Insights into the biology of Candidate Division OP3 LiM populations

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
- Dr. rer. nat. -

dem Fachbereich Biologie/Chemie der
Universität Bremen
vorgelegt von

Jana Kizina

Bremen, August 2017

Diese vorliegende Arbeit wurde in der Zeit von Juli 2013 bis August 2017 im Rahmen des Programmes „The International Max Planck Research School of Marine Microbiology“ (MarMic) in der Abteilung Mikrobiologie am Max-Planck-Institut für Marine Mikrobiologie in Bremen angefertigt.

Gutachter: Prof. Dr. Jens Harder
Gutachter: Prof. Dr. Rudolf Amann

Prüfer: Prof. Dr. Ulrich Fischer
Prüfer: Prof. Dr. Karl-Heinz Blotevogel

Tag des Promotionskolloquiums: 18. September 2017

Dedicated to my family

“If you can dream it, you can do it.”

– Walt Disney

Summary

The candidate division OP3, recently entitled *candidate phylum* Omnitrophica, is characterized by 16S rRNA gene sequences from a broad range of anoxic habitats with a broad phylogeny of up to 26% sequence dissimilarity. The 16S rRNA phylotype OP3 LiM had previously been detected in limonene-degrading, methanogenic enrichment cultures and represented small coccoid cells. Neither isolation experiments nor physiological experiments had provided insights into the metabolism of this bacterium within the complex methanogenic community. This doctoral thesis aimed at the characterization of populations of the phylotype OP3 LiM to discover its biology.

Metagenomes usually yield draft population genomes. To obtain the complete closed OP3 LiM genome, *in silico* methods were explored to improve draft assemblies. Large genomes of planctomycete strains were assembled with a variety of methods. A taxonomic classification of contig sequences was used to differentiate and separate contigs of draft assemblies into taxon-specific groups. Reassemblies of reads obtaining from mapping onto taxon-specific contigs yielded improved draft assemblies. This knowledge was used to obtain a closed genome of OP3 LiM from a metagenome of physically enriched OP3 LiM cells. Finishing the OP3 LiM genome required the combination of data of different sequencing technologies, a variety of assembly and mapping software, over 15 reassemblies with intensive manual quality controls by read and contig mapping and, finally, laboratory work with combinatorial PCR to solve the genome puzzle.

The population genome of OP3 LiM is the first closed genome of a member of *candidate phylum* Omnitrophica and comprises 1,974,501 bp with a GC content of 52.9%. Its 23S rRNA contains a group I intron. The genome offers a syntrophic life on hydrogen or formate, however, the metaproteome indicated that OP3 LiM uses glycolysis together with pyruvate oxidation as major catabolic pathway. The metaproteome also identified high levels of proteins potentially involved in the degradation of polymers as well as in the uptake of foreign nucleic acids. The genomic information was combined with observations of cells of the methanogenic community by different visualization methods.

Images of OP3 LiM required electron microscopy due to the small cell size of 0.2–0.3 µm in diameter. *In situ* hybridizations revealed two physiological stages, free-living OP3 LiM cells with low ribosome content and OP3 LiM cells attached to either bacteria or

archaea, which showed strong signals. This observation indicated a higher metabolic activity of OP3 LiM cells during the attachment and, likewise, that the bacterium utilizes surface polysaccharides as preferred substrate.

In situ hybridizations revealed that the methanogen *Methanosaeta* in the enrichment culture contained cells in the filaments that lacked DNA and rRNA suggesting that these cells lost their cellular content. We also observed faint signals of the OP3 LiM 16S rRNA in *Methanosaeta* cells. The presence of the intron RNA of the 23S rRNA of OP3 LiM was visualized in *Methanosaeta* cells devoid of DNA and rRNA. This first direct observation of an intron transfer from a bacterium to an archaeon together with metaproteomic observations indicate the lifestyle of an epibiotic bacterium for OP3 LiM. OP3 LiM is the first predatory bacterium that preys on *Archaea*. We propose to name OP3 LiM "*Candidatus* Vampirococcus archaeovorus".

Zusammenfassung

Die Kandidaten-Division OP3, neuerdings *Kandidat Phylum* Omnitrophica benannt, ist gekennzeichnet durch 16S rRNA Gensequenzen, die aus einer großen Bandbreite anoxischer Lebensräume gewonnen wurden und mit Sequenzunterschieden von bis zu 26% eine breite Phylogenie aufzeigen. Der 16S rRNA Phylotyp OP3 LiM wurde in Limonen-abbauenden methanogenen Anreicherungen nachgewiesen und ist repräsentiert durch kleine kokkoidförmige Zellen. Weder Isolationsexperimente noch physiologische Tests hatten Einblicke in den Metabolismus dieses Bakteriums innerhalb der komplexen methanogenen Gemeinschaft gegeben. Das Ziel dieser Doktorarbeit bestand in der Charakterisierung von Populationen des Phylotyps OP3 LiM, um seine Biologie aufzudecken.

Aus Metagenomen lassen sich in der Regel unvollständige „Entwurf"-Populationsgenome gewinnen. Um das vollständig geschlossene OP3 LiM Genom zu erhalten, wurden *in silico* Methoden zur Verbesserung von unvollständigen Genomen erkundet. Große Genome von Planctomycetenstämmen wurden mit einer Vielzahl von Methoden aus Sequenzen zusammengebaut. Eine taxonomische Klassifikation von Contig-Sequenzen (langen Genomstücken) wurde verwendet, um die Contigs der Genome in taxonomisch-klassifizierte Gruppen zu differenzieren und zu trennen. Sequenzen, die durch den Anlagerungsprozess („Mapping") auf taxonomisch-spezifische Contigs ausgewählt wurden, ergaben beim Genomzusammenbau verbesserte Contigs. Diese Erkenntnis wurde verwendet, um ein geschlossenes Genom von OP3 LiM aus einem Metagenom von physikalisch-angereicherten OP3 LiM Zellen zu gewinnen. Die Fertigstellung des OP3-LiM-Genoms erforderte die Kombination von Daten verschiedener Sequenzierungstechnologien, eine Vielzahl von Assemblierungs- und Mapping-Software, die Durchführung von über 15 Reassemblierungen zusammen mit intensiver manueller Qualitätskontrolle, welche durch Mapping-Verfahren der DNA-Bruchstücke und der Contigs ermöglicht wurde, und schließlich Laborarbeit mit kombinatorischer PCR für die Lösung des Genompuzzles.

Das Populationsgenom von OP3 LiM ist das erste geschlossene Genom eines Mitglieds des *Kandidat Phylums* Omnitrophica und umfasst 1.974.501 Basenpaare mit einem GC Gehalt von 52,9%. Seine 23S rRNA beinhaltet ein Gruppe-I-Intron. Das Genom ermöglicht ein syntrophes Leben auf Wasserstoff oder Formiat, jedoch wies das Metaproteom darauf hin, dass OP3 LiM hauptsächlich die Glykolyse zusammen mit der Pyruvat-Oxidation

katabol nutzt. Das Metaproteom identifizierte auch eine große Anzahl an in der Zelle häufig vorkommenden Proteinen, die potentiell an dem Polymerabbau sowie an der Aufnahme von fremder Nukleinsäure beteiligt sind. Die Genominformationen wurden mit Beobachtungen, die durch die Anwendung verschiedener Bildgebungsmethoden von den Zellen der methanogenen Gemeinschaft gemacht werden konnten, kombiniert.

Die visuelle Darstellung von OP3 LiM erforderte Elektronenmikroskopie, was auf die geringe Zellgröße von 0,2–0,3 µm im Durchmesser zurückzuführen ist. *In situ* Hybridisierungen zeigten zwei physiologische Zustände, frei-lebende OP3 LiM Zellen mit geringem Ribosomengehalt und OP3 LiM Zellen, die an Bakterien beziehungsweise Archaeen anhafteten und starke Signale aufwiesen. Diese Beobachtung deutete auf eine höhere metabolische Aktivität von OP3 LiM Zellen während der Anlagerung hin und ebenso, dass das Bakterium Oberflächen-Polysaccharide als bevorzugtes Substrat verwendet.

In situ Hybridisierungen offenbarten, dass der in der Anreicherungskultur lebende Methanbildner *Methanosaeta* Zellen in den Filamenten besitzt, denen DNA und rRNA fehlen, was auf den Verlust ihres Zellinhaltes deutet. Wir beobachteten auch schwache OP3 LiM 16S rRNA Signale in *Methanosaeta* Zellen. Die Anwesenheit der RNA des 23S rRNA Introns von OP3 LiM wurde in *Methanosaeta* Zellen, die weder DNA und rRNA aufwiesen, sichtbar gemacht. Diese erste direkte Beobachtung eines Introntransfers von einem Bakterium zu einem Archaeon, zusammen mit Beobachtungen aus dem Metaproteom, deutet auf den Lebensstil eines Epibionten für OP3 LiM hin. OP3 LiM ist das erste räuberische Bakterium, welches auf Archaeen Jagd macht. Wir schlagen den Namen "*Candidatus* Vampirococcus archaeovorus" vor.

Table of contents

Abbreviations of chapters

ANI	average nucleotide identity
BALO	*Bdellovibrio*-and-like organisms
BLAST	Basic Local Alignment Search Tool
bp	base pairs
BWA	Burrows-Wheeler Aligner
CARD	catalyzed reporter deposition
CAZymes	carbohydrate-active enzymes
CLSM	confocal laser scanning microscope
CoA	coenzyme A
CRISPR	clustered regularly interspaced short palindromic repeats
cryo-TEM	cryogenetic TEM
DAPI	4’,6-diamidin-2-phenylindol
DNA	deoxyribonucleic acid
dsDNA	double-stranded DNA
EDTA	ethylenediaminetetraacetic acid
FA	formamide
FACS	fluorescence-activated cell sorting
FDR	false discovery rate
FISH	fluorescence *in situ* hybridization
GSP	general secretory pathway
HCR	hybridization chain reaction
HGT	horizontal gene transfer
HMN	2,2,4,4,6,8,8-heptamethylnonane
HRP	horseradish peroxidase
IMG	Integrated Microbial Genomes
IMG/M	Integrated Microbial Genomes & Microbiomes
IVS	intervening sequence
LPS	lipopolysaccharide
Mb	mega base pairs
MIxS	Minimum Information about any (x) Sequence

NER	nucleotide excision repair
NGS	next-generation sequencing
NSAF	normalized spectral abundance factor
OD	optical density
OP	Obsidian Pool
ORF	open reading frame
ori	origin of replication
OTU	operational taxonomic unit
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PVC	*Planctomycetes*/*Verrucomicrobia*/*Chlamydiae*
RAST	Rapid Annotations using Subsystems Technology
rDNA	ribosomal deoxyribonucleic acid
RM	restriction modification
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
RT	room temperature
S	Svedberg
SEM	scanning electron microscopy
SDS	sodium dodecyl sulfate
SGS	second-generation sequencing
SMART	single molecule real time
ssDNA	single-stranded DNA
SSU	small subunit
TAT	twin-arginine translocation
TE	Tris and EDTA
TEM	transmission electron microscopy
TGS	third-generation sequencing
tRNA	transfer ribonucleic acid
TSC	total spectral count
USD	United States Dollar

Chapter 1

Introduction

1.1 The microbial diversity

Although invisible to the naked eye, prokaryotes constitute an essential component of the Earth´s biota (Whitman *et al.*, 1998). The prokaryotic diversity is a product of about 3.8 billion years of evolution – two billion years longer than that of eukaryotic organisms (Torsvik *et al.*, 2002), and they have continued to evolve to occupy every possible metabolic niche (von Wintzingerode *et al.*, 1997; Rinke *et al.*, 2013). 4–6 x 10^{30} bacterial and archaeal cells are estimated to inhabit the Earth (Whitman *et al.*, 1998) and outnumber the eukaryotic cells by several order of magnitudes (Schleifer, 2004). The majority of the prokaryotes occur in the open ocean (1.2 x 10^{29} cells), in soil (2.6 x 10^{29} cells), and in oceanic and terrestrial subsurfaces (3.5 x 10^{30} and 0.25–2.5 x 10^{30} cells) (Whitman *et al.*, 1998). The total organic carbon of prokaryotes on earth is enormous, approximately 60% to 100% of the total organic carbon found in plants (Whitman *et al.*, 1998). In addition, prokaryotes contain large amounts of N, P, and other essential nutrients (Whitman *et al.*, 1998). *Bacteria* and *Archaea* possess an immense metabolic diversity, and their activities are critical in processes ranging from sewage treatment to regulating the composition of the atmosphere (Ward, 2002). They are a crucial component of the biosphere because they catalyze biogeochemical cycles sustaining all life on Earth (Torsvik *et al.*, 2002). Therefore, discovering and understanding the diversity of microbial communities (the number of species and their relative abundances) is a high priority in ecology (Ward, 2002).

Although we are aware of the central role of microbes in in biotic processes, we have known very little about their actual diversity (Schleifer, 2004). Estimating the microbial diversity is a persisting challenge by the fact that the development of a reliable classification based on morphological traits as these for higher eukaryotes has been difficult due to their relative simplicity (Schloss & Handelsman, 2004; Prakash *et al.*, 2007). For the recognition of prokaryotic species, the isolation of an organism in pure culture is an indispensable requisite (Rosselló-Mora & Amann, 2001). Currently, 14,800 species are validly published (http://www.dsmz.de/bacterial-diversity/prokaryotic-nomenclature-up-to-date). Molecular

techniques indicate that the hitherto classified prokaryotes account a small portion of the real microbial diversity (Rosselló-Mora & Amann, 2001). More than 99% of the global population of prokaryotes has yet not been isolated by using standard techniques (Hugenholtz *et al.*, 1998b; Schleifer, 2004).

Consequently, cultivation-independent methods, based mainly on molecular retrieval of 16S rRNA gene (rDNA) sequences, have become the most important tools for detection and identification of microbes (Derakshani *et al.*, 2001). Since Woese and Fox (1977) first proposed the 16S rRNA gene as a phylogenetic tool to describe the evolutionary relationships among organisms and Pace *et al.* (1985) described its use for classifying uncultivable microorganisms in the environment (Schloss & Handelsman, 2004), the perspective on microbial diversity has improved enormously (Hugenholtz *et al.*, 1998b).

Comparative sequence analysis of small subunit ribosomal ribonucleic acids (SSU rRNA; 16S rRNA in prokaryotes, 18S rRNA in eukaryotes) indicates that the living organisms comprise at least three primary domains, *Bacteria*, *Archaea*, and *Eucarya* (Schleifer, 2004). In addition, the use of the rRNA approach in exploring uncultured prokaryotes in natural samples has given valuable insights into prokaryotic diversity (Rosselló-Mora & Amann, 2001). The usage of new tools unveiled a widespread distribution and unexpected diversity by the discovery of new uncultured linages of *Archaea* (Auguet *et al.*, 2010). Contrary to previous belief that species of this domain inhabited only the extreme environments (e.g., extreme, hot, saline or strictly anaerobic habitats) (Amann, 2000), a high numbers of novel and unexpected "non-extreme" archaeal phenotypes were discovered in habitats like soil, seawater, etc. (Sharma *et al.,* 2005). Many of the uncultivated phyla are found in diverse habitats, and some are extraordinarily abundant (Rappé & Giovannoni, 2003). Today, 2 million nearly complete sequences are known on the 16S rRNA level, thereof 650,000 with less than 99% identity to other 16S rRNA gene sequences (SILVA database, release 128, Sept. 2016).

Over the past decades the species census estimates have changed, but remains still controversial (Amann & Rosselló-Móra, 2016). The number of bacterial species in the world was previously estimated to range from 10^7 to 10^9 (Schloss & Handelsman, 2004). According to a new estimate, which based on the combination of scaling laws with a model of biodiversity, there are even about one trillion (10^{12}) microbial species on Earth, and 99.999% of them have yet to be discovered (Locey & Lennon, 2016). The study of Schloss *et al.* (2016) provides evidence that the PCR-based census of species might stop at a few millions (Amann & Rosselló-Móra, 2016).

The microorganism diversity is hidden in the "rare biosphere". Yarza *et al.* (2014) revealed that there are likely more than 1,000 phylum-level clades that lack any cultured representatives (Amann & Rosselló-Móra, 2016). This "unculturable" bacterial diversity presents a vast gene pool for biotechnological exploitation and poses a major challenge for microbiologists to understand their phylogenetic relationship and ecological significances (Sharma *et al.*, 2005).

Sequencing of bacterial genome sequences is now a standard procedure (Land *et al.*, 2015). However, because the majority of microbes are "unculturable", classical methods for cloning and sequencing the entire genome of every species in a population are unsuitable. This is reflected in the biased representation of SSU rRNA-based phylogeny according to which more than 88% of all microbial strains belong to only four bacterial phyla, the *Proteobacteria*, *Firmicutes*, *Actinobacteria*, and *Bacteroidetes* (Rinke *et al.*, 2013).

Metagenomics technologies relying on environmental shotgun sequencing usually offer the possibility of sampling from a community and investigating microbes in their natural environment (Pachter, 2007). They have become an indispensable tool for studying the diversity and metabolic potential of environmental microbes, whose bulk is as yet non-cultivable (Teeling & Glöckner, 2012). Due to new genomic sampling of previously enigmatic or unknown microbial lineages, the tree of life has dramatically expanded revealing that large fraction of diversity is currently only accessible via cultured-independent genome-resolved approaches (Hug *et al.*, 2016).

But the ever-increasing amount and complexity of generated sequences has also large implications for analysis of this data. Therefore, the ability of bioinformatics to analyze, compare, interpret, and visualize the vast increase in bacterial genomes, transcriptomes, proteomes, and metatranscriptomes is valiantly trying to keep up with these developments (Paulino *et al.*, 2015). New software packages are available for this problem (e.g., Metawatt (Strous *et al.*, 2012), CONCOCT (Alneberg *et al.*, 2014), GroopM (Imelfort *et al.*, 2014) or Anvi'o (Eren *et al.*, 2015). Nevertheless, it is likely that there will be a continued demand for good bioinformatics tools. The reconstruction of a complete and fully automated assembly of genomes using next-generation sequencing (NGS) short read sequences remains a significant bioinformatics challenge (Paulino *et al.*, 2015). The technology of long-read sequencing now offers different alternatives to solve these genome assembly problems (e.g., repeated elements or segmental duplications involved in complex regions) (Madoui *et al.*, 2015).

The number of sequenced genomes has continued to increase dramatically in the last 10 years. The Integrated Microbial Genomes & Microbiomes (IMG/M) system contains a

total of 47,516 (among them 40,894 public) archaeal, bacterial and eukaryotic genomes currently available in July 2016 (Chen *et al.*, 2017). These genomic data represent an over 300% increase since September 2013 (Markowitz *et al.*, 2014). In addition, IMG also includes 5,185 (3,907 public) viral genomes, 1,220 (1,192 public) plasmids and 1,196 (1,192 public) genome fragments (Chen *et al.*, 2017). This information has expanded the catalog of microbial taxa by orders of magnitude, and has had a major impact on the understanding of microbial diversity (Locey & Lennon, 2016). In summary, the last decade has seen an explosion of sequence information from the environment, and it is time to study especially the so-far not characterized phyla in more detail.

1.2 Candidate Division OP3

The candidate division OP3 was first recognized in a culture-independent molecular phylogenetic survey carried out on the bacterial community in the Obsidian Pool (OP), a hot spring (75°C to 95°C) located in the Yellowstone National Park (Fig. 1). This unique ecosystem, rich in iron, sulfide, carbon dioxide and hydrogen, is dominated by lithotrophic communities and a fertile ground for discovery of novel microbial diversity (Hugenholtz *et al.*, 1998a; Kumar & Saravanan, 2010). In the study by Hugenholtz *et al.* (1998a), 12 novel division level bacterial lineages were detected in sediment on the basis of 16S rRNA gene sequences (Wagner & Horn, 2006; Glöckner *et al.*, 2010). One of these candidate divisions was named Obsidian Pool 3 (OP3) (Hugenholtz *et al.*, 1998a).

Fig. 1. Obsidian Pool, The Yellowstone National Park (from Kumar & Saravanan, 2010).

Phylogenetic analysis of 16S rRNA genes assigned OP3 to the *Planctomycetes*/*Verrucomicrobia*/*Chlamydiae* (PVC) superphylum. This assemblage is

formed by four bacteria phyla, comprising the *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* and *Lentisphaerae*, together with the candidate divisions *Poribacteria* and OP3 (Fig. 2) (Wagner & Horn, 2006).

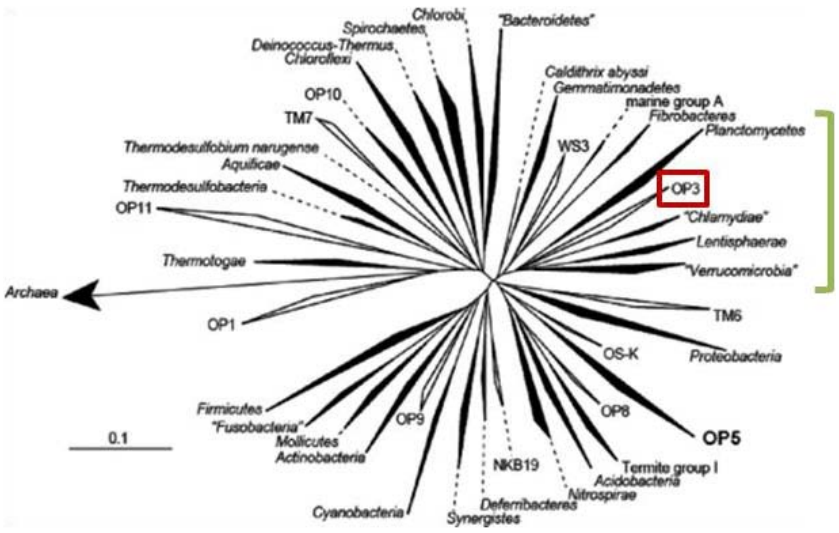


Fig. 2. Phylogenetic relationship based on 16S rRNA gene sequences showing major recognized bacterial phyla. Solid wedges indicate phyla with cultured representatives. Open wedges indicate candidate phyla. OP3 is highlighted with red frame. PVC superphylum is highlighted with green colour range (adapted from Kumar & Saravanan, 2010).

On basis of their monophyletic cluster in 16S rRNA-based phylogenetic trees, a common ancestor was speculated. Despite this proposal, the members of the PVC superphylum differ greatly within the superphylum with respect to the life-style, physiology, and ecology. Microorganisms of this group possess dramatically different lifestyles and colonize sharply contrasting habitats (Wagner & Horn, 2006). Each phylum includes members that attracted significant research interest (Lagkouvardos *et al.*, 2014). Members are important in carbon and nitrogen cycling (e.g., *Rhodopirellua* and anammox bacteria, such as "*Candidatus* Kuenenia" species) (Gupta *et al.*, 2012; Lagkouvardos *et al.*, 2014), as pathogens or symbionts (e.g., *Chlamydia* species), or as environmental microbes in soil and marine habitats (e.g., *Verrumicrobia*) (Fuerst, 2013; Lagkouvardos *et al.*, 2014).

The initial description of OP3 was based on a single 16S rRNA gene sequence retrieved from OP sediment (Hugenholtz *et al.*, 1998a). However, soon a diverse range of habitat was identified for members of the OP3 phylum based on the presence of 16S rRNA sequences in clone libraries of environmental samples. Most of these sequences were obtained from anoxic environments including various marine habitats, such as water column (Madrid *et*

al., 2001), hydrothermal fluids (Pisapia *et al.*, 2017), or sediments (Stott *et al.*, 2008; Kolinko *et al.*, 2012), hypersaline wastewater (Lefebvre *et al.,* 2006), wetlands (Dedysh, 2011), flooded paddy soil (Derakshani *et al.*, 2001), and waste water treatment plants and methanogenic bioreactors (Chouari *et al.*, 2005). In example (Fig. 3), OP3 was observed in the microbial community present in monoterpene degrading methanogenic enrichment cultures, a complex syntrophic community (Rotaru *et al.*, 2012).

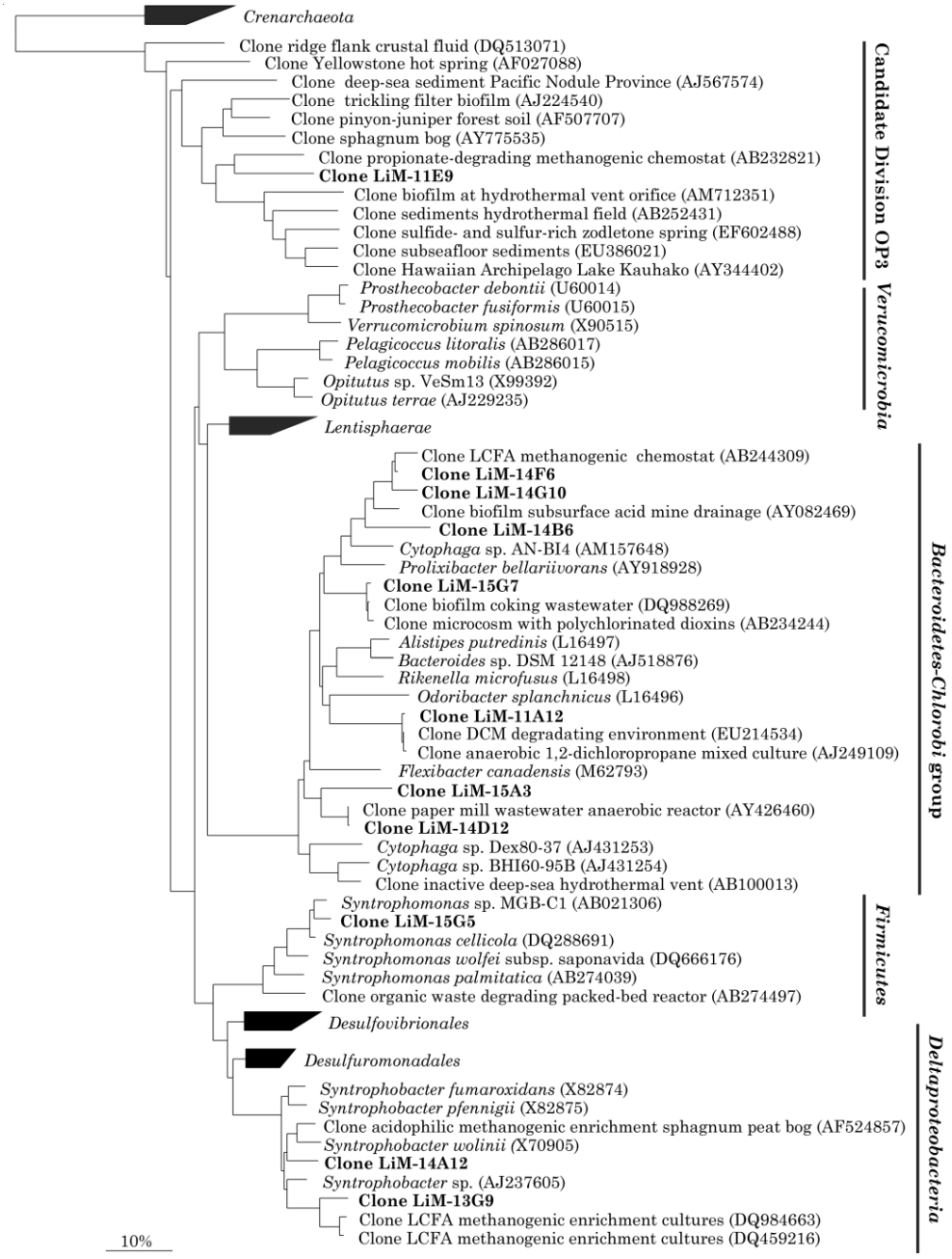


Fig. 3. Maximum parsimony tree of *Bacteria* 16S rRNA gene sequences retrieved from limonene degrading methanogenic enrichment cultures (from Rotaru *et al.*, 2012).

First phylogenetic analysis of all available OP3-related 16S rRNA gene sequences (SILVA database, release 100, checked November 2009) had shown that OP3 is represented by at least five different subdivision level lineages (Glöckner *et al.*, 2010). At least members of one of these sub-lineages (sub-lineage V) may possess a facultative anaerobic lifestyle (Ivanova & Dedysh, 2012) based on OP3-related bacteria recently detected in both oxic and anoxic peat layers from acidic wetlands (Ivanova & Dedysh, 2012). Currently, the SILVA database contains 2,960 *candidate phylum* Omnitrophica (*Omnitrophica*, renamed by Rinke *et al.* (2013)) SSU sequences, all defined as uncultured and obtained from diverse habitats (SILVA database; release 128 of September 2016; https://www.arb-silva.de).

16S rRNA gene sequences, recovered from a wide range of anoxic habitats, revealed a phylogenetical broad phylum of OP3. First genomic insights into the biology of the OP3 bacteria were obtained from metagenomic studies including the analysis of large metagenomic cosmid clones (Glöckner *et al.*, 2010), single cell genomes (Kolinko *et al.*, 2012; Rinke *et al.*, 2013), and metagenomic population draft genomes (Speth *et al.*, 2016). Currently, 87 population genome bins / draft genomes for *candidate phylum* Omnitrophica, obtained from diverse habitats, are available from GenBank (June 2017; https://www.ncbi.nlm.nih.gov) (Appendix Table S1).

Metagenomic cosmids and draft assemblies identified OP3 genes potentially involved in anaerobic respiration (Glöckner *et al.*, 2010). The potential of nitrate respiration was described for OP3. A gene encoding a nitrate reductase was identified in the most complete OP3 draft genome (assembly accession number ASM156711v1) to date, retrieved from sludge of a wastewater treatment plant (Speth *et al.*, 2016). OP3 is capable of fixing carbon via the reductive acetyl-CoA cycle (Wood-Ljungdahl pathway) (Rinke *et al.*, 2013). OP3 genomes share a high proportion of orthologal genes with members of the *Deltaproteobacteria*, rather than with those of the PVC superphylum (Glöckner *et al.*, 2010). One OP3 genome has genes for magnetotactic behavior (Kolinko *et al.,* 2012), a representation of the high diversity of OP3.

The genome-based information was also used in the construction of phylogenetic trees. The limited resolution of a single gene, such as the widely used 16S rRNA gene, was recently replaced by a set of genes with less degree of conservation (e.g., a set of ribosomal protein sequences) (Hug *et al.*, 2016). This should reduce the uncertainty of the branching in the trees. The concatenated genes placed the candidate division OP3 at the root of the PVC superphylum, before the division of *Chlamydiae* and *Planctomycetes* in maximum likelihood analyses (Fig. 4) (Rinke *et al.*, 2013; Anantharaman *et al.*, 2016; Hug *et al.*, 2016).

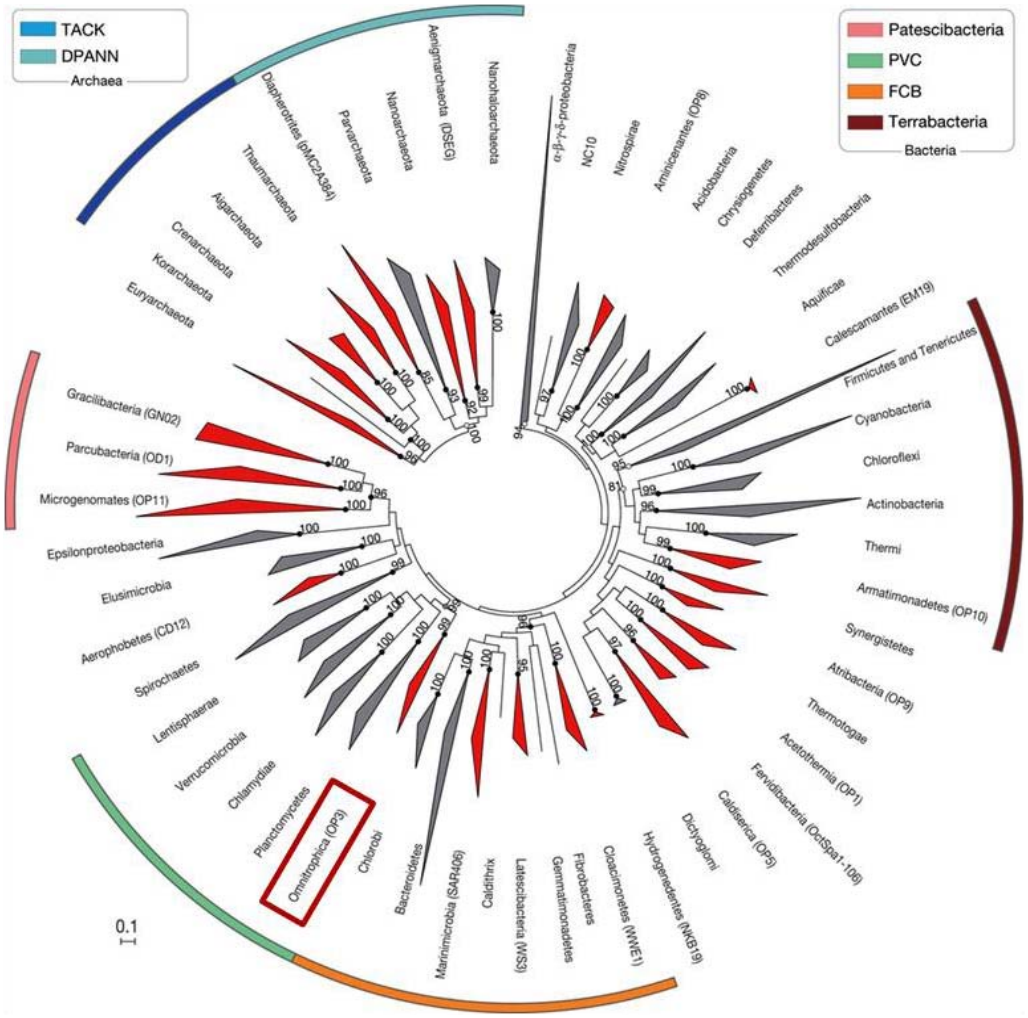


Fig. 4. Phylogenetic trees are based on up to 38 marker genes (Wu *et al.*, 2013) and sequences are collapsed at the phylum level. OP3 is highlighted with red frame. The PVC superphylum is highlighted with green color range (adapted from Rinke *et al.*, 2013).

1.3 Anaerobic monoterpene degrading methanogenic enrichments

1.3.1 Methanogenic degradation of complex organic matter

The degradation of complex biomass in anoxic habitats typically proceeds in several steps by a complex microbial community (Fig. 5). These steps include primary fermentations, followed by further oxidation by sulfate reduction or iron reduction, or by coupling primary fermentations with secondary fermentations to methanogenesis as terminal electron accepting process (Schink, 2006).

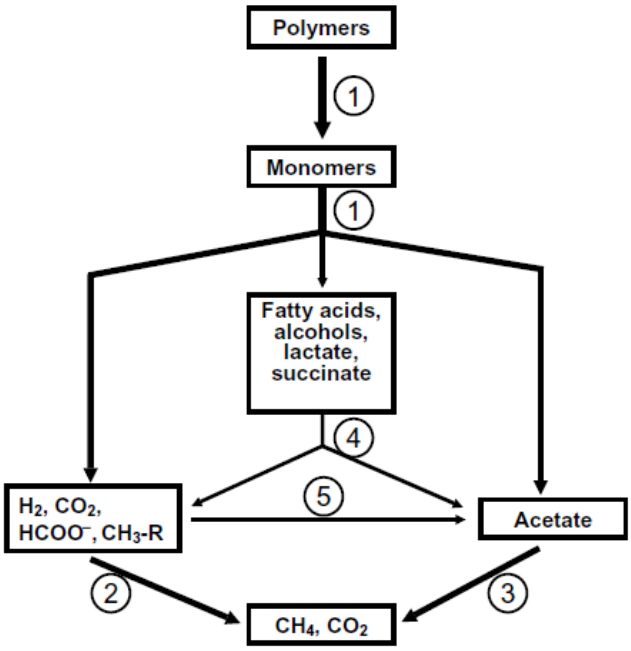


Fig. 5. Carbon and electron flow in the methanogenic degradation of complex biomass. Involved are: (1) primary fermentative bacteria; (2) hydrogenoxidizing methanogens; (3) acetate-cleaving methanogens; (4) secondary fermenting bacteria (syntrophs); and (5) homoacetogenic bacteria (from Schink, 2006).

The polymeric substrates (polysaccharides, proteins, lipids, and nucleic acids) are hydrolyzed to oligo- and monomers (sugars, amino acids, purines, pyrimidines, fatty acids, and glycerol) by primary fermenters. This metabolic group of bacteria ferments the monomers further to fatty acids, branched-chain fatty acids, succinate, alcohols, lactate, etc. Some of these products (acetate, CO_2, H_2, other one-carbon compounds) can be used directly by methanogens converting them to methane and CO_2. For the methanogenic degradation of other products of primary fermenters (fatty acids longer than two carbon atoms, alcohol longer than one carbon atom, branched-chain or aromatic fatty acids) secondary fermenters are needed to convert them to acetate, CO_2, H_2, and formate. These compounds are subsequently used by methanogens (Schink & Stams, 2006), which are divided into two groups: the hydrogenotrophic methanogens and the acetotrophic methanogens. They complete the process (Sieber *et al.*, 2010). The function of homoacetogenic bacteria in the overall process is less understood (Schink & Stams, 2006). These organisms can catalyze the reductive synthesis of acetate from CO_2. They constitute a very phylogenetically diverse bacteriological group, present in both oxic and anoxic environments (Drake *et al.*, 2002). Owing their metabolic versatility, they can participate in fermentation processes and compete

with primary fermenters for monomeric compounds and with secondary fermenters for fermentation products in anoxic environments. In certain environments (low pH or low temperature) they may even successfully compete with hydrogenotrophic methanogens (Schink & Stams, 2006).

In sediments or in well-balanced sludge digesters in which the hydrogen-utilizing community maintains a low hydrogen partial pressure the degradation of compounds, such as sugars and polysaccharides, may proceed nearly exclusively through acetate and hydrogen (bold arrows in Fig. 5), and reduced side products play only a minor role (Schink, 2006; Schink & Stams, 2006). Excessive production of reduced side products, however, is typically found in pure cultures or in unbalanced reactors. There, large amounts of easily fermentable substrates boost acid formation and inhibit hydrogenotrophic methanogens due to a drop in pH (< 6.0), or the presence of toxic compounds causes the inhibition. The degree of mutual dependence among the different metabolic groups of primary fermenters, secondary fermenters, and methanogens can vary considerably, from total independence to the entire dependence of fermenters on the cooperation with methanogens which consume the catabolic products of the secondary fermenters. This product consumption makes the catabolism of the fermenter thermodynamically feasible and is called syntrophic interaction (Schink & Stams, 2006).

1.3.2 Degradation of monoterpenes under methanogenic conditions

Terpenes are natural hydrocarbons, which are built from isoprene (C_5H_8) units. These compounds are the largest class of plant secondary metabolites (van der Werf *et al.*, 1999). Monoterpenes ($C_{10}H_{16}$) consist of two linked isoprene units. They belong to the diverse group of terpenoids, of which more than 55,000 structures are known to date. Monoterpenes comprise acyclic, monocyclic and bicyclic structures, as well as oxygenated forms (monoterpenoids) (Marmulla & Harder, 2014). They are widely distributed in nature (van der Werf *et al.*, 1999) and are used in the food, flavor, and fragrance industries due to their odorous properties (Lüddeke *et al.,* 2012). These components are ubiquitous present in leaves and fruits. Transport processes, such as the transport of decayed plant material by rain, runoff and wind into river and lakes (Rotaru, 2009) cause the presence of monoterpenes in anoxic habitats (Harder & Foss, 1999). The hydrophobic character of monoterpenes causes cell toxicity (Marmulla *et al.*, 2016). Below toxic concentrations, they can serve as carbon and

energy source for microorganisms in aerobic and anaerobic respiration mode (Marmulla & Harder, 2014).

Since the last two decades more and more bacteria were identified that oxidized aliphatic or aromatic hydrocarbons in the absence of oxygen with nitrate, sulfate, or other electron acceptors (Widdel *et al.*, 2010). Anaerobic degradation of alkenes was first observed with squalene ($C_{30}H_{50}$) and the model alkene, 1-hexadecene, enrichment cultures, which converted these alkenes to methane and carbon dioxide (Widdel *et al.*, 2010). Later, the mineralization of monoterpenes in denitrifying bacteria and methanogenic communities was discovered (Marmulla & Harder, 2014). In many deep sediments, oxygen, nitrate, ferric iron and sulfate are depleted. Then, only methanogenic communities can degrade the remaining organic matter (Zengler *et al.*, 1999). In previous studies the degradation of alkenoic monoterpenes, such as bicyclic monoterpenes, 2-carene and α-pinene, under methanogenic conditions was described (Harder & Foss, 1999).

Limonene ($C_{10}H_{16}$), the most abundant monocyclic monoterpene, is formed by more than 300 plants. It represents the major constituent of essential oils from citrus plants (van der Werf *et al.*, 1999; Marmulla & Harder, 2014). This compound can be degraded under denitrifying conditions (Foss & Harder, 1998). The review of Marmulla & Harder (2014) shows an overview on the transformations of limonene provided by a number of older review articles. These described components including perillyl alcohol, perillic acid, p-menth-1-ene-6,8-diol, alpha-terpineol, carveol and carvone, and limonene-1,2-diol (Marmulla & Harder, 2014) are only intermediate products in the transformation of limonene. Limonene degradation in the absence of inorganic electron acceptor is less exergonic when compared to aerobic or other anaerobic respiratory processes. Consequently, the microorganisms adapt to the exploitation of minimal energy spans by establishing syntrophic interactions (Rotaru, 2009; Stams & Plugge, 2009). Rotaru (2009) proposed a model of limonene degradation. The monocyclic monoterpene could be degraded to different fatty acids and alcohols by fermenters, which then break them down to smaller molecules. Smaller fatty acids could use as substrates by syntrophic bacteria and their catabolic end-products (acetate, formate, H_2, and CO_2) would feed the methanogens. Acetate could also be produced by homoacetogens.

1.3.3 Limonene-degrading methanogenic cultures

Members of *candidate phylum* Omnitrophica were detected in 16S rRNA gene clone libraries prepared from methanogenic enrichment cultures that utilized limonene as carbon and energy

source (Fig. 3, Rotaru *et al.*, 2012). This phylotype was called OP3 LiM, for limonene and methane.

The methanogenic cultures were established earlier with activated sludge from waste water treatment plant (Harder & Foss, 1999) and transferred annually in fresh water methanogenic media, overlaid with 2,2,4,6,8,8-heptamethylnonane (HMN) as organic carrier phase for limonene. However, these methanogenic enrichments were established previously on α-pinene and 2-carene and after five successive transfers a stable community capable of limonene degradation was obtained (Rotaru, 2009). Cultures were branched into twelve lines and transferred once a year. They produced methane for up to two and a half years (Fig. 6).

The composition and structure of the community was analyzed using the full cycle rRNA approach (Amann *et al.*, 1995) and catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH) experiments. Microscopic images identified numerous morphotypes in the enrichment, from small cocci and vibrio to long filaments. *Bacteria* represented 40% and *Archaea* represented 33% of the microbial community, as detected by CARD-FISH using the domain specific probes, EUB338 I + VI and ARCH-915 (Rotaru *et al.*, 2012). *Archaea* 16S rRNA gene sequences were related to microorganisms from the order *Methanomicrobiales* and *Methanosarcinales* (Rotaru *et al.*, 2012). *Bacteria* 16S rRNA gene sequences showed 16S similarity to representatives of the lineages: *Bacteroidetes*, *Deltaproteobacteria*, Candidate Division OP3, and *Firmicutes* (Fig. 3).

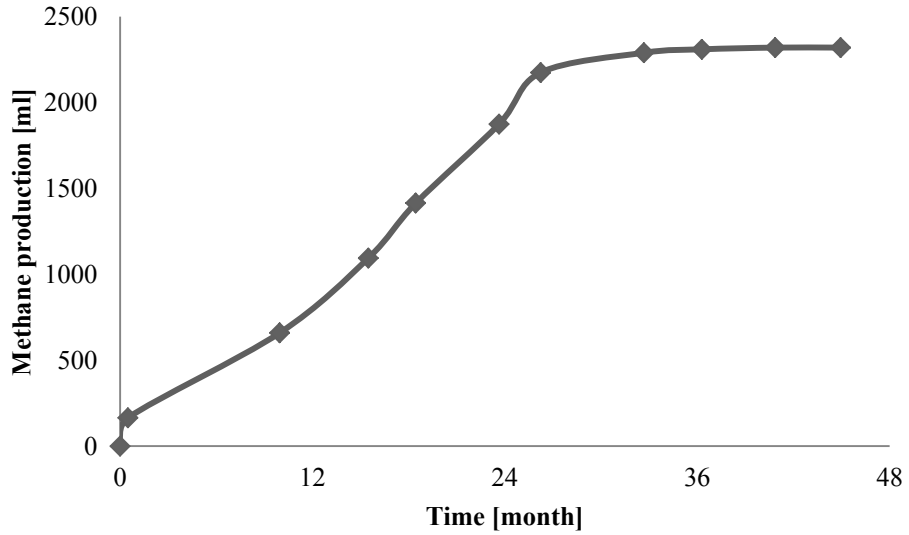


Fig. 6. Accumulated amount of methane production of the enrichment culture MM-214 (July 2010 to April 2014).

An OP3-specific probe [OP3-565] was developed and used for the visualization and quantification of OP3 LiM cells by CARD-FISH. The studies revealed a presence of 18% OP3 cells in the enrichment during the exponential growth (Rotaru *et al.*, 2012). These members of *candidate phylum* Omnitrophica were small and round-shaped cells, located either free-living or attached to larger cells (Fig. 7). All knowledge about OP3 candidate division so far has been derived exclusively from analysis of 16S rRNA gene clone sequences and metagenomic cosmid clones (Glöckner *et al.*, 2010; Kolinko *et al.*, 2012). The corresponding bacteria have remained undetected and consequently, nothing was known about the morphology of OP3 (Kolinko *et al.,* 2012). Thus, this was the first time that a member of this phylum had been visualized (Rotaru *et al.*, 2012).

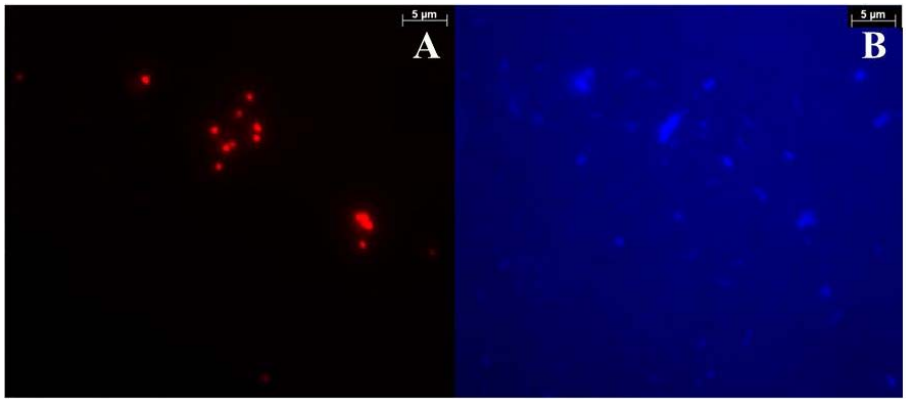


Fig. 7. Microscopic images of samples from methanogenic enrichment cultures thriving on limonene, as visualized by (A) epifluorescence microscopy with probe OP3-565, and (B) the same microscopic field visualized by DAPI staining. Scale bars, 5 µm (from Rotaru *et al.*, 2012).

