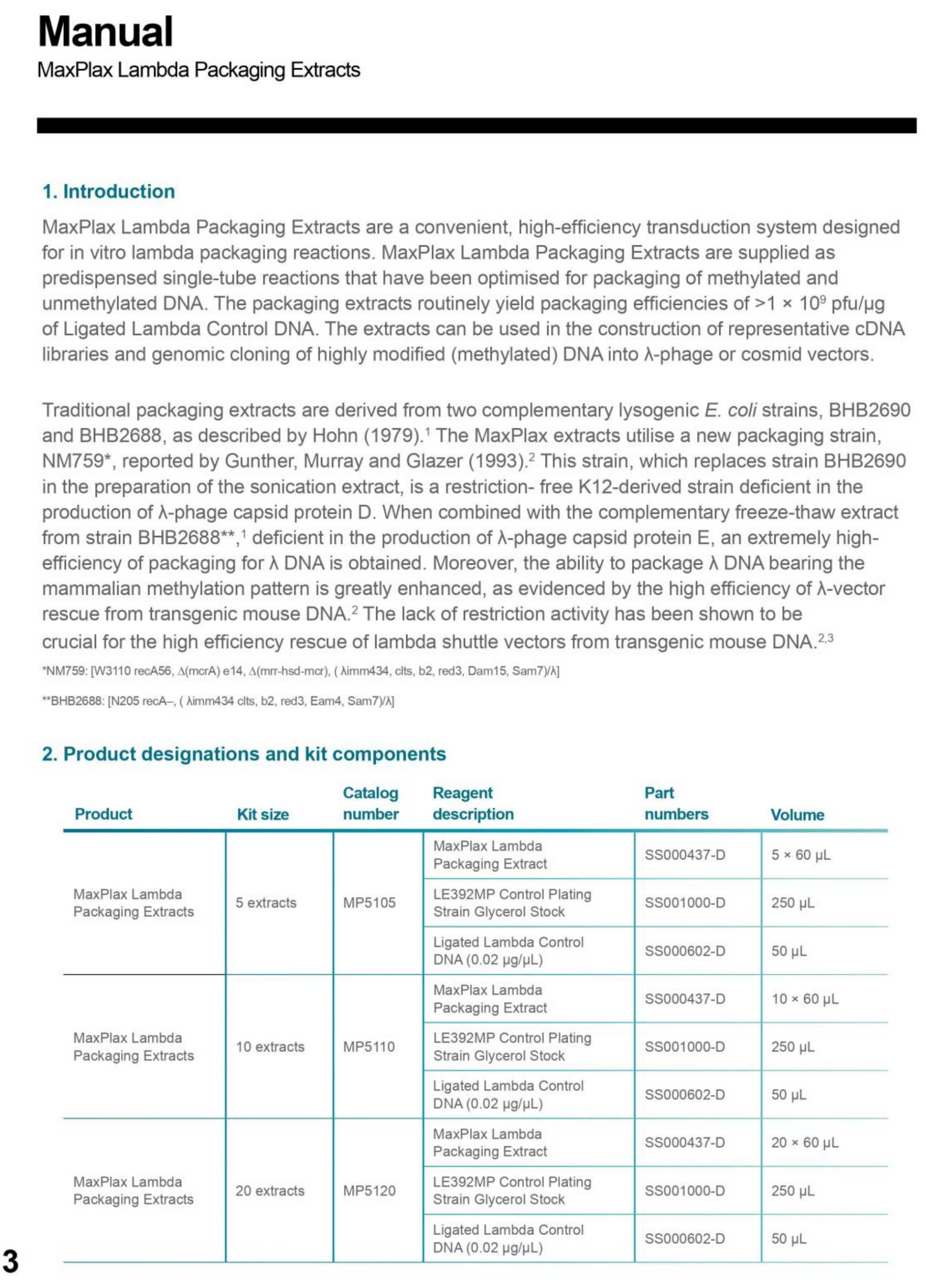
BIOSEARCH™
TECHNOLOGIES
GENOMIC ANALYSIS BY LGC



# Manual

MaxPlax Lambda Packaging Extracts

## Contents



## 1. Introduction

MaxPlax Lambda Packaging Extracts are a convenient, high-efficiency transduction system designed for in vitro lambda packaging reactions. MaxPlax Lambda Packaging Extracts are supplied as predispensed single-tube reactions that have been optimised for packaging of methylated and unmethylated DNA. The packaging extracts routinely yield packaging efficiencies of >1 × $10^9$ pfu/µg of Ligated Lambda Control DNA. The extracts can be used in the construction of representative cDNA libraries and genomic cloning of highly modified (methylated) DNA into λ-phage or cosmid vectors.

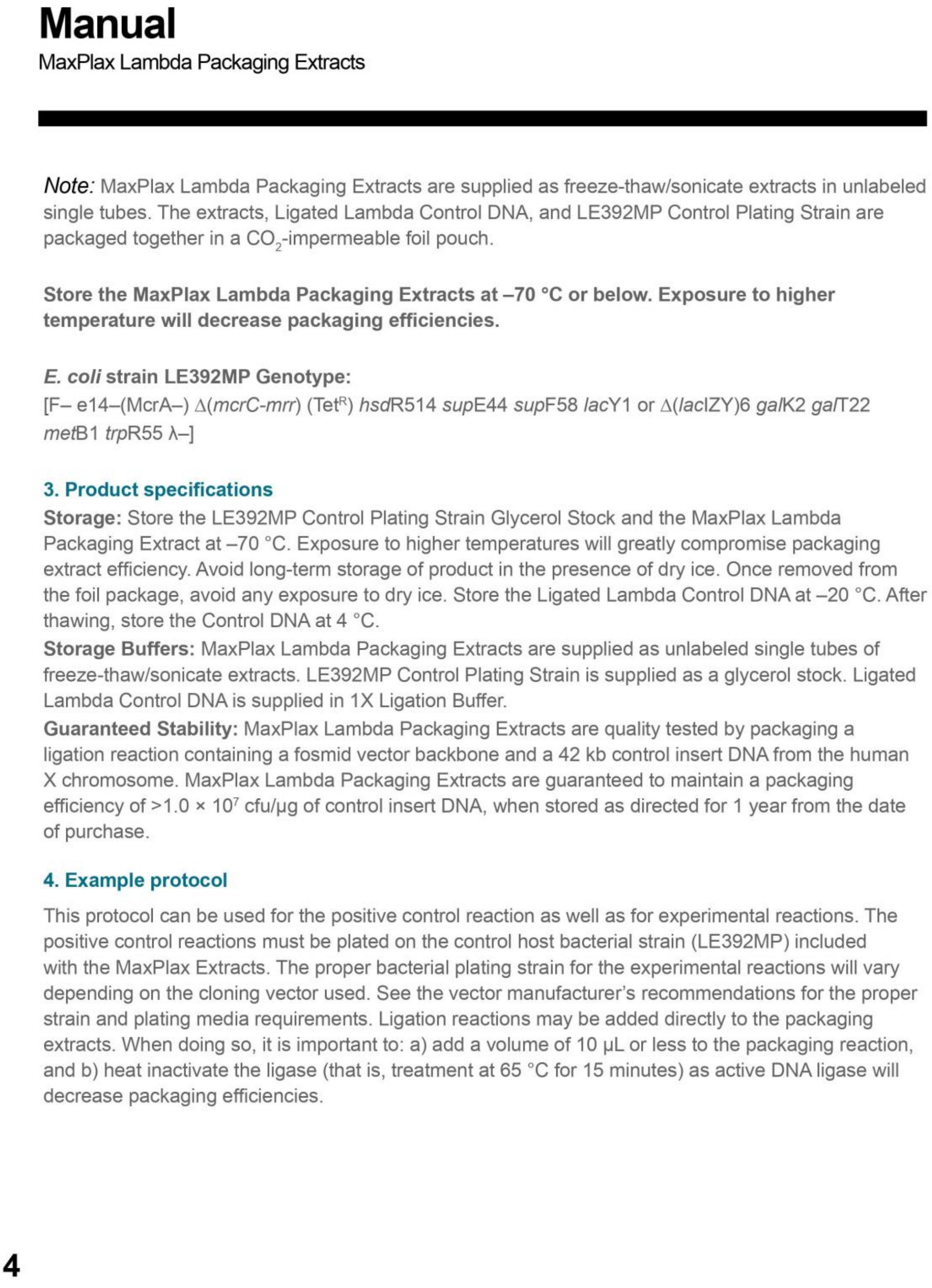
Traditional packaging extracts are derived from two complementary lysogenic *E. coli* strains, BHB2690 and BHB2688, as described by Hohn (1979).[1] The MaxPlax extracts utilise a new packaging strain, NM759*, reported by Gunther, Murray and Glazer (1993).[2] This strain, which replaces strain BHB2690 in the preparation of the sonication extract, is a restriction- free K12-derived strain deficient in the production of λ-phage capsid protein D. When combined with the complementary freeze-thaw extract from strain BHB2688**,[1] deficient in the production of λ-phage capsid protein E, an extremely high-efficiency of packaging for λ DNA is obtained. Moreover, the ability to package λ DNA bearing the mammalian methylation pattern is greatly enhanced, as evidenced by the high efficiency of λ-vector rescue from transgenic mouse DNA.[2] The lack of restriction activity has been shown to be crucial for the high efficiency rescue of lambda shuttle vectors from transgenic mouse DNA.[2,3]

*NM759: [W3110 recA56, Δ(mcrA) e14, Δ(mrr-hsd-mcr), ( λimm434, clts, b2, red3, Dam15, Sam7)/λ]

**BHB2688: [N205 recA–, ( λimm434 clts, b2, red3, Eam4, Sam7)/λ]

## 2. Product designations and kit components

| Product | Kit size | Catalog number | Reagent description | Part numbers | Volume |
|---|---|---|---|---|---|
| MaxPlax Lambda Packaging Extracts | 5 extracts | MP5105 | MaxPlax Lambda Packaging Extract | SS000437-D | 5 × 60 µL |
| | | | LE392MP Control Plating Strain Glycerol Stock | SS001000-D | 250 µL |
| | | | Ligated Lambda Control DNA (0.02 µg/µL) | SS000602-D | 50 µL |
| MaxPlax Lambda Packaging Extracts | 10 extracts | MP5110 | MaxPlax Lambda Packaging Extract | SS000437-D | 10 × 60 µL |
| | | | LE392MP Control Plating Strain Glycerol Stock | SS001000-D | 250 µL |
| | | | Ligated Lambda Control DNA (0.02 µg/µL) | SS000602-D | 50 µL |
| MaxPlax Lambda Packaging Extracts | 20 extracts | MP5120 | MaxPlax Lambda Packaging Extract | SS000437-D | 20 × 60 µL |
| | | | LE392MP Control Plating Strain Glycerol Stock | SS001000-D | 250 µL |
| | | | Ligated Lambda Control DNA (0.02 µg/µL) | SS000602-D | 50 µL |



*Note:* MaxPlax Lambda Packaging Extracts are supplied as freeze-thaw/sonicate extracts in unlabeled single tubes. The extracts, Ligated Lambda Control DNA, and LE392MP Control Plating Strain are packaged together in a $CO_2$-impermeable foil pouch.

**Store the MaxPlax Lambda Packaging Extracts at –70 °C or below. Exposure to higher temperature will decrease packaging efficiencies.**

***E. coli*** **strain LE392MP Genotype:**
[F– e14–(McrA–) Δ(*mcrC-mrr*) (Tet$^R$) *hsd*R514 *sup*E44 *sup*F58 *lac*Y1 or Δ(*lac*IZY)6 *gal*K2 *gal*T22 *met*B1 *trp*R55 λ–]

## 3. Product specifications

**Storage:** Store the LE392MP Control Plating Strain Glycerol Stock and the MaxPlax Lambda Packaging Extract at –70 °C. Exposure to higher temperatures will greatly compromise packaging extract efficiency. Avoid long-term storage of product in the presence of dry ice. Once removed from the foil package, avoid any exposure to dry ice. Store the Ligated Lambda Control DNA at –20 °C. After thawing, store the Control DNA at 4 °C.

**Storage Buffers:** MaxPlax Lambda Packaging Extracts are supplied as unlabeled single tubes of freeze-thaw/sonicate extracts. LE392MP Control Plating Strain is supplied as a glycerol stock. Ligated Lambda Control DNA is supplied in 1X Ligation Buffer.

**Guaranteed Stability:** MaxPlax Lambda Packaging Extracts are quality tested by packaging a ligation reaction containing a fosmid vector backbone and a 42 kb control insert DNA from the human X chromosome. MaxPlax Lambda Packaging Extracts are guaranteed to maintain a packaging efficiency of >1.0 × $10^7$ cfu/µg of control insert DNA, when stored as directed for 1 year from the date of purchase.

## 4. Example protocol

This protocol can be used for the positive control reaction as well as for experimental reactions. The positive control reactions must be plated on the control host bacterial strain (LE392MP) included with the MaxPlax Extracts. The proper bacterial plating strain for the experimental reactions will vary depending on the cloning vector used. See the vector manufacturer's recommendations for the proper strain and plating media requirements. Ligation reactions may be added directly to the packaging extracts. When doing so, it is important to: a) add a volume of 10 µL or less to the packaging reaction, and b) heat inactivate the ligase (that is, treatment at 65 °C for 15 minutes) as active DNA ligase will decrease packaging efficiencies.

### Solutions

**Phage Dilution Buffer**

| | | |
|---|---|---|
| 10 | mM | Tris-HCl (pH 8.3) |
| 100 | mM | NaCl |
| 10 | mM | $MgCl_2$ |

**LB Broth (1 Liter)**

| | | |
|---|---|---|
| 10 | g | Bacto-tryptone |
| 5 | g | Bacto-yeast extract |
| 10 | g | NaCl |

Adjust pH to 7.0 with NaOH

**LB Plates**

LB Broth with 1.5% (w/v) Bacto-agar

**LB Top Agar**

LB Broth with 0.7% (w/v) Bacto-agar

### Plating bacteria preparation:

1. The day before performing the packaging reactions, inoculate 50 mL of supplemented (10 mM $MgSO_4$) LB broth with a single colony of the plating bacterial strain and shake overnight at 37 °C.
2. The day of the packaging reactions, inoculate 50 mL of supplemented (10 mM $MgSO_4$ + 0.2% maltose) LB broth with 5 mL of the overnight culture and shake at 37 °C to an $OD_{600}$ = 0.8-1.0. Store the cells at 4 °C until needed; cells may be stored for up to 72 hours.

### Plating bacteria preparation:

1. Thaw the appropriate number of packaging extracts at room temperature. For every two packaging reactions, thaw one extract then place on ice.
2. When thawed, immediately transfer half (25 µL) of each packaging extract to a second 1.5-mL tube and place on ice.
3. Add the substrate DNA (10 µL [0.2 µg] of the control DNA) to a tube containing 25 µL of extract. If performing an odd number of packaging reactions, the remaining 25 µL of extract can be refrozen at -70 °C.
4. Mix by pipetting several times; avoid the introduction of air bubbles. Return all of the contents to the bottom of the tube by brief centrifugation if necessary.
5. Incubate the reaction(s) at 30 °C for 90 minutes.
6. At the end of this incubation, add the additional 25 µL of thawed extract to each reaction tube at 30 °C (If performing two packaging reactions, thaw another tube of extract and add 25 µL to each tube.) and incubate the reaction(s) for an additional 90 minutes at 30 °C.
7. Add 500 µL of phage dilution buffer and mix by gentle vortexing.
8. Add 25 µL of chloroform and mix by gentle vortexing (store at 4 °C).
9. Assay the packaged phage by titering on the appropriate bacterial strain (LE392MP for the control).