Because attempts to isolate OP3 cells as strains failed (Rotaru, 2009), physical separations of bulk biomass was attempted. Anaerobic density gradient centrifugation was performed to separate the diverse microbial community and to provide highly enriched fractions with living cells. After two consecutive density gradient centrifugations, fractions containing over 80% OP3 LiM cells were obtained as revealed by CARD-FISH experiments (Kulkarni, 2010). Cells of the methanogenic enrichment cultures and of enriched OP3 LiM fractions were visualized by transmission electron microscopy (TEM) (Jordan, 2013). Small OP3-like cells both as free-living cells and cells attached to various types of larger rod-shaped cells were tentatively identified. This cell morphology was abundantly represented in fractionated materials. In addition, different cell sizes of the small cocci were observed for the

attached-living cells. Cells of this morphology both from fraction and from cultures consisted of a center and a sort of floating coat around (Fig. 8) (Jordan, 2013).

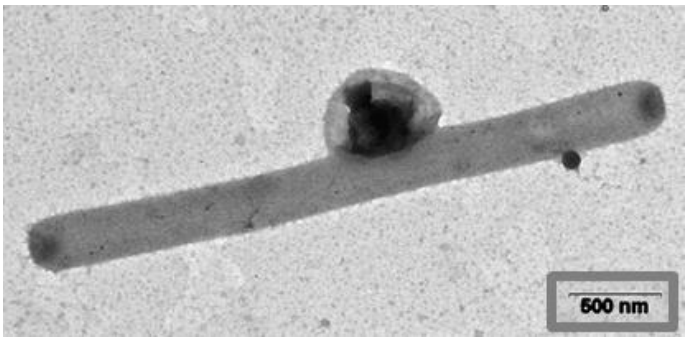


Fig. 8. TEM image of a methanogenic enrichment culture. A direct contact between several small coccoid cells of different cell sizes and a long rod-shaped cell exists.

The analysis of the genome of OP3 LiM had been started by the enrichment of OP3 LiM cells by physical cell separation. Genomic DNA was extracted and the OP3 LiM-enriched metagenome was sequenced by Roche 454 Titanium pyrosequencing technology (454 GS FLX) by the Max Planck-Genome-centre Cologne. An assembly of 454 reads by Newbler v. 2.3 resulted in a draft metagenome of 5,779 contigs of 16,026,544 bp in size. An initial analysis of the ten largest contigs, using the software JSpecies (Richter & Rosselló-Móra, 2009), indicated that at least five of these contigs were likely fragments of the OP3 genome. These contigs spanned a length of 1,650,908 bp (Jordan, 2013).

1.4 Aims of the study

Previous research has shown a complex interaction of microbes in the limonene-degrading enrichment cultures. Attempts to isolate OP3 LiM cells had failed. Previous studies had tested metabolic roles for OP3 LiM, but did not provide conclusive experimental evidence. Thus, the exploration of the biology of OP3 LiM cells promised insight into a so-far undisclosed biology.

The overarching goal of this thesis has been the understanding of the biology of OP3 LiM. Genome information has enormous impact on the view of an organism. Therefore, the determination of the complete closed genome sequence of this small bacterium was a major aim of this thesis. The step from a draft genome to a closed genome was expected to require a variety of assembly methods and manual quality controls to solve the genome puzzle. This bioinformatic approach was also to be applied to some planctomycete genomes. The strategy foresaw eventual laboratory work with PCR amplifications to resolve uncertainties in the genome draft. The complete genome was then basis for a comprehensive OP3 LiM study with metatranscriptomes and metaproteomes. For these experiments, a size-fractionation by centrifugation was developed during this thesis work.

OP3 LiM cells were further characterized by the visualization of the cells in enrichment cultures. The application of visualization approaches was optimized and indicated the physiological stages of OP3 LiM cells. The information obtained by the experimental work in this thesis was expected to provide a deep insight into the biology of OP3 LiM.

1.5 Manuscript and publication outline

Chapter 2

Permanent draft genome of '*Rhodopirellula islandica*' strain K833.

Jana Kizina, Marina Žure, Erhard Rhiel, Colin Bernhard Munn, Michael Richter, Jens Harder

Published in Mar Genomics 24 (2015):249–251.

Autor´s contributions:

J.K. performed the bioinformatics analyses under supervision of M.R., M.Z. performed the growth experiments supervised by C.B.M. and J.H., E.R. performed the electron microscopy, J.H. designed and supervised the study and wrote the manuscript together with all authors.

Contribution of J.K.:

Experimental concept and design: 0%; Experimental work: 60%; Data analysis and interpretation: 60%; Preparation of figures and tables: 40%; Writing of the manuscript: 10%.

Chapter 3

***In silico* detection of taxon-unrelated contigs and reassembling of taxon-specific reads improve draft genomes of strains.**

Jana Kizina, Jens Harder

In preparation for submission to J Microbiol Methods.

Autor´s contributions:

J.K. designed the method development and performed the bioinformatics analyses, J.K. and J.H. wrote the manuscript.

Contribution of J.K.:

Experimental concept and design: 100%; Experimental work: 100%; Data analysis and interpretation: 100%; Preparation of figures and tables: 95%; Writing of the manuscript: 70%.

Chapter 4

Optimization of CARD-FISH for *Methanosaeta* cells.

Jana Kizina, Jens Harder

Manuscript in preparation.

Autor´s contributions:

J.K. designed and performed the experiments, J.K. and J.H. wrote the manuscript.

Contribution of J.K.:

Experimental concept and design: 90%; Experimental work: 100%; Data analysis and interpretation: 90%; Preparation of figures and tables: 100%; Writing of the manuscript: 90%.

Chapter 5

"*Candidatus* Vampirococcus archaeovorus" and a transfer of a bacterial intron ribonucleic acid into archaeal cells.

Jana Kizina, Sebastian Jordan, Anastasia Resteu, Christina Probian, Sten Littmann, Stefanie Markert, Erhard Rhiel, Kurt Stüber, Thomas Schweder, Richard Reinhardt, Michael Richter, Jens Harder

In preparation for submission to Appl Environ Microbiol.

Autor´s contributions:

J.H. and J.K. designed the study, C.P. and J.K. maintained cultures, S.J. performed Percoll gradients and prepared DNA enriched in OP3 LiM, J.K. and A.R. performed size fractionation, A.R. prepared DNA and RNA of size-fractionated samples, J.K. performed the bioinformatics analyses, initially supervised by M.R., K.S. and R.R. performed sequencing and prepared the first Newbler assembly, E.R. contributed EM graphs, J.K. and S.L. prepared SEM graphs, J.K. performed *in situ* hybridizations, J.K., S.M. and T.S. performed the proteomic analysis, J.K. and J.H. annotated the genome and interpreted the proteomic results, J.H. supervised the study and J.K. and J.H. wrote the manuscript.

Contribution of J.K.:

Experimental concept and design: 30%; Experimental work: 80%; Data analysis and interpretation: 80%; Preparation of figures and tables: 95%; Writing of the manuscript: 50%.

References

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NL, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146.

Amann R. 2000. Who is out there? Microbial aspects of biodiversity. Syst Appl Microbio 23:1–8.

Amann R, Rosselló-Móra R. 2016. After all, only millions? MBio 7:e00999–16.

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

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun 7:13219.

Auguet JC, Barberan A, Casamayor EO. 2010. Global ecological patterns in uncultured *Archaea*. ISME J 4:182–190.

Chen I-MA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, Ratner A, Huang J, Andersen E, Huntemann M, Varghese N, Hadjithomas M, Tennessen K, Nielsen T, Ivanova NN, Kyrpides NC. 2017. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res 45:D507–D516.

Chouari R, Le Paslier D, Dauga C, Daegelen P, Weissenbach J, Sghir A. 2005. Novel major bacterial candidate division within a municipal anaerobic sludge digester. Appl Environ Microbiol 71:2145–2153.

Dedysh SN. 2011. Cultivating uncultured bacteria from northern wetlands: knowledge gained and remaining gaps. Front Microbiol 2:184.

Derakshani M, Lukow T, Liesack W. 2001. Novel bacterial lineages at the (sub)division level as detected by signature nucleotide-targeted recovery of 16S rRNA genes from bulk soil and rice roots of flooded rice microcosms. Appl Environ Microbiol 67:623–631.

Drake HL, Küsel K, Matthies C. 2002. Ecological consequences of the phylogenetic and physiological diversities of acetogens. Antonie Van Leeuwenhoek 81:203–213.

Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.

Foss S, Harder J. 1998. *Thauera linaloolentis* sp. nov. and *Thauera terpenica* sp. nov., isolated on oxygen-containing monoterpenes (linalool, menthol, and eucalyptol) and nitrate. Syst Appl Microbio 21:365–373.

Fuerst JA. 2013. The PVC superphylum: exceptions to the bacterial definition? Antonie Van Leeuwenhoek 104: 451–466.

Glöckner J, Kube M, Shrestha PM, Weber M, Glöckner FO, Reinhardt R, Liesack W. 2010. Phylogenetic diversity and metagenomics of candidate division OP3. Environ Microbiol 12:1218–1229.

Gupta RS, Bhandari V, Naushad HS. 2012. Molecular signatures for the PVC clade (*Planctomycetes*, *Verrucomicrobia*, *Chlamydiae*, and *Lentisphaerae*) of *Bacteria* provide insights into their evolutionary relationships. Front Microbiol 3:372.

Harder J, Foss S. 1999. Anaerobic formation of the aromatic hydrocarbon p-cymene from monoterpenes by methanogenic enrichment cultures. Geomicrobiol J 16:295–305.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:16048.

Hugenholtz P, Goebel BM, Pace NR. 1998a. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 180:4765–4774.

Hugenholtz P, Pitulle C, Hershberger KL, Pace NR. 1998b. Novel division level bacterial diversity in a Yellowstone hot spring. J Bacteriol 180:366–376.

Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. 2014. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ 2:e603.

Ivanova AO, Dedysh SN. 2012. Abundance, diversity, and depth distribution of planctomycetes in acidic northern wetlands. Front Microbiol 3:5.

Jordan S. 2013. Genomische und physiologische Analysen von Kandidatengenus OP3 Zellen aus methanogenen Anreicherungen. Bachelor thesis, University of Applied Sciences Aachen, Germany.

Kolinko S, Jogler C, Katzmann E, Wanner G, Peplies J, Schüler D. 2012. Single-cell analysis reveals a novel uncultivated magnetotactic bacterium within the candidate division OP3. Environ Microbiol 14:1709–1721.

Kulkarni U. 2010. Characterization of bacterial candidate division OP3. Master thesis, Universität Bremen, Germany.

Kumar RM, Saravanan VS. 2010. Candidate OP phyla: importance, ecology and cultivation prospects. Indian J Microbiol 50:474–477.

Lagkouvardos I, Jehl MA, Rattei T, Horn M. 2014. Signature protein of the PVC superphylum. Appl Environ Microbiol 80:440–445.

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

Lefebvre O, Vasudevan N, Thanasekaran K, Moletta R, Godon JJ. 2006. Microbial diversity in hypersaline wastewater: the example of tanneries. Extremophiles 10:505–513.

Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. Proc Natl Acad Sci U S A 113:5970–5975.

Lüddeke F, Wülfing A, Timkea M, Germer F, Weber J, Dikfidan A, Rahnfeld T, Linder D, Meyerdierks A, Harder J. 2012. Geraniol and geranial dehydrogenases induced in anaerobic monoterpene degradation by *Castellaniella defragrans*. Appl Environ Microbiol 78:2128–2136.

Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, Lemainque A, Wincker P, Aury JM. 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics 16:327.

Madrid VM, Taylor GT, Scranton MI, Chistoserdov AY. 2001. Phylogenetic diversity of bacterial and archaeal communities in the anoxic zone of the Cariaco Basin. Appl Environ Microbiol 67:1663–1674.

Markowitz VM, Chen IA, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang J, Woyke T, Huntemann M, Anderson I, Billis K, Varghese N, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC. 2014. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res 42:D560–D567.

Marmulla R, Harder J. 2014. Microbial monoterpene transformations-a review. Front Microbiol 5:346.

Marmulla R, Šafarić B, Markert S, Schweder T, Harder J. 2016. Linalool isomerase, a membrane-anchored enzyme in the anaerobic monoterpene degradation in *Thauera linaloolentis* 47Lol. BMC Biochem 17:6.

Pace NR, Stahl DA, Lane DJ, Olsen GJ. 1985. The analysis of natural microbial populations by ribosomal RNA sequences. Am Soc Microbiol News 51:4–12.

Pachter L. 2007. Interpreting the unculturable majority. Nat Methods 4:479–480.

Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD, Birol I. 2015. Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics 16:230.

Pisapia C, Gérard E, Gérard M, Lecourt L, Lang SQ, Pelletier B, Payri CE, Monnin C, Guentas L, Postec A, Quéméneur M, Erauso G, Ménez B. 2017. Mineralizing filamentous bacteria from the Prony Bay hydrothermal field give new insights into the functioning of serpentinization-based subseafloor ecosystems. Front Microbiol 8:57.

Prakash O, Verma M, Sharma P, Kumar M, Kumari K, Singh A, Kumari H, Jit S, Gupta SK, Khanna M, Lal R. 2007. Polyphasic approach of bacterial classification – an overview of recent advances. Indian J Microbiol 47:98–108.

Rappé MS, Giovannoni SJ. 2003. The uncultured microbial majority. Annu Rev Microbiol 57:369–394.

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

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA,. Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437.

Rosselló-Mora R, Amann R. 2001. The species concept for prokaryotes. FEMS Microbiol Rev 25:39–67.

Rotaru AE. 2009. Anaerobic degradation of limonene and p-xylene in freshwater enrichment cultures. PhD thesis, Universität Bremen, Germany.

Rotaru AE, Schauer R, Probian C, Mussmann M, Harder J. 2012. Visualization of candidate division OP3 cocci in limonene-degrading methanogenic cultures. J Microbiol Biotechnol 22:457–461.

Schink B. 2006. Syntrophic associations in methanogenic degradation. Prog Mol Subcell Biol 41:1–19.

Schink B, Stams AJM. 2006. Syntrophism among prokaryotes, p 309–335. *In* Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (ed), The prokaryotes: prokaryotic communities and ecophysiology, vol 2. Springer, New York.

Schleifer K. 2004. Microbial diversity: facts, problems and prospects. Syst Appl Microbiol 27:3–9.

Schloss PD, Handelsman J. 2004. Status of the microbial census. Microbiol Mol Biol Rev 68:686–691.

Schloss PD, Girard RA, Martin T, Edwards J, Thrash JC. 2016. Status of the archaeal and bacterial census: an update. MBio 7:e00201–16.

Sharma R, Ranjan R, Kapardar RK, Grover A. 2005. 'Unculturable' bacterial diversity: an untapped resource. Current science 89:72–77.

Sieber J, McInerney MJ, Plugge CM, Schink B, Gunsalus RP. 2010. Methanogenesis: syntrophic metabolism, p 337–355. *In* Timmis KN (ed), Handbook of hydrocarbon and lipid microbiology, vol 1. Springer, Berlin Heidelberg.

Speth DR, In ´t Zandt MH, Guerrero-Cruz S, Dutilh BE, Jetten MSM. 2016. Genome-based microbial ecology of anammox granules in a full-scale wastewater treatment system. Nat Commun 7:11172.

Stams AJM, Plugge CM. 2009. Electron transfer in syntrophic communities of anaerobic bacteria and archaea. Nat Rev Microbiol 7:568–577.

Stott MB, Saito JA, Crowe MA, Dunfield, PF, Hou S, Nakasone E, Daughney CJ, Smirnova AV, Mountain BW, Takai K, Alam M. 2008. Culture-independent characterization of a novel microbial community at a hydrothermal vent at Brothers volcano, Kermadec arc, New Zealand. J Geophys Res 113:B08S06.

Strous M, Kraft B, Bisdorf R, TegetMeyer H. 2012. The binning of metagenomic contigs for microbial physiology of mixed cultures. Front Microbiol 3:410.

Teeling H, Glöckner FO. 2012. Current opportunities and challenges in microbial metagenome analysis -- a bioinformatic perspective. Brief Bioinform 13:728–742.

Torsvik V, Øvreås L, Thingstad TF. 2002. Prokaryotic diversity--magnitude, dynamics, and controlling factors. Science 296:1064–1066.

van der Werf M, Swarts HJ, de Bont JAM. 1999. *Rhodococcus erythropolis* DCL14 contains a novel degradation pathway for limonene. Appl Environ Microbiol 65:2092–2102.

von Wintzingerode F, Göbel UB, Stackebrandt E. 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev 21:213–229.

Wagner M, Horn M. 2006. The *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opin Biotechnol 17:241–249.

Ward BB. 2002. How many species of prokaryotes are there? Proc Natl Acad Sci U S A 99:10234–10236.

Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: The unseen majority. Proc Natl Acad Sci U S A 95:6578–6583.

Widdel F, Knittel K, Galushko A. 2010. Anaerobic hydrocarbon degrading microorganisms: an overview, p 1997–2021. *In* Timmis KN (ed), Handbook of hydrocarbon and lipid microbiology, vol 1. Springer, Berlin Heidelberg.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088–5090.

Wu D, Jospin G, Eisen JA. 2013. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS ONE 8:e77033.

Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol 12:635–645.

Zengler K, Richnow HH, Rossello´-Mora R, Michaelis W, Widdel F. 1999. Methane formation from long-chain alkanes by anaerobic microorganisms. Nature 401:266–269.

Appendix of Chapter 1:

Supplementary table

Table S1. Overview on currently available population genome bins / draft genomes for *Candidatus phylum* Omnitrophica from GenBank (June 2017; https://www.ncbi.nlm.nih.gov).

Organism/Name	Assembly accession no.	Size (Mb)	Source of DNA	Latitude/Longitude	Habitat	Sample characteristics
Candidatus Omnitrophus fodinae SCGC AAA011-A17	ASM40594v1	2.03708	single cell isolate	44°20'46"N, 103°45'39"W	water	
Omnitrophica bacterium SCGC AAA011-L16	ASM39812v1	0.46857	single cell isolate	44°20'46"N, 103°45'39"W	water	
Omnitrophica bacterium SCGC AAA252-B19	ASM39808v1	0.51897	single cell isolate	49°40'30"N, 124°2'2.4"W	brackish water	
Omnitrophica bacterium SCGC AAA257-O07	ASM40298v1	0.76785	single cell isolate	38°28'59.54"N, 21°19'17.44"E	sediment	
Candidatus Omnitrophus magneticus	ASM95409v1	3.14590	single cell isolate	47.869440 N 12.398015 E	freshwater	
Candidatus Omnitrophica bacterium CG1_02_40_15	ASM187181v1	1.16900	metagenome	38.56 N 110.8 W	subsurface	
Candidatus Omnitrophica bacterium CG1_02_41_171	ASM187186v1	0.92485	metagenome	38.56 N 110.8 W	subsurface	
Candidatus Omnitrophica bacterium CG1_02_43_210	ASM187277v1	1.28923	metagenome	38.56 N 110.8 W	subsurface	
Candidatus Omnitrophica bacterium CG1_02_44_16	ASM187399v1	1.55975	metagenome	38.56 N 110.8 W	subsurface	
Candidatus Omnitrophica bacterium CG1_02_46_14	ASM187402v1	1.19141	metagenome	38.56 N 110.8 W	subsurface	
Candidatus Omnitrophica bacterium CG1_02_49_10	ASM187304v1	0.90391	metagenome	38.56 N 110.8 W	subsurface	
Candidatus Omnitrophica bacterium CG1_02_49_16	ASM187210v1	0.95625	metagenome	38.56 N 110.8 W	subsurface	
Omnitrophica WOR_2 bacterium SM23_29	ASM130364v1	2.04325	metagenome	34.7414 N 77.1234 W	sediment	
Omnitrophica WOR_2 bacterium SM23_72	ASM130413v1	2.73046	metagenome	34.7414 N 77.1234 W	sediment	
Omnitrophica bacterium GWA2_41_15	ASM180397v1	1.05600	metagenome	39.5369 N 107.7828 W	subsurface	pH 7.25
Omnitrophica bacterium GWA2_50_21	ASM180400v1	0.92178	metagenome	39.5369 N 107.7828 W	subsurface	pH 7.25
Omnitrophica bacterium GWA2_52_12	ASM180530v1	1.03041	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.25

Table S1. – continued from previous page.

Organism/Name	Assembly accession no.	Size (Mb)	Source of DNA	Latitude/Longitude	Habitat	Sample characteristics
Omnitrophica bacterium GWA2_52_8	ASM180402v1	2.19580	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.25
Omnitrophica bacterium RBG_13_46_9	ASM180404v1	2.04538	metagenome	39.5369 N 107.7828 W	subsurface-sediment	pH 7.11
Omnitrophica bacterium RIFCSPHIGHO2_02_FULL_45_28	ASM180532v1	3.07542	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica bacterium RIFCSPHIGHO2_02_FULL_46_11	ASM180405v1	1.79055	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 8.36
Omnitrophica bacterium RIFCSPHIGHO2_02_FULL_46_20	ASM180406v1	1.31247	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 9.36
Omnitrophica bacterium RIFCSPHIGHO2_02_FULL_49_9	ASM180410v1	0.72415	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 10.36
Omnitrophica bacterium RIFCSPHIGHO2_02_FULL_51_18	ASM180412v1	1.52665	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 11.36
Omnitrophica bacterium RIFCSPHIGHO2_02_FULL_63_14	ASM180414v1	1.24372	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 12.36
Omnitrophica bacterium RIFCSPHIGHO2_12_FULL_44_12	ASM180415v1	2.63684	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 13.36
Omnitrophica bacterium RIFCSPLOWO2_01_FULL_45_10	ASM180416v1	1.43378	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.28
Omnitrophica bacterium RIFCSPLOWO2_01_FULL_45_10b	ASM180420v1	1.84359	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 8.28
Omnitrophica bacterium RIFCSPLOWO2_01_FULL_45_24	ASM180536v1	1.33392	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 9.28
Omnitrophica bacterium RIFCSPLOWO2_01_FULL_50_24	ASM180422v1	1.20078	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 10.28
Omnitrophica bacterium RIFCSPLOWO2_02_FULL_44_11	ASM180424v1	3.14847	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 11.28
Omnitrophica bacterium RIFCSPLOWO2_02_FULL_45_16	ASM180425v1	1.35832	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 12.28
Omnitrophica bacterium RIFCSPLOWO2_12_FULL_44_17	ASM180428v1	3.02012	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 13.28

Table S1. – continued from previous page.

Organism/Name	Assembly accession no.	Size (Mb)	Source of DNA	Latitude/Longitude	Habitat	Sample characteristics
Omnitrophica bacterium RIFCSPLOWO2_12_FULL_45_13	ASM180430v1	1.28991	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 14.28
Omnitrophica bacterium RIFCSPLOWO2_12_FULL_50_11	ASM180431v1	2.23535	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 15.28
Omnitrophica bacterium RIFOXYB12_FULL_50_7	ASM180434v1	2.06216	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 6.69
Omnitrophica WOR_2 bacterium GWA2_37_7	ASM180436v1	1.60195	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.25
Omnitrophica WOR_2 bacterium GWA2_44_7	ASM180438v1	1.55160	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.25
Omnitrophica WOR_2 bacterium GWA2_45_18	ASM180439v1	2.33574	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.25
Omnitrophica WOR_2 bacterium GWA2_47_8	ASM180537v1	2.40036	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.25
Omnitrophica WOR_2 bacterium GWA2_53_43	ASM180539v1	1.02566	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.25
Omnitrophica WOR_2 bacterium GWA2_63_20	ASM180441v1	1.42040	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.25
Omnitrophica WOR_2 bacterium GWB2_45_9	ASM180444v1	0.77660	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 6.88
Omnitrophica WOR_2 bacterium GWC2_44_8	ASM180446v1	1.71475	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.39
Omnitrophica WOR_2 bacterium GWC2_45_7	ASM180540v1	1.34007	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 8.39
Omnitrophica WOR_2 bacterium GWF2_38_59	ASM180447v1	2.52572	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.75
Omnitrophica WOR_2 bacterium GWF2_43_52	ASM180544v1	2.47217	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 8.75
Omnitrophica WOR_2 bacterium GWF2_63_9	ASM180448v1	1.46283	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 9.75
Omnitrophica WOR_2 bacterium RBG_13_41_10	ASM180546v1	1.29887	metagenome	39.5369 N 107.7828 W	subsurface-sediment	pH 7.11

Table S1. – continued from previous page.

Organism/Name	Assembly accession no.	Size (Mb)	Source of DNA	Latitude/Longitude	Habitat	Sample characteristics
Omnitrophica WOR_2 bacterium RBG_13_44_8	ASM180547v1	0.82768	metagenome	39.5369 N 107.7828 W	subsurface-sediment	pH 7.11
Omnitrophica WOR_2 bacterium RBG_13_44_8b	ASM180548v1	1.38070	metagenome	39.5369 N 107.7828 W	subsurface-sediment	pH 7.11
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_01_FULL_48_9	ASM180552v1	2.05604	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_01_FULL_49_10	ASM180554v1	1.08626	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_01_FULL_52_10	ASM180556v1	0.89811	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_45_21	ASM180557v1	1.31244	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_46_37	ASM180560v1	0.76434	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_48_11	ASM180561v1	2.01293	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_50_17	ASM180564v1	1.92421	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_52_10	ASM180565v1	1.56453	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_63_39	ASM180568v1	1.39510	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_67_20	ASM180569v1	0.84470	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_02_FULL_68_15	ASM180572v1	1.82052	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPHIGHO2_12_FULL_64_13	ASM180573v1	1.01448	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.36
Omnitrophica WOR_2 bacterium RIFCSPLOWO2_01_FULL_41_12	ASM180576v1	0.70425	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.28
Omnitrophica WOR_2 bacterium RIFCSPLOWO2_02_FULL_45_28	ASM180577v1	0.90900	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.28

Table S1. – continued from previous page.

Organism/Name	Assembly accession no.	Size (Mb)	Source of DNA	Latitude/Longitude	Habitat	Sample characteristics
Omnitrophica WOR_2 bacterium RIFCSPLOWO2_02_FULL_50_19	ASM180580v1	1.25195	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.28
Omnitrophica WOR_2 bacterium RIFCSPLOWO2_02_FULL_63_16	ASM180581v1	1.70241	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.28
Omnitrophica WOR_2 bacterium RIFCSPLOWO2_12_FULL_46_30	ASM180584v1	0.86563	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.28
Omnitrophica WOR_2 bacterium RIFCSPLOWO2_12_FULL_50_9	ASM180585v1	1.23654	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.28
Omnitrophica WOR_2 bacterium RIFCSPLOWO2_12_FULL_51_24	ASM180588v1	0.92580	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.28
Omnitrophica WOR_2 bacterium RIFCSPLOWO2_12_FULL_51_8	ASM180589v1	0.93623	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.28
Omnitrophica WOR_2 bacterium RIFCSPLOWO2_12_FULL_63_16	ASM180592v1	1.65354	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.28
Omnitrophica WOR_2 bacterium RIFOXYA12_FULL_38_10	ASM180594v1	2.27093	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.67
Omnitrophica WOR_2 bacterium RIFOXYA2_FULL_38_17	ASM180596v1	2.67715	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.67
Omnitrophica WOR_2 bacterium RIFOXYA2_FULL_45_12	ASM180597v1	0.96326	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.67
Omnitrophica WOR_2 bacterium RIFOXYB2_FULL_38_16	ASM180599v1	2.74648	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 6.69
Omnitrophica WOR_2 bacterium RIFOXYB2_FULL_45_11	ASM180602v1	0.94837	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 6.69
Omnitrophica WOR_2 bacterium RIFOXYC2_FULL_38_12	ASM180604v1	2.75002	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.57
Omnitrophica WOR_2 bacterium RIFOXYC2_FULL_43_9	ASM180605v1	2.03736	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.57
Omnitrophica WOR_2 bacterium RIFOXYC2_FULL_45_15	ASM180607v1	1.20127	metagenome	39.5369 N 107.7828 W	subsurface-groundwater	pH 7.39
Omnitrophica bacterium OLB16	ASM156711v1	4.05294	metagenome	52.048138 N 6.144261 E	sludge from waste water treatment plant	pH 7.80; 35°C

Table S1. – continued from previous page.

Organism/Name	Assembly accession no.	Size (Mb)	Source of DNA	Latitude/Longitude	Habitat	Sample characteristics
Candidatus Omnitrophica bacterium 4484_100	ASM204978v1	1.33066	metagenome	27.0388 N 111.24560 W	marine sediment from deep-sea hydrothermal vent	
Candidatus Omnitrophica bacterium 4484_171	ASM204989v1	1.26784	metagenome	27.0388 N 111.24560 W	marine sediment from deep-sea hydrothermal vent	
Candidatus Omnitrophica bacterium 4484_49	ASM208526v1	0.89131	metagenome	27.0388 N 111.24560 W	deep-sea hydrothermal vent sediments	
Candidatus Omnitrophica bacterium 4484_213	ASM208502v1	0.93261	metagenome	27.0388 N 111.24560 W	deep-sea hydrothermal vent sediments	
Candidatus Omnitrophica bacterium 4484_70.1	ASM208531v1	1.70116	metagenome	27.0388 N 111.24560 W	deep-sea hydrothermal vent sediments	
Candidatus Omnitrophica bacterium 4484_70.2	ASM208534v1	1.65032	metagenome	27.0388 N 111.24560 W	deep-sea hydrothermal vent sediments	

Chapter 2

Permanent draft genome of '*Rhodopirellula islandica*' strain K833

Jana Kizina, Marina Žure, Erhard Rhiel, Colin Bernhard Munn, Michael Richter, Jens Harder

Mar Genomics 24 (2015):249–251

Permanent draft genome of '*Rhodopirellula islandica*' strain K833☆

Jana Kizina [a], Marina Žure [a], Erhard Rhiel [c], Colin Bernhard Munn [d], Michael Richter [b], Jens Harder [a,*]

[a] *Max Planck Institute for Marine Microbiology, Department of Microbiology, D-28359 Bremen, Germany*

[b] *Max Planck Institute for Marine Microbiology, Microbial Genomics and Bioinformatics Research Group, D-28359 Bremen, Germany*

[c] *Institute for Chemistry and Biology of the Marine Environment, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany*

[d] *School of Marine Sciences and Engineering, University of Plymouth, Plymouth PL4 8AA, United Kingdom*

☆ Competing interest: The authors have declared that no competing interests exist.

* Corresponding author.
E-mail address: jharder@mpi-bremen.de (J. Harder).

Keywords: *Planctomycetes*, Cold adaptation

ABSTRACT

The planctomycete strain K833 was isolated from cold waters at the coast of Island and is tentatively named '*Rhodopirellula islandica*'. It has a lower temperature range for growth than other genome-sequenced *Rhodopirellula* strains affiliating to *Rhodopirellula baltica* and '*Rhodopirellula europaea*'. The permanent draft genome of strain K833 was obtained as part of a larger study on the biogeography of *Rhodopirellula* species in European marine waters. The genome consists of 55 contigs with a genome size of 7,433,200 bp. With an average nucleotide identity of 81% to related genomes of *R. baltica* and '*R. europaea*' and more than 4000 common genes, it will be a valuable source for the study of temperature adaptation of planctomycete genomes.

2.1 Introduction

Rhodopirellula is a genus of marine *Planctomycetes* which are unusual bacteria lacking a murein saculus in the membrane. *Planctomycetes* live frequently attached to surfaces. They are abundant in the particulate fractions of marine ecosystems and are considered as important participants in the global carbon and nitrogen cycles. *Rhodopirellula baltica* SH1^{T} was the source for the first planctomycete genome (Glöckner et al. 2003). A collection of 70 *Rhodopirellula* strains obtained from different European seas (Winkelmann and Harder, 2009) revealed 13 distinct operational taxonomic units (OTUs) (Winkelmann et al., 2010). Eight strains were sequenced which covered sample sites from the Baltic Sea to the Mediterranean Sea (Klindworth et al., 2014, Richter et al., 2014a,b, Richter-Heitmann et al., 2014, Wegner et al., 2014). '*R. islandica*' strain K833 (=JCM 17612 = DSM 24040) was isolated from a water sample on the coast of Sandgerdi, Island (64.0356 N 22.6986 W) (Winkelmann and Harder, 2009) and features in electron micrographs the typical structures of the genus: swarming motile cells with a flagellum, attached-living cells often form rosettes, the reproductive pole also features small flagella-like structures (Fig. 1).

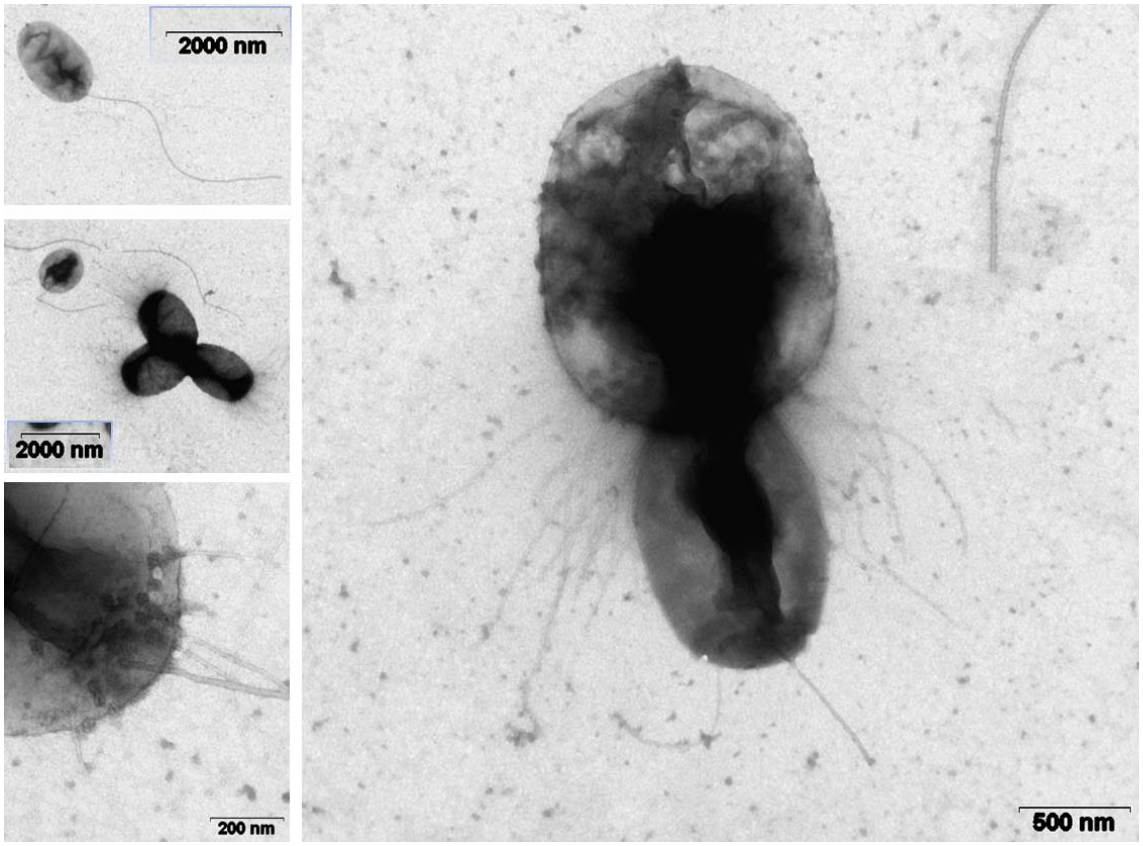


Fig. 1. Electron micrographs of cells of '*R. islandica*' strain K833: free-living swarmer cell (upper left), sessile cells (middle left), budding cell (right) and reproductive pole of a cell in transition from swarming to sessile life style (lower left). Preparation and EM conditions were as described (Hahnke et al., 2014).

In a comparative growth study performed in triplicates, cells pre-cultured on M13a agar plates were inoculated into 250 ml flasks that contained 50 ml of M13 medium in artificial seawater at ~35 per mille salinity. The culture flasks were incubated rotating at 50 rpm in the dark at 14, 21, 28 or 37 °C (Fig. 2). Strain K833 did not grow at 37 °C in contrast to the other strains, whereas *R. baltica* SH1^{T} grew and '*R. europaea*' 6C had the highest growth rate. At 14 °C, strain 6C had a long lag phase before growth started and strain K833 showed the highest growth rate of the strains (Fig. 2). Thus, the strains are a valuable source to investigate the genomic features of temperature adaptation in microorganisms.

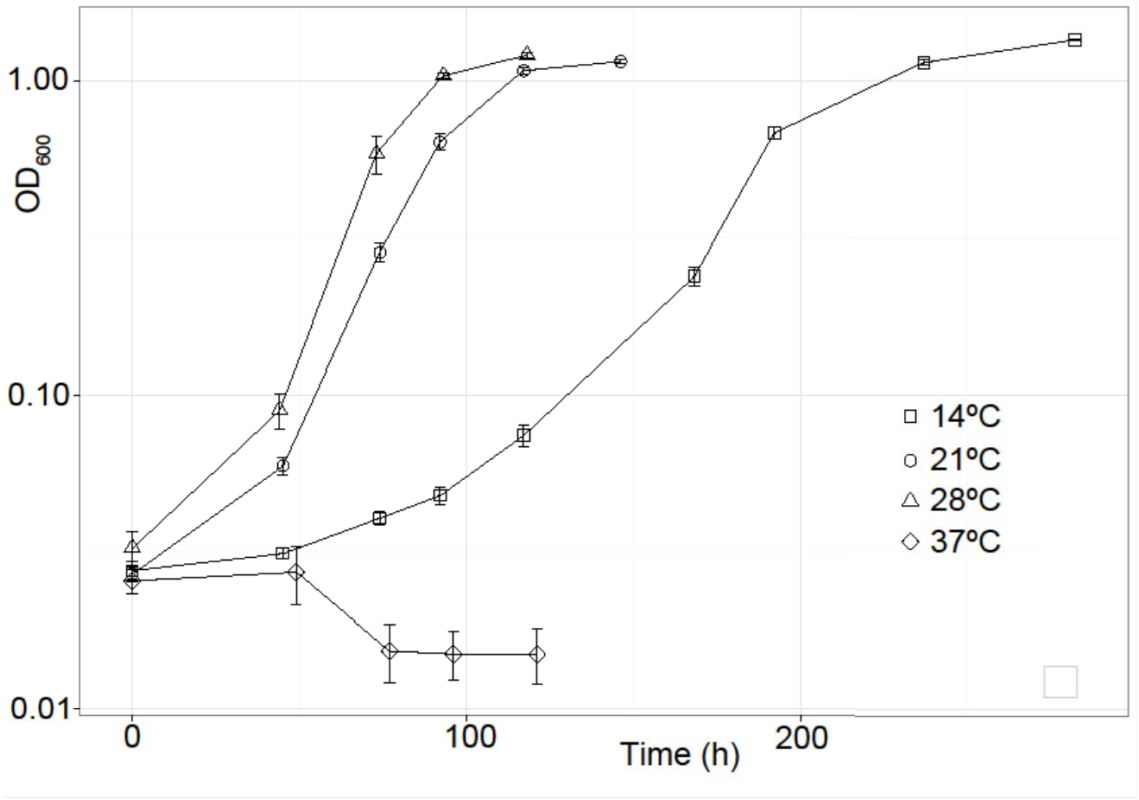


Fig. 2. Effect of temperature on growth of '*R. islandica*' K833. Error bars indicate the standard deviation of the mean of optical density (OD) for three cultures.

2.2 Data description

Genomic DNA of strain K833 was sequenced by the Illumina MiSeq technology performed by the Max Planck-Genome-centre Cologne, Germany ((http://mpgc.mpipz.mpg.de/home/). 5,494,521 paired-end reads of 250 bp were dynamically trimmed with SolexaQA v.2.2. (Cox et al. 2010) and normalized with Khmer 1.0 (Crusoe et al. 2014). 1,462,500 high-quality reads were assembled with SPades (Bankevich et al. 2012). Contigs were *de-novo* assembled in Geneious R8 (Biomatters, Auckland, New Zealand) to remove duplications and reads were mapped onto contigs with BBtools to identify possible contig elongations. The mapping reads were reassembled using the first assembly as trusted assembly in Spades. After six rounds, the assembly was stable and CheckM 0.9 indicated a completeness of 99.93% with a contamination value of 0.0% (Parks et al. 2014). The genome was annotated in RAST (Aziz et al. 2008) (Table 1).

Table 1

Genome and environmental features.

Item	Description
MIGS data	
Investigation_type	Bacteria_archaea
Project_name	K833
Collected_by	Julia Strahl
Collection_date	18-May-2005
Lat_lon	64.0356 N 22.6986 W
Depth	0 m
Alt_elev	NA
Country	Iceland
Environment	Marine water
Ref_biomaterial	http://dx.doi.org/10.1128/AEM.01525-09
Biotic_relationship	Free living
Trophic_level	Heterotroph
Rel_to_oxygen	Aerobe
Isol_growth_condt	http://www.ncbi.nlm.nih.gov/pubmed/19303037
Sequencing_meth	Illumina MiSeq
Num_replicons	NA
Assembly	SPAdes 3.5
Finishing_strategy	Draft
Annot_source	RAST
Estimated_size	7,433,200
Biome	ENVO:00000569
Feature	ENVO:02000049
Material	ENVO:00002150
Geo_loc_name	Sandgerdi, Iceland
Sample-material	Surface water from the beach
Source_mat_id	DSM 24040
Temp	7 °C
Salinity	34,6 PSU
Motility	Yes
Genome assembly data	
Assembly method	SPAdes v. 3.5.
Assembly name	K833
Genome coverage	370×
Sequencing technology	Illumina MiSeq

The genome encodes 6851 proteins, 54 tRNAs and 3 rRNAs. These values are in the range of previous observations for genomes of *Rhodopirellula* strains, with over 7 Mb and over 6000 predicted open reading frames each, and reflect the complex lifestyle of the planctomycetes. The average nucleotide identity between strains was determined by JSpecies (Richter and Rosselló-Móra, 2009) and is 81.2% between strain K833 and strain SH1^{T}, 81.4%

between strain K833 and strain 6C, and 88.1% between strain SH1^T and strain 6C, indicating a very close relationship between the three species. Pairwise analysis by reciprocal best match BLAST in RAST revealed 4241 genes present in both strain K833 and *R. baltica* SH1^T. This high number reflects the close relation between the two species as predicted by 16S rDNA and ANI analysis (Table 2). The sessile lifestyle of planctomycetes comprises life in oxygen-limited biofilms. The K833 genome codes for the synthesis of menaquinone, the typical quinone of microaerophiles, and a menaquinone (vitamin K)-dependent gamma-carboxylase that is not present in *R. baltica* SH1^T. The genome of strain K833 codes not only for cbb_3-type cytochrome c and cytochrome d terminal oxidases — well known for their high affinity to molecular oxygen—, but also for a periplasmatic nitrate reduction pathway as alternative electron acceptor which is lacking in *R. baltica* SH1^T. Thus, the bacterium is well adapted to microoxic-anoxic transition zones that occur frequently in coastal sediments. A feature of these transition zones is fermentation yielding an acidification. In contrast to the *R. baltica* SH1^T, the K833 genome contains a glutamate decarboxylase (EC 4.1.1.15) conferring acid-resistance. It has also some unique sulfatases and glycosyl hydrolases. Unusual is the presence of *traG* and *traI* genes within a region encoding hypothetical proteins, indicating a potential for genetic transfer.

Table 2

Genome features.

	'*R. islandica*' strain K833
Size (bp)	7,433,200
Contigs	55
GC content (%)	57.2
Coding (%)	88
CDS	6851
rRNA genes	3
tRNA genes	54
ANI [%][a,b]	81.2%
16S rDNA identity[a]	99.54%

[a] Compared to the type strain *Rhodopirellula baltica* SH1^T (Glöckner et al., 2003).

[b] Calculated by using the software JSpecies (Richter and Rosselló-Móra, 2009).

Nucleotide sequence accession number

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession LECT00000000. The version described in this paper is version LECT01000000. The sequence associated contextual (meta)data are MIxS compliant (Table 1, Yilmaz et al., 2011).

Acknowledgments

We thank the Max Planck-Genome-centre Cologne (http://mpgc.mpipz.mpg.de/home/) for performing the sequencing. This study was funded by the Max Planck Society. Marina Zure was supported through a MARES Grant. MARES is a Joint Doctorate Programme selected under Erasmus Mundus coordinated by Ghent University (FPA 2011-0016). See www.mares-eu.org for extra information.

References

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O., 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics 9, 75.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.-V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Cox, M.P., Peterson, D.A., Biggs, P.J., 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11, 485.

Crusoe, M., Edvenson, G., Fish, J., Howe, A., McDonald, E., Nahum, J., Nanlohy, K., Ortiz-Zuazaga, H., Pell, J., Simpson, J., Scott, C., Srinivasan, R.R., Zhang, Q., Brown, C.T., 2014. The Khmer software package: enabling efficient sequence analysis. http://dx.doi.org/10.6084/m9.figshare.979190.

Glöckner, F.-O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, A., Borzym, K., Heitmann, K., Rabus, R., Schlesner, H., Amann, R., Reinhardt, R., 2003. Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. Proc. Natl. Acad. Sci. USA. 100, 8298–8303.

Hahnke, R.L., Bennke, C.M., Fuchs, B.M., Mann, A.-J., Rhiel, E., Teeling, H., Amann, R. Harder, J., 2014. Dilution cultivation of marine heterotrophic bacteria abundant after a spring phytoplankton bloom in the North Sea. Environ. Microbiol. http://dx.doi.org/10.1111/1462–2920.12479.

Klindworth, A., Richter, M., Richter-Heitmann, T., Wegner, C.-E., Frank, C.S., Glöckner, F.O. Harder, J., 2014. Permanent draft genome of *Rhodopirellula rubra* SWK7. Mar. Gen. 13, 11–12.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2014. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 15. http://dx.doi.org/10.1101/gr.186072.114.

Richter, M., Richter-Heitmann, T., Klindworth, A., Wegner, C.-E., Frank, C.S. Harder, J., Glöckner, F.O., 2014a. Permanent draft genomes of the *Rhodopirellula maiorica* strain SM1. Mar. Gen. 13, 19–20.

Richter, M., Richter-Heitmann, T., Klindworth, A., Wegner, C.-E., Frank, C.S. Harder, J., Glöckner, F.O., 2014b. Permanent draft genomes of the three *Rhodopirellula baltica* strains SH28, SWK14 and WH47. Mar. Gen. 13, 13–14.

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

Richter-Heitmann, T., Richter, M., Klindworth, A., Wegner, C.-E., Frank, C.S., Glöckner, F.O. Harder, J., 2014. Permanent draft genomes of the two *Rhodopirellula europaea* strains 6C and SH398. Mar. Gen. 13, 15–16.

Wegner, C.-E., Richter, M., Richter-Heitmann, T., Klindworth, A., Frank, C.S., Glöckner, F.O. Harder, J., 2014. Permanent draft genome of *Rhodopirellula sallentina* SM41. Mar. Gen. 13, 17–18.

Winkelmann, N., Harder, J., 2009. An improved isolation method for attached-living *Planctomycetes* of the genus *Rhodopirellula*. J. Microbiol. Meth. 77, 276–284.

Winkelmann, N., Jaekel, U., Meyer, C., Serrano, W., Rachel, R., Rosselló-Móra, R., Harder, J., 2010. Determination of the diversity of *Rhodopirellula* isolates from European seas by multi-locus sequence analysis. Appl. Envir. Microbiol. 76, 776–785.

Yilmaz, P., et al., 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat. Biotechnol. 29, 415–420.

Chapter 3

In silico detection of taxon-unrelated contigs and reassembling of taxon-specific reads improve draft genomes of strains

Jana Kizina, Jens Harder

In preparation for submission to J Microbiol Methods

In silico detection of taxon-unrelated contigs and reassembling of taxon-specific reads improve draft genomes of strains

Jana Kizina, Jens Harder*

Department of Microbiology, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany

* Corresponding author: Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany. Tel.: +49 421 2028 750; Fax: +49 421 2028 580.
E-mail address: jharder@mpi-bremen.de (J. Harder).

Keywords: draft assembly; next-generation sequencing; reassembly; taxonomic assignment

Highlights

- The *in silico* approach improves short read *de novo* draft assemblies of strains
- Binning differentiates sequences by phylogeny
- Assembly of taxon-specific reads have fewer contigs with longer sequence lengths

Abstract

The next-generation sequencing (NGS) technology Illumina produces high-throughput sequence information at low costs. But the short read length limits the genome assembly quality. Many *de novo* draft assemblies remain highly fragmented. In this contribution, we describe an *in silico* approach to improve microbial draft genome assemblies by iterative rounds of taxonomic evaluation of contigs by the software Metawatt, read mapping to taxonomically coherent contigs, and reassembly of the recruited reads. This taxonomically guided approach reduced the numbers of contigs substantially and yielded contamination-free permanent draft genomes.

3.1 Introduction

The complete genome sequence of an organism provides an invaluable resource to the wider research community and is the basis for comparative and evolutionary genomics studies (Tsai *et al.*, 2010). The capillary electrophoresis sequencing technology using Sanger´s dideoxynucleotide synthesis termination yields about 1.5 Mb/day of high-quality reads with average length of 500–800 bases. However, the fragments of DNA to be sequenced must first be cloned and the resulting libraries must be maintained (Mavromatis *et al.*, 2012). The introduction of next-generation sequencing (NGS) in the last decade has accelerated the study of microbial genomes (Koren & Phillippy, 2015). 454 pyrosequencing and the widely used second-generation sequencing (SGS) technology Illumina sequencing-by-synthesis do not require cloning (Mavromatis *et al.*, 2012). The most advanced chemistry/platform combination of 454 pyrosequencing (GS FLX + System with the GS FLX Titanium Sequencing Kit XL+) can produce ~1 million reads per run in 23 h with reads up to 1,000 bases in length and an average read length of 700 bases (Hodkinson & Grice, 2015). The costs decreased from 500 United States Dollars (USD) per million bases for Sanger sequencing to 8.57 USD/Mb for 454 systems (Rhoads & Au, 2015). A typical technological error was the frequent misidentification of the length of homopolymers (Hodkinson & Grice, 2015). 454 Life Sciences (Roche, Basel, Switzerland) stopped the support of the platform after 2016 due to the substantially lower costs for the Illumina technology (Hodkinson & Grice, 2015). The Illumina technology dominates currently the market due to the production of high-throughput reads at moderate cost (El-Metwally *et al.*, 2013; Reuter *et al.*, 2015). The Illumina MiSeq platform provides long reads with 300 bases, also as paired-end reads. The HiSeq 2500 is the platform with greatest performance output, producing 8 billion fragments in a paired-end fashion with 125 bases for each read in 7 h to 60 h. The costs dropped to 0.03 USD/Mb (Rhoads & Au, 2015). Arrays of HiSeq machines, such as HiSeq X Ten (released 2014), provide an even higher throughput (Hodkinson & Grice, 2015).

The disadvantage of the second generation technologies is primarily the short length of reads which makes assembly a complex and difficult challenge (Henson *et al.*, 2012). The performance of *de novo* assembly software depends heavily on the sequence length, depth of sequence coverage, fragment size of the templates that are sequenced and the types of sequence errors specific to each technology (Tsai *et al.*, 2010). No single assembler is optimal in every possible quality metric (Wences & Schatz, 2015) and consequently, draft genome assemblies vary in their quality (Tsai *et al.*, 2010). The assembled data generally are highly

fragmented with many gaps (Tsai *et al.*, 2010) due to difficulties in assembling repeat regions and to sequencing biases. Moreover, the assembled fragments frequently contain errors, either due to sequencing artefacts or to the incorrect reconstruction of repeats (Nagarajan *et al.*, 2010). Therefore, to obtain a closed genome directly with an Illumina dataset of high coverage is rather unlikely (El-Metwally *et al.*, 2013).

Technical errors in the assembled sequence can be automatically corrected due to a high level of coverage. Assembly software is designed to consider NGS-platform specific sequence error profiles. However, the resulting assemblies are usually more fragmented (Nagarajan *et al.*, 2010). The closing of the genome sequence requires a time- and cost-intensive process of additional experiments. The standard strategy to closure gaps usually involves approaches such as directed-PCR and primer-walking in combination with dideoxy sequencing. Pairs of adjacent contigs are identified, following by the determination of genomic sequences which span the gaps between them (Tsai *et al.*, 2010; Nagarajan *et al.*, 2010). Because the closure of the complete genome can involve month of lab work and thousands of finishing experiments, the task is usually done in large genome centres (http://www.cbcb.umd.edu). The prohibitive analysis cost of finishing appears to only be justified for high-priority genomes (Nagarajan *et al.*, 2010). As a result, increasingly large number of sequenced genomes remains unfinished, at a "permanent draft" stage, which are used for subsequent analysis (Mavromatis *et al.*, 2012; Paulino *et al.*, 2015).

An alternative approach is the third-generation sequencing (TGS) technology capable to provide much longer reads (Rhoads & Au, 2015). The PacBio RS was the first commercially available long read sequencer (Koren & Phillippy, 2015), enabling single molecule real time (SMART) sequencing (Quail *et al.*, 2012). RS II (P6-C4) generates reads with average length of 1.0–1.5 x 10^4 bp (Rhoads & Au, 2015). 3.5–7.5 x 10^4 reads can be produced per run in 0.4 h to 6 h (Rhoads & Au, 2015). The long read lengths make the technology well-suited for unsolved problems in genome assembly, for example in complex regions involving repeated elements or segmented duplications (Rhoads & Au, 2015; Madoui *et al.*, 2015). PacBio sequencing offers much longer read length than SGS methods but is hindered by a lower throughput, higher error rates, and higher costs per base (Rhoads & Au, 2015), which are 0.4–0.8 USD/Mb using the RS II P6-C4 system (Rhoads & Au, 2015). Therefore, Illumina as most cost-economic sequencing method is expected to be intensively used in the near future.

Often draft assemblies are sufficient for many genomic analyses, especially if complete sequences of closely related organisms are available (Nagarajan *et al.*, 2010).

However, long contigs are essential for most downstream applications, such as annotation and variant calling (Ekblom & Wolf, 2014). Consequently, obtaining better draft genomes is a common goal (Paulino *et al.*, 2015). There is currently an urgent need for tools to efficiently improve the draft assemblies without the additional cost of manual finishing (Tsai *et al.*, 2010; Paulino *et al.*, 2015).

In this study, we proposed a strategy for *in silico* improving of *de novo* draft assemblies utilizing the same set of Illumina short reads that were used to generate the original *de novo* assembly. Metawatt-3.5.2 binning was introduced as taxonomic analysis tool to differentiate contaminations from the "organism" bin that was then used as guided assembly for the next round. Iterative rounds of targeted read mapping to taxonomically coherent contigs and the *de novo* reassembly of the mapped subset of reads are used to improve the assembly. Contigs are again taxonomically classified. A last improvement step is a final *de novo* assembly of extended and overlapping contigs using Geneious and a manual inspection of the resulting draft genome.

3.2 Materials and methods

3.2.1 Bacterial strains

This study used '*Rhodopirellula bahusiensis*' strain SWK21 (= DSM 24079), which was isolated from the surface of a macroalgae sampled at Tjärnö, Sweden (58.8764 N 11.1447 E), and '*Rhodopilula apulia*' strain SM50 (= DSM 24084), which was isolated from a mixed sediment water sample originating from San Cataldo, Italy (40.3861 N 18.3055 E) (Winkelmann & Harder, 2009).

3.2.2 DNA extraction and sequencing

The genomic DNA of both strains was extracted using the FastDNA SPIN Kit for Soil (MP Biomedicals, Germany). Sequencing was performed by the Illumina MiSeq technology (2 x 300 bp) for genomic DNA of strain SWK21 and by the Illumina HiSeq technology (2 x 150 bp) for genomic DNA of strain SM50 (Fig. 1) by the Max Planck-Genome-centre Cologne (http://mpgc.mpipz.mpg.de/home/).

3.2.3 Preprocessing of Illumina raw reads

Obtained raw reads were quality checked with FastQC (www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) and dynamically trimmed with SolexaQA v.3.1.4 (Cox *et al.,* 2010) (Fig. 1). The default quality cutoff was p = 0.05. To perform filtering on read lengths, the "lengthsort" command of SolexaQA v.3.1.4 was performed after the initial trimming command. The paired-read files were combined, passed single reads were added and then normalized with Khmer 2.0 (Crusoe *et al.*, 2014) (Fig. 1A). The program was run twice, once with coverage (C) = 20 and then run again on the new file with C = 5, both passes with a kmer size of 20. Paired sequences and orphaned sequences were separated. Each file was treated separately using Trimmomatic (version 0.32) (Bolger *et al.*, 2014) (Fig. 1A). This trimming tool was used to clean thoroughly the beginnings and the ends of reads. A two-step procedure was implemented. In short, the first step included quality trimming functionality with a sliding window (SLIDINGWINDOW:4:15), cutting of adapters (ILLUMINACLIP:adapters.fa:2:30:10), removal of low quality bases off the start (LEADING:3) and the end (TRAILING:3) of a read based on a threshold, and dropping of short reads (MINLEN:36), using phred + 33 quality score. Specified number of bases was then removed from the start (HEADCROP) and the end (CROP) of the reads in a second step. The usage either of one option or both, and the specific number of trimmed bases were determined individually for each file by the consideration of summary graphs provided by FastQC.

3.2.4 *De novo* assembly

Preprocessed Illumina reads were assembled *de novo* using SPAdes-3.8.0 (Bankevich *et al.*, 2012) (Fig. 1A). The SPAdes script "spades.py" was performed with the option for single cell data (--sc). In addition, we allowed the assembler to incorporate its own error correction routines to do further correction in addition to the pre-processing of reads by setting the option "--careful".

3.2.5 Improving of draft genome assembly completeness

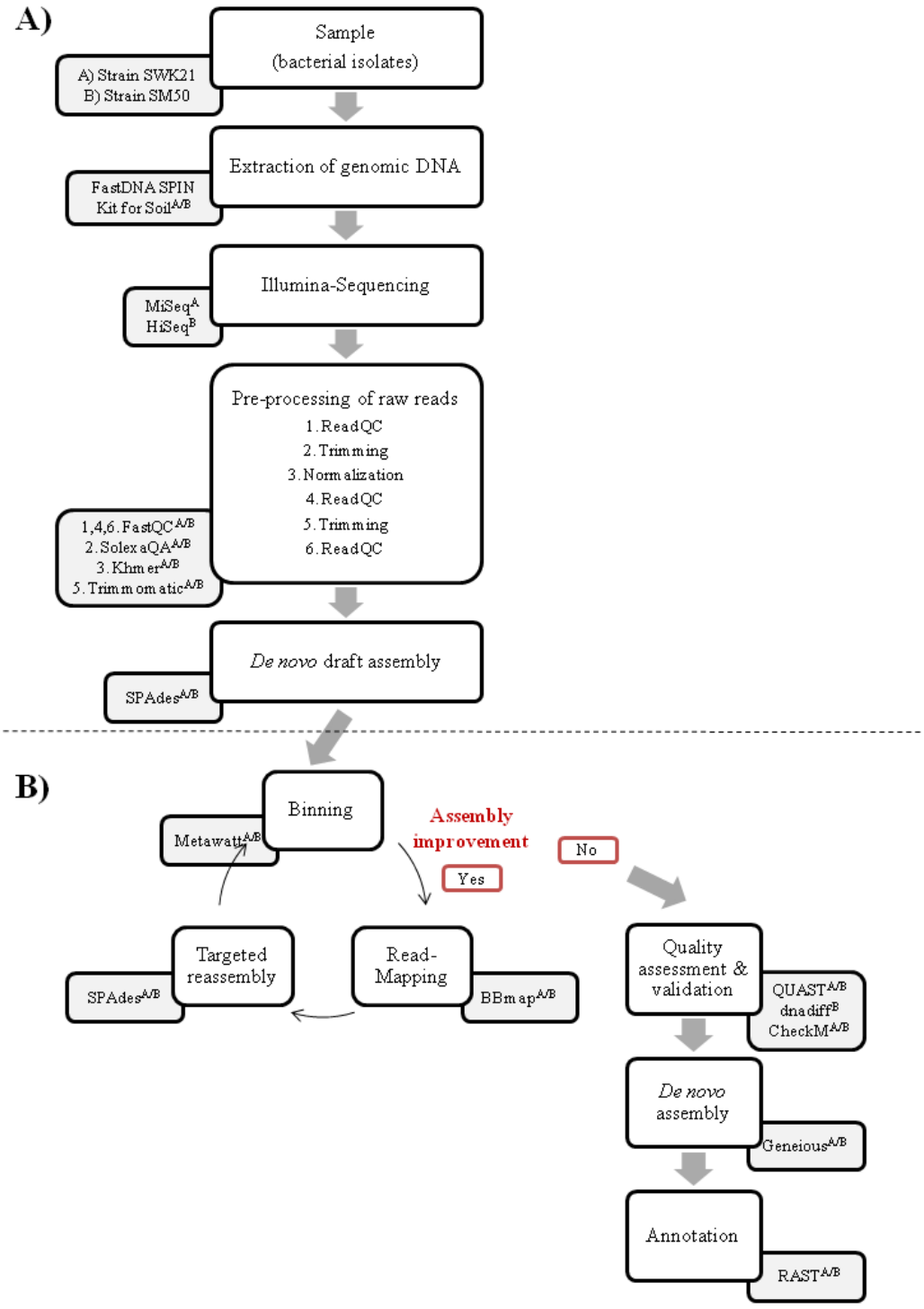


Fig. 1. Workflow for improvement of *de novo* draft assemblies. A) Steps to generate the original *de novo* draft assemblies. B) Steps of the developed polishing approach. Details of the individual steps are explained in the text. Individual tool was applied on data of [A] strain SWK21, [B] strain SM50, or [A/B] both strains. QC = quality control.

The binning function of Metawatt-3.5.2 (Strous *et al.*, 2012) was used to obtain a taxonomic affiliation of the contigs of the first assembly. The classification allowed the differentiation of a set of taxonomically coherent contigs in an "organism" bin from contigs without a taxonomic affiliation on the basis of multiple criteria, such as strong deviations of GC content, tetranucleotide frequencies and coverage. The contigs without taxonomic affiliation were grouped by Metawatt-3.5.2 into an "unbinned" bin. Additionally, the taxonomic profiling allowed a verification of the expected genome affiliation, if the target genome was highly similar to a reference genome in the database used. The bin of the target organism was picked using the taxonomic level "order". The strategy of binning and targeted reassembly has been introduced by Meier *et al.* (2016) for binning refinements of metagenomes. For the assembly of genomes of strains, the procedure was further developed.

The preprocessed Illumina reads were aligned to the target bin using BBmap of the BBmap package (version 35.14, http://sourceforge.net/projects/bbmap/). The mapped reads were *de novo* reassembled again, usually improving the initial draft assembly. The same datasets of reads were used for each round of mapping. For the first round, the minimum identity value was set relatively low with 80%. The read mapping procedure was performed separately for each paired-end read file and single read file of a genome project. All the reads of the different procedures that mapped were reassembled together with SPAdes-3.8.0. The option for single cell data (--sc) was included. Again, the binning function of Metawatt-3.5.2 was used to pick the new draft assembly. Taxonomically non-related contigs, identified as "unbinned" in Metawatt-3.5.2, were not included in the next round of mapping. The summarized statistic of the input data (number of contigs, bin size (nt), and N50 contig length) were used as quality parameter to assess the new draft assembly. An optimization was reached by the realization of at least one of these criteria. In case of an increased degree of quality, the reassembly round consisting of mapping, reassembling, and taxonomic analysis by binning, was then run repeatedly with the newly defined bin as reference. For all iterations, the minimum identity value for read mapping was set of 95% and the mapped reads were reassembled with the SPAdes assembler, supported by the usage of "trusted contigs". All contigs, longer than 10,000 bp, of the new picked bin were selected.

If the defined quality parameters in Metawatt indicated no further optimization of the draft assembly, the procedure of iterative rounds was stopped. Then quality assessment and validation were performed to find the core bin of highest quality for further improvement. The draft assemblies both before the improvement by iterative reassembly rounds and of each taxon-specific reassembly were analysed by the web interface of QUAST (Gurevich *et al.*,

2013), which was run with default parameters. The resulted report of the multiple assemblies supported the selection of the bin with highest quality. Because assembly metrics, such as the N50 size and contig number, may not always correlate well with actual quality of the assembly (Utturkar *et al.*, 2014), the following other tools were incorporated. Detailed information about differences between two assemblies was provided using dnadiff (Phillippy *et al.*, 2008) from MUMmer v.3.23 (Kurtz *et al.*, 2004). The accuracy of completeness and contamination of the polished assemblies were estimated using CheckM 0.9.7 (Parks *et al.*, 2015). Completeness was also indicated in Metawatt-3.5.2.

The commercial bioinformatics program Geneious (Biomatters Ltd., Auckland, New Zealand; Kearse *et al.,* 2012) was used to find overlapping positions of contigs from the selected taxon-specific bin of highest quality. A "*de novo* assembly" of the contigs was performed using custom sensitivity settings. High stringent parameters (no allowed gaps or mismatches per read, a minimum overlap identity of 100%, a maximum gap size of 2, and a maximum ambiguity of 1) were set to find only true overlaps. The produced contigs were inspected manually and consensus sequences were combined with the non-assembled contigs to create the final permanent draft genome.

3.2.6 Draft genome annotation and comparative analyses

The updated draft assemblies were automatically annotated by the Rapid Annotations using Subsystems Technology (RAST) server (Aziz *et al.*, 2008). Calculations of average nucleotide identities (ANIs) with published whole and draft *Rhodopirellula* genomes were determined by JSpecies (Richter & Rosselló, 2009).

3.3 Results

We developed an approach to improve draft assemblies obtained from Illumina reads. Our approach consisted of an iteration of taxon-specific binning of contigs, mapping of reads against the taxon-specific set of contigs, and reassembling of the reads that were recruited by mapping (Fig. 1B). *Planctomycetes* are well known for their large genomes and we optimized two *de novo* draft assemblies for the genomes of '*Rhodopirellula bahusiensis*' strain SWK21 and '*Rhodopilula apulia*' strain SM50.

3.3.1 Example 1: '*Rhodopirellula bahusiensis'* strain SWK21

3.3.1.1 De novo draft assembly

Illumina MiSeq sequencing of the genomic DNA of strain SWK21 generated a total of 9,030,614 (x 2) paired-end reads (Fig. 3). 1,858,082 read pairs and 12,770 single reads were obtained after trimming and normalization using SolexaQA v.3.1.4 and Khmer 2.0, respectively. The preprocessed data were trimmed using Trimmomatic v.0.32 to remove especially the beginnings and ends of reads with low quality, resulting in 1,851,828 read pairs and 12,580 singletons (Fig. 2; Fig. 3).

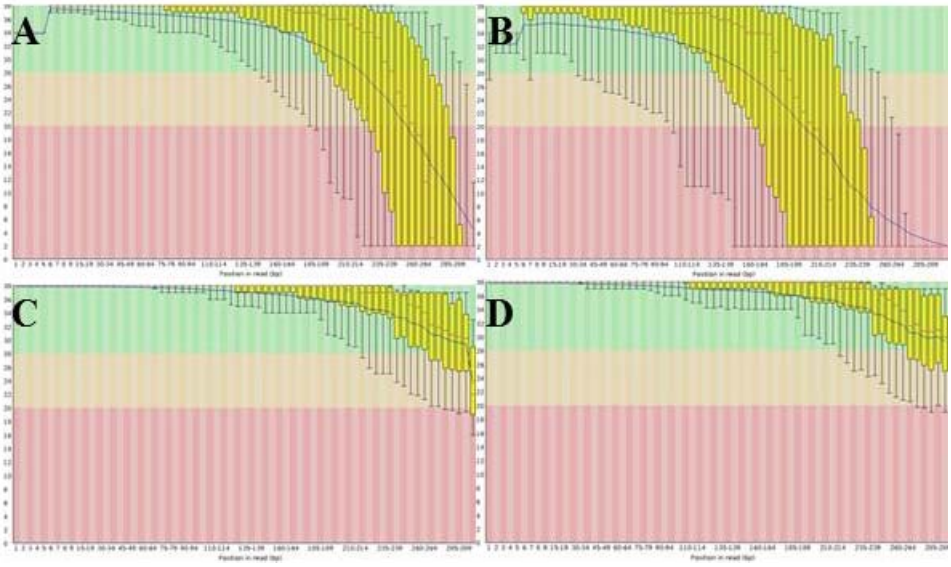


Fig. 2. Overview of the range of quality values across all bases at each position in the FastQ file of A) raw reads in forward direction, B) raw reads in reverse direction, C) final preprocessed paired-end reads, and D) final preprocessed singletons of one performed sequencing run, provided by FastQC. The x-axis describes the position in read (bp), the y-axis provides the quality score.

These high-quality reads were assembled *de novo* using SPAdes V3.8.0. The mismatch correction of the assembler failed for unknown reasons. The assembly consisted of 236 contigs with a total size of 7.91 Mb (Fig. 3).

3.3.1.2 Polishing of draft assembly by targeted reassembling

The binning of the assembly in Metawatt-3.5.2 with a taxonomic profile of the rank "order" assigned 95 contigs in a "*Planctomycetales*" bin and 115 contigs in an "unbinned" bin. The latter had a low N50 size (Table S1) indicating short contigs. The "*Planctomycetales*" bin served as reference for the mapping and in the second and following iterative rounds it served as "guided assembly" for the assembly of recruited reads. Three iterative rounds showed an improvement in at least one of the quality parameters: assembly size, contig number, or N50 size (Fig. 3). Each reassembly yielded the same taxonomic profiling in Metawatt with the classification of a single taxon-specific bin (Table S1). The "*Planctomycetales*" bin improved to 75 contigs with a total size of 7.79 Mb. This bin had the highest quality according to QUAST calculations (Fig. 3).

The contigs of the SPAdes assembly were reassembled using a stringent *de novo* assembly by custom settings in Geneious to identify overlapping regions of contigs. Manual inspection of the assembled contigs confirmed that 18 of 75 contigs were correctly assembled into six new contigs. The final permanent draft genome for strain SWK21 comprises 63 contigs (Fig. 3) and represents a nearly complete genome of 99.93% and of 89.4% completeness according to CheckM and Metawatt-3.5.2, respectively. This final draft assembly was annotated using RAST (Table 1). ANIs with published complete and draft genomes of the genus *Rhodopirellula* confirmed a novel species in the genus *Rhodopirellula* (Table S3).

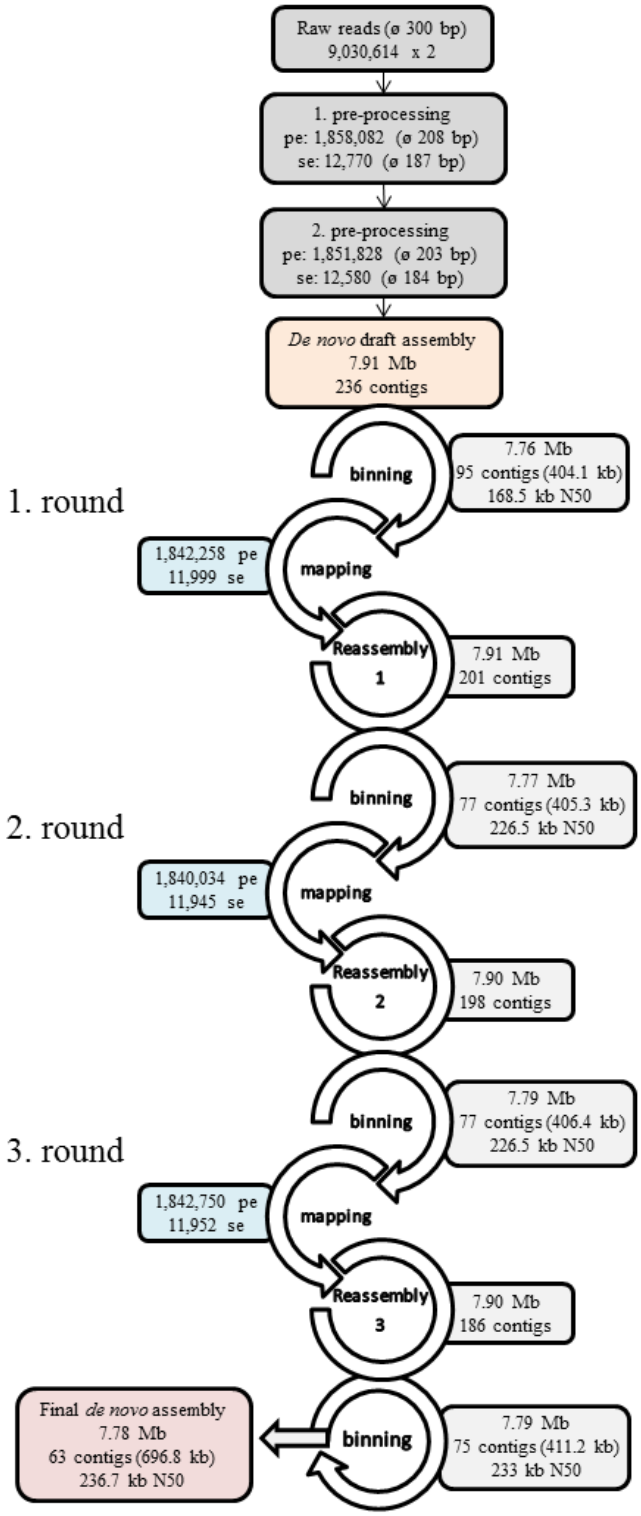


Fig. 3. Overview of the iterative approach to obtain a permanent draft genome for strain SWK21 with details to raw and preprocessed reads (dark gray boxes), initial *de novo* draft assembly (orange box), reassemblies or taxon-specific bins (light gray boxes), mapped read data (blue boxes), and permanent draft genome of strain SWK21 (pink box). Statistics were obtained by Metawatt and QUAST. Behind the contig number, the largest contig size is provided in brackets. pe: paired-end. se: singletons.

3.3.2 Example 2: '*Rhodopilula apulia'* strain SM50

3.3.2.1 De novo draft assembly

'*Rhodopilula apulia*' strain SM50 represents a novel genus and was selected based on the taxonomic distance to known genomes as study case. The Illumina HiSeq sequencing of the genomic DNA of strain SM50 yielded 18,135,378 (x 2) paired-end reads with an average length of 150 bp (Fig. 4). The pre-processing resulted in 2,625,722 read pairs and 41,906 singletons using SolexaQA v.3.1.4 and Khmer 2.0. Trimming using Trimmomatic v.0.32 reduced the read number to 2,608,370 read pairs and 32,178 singletons. SPAdes V3.8.0 assembled the reads into 1,684 contigs with a total size of 10.28 Mb (Fig. 4).

3.3.2.2 Improvement of the draft assembly

Metawatt assigned the contigs of the assembly to seven taxon-specific bins (Table S2). 1,252 short contigs remained as unbinned. The taxon bin "*Planctomycetales*" was used as reference for read mapping and reassembling. Three iterative cycles were performed to improve the assembly. Metawatt always clustered the contigs of the actual reassembly in several bins. As we only selected the *"Planctomycetales"* bin for the next iterative cycle, it was interesting to obtain for contigs taxonomic assignments outside of *"Planctomycetales"*. The third iterative cycle yielded an increase number of contigs, a lower total base number and a lower N50. Dnadiff was used to compare the *"Planctomycetales"* bins of the second and third cycle. 99.74% bases of the bin obtained after the second performed round were aligned to the other bin. Consequently, the bins of the second cycle were selected as high-polished draft assembly (Fig. 4).

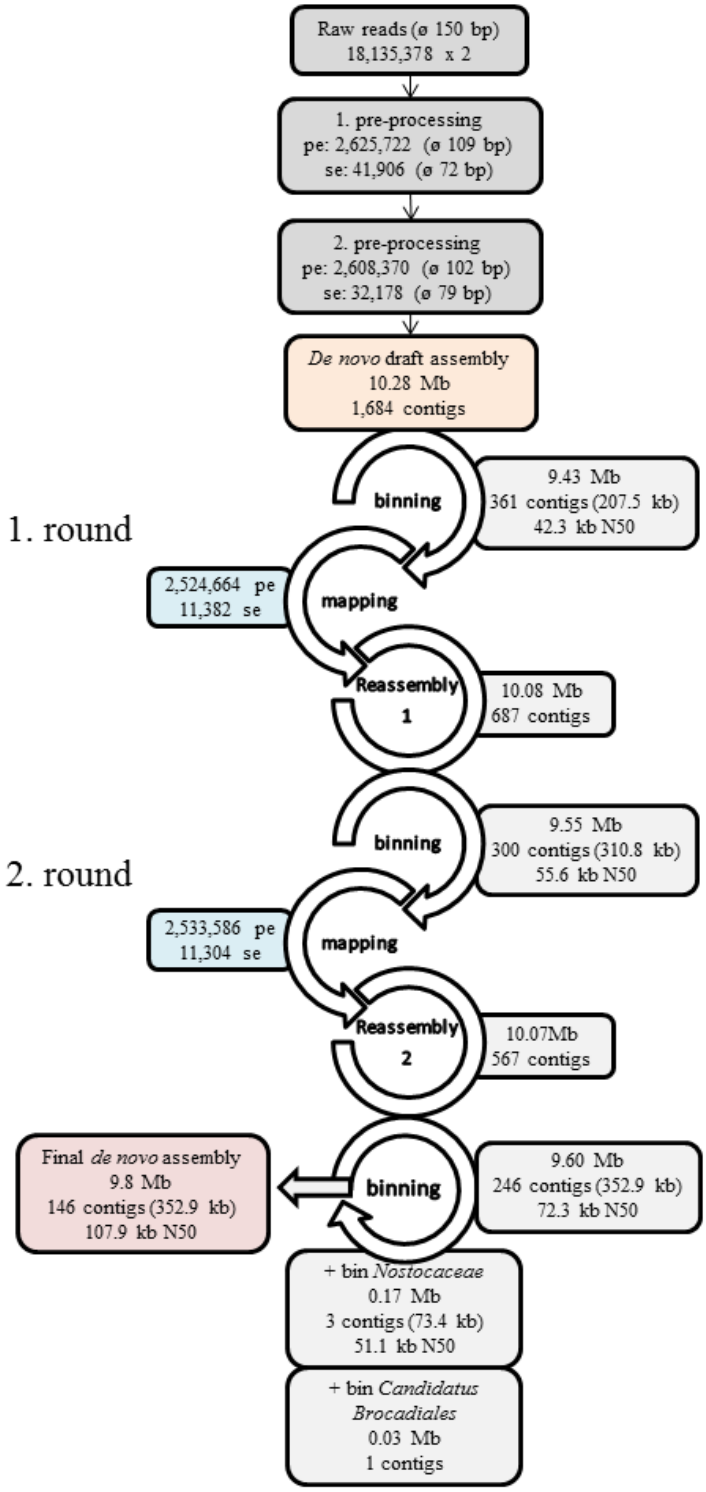


Fig. 4. Overview of the iterative approach to obtain a permanent draft genome for strain SM50. Detailed information of raw and preprocessed reads (dark gray boxes), the initial draft assembly (orange box), reassemblies or selected bins (light gray boxes), mapped read data (blue boxes), and the permanent draft genome (pink box) are shown. Statistics were obtained by Metawatt and QUAST. The largest contig size is shown in brackets behind the contig number. pe: paired-end. se: singletons.

For the *de novo* assembly of contigs in Geneious, we considered the taxonomic position of strain SM50 as representative for a novel genus and included also the contig of "*Nostocaceae*" and the three contigs of "*Candidatus Brocadiales*" (Table S2). The assembly aligned 163 of 250 contig sequences to 59 new contigs, which were confirmed by manual inspection. One "*Nostocaceae*" contig and one "*Candidatus Brocadiales*" contig assembled with "*Planctomycetales*" contigs, thus suggesting that they indeed belong to the genome of SM50. The permanent draft genome of strain SM50 consisted finally of 146 contigs (Fig. 4), representing a nearly complete genome of 99.93% and of 91.1% using CheckM and Metawatt-3.5.2, respectively. Annotation was performed in RAST (Table 1) and the calculation of ANI values with published complete and permanent draft genomes of the genus *Rhodopirellula* were implemented in the JSpecies software (Table S4).

Table 1. Annotation results from RAST, the completeness from CheckM ([1]) and Metawatt ([2]), and contamination from CheckM ([1]) for the final improved draft assemblies.

	Strain SWK21	**Strain SM50**
Draft genome size (bp)	7,784,569	9,795,212
Contig number	63	146
GC content	55.8	59.0
N50	236,729	107,901
L50	10	31
Number of coding sequences	7,413	7,925
Number of RNAs	67	82
Closest neighbours with highest scores (score)	*Pirellula* sp. 1 (542)* *Rhodopirellula baltica* SH1^T (536)	*Pirellula* sp. 1 (546)* *Rhodopirellula baltica* SH1^T (535)
Completeness (%)	99.93[1]/89.4[2]	99.93[1]/91.1[2]
Contamination (%)	1.16[1]	0.00[1]

* *Pirellula* sp. 1 (presently *Rhodopirellula baltica* SH1^T) is validly described as type strain of *Rhodopirellula baltica* gen. nov., sp. nov. (Schlesner *et al.*, 2004).

3.4 Discussion

A complete and fully automated assembly of genomes using SGS short reads remains a significant bioinformatics challenge. Areas of low sequence coverage, repetitive elements and short read length often result in highly fragmented *de novo* draft assemblies (Paulino *et al.*, 2015). Manual finishing is laborious and slow (Tsai *et al.*, 2010). TGS technology, such as PacBio RS, capable of much longer reads and faster runs than SGS, offers an alternative approach to overcome many of the limitations (Rhoads & Au, 2015). But the advantages of PacBio and SGS are complementary. A lower throughput, higher error rate, and higher costs hindered PacBio to outperform Illumina sequencing technology. Thus, it is critical to develop tools to efficiently improve the short read *de novo* draft assemblies (Rhoads & Au, 2015; Paulino *et al.*, 2015).

We developed an *in silico* approach to increase the quality of microbial *de novo* draft assemblies. By applying this strategy, original genomic assemblies are substantially improved by targeted reassembly, often with a trusted assembly as starting point. This strategy of aligning sequences against contigs ends and the generation of *de novo* reassemblies is not a new idea. However, the implementation of iterative rounds of reassembling is mainly based on the usage of additional data from other sequencing technologies. The generation of reads with increased length can increase the fraction of the genomes that can be fully resolved (Koren & Phillippy, 2015; Sangwang *et al.*, 2016). Such hybrid sequencing strategies combine the strengths of SGS (high-throughput and high sequencing accuracy) and TSG (long reads) (Rhoads & Au, 2015). However, this finishing procedure also results in additional costs.

We were interested to improve *de novo* draft assemblies without the generation of new sequences and also without an available reference genome. Our approach resolved gaps by a separation of contaminating reads with a taxonomic classifier. Mapping of reads to the taxonomically coherent assembly enabled a reassembly that usually established new linkage information. The absence of contaminating reads may be a crucial factor for the success of the reassembly caused by the reduction of number of positions to which a read can be aligned.

The integration of a taxonomic binning software has been recently described for metagenomes (Meier *et al.*, 2016; Laso-Pérez *et al.*, 2016; Sedlar *et al.*, 2017). We applied binning on single-genome draft assemblies which allowed the differentiation and separation of non-phylogenetically related contigs which consequently increased the quality of the assembly. The unbinned contigs are very short sequences with low information content.

Genome projects that target the reconstruction of complete genomes will require additional efforts beyond the described procedure on Illumina data sets. The longer reads of TGS can range across long regions of repetitive sequences. Because of the high cost, the important first step is to thoroughly consider whether a closed genome sequence is necessary for addressing the biological question at hand (Ekblom & Wolf, 2014). For many downstream analyses high-quality permanent draft genomes are often sufficient and can often be obtained by the application of few bioinformatics tools. Nagarajan *et al.* (2010) called it "a small bag of tricks that can make the task less daunting".

We selected the strain SWK21, closely related to the genus *Rhodopirellula*, and the strain SM50, a distantly related strain, to demonstrate our approach. With a 16S rRNA gene identity of over 98.7% to *Rhodopirellula baltica* SH1^T (Table S3), strain SWK21 was assigned to the genus *Rhodopirellula* (Winkelmann & Harder, 2009; Frank, 2011). Functional genes as phylogenetic marker (*carB* gene) clearly showed the formation of a different lineage in the phylogenetic tree and consequently a new species tentatively named '*Rhodopirellula bahusiensis*' (Žure, 2015). Based on a gene identity of 93.7% for the complete 16S rRNA gene to *Rhodopirellula baltica* SH1^T (Table S4), the strain SM50 represents a novel genus (Winkelmann & Harder, 2009) and is tentatively named '*Rhodopilula apulia*' (Žure *et al.*, 2015). To assess the extent of their genetic relatedness to *Rhodopirellula baltica* SH1^T, nearly complete genomes are needed. The final permanent draft genomes confirmed previous insights in the phylogeny of planctomycetes (Table S3; Table S4).

The demonstration of our strategy on two different bacterial strains also represents the flexibility of the strategy: each step of the approach can be adapted to the individual datasets. This is important because no single approach is the best for all applications (Koren & Phillippy, 2015). '*Rhodopilula apulia*' strain SM50 represents a novel genus of bacterial phylum *Planctomycetes*. Binning resulted in the taxonomic assignment of assembled contigs outside of "*Planctomycetales*" using Metawatt. The taxonomic profiling can failed for target genomes without high similarity to a reference genome in the database used. Bioinformatics tools – not only assemblers – have weaknesses and all results should be considered critically. Therefore, we performed a final *de novo* assembly of extended contigs using Geneious as well as the manually inspection of overlapping regions. This last improvement step clarifies the affiliation.

In addition, Geneious can help to overcome assembly problems based on the usage of NGS data. Obtained reads may have arisen from two different copies of the same genome sequence, which complicates assemblies. If the repetitive region longer than the longest read,

the assembled sequences end at the boundaries of such repeats (Henson *et al.*, 2012). The visual inspection of mapping results in Geneious can contribute to an understanding of these assembly problems. Therefore, gains in ideal assembly quality can be made with the extension of contigs by read mapping and a final assembling in Geneious.

Our approach is utilizable for sequencing projects targeting a complete genome. The effort of manual finishing can be reduced by supporting the initial closing of many gaps using our *in silico* strategy. Furthermore, the integration of a binning step allows using the strategy for both genomic and metagenomic data. The organism of interest can be initially picked from draft metagenome assemblies and then improved.

3.5 Conclusion

We developed a bioinformatics approach to substantially improve draft genome assemblies by the identification of taxonomically related contig sequences using metagenome binning software and a refinement based on the taxon-specific assembly. Thereby, taxonomically unrelated contigs were removed and the reassembly of recruited reads provided longer contigs. The iterative cycle can be performed until no further improvement. The obtained contigs were finally used for an assembly to identify overlapping contigs and, after manual curation, to obtain a permanent draft genome.

Acknowledgments

We thank Christina Probian for growth of the cells and DNA preparation. This study was funded by the Max Planck Society.

References

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Cox MP, Peterson DA, Biggs, PJ. 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485.

Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, Charbonneau A, Constantinides B, Edvenson G, Fay S, Fenton J, Fenzl T, Fish J, Garcia-Gutierrez L, Garland P, Gluck J, González I, Guermond S, Guo J, Gupta A, Herr JR, Howe A, Hyer A, Härpfer A, Irber L, Kidd R, Lin D, Lippi J, Mansour T, McA'Nulty P, McDonald E, Mizzi J, Murray KD, Nahum JR, Nanlohy K, Nederbragt AJ, Ortiz-Zuazaga H, Ory J, Pell J, Pepe-Ranney C, Russ ZN, Schwarz E, Scott C, Seaman J, Sievert S, Simpson J, Skennerton CT, Spencer J, Srinivasan R, Standage D, Stapleton JA, Steinman SR, Stein J, Taylor B, Trimble W, Wiencko HL, Wright M, Wyss B, Zhang Q, Zyme E, Brown CT. 2014. The khmer software package: enabling efficient sequence analysis. F1000Res 4:900.

Frank CS. 2011. Polyphasische Taxonomie, Kerngenom und Lebenszyklus von *Rhodopirellula*-Stämmen. PhD thesis, Universität Bremen, Germany.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075.

Henson J, Tischler G, Ning Z. 2012. Next-generation sequencing and large genome assemblies. Pharmagenomics 13:901–915.

Hodkinson BP, Grice EA. 2015. Next-generation sequencing: a review of technologies and tools for wound microbiome research. Adv Wound Care (New Rochelle) 4:50–58.

Ekblom R, Wolf, JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042.

El-Metwally S, Hamza T, Zakari M, Helmy M. 2013. Next-generation sequence assembly: four stages of data processing and computational challenges. PLoS Comput Biol 9:e1003345.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Koren S, Phillippy AM. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 23:110–120.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12.

Laso-Pérez R, Wegener G, Knittel K, Widdel F, Harding KJ, Krukenberg V, Meier DV, Richter M, Tegetmeyer HE, Riedel D, Richnow H, Adrian L, Reemtsma T, Lechtenfeld OJ, Musat F. 2016. Thermophilic archaea activate butane via alkyl-coenzyme M formation. Nature 539:396–401.

Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, Lemainque A, Wincker P, Aury JM. 2015. Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics 16:327.

Mavromatis K, Land ML, Brettin TS, Quest DJ, Copeland A, Clum A, Goodwin L, Woyke T, Lapidus A, Klenk HP, Cottingham RW, Kyrpides NC. 2012. The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. PLoS One 7:e48837.

Meier DV, Bach W, Girguis PR, Gruber-Vodicka HR, Reeves EP, Richter M, Vidoudez C, Amann R, Meyerdierks A. 2016. Heterotrophic *Proteobacteria* in the vicinity of diffuse hydrothermal venting. Environ Microbiol 18:4348–4368.

Nagarajan N, Cook C, Di Bonaventura MP, Ge H, Richards A, Bishop-Lilly KA, DeSalle R, Read TD, Pop M. 2010. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. BMC Genomics 11:242.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

Paulino D, Warren, RL, Vandervalk BP, Raymond A, Jackman SD, Birol I. 2015. Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics 16:230.

Phillippy AM, Schatz MC, Pop M. 2008. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 9:R55.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341.

Reuter JA, Spacek D, Snyder MP. 2015. High-throughput sequencing technologies. Mol Cell 58:586–597.

Rhoads A, Au KF. 2015. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13:278–289.

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

Sangwan N, Xia F, Gilbert JA. 2016. Recovering complete and draft population genomes from metagenome datasets. Microbiome 4:8.

Schlesner H, Rensmann C, Tindall BJ, Gade D, Rabus R, Pfeiffer S, Hirsch P. 2004. Taxonomic heterogeneity within the *Planctomycetales* as derived by DNA-DNA hybridization, description of *Rhodopirellula baltica* gen. nov., sp. nov., transfer of *Pirellula marina* to the genus *Blastopirellula* gen. nov. as *Blastopirellula marina* comb. nov. and emended description of the genus *Pirellula*. Int J Syst Evol Microbiol 54:1567–80.

Sedlar K, Kupkova K, Provaznik I. 2017. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Comput Struct Biotechnol J 15:48–55.

Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. 2012. The binning of metagenomic contigs for microbial physiology of mixed cultures. Front Microbiol 3:410.

Tsai IJ, Otto TD, Berriman M. 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol 11:R41.

Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, Brown SD. 2014. Evaluation and validation of *de novo* and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics 30:2709–2716.

Wences AH, Schatz MC. 2015. Metassembler: merging and optimizing *de novo* genome assemblies. Genome Biol 16:207.

Winkelmann N. 2009. Isolierung und genetische Charakterisierung von Stämmen des Planctomycetengenus *Rhodopirellula* und die Biogeographie in europäischen Gewässern. PhD thesis, Universität Bremen, Germany.

Winkelmann N, Harder J. 2009. An improved isolation method for attached living *Planctomycetes* of the genus *Rhodopirellula*. J Microbiol Methods 77:276–284.

Žure M. 2015. Biogeography of *Rhodopirellula* in European coastal sediments. PhD thesis, Universität Bremen, Germany.

Žure M, Munn CB, Harder J. 2015. Diversity of *Rhodopirellula* and related planctomycetes in a North Sea coastal sediment employing *carB* as molecular marker. FEMS Microbiol Lett 362:fnv127.

Appendix of Chapter 3:

Supplementary tables

Table S1. Overview about properties of input assemblies and their produced bins for the strain SWK21.

Binning character	**Taxon**	**Size in Mb (Largest contig in kb)**	**Contig number**	**N50**
Initial assembly				
Original	unbinned	7.91 (404.1)	236	168,554
After binning	*Planctomycetales*	7.76 (404.1)	95	168,554
	unbinned	0.07 (4.1)	115	1,245
First Reassembling				
Reassembly	unbinned	7.91 (405.3)	201	226,510
After binning	*Planctomycetales*	7.77 (405.3)	77	226,510
	unbinned	0.07 (3.7)	106	1,030
Second Reassembling				
Reassembly	unbinned	7.90 (406.4)	198	226,510
After binning	*Planctomycetales*	7.79 (406.4)	77	226,510
	unbinned	0.07 (3.5)	102	993
Third Reassembling				
Reassembly	unbinned	7.90 (411.2)	186	232,953
After binning	*Planctomycetales*	7.79 (411.2)	75	232,953
	unbinned	0.07 (3.5)	97	1,027

Table S2. Overview about properties of input assemblies and their produced bins for the strain SM50.