### Titering phage extracts:

1. Make serial dilutions of the packaged phage in phage dilution buffer. Use $10^{-5}$ and $10^{-6}$ dilutions for the control reactions.
   $10^{-2}$ dilution is 10 µL of packaged phage particles into 990 µL of phage dilution buffer; vortex mix.
   $10^{-4}$ dilution is 10 µL of 10-2 dilution into 990 µL phage dilution buffer; vortex mix.
   $10^{-5}$ dilution is 100 µL of $10^{-4}$ dilution into 900 µL phage dilution buffer; vortex mix.
   $10^{-6}$ dilution is 10 µL of $10^{-4}$ dilution into 990 µL phage dilution buffer; vortex mix.
2. Add 100 µL of the appropriate serial dilutions to 100 µL of prepared plating bacteria (use LE392MP for the control reactions) and incubate for 15 minutes at 37 °C.
3. Add 3.0 mL of melted supplemented (10 mM $MgSO_4$) LB top agar (cooled to ~48 °C). Vortex gently and pour onto pre-warmed (37 °C) LB plates. Allow the top agar to solidify and then incubate inverted overnight at 37 °C.
4. Count the plaques and determine the titer (pfu/mL) and packaging efficiency (See sample calculations).

### Sample calculations:

If there were 110 plaques on a 10-6 dilution plate, then the titer, pfu/mL, (where pfu represents plaque forming units) of this reaction would be:

$$\frac{\text{(\# of plaques) (dilution factor) (1000 µL/mL)}}{\text{(volume of phage plated [µL])}} \quad \textbf{OR} \quad \frac{\text{(110 pfu) (}10^6\text{) (1000 µL/mL)}}{\text{(100 µL)}} \quad \mathbf{= 1.1 \times 10^9\ pfu/mL}$$

The packaging efficiency (pfu/µg DNA) of this reaction would be:

$$\frac{\text{(\# of plaques) (dilution factor) (total reaction vol.)}}{\text{(vol. of dilution plated) (amount of DNA packaged)}} \quad \textbf{OR} \quad \frac{\text{(110 pfu) (}10^6\text{) (550 µL)}}{\text{(100 µL) (0.2 µg)}} \quad \mathbf{= 3 \times 10^9\ pfu/µg}$$

## 5. References

1. Hohn, E.G. (1979) *Methods Enzymol.* **68,** 299.
2. Gunther, E.G. *et al.,* (1993) *Nucl. Acids Res.* **21,** 3903.
3. Kohler, S.W. *et al.,* (1990) *Nucl. Acids Res.* **18,** 3007.

## 6. Further support

If you require any further support, please do not hesitate to contact our Technical Support Team: techsupport@lgcgroup.com.

LGC

## Integrated tools.
## Accelerated science.

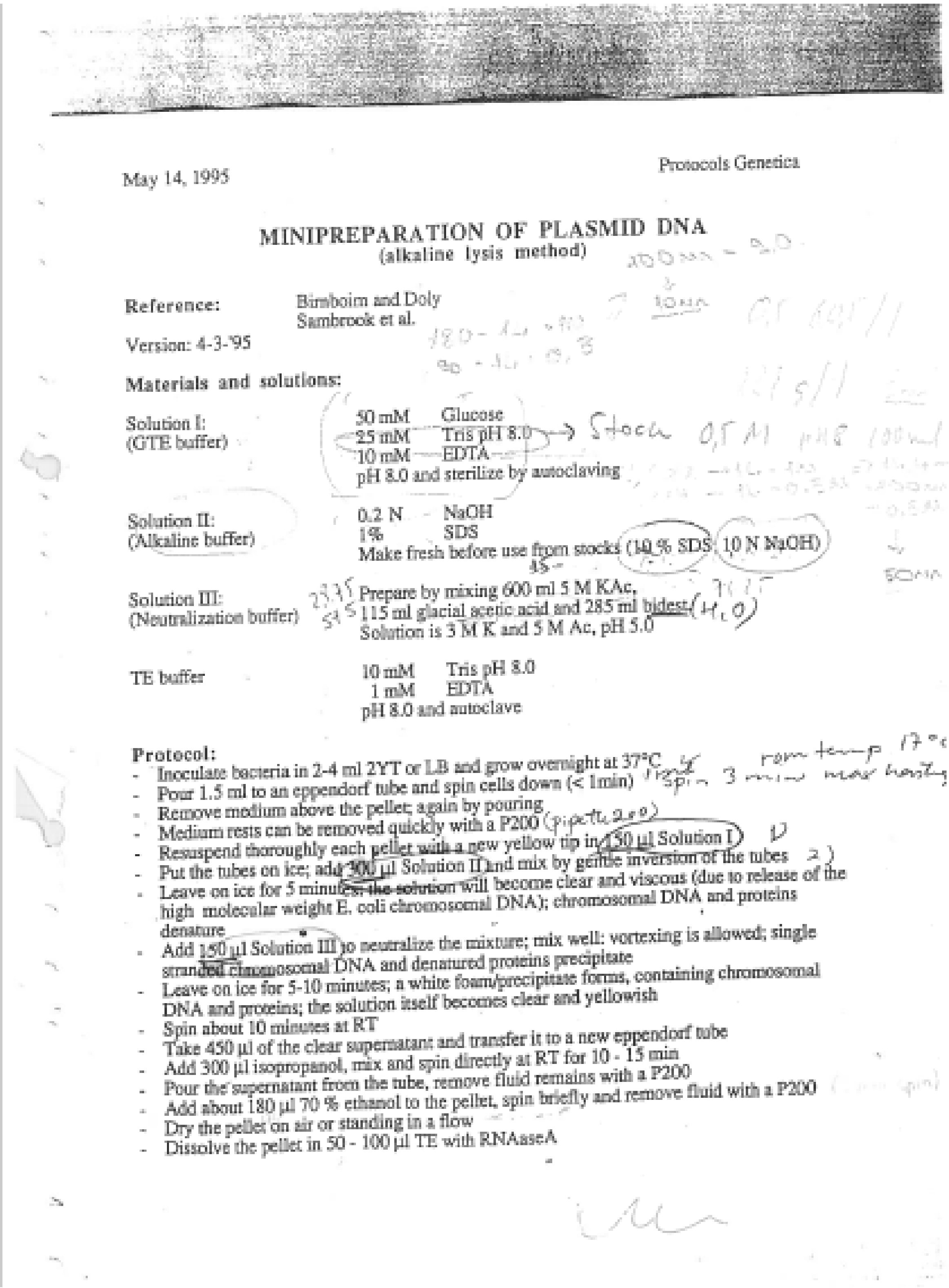
@LGCBiosearch | lucigen.com
biosearchtech.com

For Research Use Only. Not for use in diagnostic procedures.

bsi. ISO 13485 Medical Devices Quality Management
FM 598408

BIOSEARCH™
TECHNOLOGIES
GENOMIC ANALYSIS BY LGC

# Appendix 4: Minipreparation of plasmid DNA (alkaline lysis method)



May 14, 1995

Protocols Genetica

## MINIPREPARATION OF PLASMID DNA

(alkaline lysis method)

Reference: Birnboim and Doly
Sambrook et al.

Version: 4-3-'95

**Materials and solutions:**

Solution I: (GTE buffer)
50 mM Glucose
25 mM Tris pH 8.0 → Stock 0,5 M pH8 100ml
10 mM EDTA
pH 8.0 and sterilize by autoclaving

Solution II: (Alkaline buffer)
0.2 N NaOH
1% SDS
Make fresh before use from stocks (10 % SDS, 10 N NaOH)

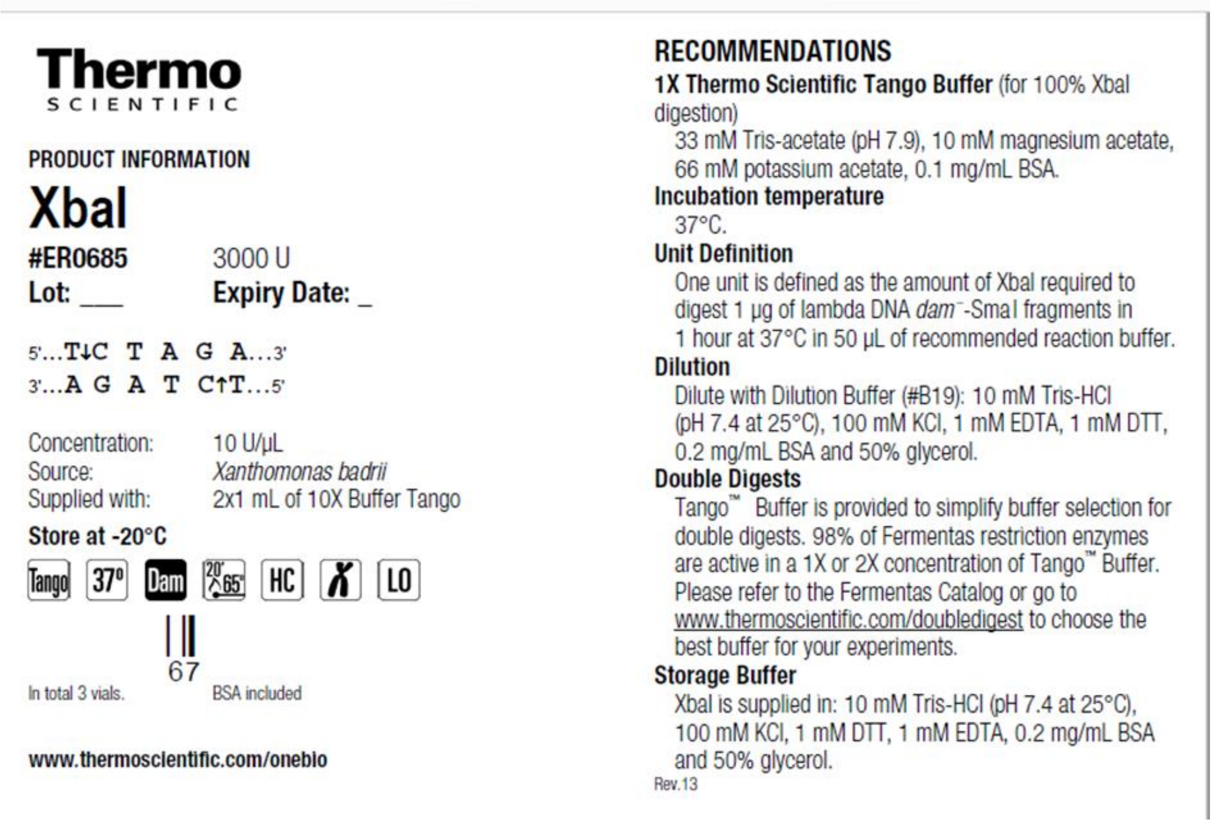
Solution III: (Neutralization buffer)
Prepare by mixing 600 ml 5 M KAc,
115 ml glacial acetic acid and 285 ml bidest (H2O)
Solution is 3 M K and 5 M Ac, pH 5.0

TE buffer
10 mM Tris pH 8.0
1 mM EDTA
pH 8.0 and autoclave

**Protocol:**

- Inoculate bacteria in 2-4 ml 2YT or LB and grow overnight at 37°C room temp 17°c
- Pour 1.5 ml to an eppendorf tube and spin cells down (< 1min) spin 3 min
- Remove medium above the pellet; again by pouring
- Medium rests can be removed quickly with a P200 (pipette 200)
- Resuspend thoroughly each pellet with a new yellow tip in 150 µl Solution I 1)
- Put the tubes on ice; add 300 µl Solution II and mix by gentle inversion of the tubes 2)
- Leave on ice for 5 minutes; ~~the solution~~ will become clear and viscous (due to release of the high molecular weight E. coli chromosomal DNA); chromosomal DNA and proteins denature
- Add 150 µl Solution III to neutralize the mixture; mix well: vortexing is allowed; single stranded chromosomal DNA and denatured proteins precipitate
- Leave on ice for 5-10 minutes; a white foam/precipitate forms, containing chromosomal DNA and proteins; the solution itself becomes clear and yellowish
- Spin about 10 minutes at RT
- Take 450 µl of the clear supernatant and transfer it to a new eppendorf tube
- Add 300 µl isopropanol, mix and spin directly at RT for 10 - 15 min
- Pour the supernatant from the tube, remove fluid remains with a P200
- Add about 180 µl 70 % ethanol to the pellet, spin briefly and remove fluid with a P200
- Dry the pellet on air or standing in a flow
- Dissolve the pellet in 50 - 100 µl TE with RNAaseA

## Appendix 5: Enzyme restriction protocol for *XbaI*



Thermo
SCIENTIFIC

PRODUCT INFORMATION

# XbaI

**#ER0685** 3000 U

**Lot: ___** **Expiry Date: _**

5'...T↓C T A G A...3'
3'...A G A T C↑T...5'

| | |
|---|---|
| Concentration: | 10 U/µL |
| Source: | *Xanthomonas badrii* |
| Supplied with: | 2x1 mL of 10X Buffer Tango |

**Store at -20°C**

Tango 37° Dam 20' 65° HC LO

67

In total 3 vials. BSA included

**www.thermoscientific.com/onebio**

**RECOMMENDATIONS**

**1X Thermo Scientific Tango Buffer** (for 100% XbaI digestion)
33 mM Tris-acetate (pH 7.9), 10 mM magnesium acetate, 66 mM potassium acetate, 0.1 mg/mL BSA.

**Incubation temperature**
37°C.

**Unit Definition**
One unit is defined as the amount of XbaI required to digest 1 µg of lambda DNA *dam*⁻-SmaI fragments in 1 hour at 37°C in 50 µL of recommended reaction buffer.

**Dilution**
Dilute with Dilution Buffer (#B19): 10 mM Tris-HCl (pH 7.4 at 25°C), 100 mM KCl, 1 mM EDTA, 1 mM DTT, 0.2 mg/mL BSA and 50% glycerol.

**Double Digests**
Tango™ Buffer is provided to simplify buffer selection for double digests. 98% of Fermentas restriction enzymes are active in a 1X or 2X concentration of Tango™ Buffer. Please refer to the Fermentas Catalog or go to www.thermoscientific.com/doubledigest to choose the best buffer for your experiments.

**Storage Buffer**
XbaI is supplied in: 10 mM Tris-HCl (pH 7.4 at 25°C), 100 mM KCl, 1 mM DTT, 1 mM EDTA, 0.2 mg/mL BSA and 50% glycerol.