Binning character	Taxon	Size in Mb (Largest contig in kb)	Contig number	N50
Initial assembly				
Original	unbinned	10.28 (207.5)	1,684	40,298
After binning	*Planctomycetales*	9.43 (207.5)	361	42,311
	Nostocaceae	0.09 (43.6)	4	39,090
	Candidatus Brocadiales	0.03 (34.1)	1	34,111
	Spartobacteria	0.07 (31.6)	3	28,021
	Lentisphaeria	0.03 (17.9)	2	17,871
	Deltaproteobacteria	0.09 (15.5)	11	10,956
	Verrucomicrobiae	0.06 (14.9)	7	11,744
	unbinned	0.33 (3.7)	1,252	302
First Reassembly				
Original	unbinned	10.08 (310.8)	687	51,624
After binning	*Planctomycetales*	9.55 (310.8)	300	55,557
	Nostocaceae	0.10 (54.3)	2	54,321
	Chroococcales	0.05 (54.0)	2	54,041
	Candidatus Brocadiales	0.03 (32.3)	1	32,280
	Spartobacteria	0.03 (31.9)	1	31,926
	Burkholderiales	0.03 (17.2)	3	7,910
	Deltaproteobacteria	0.03 (12.1)	6	10,747
	unbinned	0.12 (3.9)	348	458
Second Reassembly				
Original	unbinned	10.07 (352.9)	567	67,761
After binning	*Planctomycetales*	9.60 (352.9)	246	72,329
	Nostocaceae	0.17 (73.4)	3	51,099
	Candidatus Brocadiales	0.03 (32.3)	1	32,280
	unbinned	0.10 (3.2)	282	457

Table S3. 16S rRNA gene similarities and average nucleotide identities (ANIs) of published complete and draft *Rhodopirellula* genomes and final draft genome of strain SM50 to the final draft genome of '*Rhodopirellula bahusiensis'* SWK21.

Genome	Genome size (bp)	16S rRNA dissimilarity[1]	ANIm (%)	Aligned bases (%)	Aligned bases (bp)
***'Rhodopirellula bahusiensis'* SWK21**	7,784,569	*	*	*	*
***'Rhodopilula apulia'* SM50**	9,795,212	7.1	83.90	0.91	71,072
***Rhodopirellula baltica* SH1**	7,145,576	0.0	86.38	52.31	4,072,177
***Rhodopirellula baltica* SWK14**	7,488,930	0.0	86.59	53.49	4,163,902
***Rhodopirellula baltica* SH28**	7,149,689	0.2	86.86	54.82	4,267,688
***Rhodopirellula baltica* WH47**	7,033,319	0.0	86.43	52.93	4,120,424
***Rhodopirellula europaea* 6C**	7,191,307	0.1	88.67	56.15	4,371,385
***Rhodopirellula europaea* SH398**	7,446,194	0.1	89.41	60.57	4,715,393
***Rhodopirellula islandica* K833**	7,433,200	0.8	86.44	45.09	3,510,023
***Rhodopirellula maiorica* SM1**	8,874,084	4.8	87.48	1.43	111,681
***Rhodopirellula sallentina* SM41**	8,186,686	2.1	85.70	2.77	215,598
***Rhodopirellula sp.* SWK7**	8,777,069	2.6	85.10	2.67	208,073

[1] 16S rRNA gene dissimilarity to strain SWK21. Numbers indicate differences in percent (from Winkelmann, 2009).

Table S4. 16S rRNA gene similarities and average nucleotide identities (ANIs) of published complete and draft *Rhodopirellula* genomes and final draft genome of strain SWK21 to the final draft genome of '*Rhodopilula apulia'* SM50.

Genome	Genome size (bp)	16S rRNA dissimilarity[1]	ANIm (%)	Aligned bases (%)	Aligned bases (bp)
***'Rhodopilula apulia'* SM50**	9,795,212	*	*	*	*
***'Rhodopirellula bahusiensis'* SWK21**	7,784,569	7.1	83.93	0.73	71,707
***Rhodopirellula baltica* SH1**	7,145,576	6.4	83.20	0.53	51,891
***Rhodopirellula baltica* SWK14**	7,488,930	7.1	83.81	0.69	67,376
***Rhodopirellula baltica* SH28**	7,149,689	7.6	84.61	0.78	76,108
***Rhodopirellula baltica* WH47**	7,033,319	7.1	83.88	0.68	66,581
***Rhodopirellula europaea* 6C**	7,191,307	7.7	82.87	0.75	73,819
***Rhodopirellula europaea* SH398**	7,446,194	7.7	82.81	0.59	57,619
***Rhodopirellula islandica* K833**	7,433,200	8.0	83.62	0.72	70,278
***Rhodopirellula maiorica* SM1**	8,874,084	7.9	84.56	0.88	86,559
***Rhodopirellula sallentina* SM41**	8,186,686	6.4	82.92	0.52	50,614
***Rhodopirellula sp.* SWK7**	8,777,069	8.2	83.13	0.59	58,088

[1] 16S rRNA gene dissimilarity to SM50. Numbers indicate differences in percent (from Winkelmann, 2009).

Chapter 4

Optimization of CARD-FISH for *Methanosaeta* cells

Jana Kizina, Jens Harder

Optimization of CARD-FISH for *Methanosaeta* cells

Jana Kizina, Jens Harder*

Department of Microbiology, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany

* Corresponding author: Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany. Tel.: +49 421 2028 750; Fax: +49 421 2028 580.
E-mail address: jharder@mpi-bremen.de (J. Harder).

Abstract

Methanosaeta is a methanogenic filamentous archaeon specialized on acetate degradation and usually present in methanogenic syntrophic communities. Catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH) of *Methanosaeta* filaments was reported to be heterogeneous for field samples and enrichment cultures. This study explored strategies to improve CARD-FISH signals of the filamentous cells using the horseradish peroxidase (HRP)-labeled ARCH-915 probe. We suggest two strategies of CARD-FISH experiments with one or two probes. In a limonene-degrading methanogenic enrichment culture, we still observed for the optimized procedures a heterogeneous staining. This may be a biological phenomenon.

4.1 Introduction

Methane producing microorganisms belong to the domain *Archaea*. They are strictly anaerobic organisms (Welte & Deppenmeier, 2014) and are found in various anoxic environments, such as fresh water sediments, tundra areas, swamps and the intestinal tract of ruminants and termites as well as in man-made environments, such as rice fields, anaerobic digesters of sewage plants and biogas plants (Kubota *et al.*, 2008; Welte & Deppenmeier, 2014). Methanogens are important for global carbon fluxes (Welte & Deppenmeier, 2014). They occupy the terminal position in the anaerobic food chain for the recycling of carbon components from organic matter (Jupraputtasri *et al.*, 2005; Welte & Deppenmeier, 2014). In this process biopolymers (polysaccharides, proteins, lipids, and nucleic acids) are hydrolyzed to mainly sugars, amino acids, purines, pyrimidines, fatty acids, and glycerol by primary fermenters. This metabolic group of bacteria convert these organic compounds to simple carbonic acids (e.g., propionate, butyrate and acetate), alcohols (e.g., ethanol, propanol and butanol), and some other compounds (H_2, CO_2 and ketones). Syntrophic bacteria use these primary fermentation products as substrates to form acetate, CO_2, H_2, and formate, which are then converted to methane by methanogenic archaea (Schink & Stams, 2006; Welte & Deppenmeier, 2014). It has been estimated that approximately 74% of the global methane discharge into the atmosphere is from biological processes, mainly the activity of methanogens (Kubota *et al.*, 2008). A major intermediate is acetate. Only two methanogenic genera, *Methanosarcina* and *Methanosaeta*, have been described to use this substrate (Barber *et al.*, 2011; Welte & Deppenmeier, 2014). While members of the genus *Methanosarcina* are

versatile in their substrate range (Welte & Deppenmeier, 2014), *Methanosaeta* species are specialized on acetate degradation. They have a high affinity for this substrate. Only a steady state concentration of 7 µM to 70 µM is needed for growth, contrary to *Methanosarcina* species, which minimum threshold for acetate utilization is 0.2 mM to 1.2 mM (Jetten *et al.*, 1992). Species of the genus *Methanosaeta* form rod-shaped cells and are combined end to end in long filaments, surrounding by a sheath-like structure (Welte & Deppenmeier, 2014) (Fig. 1).

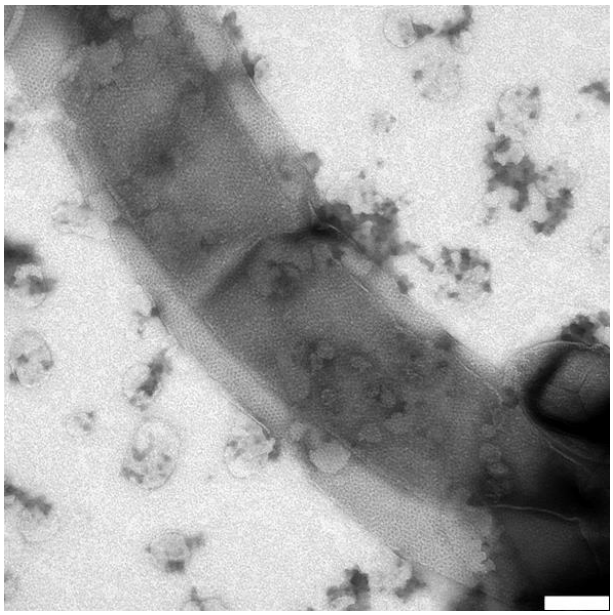


Fig. 1. Transmission electron microscopy (TEM) image of a *Methanosaeta* filament from the limonene-degrading methanogenic enrichment cultures. The individual cells inside of a proteinaceous sheath are shown. Scale bar 100 nm (Image was kindly provided by T. Fischer, 2017).

Large filaments were observed in a methanogenic enrichment culture on limonene. The phylogeny of the filaments was established in catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH) using the *Methanosaeta* specific probe MX-825 and the general domain probe for *Archaea* ARCH-915 (Rotaru, 2009). The *Methanosaeta* filaments in the enrichment culture revealed a heterogeneous staining with 4',6-diamidin-2-phenylindol (DAPI) and with CARD-FISH using the general archaeal probe. This study attempts to obtain a homogeneous staining of all cells of *Methanosaeta* filaments.

Partial staining in CARD-FISH experiments is most often caused by a lack of cell permeabilization (Amann & Fuchs, 2008). The penetration of horseradish peroxidase (HRP)-labeled probes into fixed cells is a critical step of CARD-FISH due to the size of 5–6 nm and

a molecular weight of 40 kDa of the enzyme (Amann & Fuchs, 2008; Kubota, 2013). Appropriate permeabilization protocols are essential for diffusion of HRP-labeled probes into the intact fixed cells (Kubota *et al.*, 2008; Molari & Manini, 2012). Due to differences in cell wall composition of *Bacteria* and *Archaea*, no standard protocol exists for all microbial cells. The permeabilization step has to balance the accessibility to the rRNA molecules with the possible loss of rRNA molecules or the complete lysis of the cell (Molari & Manini, 2012). Therefore, empirical optimizations often consider the specific composition of the cell wall (Amann & Fuchs, 2008).

The surface structures of methanogens include four types: pseudomurein, surface layer proteins (S-layer), methanochondroitin, and sheath (Kubota *et al.*, 2008). The genus *Methanosaeta* has a proteinaceous sheath forms the filament with the individual cells inside (Kubota *et al.*, 2008) (Fig. 1). Therefore, lysozyme which is the most commonly used enzyme for permeabilization treatment was considered to be inadequate for the detection of *Methanosaeta* filaments (Kubota, 2013).

In this study, we evaluated several different enzymatic and chemical treatments to improve the detectability of the ARCH-915 stained cells in *Methanosaeta*-like filaments. The results suggested separate protocols for the application of CARD-FISH to detect the archaeal cells of the filaments in the limonene-degrading methanogenic enrichment cultures and for the simultaneous detection of other members by double CARD-FISH. The general probe mix for *Bacteria*, EUB338 I–III, does not match the 16S rRNA of the phylotype OP3 LiM, which is an abundant member of the bacterial community in the enrichment culture (Rotaru, 2009; Rotaru *et al.*, 2012). Hence, double CARD-FISH experiments used besides Arch-915 either EUB338 I–III or the OP3 LiM specific probe OP3-565.

4.2 Materials and methods

The CARD-FISH protocol (Pernthaler *et al.*, 2004) for samples on membrane filters was performed with permeabilization, endogenous peroxidases inactivation, *in situ* hybridization, washing, and catalyzed reporter deposition (tyramide signal amplification). The permeabilization step for prokaryotic cells uses lysozyme (Fig. 2). Because this enzyme is expected to be rendered insensitive to the archaeal filaments, the permeabilization was modified to improve the penetration of HRP-labeled probes into the fixed filamentous cells by several enzymatic and chemical treatments.

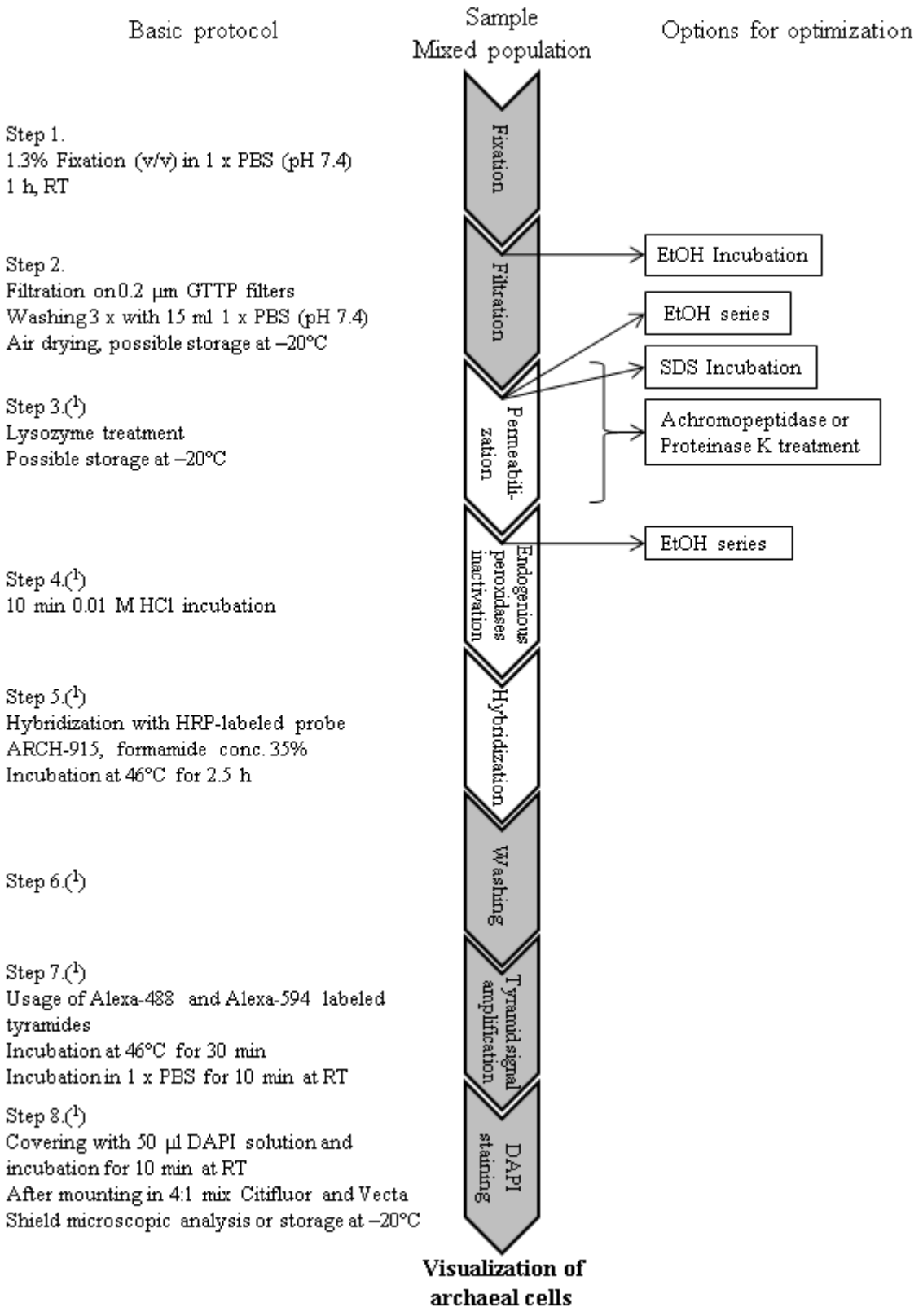


Fig. 2. Workflow of the CARD-FISH application. The sequential steps of the general CARD-FISH approach are displayed with boxes. Gray boxes represent constant steps. White boxes represent variations that are shown to the right. Arrows indicate the position of insertion. ([1]) Steps as described in the protocol (Pernthaler *et al.*, 2004).

For single CARD-FISH with ARCH-915, the protocol for formaldehyde-fixed cells (1.3% for 1 h at room temperature (RT)) was modified by (i) addition of an ethanol fixation (50% v/v for 10 min at RT) before filtration or a treatment of the cells on the filter with (ii) 0.5% sodium dodecyl sulfate (SDS) solution for 10 min at RT (Holler *et al.*, 2011) or (iii) an ethanol series (50%, 70%, and 96% v/v for 5 min each step at RT). Lysozyme treatment as described in Pernthaler *et al.* (2004) was changed to (i) 2 h incubation instead of 1 h, (ii) achromopeptidase treatment after lysozyme treatment or (iii) proteinase K treatment. Achromopeptidase (60 or 120 U/µl in 0.01 M NaCl, 0.01 M Tris/HCl, pH 8.0) was applied for 30 min at 37°C. Endogenous peroxidases were inactivated with 0.01 M HCl for 10 min at RT. Proteinase K was tested over a range of concentrations (15, 20, 25, 30, 100, or 200 µg/ml in 0.1 M Tris-HCl, 0.05 M EDTA, pH 8.0, 0.5 M NaCl). Filters were incubated for 3, 10, or 15 min at 37°C in low concentrated proteinase K solutions (15–30 µg/ml) or for 2 min at RT in proteinase K solution of concentrations of 15 µg/ml (Holler *et al.*, 2011), 100 µg/ml, and 200 µg/ml. Proteinase K activity was stopped by 0.1 M HCl for 10 min at RT. As further option, an ethanol series in 50, 70, and 96% ethanol for 5 min each step at RT was performed after the permeabilization. The modifications described by Holler *et al.* (2011) were also tested: the samples were incubated in 0.5% SDS solution for 10 min at RT, in proteinase K solution (15 µg/ml in 0.1 M Tris-HCl, 0.05 M EDTA, pH 8.0, 0.5 M NaCl) for 2 min at RT, and in 0.01 M HCl for 5 min at RT. A formamide (FA) concentration of 35% (v/v) was used for *in situ* hybridization with the HRP-labeled ARCH-915 probe (Table 1). The hybridization was performed for 2.5 h as described by Pernthaler *et al.* (2004) for coastal water samples or overnight. Washing and amplification was performed as described by Pernthaler *et al.* (2004) (Fig. 2).

For double staining of *Archaea* with probe ARCH-915 and *Bacteria* with probe mix EUB338 I–III or with the OP3 LiM specific probe OP3-565 and helper mix (Table 1), two sequential hybridizations were performed. The afore-mentioned permabilization procedures were tested. For staining of bacterial cells, steps of hybridization, washing and amplification were performed as described by Pernthaler *et al.* (2004) (Fig. 2). HRP present at the first probe used was inactivated after the signal amplification by incubation of the filter with 3% H_2O_2 for 10 min at RT. Filter sections were then washed thoroughly in MQ water and after air-drying they were stored overnight at –20°C before the second round of CARD-FISH was performed. Alexa-488 and Alexa-594 labeled tyramides were used for first and second amplification, respectively.

Table 1. HRP-labeled oligonucleotide probes and helpers (non-labeled) used in this study.

Probe name	Probe sequence (5´–3´)	Position	Target group	FA (%)[a]	Reference
EUB338 I[b]	GCTGCCTCCCGTAGGAGT	338–355	Most *Bacteria*	35	Amann *et al.*, 1990
EUB338 II[b]	GCAGCCACCCGTAGGTGT	338–355	*Planctomycetales*	35	Daims *et al.*, 1990
EUB338 III[b]	GCTGCCACCCGTAGGTGT	338–355	*Verrucomicrobiales*	35	Daims *et al.*, 1990
ARCH-915	GTGCTCCCCCGCCAATTCCT	915–934	Domain *Archaea*	35	Stahl and Amann, 1991
OP3-565	TACCTGCCCTTTACACCC	608–626[d]	Candidate OP3 LiM	30	Rotaru, 2009
H548-A[c]	AATAAATCCGAGTAACGC	590–608[d]	Candidate OP3 LiM		Kizina *et al.*, unpublished
H548-C[c]	AATCAATCCGAGTAACGC	590–608[d]	Candidate OP3 LiM		Kizina *et al.*, unpublished
H583-TC[c]	CTCCCCACTTGTCAGGCCGCC	626–647[d]	Candidate OP3 LiM		Kizina *et al.*, unpublished
H583-CT[c]	CCTCCCACTTGTCAGGCCGCC	626–647[d]	Candidate OP3 LiM		Kizina *et al.*, unpublished

[a] FA, fomamide concentration of the hybridization buffer.

[b/c] Used in a mix.

[d] OP3 LiM 16S rRNA location.

4.3 Results

To optimize the permeabilization for the cell wall of *Methanosaeta* filaments, different chemicals and enzymes (SDS, ethanol, lysozyme, achromopeptidase and proteinase K) were examined.

Briefly, an increased incubation time from 1 h to 2 h in lysozyme solution was not more effective. We confirmed the inadequate and heterogeneous detection of rRNA signals of *Methanosaeta* cells, when the lysozyme treatment was performed as described by the original protocol (Pernthaler *et al.*, 2004). Weak signals were likely due to a partial permeabilization. Both other enzymatic treatments were more successful. After the lysozyme and achromopeptidase treatment stronger hybridization signals were detected with HRP-labeled ARCH-915 probe. But non-detectable or weakly detectable archaeal filament cells were still observed. Proteinase K at high concentration (200 µg/ml) produced better CARD-FISH signals of filamentous cells than the other performed permeabilization treatments. Few signals were observed for samples prepared by the described modification by Holler *et al.* (2011) with SDS and a short proteinase K incubation. Overnight incubation increased the signal strength slightly. Chemical treatments strongly influenced the labeling of filament cells. SDS incubation led to better CARD-FISH signals independent on the used enzyme. In contrast, the effect of dehydration through an ascending series of ethanol depended on the position integrated into the workflow as well as the enzymatic treatment. Ethanol dehydration yielded disrupted cells and weak or no hybridization signals. It was partly effective, if performed or before filtration.

Overall, proteinase K was the most effective treatment for *Methanosaeta* cells. The incubation in proteinase K solution at low concentration was even adequate for the hybridization of nearly all cells of filaments (Fig. 3). The reproduction of these experiments yielded reproducibly heterogeneous hybridization signal intensities for cells of the filaments. Filaments that were completely stained were next to cells in filaments without CARD-FISH signal. It may be concluded that these cells have no rRNA (Fig. 3C).

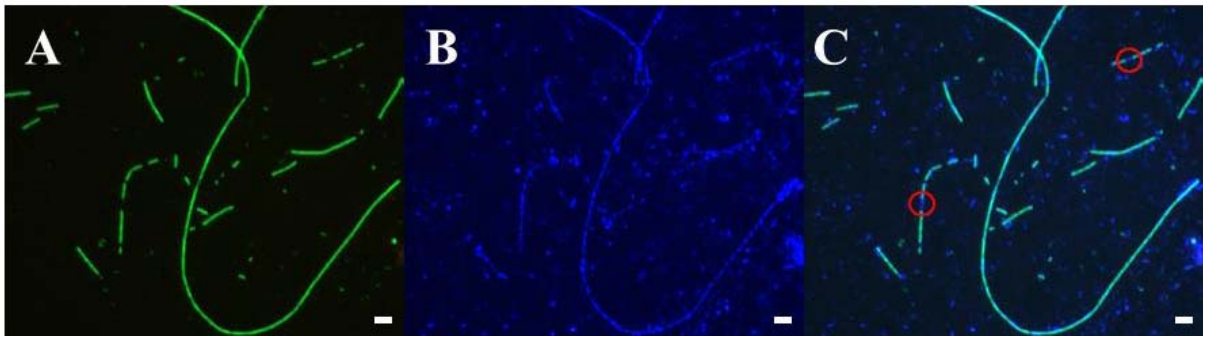
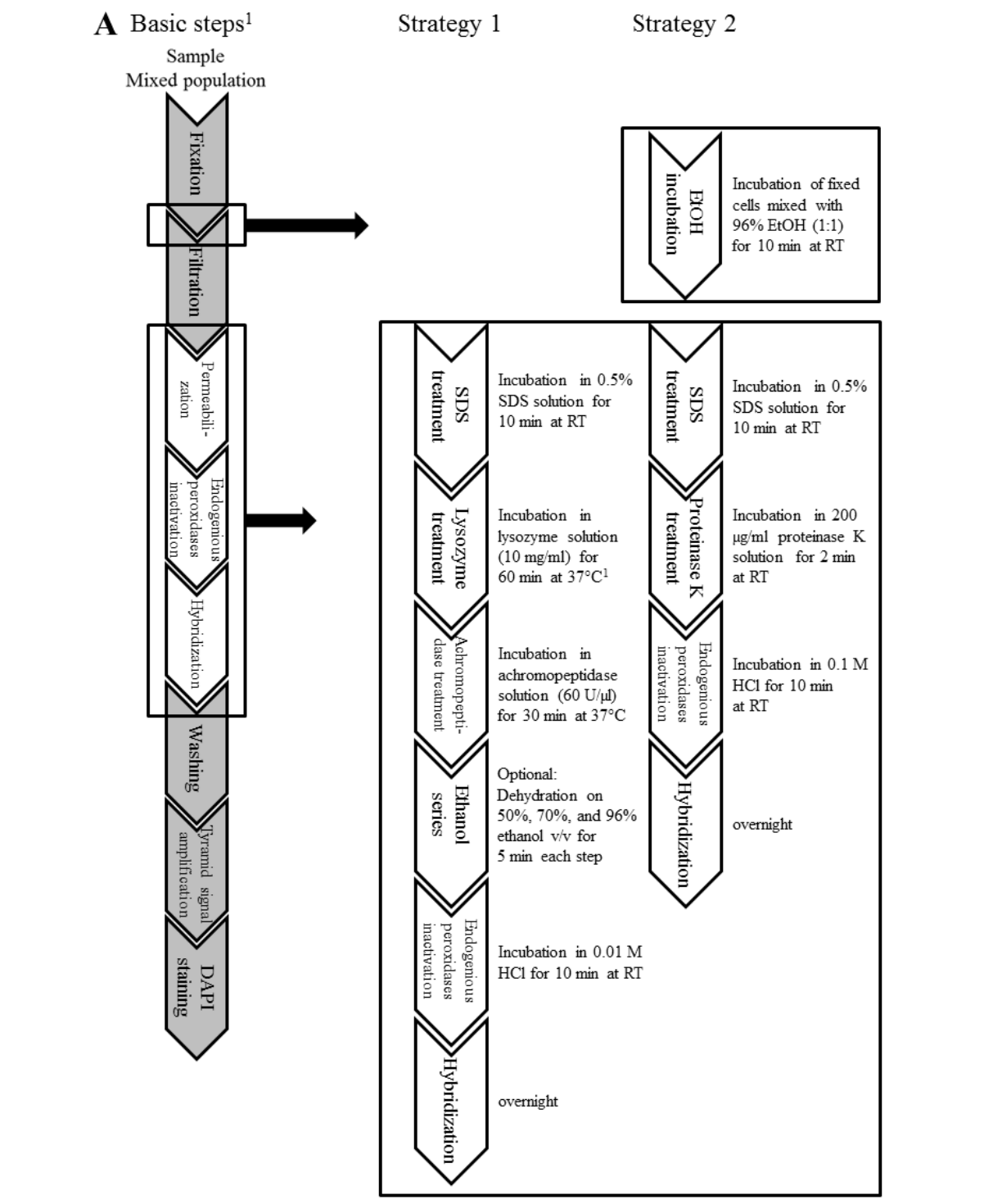


Fig. 3. Epifluorescence micrograph of CARD-FISH hybridized sample from limonene degrading methanogenic enrichment culture. The pre-treatment protocol for the filter was an ethanol dehydration (50%, 70%, and 96% v/v for 5 min each step), an incubation for 5 min at 37°C in proteinase K solution (20 µg/ml in 0.1 M Tris-HCl, 0.05 M EDTA, pH 8.0, 0.5 M NaCl), an incubation of 0.1 M HCl for 10 min at RT, and a hybridization time of 2.5 h. (A) Cells hybridized with the HRP-labeled ARCH-915 probe (green). (B) DAPI stained cells (blue). (C) Overlay of DAPI staining and ARCH-915 signals. Red circles indicate cells containing DNA (blue), but no rRNA (green). Scale bars, 5 µm.

Higher concentration of proteinase K solution (200 µg/ml) resulted in even clearer pictures, likely due to the destruction of bacterial cells in the samples. This enzymatic treatment in combination with ethanol incubation, SDS incubation, incubation in 0.1 M HCl to inactivate endogenous peroxidases, and overnight hybridization is recommended as approach for *Methanosaeta* cell detection by CARD-FISH (Fig. 4A, strategy 2; Fig. 4D).

The strong proteinase K treatment was ineffective to detect archaeal cells of smaller morphology as well as bacteria in double CARD-FISH experiments. The more gentle enzymatic treatment of lysozyme and achromopeptidase for permeabilization in combination with SDS incubation and overnight hybridization was resulting in reproducible results (Fig. 4A, strategy 1). CARD-FISH signals of small archaeal cells were observed and the permeabilization strategy was also appropriate for the detection of bacterial cells using the EUB338 I–III probe mix or the OP3 LiM specific probe OP3-565 by double CARD-FISH (Fig. 4B). The DAPI stained filaments showed archaeal CARD-FISH signals for the majority of cells in the filaments. Therefore, this modified CARD-FISH protocol is suggested for quantification of archaeal cells as well as the analysis of the microbial diversity in the methanogenic enrichment cultures on limonene. An ethanol series after permeabilization treatment showed better signals for archaeal cells (Fig. 4C). Because of observed hydrolysis of bacterial cells this step has to be omitted in double CARD-FISH application.


A Basic steps[1]
Strategy 1
Strategy 2
Sample
Mixed population
Fixation
Filtration
Permeabili-zation
Endogenious peroxidases inactivation
Hybridization
Washing
Tyramid signal amplification
DAPI staining
EtOH incubation
Incubation of fixed cells mixed with 96% EtOH (1:1) for 10 min at RT
SDS treatment
Incubation in 0.5% SDS solution for 10 min at RT
Lysozyme treatment
Incubation in lysozyme solution (10 mg/ml) for 60 min at 37°C[1]
Achromopepti-dase treatment
Incubation in achromopeptidase solution (60 U/µl) for 30 min at 37°C
Ethanol series
Optional: Dehydration on 50%, 70%, and 96% ethanol v/v for 5 min each step
Endogenious peroxidases inactivation
Incubation in 0.01 M HCl for 10 min at RT
Hybridization
overnight
SDS treatment
Incubation in 0.5% SDS solution for 10 min at RT
Proteinase K treatment
Incubation in 200 µg/ml proteinase K solution for 2 min at RT
Endogenious peroxidases inactivation
Incubation in 0.1 M HCl for 10 min at RT
Hybridization
overnight

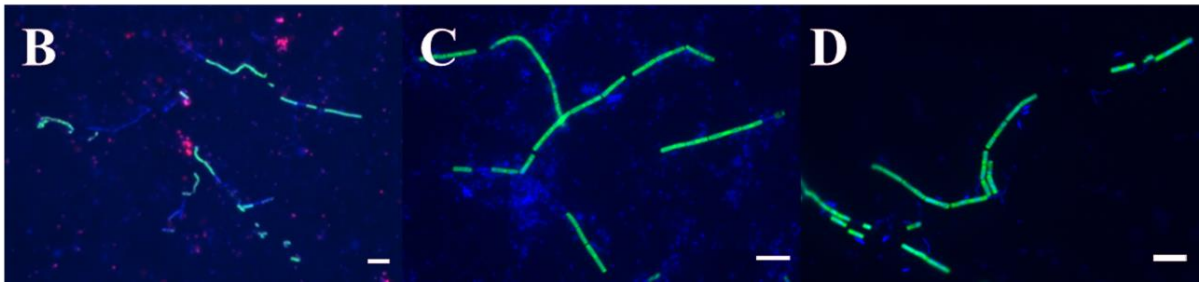

B
C
D

Fig. 4. (A) Overview on protocols for CARD-FISH of (strategy 1) simultaneous detection of archaeal and bacterial cells by double CARD-FISH and of (strategy 2) single detection of *Methanosaeta* filament cells. [1] Basic steps of the original protocol described by Pernthaler *et al.* (2004). Gray boxes show steps used in the original protocol. White boxes show steps which were modified. Modifications are displayed in black frames. (B–D) Epifluorescence micrographs of CARD-FISH hybridized sample from limonene degrading, methanogenic enrichment cultures. (B) Merged images show cells hybridized with OP3-565/helper mix in red and with archaeal probe ARCH-915 in green after performed strategy 1 for double CARD-FISH; (C) cells hybridized with archaeal probe ARCH-915 after performed strategy 1 with optional ethanol dehydration; (D) hybridized with archaeal probe ARCH-915 after performed strategy 2. (B–C) DAPI stained cells are shown in blue. Scale bars, 5 μm.

4.4 Discussion

Fluorescence *in situ* hybridization (FISH) has become a standard technique in environmental microbiology for studying the abundance of species or taxonomic groups in microbial communities. The sensitivity of FISH is improved by CARD-FISH (Ishii *et al.*, 2004; Kubota, 2013). This approach was used to characterize the microbial community of limonene-degrading methanogenic enrichment cultures. Based on the heterogeneous labeling of archaeal filament cells using the HRP-labeled ARCH-915 probe, an improved CARD-FISH method for *Methanosaeta* cells was the aim of this study.

Previous studies analyzed the composition and structure of this methanogenic enrichment culture community using the full cycle rRNA approach (Rotaru *et al.*, 2012) and CARD-FISH experiments (Rotaru *et al.*, 2012; Kizina *et al.*, unpublished). M*ethanosaeta* filaments were stained very heterogeneously. Some cells showed a weak or no signal using the general domain probe ARCH-915. This has also been noticed in other studies (Jupraputtasri *et al.*, 2005; Kubota, 2013). Lysozyme is usually used for the permeabilization of cells, but may not be effective for microorganisms with unusual cell wall structures (Ishii *et al.*, 2004), such as *Methanosaeta* which have a proteinaceous sheath outside to their individual cell envelopes (Kubota *et al.*, 2008). An incomplete permeabilization of the sheath of the filaments was considered for an insufficient probe penetration of ARCH-915-HRP into the fixed filamentous cells as cause of heterogeneous CARD-FISH signals. To improve the permeabilization, we investigated combinations of different enzymes (lysozyme, achromopeptidase, and proteinase K) and chemicals (SDS and ethanol) as well as adjusted basic steps of the CARD-FISH protocol (Pernthaler *et al.*, 2004). Although sonication may be effective, it disrupts cell-to-cell contacts and was therefore not considered. SDS incubation as

pretreatment before the enzymatic treatment as well as an overnight hybridization increased the signal intensities. As described by Kubota (2013), achromopeptidase treatment after the lysozyme treatment was effective. Achromopeptidase hydrolyzes lysyl peptide bonds. The lysozyme treatment likely improves the accessibility of the peptide bonds (Ishii *et al.*, 2004; Kubota, 2013).

Proteinase K is a suitable candidate for the permeabilization of *Methanosaeta* filaments due to the degradation of a broad spectrum of proteins (Lloyd *et al.*, 2013). Proteinase K showed the best results for *Methanosaeta* cells. Other studies also described that proteinase K is more effective for *Archaea* than lysozyme (Teira *et al.*, 2004; Molari & Manini, 2012). However, we confirmed the results of previous studies (Schönhuber *et al.*, 1999; Kubota, 2013) that this enzyme is difficult to be used reproducible at low concentrations. Proteinase K at high concentrations (200 µg/ml) was not more suitable for small archaeal cells as well as bacterial cells. Thus, the simultaneous detection of bacterial and archaeal cells by double CARD-FISH experiments needed another method.

The recombinant pseudomurein endopeptidase [PeiW] was proposed as permeabilization method for CARD-FISH application to methanogens by Kubota *et al.* (2008). This enzyme was not very effective for methanogens having methanochondroitin and a sheath. Partial staining of *Methanosaeta* after PeiW treatment was observed using HRP-labeled ARCH-915. Therefore, this enzyme was not included in our study.

CARD-FISH is a standard method for *in situ* detection of cells in environmental microbiology. An alternative amplification method is hybridization chain reaction (HCR). DNA-HCR and RNA-HCR systems are available (Yamaguchi *et al.*, 2015). *In situ* DNA-HCR was carried out using *Methanosaeta concilii* cells and the results were compared with the application of CARD-FISH. In contrast to CARD-FISH, the non-enzymatic technique was able to stain cells without permeabilization indicating a more readily entrance of the probe (Yamaguchi *et al.*, 2015). This is advantageous for *Methanosaeta* cells because of their cell wall structure. However, the images of the detection of *Methanosaeta concilii* cells using *in situ* DNA-HCR with ARCH-915-initiatorH probe and Cye3-labelled amplifier probes showed filamentous cells which are DAPI stained but not stained using the probes (Yamaguchi *et al.*, 2015). Thus, this alternative approach confirmed our best CARD-FISH results. It suggests that the unstained cells have no ribosomal RNA.

No standard protocol exists for all microbial cells (Molari & Manini, 2012). Our experiments demonstrated that due to great variations in cell wall composition among prokaryotes an approach likely cannot have a universal applicability. Therefore, we suggested

two strategies, one (Fig. 4A, strategy 2) for the application of CARD-FISH for *Methanosaeta* cells and the other (Fig. 4A, strategy 1) for the simultaneous detection of *Methanosaeta* cells and all other archaeal and bacterial cells using double CARD-FISH experiments. Nevertheless, the further development of highly reproducible cell permeabilization methods especially for the application of double CARD-FISH experiments to microbes with different cell wall structure remains a necessity.

4.5 Conclusion

In this study we suggested two strategies for the permeabilization of the *Methanosaeta* cell wall to improve successfully their visualization, either for single detection or simultaneous detection with other archaeal or bacterial cells. The best established technique still revealed cells in *Methanosaeta* filaments that contained DNA but no rRNA according to *in situ* hybridizations. We conclude that the partial staining was not caused methodologically, but may have a biological reason, e.g. lysis by phages or predation.

Acknowledgments

We thank Jörg Wulf for technical assistance. This study was funded by the Max Planck Society.

References

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

Amann RI, Krumholz L, Stahl DA. 1990. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. J Bacteriol 172:762–770.

Barber RD, Zhang L, Harnack M, Olson MV, Kaul R, Ingram-Smith C, Smith KS. 2011. Complete genome sequence of *Methanosaeta concilii*, a specialist in aceticlastic methanogenesis. J Bacteriol 193:3668–3669.

Daims H, Brühl A, Amann R, Schleifer KH, Wagner M. 1999. The domain-specific probe EUB338 is insufficient for the detection of all *Bacteria*: development and evaluation of a more comprehensive probe set. Syst Appl Microbiol 22:434–444.

Holler T, Widdel F, Knittel K, Amann R, Kellermann MY, Hinrichs K-U, Teske A, Boetius A, Wegener G. 2011. Thermophilic anaerobic oxidation of methane by marine microbial consortia. ISME J 5:1946–1956.

Ishii K, Mussmann M, MacGregor BJ, Amann R. 2004. An improved fluorescence *in situ* hybridization protocol for the identification of bacteria and archaea in marine sediments. FEMS Microbiol Ecol 50:203–213.

Jetten MSM, Stams AJM, Zehnder AJB. 1992. Methanogenesis from acetate: a comparison of the acetate metabolism in *Methanothrix soehngenii* and *Methanosarcina* spp. FEMS Microbiol Rev 88:181–198.

Jupraputtasri W, Boonapatcharoen N, Cheevadhanarak S, Chaiprasert P, Tanticharoen M, Techkarnjanaruk S. 2005. Use of an alternative *Archaea*-specific probe for methanogen detection. J Microbiol Methods 61:95–104.

Kubota K. 2013. CARD-FISH for environmental microorganisms: technical advancement and future applications. Microbes Environ 28:3–12.

Kubota K, Imachi H, Kawakami S, Nakamura K, Harada H, Ohashi A. 2008. Evaluation of enzymatic cell treatments for application of CARD-FISH to methanogens. J Microbiol Methods 72:54–59.

Lloyd KG, May MK, Kevorkian RT, Steen AD. 2013. Meta-analysis of quantification methods shows that archaea and bacteria have similar abundances in the subseafloor. Appl Environ Microbiol 79:7790–7799.

Molari M, Manini E. 2012. Reliability of CARD-FISH procedure for enumeration of *Archaea* in deep-sea surficial sediments. Curr Microbiol 64:242–250.

Pernthaler A, Pernthaler J, Amann R. 2004. Sensitive multi-color fluorescence *in situ* hybridization for the identification of environmental microorganisms, p 711–726. *In* Kowalchuk G, de Bruijn, FJ, Head IM, Akkermans ADL, van Elsas JD (ed), Molecular microbial ecology Manual, 2nd, vol 1. Kluwer Academic Publishers, Dordrecht, Boston, London.

Rotaru AE. 2009. Anaerobic degradation of limonene and p-xylene in freshwater enrichment cultures. PhD thesis, Universität Bremen, Germany.

Rotaru AE, Schauer R, Probian C, Mussmann M, Harder J. 2012. Visualization of candidate division OP3 cocci in limonene-degrading methanogenic cultures. J Microbiol Biotechnol 22:457–461.

Schink B, Stams AJM. 2006. Syntrophism among prokaryotes, p 309–335. *In* Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (ed), The prokaryotes: prokaryotic communities and ecophysiology, vol 2. Springer, New York.

Schönhuber W, Zarda B, Eix S, Rippka R, Herdman M, Ludwig W, Amann R. 1999. *In situ* identification of cyanobacteria with horseradish peroxidase-labeled, rRNA-targeted oligonucleotide probes. Appl Environ Microbiol 65:1259–1267.

Stahl DA, Amann R. 1991. Development and application of nucleic acid probes, p 205–248. *In* Stackebrandt E, Goodfellow M (ed), Nucleic acid techniques in bacterial systematics. John Wiley & Sons Ltd, Chichester, United Kingdom.

Teira E, Reinthaler T, Pernthaler A, Pernthaler J, Herndl GJ. 2004. Combining catalyzed reporter deposition-fluorescence *in situ* hybridization and microautoradiography to detect substrate utilization by bacteria and *Archaea* in the deep ocean. Appl Environ Microbiol 70:4411–4414.

Welte C, Deppenmeier U. 2014. Bioenergetics and anaerobic respiratory chains of aceticlastic methanogens. Biochim Biophys Acta 1837:1130–1147.

Yamaguchi T, Kawakami S, Hatamoto M, Imachi H, Takahashi M, Araki N, Yamaguchi T, Kubota K. 2015. *In situ* DNA-hybridization chain reaction (HCR): a facilitated *in situ* HCR system for the detection of environmental microorganisms. Environ Microbiol 17:2532–2541.

Chapter 5

”*Candidatus* Vampirococcus archaeovorus“ and a transfer of a bacterial intron ribonucleic acid into archaeal cells

Jana Kizina, Sebastian Jordan, Anastasia Resteu, Christina Probian, Sten Littmann, Stefanie Markert, Erhard Rhiel, Kurt Stüber, Thomas Schweder, Richard Reinhardt, Michael Richter, Jens Harder

In preparation for submission to Appl Environ Microbiol

”*Candidatus* Vampirococcus archaeovorus“ and a transfer of a bacterial intron ribonucleic acid into archaeal cells

Jana Kizina[a], Sebastian Jordan[a], Anastasia Resteu[a], Christina Probian[a], Sten Littmann[b], Stefanie Markert[c], Erhard Rhiel[d], Kurt Stüber[e], Thomas Schweder[c], Richard Reinhardt[e], Michael Richter[f] and Jens Harder[a]#

[a] Max Planck Institute for Marine Microbiology, Department of Microbiology, Bremen, Germany

[b] Max Planck Institute for Marine Microbiology, Department of Biogeochemistry, Bremen, Germany

[c] Department of Pharmaceutical Biotechnology, Institute for Pharmacy, University of Greifswald, Germany

[d] Institute for Chemistry and Biology of the Marine Environment, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

[e] Max Planck Genome Centre Cologne, Cologne, Germany

[f] Max Planck Institute for Marine Microbiology, Microbial Genomics and Bioinformatics Research Group, Bremen, Germany

Address correspondence to Jens Harder: jharder@mpi-bremen.de

Running title: ”*Candidatus* Vampirococcus archaeovorus“

Key words: *Omnitrophica*, predatory bacteria, LPS, methanogenic enrichment, mobile element, intron

Abstract

The *candidate phylum* Omnitrophica (candidate division OP3) is currently characterized by draft genomes from metagenomes and single cell genomes. Recently, we visualized the phylotype OP3 LiM cells in methanogenic cultures on limonene. In this study, the closed genome of OP3 LiM was obtained from a nearly clonal enrichment. Visualization experiments applying CARD-FISH showed two physiological stages of OP3 LiM cells: free-living cells had a weak fluorescence signal, whereas the strong signal of attached-living OP3 LiM cells indicated an active metabolism. OP3 LiM cells presented a characteristic surface extension in electron micrographs and attached to *Bacteria* and to *Archaea*. A group I intron in the 23S rRNA of OP3 LiM is excised from the RNA transcript by a LAGLIDADG/HNH homing endonuclease, as demonstrated by the presence of the protein and the nascent and mature RNA transcript. *Methanosaeta* filaments of the methanogenic cultures thriving on limonene had cells without rRNA or both rRNA and DNA. Some of these *Methanosaeta* cells contained the group I intron RNA of OP3 LiM, as detected in *in situ* hybridization experiments. OP3 LiM cells are very small (200–300 nm in diameter) and have highly expressed secreted proteins involved in depolymerization and uptake of macromolecules. Expressed proteins revealed that OP3 LiM uses glycolysis and conserves energy by the utilization of pyruvate via a pyruvate:ferredoxin oxidoreductase and an RNF complex (Ferredoxin:NAD oxidoreductase). Transferases allow energy conservation from the depolymerization of nucleic acids. Our study, especially the interdomain transfer of intron RNA, demonstrated a predatory activity for OP3 LiM cells and we propose to name the anaerobic predatory OP3 LiM coccus ”*Candidatus* Vampirococcus archaeovorus“.

Importance

Predatory bacteria may represent an intermediate evolutionary stage from bacteria to phages by genome reduction. So far, only a few genomes of aerobic predatory bacteria are known. In this study on a nearly clonal limonene-degrading methanogenic enrichment culture, we present the first closed genome of an anaerobic predatory bacterium. It is also the first closed genome of a member of the *candidate phylum* Omnitrophica. The predatory capacity of this epibiontic bacterium is concluded from the content of its closed genome and visualizations including CARD-FISH. The bacterium has a group I intron in its 23S rRNA gene and we demonstrated the presence of the intron RNA in *Methanosaeta* cells. This is the first visualization of an interdomain intron RNA transfer from a bacterium into an archaeon. Because OP3 LiM is the first bacterium that preys on archaea, we propose to name it “*Candidatus* Vampirococcus archaeovorus”.

5.1 Introduction

A primordial soup (1) initiated the evolution of life starting with the RNA world (2). Once the available dissolved organic carbon – the primordial soup – was depleted, energy was only available in the form of reduced inorganic compounds or as particulate organic matter in the form of cells. Geochemically produced hydrogen has been identified as potential energy source for ancient acetogenic and methanogenic organisms (3). The origin and evolution of predation has less been considered as important process in early times of the earth (4), but is now established as a widespread mode of interaction among living organisms in many ecosystems (5–6). Predators were identified in groundwater, rivers, estuaries, the open ocean, sewage, soils, plant roots, and animal feces (7–8). They have been classified as obligate (unable to grow in the absence of prey) or facultative (able to grow as a pure culture without the presence of prey) (8). Bacterial predators attack their prey in groups (*Myxobacteria*) or individually. Epibiotic species attach to the prey, and other species penetrate the periplasma or the cytoplasm (8–9). Members of the genus *Bdellovibrio* and related organisms, summarized as “*Bdellovibrio* and like organisms” (BALOs), are the most-studied group of predatory bacteria (10). BALOs include the deltaproteobacterial order *Bdellovibrionales* and *Bacteriovoracaceae* as well as the alphaproteobacterial *Micavibrio* spp. (7). They exclusively prey on gram-negative cells and have a dimorphic life cycle. Motile cells with a single polar flagellum find prey cells and attach to the outer membrane (7). After an irreversible attachment (11) invading BALOs such as *Bdellovibrio bacteriovorus* enter the prey’s periplasmic space and proliferate at the expense of the prey’s cytoplasmic content. Motile progeny cells release themselves from the remnants of the prey cell to start a new cycle (11).

Epibiontic predators remain attached to the outer membrane while nourishing on the prey (11). This lifestyle have *Bdellovibrio exovorus* sp. nov. (formerly *Bdellovibrio* sp. strain JSS), a novel predator of *Caulobacter crescentus*, and *Micavibrio aeruginosavorus* (12–13). The eukaryotic microalgae *Chlorella* has as epibiontic predator *Vampirovibrio chlorellavorus* (14). In the epibiotic strategy, the cell-to-cell contact via pili has been considered as essential element of the predation (6, 12, 15). The genome of *Micavibrio aeruginosavorus* contains a complete type I secretion system and a functional type II secretion system for protein secretion (12). Complete twin arginine translocation (TAT) and Sec transport systems indicated a functional type II secretion system that is known to assemble type IV pili (16). Several *pil* genes encoding type IV pili are dispersed in the genome of *Micavibrio aeruginosavorus*. Also *Bdellovibrio exovorus* JSS^T uses type IV pili. The genome contains the

Sec pathway and an almost full general secretory pathway (GSP) type II export system (15). The genomic potential to produce a functional flagellum and type IV pili are also present in the obligate predator *Vampirovibrio chlorellavorus* (8).

Besides the aforementioned species with available genome sequences, many reports on predatory bacteria have been published in the last century that provided an at that time state-of-the-art experimental evidence. Notably, the anaerobe *Vampirococcus* (17) has stimulated discussion on the early evolution of predators. *Vampirococcus* has been defined by microscopic cell counts and electron micrographs and has no validated standing in nomenclature (10).

In this contribution, we describe a novel anaerobic predatory bacterium with a coccal morphology. We had observed cells of candidate division OP3, now also named *Candidate phylum* Omnitrophica (18), in high abundances in a methanogenic enrichment culture thriving on limonene. The phylotype OP3 LiM originated from a 16S rRNA gene clone library. Catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH) with a specific OP3 LiM-probe [OP3-565] revealed that 18% of all cells in the enrichment culture were OP3 LiM cells (19). The images showed signals of small and round-shaped cells, located either free-living or attached to larger cells. We applied in this study physical cell separations and a range of visualization techniques as well as metagenomes, metatranscriptomes and metaproteomes to provide a first insight into the biology of OP3 LiM cells. Based on our observations, we propose a taxonomic affiliation with *Vampirococcus*.

5.2 Results

Visualizations of cells and of ribonucleic acids.

The small size of OP3 LiM cells as observed in previous CARD-FISH studies (19) suggested a separation of cell populations by density gradients. The density centrifugation of concentrated cells from the limonene-degrading methanogenic enrichment culture yielded two visible bands close to the top and to the bottom in a Percoll gradient (Fig. S1). PCR and CARD-FISH analyses specific for OP3 LiM detected in the gradient the highest abundance of OP3 cells in a macroscopically not turbid layer above the visible bottom layer. This fraction was further concentrated in a second Percoll gradient and yielded a fraction with over 80% OP3 LiM cells according to CARD-FISH. The twice-enriched cells were used for an OP3 LiM-enriched metagenome.

Transmission electron micrographs of the OP3 LiM-enriched Percoll gradient fraction of the second gradient showed a dominant morphotype of small cells surrounded by a surface structure weakly stained with uranyl acetate (Fig. 1A). The cell size was about 200 nm to 300 nm in diameter. The abundance of the morphotype in transmission electron microscopy (TEM) pictures coincided with the signal abundance of OP3 LiM cells in CARD-FISH experiments and related the small cells with a cape to the phylotype OP3 LiM. The characteristic morphotype was also detected in samples of enrichment cultures, mainly attached to larger cells (Fig. 1B–D). The latter were short rods and filaments, and they seemed to have an intact cell morphology. The attached small cocci were present in different sizes, likely presenting different growth stages.

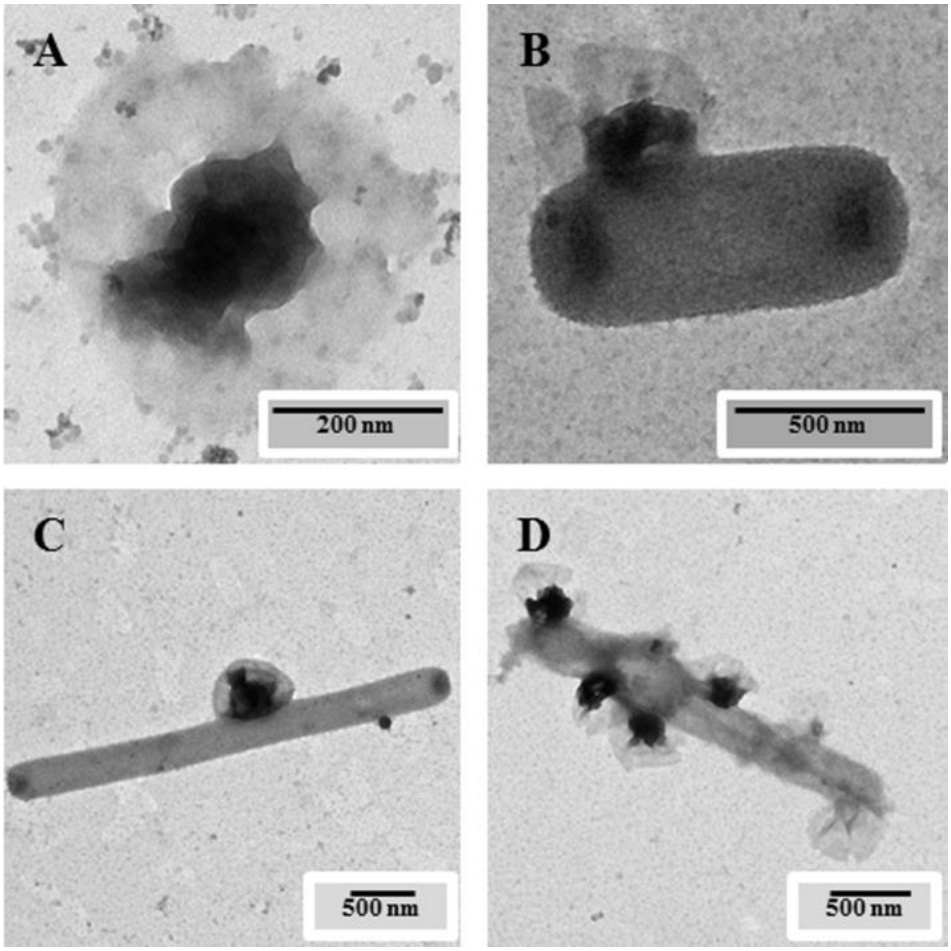


Fig. 1. Negatively stained OP3 LiM cells from methanogenic enrichment cultures: (A) single cell morphotype that was abundant in the Percoll gradient fraction containing over 80% OP3 LiM cells according to CARD-FISH and (B–D) cells attached to larger cells. The small particles in Fig. 1A are Percoll that are colloidal silica particles of 15 nm to 30 nm diameter coated with polyvinylpyrrolidone. The small single cells had a size of 200 nm to 300 nm and a cape of biological material that was less stained than the cell. In the enrichment culture, such OP3 LiM cells were attached to larger cells of different morphologies.

A second method to enrich OP3 LiM cells was differential centrifugation. Twelve lineages of limonene-degrading methanogenic enrichments originated from one microliter of inoculum transferred in 1999 and were maintained by one annual transfer of 10% v/v. Template-dilution OP3 LiM-specific PCR and CARD-FISH experiments with probe OP3-565 guided the selection of a lineage containing a high cell number of OP3 LiM. The majority of cells in the enrichment cultures pelleted with 10,000 S (Pe1). Small cells in the supernatant were collected in a 100 S-pellet that after resuspension was separated into aggregated cells (second 10,000 S-pellet (Pe2)) and free cells (supernatant of second 10,000 S centrifugation (Pe3)). Scanning electron micrographs of the different fractions confirmed the size

fractionation. The fraction of large cells was dominated by filaments and aggregated cells (Fig. S2A–D). Small cells were attached on the filaments (Fig. 2; Fig. S2–S5). The fractions Pe2 and Pe3 contained to over 99% single cells of small size (Fig. S2E–F; Fig. S3). Pe1, Pe2 and Pe3 were the source for additional metagenomes and for metatranscriptomes and -proteomes.

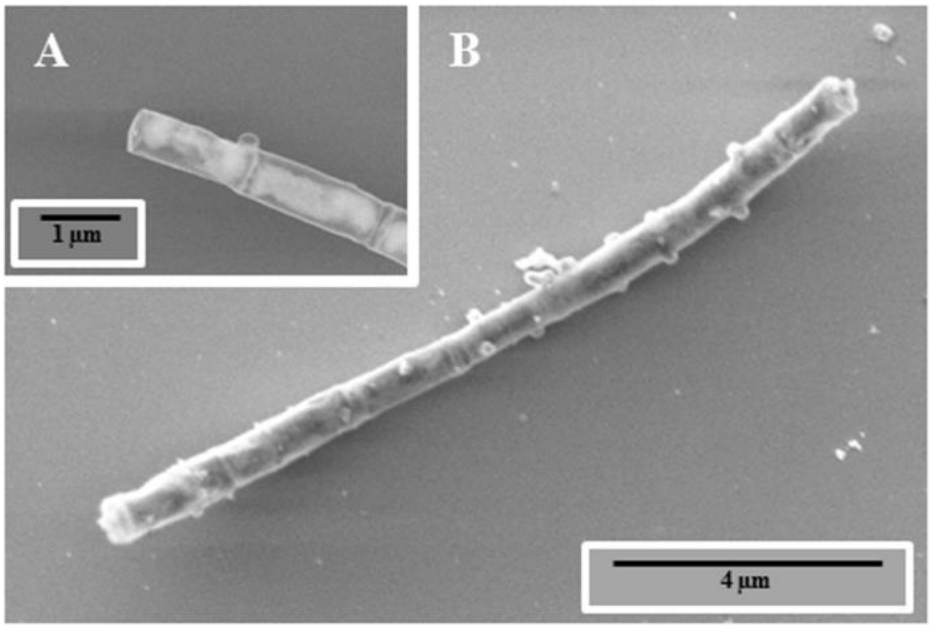


Fig. 2. Scanning electron micrograghs of a filamentous microorganism and attached small cells, assigned to *Methanosaeta* and OP3 LiM cells, respectively. Scale bar, 1 µm (A) and 4 µm (B).

CARD-FISH detection of OP3 LiM cells was improved by introducing four helper oligonucleotides (20). In cultures of the lineages, OP3 LiM cells presented up to 30% of 4’,6-diamidin-2-phenylindol (DAPI) stained cells (Fig. S6). Strong detections signals were observed for attached OP3 LiM cells, suggesting a larger ribosome content and a higher metabolic activity of the attached-living cells in comparison to free-living cells. The weaker signal suggests a state of low metabolic activity, eventually of starvation. OP3 LiM cells attached to archaeal and bacterial cells as visualized by double CARD-FISH experiments with two probes for OP3 LiM cells and for *Archaea* or for *Bacteria*, respectively.

Among the *Archaea* in the enrichment cultures, filaments of cells were conspicuous in the microscopic view and affiliated to *Methanosaeta* by CARD-FISH experiments applying either a *Methanosaetaceae*-specific probe (MX-825) (21) that targeted *in silico* all *Methanosaeta* phylotypes present in the enrichment culture (22) or a general probe for

Archaea (Fig. 3). These filaments often contained one or some cells that did not stain with DAPI, a universal stain for DNA, suggesting that they did not contain DNA (Fig. 3–5; Fig. S7–S9). In addition, CARD-FISH with the general archaeal probe ARCH-915 did not detect ribosomes in all cells of the filaments.

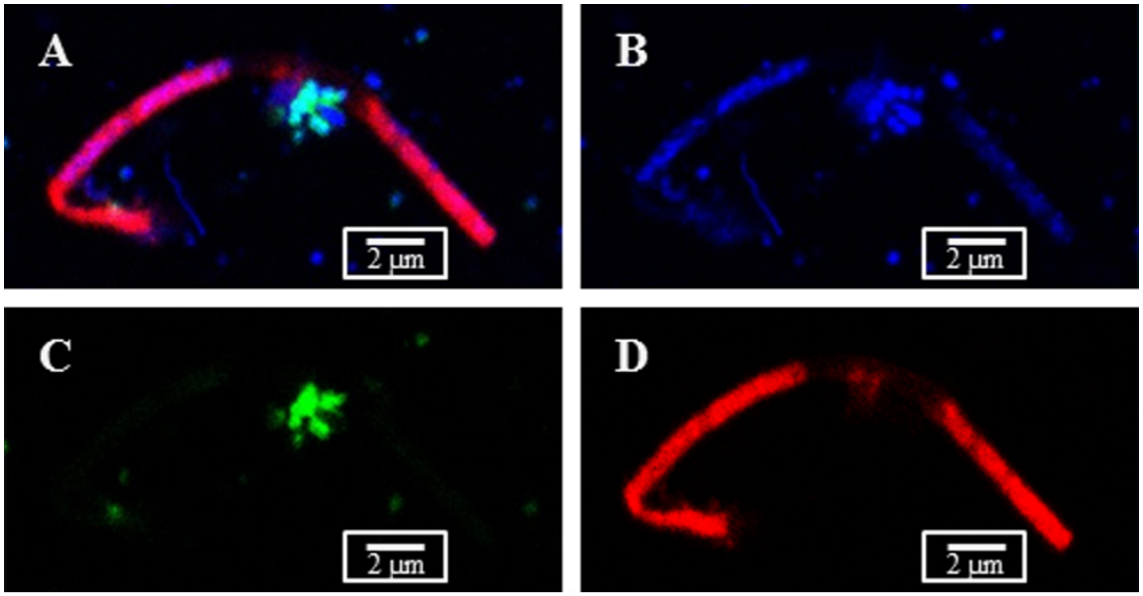


Fig. 3. Detection of *Archaea* and OP3 LiM in a methanogenic enrichment culture in CLSM images. Overlay (A) and individual signals of DNA (B), OP3 LiM (C) and *Archaea* (D) were obtained by DAPI staining and double CARD-FISH with probes OP3-565 together with helper oligonucleotides and ARCH-915, respectively. Scale bar represents 2 µm.

To exclude a methodological problem, we tested a variety of cell lysis treatments to improve the detection of archaeal rRNA in all cells of filaments. Similar to the results of Kubota *et al.* (23), experiments with a harsh treatment with proteinase K did not result in CARD-FISH signals from all cells in filaments (see also Chapter 4). The visibility of filaments in phase-contrast micrographs together with the absence of DNA and of rRNA in cells of the filaments suggested that these cells in the filament had lost their cellular content.

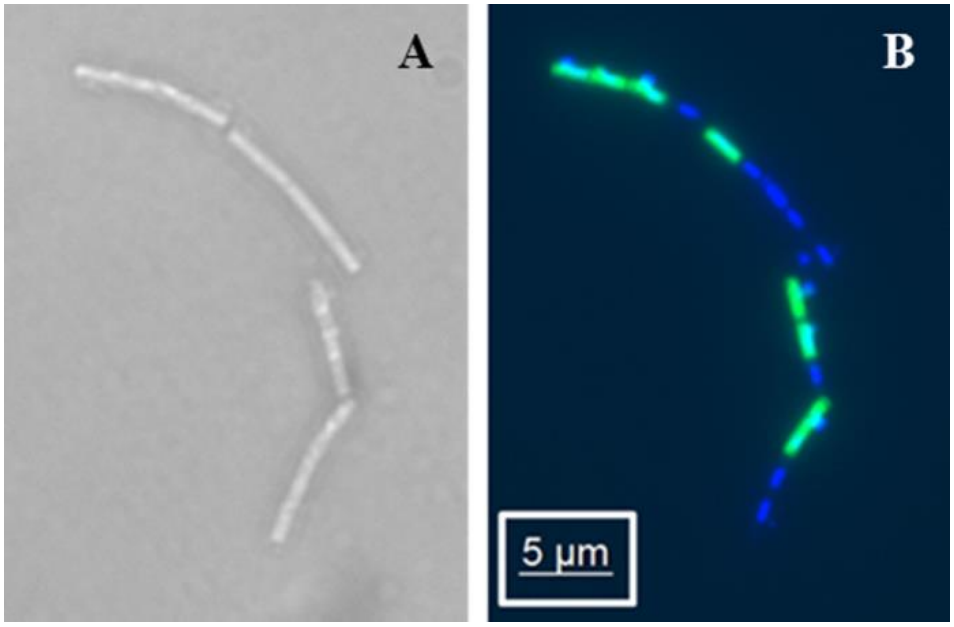


Fig. 4. *Methanosaeta* filaments in phase contrast micrograph (A) and overlay of epifluorescence micrographs (B) revealing the presence of DNA (DAPI staining in blue) and of rRNA (CARD-FISH staining with ARCH-915 in green). Scale bar for all images is 5 µm.

OP3 LiM cells were frequently attached to the filaments, either as single cell (Fig. 5) or as a group of cells (Fig. 3; Fig. S10). The *Methanosaeta* cells lacking DNA and rRNA had OP3 LiM cells on their surface. The CARD-FISH signal of probe OP3-565 showed the hybridization not only in small cells of OP3 LiM, but also in *Methanosaeta* cells. This staining was faint (Fig. 5). To exclude an archaeal target sequence for the probe, BLASTN of the probe sequence was performed against the metagenome Pe1–Pe3 and detected the 18mer target sequence of OP3 LiM and next two 15mer oligonucleotides. Sequences affiliated to *Methanosaeta* as well as the *Methanosaeta concilii* GP-6 genome had a 13mer of base-pairing with the probe OP3-565, with a GC content of 54%. The difference in length of these base-pairing regions strongly excludes a false-positive signal and suggests that the signal of the probe OP3-565 in *Methanosaeta* cells originate from the hybridization to rRNA of OP3 LiM present in the archaeal cells.

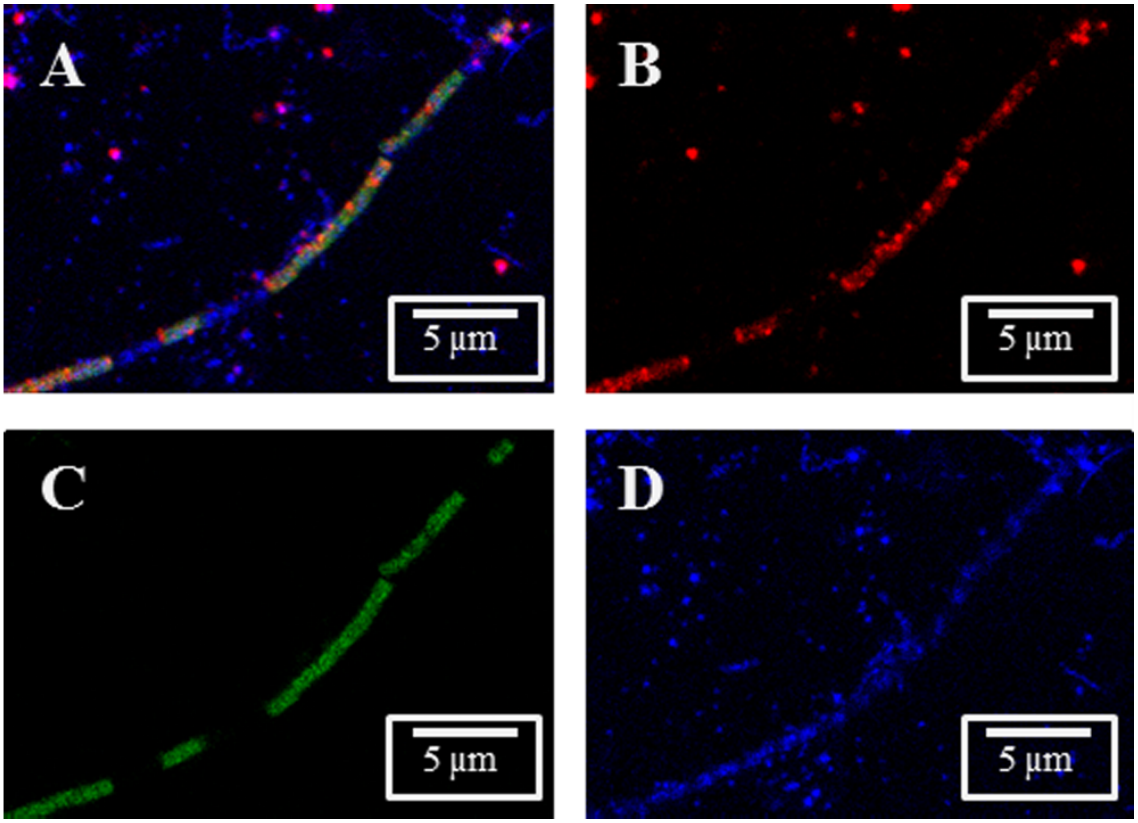


Fig. 5. Detection of rRNA from OP3 LiM by CLSM. Detection of *Archaea* and OP3 LiM in a methanogenic enrichment culture in CLSM images. Overlay (A) and individual signals of OP3 LiM rRNA (B), *Archaea* rRNA (c) and DNA (D), were obtained by double CARD-FISH with probes OP3-565 together with helper oligonucleotides and ARCH-915 as well as DAPI staining, respectively. Scale bar represents 5 µm.

The 23S rRNA of OP3 LiM contains a group I intron at position 2,061 to 2,768 bp including an open reading frame (ORF) encoding for a LAGLIDADG/HNH homing endonuclease. CARD-FISH with three probes and six helper oligonucleotides targeting the intron RNA was performed to visualize the presence of the intron in enrichment cultures, together with taxon-specific probes. The intron was detected in several cells with different morphotypes. Cells of filaments that did not contain DNA according to DAPI-staining contained the intron (Fig. 6). Intron containing cells were identified as part of an archaeal filament in double CARD-FISH experiments with the archaeal probe (Fig. S11). Together with the faint signal of OP3-LiM rRNA in filamentous cells, this observation suggested a mobility of ribonucleic acid molecules from OP3 LiM cells into *Methanosaeta* cells. Intron RNA was also detected in cells that had not the *Methanosaeta* morphology (Fig. S11–S16). Double CARD-FISH experiments together with the OP3 LiM probe confirmed that the intron was detected outside of OP3 LiM (Fig 6; Fig. S12, S15, S16).

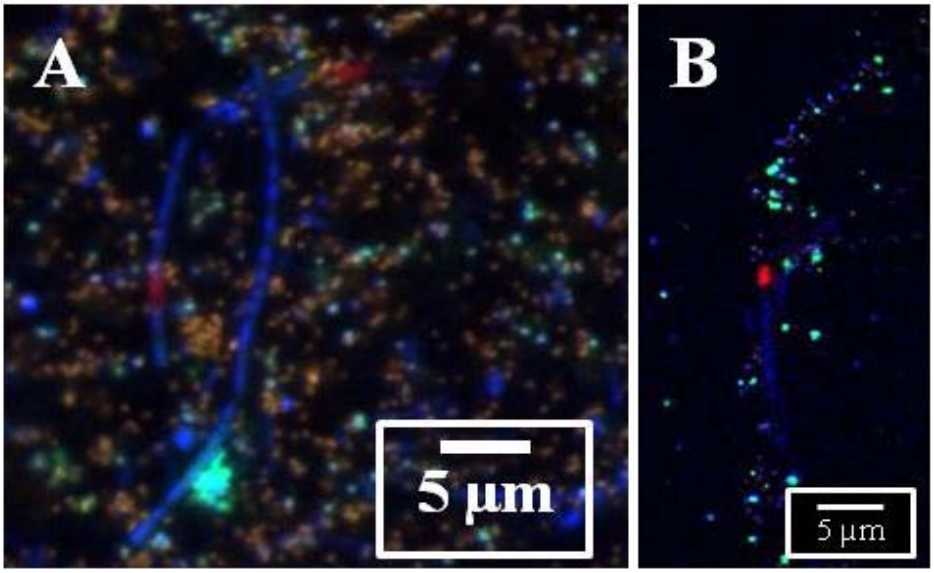


Fig. 6. *In situ* detection of intron RNA originating from OP3 LiM by light microscopy (A) and confocal laser scanning microscopy (B). DNA was stained with DAPI (blue). In a double CARD-FISH experiment, OP3 LiM rRNA was detected with OP3-565 (green) and the intron RNA was detected with three intron probes hen1-2235, hen2-2309 and hen3-2538 (red). All probes were used with flanking helper oligonucleotides. Scale bars, 5 µm.

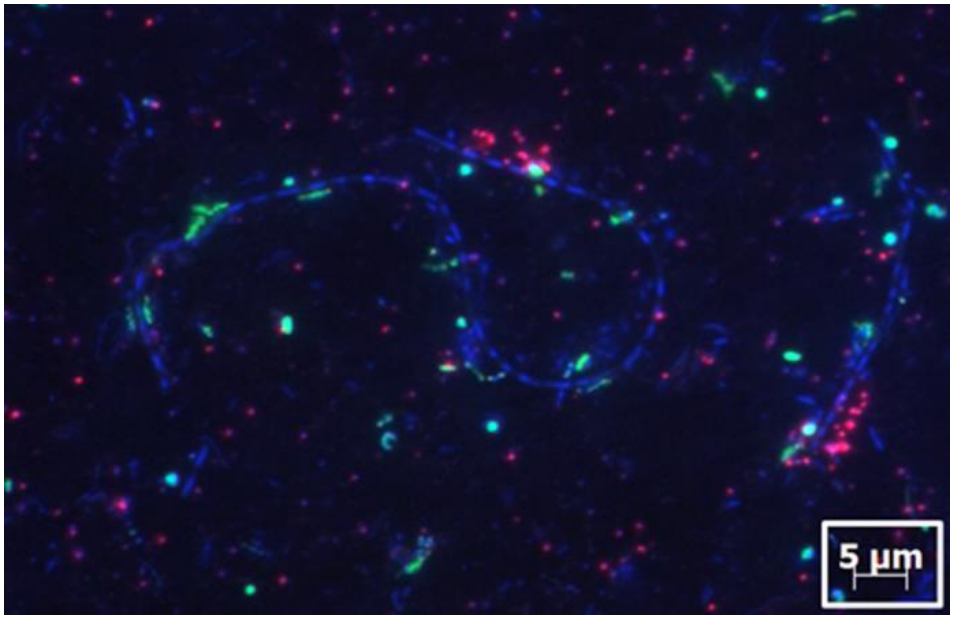


Fig. 7. Detection of *Bacteria* in a methanogenic enrichment culture. DNA is stained with DAPI (blue), the 16S rRNA probe mix EUB338 I–III revealed bacterial cells (excluding OP3 LiM) (green) and the OP3 LiM-specific probe OP3-565 detected OP3 LiM cells (red) in double CARD-FISH experiments. Scale bar, 5 µm.

Bacteria also had OP3 LiM cells on their surface (Fig. 7 and 8). The probe mix EUB338 I–III does not detect OP3 LiM due to mismatches in the target 16S rRNA sequence (19). A range of morphologies was identified as *Bacteria* including vibrios of different sizes, large coccoid cells as well as short and thin rod-shaped cells (Fig. 7 and 8). The bright CARD-FISH signal of several OP3 LiM cells attached to individual bacterial cells highlights their metabolic activity. Intron RNA was detected in one bacterial cell aggregate (Fig. S13). Other intron signals in double CARD-FISH experiments together with the EUB338 I–III probe mix did not detect the intron in a cell containing bacterial rRNA (Fig. S13 and S14).

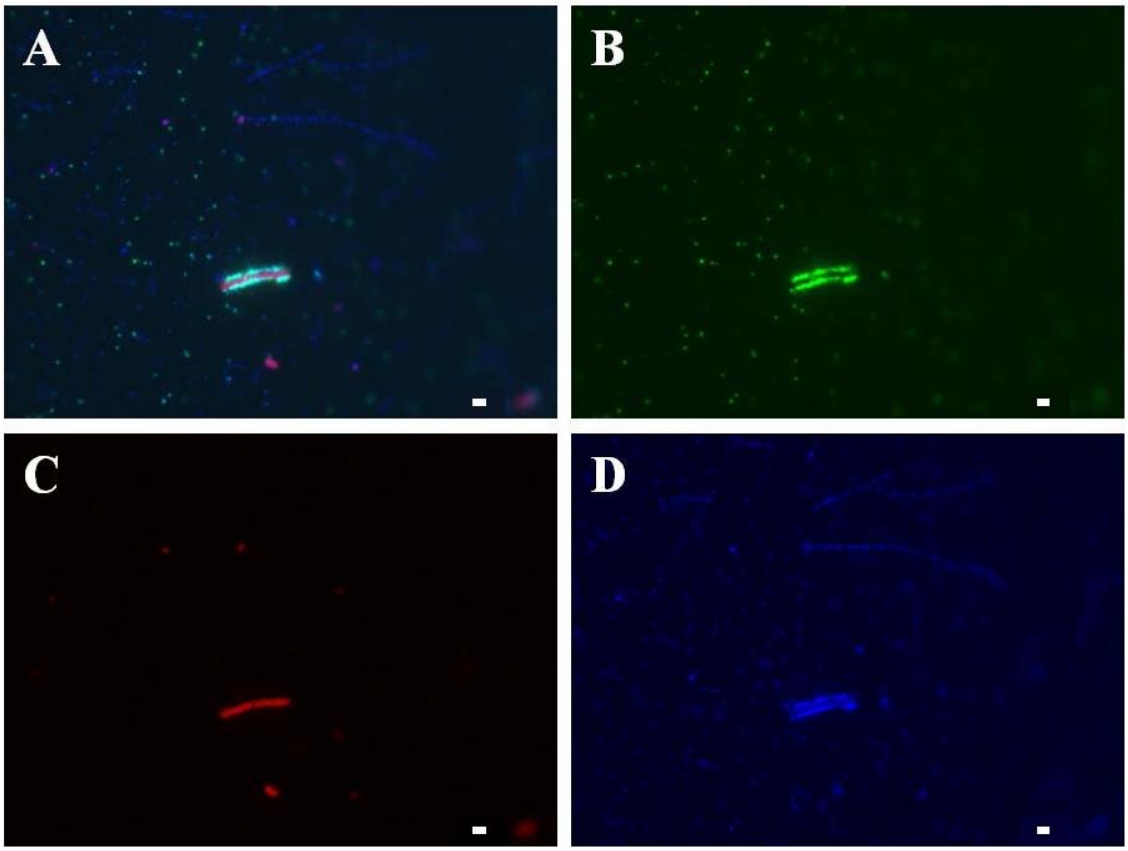


Fig. 8. Detection of OP3 LiM and *Bacteria* by double CARD-FISH. The epifluorescence micrographs show cells from culture MM-376 visualized with DAPI (blue, A and D), probe OP3-565 for OP3 LiM cells (green, A and B), and probe mix EUB338 I–III for *Bacteria* (red, A and C). Scale bars, 2 µm.

The OP3 LiM genome.

The OP3 LiM-enriched metagenome with 454 and MiSeq datasets provided a draft genome with uncertainties in repetitive regions. Combinatorial PCR reactions ordered the contigs of repetitive elements or confirmed suspicious regions including the presence of three repeats of about 2,000 bp with small sequence variations mixed with two completely identical sequences of 2,506 bp within the ORF of the very large multienzyme surface protein (Fig. S17A).

The coverage of 33.68% 454 reads and 6.97% MiSeq reads at the final OP3 LiM genome confirmed the enrichment of OP3 LiM cells.

The closed OP3 LiM genome was confirmed by mapping of metagenomes obtained from a second biological sample which was size-fractionated (Pe1, Pe2 and Pe3). Increased concentrated OP3 LiM cells in the different fractions are confirmed by the OP3 LiM genome mapping coverage of 2.77% Pe1, 5.94% Pe2 and 12.66% Pe3 high-quality reads. Sequence differences between the two biological samples were mainly homopolymers caused by 454 reads. Manual curation gave the final genome of OP3 LiM.

The OP3 LiM genome is a single, circulate chromosome of 1,974,201 bp in length (Fig. 9). We placed the genome start point at the origin of replication (*ori*) at the N-terminus of *dnaA*. The asymmetry of the nucleotide composition between leading and lagging strands indicated the first base of the start codon of *dnaA* coding for the chromosomal replication initiator protein DnaA (24) as base number one. The terminus is about 950,000 bp away from the *ori*, based on the shift points of the GC skew graphs (25). Gene orientation is highly ordered. 85% of the genes are transcribed in the direction of the DNA replication and only 15% of the genes are transcribed opposite to the direction of the DNA replication.

The genome has an average GC content of 52.9% with the overall sequence composition of the forward strand of 23.56% A, 26.59% C, 26.31% G, and 23.52% T. Rapid Annotations using Subsystems Technology (RAST) annotation revealed 2,015 protein-coding open reading frames (ORFs) with an average length of 894 bp. 37% of ORFs covered known metabolic processes in RAST-subsystems with 735 proteins, 44 tRNA genes and one rRNA operon encompassing a 5S (1,194,960 bp to 1,195,074 bp), a 23S (1,195,261 bp to 1,198,995 bp), and a 16S (1,199,780 bp to 1,201,360 bp) RNA gene. The 16S rRNA gene of OP3 LiM has a sequence identity of 99% to 16S rRNA LiM gene clones from our previous study (19), an identity of 77% to the 16S rRNA of the next related single cell genome (Genbank acc. no.

ASOC01000103), and an identity of 94% to the next related metagenome-derived OP3 population genome (Genbank acc. no. MNVR01000026).

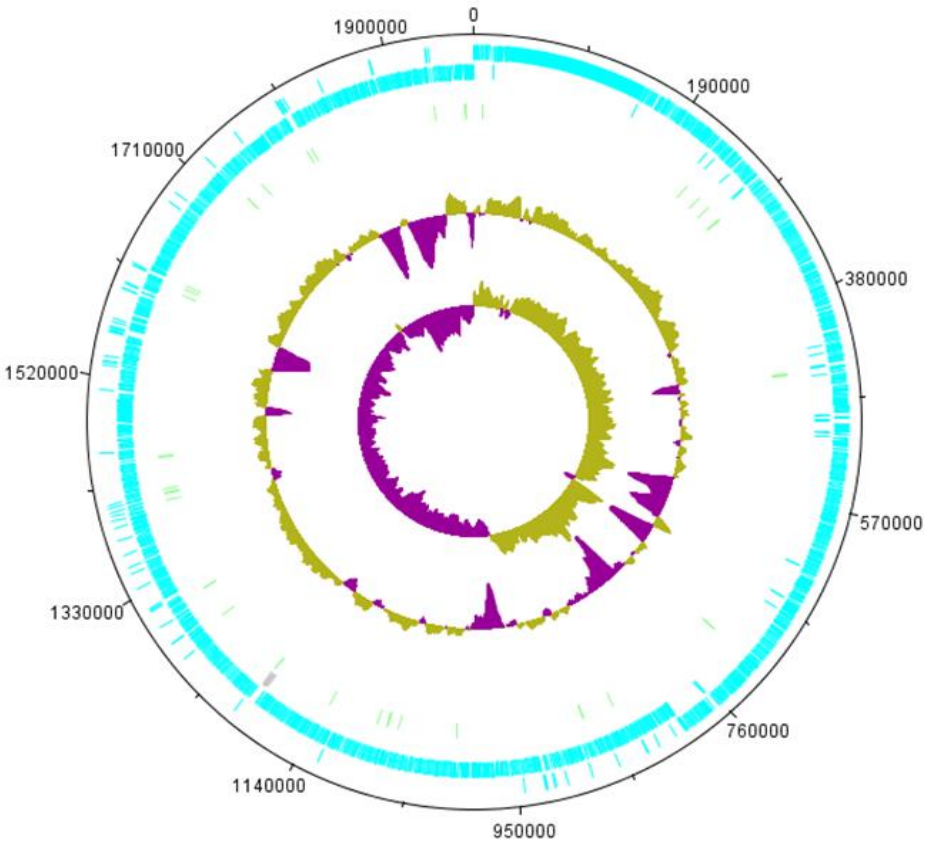


Fig. 9. Circular representation of the OP3 LiM genome. From outside to inside, the circles display: (1) Physical map with a start of the replication origin, (2) coding sequences transcribed in forward direction, (3) coding sequences transcribed in reverse direction, (4) rRNA, (5) tRNA, (6) G+C percent plot, (7) GC-skew [G–C]/[G+C] (http://www.sanger.ac.uk/science/tools/dnaplotter).

A group I intron in the 23S rRNA gene of OP3 LiM at position 2,061 bp to 2,768 bp possesses a gene for a homing endonuclease (position 2,225 bp to 2,779 bp). This intron is actively excised. The potential transcript of 3,733 bp for the 23S rRNA including the intron RNA was not present in the capillary electropherogram of the whole RNA preparation (Fig. S18). Mapping of 134,406,170 RNA-seq reads to the OP3 LiM genome revealed a high coverage of the 23S rRNA gene and a lower coverage of the intron. On average, the 23S rRNA region before the intron had a read coverage of 998,331 per base and the region behind the intron had a read coverage of 772,062 per base. Intron RNA was presented by 5,918 reads per base. Detailed analysis of the reads at the intron borders revealed the presence of both the

23S rRNA gene with intron and the mature 23S rRNA in the RNA preparations. The insertion site corresponds to the *E. coli* 23S rRNA gene position 1,917. The LAGLIDADG homing endonuclease was detected in the metaproteomes.

A group II intron was found in a region with a large number of various inserted genetic elements. This region of 6,505 bp was located from position 648,611 bp to 655,115 bp. The group II intron is located between two copies of a gene of unknown function (peg.582 and peg.585) (Fig. S17B). The genes called CHP1 and CHP2 have a nucleotide identity of 95%. In each gene is a highly-variable region of 83 bp that is flanked by repetitive sequences and is ending 28 bp in front of the stop signal (Fig. S19). On both sites of the genetic element are genes for 23S rRNA intervening sequence (IVS) proteins (122 aa; peg.581 and 209 aa; peg.586). The group II intron encodes a third IVS protein (133 aa; peg.583) and a retron-type RNA-directed DNA polymerase (reverse transcriptase) (peg.584). Downstream of this gene is a G homopolymer of 21 bases.

Defense systems against foreign nucleic acids are restriction modification (RM) systems (26). An expressed type I RM cluster (*hsdMSR*) peg.557/558/560 with an unexpressed *hsdS* copy peg.559, a type II RM (expressed restriction endonuclease peg.1905 and unexpressed modification methylase peg.1906) and an expressed subunit of type III RM system (peg.612) were found in the OP3 LiM genome.

Candidate regions for clustered regularly interspaced short palindromic repeats (CRISPR) providing acquired immunity against foreign genetic elements (27) are identified at positions 875,578 bp to 875,727 bp, 1,062,302 bp to 1,062,425 bp and 1,625,552 bp to 1,625,657 bp as questionable, because each CRISPR locus was separated by only one spacer, located on the positions 875,626 bp (54 bp in length), 1,062,335 bp (58 bp in length) and 1,625,587 bp (36 bp in length), respectively.

Expressed are genes (peg.1109 with *uvrA*, peg.1010 with *uvrB*, and peg.1508 with *uvrC*) encoding all subunits for the UvrABC endonuclease. This multienzyme complex is involved in DNA repair by nucleotide excision repair (NER). An expressed *uvrD* of the superfamily I DNA helicase (peg.988), required for uvrABC excision repair and believed to function by unwinding duplex DNA (28), is also found in the OP3 LiM genome.

Physiology of OP3 LiM.

1,298 proteins of OP3 LiM were identified in metaproteomes of Pe1, Pe2 and Pe3 and provided a view on the metabolism (Table S1). Most abundant is a 18-stranded beta-sheet outer membrane pore protein (peg.955) that facilitates the uptake and for export of macromolecules (Table S2 and S3). Three other large proteins with transmembrane helices are among the most frequently detected proteins, a GPI-membrane anchored actin/protein binding protein with KELCH repeats (499 kDa; peg.1770), a protein with unknown function (326 kDa; peg.589) and a very large multi-enzyme surface protein (peg.37). All three proteins are abundant due to their size (Table S2), not by number (Table S3).

The gene (peg.37) encodes for a very large multi-enzyme surface protein with 42 predicted transmembrane helices and has a size of 4,384 kDa. The 39,678 amino acids of the protein represent a multi-enzyme with 308 domains thereof 43 were identified as conserved domains. The degradation of polysaccharides is the most abundant function, represented by 8 glycosyltransferases, 2 glucosidases and a sugar epimerase. Three peptidases and a chaperone act on proteins. A phosphatase, a dehydrogenase, a methyltransferase and an acetyltransferase complete the identified range of degradative enzyme domains. As environmental signal sensor serves a signal receiver domain and a cellular signaling response may be induced via cyclic guanine-nucleotides by a diguanylate cyclase/phosphodiesterase (GGDEF & EAL domains). The nucleotide metabolism is influenced by 3 ATP-binding domains, a protein kinase and two nucleotide kinases, two ppGpp synthases/hydrolases as well as two ssDNA and two dsDNA binding domains. Peptides of the first 8,290 amino acids of the 39,678 amino acids multi-functional protein containing a peptidase, cellulose synthase and glycosyltransferase were clearly detected in the metaproteome.

Abundant cytosolic proteins are GroES, EF-Tu, DNA-directed RNA-polymerases, glutamine synthetase and carbamoyl phosphate synthase. OP3 LiM has the potential to conserve energy from the phosphoester bonds in nucleic acids. A polynucleotide nucleotidyl transferases (peg.1256) moves a nucleoside monophosphate from ribonucleic acid onto phosphate, thus synthesizes a nucleoside diphosphate. Another source of ATP is pyruvate phosphate dikinase (peg.972). A pyruvate:ferredoxin oxidoreductase with four expressed subunits gamma (peg.113), delta (peg.114), alpha (peg.115), and beta (peg.117) provides reducing equivalents for energy conservation via a membrane-integrated RNF-complex (ferredoxin:NAD oxidoreductase) (gene cluster peg.1598–1603). Furthermore, a proton-translocating pyrophosphatase (peg.63) is highly expressed. Enzymes acting on the 3-carbon

metabolites of the glycolysis and anabolic enzymes of the branched-amino acid biosynthesis pathway are also present in large quantities.

The periplasm contains a highly expressed peptidyl prolyl cis-trans isomerase (peg.12), a Clp protease (peg.1300) and an ATPase (peg.15) that has in addition to the CpaF/VirB11-like_ATPase domain an N-terminal CobQ/CobB/MinD/ParA nucleotide binding domain suggesting a participation in nucleic acid transfer. Although OP3 has no flagellum, an abundant protein is MotD (peg.236), a stator ring protein for a flagellum forming a pore in the inner membrane that is usually filled by a type III secretion system and accessor proteins. The second stator protein MotB (peg.1609) is also present. Other abundant proteins are an outer membrane protein assembly factor, the Sec translocase SecD/SecF (peg.1065) and the bacterial and archaeal type of the chromosome segregation protein SMC (peg.52) that binds nucleic acids. The bacterial actin MreB (peg.1197) is also highly expressed.

We searched for other proteins potentially involved in the uptake of nucleic acids or other macromolecules. Besides the aforementioned extracellular nucleotide-binding CpaF-related ATPase, three PilT ATPases (peg.695, peg.1506, and peg.1804) are present to import macromolecules. These ATPases also power depolymerization of the pilus polymer. Extension by polymerization of pilin subunits maybe driven by several expressed genes for PilB ATPases or orthologs (peg.846, peg.1097, peg.1186, peg.1188, peg.1189, peg.1248, peg.1249, and peg.1789). Of the *pil* genes, CpaB (peg.13), PilM (peg.1101), PilO (peg.1179), three PilQ (peg.1098, peg.1177, and peg.1779) and three PilZ (peg.169, peg.393, and peg.788) were expressed. A PilA-related pseudopilin PulG (peg.1182) is the most abundant pilin. OP3 LiM has an active Sec secretion system and a type II secretion / type IV pilus assembly system. Several loci contain *pil*-related genes, including one locus in front of the very large multi-enzyme surface protein that contains several export proteins of a type II secretion system.

The OP3 LiM proteomes highlighted the presence of large outer membrane pores and of depolymerizing and polymer uptake domains in the periplasm. The metabolism of OP3 LiM is heterotroph. The genome offers a syntrophic life on hydrogen or formate, with a carbon dioxide fixation on the Wood–Ljungdahl pathway using formate dehydrogenase, the reduction of a formyl group on tetrahydrofolate and methyl transfer to an acetyl-CoA synthase. But these enzymes are not highly expressed. Among the four hydrogenase, an uptake hydrogenase (hydrogen:NAD oxidoreductase) was the most expressed hydrogenase. Interestingly, a high expression of serine hydroxymethyltransferase (peg.1522) suggested

serine and glycine as precursors of C1-compounds. Glycolysis with a class II fructose-bisphosphate aldolase (peg.192) is combined with substrate level-phosphorylation and a RNF-complex for the formation of a membrane potential. Building blocks (amino acids, ribose, nucleotides, and fatty acids) are synthesized according to expressed enzymes. OP3 LiM may salvage these monomers from the prey, but does not seem to depend on salvage pathways. D-alanine (peg.1279) and D-glutamate racemases (peg.342) as well as peptidoglycan biosynthetic enzymes are expressed. Together with the presence of peptidoglycan-binding domains in several secreted proteins it suggests the presence of a murein saculus in OP3 LiM. Genes for the synthesis of the LPS core glycolipid (peg.644) suggest the well-known lipid asymmetry in the outer membrane.