Rev.13

# Appendix 6: MEGABLAST and BLASTn searches of ORFs

## Sequence 1

| | Total ORFs | Hits | No hits |
|---|---|---|---|
| MEGABLAST | 79 | 36 | 43 |
| BLASTn | 79 | 79 | 0 |

## Hits on MEGABLAST

| ORF name | Length (bp) | Organism | Accession | Pairwise identity (%) | Query coverage (%) | E-value | Identical proteins |
|---|---|---|---|---|---|---|---|
| S1 - ORF002 | 141 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | WP_072756907: hypothetical protein [Gammaproteobacteria] |
| S1 - ORF004 | 108 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | MKR19254: hypothetical protein D7V69_03850 |
| S1 - ORF007 | 771 | *Planctomycetes bacterium* | CP036432 | 76.4 | 91.70 | 2.16 × 10 -140 | WP_146600803: nucleotidyltransferase domain-containing protein |
| S1 - ORF008 | 990 | *Planctomycetes bacterium* | CP036525 | 82.0 | 93.03 | 0 | TWU40200: putative nucleotidyltransferase |
| S1 - ORF0010 | 414 | *Roseimaritima ulvae* | CP042914 | 78.8 | 91.06 | 3.19 × 10 -84 | WP_068132123: DUF393 domain-containing protein |
| S1- ORF0011 | 609 | *Rhodopirellula baltica* | BX294153 | 77.01 | 92.98 | 5.88 × 10 -121 | WP_146516088: heme-binding protein |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S1 - ORF0021 | 354 | *Planctomycetes bacterium* | CP036341 | 92.06 | 79.94 | 2.34 × 10 -110 | ACT08310: hypothetical protein Dd1591_3499 |
| S1 - ORF0028 | 384 | *Planctomycetes bacterium* | CP036525 | 91.01 | 100 | 2.10 × 10 -149 | TWT80386: hypothetical protein CA13_18010 |
| S1 - ORF0032 | 1116 | *Planctomycetes sp.* | CP011270 | 69.6 | 87.63 | 3.88 × 10 -101 | WP_068259366: slipin family protein |
| S1 - ORF0038 | 1038 | *Planctomycetes bacterium* | CP036264 | 72.3 | 100 | 6.48 × 10 -155 | WP_068132169: ADP-ribosylglycohydrolase family protein |
| S1 - ORF0042 | 477 | *Methylocystis sp.* | HE956757 | 96.8 | 6.50*** | 3.50 × 10 -02 | WP_146461115: hypothetical protein |
| S1 - ORF0047 | 396 | *Planctomycetes bacterium* | CP036526 | 95.7 | 46.97 | 3.46 × 10 -77 | WP_145418494: TROVE domain-containing protein [Planctomycetes bacterium K23_9] |
| S1 - ORF0070 | 396 | *Rhodopirellula baltica* | BX294153 | 88.9 | 100 | 2.48 × 10 -142 | WP_068260228: tryptophan-rich sensory protein |
| S1 - ORF0071 | 270 | *Rhodopirellula baltica* | BX294153 | 95.4 | 97.41 | 9.56 × 10 -114 | CAD79096: hypothetical protein RB11393 |
| S1 - ORF0074 | 108 | *Rhodopirellula baltica* | BX294153 | 96.6 | 82.41 | 9.70 × 10 -323 | - |
| S1 - ORF0076 | 357 | *Rhodopirellula baltica* | BX294153 | 85.4 | 100 | 2.37 × 10 -110 | WP_146516147: hypothetical protein |
| S1 - ORF0087 | 726 | *Planctomycetes bacterium* | CP036298 | 69.1 | 99.86 | 9.86 × 10 -75 | TWU55009: DNA alkylation repair enzyme |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| S1 - ORF0090 | 636 | *Rhodopirellula baltica* | BX294153 | 77.3 | 97.33 | 2.99 × 10 -131 | WP_068264266: HD domain-containing protein |
| S1 - ORF0094 | 279 | *Rhodopirellula baltica* | BX294153 | 87.7 | 99.28 | 1.80 × 10 -91 | WP_146516122: TIGR03643 family protein |
| S1 - ORF00102 | 792 | *Planctomycetes bacterium* | CP036262 | 74.8 | 100 | 1.23 × 10 -143 | WP_164922435: tRNA-binding protein |
| S1 - ORF00105 | 321 | *Planctomycetes bacterium* | CP036262 | 100 | 100 | 0 | WP_164922434: DUF1801 domain-containing protein |
| S1 - ORF00108 | 552 | *Planctomycetes bacterium* | CP036262 | 100 | 100 | 0 | WP_146459133: CIA30 family protein |
| S1 - ORF00114 | 123 | *Planctomycetes bacterium* | CP036262 | 100 | 100 | 0 | WP_007332630: CoA-binding protein |
| S1 - ORF00122 | 390 | *Planctomycetes bacterium* | CP036262 | 100 | 100 | 0 | WP_068264250: CoA-binding protein |
| S1 - ORF00123 | 267 | *Planctomycetes bacterium* | CP036262 | 100 | 100 | 0 | KAA1258034: hypothetical protein LF1_05490 |
| S1 - ORF00140 | 735 | *Rhodopirellula baltica* | BX294156 | 74.2 | 91.43 | 3.48 × 10 -112 | EMI45957: UvrB/UvrC protein |
| S1 - ORF00144 | 867 | *Rhodopirellula baltica* | BX294146 | 68.8 | 96.19 | 5.39 × 10 -79 | TWT75350: Alpha/beta hydrolase family protein |
| S1 - ORF00145 | 96 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  | TPC99492: lac repressor |
| S1 - ORF00146 | 660 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  | EDS05563: chloramphenicol acetyltransferase |

| | | | | | | |
|---|---|---|---|---|---|---|
| S1 - ORF00147 | 513 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | WP_171831280: tyrosine-type recombinase/integrase |
| S1 - ORF00148 | 120 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | WP_001302176: hypothetical protein |
| S1 - ORF00149 | 183 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | PTB82203: hypothetical protein |
| S1 - ORF00151 | 756 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | EEU3264372: replication initiation protein RepE |
| S1 - ORF00152 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | WP_001365321: hypothetical protein |
| S1 - ORF00155 | 1176 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | WP_001304218: plasmid-partitioning protein SopA |
| S1 - ORF00156 | 972 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | WP_059242500: ParB/RepB/Spo0J family plasmid partition protein |

## BLASTn search of «No hits» on MEGABLAST

| ORF name | Length (bp) | Identical proteins | Grade (%) | Number of results (out of 10) |
|---|---|---|---|---|
| S1 - ORF0013 | 99 | WP_162273174: hypothetical protein [Rubripirellula obstinata] | 73.5 | 10 |
| S1 - ORF0014 | 495 | WP_146516087: rhodanese-like domain-containing protein [Rubripirellula amarantea] | 92.7 | 10 |
| **S1 - ORF0016** | **102** | **-** | | |
| **S1 - ORF0017** | **96** | **-** | | |
| **S1 - ORF0018** | **99** | **-** | | |
| S1 - ORF0023 | 951 | QDU26629: Group II intron-encoded protein LtrA [Planctomycetes bacterium ETA_A8] | 86.7 | 10 |
| S1 - ORF0024 | 471 | REJ87213: group II intron reverse transcriptase/maturase [Planctomycetes bacterium] | 88.5 | 10 |
| **S1 - ORF0025** | **105** | **-** | | |
| **S1 - ORF0026** | **96** | **-** | | |
| S1 - ORF0034 | 651 | WP_068258815: RNA 2'-phosphotransferase [Rubripirellula obstinata] | 87.5 | 10 |
| S1 - ORF0037 | 660 | WP_146405649: tyrosine-protein phosphatase [Planctomycetes bacterium Poly21] | 95.4 | 10 |

| | | | | |
|---|---|---|---|---|
| S1 - ORF0041 | 264 | WP_146462253: M28 family peptidase [Rubripirellula tenax] | 42.0 | 10 |
| S1 - ORF0044 | 639 | WP_146461113: hypothetical protein [Rubripirellula tenax] | 71.7 | 8 |
| S1 - ORF0046 | 225 | TWT97173: hypothetical protein Pla100_23220 [Planctomycetes bacterium Pla100] | 41.7 | 1 |
| S1 - ORF0049 | 579 | WP_146520985: hypothetical protein [Planctomycetes bacterium Pla52n] | 94.0 | 10 |
| S1 - ORF0052 | 411 | WP_068264232: hypothetical protein [Rubripirellula obstinata] | 91.2 | 10 |
| S1 - ORF0054 | 669 | WP_068264230: metallophosphoesterase family protein [Rubripirellula obstinata] | 98.4 | 10 |
| **S1 - ORF0056** | **156** | - | | |
| S1 - ORF0060 | 1548 | MBA3583771: transposase [Gemmatimonadetes bacterium] | 74.0 | 10 |
| **S1 - ORF0061** | **111** | - | | |
| S1 - ORF0063 | 336 | WP_145348580: type II toxin-antitoxin system RelE/ParE family toxin [Planctomycetes bacterium EC9] | 74.7 | 10 |
| S1 - ORF0066 | 141 | TWU39980: hypothetical protein Q31b_32960 [Planctomycetes bacterium Q31b] | 46.3 | 2 |

| | | | | |
|---|---|---|---|---|
| **S1 - ORF0068** | **222** | - | | |
| **S1 - ORF0079** | **93** | - | | |
| S1 - ORF0080 | 195 | WP_146459139: hypothetical protein [Rubripirellula tenax] | 38.3 | 10 |
| S1 - ORF0082 | 504 | WP_146410002: DinB family protein [Planctomycetes bacterium Poly21] | 92.7 | 10 |
| S1 - ORF0086 | 225 | WP_146516129: hypothetical protein [Rubripirellula amarantea] | 85.7 | 10 |
| S1 - ORF0091 | 192 | WP_068264265: hypothetical protein [Rubripirellula obstinata] | 96.8 | 10 |
| S1 - ORF0096 | 162 | TWT51002: hypothetical protein Pla22_37780 [Rubripirellula amarantea] | 90.1 | 10 |
| S1 - ORF00100 | 390 | WP_146516120: VOC family protein [Rubripirellula amarantea] | 98.5 | 10 |
| S1 - ORF00112 | 1062 | WP_146459132: hypothetical protein [Rubripirellula tenax] | 89.9 | 10 |
| S1 - ORF00117 | 432 | WP_146462230: PIN domain-containing protein [Rubripirellula tenax] | 94.1 | 10 |
| S1 - ORF00119 | 219 | WP_146462231: hypothetical protein [Rubripirellula tenax] | 96.5 | 9 |
| **S1 - ORF00120** | **96** | - | | |
| **S1 - ORF00125** | **108** | - | | |
| S1 - ORF00128 | 282 | WP_146520985: hypothetical protein | 88.7 | 10 |

| | | | | |
|---|---|---|---|---|
| | | [Planctomycetes bacterium Pla52n] | | |
| S1 - ORF00129 | 474 | WP_068264232: hypothetical protein [Rubripirellula obstinata] | 83.8 | 10 |
| S1 - ORF00130 | 105 | RLS74515: ISAs1 family transposase, partial [Planctomycetes bacterium] | 57.4 | 6 |
| S1 - ORF00133 | 666 | OYV96015: hypothetical protein B7Z68_06230 [Acidobacteria bacterium 21-70-11] | 45.0 | 10 |
| **S1 - ORF00135** | **117** | - | | |
| S1 - ORF00136 | 456 | AJY36415: putative lipo domain protein [Burkholderia mallei] | 26.6 | 4 |
| S1 - ORF00138 | 1413 | WP_008674207: HAMP domain-containing histidine kinase [Rhodopirellula sallentina] | 88.9 | 10 |
| S1 - ORF00142 | 429 | WP_009095090: aminoacyl-tRNA hydrolase [Rhodopirellula sp. SWK7] | 92.3 | 10 |

# No hits on MEGABLAST or BLASTn: 0

# Sequence 2

| | Total ORFs | Hits | No hits |
|---|---|---|---|
| MEGABLAST | 41 | 34 | 7 |
| BLAStn | 41 | 41 | 0 |

## Hits on MEGABLAST

| ORF name | Length (bp) | Organism | Accession | Pairwise identity (%) | Query coverage (%) | E-value | Identical proteins |
|---|---|---|---|---|---|---|---|
| S2 - ORF002 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | EEV9030702 hypothetical protein: |
| S2 - ORF003 | 159 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | ADE34479: hypothetical protein |
| S2 - ORF004 | 831 | *Homo sapiens* | AC277637 | 99.4 | 39.11*** | 1.92 × 10 - 160 | WP_133689075: 3-dehydroquinate synthase |
| S2 - ORF005 | 885 | *Maribacter sp.* | CP018760 | 81.4 | 100 | 0 | TDT37185.1: 4-hydroxybenzoate polyprenyltransferase |
| S2 - ORF006 | 909 | *Maribacter sp.* | CP018760 | 89.5 | 100 | 0 | WP_084135053.1: TatD family hydrolase |
| S2 - ORF008 | 582 | *Maribacter sp.* | CP018760 | 86.9 | 100 | 0 | WP_073245624: tRNA-(ms[2]io[6]A)-hydroxylase |
| S2 - ORF009 | 891 | *Maribacter sp.* | CP018760 | 84.4 | 100 | 0 | TDT37181.1: exopolyphosphatase/guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S2 - ORF0010 | 2068 | *Maribacter sp.* | CP011318 | 84.4 | 100 | 0 | WP_133689069.1: polyphosphate kinase 1 |
| S2 - ORF0012 | 652 | *Maribacter sp.* | CP018760 | 82.7 | 100 | 0 | TDT37178.1: pyridoxamine 5'-phosphate oxidase |
| S2 - ORF0013 | 909 | *Maribacter sp.* | CP018760 | 81.6 | 100 | 0 | WP_133689067.1: ribonuclease Z |
| S2 - ORF0015 | 927 | *Maribacter sp.* | CP018760 | 85.8 | 100 | 0 | WP_133689065.1: aspartate carbamoyltransferase catalytic subunit |
| S2 - ORF0016 | 549 | *Maribacter sp.* | CP018760 | 84.0 | 100 | 1.61 × 10 - 165 | WP_133689064.1: bifunctional pyr operon transcriptional regulator/uracil phosphoribosyltransferase PyrR |
| S2 - ORF0017 | 732 | *Maribacter sp.* | CP018760 | 82.6 | 99.86 | 0 | TDT37173.1: DNA-binding LytR/AlgR family response regulator |
| S2 - ORF0018 | 1042 | *Maribacter sp.* | CP011318 | 79.3 | 97.0 | 0 | WP_133689062.1: histidine kinase |
| S2 - ORF0020 | 996 | *Maribacter sp.* | LT629754 | 99.60 | 80.9 | 0 | WP_133689060.1 galactose oxidase |
| S2 - ORF0021 | 1344 | *Maribacter sp.* | CP011318 | 75.9 | 100 | 0 | TDT37169.1: uncharacterized protein DUF4270 |
| S2 - ORF0023 | 1839 | *Maribacter sp.* | LT629754 | 93.6 | 100 | 0 | WP_073245651.1:30S ribosomal protein S1 |
| S2 - ORF0024 | 383 | *Maribacter sp.* | CP018760 | 87.4 | 100 | 8.88 × 10 - 129 | HDZ03562.1: LysM peptidoglycan-binding domain-containing protein |
| S2 - ORF0025 | 699 | *Maribacter sp.* | CP018760 | 82.0 | 100 | 0 | WP_073245656.1: (d)CMP kinase |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| S2 - ORF0027 | 2457 | *Maribacter sp.* | LT629754 | 84.7 | 100 | 0 | TDT37164.1: ATP-dependent Lon protease |
| S2 - ORF0030 | 555 | *Maribacter sp.* | CP011318 | 93.9 | 100 | 0 | SHK43472.1: RNA polymerase sigma-70 factor, ECF subfamily |
| S2 - ORF0031 | 564 | *Maribacter sp.* | LT629754 | 82.1 | 100 | 9.78 × 10 - 156 | WP_073245662.1: hypothetical protein |
| S2 - ORF0032 | 1096 | *Maribacter sp.* | CP018760 | 80.7 | 100 | 0 | WP_073245665.1: hypothetical protein |
| S2 - ORF0034 | 1371 | *Maribacter sp.* | LT629754 | 80.5 | 100 | 0 | TDT37159.1: UMF1 family MFS transporter |
| S2 - ORF0035 | 816 | *Maribacter sp.* | CP011318 | 84.3 | 100 | 0 | WP_133689050.1: M48 family metallopeptidase |
| S2 - ORF0036 | 1089 | *Maribacter sp.* | LT629754 | 81.2 | 100 | 0 | WP_073245671.1: glycoside hydrolase family 31 protein |
| S2 - ORF0037 | 123 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S2 - ORF0039 | 660 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S2 - ORF0040 | 513 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S2 - ORF0041 | 120 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S2 - ORF0043 | 756 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S2 - ORF0044 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| S2 - ORF0045 | 1167 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S2 - ORF0046 | 972 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |

# BLASTn search of «No hits» on MEGABLAST

| **ORF name** | **Length (bp)** | **Identical proteins** | **Grade (%)** | **Number of results (out of 10)** |
|---|---|---|---|---|
| S2 - ORF007 | 852 | WP_073245622: EboA domain-containing protein [Maribacter aquivivus] | 94.7 | 10 |
| S2 - ORF0011 | 246 | WP_133689068: histidine phosphatase family protein [Maribacter spongiicola] | 96.9 | 10 |
| S2 - ORF0014 | 333 | WP_133689066: ribonuclease Z [Maribacter spongiicola] | 98.6 | 10 |
| S2 - ORF0019 | 342 | WP_133689061: DUF4907 domain-containing protein [Maribacter spongiicola] | 93.4 | 10 |
| S2 - ORF0022 | 1251 | TDT37168: long-subunit fatty acid transport protein [Maribacter spongiicola] | 97.8 | 10 |
| **S2 - ORF0028** | **120** | - |  |  |
| S2 - ORF0033 | 735 | WP_036158526: DUF2807 domain-containing protein [Maribacter forsetii] | 93.6 | 10 |

## No hits on Megablast or Blastn: 0

# Sequence 3

| | Total ORFs | Hits | No hits |
|---|---|---|---|
| MEGABLAST | 56 | 19 | 37 |
| BLASTn | 56 | 45 | 11 |

## Hits on MEGABLAST

| ORF name | Length (bp) | Organism | Accession | Pairwise identity (%) | Query coverage (%) | E-value | Protein |
|---|---|---|---|---|---|---|---|
| S3 - ORF001 | 972 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S3 - ORF003 | 1176 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S3 - ORF004 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S3 - ORF006 | 756 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S3 - ORF008 | 183 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S3 - ORF009 | 513 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| S3 - ORF0010 | 660 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S3 - ORF0011 | 96 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S3 - ORF0012 | 2885 | *Rubinisphaera brasiliensis* | CP002546 | 69.1 | 24.15**** | 2.93 × 10 -63 | WP_180683958.1: ammonium transporter |
| S3 - ORF0022 | 1804 | *Roseimaritima ulvae* | CP042914 | 74.4 | 99.22 | 0 | WP_164100682.1: elongation factor 4 |
| S3 - ORF0037 | 894 | *Planctomycetes bacterium* | CP036432 | 68.8 | 55.82***** | 4.58 × 10 -42 | WP_146579417.1: histone deacetylase |
| S3 - ORF0055 | 1270 | *Crateriforma conspicua* | CP036319 | 69.6 | 85.25 | 5.56 × 10 -106 | WP_164103645.1: beta-ketoacyl-ACP synthase II |
| S3 - ORF0056 | 246 | *Planctomycetes bacterium* | CP036262 | 84.1 | 97.15 | 1.77 × 10 -65 | WP_164103644.1: acyl carrier protein |
| S3 - ORF0059 | 763 | *Planctomycetes bacterium* | CP036292 | 74.9 | 95.39 | 3.60 × 10 -131 | WP_164103643.1: 3-oxoacyl-[acyl-carrier-protein] reductase |
| S3 - ORF0070 | 1343 | *Planctomycetes bacterium* | CP036525 | 69.7 | 43.62**** | 5.68 × 10 -62 | WP_164100385.1: 3-phosphoshikimate 1-carboxyvinyltransferase |
| S3 - ORF0080 | 788 | *Planctomycetes bacterium* |  | 79.7 | 88.93 | 1.69 × 10 -173 | WP_164102930.1: FliA/WhiG family RNA polymerase sigma factor |
| S3 - ORF0084 | 117 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |

| | | | | | |
|---|---|---|---|---|---|
| S3 - ORF0086 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S3 - ORF0088 | 141 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |

# BLASTn search of «No hits» on MEGABLAST

| ORF name | Length (bp) | Identical proteins | Grade (%) | Number of results (out of 10) |
|---|---|---|---|---|
| **S3 - ORF0015** | **1026** | **-** | | |
| **S3 - ORF0016** | **531** | **-** | | |
| S3 - ORF0018 | 615 | WP_052031530: class I SAM-dependent methyltransferase [Rhodopirellula maiorica] | 79.8 | 10 |
| S3 - ORF0020 | 618 | EMI42599: putative membrane protein [Rhodopirellula sp. SWK7] | 77.1 | 10 |
| S3 - ORF0024 | 1086 | WP_145098080: ROK family protein [Planctomycetes bacterium Poly24] | 81.9 | 10 |
| S3 - ORF0026 | 345 | MAI72440: hypothetical protein [Rhodopirellula sp.] | 65.3 | 10 |
| **S3 - ORF0027** | **102** | **-** | | |
| S3 - ORF0029 | 2808 | WP_154900441: circularly permuted type 2 ATP-grasp protein [Planctomycetes bacterium CA11] | 70.8 | 10 |
| S3 - ORF0030 | 885 | WP_145262651: transglutaminase family protein [Planctomycetes bacterium Pan216] | 76.6 | 10 |
| S3 - ORF0031 | 792 | WP_145419616: GIY-YIG nuclease family protein [Planctomycetes bacterium K23_9] | 74.8 | 10 |
| S3 - ORF0032 | 1101 | QDV58998: hypothetical protein Mal33_50230 [Planctomycetes bacterium Mal33] | 77.8 | 10 |

| | | | | |
|---|---|---|---|---|
| S3 - ORF0034 | 984 | WP_164102765: D-2-hydroxyacid dehydrogenase [Roseimaritima sp. JC640] | 85.0 | 10 |
| S3 - ORF0035 | 1155 | WP_164102793: protein kinase [Roseimaritima sp. JC640] | 89.5 | 10 |
| S3 - ORF0038 | 1227 | WP_075084119: endonuclease/exonuclease/phosphatase family protein [Mariniblastus fucicola] | 74.6 | 10 |
| S3 - ORF0039 | 456 | WP_164101729: response regulator [Roseimaritima sp. JC640] | 90.6 | 10 |
| **S3 - ORF0041** | **102** | **-** | | |
| **S3 - ORF0043** | **141** | **-** | | |
| S3 - ORF0045 | 1428 | WP_164104070: MFS transporter [Roseimaritima sp. JC640] | 78.5 | 10 |
| **S3 - ORF0046** | **114** | **-** | | 10 |
| S3 - ORF0047 | 216 | PHR99503: IS5/IS1182 family transposase [Blastopirellula sp.] | 52.9 | 10 |
| **S3 - ORF0048** | **129** | **-** | | |
| S3 - ORF0051 | 840 | WP_164104021: purine-nucleoside phosphorylase [Roseimaritima sp. JC640] | 89.1 | 10 |
| **S3 - ORF0052** | **126** | **-** | | |
| S3 - ORF0053 | 1458 | WP_146406530: sulfatase-like hydrolase/transferase [Planctomycetes bacterium Poly21] | 83.6 | 10 |
| S3 - ORF0060 | 921 | WP_164103642: ACP S-malonyltransferase [Roseimaritima sp. JC640] | 84.9 | 10 |
| S3 - ORF0062 | 96 | GDX91350: 50S ribosomal protein L32 [Planctomycetia bacterium] | 44.2 | 2 |
| S3 - ORF0063 | 108 | WP_153556032: 50S ribosomal protein L32 [Roseimaritima sp. JC651] | 52.8 | 10 |
| S3 - ORF0065 | 3069 | WP_165225646: protein kinase [Aquisphaera sp. JC669] | 71.5 | 10 |

|  |  |  |  |  |
|---|---|---|---|---|
| **S3 - ORF0066** | **96** | - |  |  |
| S3 - ORF0068 | 1203 | WP_008662881: hypothetical protein [Rhodopirellula europaea] | 74.1 | 10 |
| S3 - ORF0069 | 1233 | WP_145282472: PQQ-binding-like beta-propeller repeat protein [Planctomycetes bacterium Mal33] | 75.9 | 10 |
| **S3 - ORF0072** | **102** | - |  |  |
| S3 - ORF0073 | 1635 | WP_145350034: von Willebrand factor type A domain-containing protein [Planctomycetes bacterium FF011L] | 76.5 | 10 |
| **S3 - ORF0075** | **138** | - |  |  |
| S3 - ORF0076 | 462 | WP_145342310: low molecular weight phosphotyrosine protein phosphatase [Planctomycetes bacterium EC9] | 78.5 | 10 |
| S3 - ORF0078 | 276 | WP_146592785: hypothetical protein [Planctomycetes bacterium Pla52o] | 39.1 | 7 |
| S3 - ORF0083 | 1173 | WP_008693703: type II and III secretion system protein [Rhodopirellula maiorica] | 74.2 | 10 |

## No hits on MEGABLAST or BLASTn

-S3 - ORF0015

- S3 - ORF0016
- S3 - ORF0027
- S3 - ORF0041
- S3 - ORF0043
- S3 - ORF0046
- S3 - ORF0048
- S3 - ORF0052

- S3 - ORF0066
- S3 - ORF0072
- S3 - ORF0075

# Sequence 4

| | Total ORFs | Hits | No hits |
|---|---|---|---|
| MEGABLAST | 51 | 15 | 36 |
| BLASTn | 51 | 40 | 11 |

## Hits on MEGABLAST

| ORF name | Length (bp) | Organism | Accession | Pairwise identity (%) | Query coverage (%) | E-value | Protein |
|---|---|---|---|---|---|---|---|
| S4 - ORF001 | 972 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S4 - ORF002 | 1167 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S4 - ORF005 | 756 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S4 - ORF007 | 126 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S4 - ORF009 | 120 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S4 - ORF0010 | 513 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S4 - ORF0012 | 660 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S4 - ORF0032 | 1824 | *Thermoleptolyngbya sp.* | CP053661 | 70.1 | 68.86** | 1.24e-160 | WP_017304671.1: NAD+ synthase |
| S4 - ORF0035 | 699 | *Geitlerinema sp.* | CP003591 | 71.6 | 81.26 | 5.98e-71 | OJJ27538.1: ribosomal subunit interface protein |
| S4 - ORF0037 | 681 | *Oscillatoria acuminata* | CP003607 | 71.4 | 63.44*** | 3.93e-54 | NJN32795.1: lipoyl(octanoyl) transferase LipB |
| S4 - ORF0047 | 1920 | *Thermoleptolyngbya sp.* | CP053661 | 78.2 | 90.52 | 0 | NJN31524.1: threonine--tRNA ligase |
| S4 - ORF0062 | 807 | *Thermoleptolyngbya sp.* | CP053661 | 69.7 | 53.78** | 1.44e-41 | NJN56932.1: FAD-dependent oxidoreductase |
| S4 - ORF0065 | 732 | *Leptolyngbya sp.* | AP017367 | 77.3 | 31.28*** | 1.58e-40 | BAY80878.1: Mo-dependent nitrogenase-like protein |
| S4 - ORF0079 | 258 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S4 - ORF0080 | 159 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |

## BLASTn search of «No hits» on MEGABLAST

| ORF name | Length (bp) | Identical proteins | Grade (%) | Number of results (out of 10) |
|---|---|---|---|---|
| S4 - ORF0015 | 1080 | NJN29057: aromatic ring-hydroxylating dioxygenase subunit alpha [Synechococcales cyanobacterium RM1_1_8] | 92.0 | 10 |
| S4 - ORF0016 | 1959 | NJM57392: FAD-dependent oxidoreductase [Synechococcales cyanobacterium RU_4_20] | 88.0 | 10 |
| **S4 - ORF0018** | **114** | - | | |
| S4 - ORF0019 | 1827 | NJN32773: alpha/beta hydrolase [Synechococcales cyanobacterium RM1_1_8] | 75.4 | 10 |
| **S4 - ORF0020** | **99** | - | | |
| **S4 - ORF0021** | **168** | - | | |
| S4 - ORF0022 | 1347 | BAZ30658: putative nicotinate phosphoribosyltransferase [Cylindrospermum sp. NIES-4074] | 83.3 | 10 |
| S4 - ORF0026 | 645 | TVQ16582: nicotinate-nucleotide adenylyltransferase [Leptolyngbya sp. DLM2.Bin15] | 73.0 | 10 |
| S4 - ORF0030 | 756 | NJM57395: NUDIX hydrolase [Synechococcales cyanobacterium RU_4_20] | 88.0 | 10 |
| **S4 - ORF0033** | **99** | - | | |
| S4 - ORF0038 | 285 | NJN32796: DUF427 domain-containing protein [Synechococcales cyanobacterium RM1_1_8] | 88.8 | 10 |

| | | | | |
|---|---|---|---|---|
| S4 - ORF0039 | 825 | NJN32797: phytoene/squalene synthase family protein [Synechococcales cyanobacterium RM1_1_8] | 94.9 | 10 |
| **S4 - ORF0040** | **168** | **-** | | |
| **S4 - ORF0041** | **102** | **-** | | |
| S4 - ORF0042 | 942 | WP_009768167: M48 family metalloprotease [Oscillatoriales cyanobacterium JSC-12] | 70.9 | 10 |
| S4 - ORF0043 | 561 | NJN31522: Uma2 family endonuclease [Synechococcales cyanobacterium RM1_1_8] | 92.5 | 10 |
| S4 - ORF0044 | 426 | WP_058030484: hypothetical protein [Pseudoalteromonas phenolica] | 74.6 | 10 |
| S4 - ORF0045 | 663 | NJN31523: Uma2 family endonuclease [Synechococcales cyanobacterium RM1_1_8] | 94.2 | 10 |
| S4 - ORF0048 | 399 | WP_168570017: DUF2605 domain-containing protein [Oxynema sp. AP17] | 70.9 | 10 |
| S4 - ORF0049 | 333 | NJN31526: DUF2973 domain-containing protein [Synechococcales cyanobacterium RM1_1_8] | 65.1 | 10 |
| **S4 - ORF0051** | **147** | **-** | | |
| S4 - ORF0052 | 1302 | NJN30652: response regulator [Synechococcales cyanobacterium RM1_1_8] | 72.6 | 10 |
| **S4 - ORF0053** | **414** | **-** | | |
| S4 - ORF0055 | 1233 | WP_068514664: FIST C-terminal domain-containing protein [Leptolyngbya sp. O-77] | 80.0 | 10 |

|  |  |  |  |  |
|---|---|---|---|---|
| S4 - ORF0056 | 228 | WP_066349091: Calvin cycle protein CP12 [Geminocystis sp. NIES-3708] | 84.9 | 10 |
| S4 - ORF0058 | 603 | WP_068789457: DUF3177 family protein [Phormidium willei] | 78.9 | 10 |
| S4 - ORF0061 | 237 | NJM48050: hypothetical protein [Alkalinema sp. RU_4_3] | 80.1 | 10 |
| S4 - ORF0064 | 822 | WP_068510336: FAD-dependent oxidoreductase [Leptolyngbya sp. O-77] | 81.1 | 10 |
| **S4 - ORF0067** | **108** | **-** |  |  |
| S4 - ORF0069 | 426 | WP_162398825: hypothetical protein [Nostoc sp. B(2019)] | 72.5 | 10 |
| **S4 - ORF0072** | **156** | **-** |  |  |
| S4 - ORF0073 | 537 | NJL85273: porin family protein [Leptolyngbyaceae cyanobacterium SM1_1_3] | 68.0 | 10 |
| S4 - ORF0074 | 1218 | WP_146133632: tetratricopeptide repeat protein, partial [filamentous cyanobacterium Phorm 46] | 60.2 | 10 |
| **S4 - ORF0075** | **99** | **-** |  |  |
| S4 - ORF0076 | 1101 | NJR68231: signal peptidase I [Synechococcales cyanobacterium CRU_2_2] | 71.3 | 10 |
| S4 - ORF0078 | 975 | NJN32371: nuclear transport factor 2 family protein [Synechococcales cyanobacterium RM1_1_8] | 70.4 | 10 |