5.3 Discussion

The characterization of the phylotype OP3 LiM indicated a chemoheterotrophic metabolism confirming suggestions from other studies (18, 29). Major nutrient for the cell is likely the organic matter presenting the surface of microorganisms. Cellulosomes of clostridia have been described as extracellular, surface-associated “highly efficient nanomachines” that had evolved to perfect depolymerisation multi-enzyme complexes for plant cell wall complex carbohydrates (30). Similar to plant polysaccharides, archaeal and bacterial surface polysaccharides represent an energy-rich growth substrate and OP3 LiM has enzymes for their utilisation. The bacterium has expressed extracellular enzyme domains for the depolymerisation of polysaccharides and proteins. This attack on the surface of the prey microorganism may open an entry to the cytosol of the prey. The presence of excreted ATPases in OP3 LiM is unusual for microorganisms, usually ATP-consuming enzymes are restricted to the cytosol. The purpose of the ATPases in the attack of OP3 LiM may be a depletion of the energy charge of the prey cell that paralyses the prey metabolism. Then OP3 LiM cells can exhaust the prey´s macromolecular content.

OP3 LiM is a coccus of about 0.2 µm to 0.3 µm in diameter. This “small eats big” is usual for bacterial predators (6). The size also means that OP3 LiM cells are not or barely visible in phase-contrast microscopes which have an optical resolution of 0.3 µm. The world of such very small bacteria has to be explored by optical detection of nucleic acids by *in situ* hybridization that identifies the phylotype together with phase-contrast and by electron microscopical images. Small bacteria are expected to have small genomes. The closed genome of OP3 LiM was obtained from a close-to-clonal population because the methanogenic enrichment culture originated from a dilution-to-extinction culture with an inoculum of one microliter prepared in 1999 and is kept at 3–4 annual generations by one annual transfers of 10% v/v. The genome is small and highly organized. The leading strand contains after the genes involved in DNA replication the genes for a secretion pathway and the very large multi-enzyme surface protein involved in polymer degradation. This is the expected position for highly expressed proteins minimizing collisions of replicating and transcribing polymerases (31–32).

The genome includes several mobile elements. A group I intron in the 23S rRNA gene has a gene for a homing endonuclease in the peripheral stem-loop regions of the group I ribozyme (33–34). The identified insertion position 1,917 (referring to the *E. coli* sequence) has been reported for other 23S rRNA introns in *Coxiella burnetii* (35), *Thermotoga*

subterranean (35), *Thiomargarita* sp. NAM092 (36) and *Synecoccus* sp. C9 (33). 23S rRNA introns seem also to be frequent in groundwater-associated bacteria, with the 23S rRNA position 1,917 as a preferred insertion site (37). The intron excision is an essential process for OP3 LiM. We investigated the fate of the intron RNA in *in situ* hybridization studies and detected huge amounts in cells that lacked DNA and rRNA. The filamentous morphology of *Methanosaeta* revealed that individual cells in the filaments had accumulated the intron RNA. So far, an interdomain horizontal gene transfer (HGT) by conjugation has been demonstrated by a plasmid transfer from *Bacteria* to *Archaea* (38). Although the mechanism of transfer from the bacterium OP3 LiM into the dead archaeal cell with a nearly intact cell envelope is not understood on the molecular level, it demonstrated the mobility of nucleic acids between predator and prey cells. A second experimental observation, the presence of small amounts of OP3 LiM rRNA in cells of *Methanosaeta*, supported the presence of a polynucleotide export mechanism in OP3 LiM or the presence of a cytosolic continuum linking both cytosols of prey and predator. As coccal cells usually have no bacterial actin MreB, the presence of this protein in OP3 LiM suggests the presence of a MreB foothold in OP3 LiM cells as structural ground for the transport processes between the cells, e.g. the energy-driven uptake of proteins or nucleic acids by several PilT ATPases. Several expressed genes of both ATPases, PilT and PilB, are present in the genome of OP3 LiM suggesting that the bacterium may be capable of undergoing cycles of extension and retraction (39), also for gliding on the prey surface.

A reverse transcriptase is the hallmark of a group II intron. The high sequence microdiversity of the potential insertion sites was surprising in view of the very low single-nucleotide polymorphism of the population genome. We attribute this observation tentatively to an active retro-transcriptase, an RNA-directed DNA polymerase, in the OP3 LiM population.

This study has provided a closed genome of high quality of the OP3 LiM population in limonene-degrading methanogenic enrichment cultures. The genome has been confirmed by recruitment plots from additional metagenomes of the enrichment cultures. The 16S rRNA gene sequence indicated a large phylogenetic distance to the next described candidatus species or species. The phylogenetic probe OP3-565 showed the morphology of the identified cells and guided the physical enrichment of OP3 LiM cells. Metaproteomes provided insights to infer the putative metabolism and the observed transfer of nucleic acids into the prey cell revealed the lifestyle of OP3 LiM. Based on these observations, we propose to name the OP3 LiM cells *"Candidatus* Vampirococcus archaeovorus".

5.4 Description of “*Candidatus* Vampirococcus archaeovorus”

“Candidatus Vampirococcus archaeovorus” [N.L. n. *vampirum* (from Hung. *vampir*), vampire; N.L. masc. n. *coccus* (from Gr. masc. n. *kokkos*, grain, seed), berry, coccus; N.L. masc. n. *Vampirococcus*, a vampire-like coccus; Gr. masc. adj. *archaios* (Latin transliteration *archaeos*), ancient; L. v. *voro*, to eat, to devour; N.L. masc. adj. *archaeovorus*, archaea (ancient microorganisms) devouring.

Represented by the phylotype OP3 LiM and its genome (Genbank acc. no. CP019384). OP3 LiM was found in a limonene-degrading methanogenic enrichment culture. Cells are coccoid shaped with diameters around 0.2 µm to 0.3 µm and occur free-living and attached to microorganisms. Phylogenetic analyses of the 16S rRNA gene indicate that this strain is different from all other recognized candidate genera or genera and belongs to the *candidate phylum* Omnitrophica, formerly candidate division OP3. The bacterium is maintained in slowly growing *Methanosaeta*-rich methanogenic enrichment cultures in freshwater medium at 28°C.

5.6 Materials and methods

Cultivation of methanogenic enrichment cultures.

The methanogenic enrichment culture originated from a wastewater sample taken in 1997 (40). From a dilution-to-extinction series prepared in 1999, a culture that had been inoculated with one microliter enrichment culture became the origin of all cultures investigated in this study. Twelve lineages were established in 2005 and maintained with an annual transfer of 10% v/v inoculum. The cultures contained 300 ml fresh water methanogenic media including 2 mM acetate and 1 mM cysteine, 30 ml 2,2,4,6,8,8-heptamethylnonane (HMN), and 1.5 ml of *R*-(+)-limonene in 500 ml borosilicate bottles (19,40). OP3 LiM-specific PCR and CARD-FISH were performed to select lineages for the experiments.

Cell separation by Percoll density gradient centrifugation.

Cell biomass from 100 ml of enrichment culture was pelleted at 15,500x *g* for 10 min in a Beckman ultracentrifuge rotor 70.1 Ti (Beckman, Palo Alto, CA). The pellet was resuspended in a mixture of 45 ml Percoll (GE Healthcare, Freiburg, Germany) and 5 ml 1.5 M NaCl. Portions of 10 ml suspension were centrifuged at 36,680x *g* for 60 min in the aforementioned rotor. Gradient fractions of 1 ml size were assayed for the presence of OP3-LiM cells by applying OP3 LiM-specific PCR with a template dilution series and CARD-FISH with the probe OP3-565. Fractions enriched in OP3 LiM cells had densities of 1.05 ± 0.05 g/cm^3 and were combined from several separations. The enriched fractions were further concentrated with a second gradient centrifugation using the aforementioned conditions. OP3 LiM cells were visible at a density of 1.05 g/cm^3 and were collected for an OP3 LiM-enriched metagenome.

Cell separation by differential centrifugation.

For the separation of large cells and aggregates from small cells in enrichment cultures, a Beckmann SW28 ultracentrifuge rotor was used at 7,600 rpm (7,643x *g*) for 20 min, corresponding to a sedimentation coefficient of 10,000 S. Each pellet (Pe1) was resuspended in 1 ml 10 mM Tris, 1 mM EDTA, pH 8.0 (TE). The supernatant was centrifuged at 27,000 rpm (96,467x *g*) for 160 min, corresponding to a sedimentation coefficient of 100 S. The pellet of each tube was resuspended in 0.5 ml TE and a 10,000 S pellet of aggregated

cells and a few large cells (Pe2) was obtained at 12,400 rpm (16,331x *g*, 10,000 S) for 3 min. Cell suspensions (Pe1 and Pe2) and the supernatant of the last centrifugation (Pe3) were stored at –80°C.

Transmission electron microscopy (TEM).

For negative-staining transmission electron microscopy, bacterial cultures were absorbed onto carbon film, washed in 20 mM Tris/HCL, 1 mM EDTA, pH 6.9, and stained for 1 min with 4% (w/v) aqueous uranyl acetate (41). After transfer onto copper grids and air drying, samples were examined in a Zeiss EM109 TEM (Zeiss, Oberkochen, Germany) operated at 80 kV and calibrated magnifications. Cells were analyzed using the program MeasureIT (Olympus Soft Imaging System GmbH, Münster, Germany).

Scanning electron microscopy (SEM).

Cells fixed with formaldehyde (1.3% w/v in 1 x phosphate buffered saline (PBS) (pH 7.4)) were spotted on silicon wafers (5 x 7 mm, PLANO, Wetzlar, Germany) and incubated for 60 min at room temperature (RT). Then cells were dehydrated in an ethanol series (30%, 50%, 70%, 80%, and 96% v/v, each for 10 min) and critical point dried (LEICA EM CPD 300, Leica, Wien, Österreich). Micrographs were obtained with a Quanta FEG 250 SEM (FEI, Hillsboro, OR).

Extraction of nucleic acids.

OP3 LiM-enriched fractions were extracted for genomic DNA according to Martín-Platero *et al.* (42). RNA and DNA from fractions of the differential centrifugation (Pe1, Pe2, Pe3) were obtained with the RNA PowerSoil® Total RNA Isolation Kit and the RNA PowerSoil® DNA Elution Accessory Kit (Mo Bio, Laboratories, Carlsbad, CA). After in-house quality control by spectroscopy and agarose gel electrophoresis, sample quality control by capillary electrophoresis and sequencing was performed by the Max Planck-Genome-centre Cologne, Germany (http://mpgc.mpipz.mpg.de/home/). For PCR analyses, DNA was extracted from biomass of 1ml-samples of enrichment culture using the FastDNA SPIN Kit for Soil (MP Biomedicals, Santa Ana, CA) according to the manufacturer’s instructions. The nucleic acid was dissolved overnight at 4°C and the solution was clarified by centrifugation at 14,000 rpm for 5 min.

OP3 LiM-specific PCR.

OP3 LiM was detected using as template 1.0 µl (a few ng) extracted genomic DNA or 1.0 µl culture of an $OD_{600\ nm}$ of 0.2–0.5 with freeze-thaw fractured cells (43) together with 1.7 µM of each primer OP3-565F and OP3-1481R (Table 1), 15.0 µl 2 x GoTaq mastermix (Promega, Madinson, WI) and 13 µl water. The PCR protocol was 4 min at 94°C, 30 cycles of 94°C for 1 min, 62°C for 1 min and 72°C for 3 min, and finally 72°C for 10 min. After amplicon analysis by separation in a 1% w/v agarose gel and ethidium bromide staining, sequences were obtained from dideoxynucleotide-terminated oligonucleotides. The sequencing reaction was performed for 60 cycles with an initial denaturing step of 20 sec at 96°C, a denaturation of 10 sec at 96°C, 5 sec at 62°C, and 4 min at 62°C. The products were purified by molecular sieve chromatography using Sephadex G50 superfine (GE Healthcare Life Sciences, Freiburg, Germany) and were separated on an ABI Prism 3130 XL Genetic Analyzer (Applied Biosystems, Foster City, CA, USA).

Table 1. PCR primers used in this study.

Primer name	Sequence (5´–3´)	Position	Application	Reference
OP3-565F	GGGTGTAAAGGGCAGGTA	608–626[a]	OP3-specific 16S rDNA forward primer	This study
OP3-1481R	TACGACTTAGCGCCAGTC	1525–1543[a]	OP3-specific 16S rDNA reverse primer	This study
8-27F	AGAGTTTGATCCTGGCTCAG	8–27[b]	Bacterial universal 16S rDNA forward primer	(44)
907R	CCGTCAATTCMTTTGAGTTT	907–926[b]	Bacterial universal 16S rDNA reverse primer	(45)

[a] OP3 LiM 16S rRNA location.

[b] *E. coli* 16S rRNA location.

Fluorescence *in situ* hybridization of 16S rRNA.

For cell identity visualization, we used the CARD-FISH technique (46). For brighter signals of OP3 LiM cells, clone sequences (Genbank acc. no. FN646451.1, FN646447.1, FN646441.1, FN646440.1, and FN646435.1) were used to manually design partially degenerated helper oligonucleotides, two for each adjacent side of probe OP3-565 (Table 2) and used at probe concentration.

Table 2. Helpers (non-labeled) used and designed in this study.

Helper name	Helper sequence (5´–3´)	Position
H548-A[a]	AATAAATCCGAGTAACGC	590–608[c]
H548-C[a]	AATCAATCCGAGTAACGC	590–608[c]
H583-TC[a]	CTCCCCACTTGTCAGGCCGCC	626–647[c]
H583-CT[a]	CCTCCCACTTGTCAGGCCGCC	626–647[c]
Hhen1-2210[b]	TCTGGTTTCATAAAGACCTCCTTCT	2210–2235[d]
Hhen1-2255[b]	TCCTTCTCCGTCGGCAAAAC	2255–2275[d]
Hhen2-2288[b]	AGTCTTGCCGTTGTCTGAAGG	2288–2309[d]
Hhen2-2329[b]	CTGGGAGATGTTGAAACAAAGAGA	2329–2353[d]
Hhen3-2516[b]	CAGAATTTCGCAAAATCCCGTT	2516–2538[d]
Hhen3-2561[b]	GCTACTCTTAAGATGTTCCCCCT	2561–2584[d]

[a/b] Used in a mix.

[c] OP3 LiM 16S rRNA location.

[d] OP3 LiM 23S rRNA location.

One ml of enrichment cultures was fixed with formaldehyde (1.3% w/v in 1 x PBS (pH 7.4)) for 60 min at RT. Fixed cultures were filtered on 0.2 µm GTTP filters (Millipore, Darmstadt, Germany) and three times washed with 15 ml 1 x PBS (pH 7.4). After air-drying, permeabilization was performed for 60 min at 37°C by lysozyme treatment (10 mg/m1). Probes (Table 3) were hybridized for 160 min at 46°C. After staining with DAPI (1 µg/ml), cells were visualized using an epifluorescence microscope (Nikon Eclipse 50i; Nikon, Tokyo, Japan) or a confocal laser scanning microscope (CLSM) (Zeiss LSM 780, Zeiss, Oberkochen, Germany).

The visualization of two populations required two separate CARD-FISH experiments with two probes and differently labeled tyramides for the first and second signal amplification, respectively. The order of probes and tyramides was optimized. Horseradish peroxidase (HRP) present at the first probe was inactivated after the signal amplification by incubation of the filter with 3% H_2O_2 for 10 min at RT. After rinsing with 1 L MQ water and air-drying filters were stored overnight at –20°C and the second CARD-FISH experiment was performed the next day.

Table 3. HRP-labeled oligonucleotide probes used in this study.

Probe name	Probe sequence (5´–3´)	Position	Target group	FA (%)[a]	Reference
EUB338 I[b]	GCTGCCTCCCGTAGGAGT	338–355	Most *Bacteria*	35	(47)
EUB338 II[b]	GCAGCCACCCGTAGGTGT	338–355	*Planctomycetales*	35	(48)
EUB338 III[b]	GCTGCCACCCGTAGGTGT	338–355	*Verrucomicrobiales*	35	(48)
ARCH-915	GTGCTCCCCCGCCAATTCCT	915–934	Domain *Archaea*	35	(49)
OP3-565	TACCTGCCCTTTACACCC	608–626[d]	Candidate OP3 LiM	30	(19)
hen1-2235[c]	CCGCCAAGTAGTAGCCGATT	2235–2255[e]	Intron in 23S rRNA of OP3 LiM	20	This study
hen2-2309[c]	AACTTTCCACGGTGACTTGT	2309–2329[e]	Intron in 23S rRNA of OP3 LiM	20	This study
hen3-2538[c]	TCTTTATCATTTCTGCCAGTTCG	2538–2561[e]	Intron in 23S rRNA of OP3 LiM	20	This study

[a] FA, fomamide concentration of the hybridization buffer.

[b/c] Used in a mix.

[d] OP3 LiM 16S rRNA location.

[e] OP3 LiM 23S rRNA location.

CARD-FISH detection of intron RNA.

The group I intron in the 23S rRNA gene of OP3 LiM (position 2,061 to 2,768 of 23S rRNA gene) includes a gene for a LAGLIDADG/HNH homing endonuclease. Targeting within the functional gene, probe and flanking helper oligonucleotides were designed using Primer3 v. 4.0.0 (http://sourceforge.net/projects/primer3/) (Tables 2 and 3). Specificity was tested against the OP3 genome; *Methanosaeta* genomes (Genbank acc. no. NC_015416.1, CP003117.1, CP000477.1, and LKUG01000745.1) as well as population genomes of members of the limonene-degrading, methanogenic enrichment cultures (*Aminiphilus*, *Anaerolinea*, *Coriobacterineae*, *Desulfovibrionaceae*, *Marinilabiliaceae*, *Mesotoga*, *Methanoculleus*, *Methanoregula*, *Methanosaeta*, *Methanospirillum*, *Prolixibacteraceae*, *Synergistaceae*, *Syntrophaceae*, *Syntrophobacteraceae*, *Thermoanaerobaculum*, and *Treponema*), obtained by Metawatt-3.5.2 binning (50), in Geneious R9 (Biomatters, Auckland, New Zealand). After optimization, a formamide concentration of 20% was used at

concentrations of 0.17 ng/µl for each of the oligonucleotide in the mixture of three probes and six helpers.

Genome assembly from an OP3 LiM-enriched metagenome.

The OP3 LiM genome was obtained from two biological samples. An OP3 LiM-enriched sample was sequenced by 454 Titanium pyrosequencing technology (450 bp reads, 454 GS FLX, Roche, Basel, Switzerland) as well as by MiSeq technology (2x 250 bp reads, Illumina, San Diego, CA, USA) by the Max Planck-Genome-centre Cologne, Germany (http://mpgc.mpipz.mpg.de/home/). Three cell size-fractionated samples (Pe1, Pe2, Pe3) provided DNA and RNA for metagenomes and metatranscriptomes using the Illumina HiSeq instrument (1 x 150 bp reads for RNA and 2 x 250 bp reads for DNA) (Tab. S4 and S5; Fig. S20).

454 pyrosequencing yielded 491,907 reads. An assembly of 454 reads was performed by Newbler v. 2.3 (51) and resulted in a metagenome of 5,779 contigs with 16,026,544 bp. Analysis of tetranucleotide frequencies using JSpecies (52) indicated that five of the six largest contigs covering 1,650,908 bp were part of the OP3 LiM genome. For further analyses, the raw reads of the 454 sequencing were processed in mothur-1.29.1 (53) resulting in 426,697 quality-controlled reads.

A MiSeq read set of the same biological sample was used to finish the genome. 9,888,618 paired-end reads were quality-controlled using FastQC (www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Raw reads were processed with dynamic trimming with SolexaQA v.2.2. (54) and normalization with Khmer 1.0 (Fig. S21, Workflow 1) (55). The assembly by SPAdes-3.1.0 (56) contained 28,618 contigs with 61,364,565 bp. The assembly was inspected using QUAST v-2.3 (57).

Contigs of the Newbler assembly were grouped by Metawatt-2.1. 22 contigs of two “*Planctomycetales*” bins with 1,949,258 bp were selected as target of a mapping of processed 454 and MiSeq reads within Geneious R8 (Biomatters, Auckland, New Zealand). In addition, MiSeq raw reads were also processed in different ways with tools of the BBMap package (version 32.27, http://sourceforge.net/projects/bbmap/) (Fig. S21, Workflow 2–4). Contigs of the two “*Planctomycetales*” bins were extended by read mapping using 454 and MiSeq quality-controlled reads and then *de novo* assembled by Geneious R8 (Biomatters, Auckland, New Zealand). The assembly was improved by fourteen rounds of read mapping with BBmap (version 32.27), assembly of the mapping reads in SPAdes-3.5 taking the actual assembly as

trusted and binning of contigs with Metawatt-2.1. When the improvement in the assembly became asymptotic, a MiSeq read dataset obtained with a different processing method was used for the next mapping. Almost all bin information was assembled in one contig. Finally, assemblies were based on the Newbler or the first Geneious assembly as trusted assembly to correct errors introduced in the assembly process. The comparison of the final assemblies revealed a linear presentation of a circular genome with different start points.

Manual visual inspection of mapping results of 454 and MiSeq reads obtained with Geneious R8 and with the Burrows-Wheeler Aligner (BWA) algorithm (Sequencher 5.3, Gene Codes, Avis Drive, USA) revealed questionable regions in the genome. These regions were often repetitive elements which were in general identified using REPuter (58) and dotplot visualization in Geneious R8 (Biomatters, Auckland, New Zealand). Flanking sequences of repetitive elements were manually identified in the visualization of sequence diversity of mapped reads. This indicated a false-positive assembly of two or more repetitive elements into one in the contig. Sequence comparison using dotplot visualization identified repetitive elements of the draft genome also in two small contigs of the 454 assembly assigned by Metawatt-2.1 to the OP3 LiM bin. Both contigs were *de novo* assembled from 454 reads that mapped to the Newbler contigs using Sequencher 5.3 and Geneious R8 and several mapping rounds to verify and extent the Newbler assembly for these two contigs. *In silico* read walking from one flanking sequence to the other site across the repetitive elements failed due to read length shortage (repetitive elements were longer than 454 reads) and additional repetitive elements within the genetic content between the repetitive elements.

To clarify the physical sequence order around the triple repetitive element with a second repetitive element within the region, combinatory PCR reactions were performed with primers developed with Primer3 v.4.0.0 and located on the flanking sites of repetitive elements. PCR conditions were 4 min at 94°C, 41 cycles of 94°C for 1 min, 55°C for 1 min and 72°C for 4 min, and finally 72°C for 10 min. Amplicons were using the PCR Purification Kit or from agarose gels using the Gel Extraction Kit (both Quiagen, Hilden, Germany). Amplicons were sequenced using 60 cycles at 96°C for 10 sec, 58°C for 5 sec with ramping of 1°C per second and 60°C for 4 min. Acquired sequences were edited by Finch TV version 1.4.0. The integration of the genetic content between the repetitive elements was manually assembled into the large contig on the basis of *in silico* and *in vitro* results, thus providing a closed genome of OP3 LiM.

Confirmation of the OP3 LiM genome using three metagenome.

Size-fractionated cells as present in Pe1, Pe2 and Pe3 were used to obtain three metagenomes using Illumina HiSeq technology. The sequencing yielded for Pe1 42,649,735, for Pe2 21,648,241 and for Pe3 10,904,818 paired-end reads (Tab. S4). FastQC provided a quality-control. Processing used dynamical trimming with SolexaQA v.3.1.4 and normalization with Khmer 1.4.1. Low quality regions of these reads were thoroughly cleaned using Trimmomatic version 0.32 (59) using paired end mode as well as single end mode. Trimming included cutting of adapters, removal of low quality bases based on a threshold and dropping of short reads. Specified number of bases was then removed from the start and the end of the reads, guided by FastQC visualizations. The processing gave 8,222,350 paired-end reads and 1,160,518 single reads of Pe1, 3,674,326 paired-end reads and 719,609 single reads of Pe2, and 1,610,392 paired-end reads and 383,400 single reads of Pe3 (Tab. S5). Assembly used SPAdes-3.8.0 (56) with the option “metaspades.py”. A *de novo* assembly of all three metagenomes was also performed using script “spades.py”, because the option “metaspased.py” allowed only a single library with paired-end reads (Tab. S6).

The OP3 LiM genome was verified by mapping processed reads onto the OP3 LiM genome in Geneious R9 and visual inspection. OP3 LiM bins of the Pe1, Pe2, and Pe3 assemblies as well as the combined dataset Pe123 were obtained by contig sorting in Metawatt-3.5.2. These bins were compared with the genome using dnadiff from MUMmer v3.23 (60) as well as mapping to the OP3 LiM genome.

Pilon (61) was used to automatically improve the OP3 LiM genome. Read sets from all sequencing platforms were aligned individually as well as a pool of HiSeq data of Pe1, Pe2, and Pe3 to the OP3 LiM genome using the BWA-MEM algorithm in Sequencher 5.3. The generated SAM files were converted into BAM files, sorted in coordinate order and indexed using SAMtools (version 0.1.19) (62). Pilon analyses were performed with the OP3 LiM sequence as input genome and the arguments “change”, “vcf”, “tracks”, “variant”, “fix” and “all duplicates”. The analyses results were analyzed manually and the OP3 LiM genome was manually edited to match the HiSeq reads from the Pe1, Pe2, and Pe3 metagenomes.

Genome annotation.

The OP3 LiM genome was annotated using several pipelines and manual annotations. The NCBI Prokaryotic Genome Annotation (63) was refined using JCoast (64) and Geneious with results of an in-house annotation based on GenDB (65), results of the RAST project with

genome ID 6666666.225137 (66) and online resources of NCBI. CRISPRs structures were searched using CRISPRFinder (67). A closer view to the composition of the nucleotides was given by using Artemis Release 16.0.0 (68). RNAmmer-1.2 was used to predict 5S, 16S, and 23S rRNA genes in the genome sequence (69). The number of tRNA genes was identified using ARAGON v1.2.38 (70).

Linearization at *ori*.

The OP3 LiM genome was linearized at the origin of replication (*ori*), with the start codon of the chromosomal replication initiator protein DnaA gene as base 1. The GC-skew (71–72) and the pattern of ORF-orientation supported this decision, as analyzed with GenSkew (http://genskew.csb.univie.ac.at/) and ORF prediction programs. However, the genome contains a homopolymer of 21 guanines as potential telomeres.

Metatranscriptomes.

RNA sequencing with Illumina HiSeq technology yielded for Pe1 12,998,858 total RNA reads and 57,878,109 rRNA-depleted RNA reads, for Pe2 7,063,485 total RNA reads and 56,879,058 rRNA-depleted RNA reads, and for Pe3 10,025,317 total RNA reads of sequencing project 1906 and for Pe1 36,052,513 and 14,307,025 total RNA reads and 60,205,817 rRNA-depleted RNA reads of sequencing project 1843 (Tab. S4). All RNA reads of project 1906 were merged providing 144,844,827 raw reads. Guided by FastQC visualizations, these reads were processed with BBDuk of the BBmap package (version 32.27) for adapter trimming, quality trimming and contaminant filtering. Next, the reads were trimmed with SolexaQA v.3.1.4 and Trimmomatic version 0.32 including options to cut adapters, to remove low quality bases below a threshold and to drop short reads in a first round and with options for removal of a specified number of bases from the start and the end of the reads in a second round. 134,406,170 high-quality RNA reads were obtained and used for mapping to the OP3 LiM genome. The RNA seq reads of project 1843 were trimmed with SolexaQA v.3.1.4 (Tab. S5).

Construction of reference database for the OP3 LiM proteome.

Two metagenomes from the OP3 LiM-enriched cells were obtained, a Newbler assembly on 454 reads and a SPAdes assembly on MiSeq reads. After removal of contigs mapping to the OP3 LiM genome, all contigs were *de novo* assembled in Geneious R9.

Contigs with a nucleotide identity of over 96.2% were manually inspected and duplicate contigs from the combined assemblies were removed. After addition of the OP3 LiM genome to the metagenome, an in-house GenDB database project was built. JCoast was used to create a faa-file of 92,294 coding sequences for the identification of proteins in the metaproteome.

Metaproteomic analysis.

Cells obtained by differential centrifugation (Pe1, Pe2, and Pe3) were extracted and the protein content was determined by photometric measurement of the absorption at 595 nm using Nanoquant solution. To increase the protein concentration, necessary for direct loading on a gel, samples of Pe1 and Pe2 were concentrated to a volume of 18 µl in vacuum centrifuge (Eppendorf Concentrator plus). The concentration of Pe3 cells was increased by centrifugation, a Beckmann ultracentrifuge rotor Ti 50.2 was used at 45,000 rpm (184,048x *g*, 100 S) for 55 min. Proteins of the samples were then separated by size in an SDS-PAGE (Fig. S22). Each gel lane was cut into 10 equal pieces, which proteins were digested with trypsin. Peptides were eluted in an ultrasonic bath for 15 min, concentrated and finally purified with ZipTips with C18 resin (Millipore, Billerica, Mass). The peptide mix was separated on a Nano HPLC (Easy-nLCII HPLC system, Thermo Fisher Scientific, Dreieich, Germany) and analyzed by MS/MS in an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific) (73). For protein identification, tandem mass spectra were extracted using the Sorcerer™-SEQUEST® platform (Sage-N Research, CA; version 3.5) searching the MS/MS data against a metaproteome database containing the aforementioned ORFs and common laboratory contaminants. Search parameters were parent ion tolerance of 10 ppm, fragment ion mass tolerance of 1.00 Da, and oxidation of methionine (15.99 Da) as variable modification (max. three modifications per peptide). MS/MS-based peptide and protein identifications were validated with Scaffold V4.4.8 (www.proteomesoftware.com). Peptide false discovery rates (FDRs) were set to 1%, and protein FDRs were set to 5% throughout all experiments. Quantification of each protein was considered as total spectral counts (TSCs) for each protein, because the molecular weight of ORFs varied largely. A normalization incorporating the molecular weight to yield a relative molecule number used the normalized spectral abundance factors (NSAF) (74) that were normalized by all TSC assigned to OP3 LiM.

Acknowledgments

We thank Daniela Tienken, Swantje Lilienthal, Dimitri Meier, Christian Quast and Andreas Ellrott for technical assistance. We also thank Rudolf Amann and Marcel Kuypers for access to imaging facilities. This study was funded by the Max Planck Society.

References

1. Bada JL. 2013. New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. Chem Soc Rev 42:2186–2196.
2. Robertson MP, Joyce GF. 2012. The origins of the RNA world. Cold Spring Harb Perspect Biol 4:a003608.
3. Martin WF, Sousa FL, Lane N. 2014. Evolution. Energy at life's origin. Science 344:1092–1093.
4. Bengtson S. 2002. Origins and early evolution of predation. The Paleontological Society Papers 8:289–318.
5. Jurkevitch E. 2007. A brief history of short bacteria: a chronicle of *Bdellovibrio* (and like organisms) research, p 1–9. *In* Jurkevitch E (ed), Predatory prokaryotes, Springer, Berlin Heidelberg.
6. Pérez J, Moraleda-Muñoz A, Marcos-Torres FJ, Muñoz-Dorado J. 2016. Bacterial predation: 75 years and counting! Environ Microbiol 18:766–779.
7. Jurkevitch E. 2007. Predatory behaviors in bacteria–diversity and transitions. Microbe 2:67–73.
8. Soo RM, Woodcroft BJ, Parks DH, Tyson GW, Hugenholtz P. 2015. Back from the dead; the curious tale of the predatory cyanobacterium *Vampirovibrio chlorellavorus*. PeerJ 3:e968.
9. Tudor JJ, McCann MP. 2007. Genomic analysis and molecular biology of predatory prokaryotes, p 153–189. *In* Jurkevitch E (ed), Predatory prokaryotes. Springer, Berlin Heidelberg.
10. Jurkevitch E, Davidov Y. 2006. Phylogenetic diversity and evolution of predatory prokaryotes. p 11–56. *In* Jurkevitch E (ed), Predatory prokaryotes. Springer, Berlin Heidelberg.
11. Mahmoud KK, Koval SF. 2010. Characterization of type IV pili in the life cycle of the predator bacterium *Bdellovibrio*. Microbiology 156:1040–1051.
12. Wang Z, Kadouri DE, Wu M. 2011. Genomic insights into an obligate epibiotic bacterial predator: *Micavibrio aeruginosavorus* ARL-13. BMC Genomics 12:453.
13. Koval SF, Hynes SH, Flannagan RS, Pasternak Z, Davidov Y, Jurkevitch E. 2013. *Bdellovibrio exovorus* sp. nov., a novel predator of *Caulobacter crescentus*. Int J Syst Evol Microbiol 63:146–151.

14. Ganuza E, Sellers CE, Bennett BW, Lyons EM, Carney LT. 2016. A novel treatment protects *Chlorella* at commercial scale from the predatory bacterium *Vampirovibrio chlorellavorus*. Front Microbiol 7:848.
15. Pasternak Z, Njagi M, Shani Y, Chanyi R, Rotem O, Lurie-Weinberger MN, Koval S, Pietrokovski S, Gophna U, Jurkevitch E. 2014. In and out: an analysis of epibiotic vs periplasmic bacterial predators. ISME J 8:625–635.
16. Hazes B, Frost L. 2008. Towards a systems biology approach to study type II/IV secretion systems. Biochim Biophys Acta1778:1839–1850.
17. Guerrero R, Pedros-Alio C, Esteve I, Mas J, Chase D, Margulis L. 1986. Predatory prokaryotes: predation and primary consumption evolved in bacteria. Proc Natl Acad Sci U S A 83:2138–2142.
18. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA,. Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437.
19. Rotaru AE, Schauer R, Probian C, Mussmann M, Harder J. 2012. Visualization of candidate division OP3 cocci in limonene-degrading methanogenic cultures. J Microbiol Biotechnol 22:457–461.
20. Fuchs BM, Glöckner FO, Wulf J, Amann R. 2000. Unlabeled helper oligonucleotides increase the in situ accessibility to 16S rRNA of fluorescently labeled oligonucleotide probes. Appl Environ Microbiol 66:3603–3607.
21. Raskin L, Stromley JM, Rittmann BE, Stahl DA. 1994. Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. Appl Environ Microbiol 60:1232–1240.
22. Rotaru AE. 2009. Anaerobic degradation of limonene and p-xylene in freshwater enrichment cultures. PhD thesis, Universität Bremen, Germany.
23. Kubota K, Imachi H, Kawakami S, Nakamura K, Harada H, Ohashi A. 2008. Evaluation of enzymatic cell treatments for application of CARD-FISH to methanogens. J Microbiol Methods 72:54–59.
24. Wolański M, Donczew R, Zawilak-Pawlik A, Zakrzewska-Czerwińska J. 2015. oriC-encoded instructions for the initiation of bacterial chromosome replication. Front Microbiol 5:735.

25. Frank AC, Lobry JR. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238:65–77.
26. Vasu K, Nagaraja V. 2013. Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol Mol Biol Rev 77:53–72.
27. Horvath P, Barrangou R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. Science 327:167–170.
28. Runyon, G.T., Bear, D.G., Lohman, T.M. 1990. *Escherichia coli* helicase II (UvrD) protein initiates DNA unwinding at nicks and blunt ends. Proc Natl Acad Sci U S A 87:6383–6387.
29. Morrison JM1, Baker KD, Zamor RM, Nikolai S, Elshahed MS, Youssef NH. 2017. Spatiotemporal analysis of microbial community dynamics during seasonal stratification events in a freshwater lake (Grand Lake, OK, USA). PLoS One 12:e0177488.
30. Fontes CM, Gilbert HJ. 2010. Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. Annu Rev Biochem 79:655–681.
31. Mao X, Zhang H, Yin Y, Xu Y. 2012. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. Nucleic Acids Res 40:8210–8218.
32. Chen X, Zhang J. 2013. Why are genes encoded on the lagging strand of the bacterial genome? Genome Biol Evol 5:2436–2439.
33. Haugen P, Bhattacharya D, Palmer JD, Turner S, Lewis LA, Pryer KM. 2007. Cyanobacterial ribosomal RNA genes with multiple, endonuclease-encoding group I introns. BMC Evol Biol 7:159.
34. Nesbø CL, Doolittle WF. 2003. Active self-splicing group I introns in 23S rRNA genes of hyperthermophilic bacteria, derived from introns in eukaryotic organelles. Proc Natl Acad Sci U S A 100:10806–10811.
35. Raghavan R, Miller SR, Hicks LD, Minnick MF. 2007. The unusual 23S rRNA gene of *Coxiella burnetii*: two self-splicing group I introns flank a 34-base-pair exon, and one element lacks the canonical omegaG. J Bacteriol 189:6572–6579.
36. Flood BE, Fliss P, Jones DS, Dick GJ, Jain S, Kaster AK, Winkel M, Mußmann M, Bailey J. 2016. Single-cell (meta-)genomics of a dimorphic *Candidatus Thiomargarita nelsonii* reveals genomic plasticity. Front Microbiol 7:603.

37. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain *Bacteria*. Nature 523:208–211.
38. Dodsworth JA, Li L, Wei S, Hedlund BP, Leigh JA, de Figueiredo P. 2010. Interdomain conjugal transfer of DNA from bacteria to archaea. Appl Environ Microbiol 76:5644–5647.
39. Bischof LF, Friedrich C, Harms A, Søgaard-Andersen L, van der Does C. 2016. The Type IV Pilus assembly ATPase PilB of *Myxococcus xanthus* interacts with the inner membrane platform protein PilC and the nucleotide-binding protein PilM. J Biol Chem 291:6946–6957.
40. Harder J, Foss S. 1999. Anaerobic formation of the aromatic hydrocarbon p-cymene from monoterpenes by methanogenic enrichment cultures. Geomicrobiol J 16:295–305.
41. Valentine RC, Shapiro BM, Stadtman ER. 1968. Regulation of glutamine synthetase. XII. Electron microscopy of the enzyme from *Escherichia coli*. Biochemistry 7:2143–2152.
42. Martín-Platero AM, Valdivia E, Maqueda M, Martínez-Bueno M. 2007. Fast, convenient, and economical method for isolating genomic DNA from lactic acid bacteria using a modification of the protein "salting-out" procedure. Anal Biochem 366:102–104.
43. Hahnke RL, Harder J. 2013. Phylogenetic diversity of *Flavobacteria* isolated from the North Sea on solid media. Syst Appl Microbiol 36:497–504.
44. Lane DJ. 1991. 16S/23S rRNA sequencing, p 115–175. *In* Stackebrandt E, Goodfellow M (ed), Nucleic acid techniques in bacterial systematics. John Wiley & Sons, New York.
45. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A 82:6955–6959.
46. Pernthaler A, Pernthaler J, Amann R. 2004. Sensitive multi-color fluorescence *in situ* hybridization for the identification of environmental microorganisms, p 711–726. *In* Kowalchuk G, de Bruijn, FJ, Head IM, Akkermans ADL, van Elsas JD (ed), Molecular microbial ecology Manual, 2nd, vol 1. Kluwer Academic Publishers, Dordrecht, Boston, London.

47. Amann RI, Krumholz L, Stahl DA. 1990. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. J Bacteriol 172:762–770.

48. Daims H, Brühl A, Amann R, Schleifer KH, Wagner M. 1999. The domain-specific probe EUB338 is insufficient for the detection of all *Bacteria*: development and evaluation of a more comprehensive probe set. Syst Appl Microbiol 22:434–444.

49. Stahl DA, Amann R. 1991. Development and application of nucleic acid probes in bacterial systematics, p 205–248. *In* Stackebrandt E, Goodfellow M (ed), Nucleic acid techniques in bacterial systematics. John Wiley & Sons Ltd., Chichester, England.

50. Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. 2012. The binning of metagenomic contigs for microbial physiology of mixed cultures. Front Microbiol 3:410.

51. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

52. Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci USA 106:19126–19131.

53. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541.

54. Cox MP, Peterson DA, Biggs, PJ. 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485.

55. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, Charbonneau A, Constantinides B, Edvenson G, Fay S, Fenton J, Fenzl T, Fish J, Garcia-Gutierrez L, Garland P, Gluck J, González I, Guermond S, Guo J, Gupta A, Herr JR, Howe A, Hyer A, Härpfer A, Irber L, Kidd R, Lin D, Lippi J, Mansour T, McA'Nulty P, McDonald E, Mizzi J, Murray KD, Nahum JR, Nanlohy K, Nederbragt

AJ, Ortiz-Zuazaga H, Ory J, Pell J, Pepe-Ranney C, Russ ZN, Schwarz E, Scott C, Seaman J, Sievert S, Simpson J, Skennerton CT, Spencer J, Srinivasan R, Standage D, Stapleton JA, Steinman SR, Stein J, Taylor B, Trimble W, Wiencko HL, Wright M, Wyss B, Zhang Q, Zyme E, Brown CT. 2014. The khmer software package: enabling efficient sequence analysis. F1000Res 4:900.

56. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.
57. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075.
58. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633–4642.
59. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.
60. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12.
61. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963.
62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.
63. Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira CD, Kyrpides N, Madupu R, Markowitz V, Tatusova T, Thomson N, White O. 2008. Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation. OMICS 12:137–141.
64. Richter M, Lombardot T, Kostadinov I, Kottmann R, Duhaime MB, Peplies J, Glöckner FO. 2008. JCoast – a biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes. BMC Bioinformatics 9:177.

65. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Pühler A. 2003. GenDB–an open source genome annotation system for prokaryote genomes. Nucleic Acids Res 31:2187–2195.

66. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75.

67. Grissa I, Vergnaud G, Pourcel C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35:W52–W57.

68. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944–945.

69. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108.

70. Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32:11–16.

71. Bao Q, Tian Y, Li W, Xu Z, Xuan Z, Hu S, Dong W, Yang J, Chen Y, Xue Y, Xu Y, Lai X, Huang L, Dong X, Ma Y, Ling L, Tan H, Chen R, Wang J, Yu J, Yang H. 2002. A complete sequence of the *T. tengcongensis* genome. Genome Res 12:689–700.

72. Necşulea A, Lobry JR. 2007. A new method for assessing the effect of replication on DNA base composition asymmetry. Mol Biol Evol 24:2169–2179.

73. Heinz E, Williams TA, Nakjang S, Noël CJ, Swan DC, Goldberg AV, Harris SR, Weinmaier T, Markert S, Becher D, Bernhardt J, Dagan T, Hacker C, Lucocq JM, Schweder T, Rattei T, Hall N, Hirt RP, Embley TM. 2012. The genome of the obligate intracellular parasite *Trachipleistophora hominis*: new insights into microsporidian genome dynamics and reductive evolution. PLoS Pathog 8:e1002979.

74. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. 2006. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. J Proteome Res 5:2339–2347.

Appendix of Chapter 5:

Supplementary materials

Supplementary figures.

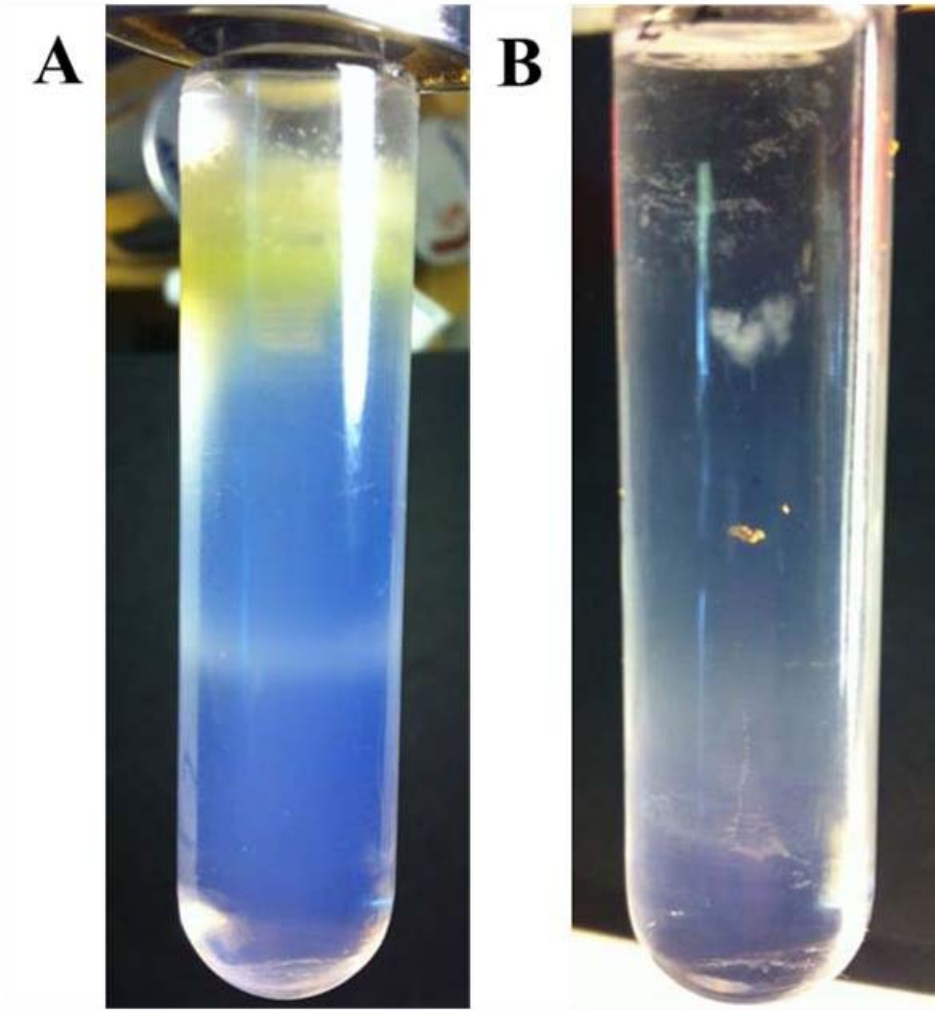


Fig. S1. Distribution of biomass along first and second Percoll density gradients. Concentrated biomass from the limonene methanogenic enrichment culture was separated over a Percoll density gradient by centrifugation (A). OP3 LiM-specific PCR and CARD-FISH detected OP3 LiM cells above the bottom band in the gradient. The OP3 LiM-fractions of ten Percoll gradients were combined and separated in a second Percoll gradient (B).