## No hits on Megablast or Blastn

- S4 - ORF0018

- S4 - ORF0020

- S4 - ORF0021

- S4 - ORF0033

- S4 - ORF0040

- S4 - ORF0041

- S4 - ORF0051

- S4 - ORF0053

- S4 - ORF0067

- S4 - ORF0072

- S4 - ORF0075

# Sequence 5

| | Total ORFs | Hits | No hits |
|---|---|---|---|
| MEGABLAST | 62 | 59 | 3 |
| BLASTn | 62 | 56 | 6 |

## Hits on MEGABLAST

| ORF name | Length (bp) | Organism | Accession | Pairwise identity (%) | Query coverage (%) | E-value | Protein |
|---|---|---|---|---|---|---|---|
| S5 - ORF002 | 972 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S5 - ORF003 | 1167 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S5 - ORF004 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S5 - ORF007 | 756 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S5 - ORF008 | 96 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S5 - ORF009 | 126 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| S5 - ORF0010 | 210 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S5 - ORF0011 | 120 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S5 - ORF0012 | 513 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S5 - ORF0013 | 93 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S5 - ORF0015 | 660 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S5 - ORF0018 | 96 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |  |  |
| S5 - ORF0021 | 96 | *Klebsiella sp.* | CP056483 | 100 | 100 | 5.91e-40 | - |
| S5 - ORF0022 | 513 | *Klebsiella sp.* | CP056483 | 99.8 | 99.03 | 0 | STW27926: Signal recognition particle |
| S5 - ORF0023 | 792 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | WP_139539422: cytochrome c biogenesis protein CcsA |
| S5 - ORF0025 | 1287 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | WP_049089916: HlyC/CorC family transporter |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S5 - ORF0026 | 591 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | WP_110277451: MULTISPECIES: nucleotide exchange factor GrpE [Klebsiella] |
| S5 - ORF0027 | 99 | *Klebsiella sp.* | CP056483 | 100 | 100 | 1.44e-41 | EGK58523: NAD(+) kinase |
| S5 - ORF0028 | 807 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | WP_004104646: NAD(+) kinase [Klebsiella] |
| S5 - ORF0030 | 1662 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | QLO35732: DNA repair protein RecN |
| S5 - ORF0032 | 141 | *Klebsiella sp.* | CP056483 | 100 | 100 | 3.79e-64 | WP_160886650: hypothetical protein |
| S5 - ORF0033 | 342 | *Klebsiella sp.* | CP056483 | 100 | 100 | 8.38e-173 | WP_112217070: outer membrane protein assembly factor BamE |
| S5 - ORF0035 | 291 | *Klebsiella sp.* | CP056483 | 100 | 100 | 3.40e-145 | QLO35730: RnfH family protein |
| S5 - ORF0037 | 477 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | WP_112217068: type II toxin-antitoxin system RatA family toxin |
| S5 - ORF0038 | 483 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | WP_110277446: MULTISPECIES: SsrA-binding protein SmpB |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S5 - ORF0039 | 102 | *Klebsiella sp.* | CP056483 | 100 | 100 | 3.54e-43 | - |
| S5 - ORF0041 | 1284 | *Pantoea sp.* | CP009866 | 94.7 | 100 | 0 | WP_163525671: tyrosine-type recombinase/integrase |
| S5 - ORF0044 | 1821 | *Serratia marcescens* | AP021873 | 92.0 | 100 | 0 | WP_163525672: DUF4365 domain-containing protein |
| S5 - ORF0047 | 1446 | *Cronobacter malonaticus* | CP013940 | 78.2 | 96.89 | 0 | WP_163525673: SIR2 family protein |
| S5 - ORF0048 | 501 | *Cronobacter malonaticus* | CP013940 | 79.4 | 100 | 7.08e-119 | WP_163525674: TIR domain-containing protein |
| S5 - ORF0049 | 1479 | *Salmonella enterica* | CP053332 | 82.4 | 100 | 0 | WP_163525675: relaxase/mobilization nuclease domain-containing protein |
| S5 - ORF0050 | 372 | *Klebsiella grimontii* | CP055309 | 93.8 | 100 | 1.46e-157 | WP_163525676: plasmid mobilization relaxosome protein MobC |
| S5 - ORF0051 | 270 | *Raoultella ornithinolytica* | CP038281 | 93.7 | 100 | 1.73e-110 | WP_163525677: hypothetical protein |
| S5 - ORF0052 | 153 | *Raoultella ornithinolytica* | CP038281 | 96.7 | 100 | 1.47e-63 | WP_169050475: hypothetical protein |
| S5 - ORF0053 | 303 | *Raoultella ornithinolytica* | CP038281 | 97.4 | 100 | 7.85e-141 | WP_163525678: hypothetical protein |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S5 - ORF0056 | 276 | *Raoultella ornithinolytica* | CP038281 | 100 | 100 | 4.45e-137 | WP_063407905: hypothetical protein |
| S5 - ORF0058 | 288 | *Serratia fonticola* | CP011254 | 99.3 | 100 | 7.41e-141 | WP_163525679: hypothetical protein |
| S5 - ORF0059 | 846 | *Serratia fonticola* | CP011254 | 98.2 | 100 | 0 | WP_163525680: replication initiation protein |
| S5 - ORF0061 | 258 | *Serratia fonticola* | CP011254 | 96.9 | 100 | 1.75e-116 | WP_163525681: helix-turn-helix transcriptional regulator |
| S5 - ORF0063 | 228 | *Serratia fonticola* | CP011254 | 98.2 | 100 | 1.09e-105 | WP_097736840: hypothetical protein |
| S5 - ORF0066 | 204 | *Serratia fonticola* | CP011254 | 97.5 | 100 | 4.33e-91 | WP_163525682: AlpA family phage regulatory protein |
| S5 - ORF0067 | 150 | *Serratia fonticola* | CP011254 | 96.0 | 100 | 7.44e-61 | - |
| S5 - ORF0068 | 114 | *Serratia fonticola* | CP011254 | 98.2 | 100 | 6.62e-47 | - |
| S5 - ORF0071 | 114 | *Klebsiella sp.* | CP056483 | 100 | 100 | 1.28e-49 | EHM50700: hypothetical protein HMPREF0880_00885 |
| S5 - ORF0074 | 1248 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | WP_032453750: tyrosine-type recombinase/integrase |
| S5 - ORF0077 | 645 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | QLO35713: hypothetical protein HV213_07645 |
| S5 - ORF0079 | 126 | *Klebsiella sp.* | CP056483 | 100 | 100 | 4.51e-56 | WP_162285832: NIPSNAP family protein |
| S5 - ORF0084 | 5730 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | QLO35712: ATP-binding protein |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S5 - ORF0085 | 510 | *Klebsiella sp.* | CP033824 | 100 | 100 | 0 | WP_117077281: hypothetical protein |
| S5 - ORF0086 | 126 | *Klebsiella sp.* | CP056483 | 98.4 | 100 | 2.34e-53 | EMF07362: IS1400 transposase A |
| S5 - ORF0087 | 513 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | QLO35724: IS3 family transposase |
| S5 - ORF0088 | 249 | *Klebsiella sp.* | CP056483 | 100 | 100 | 1.79e-122 | VGA88629: integrase catalytic subunit |
| S5 - ORF0089 | 153 | *Klebsiella sp.* | CP056483 | 100 | 100 | 1.29e-70 | EZJ92286: phage integrase family protein |
| S5 - ORF0090 | 813 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | QLO35709: GntR family transcriptional regulator |
| S5 - ORF0092 | 438 | *Klebsiella sp.* | CP056483 | 100 | 100 | 0 | QLO35708: PTS fructose transporter subunit IIA |
| S5 - ORF0094 | 834 | *Klebsiella sp.* | CP056483 | 100 | 62.11*** | 0 | QLO35707: PTS sugar transporter subunit IIC |
| S5 - ORF0095 | 159 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S5 - ORF0097 | 114 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S5 - ORF0098 | 171 | *Escherichia coli* | CP056263 | 100 | 72.51 | 8.25e-55 | |

## BLASTn search of «No hits» on MEGABLAST

| ORF NAME | LENGTH (BP) | IDENTICAL PROTEINS | GRADE (%) | NUMBER OF RESULTS (OUT OF 10) |
|---|---|---|---|---|
| **S5 - ORF0046** | **99** | **-** | | |
| **S5 - ORF0062** | **105** | **-** | | |
| S5 - ORF0069 | 102 | EAA4525289: hypothetical protein | 47.7 | 2 |

## No hits on Megablast or Blastn

- S5 - ORF0021
- S5 - ORF0039
- S5 - ORF0046
- S5 - ORF0062
- S5 - ORF0067
- S5 - ORF0068

# Sequence 6

| | **Total ORFs** | **Hits** | **No hits** |
|---|---|---|---|
| MEGABLAST | 57 | 11 | 46 |
| BLASTn | 57 | 40 | 17 |

## Hits on MEGABLAST

| ORF name | Length (bp) | Organism | Accession | Pairwise identity (%) | Query coverage (%) | E-value | Protein |
|---|---|---|---|---|---|---|---|
| S6 - ORF001 | 855 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S6 - ORF002 | 1167 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S6 - ORF005 | 756 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S6 - ORF006 | 126 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S6 - ORF007 | 183 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S6 - ORF008 | 513 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S6 - ORF009 | 660 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S6 - ORF0016 | 1491 | *Halomicronema hongdechloris* | CP021983 | 83.9 | 99.93 | 0 | WP_096731512: IS1380 family transposase |
| S6 - ORF0028 | 585 | *Geobacter anodireducens* | CP014963 | 67.10 | 74.36 | 6.90e-25 | WP_153293456: zeta toxin family protein |
| S6 - ORF0056 | 1770 | *Delftia tsuruhatensis* | CP045291 | 79.3 | 5.20***** | 2.84e-10 | WP_007305794: RNB domain-containing ribonuclease *** |
| S6 - ORF0085 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |

# BLASTn search of «No hits» on MEGABLAST

| ORF name | Length (bp) | Identical proteins | Grade (%) | Number of results (out of 10) |
|---|---|---|---|---|
| **S6 - ORF0012** | **528** | - | | |
| S6 - ORF0013 | 174 | AFY93165: hypothetical protein Cha6605_2070 [Chamaesiphon minutus PCC 6605] | 44.0 | 7 |
| S6 - ORF0014 | 3393 | NEQ48243: hypothetical protein [Leptolyngbya sp. SIOISBB] | 68.6 | 10 |
| **S6 - ORF0018** | **612** | - | | |
| S6 - ORF0019 | 120 | NJR70312: hypothetical protein [Synechococcales cyanobacterium CRU_2_2] | 60.3 | 1 |
| S6 - ORF0020 | 540 | NJN32002: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 55.2 | 10 |
| S6 - ORF0021 | 261 | NJR67456: hypothetical protein [Synechococcales cyanobacterium CRU_2_2] | 83.1 | 1 |
| S6 - ORF0023 | 399 | NJN31714: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 80.7 | 1 |
| **S6 - ORF0024** | **105** | - | | |
| S6 - ORF0025 | 690 | NJN31713: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 70.6 | 2 |
| S6 - ORF0027 | 216 | WP_008317590: hypothetical protein [Leptolyngbya sp. PCC 6406] | 49.7 | 10 |

| | | | | |
|---|---|---|---|---|
| S6 - ORF0030 | 1125 | NJN31710: DGQHR domain-containing protein [Synechococcales cyanobacterium RM1_1_8] | 86.7 | 10 |
| S6 - ORF0032 | 699 | NJN29327: DNA repair protein RadC [Synechococcales cyanobacterium RM1_1_8] | 61.5 | 10 |
| S6 - ORF0033 | 180 | WP_017326153: DUF86 domain-containing protein [Synechococcus sp. PCC 7336] | 49.6 | 10 |
| S6 - ORF0036 | 534 | NJN32748: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 69.2 | 10 |
| S6 - ORF0037 | 762 | NEQ28405: ISKra4 family transposase [Microcoleus sp. SIO2G3] | 85.6 | 10 |
| S6 - ORF0038 | 210 | NJR71155: hypothetical protein [Synechococcales cyanobacterium CRU_2_2] | 51.5 | 10 |
| S6 - ORF0040 | 405 | NJN32748: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 81.7 | 3 |
| S6 - ORF0042 | 864 | NJN32749: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 75.4 | 10 |
| S6 - ORF0043 | 552 | NJN32750: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 13.0 | 1 |
| **S6 - ORF0045** | **99** | - | | |
| **S6 - ORF0047** | **123** | - | | |
| **S6 - ORF0048** | **150** | - | | |
| S6 - ORF0049 | 213 | RPI65896: protein-L-isoaspartate O- | 33.1 | 3 |

|  |  |  |  |  |
|---|---|---|---|---|
|  |  | methyltransferase, partial [Ignavibacteriae bacterium] |  |  |
| S6 - ORF0052 | 375 | NJM58486: hypothetical protein [Synechococcales cyanobacterium RU_4_20] | 34.0 | 1 |
| S6 - ORF0054 | 1089 | NJM58485: relaxase/mobilization nuclease domain-containing protein [Synechococcales cyanobacterium RU_4_20] | 73.2 | 10 |
| S6 - ORF0058 | 1533 | NJN30180: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 87.7 | 10 |
| **S6 - ORF0060** | **258** | **-** |  |  |
| S6 - ORF0062 | 390 | NJR70879: hypothetical protein [Synechococcales cyanobacterium CRU_2_2] | 50.8 | 2 |
| S6 - ORF0063 | 603 | NJR70880: hypothetical protein [Synechococcales cyanobacterium CRU_2_2] | 87.1 | 10 |
| **S6 - ORF0064** | **96** | **-** |  |  |
| **S6 - ORF0065** | **96** | **-** |  |  |
| S6 - ORF0066 | 186 | NJN30184: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 63.4 | 2 |
| S6 - ORF0068 | 255 | NJN30185: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 83.3 | 2 |
| **S6 - ORF0069** | **171** | **-** |  |  |
| **S6 - ORF0070** | **228** | **-** |  |  |
| **S6 - ORF0071** | **234** | **-** |  |  |

|  |  |  |  |  |
|---|---|---|---|---|
| S6 - ORF0072 | 321 | NJN30946: hypothetical protein [Synechococcales cyanobacterium RM1_1_8] | 83.5 | 3 |
| **S6 - ORF0073** | **126** | **-** |  |  |
| **S6 - ORF0074** | **129** | **-** |  |  |
| S6 - ORF0075 | 1803 | NJM58305: hypothetical protein [Synechococcales cyanobacterium RU_4_20] | 87.3 | 10 |
| **S6 - ORF0077** | **204** | - |  |  |
| **S6 - ORF0079** | **99** | - |  |  |
| S6 - ORF0081 | 192 | WP_144969979: 4a-hydroxytetrahydrobiopterin dehydratase [Bremerella volcania] | 86.5 | 10 |
| **S6 - ORF0082** | **168** | **-** |  |  |
| S6 - ORF0083 | 273 | WP_166276346: acetyltransferase [Aphanocapsa montana] | 70.7 | 10 |

# No hits on MEGABLAST or BLASTn

- S6 - ORF0012
- S6 - ORF0018
- S6 - ORF0024
- S6 - ORF0045
- S6 - ORF0047
- S6 - ORF0048
- S6 - ORF0060

- S6 - ORF0064
- S6 - ORF0065
- S6 - ORF0069
- S6 - ORF0070
- S6 - ORF0071
- S6 - ORF0073
- S6 - ORF0074
- S6 - ORF0077
- S6 - ORF0079
- S6 - ORF0082

# Sequence 7 – contigs 1, 2, and 3

| | Total ORFs | Hits | No hits |
|---|---|---|---|
| MEGABLAST | 44 | 23 | 21 |
| BLASTn | 44 | 39 | 5 |

## Hits on MEGABLAST

| ORF name | Length (bp) | Organism | Accession | Pairwise identity (%) | Query coverage (%) | E-value | Protein |
|---|---|---|---|---|---|---|---|
| S7 - ORF0019 | 1023 | *Planctomycetes bacterium* | CP036526 | 74.2 | 88.95 | 2.23e-154 | EMI44140.1: methanol dehydrogenase regulatory protein |
| S7 - ORF0026 | 1758 | *Rhodopirellula baltica* | BX294135 | 76.90 | 93.17 | 0 | EMI42399.1: glutaminyl-tRNA synthetase |
| S7 - ORF0035 | 1089 | *Planctomycetes bacterium* | CP036264 | 67.40 | 87.33 | 1.01e-76 | WP_146407813.1: S-methyl-5-thioribose-1-phosphate isomerase |
| CONTIG 2 S7 - ORF005 | 426 | Cloning vector | EU140752 | 100 | 7.28*** | 7.27e-04 | NNF09180.1: tandem-95 repeat protein |
| CONTIG 2 S7 - ORF007 | 261 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| CONTIG 2 S7 - ORF009 | 120 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |

| | | | | | |
|---|---|---|---|---|---|
| CONTIG 2 S7 - ORF0010 | 99 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| CONTIG 2 S7 - ORF0011 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| CONTIG 2 S7 - ORF0012 | 975 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| CONTIG 2 S7 - ORF0014 | 1161 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| CONTIG 2 S7 - ORF0015 | 96 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| CONTIG 2 S7 - ORF0016 | 93 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| CONTIG 2 S7 - ORF0017 | 126 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| CONTIG 2 S7 - ORF0019 | 756 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| CONTIG 2 S7 - ORF0020 | 147 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| CONTIG 2 S7 - ORF0021 | 117 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| CONTIG 2 S7 - ORF0022 | 330 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CONTIG 2 S7 - ORF0023 | 660 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| CONTIG 2 S7 - ORF0028 | 1704 | *Rhodopirellula baltica* | BX294135 | 71.9 | 98.88 | 0 | WP_008686651.1: type II and III secretion system protein |
| CONTIG 3 S7-ORF 001 | 1323 | *Planctomycetes bacterium* | WP_167546599 | 47.1 | 88.89 | 1.61e-110 | mucoidy inhibitor MuiA family protein |
| CONTIG 3 S7-ORF 002 | 558 | *Rhodopirellula europaea* | WP_037251127 | 84.9 | 99.46 | 4.94e-112 | Uma2 family endonuclease |
| CONTIG 3 S7-ORF 004 | 666 | *Planctomycetaceae bacterium* | HBE70685 | 42.7 | 43.24 | 1.23e-01 | hypothetical protein [Planctomycetaceae bacterium] |
| CONTIG 3 S7-ORF 005 | 5388 | *Verrucomicrobia bacterium* | PYI87545 | 43.5 | 81.85 | 0 | hypothetical protein DME26_05830, partial |

# BLASTn search of «No hits» on MEGABLAST

| ORF name | Length (bp) | Identical proteins | Grade (%) | Number of results (out of 10) |
|---|---|---|---|---|
| S7 - ORF001 | 660 | WP_150074029: mucoidy inhibitor MuiA family protein [Rhodopirellula sp. JC645] | 73.2 | 10 |
| **S7 - ORF002** | **102** | - | | |
| S7 - ORF003 | 975 | EMI22299: secreted protein [Rhodopirellula maiorica SM1] | 68.8 | 10 |
| S7 - ORF005 | 2433 | TWU45271: Periplasmic beta-glucosidase precursor [Planctomycetes bacterium Q31b] | 91.7 | 10 |
| S7 - ORF007 | 1497 | WP_009100429: TRAP transporter large permease [Rhodopirellula sp. SWK7] | 90.2 | 10 |
| S7 - ORF009 | 462 | WP_008673152: TRAP transporter small permease [Rhodopirellula sallentina] | 83.3 | 10 |
| S7 - ORF0011 | 1071 | WP_009100433: TRAP transporter substrate-binding protein [Rhodopirellula sp. SWK7] | 85.2 | 10 |
| **S7 - ORF0015** | **105** | - | | |
| S7 - ORF0016 | 327 | WP_146593823: hypothetical protein [Planctomycetes bacterium Pla52o] | 30.7 | 1 |
| **S7 - ORF0018** | **213** | - | | |
| S7 - ORF0021 | 831 | WP_146407290: SDR family oxidoreductase | 90.2 | 10 |

| | | | | |
|---|---|---|---|---|
| | | [Planctomycetes bacterium Poly21] | | |
| S7 - ORF0023 | 1047 | WP_146407289: biotin synthase BioB [Planctomycetes bacterium Poly21] | 94.0 | 10 |
| S7 - ORF0029 | 1389 | WP_146407245: transcriptional regulator [Planctomycetes bacterium Poly21] | 83.5 | 10 |
| S7 - ORF0030 | 537 | TWT74243: bifunctional nicotinamide mononucleotide adenylyltransferase/ADP-ribose pyrophosphatase [Rhodopirellula solitaria] | 84.8 | 10 |
| S7 - ORF0032 | 1152 | EMI42754: hypothetical protein RRSWK_04943 [Rhodopirellula sp. SWK7] | 79.4 | 10 |
| **S7 - ORF0033** | **153** | - | | |
| S7 - ORF0036 | 264 | WP_146392717: hypothetical protein [Rhodopirellula solitaria] | 83.2 | 10 |
| CONTIG 2 S7 - ORF003 | 1716 | WP_081796966: tandem-95 repeat protein [Bacillus ndiopicus] | 78.7 | 10 |
| CONTIG 2 S7 - ORF0031 | 930 | WP_146392717: hypothetical protein [Rhodopirellula solitaria] | 79.2 | 10 |
| CONTIG 2 S7 - ORF0034 | 402 | WP_146579500: hypothetical protein [Planctomycetes bacterium Pla100] | 17.5 | 5 |
| **CONTIG 3 S7 - ORF003** | **201** | - | | |

## No hits on MEGABLAST or BLASTn

- S7 - ORF002

- S7 - ORF0015

- S7 - ORF0018

- S7 - ORF0033

- C3 S7- orf003

# Sequence 8

| | Total ORFs | Hits | No hits |
|---|---|---|---|
| MEGABLAST | 41 | 36 | 5 |
| BLASTn | 41 | 41 | 0 |

## Hits on MEGABLAST

| ORF name | Length (bp) | Organism | Accession | Pairwise identity (%) | Query coverage (%) | E-value | Protein |
|---|---|---|---|---|---|---|---|
| S8 - ORF001 | 2193 | *Sorangium cellulosum* | CP012672 | 72.4 | 27.63**** | 1.15e-79 | NBB72571.1: response regulator AND<br>WP_109839367.1: endo alpha-1,4 polygalactosaminidase |
| S8 - ORF004 | 2565 | *Rhodothermaceae bacterium* | CP020382 | 68.00 | 45.69 | 5.73e-97 | NNF59162.1: serine/threonine protein kinase |
| S8 - ORF005 | 1419 | *Candidatus Snodgrassella* | JQ966978 | 100 | 22.48*** | 1.16e-159 | ALG05222.2: dihydrodipicolinate synthetase |
| S8 - ORF0010 | 159 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S8 - ORF0011 | 99 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S8 - ORF0013 | 114 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S8 - ORF0014 | 114 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| S8 - ORF0017 | 105 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0019 | 126 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0023 | 972 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0028 | 1167 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0029 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0031 | 99 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0032 | 129 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0034 | 105 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0035 | 93 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0040 | 756 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0041 | 96 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0042 | 96 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0043 | 108 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0044 | 108 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0047 | 120 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |
| S8 - ORF0048 | 144 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S8 - ORF0052 | 513 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S8 - ORF0053 | 105 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S8 - ORF0054 | 105 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S8 - ORF0055 | 135 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S8 - ORF0056 | 195 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S8 - ORF0060 | 153 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S8 - ORF0063 | 2487 | *Homo sapiens* | AC275383 | 100 | 1.53**** | 7.26e-07 | WP_065918146.1: beta-propeller fold lactonase family protein |
| S8 - ORF0068 | 639 | *Azotobacter chroococcum* | CP011835 | 88.2 | 83.41 | 0 | WP_095513578.1: phosphomethylpyrimidine synthase ThiC |
| S8 - ORF0069 | 1506 | *Pseudomonas oryzae* | LT629751 | 77.10 | 81.54 | 0 | MBC14677.1: phosphomethylpyrimidine synthase ThiC |
| S8 - ORF0071 | 2895 | *Bradymonadales bacterium* | CP042468 | 73.60 | 26.46**** | 1.10e-125 | MAQ93961.1: ABC transporter substrate-binding protein**** |
| S8 - ORF0072 | 1248 | *Gemmatirosa kalamazoonesis* | CP007128 | 71.20 | 52.08 | 7.87e-79 | WP_095515937.1: serine/threonine protein kinase |
| S8 - ORF0075 | 2127 | *Gemmatirosa kalamazoonesis* | CP007128 | 71.80 | 21.72 | 1.55e-52 | HIG75890.1: sigma-70 family RNA polymerase sigma factor*** |
| S8 - ORF0080 | 1623 | *Luteitalea pratensis* | CP015136 | 68.20 | 37.71 | 4.39e-39 | HHS29952.1: HAMP domain-containing histidine kinase |

## BLASTn search of «No hits» on MEGABLAST

| ORF NAME | LENGTH (BP) | IDENTICAL PROTEINS | GRADE (%) | NUMBER OF RESULTS (OUT OF 10) |
|---|---|---|---|---|
| S8 - ORF0064 | 1923 | WP_143097359: S41 family peptidase [Myxococcus fulvus] | 58.2 | 10 |
| S8 - ORF0066 | 2973 | WP_094548663: TonB-dependent receptor [Rubricoccus marinus] | 69.9 | 10 |
| S8 - ORF0067 | 1455 | NOS85174: RNA polymerase sigma factor [Ignavibacteria bacterium] | 39.7 | 10 |
| S8 - ORF0074 | 1587 | WP_095514951: serine/threonine protein kinase [Rubrivirga sp. SAORIC476] | 70.1 | 10 |
| S8 - ORF0077 | 2310 | MYF63479: sulfatase-like hydrolase/transferase [Rhodothermaceae bacterium] | 43.3 | 10 |

## No hits on MEGABLAST or BLASTn: 0

# Sequence 9

| | Total ORFs | Hits | No hits |
|---|---|---|---|
| MEGABLAST | 40 | 24 | 16 |
| BLASTn | 40 | 39 | 1 |

## Hits on MEGABLAST

| ORF name | Length (bp) | Organism | Accession | Pairwise identity (%) | Query coverage (%) | E-value | Protein |
|---|---|---|---|---|---|---|---|
| S9 - ORF003 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S9 - ORF004 | 972 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S9 - ORF005 | 1176 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S9 - ORF008 | 756 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S9 - ORF009 | 96 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S9 - ORF0010 | 126 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S9 - ORF0011 | 120 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S9 - ORF0012 | 513 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S9 - ORF0013 | 660 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| S9 - ORF0017 | 2334 | *Granulosicoccus antarcticus* | CP018632 | 68.30 | 45.59*** | 1.23e-98 | WP_092833195.1: N,N-dimethylformamidase large subunit |
| S9 - ORF0027 | 216 | *Roseobacter litoralis* | CP002623 | 80.60 | 99.54 | 6.46e-45 | MPT24899.1: acetamidase/formamidase family protein |
| S9 - ORF0028 | 1002 | *Rhizobium sp.* | CP058351 | 73.30 | 100 | 1.06e-164 | WP_085419899.1: acetamidase/formamidase family protein |
| S9 - ORF0030 | 747 | *Labrenzia sp.* | CP045380 | 71.30 | 93.71 | 1.01e-93 | WP_180285929.1: urea ABC transporter ATP-binding protein UrtD |
| S9 - ORF0031 | 1227 | *Marinobacter hydrocarbonoclasticus* | FO203363 | 72.1 | 79.38 | 6.76e-143 | MAS04463.1: urea ABC transporter permease subunit UrtC |
| S9 - ORF0032 | 927 | *Bradyrhizobium guangdongense* | CP030051 | 71.00 | 97.52 | 8.57e-115 | MAS04462.1: urea ABC transporter permease subunit UrtB |
| S9 - ORF0035 | 1206 | *Confluentimicrobium sp.* | CP010869 | 77.70 | 94.69 | 0 | WP_153772203.1: urea ABC transporter substrate-binding protein |
| S9 - ORF0038 | 3435 | *Labrenzia sp.* | CP045380 | 67.20 | 37.73*** | 1.49e-99 | QFT33353.1: Sensory/regulatory protein RpfC |
| S9 - ORF0041 | 1803 | *Granulosicoccus antarcticus* | CP018632 | 75.30 | 99.67 | 0 | ASJ74785.1: putative oxidoreductase CzcO |
| S9 - ORF0049 | 549 | *Granulosicoccus antarcticus* | CP018632 | 77.5 | 95.63 | 8.92e-112 | WP_088919471.1: CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase |
| S9 - ORF0050 | 1854 | *Granulosicoccus antarcticus* | CP018632 | 74.80 | 97.09 | 0 | NND91485.1: excinuclease ABC subunit UvrC |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S9 - ORF0052 | 972 | *Granulosicoccus antarcticus* | CP018632 | 74.50 | 99.38 | 6.05e-174 | NND91483.1: ketoacyl-ACP synthase III |
| S9 - ORF0053 | 1038 | *Granulosicoccus antarcticus* | CP018632 | 74.20 | 96.34 | 1.25e-176 | WP_088919474.1: phosphate acyltransferase PlsX |
| S9 - ORF0055 | 201 | *Granulosicoccus antarcticus* | CP018632 | 75.80 | 96.52 | 1.14e-28 | NND91481.1: 50S ribosomal protein L32 |
| S9 - ORF0057 | 279 | *Escherichia coli* | CP055251 | 100 | 47.67 | 1.92e-59 | WP_113455136.1: terminase small subunit |