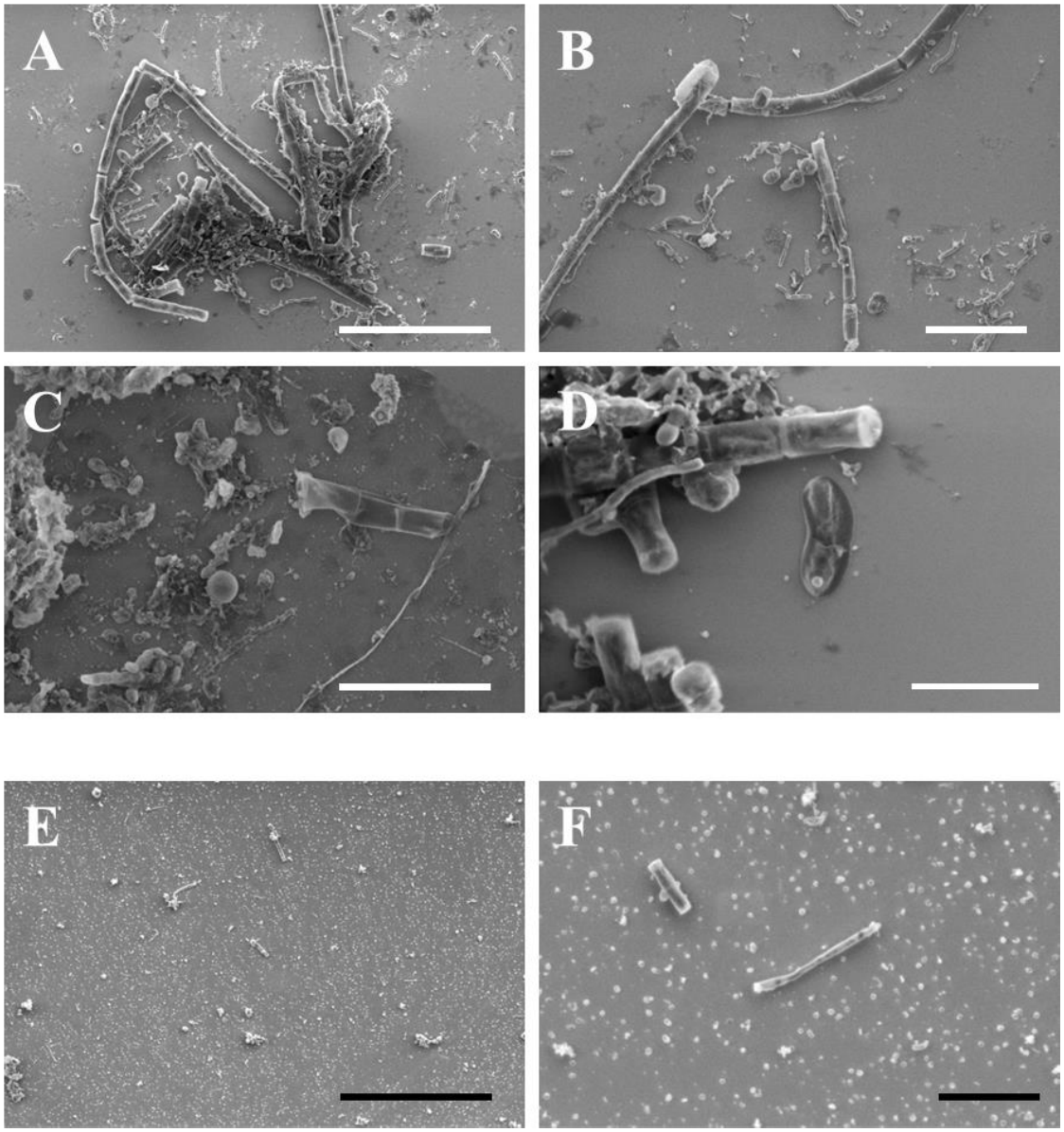


Fig. S2. Scanning electron microscopy (SEM) images of the methanogenic enrichment community from culture MM-375. Cells of the culture (A–D) were separated by differential centrifugation. The supernatant of a 10,000 S centrifugation (E–F) revealed the removal of nearly all aggregates, filaments and large cells. Scale bar: 10 µm (A), 5 µm (B), 4 µm (C), 2 µm (D), 30 µm (E) and 5 µm (F).

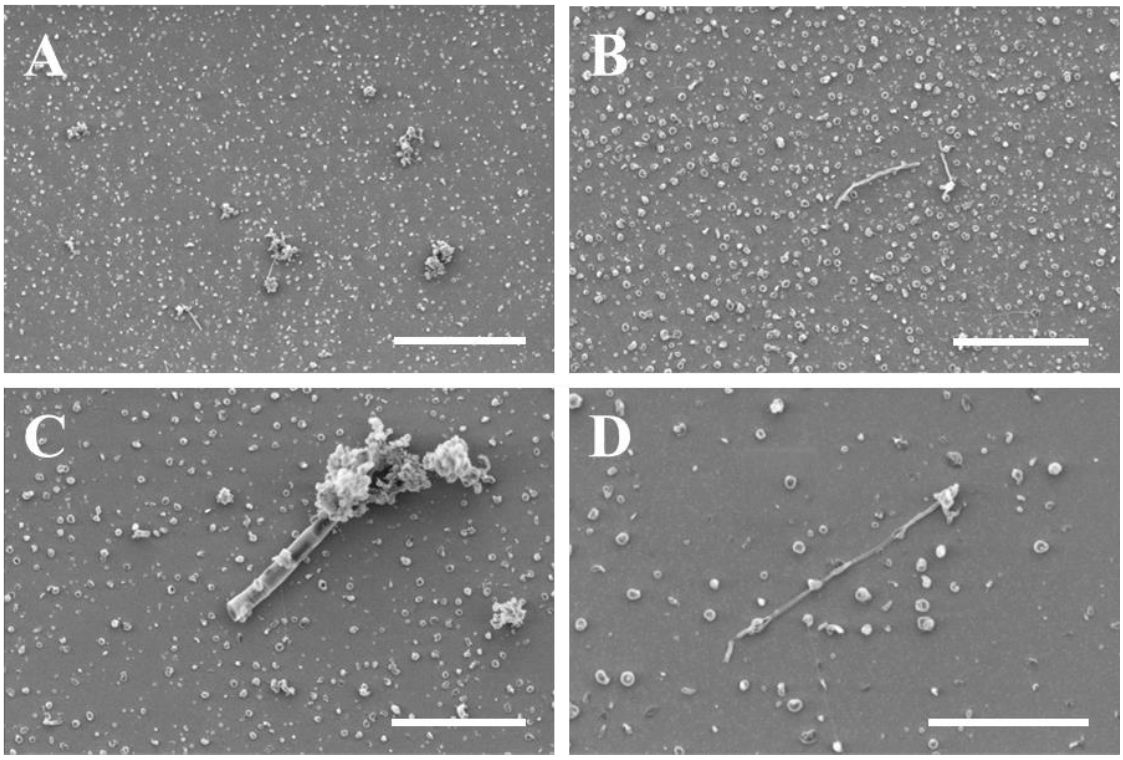


Fig. S3. SEM images of cell suspensions prepared from the 10,000 S–100 S pellet originating from culture MM-375. Scale bar: 10 µm (A), 5 µm (B–C) and 3 µm (D).

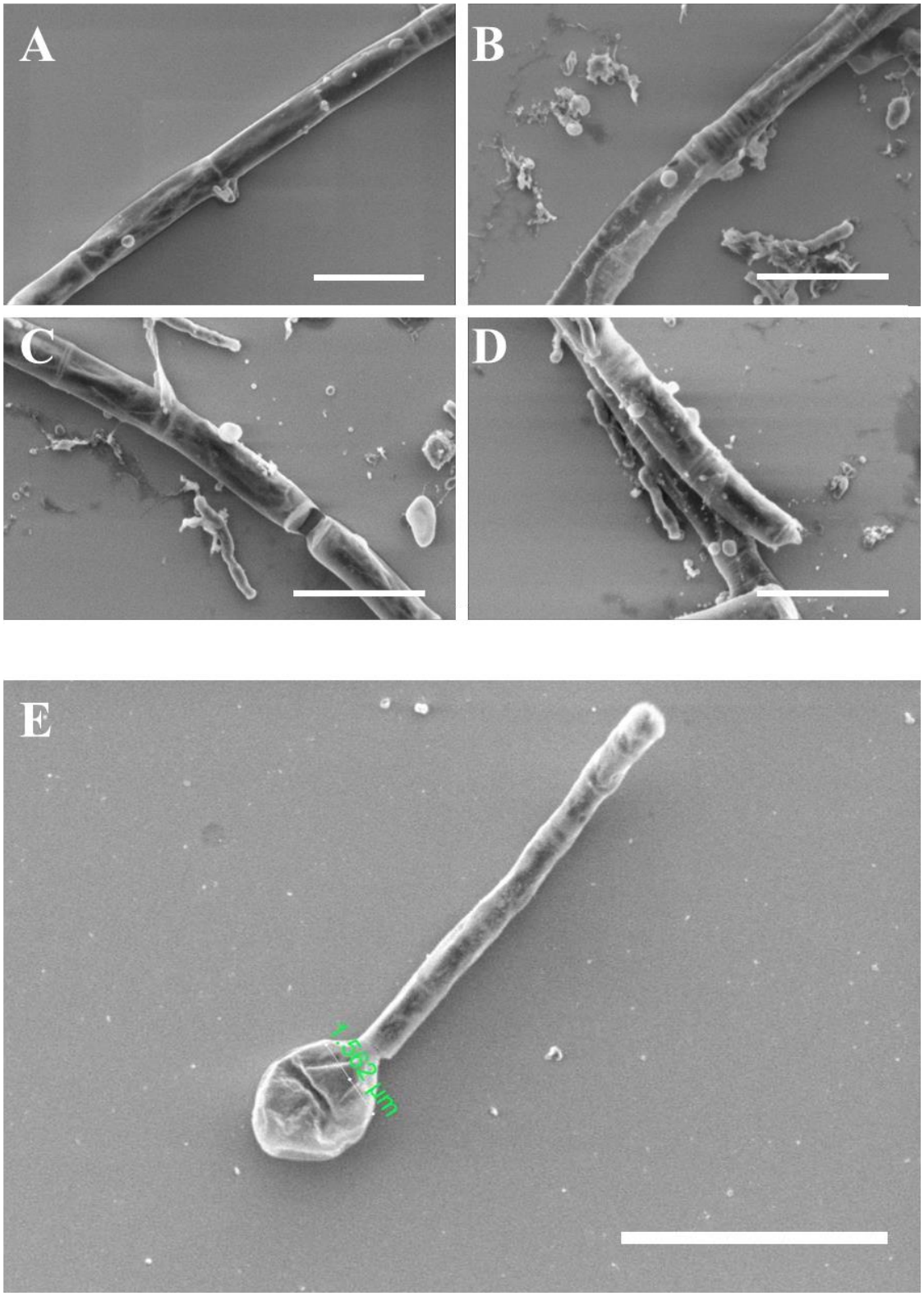


Fig. S4. SEM images of cells affiliated to *Methanosaeta* originating from the limonene methanogenic enrichment culture MM-375 (A–D) and cell suspensions prepared from the 10,000 S–100 S pellet (E). Scale bar: 2 µm (A–D) and 5 µm (E).

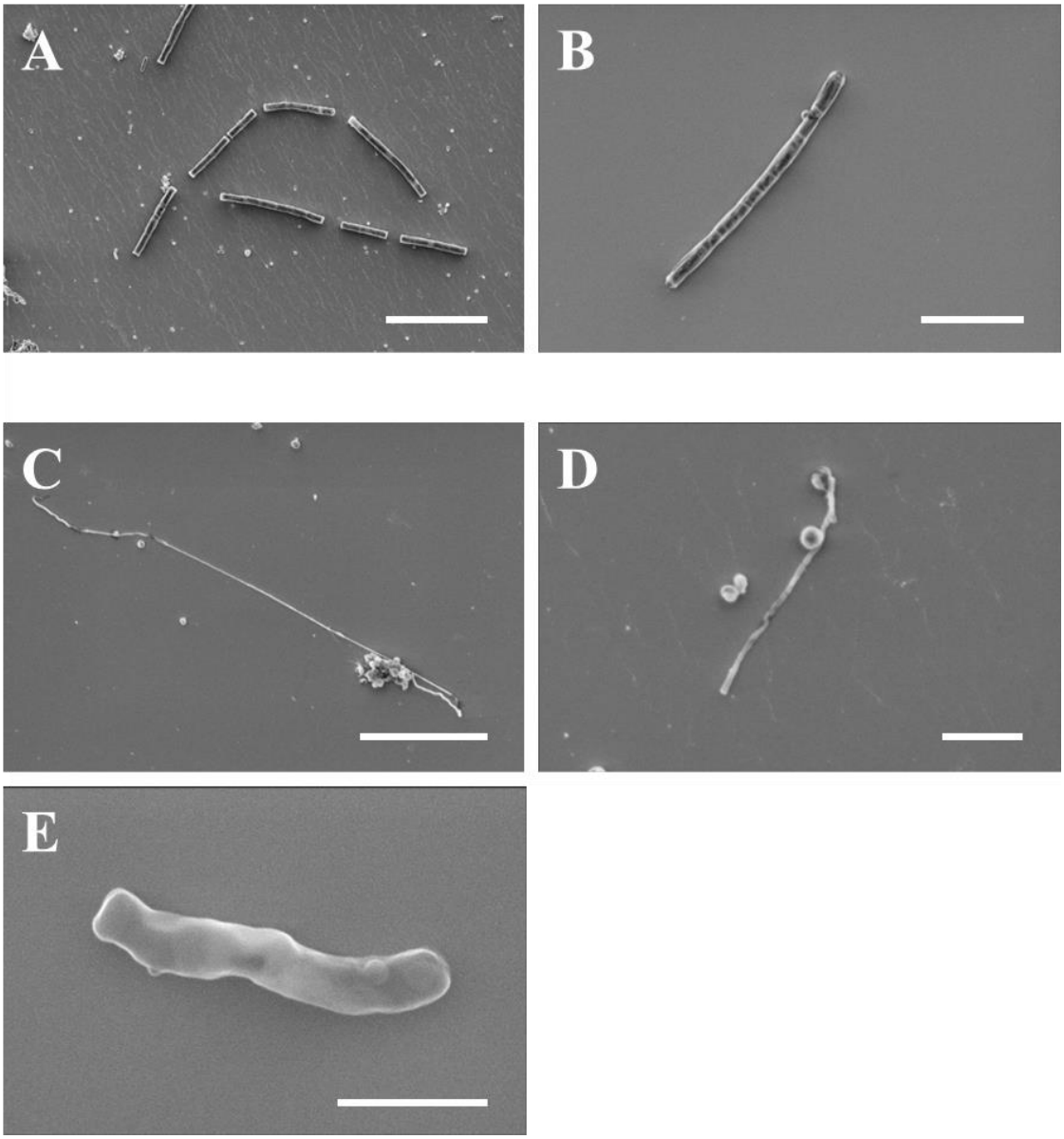


Fig. S5. SEM images of cells originating from the enrichment cultures MM-375 (A, C–E) and MM-212 (B). Scale bar: 10 µm (A), 5 µm (B–C), 1 µm (D) and 500 nm (E).

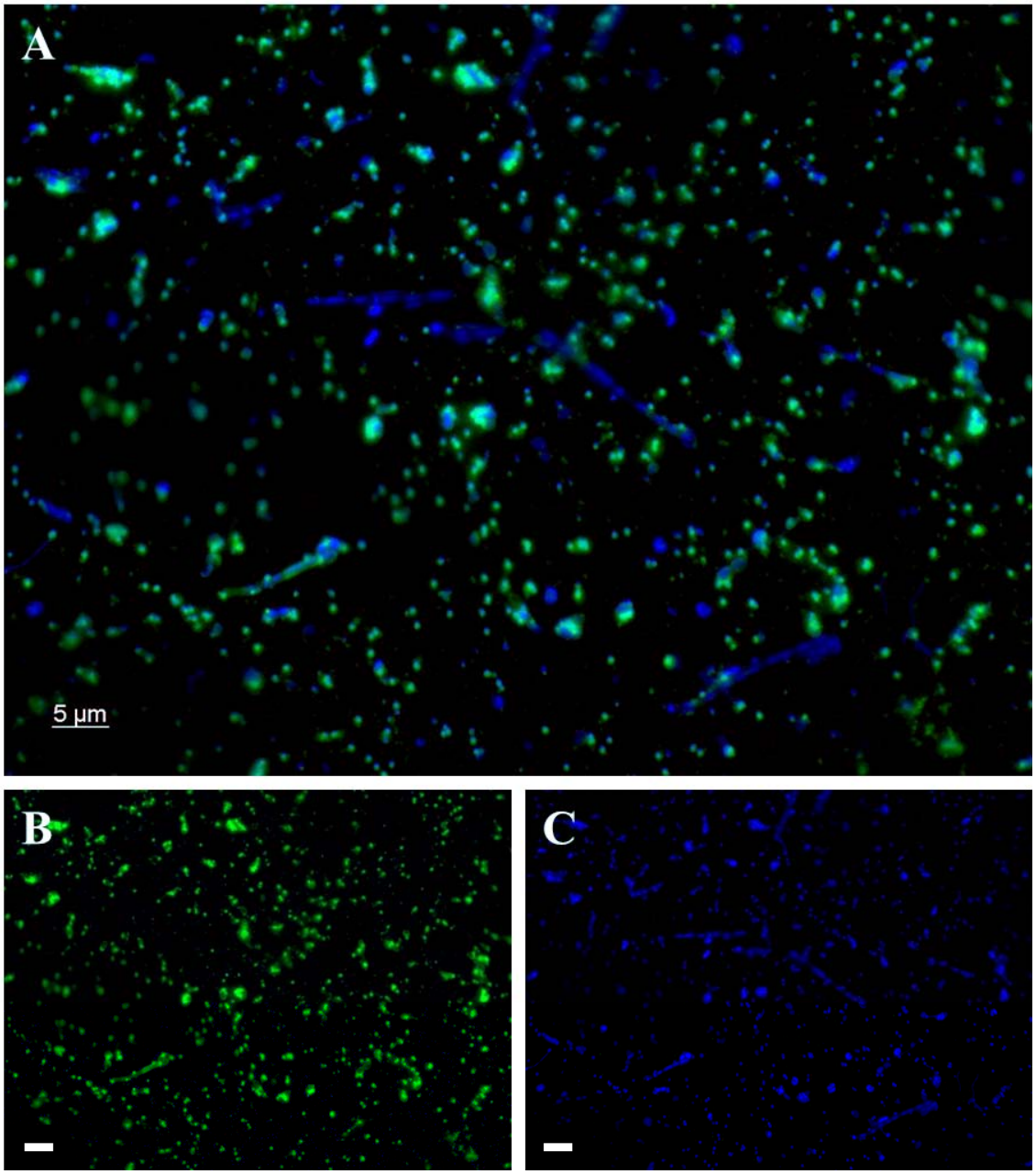


Fig. S6. Detection of OP3 LiM cells by CARD-FISH. Cells from the cultures MM-252 were detected using the probe OP3-565 with helper oligonucleotides (green, A and B) and DAPI staining (blue, A and C). Scale bars, 5 µm.

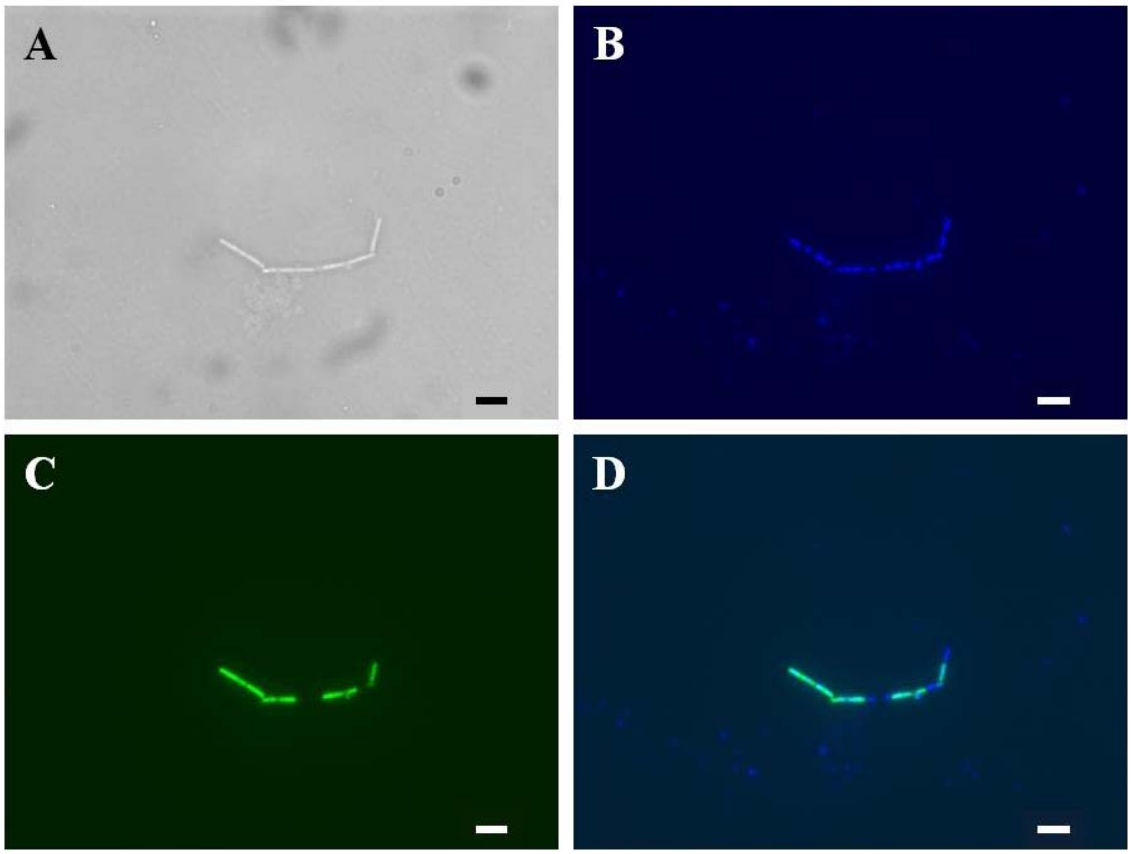


Fig. S7. Micrographs of a filament from culture MM-376. CARD-FISH was performed for the detection of archaeal cells (ARCH-915; Alexa-488) on slides. Lysozyme treatment was used for the permeabilization. (A) Phase-contrast micrograph, (B) DNA detection by DAPI staining, (C) rRNA detection by probe ARCH-915, and (D) an overlay of epifluourescence micrographs. Scale bars, 5 µm.

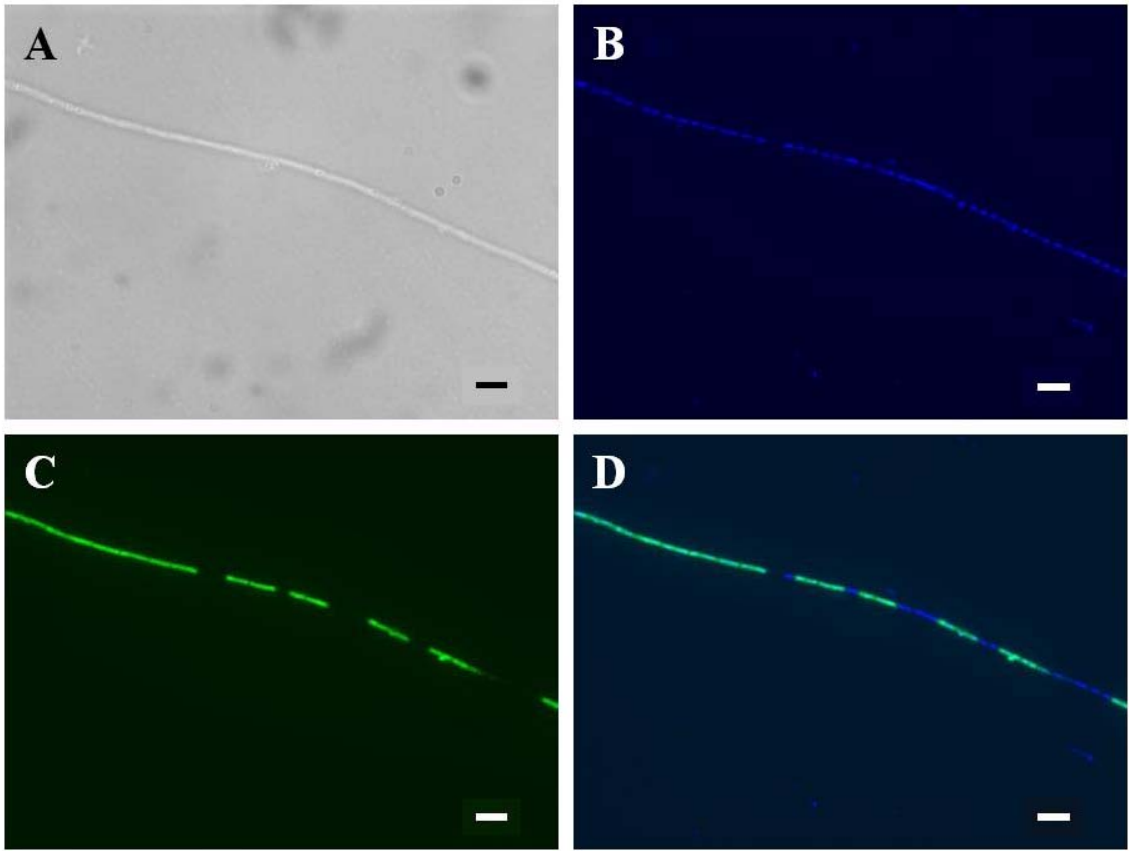


Fig. S8. Micrographs of a filament from culture MM-376. Cells were permeabilized by lysozyme and CARD-FISH was performed for the detection of archaeal cells (ARCH-915; Alexa-488) on slides. (A) Phase-contrast micrograph, (B) DNA detection by DAPI staining, (C) rRNA detection by probe ARCH-915, and (D) an overlay of epifluourescence micrographs. Scale bars, 5 µm.

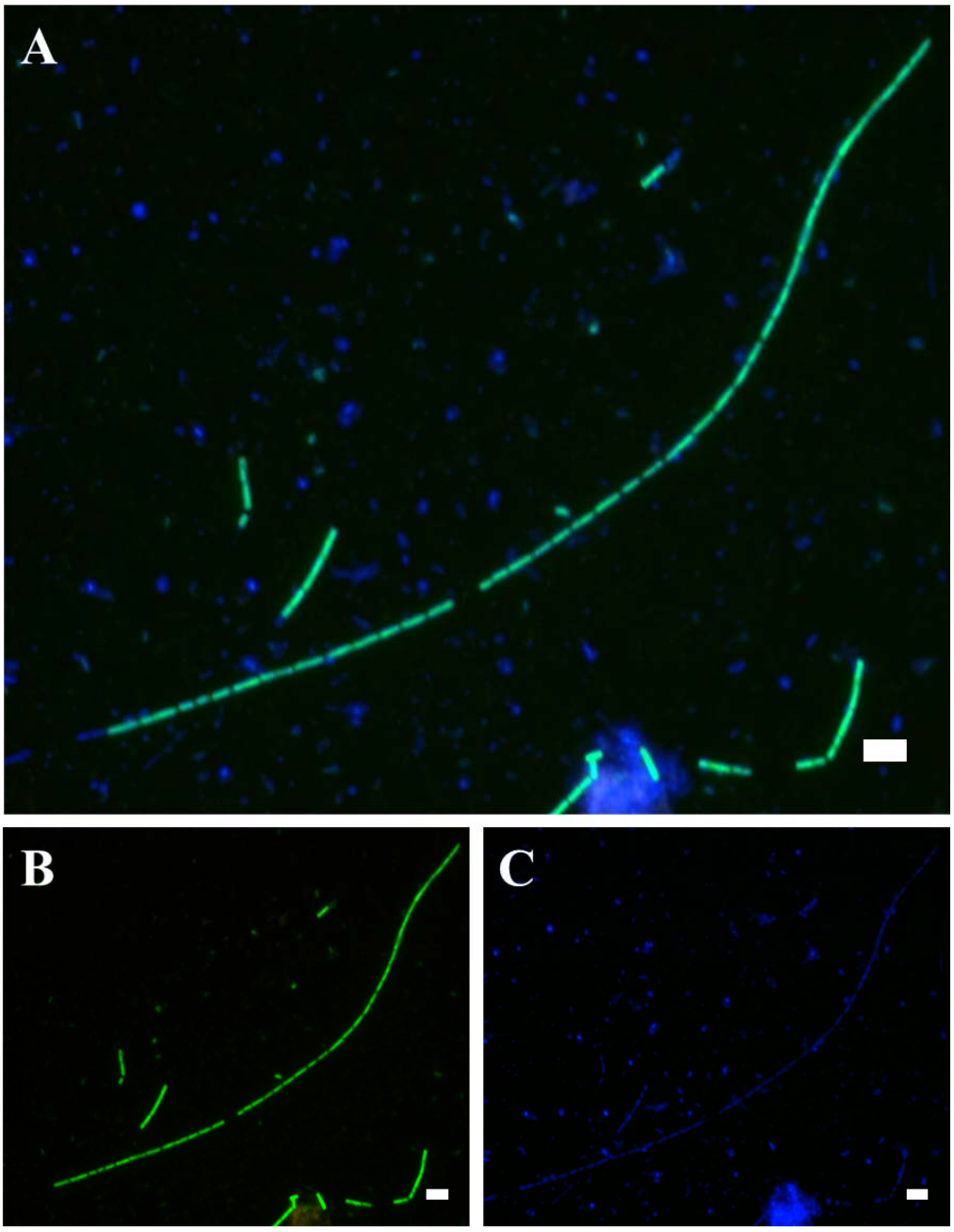


Fig. S9. Epifluorescence micrographs of cells and filaments from culture MM-376. The CARD-FISH protocol (46) for samples on membrane filters was modified by the performance of an ethanol dehydration (50%, 70%, and 96% v/v for 5 min each step), an incubation for 5 min at 37 °C in proteinase K solution (20 µg/ml in 0.1 M Tris-HCl, 0.05 M EDTA, pH 8.0, 0.5 M NaCl), an incubation of 0.1 M HCl for 10 min at RT, and a hybridization time of 2.5 h. Cellular rRNA was detected by CARD-FISH with the horseradish peroxidase (HRP)-labeled ARCH-915 probe (green, A and B). DNA was stained with DAPI (blue, A and C). (A) is an overlay of (B) and (C). Scale bars, 5 µm.

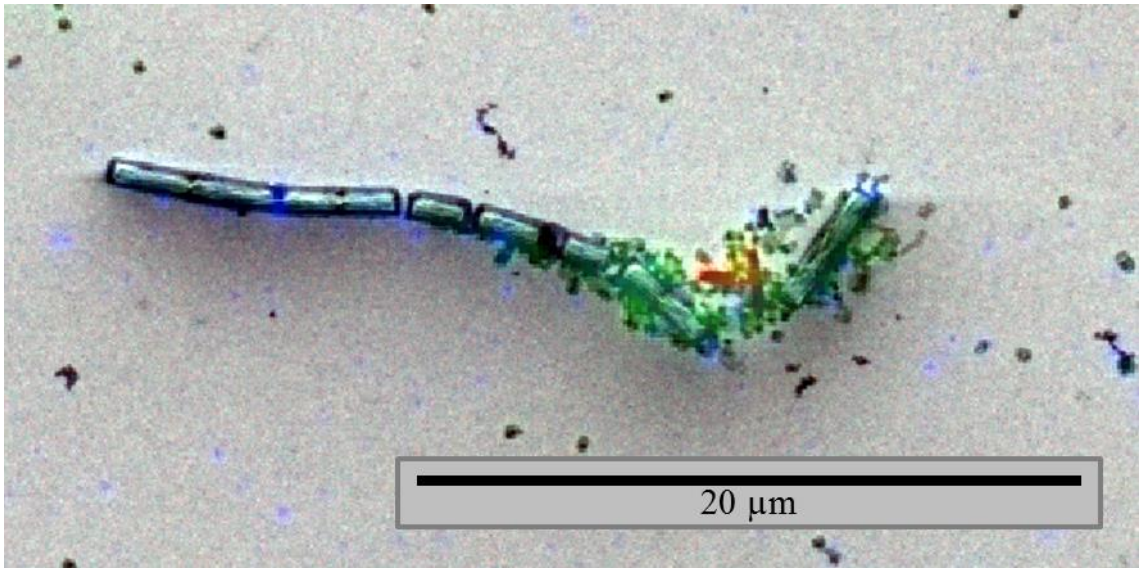


Fig. S10. Correlative light and electron microscopy image of the detection of OP3 LiM and *Archaea*. A Quanta FEG 250 SEM equipped with the DELMIC SECOM platform was used to visualize a double CARD-FISH sample of the limonene methanogenic enrichment culture. OP3 LiM rRNA was detected with OP3-595 (red, Alexa-594 as label on the tyramide). Archaeal cells were detected with ARCH-915 (green; Alexa-488). DNA was stained by DAPI (blue).

Formaldehyde-fixed cells (1.3% v/v for 1 h at RT) of the limonene methanogenic enrichment cultures were centrifuged at 13,000 rpm for 10 min. The pellet was washed twice in 500 µl 1 x phosphate buffered saline (PBS), twice in 200 µl MilliQ water and resuspended in 300 µl MilliQ water. The suspension was diluted with MilliQ water (20% v/v sample) and 80 µl was placed on an indium-tin oxide (ITO)-coated cover glass (DELMIC B.V., JA Delft, The Netherlands). After air drying CARD-FISH was performed.

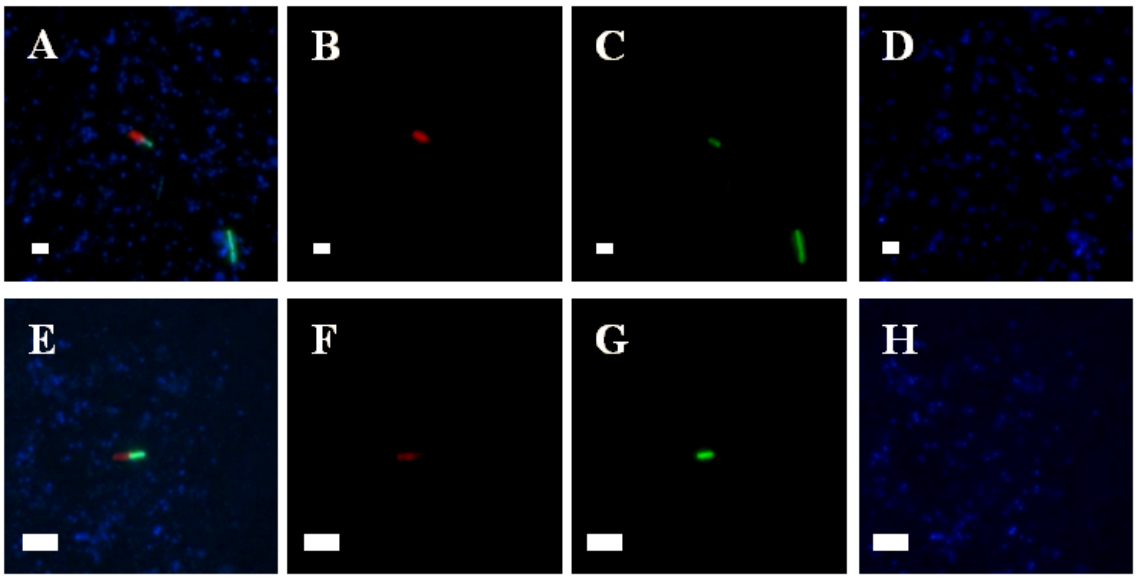


Fig. S11. Detection of intron RNA in double CARD-FISH experiments using the intron specific probe/helper-mix and ARCH-915 on cells from culture MM-376. Overlay of fluorescence micrographs (A, E) and individual graphs reveal the presence of DNA by DAPI (blue, A, D, E, H), of intron RNA by three intron probes hen1-2235, hen2-2309 and hen3-2538 (red, A, B, E, F) and of *Archaea* by probe ARCH-915 (green, A, C, E, G). Scale bars, 2 µm (A–D) and 5 µm (E–H).

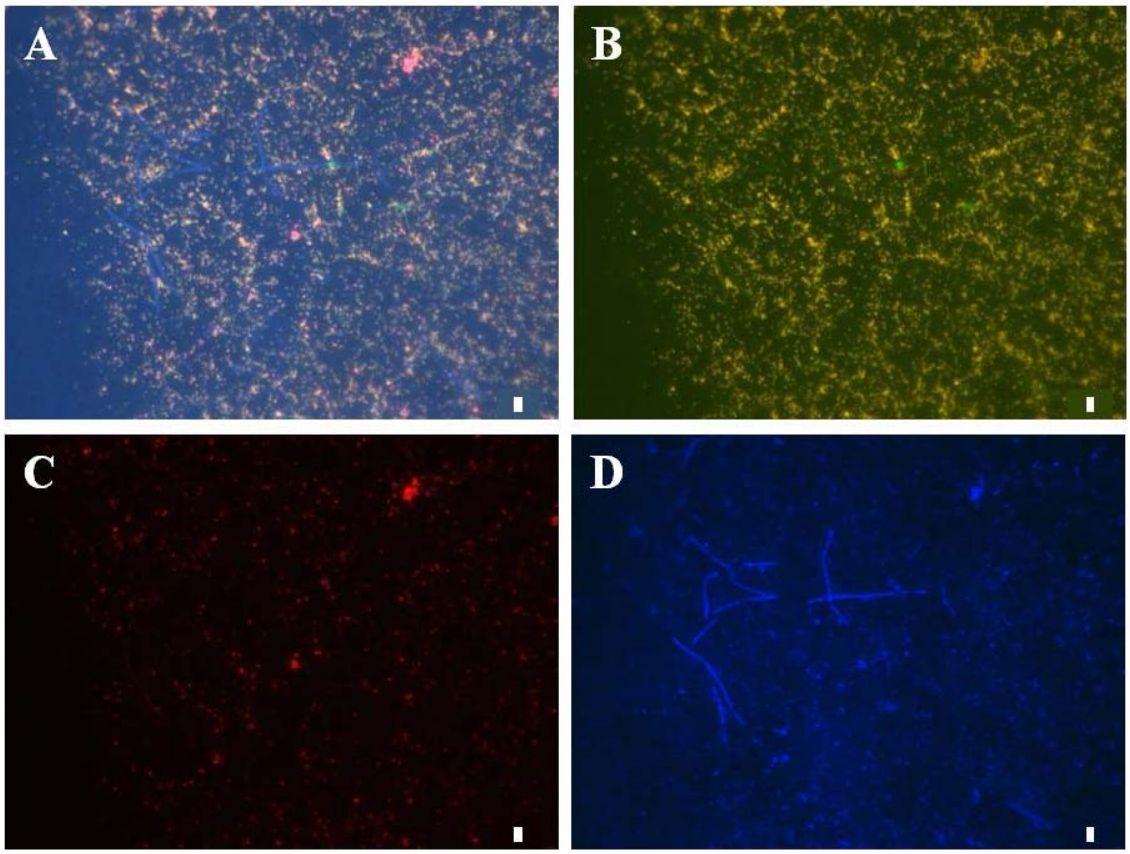


Fig. S12. *In situ* detection of intron RNA and rRNA originating from OP3 LiM by double CARD-FISH and epifluorescence microscopy. RNA was visualized with OP3-565 (green, A and B) and with three intron probes hen1-2235, hen2-2309 and hen3-2538 (red, A and C). All probes were used with flanking helper oligonucleotides. Cells originating from culture MM-376 were stained with DAPI for the presence of DNA (blue, A and D). Scale bars, 2 µm.

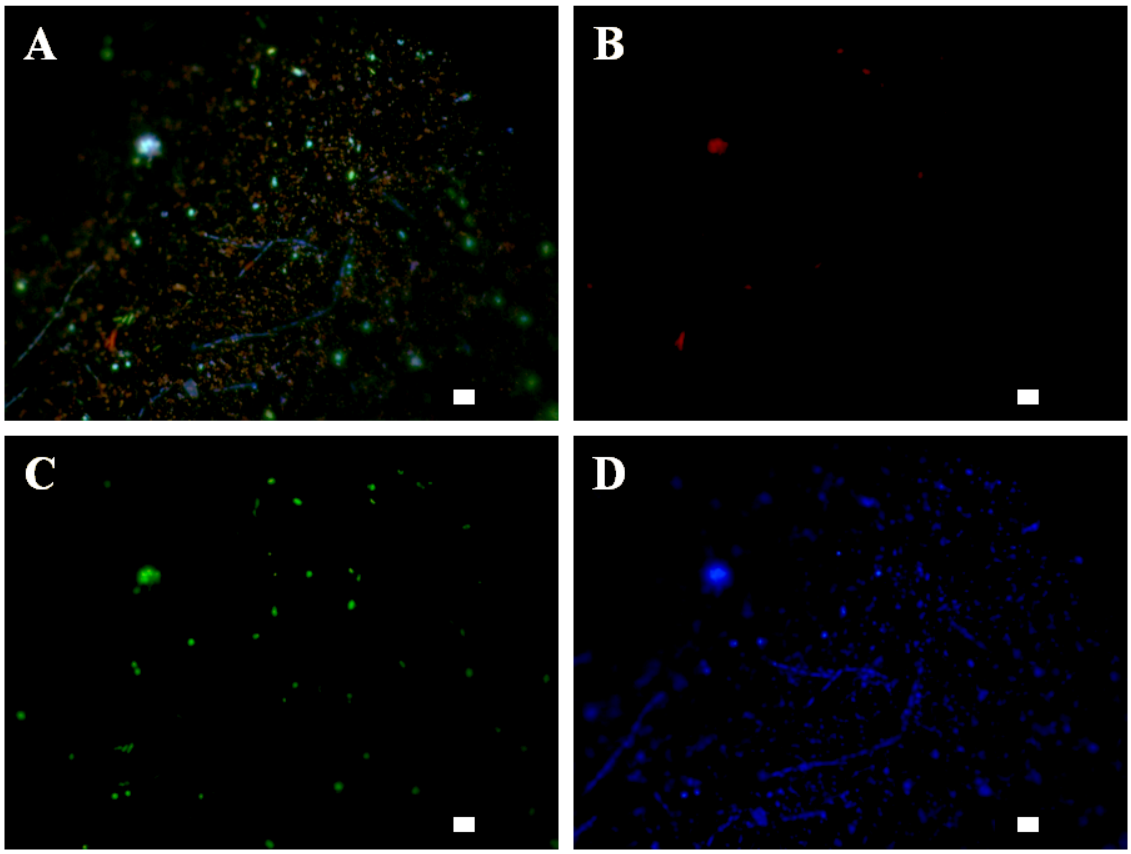


Fig. S13. *In situ* detection of intron RNA and bacterial rRNA by double CARD-FISH and epifluorescence microscopy. Intron RNA was visualized with three intron probes hen1-2235, hen2-2309 and hen3-2538 (red, A and B) with helper oligonucleotides. The probe mix EUB338 I–III (green, A and C) and DAPI (blue, A and D) were applied to detect *Bacteria* and DNA. Scale bars, 5 µm.

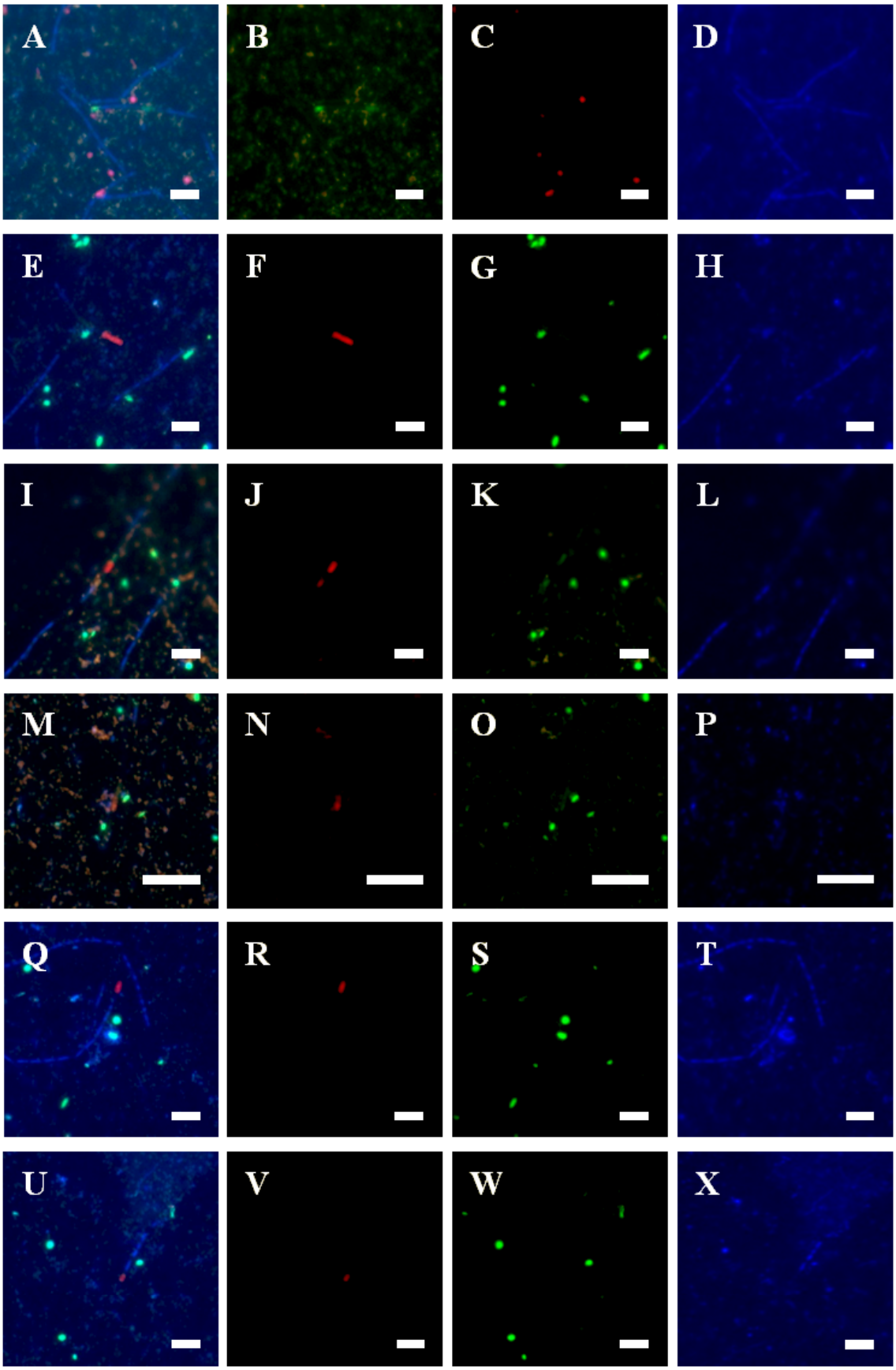

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X

Fig. S14. Detection of intron RNA in double CARD-FISH experiments using the intron specific probe/helper-mix and probe mix EUB I–III on cells from culture MM-376. Overlay of fluorescence micrographs (first column) and individual graphs reveal the presence of DNA by DAPI (blue, first and fourth column), of intron RNA by three intron probes hen1-2235, hen2-2309 and hen3-2538 (first and second column) and of *Bacteria* (first and third column). The detection of intron RNA was possible with Alexa-488 (green, A, B) as well as Alexa-594 (red, E, F, I, J, M, N, Q, R, U, V). Scale bars, 5 µm (A–L & Q–X) and 10 µm (M–P).