# BLASTn search of «No hits» on MEGABLAST

| **ORF name** | **Length (bp)** | **Identical proteins** | **Grade (%)** | **Number of results (out of 10)** |
|---|---|---|---|---|
| S9 - ORF0019 | 765 | WP_088920696: sulfite exporter TauE/SafE family protein [Granulosicoccus antarcticus] | 80.4 | 10 |
| S9 - ORF0020 | 597 | WP_125926863: MULTISPECIES: LEA type 2 family protein [Pseudomonas] | 65.5 | 10 |
| S9 - ORF0023 | 186 | WP_155846134: hypothetical protein [Celeribacter ethanolicus] | 46.8 | 10 |
| S9 - ORF0024 | 1395 | RZO33341: ammonium transporter [SAR116 cluster bacterium] | 86.1 | 10 |
| S9 - ORF0025 | 1818 | HIC47449: transporter substrate-binding domain- | 72.4 | 10 |

|  |  |  |  |  |
|---|---|---|---|---|
|  |  | containing protein [Methylophaga sp.] |  |  |
| S9 - ORF0026 | 225 | WP_136385948: zinc ribbon domain-containing protein [Azoarcus sp. CC-YHH848] | 49.3 | 10 |
| S9 - ORF0029 | 690 | NVK20323: urea ABC transporter ATP-binding subunit UrtE [Methylocystaceae bacterium] | 86.7 | 10 |
| S9 - ORF0039 | 918 | WP_152511548: MULTISPECIES: response regulator [unclassified Labrenzia] | 74.2 | 10 |
| S9 - ORF0043 | 1311 | NDC09553: acetamidase/formamidase family protein [Oxalobacteraceae bacterium] | 86.2 | 10 |
| S9 - ORF0044 | 498 | OED37980: hypothetical protein AB833_21610 [Chromatiales bacterium (ex Bugula neritina AB1)] | 78.3 | 10 |
| **S9 - ORF0045** | **153** | - |  |  |
| S9 - ORF0046 | 144 | WP_146366491: ABC transporter permease [Litoreibacter sp. LN3S51] | 47.7 | 10 |
| S9 - ORF0047 | 282 | WP_142905356: helix-turn-helix domain-containing protein [Exilibacterium tricleocarpae] | 87.1 | 10 |
| S9 - ORF0048 | 1260 | RLA50911: type II toxin-antitoxin system HipA family toxin [Gammaproteobacteria bacterium] | 90.6 | 10 |
| S9 - ORF0051 | 648 | WP_107940746: UvrY/SirA/GacA family response regulator | 73.4 | 10 |

| | | | | |
|---|---|---|---|---|
| | | transcription factor [Stenotrophobium rhamnosiphilum] | | |
| S9 - ORF0056 | 582 | WP_157736169: DUF177 domain-containing protein [Granulosicoccus antarcticus] | 76.6 | 10 |

## No hits on MEGABLAST or BLASTn:

- S9 - ORF0045

# Sequence 10

| | Total ORFs | Hits | No hits |
|---|---|---|---|
| MEGABLAST | 56 | 24 | 32 |
| BLASTn | 56 | 46 | 10 |

## Hits on MEGABLAST

| ORF name | Length (bp) | Organism | Accession | Pairwise identity (%) | Query coverage (%) | E-value | Protein |
|---|---|---|---|---|---|---|---|
| S10 - ORF001 | 972 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S10 - ORF002 | 1158 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S10 - ORF003 | 102 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S10 - ORF005 | 756 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S10 - ORF006 | 183 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S10 - ORF007 | 117 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S10 - ORF008 | 513 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S10 - ORF0010 | 660 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S10 - ORF0011 | 96 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S10 - ORF0012 | 624 | Uncultured bacterium | MG458673 | 93.7 | 17.63*** | 3.58e-35 | WP_007329889: flagellar basal body P-ring formation protein FlgA |
| S10 - ORF0013 | 804 | *Rhodopirellula baltica* | BX294154 | 77.5 | 99.75 | 1.42e-174 | MAP08669: flagellar basal-body rod protein FlgG [Rhodopirellula sp.] |
| S10 - ORF0026 | 2106 | *Rhodopirellula baltica* | BX294144 | 65.8 | 66.05 | 3.16e-80 | WP_007334065: CPBP family intramembrane metalloprotease [Rhodopirellula baltica] |
| S10 - ORF0035 | 381 | *Rhodopirellula baltica* | BX294151 | 81.0 | 86.88 | 5.25e-81 | WP_009099718: DUF3467 domain-containing protein |
| S10 - ORF0036 | 2238 | *Rhodopirellula baltica* | BX294151 | 77.4 | 99.91 | 0 | TWT56412: putative peptide zinc metalloprotease protein YydH [Rhodopirellula solitaria] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| S10 - ORF0040 | 2052 | *Rhodopirellula baltica* | BX294151 | 73.8 | 98.98 | 0 | WP_008679296: HlyD family efflux transporter periplasmic adaptor subunit |
| S10 - ORF0047 | 1452 | *Rhodopirellula baltica* | BX294144 | 74.1 | 99.04 | 0 | WP_083904879: RNA polymerase factor sigma-54 |
| S10 - ORF0054 | 1881 | *Rhodopirellula baltica* | BX294144 | 77.1 | 58.37** | 0 | WP_146389752: DNA polymerase III subunit gamma/tau |
| S10 - ORF0061 | 1032 | *Rhodopirellula baltica* | BX294155 | 74.6 | 70.54 | 6.89e-123 | WP_144059093: TraB/GumN family protein |
| S10 - ORF0067 | 3492 | *Rhodopirellula baltica* | BX294143 | 74.5 | 72.37 | 0 | EMI45813: transcription-repair coupling factor |
| S10 - ORF0069 | 1041 | *Rhodopirellula baltica* | BX294143 | 73.8 | 97.31 | 7.92e-173 | WP_173403143: Gfo/Idh/MocA family oxidoreductase |
| S10 - ORF0072 | 1005 | *Rhodopirellula baltica* | BX294140 | 78.4 | 99.00 | 0 | WP_044251760: sugar phosphate isomerase/epimerase |
| S10 - ORF0081 | 579 | *Rhodopirellula baltica* | BX294141 | 75.3 | 71.50 | 5.59e-64 | WP_008681921: Flp family type IVb pilin |
| S10 - ORF0098 | 159 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |
| S10 - ORF00101 | 114 | pCC2FOS fosmid vector | EU140752.1 | 100 | 100 | | |

## BLASTn search of «No hits» on MEGABLAST

| ORF name | Length (bp) | Identical proteins | Grade (%) | Number of results (out of 10) |
|---|---|---|---|---|
| S10 - ORF0014 | 753 | WP_009102573: flagellar hook basal-body protein [Rhodopirellula sp. SWK7] | 92.6 | 10 |
| S10 - ORF0017 | 465 | WP_044256526: hypothetical protein [Rhodopirellula sp. SWK7] | 92.5 | 10 |
| **S10 - ORF0019** | **96** | **-** | | |
| **S10 - ORF0020** | **159** | **-** | | |
| **S10 - ORF0021** | **120** | **-** | | |
| S10 - ORF0023 | 522 | TWT87892: Inosine-5'-monophosphate dehydrogenase [Planctomycetes bacterium Pla100] | 90.2 | 10 |
| S10 - ORF0025 | 1701 | WP_146407640: hypothetical protein [Planctomycetes bacterium Poly21] | 78.0 | 10 |
| S10 - ORF0027 | 723 | WP_044302358: ATP-binding cassette domain-containing protein [Rhodopirellula sallentina] | 93.1 | 10 |
| **S10 - ORF0028** | **93** | **-** | | |
| **S10 - ORF0031** | **93** | **-** | | |
| S10 - ORF0032 | 729 | EMI55677: protein containing DUF556 [Rhodopirellula sallentina SM41] | 72.0 | 10 |

| | | | | |
|---|---|---|---|---|
| S10 - ORF0038 | 945 | WP_085981221: efflux RND transporter periplasmic adaptor subunit [Rhodopirellula sp. SWK7] | 81.0 | 10 |
| **S10 - ORF0041** | **117** | - | | |
| S10 - ORF0043 | 516 | TWT56409: hypothetical protein CA85_42220 [Rhodopirellula solitaria] | 78.8 | 10 |
| S10 - ORF0046 | 711 | WP_146409005: alpha/beta hydrolase [Planctomycetes bacterium Poly21] | 76.7 | 10 |
| **S10 - ORF0048** | **114** | - | | |
| S10 - ORF0050 | 600 | WP_146405494: recombination protein RecR [Planctomycetes bacterium Poly21] | 96.2 | 10 |
| S10 - ORF0052 | 387 | WP_008687678: YbaB/EbfC family nucleoid-associated protein [Rhodopirellula sallentina] | 84.7 | 10 |
| S10 - ORF0055 | 648 | WP_099261283: DUF2141 domain-containing protein [Rhodopirellula bahusiensis] | 73.6 | 10 |
| S10 - ORF0060 | 963 | WP_146392977: terpene cyclase/mutase family protein [Rhodopirellula solitaria] | 86.8 | 10 |
| S10 - ORF0063 | 801 | WP_146582358: SDR family oxidoreductase | 95.3 | 10 |

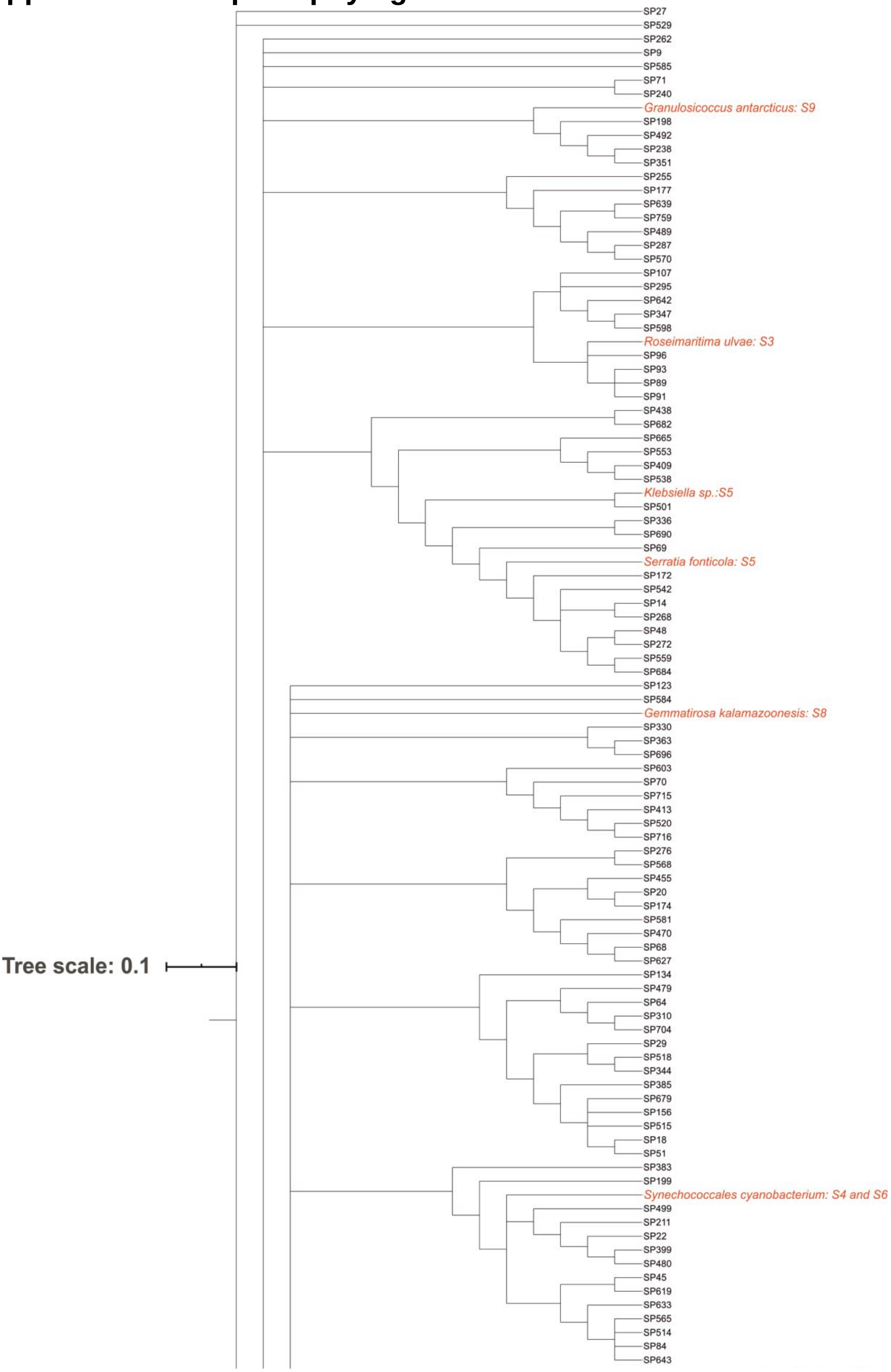
| | | | | |
|---|---|---|---|---|
| | | [Planctomycetes bacterium Pla100] | | |
| S10 - ORF0073 | 141 | WP_146392844: DUF559 domain-containing protein [Rhodopirellula solitaria] | 55.2 | 10 |
| S10 - ORF0074 | 141 | WP_146459274: DUF559 domain-containing protein [Rubripirellula tenax] | 38.6 | 1 |
| **S10 - ORF0075** | **129** | - | | |
| S10 - ORF0078 | 543 | WP_008681919: hypothetical protein [Rhodopirellula sallentina] | 69.3 | 10 |
| S10 - ORF0083 | 111 | KLU06522: putative transmembrane protein [Rhodopirellula islandica] | 59.8 | 10 |
| S10 - ORF0086 | 558 | WP_009100162: prepilin peptidase [Rhodopirellula sp. SWK7] | 95.7 | 10 |
| **S10 - ORF0089** | **135** | - | | |
| S10 - ORF0091 | 1089 | WP_044254922: Flp pilus assembly protein CpaB [Rhodopirellula sp. SWK7] | 86.3 | 10 |
| S10 - ORF0094 | 1785 | WP_009100158: pilus assembly protein N-terminal domain-containing protein [Rhodopirellula sp. SWK7] | 84.0 | 10 |
| **S10 - ORF0096** | **111** | - | | |

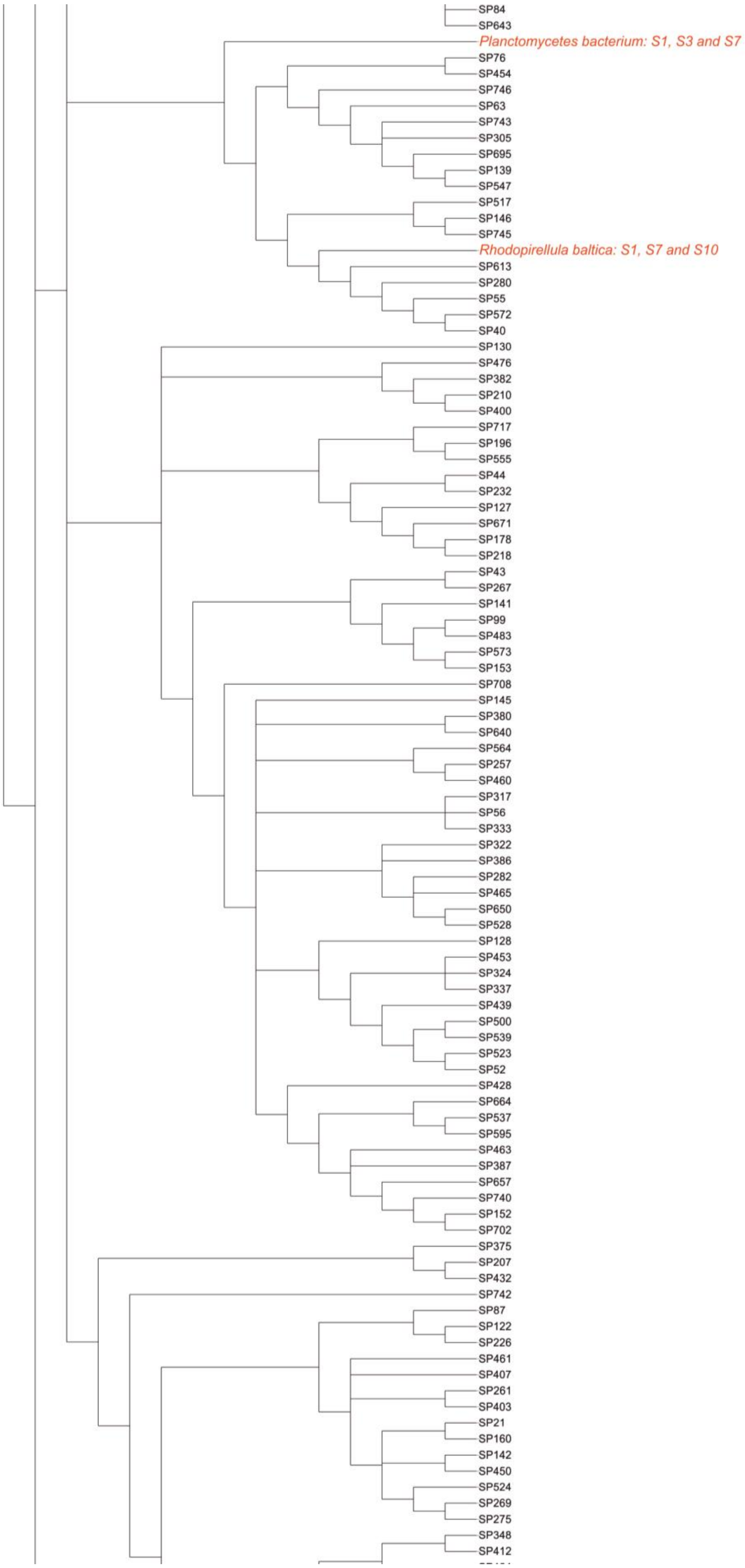
| | | | | |
|---|---|---|---|---|
| S10 - ORF0097 | 561 | WP_009100154: MinD/ParA family protein [Rhodopirellula sp. SWK7] | 90.2 | 10 |

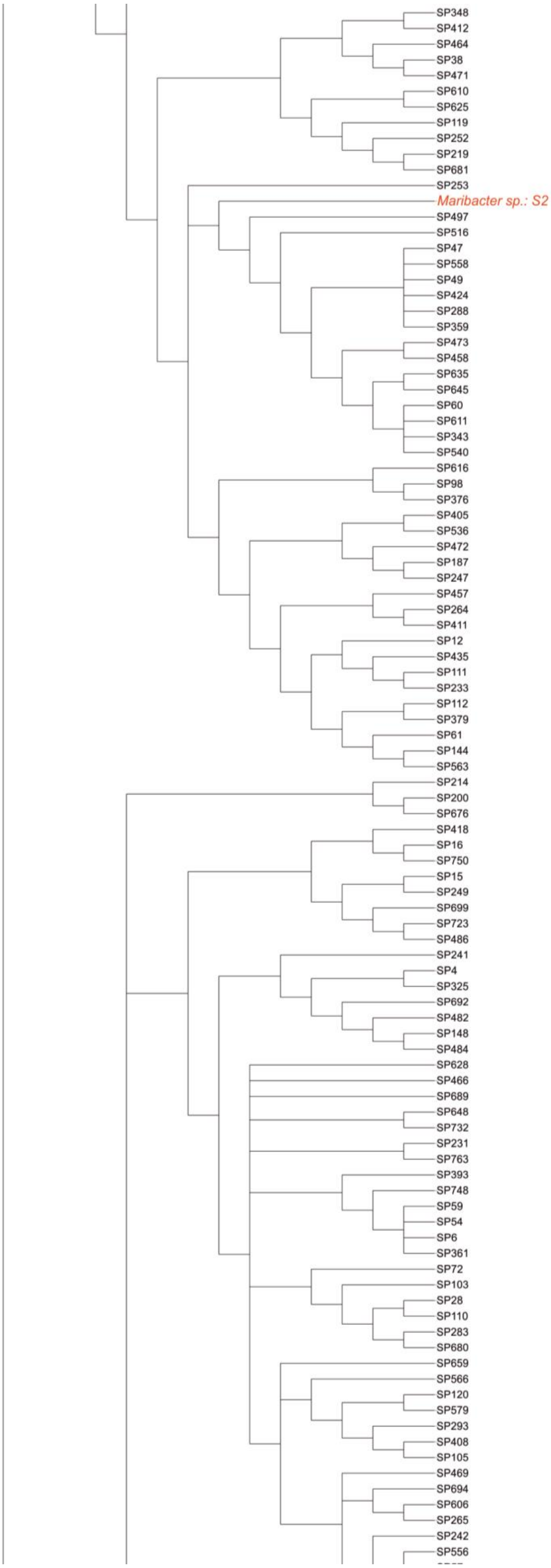
# No hits on MEGABLAST or BLASTn

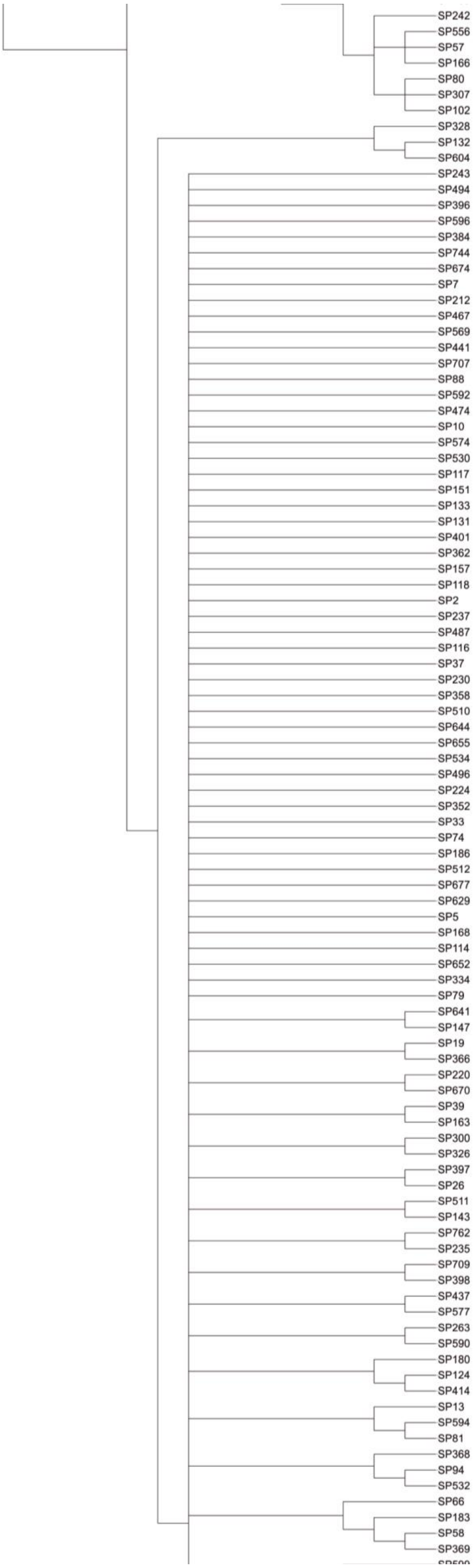
- S10 - ORF0019
- S10 - ORF0020
- S10 - ORF0021
- S10 - ORF0028
- S10 - ORF0031
- S10 - ORF0041
- S10 - ORF0048
- S10 - ORF0075
- S10 - ORF0089
- S10 - ORF0096

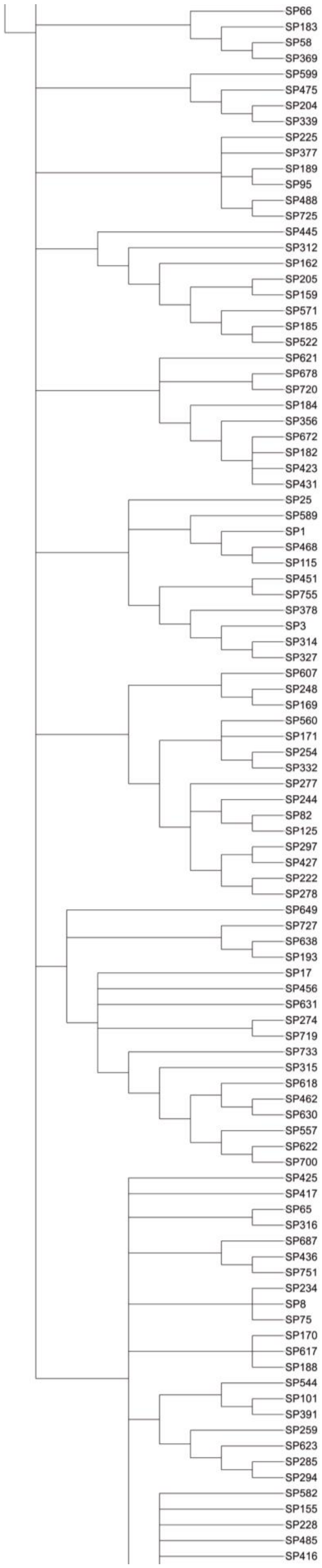
# Appendix 7: Complete phylogenetic tree

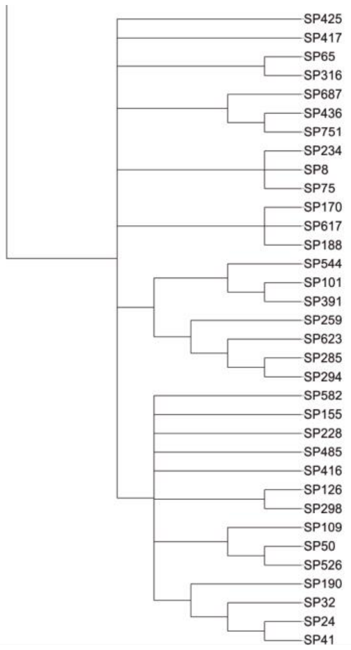



SP84
SP643
Planctomycetes bacterium: S1, S3 and S7
SP76
SP454
SP746
SP63
SP743
SP305
SP695
SP139
SP547
SP517
SP146
SP745
Rhodopirellula baltica: S1, S7 and S10
SP613
SP280
SP55
SP572
SP40
SP130
SP476
SP382
SP210
SP400
SP717
SP196
SP555
SP44
SP232
SP127
SP671
SP178
SP218
SP43
SP267
SP141
SP99
SP483
SP573
SP153
SP708
SP145
SP380
SP640
SP564
SP257
SP460
SP317
SP56
SP333
SP322
SP386
SP282
SP465
SP650
SP528
SP128
SP453
SP324
SP337
SP439
SP500
SP539
SP523
SP52
SP428
SP664
SP537
SP595
SP463
SP387
SP657
SP740
SP152
SP702
SP375
SP207
SP432
SP742
SP87
SP122
SP226
SP461
SP407
SP261
SP403
SP21
SP160
SP142
SP450
SP524
SP269
SP275
SP348
SP412


SP348
SP412
SP464
SP38
SP471
SP610
SP625
SP119
SP252
SP219
SP681
SP253
Maribacter sp.: S2
SP497
SP516
SP47
SP558
SP49
SP424
SP288
SP359
SP473
SP458
SP635
SP645
SP60
SP611
SP343
SP540
SP616
SP98
SP376
SP405
SP536
SP472
SP187
SP247
SP457
SP264
SP411
SP12
SP435
SP111
SP233
SP112
SP379
SP61
SP144
SP563
SP214
SP200
SP676
SP418
SP16
SP750
SP15
SP249
SP699
SP723
SP486
SP241
SP4
SP325
SP692
SP482
SP148
SP484
SP628
SP466
SP689
SP648
SP732
SP231
SP763
SP393
SP748
SP59
SP54
SP6
SP361
SP72
SP103
SP28
SP110
SP283
SP680
SP659
SP566
SP120
SP579
SP293
SP408
SP105
SP469
SP694
SP606
SP265
SP242
SP556


SP242
SP556
SP57
SP166
SP80
SP307
SP102
SP328
SP132
SP604
SP243
SP494
SP396
SP596
SP384
SP744
SP674
SP7
SP212
SP467
SP569
SP441
SP707
SP88
SP592
SP474
SP10
SP574
SP530
SP117
SP151
SP133
SP131
SP401
SP362
SP157
SP118
SP2
SP237
SP487
SP116
SP37
SP230
SP358
SP510
SP644
SP655
SP534
SP496
SP224
SP352
SP33
SP74
SP186
SP512
SP677
SP629
SP5
SP168
SP114
SP652
SP334
SP79
SP641
SP147
SP19
SP366
SP220
SP670
SP39
SP163
SP300
SP326
SP397
SP26
SP511
SP143
SP762
SP235
SP709
SP398
SP437
SP577
SP263
SP590
SP180
SP124
SP414
SP13
SP594
SP81
SP368
SP94
SP532
SP66
SP183
SP58
SP369


SP66
SP183
SP58
SP369
SP599
SP475
SP204
SP339
SP225
SP377
SP189
SP95
SP488
SP725
SP445
SP312
SP162
SP205
SP159
SP571
SP185
SP522
SP621
SP678
SP720
SP184
SP356
SP672
SP182
SP423
SP431
SP25
SP589
SP1
SP468
SP115
SP451
SP755
SP378
SP3
SP314
SP327
SP607
SP248
SP169
SP560
SP171
SP254
SP332
SP277
SP244
SP82
SP125
SP297
SP427
SP222
SP278
SP649
SP727
SP638
SP193
SP17
SP456
SP631
SP274
SP719
SP733
SP315
SP618
SP462
SP630
SP557
SP622
SP700
SP425
SP417
SP65
SP316
SP687
SP436
SP751
SP234
SP8
SP75
SP170
SP617
SP188
SP544
SP101
SP391
SP259
SP623
SP285
SP294
SP582
SP155
SP228
SP485
SP416


SP425
SP417
SP65
SP316
SP687
SP436
SP751
SP234
SP8
SP75
SP170
SP617
SP188
SP544
SP101
SP391
SP259
SP623
SP285
SP294
SP582
SP155
SP228
SP485
SP416
SP126
SP298
SP109
SP50
SP526
SP190
SP32
SP24
SP41