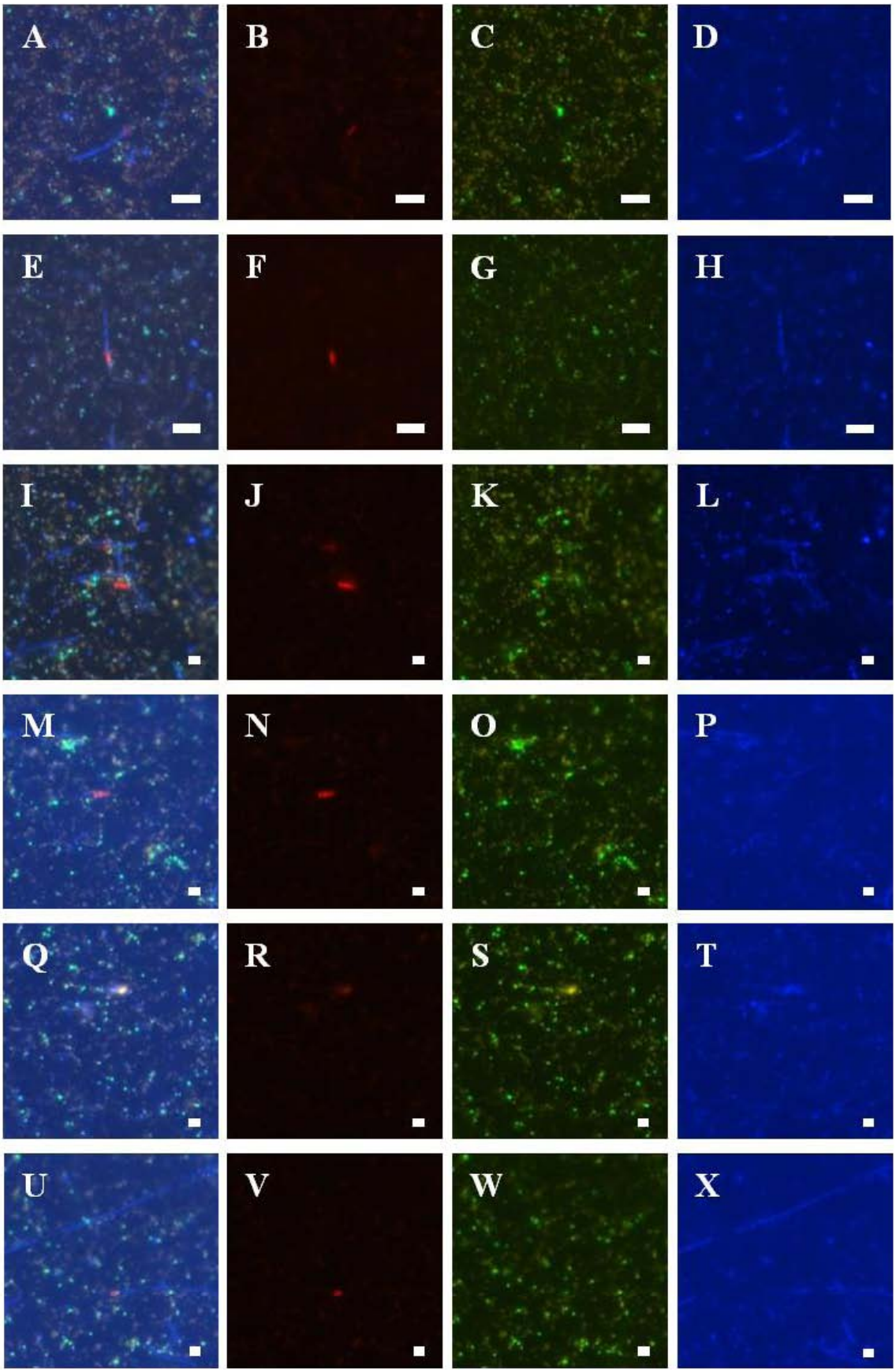

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X

Fig. S15. Gallery of intron RNA CARD-FISH signals in double CARD-FISH experiments using the intron specific probe/helper-mix and probe OP3-565 on cells from culture MM-376. Overlay of fluorescence micrographs (first column) and individual graphs reveal the presence of DNA by DAPI (blue, first and fourth column), of intron RNA by three intron probes hen1-2235, hen2-2309 and hen3-2538 (red, first and second column) and of OP3 LiM (green, first and third column). Scale bars, 5 µm (A-H) and 2 µm (I-X).

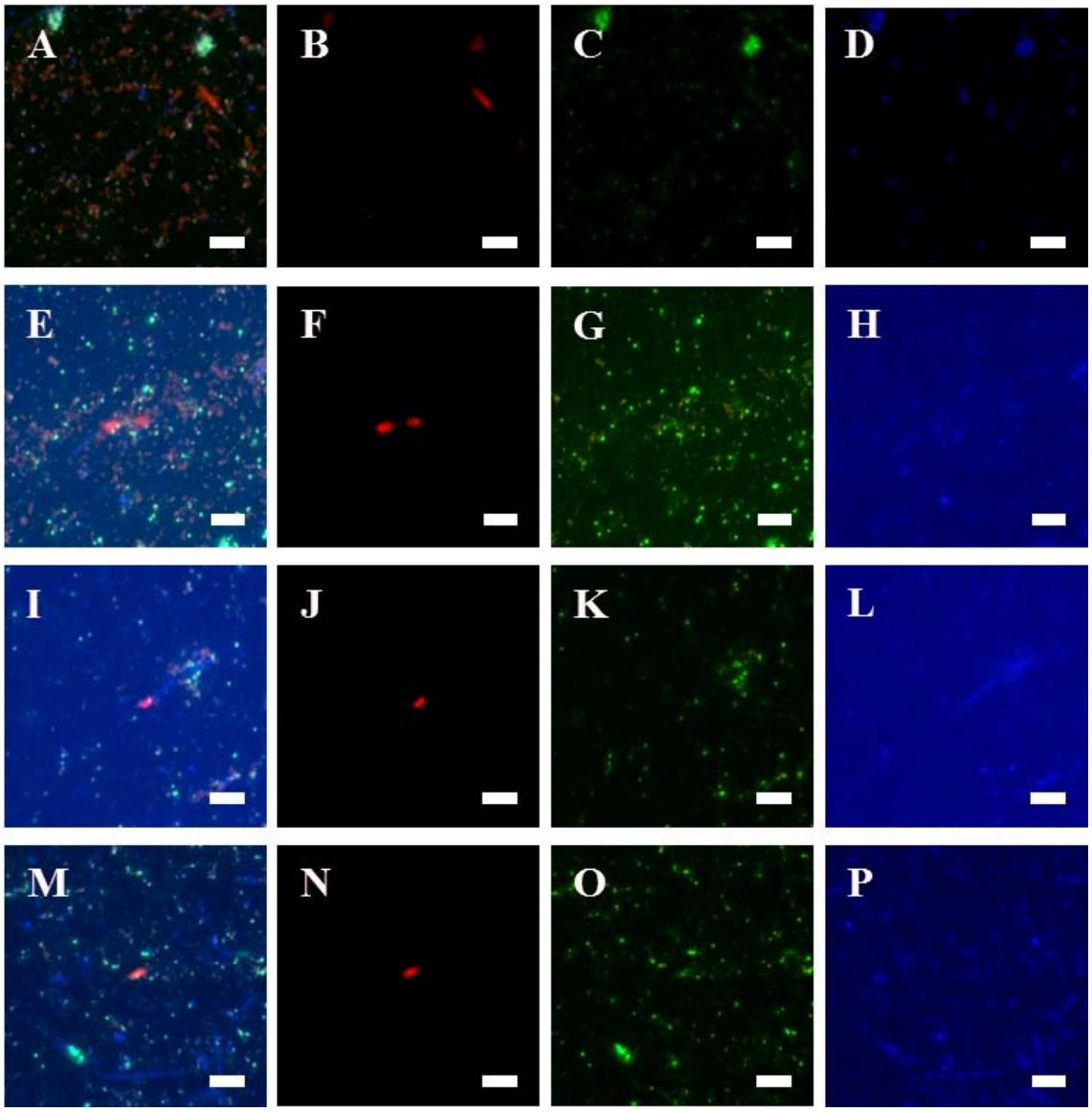


Fig. S16. Continuation of the gallery of intron RNA CARD-FISH signals in double CARD-FISH experiments using the intron specific probe/helper-mix and probe OP3-565 on cells from culture MM-376. Overlay of fluorescence micrographs (first column) and individual graphs reveal the presence of DNA by DAPI (blue, first and fourth column), of intron RNA by three intron probes hen1-2235, hen2-2309 and hen3-2538 (red, first and second column) and of OP3 LiM (green, first and third column). Scale bars, 5 µm.

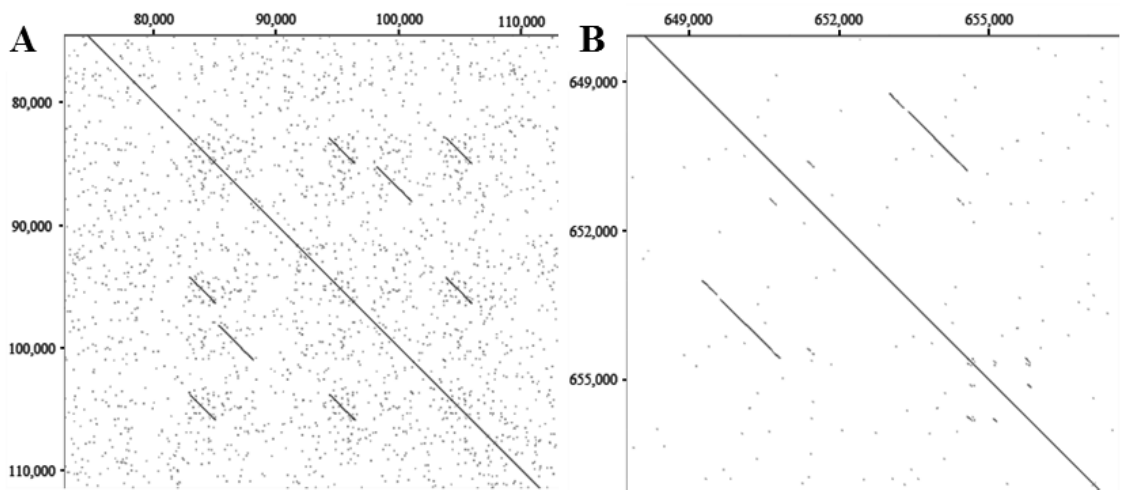


Fig. S17. Dot plot presentations of repetitive regions within the OP3 LiM genome. (A) Interlaced repetitive elements in the ORF for the very large surface protein, a gene of 119,034 bp. This assembly was finally confirmed by performed combinatorial PCR reactions. (B) Duplicated genes of unknown function (CHP1; peg.582 and CHP2; peg.585) located in a region of multiple insertions. Axes provide the position of the repeats within the OP3 LiM genome sequence. Dot plot calculations were performed in Geneious 10.1.3 with “Low sensitivity/Fast” settings, word size of 10, and title size of 100,000.

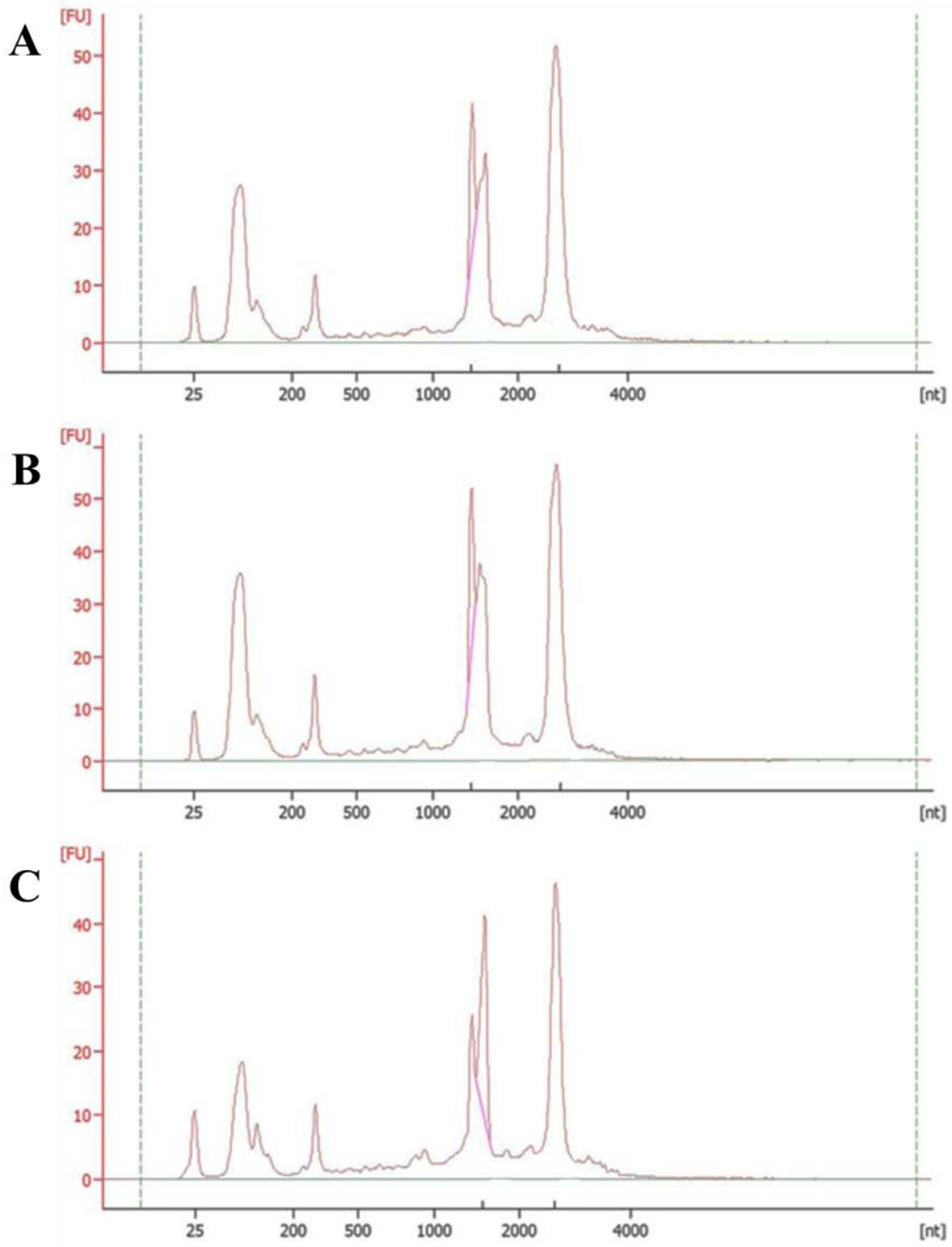


Fig. S18. Size distribution of total RNA preparations. RNAs from different Pe1 pellets used for library no. 1906.A (A) and for library no. 1906.B (B). (C) shows RNA from Pe2 used for library no. 1906.D. The x-axis describes the length of the RNA fragments in nucleotides [nt], the y-axis shows the fluorescence in fluorescence units [FU].

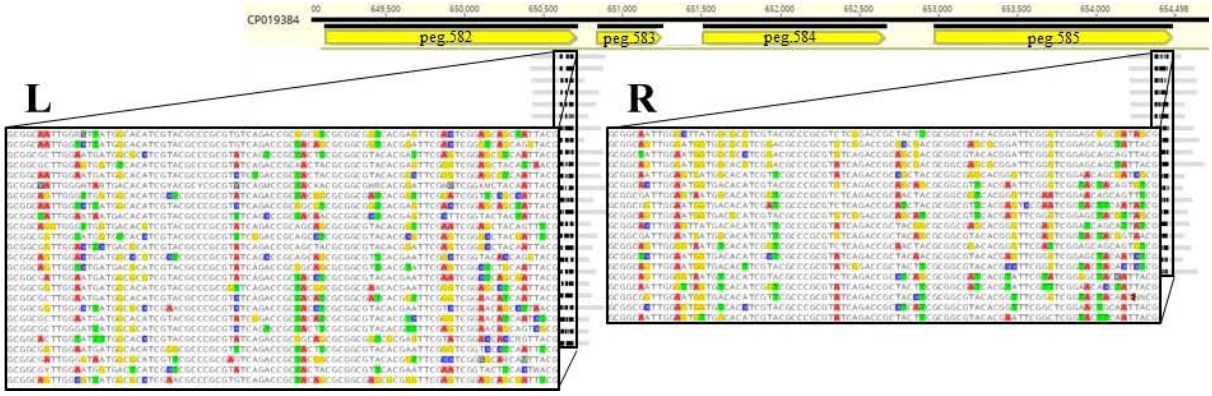


Fig. S19. Representation of high-variable regions of 83 bp length in the duplicated genes CHP1 (peg.582) and CHP2 (peg.585) representing a diversity of 44 sequences with multiple coverage in the metagenomes. These regions were present in 1,012 reads of the combined metagenome Pe1, Pe2 and Pe3. *De novo* assembly of the 1,012 reads resulted in 44 contigs of which 25 mapped to CHP1 (L) and 19 mapped to CHP2 (R). Geneious 10.1.3 were used for mapping (bbmap mapper plugin with “Normal Sensitivity” settings for mapping of metagenome reads onto the genome and Geneious mapper with “High sensitivity/Medium” settings for mapping of contigs onto the genome) and *de novo* assembly of the 1,012 reads (with “High sensitivity/Medium” settings).

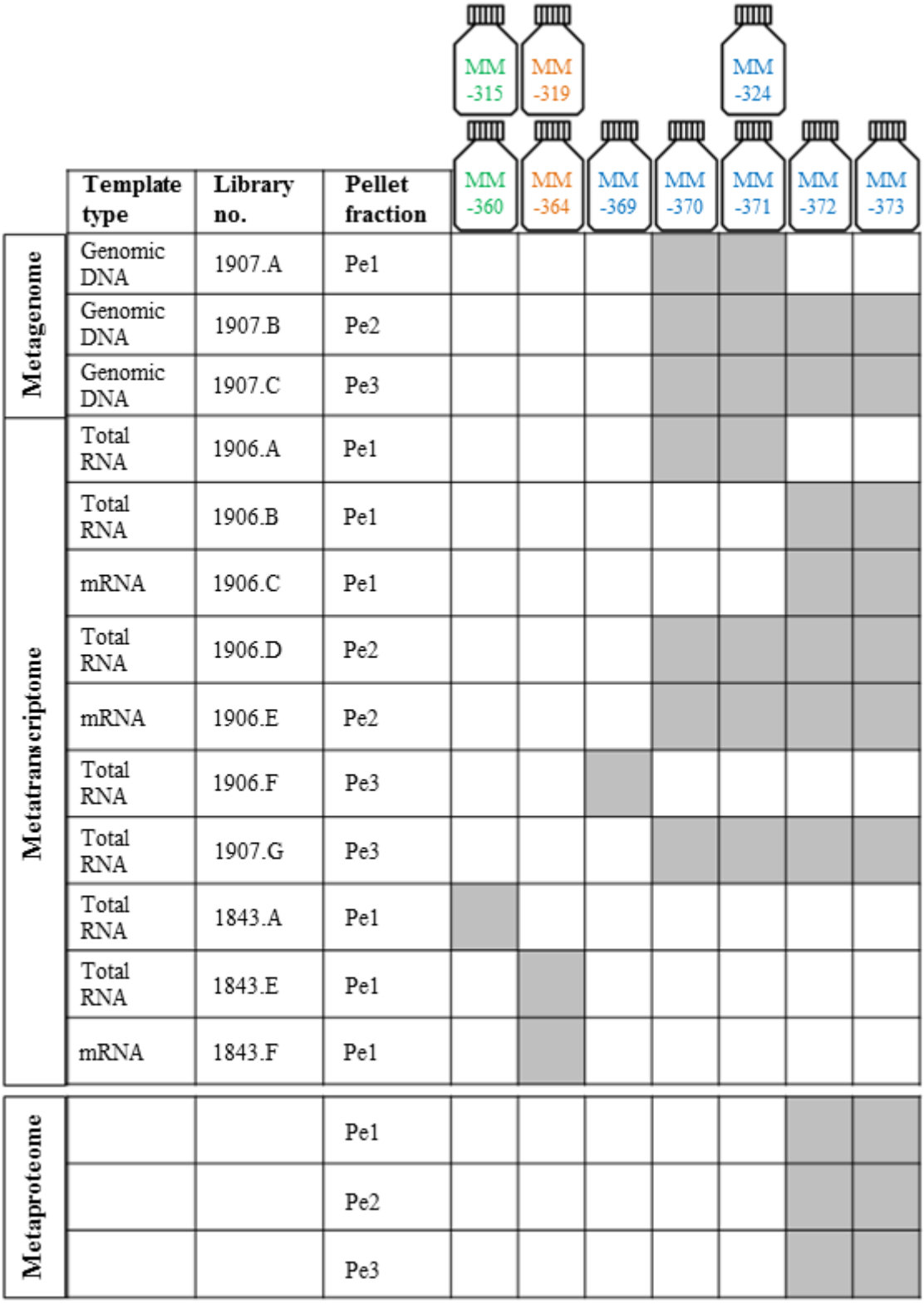


Fig. S20. Overview on culture materials selected for the individual "omics" studies. Precultures MM-315, MM-319 and MM-324 were inoculated in May, 2014 and used as inoculum for the cultures MM-360, -364, -369, -370, -371, -372, -373 in January, 2015. Biomass was harvested for "omics" studies in October/November, 2015.

Workflow 1

Dynamical trimming – SolexaQA

```
perl DynamicTrim.pl read1.fastq read2.fastq
```

Length filtering – SolexaQA

```
perl LengthSort.pl read1.trimmed.fastq read2.trimmed.fastq -l
<lowest median value>
```

File combining – Khmer

```
Interleave-reads.py read1.trimmed.paired1.fastq
read2.trimmed.paired2.fastq > trimmed.combined.fastq
```

Concatenate

Add single reads

```
cat read1.trimmed.single >> trimmed.combined.fastq
```

Digital normalization – Khmer

Adapted from https://khmer-protocols.readthedocs.io

```
phyton normalize-by-median.py -k 20 -C 20 -N 8 -x 5e8 --savetable
trimmed.combined.kh trimmed.combined.fastq
```

```
phyton filter-abund.py trimmed.combined.kh
trimmed.combined.fastq.keep
```

```
phyton normalize-by-median.py -k 20 -C 5 -N 8 -x 5e8
trimmed.combined.fastq.keep.abundfilt
```

Read extraction to get a paired read file and singleton file

```
phyton extract-paired-reads.py trimmed.combined.fastq.keep.abundfilt
```

Workflow 2

Adapter trimming – BBduk

```
bbduk.sh in1=read1.fq in2=read2.fq out=adapter_trim1.fq
out2=adapter_trim2.fq ref=/.../bbmap/resources/truseq.fa.gz ktrim=r
mink=12 k=28
```

```
bbduk.sh in1= adapter_trim1.fq in2= adapter_trim2.fq
out=qual_trim1.fq out2=qual_trim2.fq qtrim=rl qtrim=10 minlength=25
```

Normalization & error correction – BBnorm

```
bbnorm.sh in= qual_trim1.fq out=corrected1.fq ecc=t hist=hist1.txt
histout=histout1.txt
```

Workflow 3

Adapted from suggestion of Brian Bushnell (http://seqanswers.com/)

Removal of phiX, and other contaminat reads – BBduk

```
bbduk.sh in1=read1.fq in2=read2.fq out=clean.fq
ref=/.../bbmap/resources/phix174_ill.ref.fa.gz hdist=1 k=31
```

Adapter trimming – BBduk

```
bbduk.sh in=clean.fq out=adapter_trim.fq
ref=/.../bbmap/resources/truseq.fa.gz ktrim=r mink=11 k=25
```

Normalization & error correction – BBnorm

```
bbnorm.sh in= adapter_trim.fq out=corrected.fq ecc=t
```

Merging of data – BBMerge

Merged paired-end reads into single reads by overlap detection.

```
bbmerge.sh in= corrected.fq out=merged.fq outu=unmerged.fq
```

Quality-trimming of unmerged reads – BBduk

```
bbduk.sh in=unmerged.fq out=qtrimmed.fq trimq=10 qtrim=rl
minlength=50
```

Concatenate

```
cat merged.fq qtrimmed.fq >> all.fa
```

Workflow 4

Merging of data – BBMerge

Merged paired-end reads into single reads by overlap detection.

```
bbmerge.sh in1=read1.fq in2=read2.fq out=merged.fq outu=unmerged.fq
trimq=10 qtrim=t
```

Trimming of merged and unmerged data separately. Change the output file format to Fasta.

```
bbtrim.sh in=pe.fq out=pe.fa
ref=/.../bbmap/resources/phix174_ill.ref.fa.gz,/.../bbmap/resources/
truseq.fa.gz trimq=20 qtrim=t
```

```
bbtrim.sh in=se.fq out=se.fa
ref=/.../bbmap/resources/phix174_ill.ref.fa.gz,/.../bbmap/resources/
truseq.fa.gz trimq=20 qtrim=t
```

Combining

```
cat pe.fa se.fa > all.fa
```

Fig. S21. Workflows for MiSeq read processing.

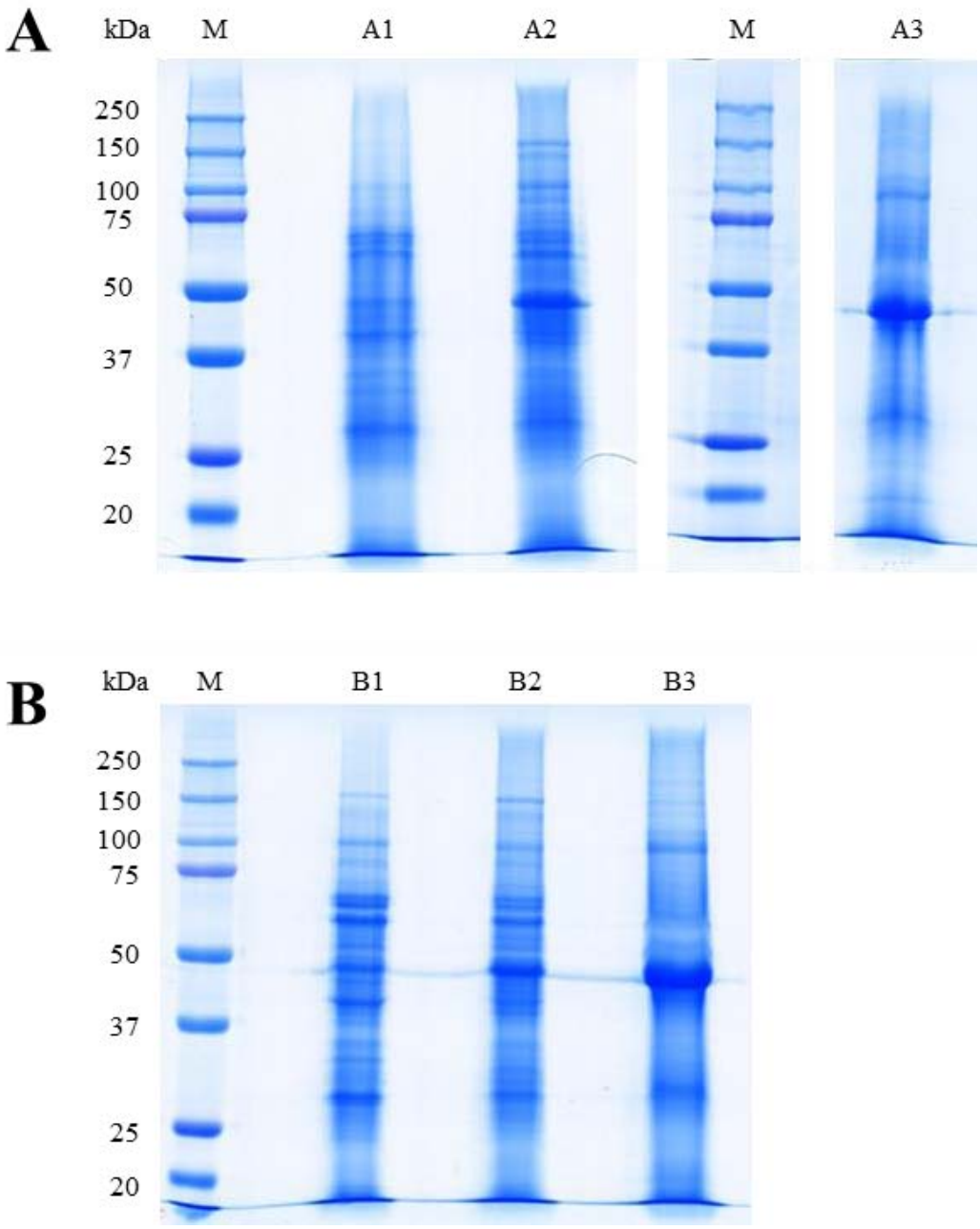


Fig. S22. Proteins present in biomass obtained by differential centrifugation from two limonene methanogenic enrichment cultures MM-372 (A) and MM-373 (B). Coomassie Brilliant Blue stained protein after separation by size in an SDS-PAGE. M is the molecular weight marker. A1 and B1 originated from fraction Pe1, A2 and B2 from Pe2, and A3 and B3 from Pe3.

Supplementary tables.

Tab. S1. Overview on total spectral counts (TSCs) of total and OP3 related proteins and proportion of OP3 related proteins in percent and NSAF.

	Proteins	**MM-372**			**MM-373**		
		Pe1	**Pe2**	**Pe3**	**Pe1**	**Pe2**	**Pe3**
TSCs[1]	4,704 (all)	66,386	64,092	86,46	70,327	95,496	88,676
	1,298 (OP3 LiM related)	44,755	19,359	53,165	23,486	56,078	60,998
% TSCs[2]	27.59	67.42	30.21	61.49	33.40	58.72	68.79
total NSAF[3]		66.31	23.23	56.33	26.06	52.75	66.64

[1] TSC: total spectral count.

[2] OP3 related percentage of TSCs.

[3] NSAF: normalized spectral abundance factor of 1,298 OP3 LiM proteins.

Tab. S2. Mass-abundant proteins identified by SDS-page in combination with HPLC-MS/MS.

Annotation	MW (kDa)	TSC in 372Pe1[1]	TSC in 372Pe3[1]	Feature ID[2]
18-strand beta-barrel outer membrane pore	53	2477	1843	peg.955
co-chaperone GroES CDS	58	933	1289	peg.1005
GPI-membrane anchored actin binding protein with KELCH repeats	499	837	828	peg.1770
translation elongation factor Tu CDS	44	775	991	peg.661
Pyruvate,phosphate dikinase (EC 2.7.9.1)	101	598	814	peg.972
very large multi-enzyme surface protein	4384	586	159	peg.37
DNA-directed RNA polymerase subunit beta CDS	128	517	694	peg.656
polyribonucleotide nucleotidyltransferase CDS	77	386	403	peg.1256
DNA-directed RNA polymerase subunit beta' CDS	152	370	646	peg.657
flagellar motor protein MotD	24	359	375	peg.236
expressed protein CDS	326	343	416	peg.589
Clp protease, ATP-binding subunit	91	340	383	peg.1300
peptidyl-prolyl cis-trans isomerase	35	334	309	peg.12
outer membrane protein assembly factor BamA CDS	87	316	273	peg.142
ATPase for pilus assembly/DNA-transfer CpaF CDS	82	292	359	peg.15
expressed DUF4139 protein, secreted	63	291	236	peg.189
glyceraldehyde-3-phosphate dehydrogenase	36	272	355	peg.902
transcription termination factor Rho CDS	47	259	283	peg.277
phosphoglycerate dehydrogenase CDS	57	255	321	peg.231
chromosome segregation protein SMC chromosome segregation protein SMC, archaeal type CDS	45	253	235	peg.52
glutamine synthetase type III CDS	83	252	322	peg.1443
DNA-directed RNA polymerase subunit alpha	37	242	278	peg.691
putative exosortase-affiliated protein	13	226	297	peg.152
expressed membrane protein CDS	28	219	188	peg.1614
transmembrane receptor with multiple PAS sensor domains and histidine kinase domain	170	218	297	peg.1028
molecular chaperone DnaK CDS	68	206	281	peg.633
DPS-like protein/bacterioferritin within the ferritin-like superfamily	18	200	191	peg.121
Pyruvate:ferredoxin oxidoreductase, alpha subunit (EC 1.2.7.1)	43	194	318	peg.115
bifunctional preprotein translocase subunit SecD/SecF	80	186	239	peg.1065
actin homologue MreB CDS	39	186	198	peg.1197
carbamoyl phosphate synthase large subunit CDS	118	183	329	peg.365
phosphoribosylpyrophosphate synthetase CDS	35	173	174	peg.133
phosphoglycerate kinase	43	171	196	peg.868
glutamate dehydrogenase CDS	49	169	170	peg.369
periplasmic-binding component of ABC transport system specific for sn-glycerol-3-phosphate	52	167	194	peg.41
ketol-acid reductoisomerase CDS	36	166	197	peg.196
pyruvate:ferredoxin oxidoreductase, gamma subunit (EC 1.2.7.1)	20	165	233	peg.113
secretion system protein with secretin and TonB domains	58	163	167	peg.14
short chain acyl-CoA dehydrogenases	42	163	145	peg.859
S-adenosylmethionine synthetase	42	157	191	peg.187
dihydroxy-acid dehydratase CDS	59	156	199	peg.193
pilin related to pseudopilin PulG	15	155	117	peg.1182
fructose-bisphosphate aldolase class-II	49	151	178	peg.192
alkaline phosphatase with homology to phosphomutase	46	151	136	peg.1454
30S ribosomal protein S2 CDS	33	149	165	peg.176
secreted protein containing glycoside hydrolase family 2, TIM barrel domain	94	145	168	peg.45
recombinase RecA	39	144	187	peg.1078
LPS assembly outer membrane protein LptD	80	143	100	peg.644
Translation elongation factor EF-G, a GTPase	71	141	192	peg.1700
chromosome segregation protein SMC, common bacterial type	124	139	223	peg.856
multidrug efflux pump subunit AcrB	116	138	192	peg.1581
translation initiation factor IF-2	79	138	183	peg.250

Tab. S2. – continued from previous page.

Annotation	MW (kDa)	TSC in 372Pe1[1]	TSC in 372Pe3[1]	Feature ID[2]
class V aminotransferase CDS	41	135	164	peg.230
6-phosphofructokinase	37	134	133	peg.809
PilT, pilus retraction ATPase	39	131	141	peg.695
expressed membrane protein CDS	78	131	129	peg.31
flagellar motor protein MotB	23	131	120	peg.1690
gliding motility-associated ABC transporter substrate-binding protein GldG	94	129	168	peg.1833
protein with domain of unknown function (DUF4912)	40	129	118	peg.238
proton-translocating pyrophosphatase CDS	72	128	147	peg.63
ATP-dependent Zn proteases FtsH	69	127	155	peg.1037
inosine 5-monophosphate dehydrogenase	41	127	157	peg.1004
transcription termination control protein NusA	43	126	155	peg.249
carbamoyl phosphate synthase large subunit CarB CDS	120	125	176	peg.1750
expressed protein CDS	43	122	320	peg.46
DNA topoisomerase tpye II (Gyrase) subunit B CDS	91	122	187	peg.4
3-deoxy-7-phosphoheptulonate synthase	37	121	130	peg.834
DNA gyrase subunit A CDS	91	115	170	peg.8
GMP synthase [glutamine-hydrolyzing], ATP pyrophosphatase subunit B	36	115	115	peg.366
Phosphoenolpyruvate-protein kinase (PTS system EI component in bacteria)	65	114	154	peg.185
ATP phosphoribosyltransferase CDS	33	114	103	peg.125
phosphodiesterase with HD domain of ribonuclease Y	59	113	144	peg.1985
expressed protein with unknown function	93	113	116	peg.580
short-chain dehydrogenase/reductase (SDR)	28	112	128	peg.1661
PilT-related nuclease/ATPase with N-terminal membrane domain (COG4956)	37	111	95	peg.1295
secretion system ATPase related to VirB11/CpaF	52	110	128	peg.1423
type I site-specific deoxyribonuclease, HsdR family	117	108	152	peg.560
expressed protein with unknown function	124	107	141	peg.320
type II secretory pathway ATPase GspE/PulE or type IV pilus assembly pathway ATPase PilB	43	107	144	peg.1249
phosphohexomutase superfamily	54	104	107	peg.1167
Type II secretory pathway component GspD/PulD (secretin) with additional STN domain	56	104	87	peg.1098
GMP synthase (glutamine amidotransferase II)	62	103	193	peg.1003
homocitrate synthase	46	102	116	peg.233
ATP-dependent protease/chaperone ClpB CDS	97	101	162	peg.629
30S ribosomal protein S4	24	100	119	peg.690
signal transducer with HAMP signal, GAF transducer and GGDEF (diguanylate cyclase) domains	51	99	98	peg.1190
bifunctional UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase/(3R)-hydroxymyristoyl-[acyl-carrier-protein] dehydratase	48	99	106	peg.145
Type II secretory pathway component GspD/PulD (secretin)	57	95	59	peg.1177
2-isopropylmalate synthase CDS	56	92	122	peg.197
S-adenosyl-L-homocysteine hydrolase	46	91	163	peg.188
cellobiose phosphorylase	93	90	127	peg.40
signal transducer with PAS sensor, GGDEF (diguanylate cyclase) and metal dependent phosphohydrolase domains	85	87	113	peg.319
ATP-dependent Zn-protease (COG0465)	62	85	115	peg.1595
alanyl-tRNA synthetase	96	84	148	peg.1076
ribonuclease E or G	57	84	130	peg.1174
isoleucine--tRNA synthetase	106	79	124	peg.1204
sensory signal transduction histidine kinase with PAS domain	114	76	138	peg.495
expressed protein with unknown function	53	75	134	peg.356
DNA polymerase I CDS	97	68	105	peg.274
enolase (phosphopyruvate hydratase	46	51	161	peg.1594

[1] 372Pe1 and 373Pe3 selected as representatives for all fraction samples.

[2] Feature ID´s are taken from the RAST annotation project with genome ID 6666666.225137.

Tab. S3. Molecule-abundant proteins identified by SDS-page in combination with HPLC-MS/MS.

Annotation	MW (kDa)	NSAF of 372Pe1[1]	NSAF of 372Pe3[1]	Feature ID[2]
18-strand beta-barrel outer membrane pore	53	4,012	2,557	peg.955
translation elongation factor Tu CDS	44	1,512	1,656	peg.661
putative exosortase-affiliated protein	13	1,492	1,680	peg.152
co-chaperone GroES CDS	58	1,381	1,634	peg.1005
flagellar motor protein MotD	24	1,284	1,149	peg.236
DPS-like protein/bacterioferritin within the ferritin-like superfamily	18	0,954	0,780	peg.121
pilin related to pseudopilin PulG	15	0,887	0,574	peg.1182
peptidyl-prolyl cis-trans isomerase	35	0,819	0,649	peg.12
pyruvate:ferredoxin oxidoreductase, gamma subunit (EC 1.2.7.1)	20	0,708	0,857	peg.113
expressed membrane protein CDS	28	0,671	0,494	peg.1614
glyceraldehyde-3-phosphate dehydrogenase	36	0,649	0,725	peg.902
DNA-directed RNA polymerase subunit alpha	37	0,561	0,552	peg.691
pyruvate phosphate dikinase (EC 2.7.9.1)	101	0,508	0,593	peg.972
flagellar motor protein MotB	23	0,489	0,384	peg.1690
chromosome segregation protein SMC, archaeal type CDS	45	0,483	0,384	peg.52
expressed protein with unknown function	19	0,474	0,360	peg.32
transcription termination factor Rho CDS	47	0,473	0,443	peg.277
polyribonucleotide nucleotidyltransferase CDS	77	0,430	0,385	peg.1256
phosphoribosylpyrophosphate synthetase CDS	35	0,424	0,366	peg.133
secreted protein containing peptidoglycan binding-like domain	14	0,411	0,242	peg.1626
actin homologue MreB CDS	39	0,409	0,373	peg.1197
Outer membrane chaperone Skp (OmpH)	21	0,401	0,427	peg.143
expressed DUF4139 protein, secreted	63	0,396	0,275	peg.189
ketol-acid reductoisomerase CDS	36	0,396	0,402	peg.196
30S ribosomal protein S2 CDS	33	0,388	0,368	peg.176
pyruvate:ferredoxin oxidoreductase, alpha subunit (EC 1.2.7.1)	43	0,387	0,544	peg.115
phosphoglycerate dehydrogenase CDS	57	0,384	0,414	peg.231
LemA/RetS hybrid sensor kinase-response regulator protein	21	0,380	0,427	peg.1440
30S ribosomal protein S4	24	0,358	0,365	peg.690
DNA-directed RNA polymerase subunit beta CDS	128	0,347	0,399	peg.656
short-chain dehydrogenase/reductase (SDR)	28	0,343	0,336	peg.1661
expressed protein with unknown function	22	0,343	0,264	peg.810
expressed secreted protein of unknown function	9	0,343	0,172	peg.216
phosphoglycerate kinase	43	0,341	0,335	peg.868
short chain acyl-CoA dehydrogenases	42	0,333	0,254	peg.859
ABC-type transporter Mla maintaining outer membrane lipid asymmetry, periplasmatic component MlaD	23	0,328	0,307	peg.1297
S-adenosylmethionine synthetase	42	0,321	0,334	peg.187
Clp protease, ATP-binding subunit	91	0,321	0,309	peg.1300
recombinase RecA	39	0,317	0,353	peg.1078
outer membrane protein assembly factor BamA CDS	87	0,312	0,231	peg.142
6-phosphofructokinase	37	0,311	0,264	peg.809
rubrerythrin	22	0,308	0,411	peg.1840
ATPase for pilus assembly/DNA-transfer CpaF CDS	82	0,306	0,322	peg.15
membrane or secreted protein containing peptidoglycan-binding Lysin subgroup domain (LysM)	19	0,298	0,352	peg.995
ATP phosphoribosyltransferase CDS	33	0,297	0,229	peg.125
glutamate dehydrogenase CDS	49	0,296	0,255	peg.369
short-chain dehydrogenase/reductase (classical (c) SDR)	28	0,294	0,257	peg.1582
PTS fructose/mannitol-specific subunit IIA [OP3_merged]	17	0,293	0,260	peg.182
DNA-binding protein similar to integration host factor beta subunit	10	0,292	0,243	peg.1000
30S ribosomal protein S7 CDS	18	0,291	0,253	peg.659
PilT, pilus retraction ATPase	39	0,288	0,266	peg.695
class V aminotransferase CDS	41	0,283	0,294	peg.230
alkaline phosphatase with homology to phosphomutase	46	0,282	0,217	peg.1454
3-deoxy-7-phosphoheptulonate synthase	37	0,281	0,258	peg.834

Tab. S3. – continued from previous page.

Annotation	MW (kDa)	NSAF of 372Pe1[1]	NSAF of 372Pe3[1]	Feature ID[2]
protein with domain of unknown function (DUF4912)	40	0,277	0,217	peg.238
periplasmatic-binding component of ABC transport system specific for sn-glycerol-3-phosphate	52	0,276	0,274	peg.41
expressed secreted protein with TMH and unknown function	15	0,275	0,240	peg.1095
GMP synthase [glutamine-hydrolyzing], ATP pyrophosphatase subunit B	36	0,274	0,235	peg.366
expressed secreted protein of unknown function	18	0,272	0,253	peg.1619
inosine 5-monophosphate dehydrogenase	41	0,266	0,282	peg.1004
fructose-bisphosphate aldolase class-II	49	0,265	0,267	peg.192
glutamine synthetase type III CDS	83	0,261	0,285	peg.1443
molecular chaperone DnaK CDS	68	0,260	0,304	peg.633
PilT-related nuclease/ATPase with N-terminal membrane domain (COG4956)	37	0,258	0,189	peg.1295
transcription termination control protein NusA	43	0,252	0,265	peg.249
DNA-binding transcriptional regulator, LacI/PurR family	32	0,247	0,216	peg.970
expressed protein CDS	43	0,244	0,547	peg.46
50S ribosomal protein L5 CDS	22	0,242	0,301	peg.675
expressed protein with unknown function	32	0,241	0,184	peg.590
D-alanine/D-glutamate or branched-chain amino acid aminotransferase	32	0,241	0,126	peg.1303
secretion system protein with secretin and TonB domains	58	0,241	0,212	peg.14
secreted protein containing molecular chaperone IbpA, HSP20 family	23	0,239	0,304	peg.1363
nicotinamidase/pyrazinamidase	31	0,235	0,187	peg.2006
nucleoside diphosphate kinase	19	0,235	0,201	peg.2007
Lipoprotein-anchoring transpeptidase ErfK/SrfK	30	0,232	0,213	peg.1682
pyruvate:ferredoxin oxidoreductase, beta subunit (EC 1.2.7.1)	31	0,230	0,235	peg.117
expressed secreted protein of unknown function	12	0,229	0,288	peg.217
dihydroxy-acid dehydratase CDS	59	0,227	0,248	peg.193
30S ribosomal protein S5 CDS	18	0,224	0,225	peg.680
protein with TMH, SAF and Flp pilus assembly protein RcpC domains	34	0,222	0,162	peg.13
membrane-associated protein with peptidoglycan-binding domains OmpA and LysM	37	0,220	0,233	peg.768
transcriptional regulator	27	0,219	0,210	peg.840
protein with tandem amino-acid binding ACT domain	14	0,215	0,473	peg.824
type II secretory pathway ATPase GspE/PulE or type IV pilus assembly pathway ATPase PilB	43	0,214	0,246	peg.1249
DNA-directed RNA polymerase subunit beta' CDS	152	0,209	0,312	peg.657
type II restriction endonuclease, TdeIII	32	0,207	0,168	peg.1905
adenine phosphoribosyltransferase [OP3_merged]	19	0,203	0,236	peg.969
bifunctional preprotein translocase subunit SecD/SecF	80	0,200	0,220	peg.1065
secretion system ATPase related to VirB11/CpaF	52	0,182	0,181	peg.1423
ATP-dependent Clp protease proteolytic subunit	22	0,172	0,241	peg.229
Translation elongation factor EF-G, a GTPase	71	0,170	0,199	peg.1700
S-adenosyl-L-homocysteine hydrolase	46	0,170	0,261	peg.188
cochaperone GroES (HSP10)	11	0,156	0,274	peg.1006
expressed protein with unknown function	9	0,153	0,180	peg.1110
proton-translocating pyrophosphatase CDS	72	0,153	0,150	peg.63
GMP synthase (glutamine amidotransferase II)	62	0,143	0,229	peg.1003
carbamoyl phosphate synthase large subunit CDS	118	0,133	0,205	peg.365
dihydrodipicolinate synthase	31	0,119	0,221	peg.1117
phosphopyruvate hydratase = enolase CDS	46	0,095	0,257	peg.1594
periplasmatic solute-binding protein of ABC transporter	28	0,083	0,305	peg.1815

[1] 372Pe1 and 373Pe3 selected as representatives for all fraction samples.

[2] Feature ID´s are taken from the RAST annotation project with genome ID 6666666.225137.

Tab. S4. Overview on raw data of next generation sequencing projects.

NGS technique	Biological sample		Project no.	Sequencer	Sequencing run conditions	Template type	Library no.	Run no.	Read no.	Read length (bp)	Total bases no.
454 sequencing	enriched OP3 LiM sample		385	454 GS FLX	Single reads	DNA	385.A	2	491,907	450	224,261,492
Illumina sequencing			1221	MiSeq	Paired-end reads	DNA	1221.A	2	9,888,618 x 2	250	4,963,092,232
	cell sized fractionated sample	Pe1	1907	HiSeq	Paired-end reads	DNA	1907.A	1	42,649,735 x 2	250	21,371,026,667
		Pe2	1907		Paired-end reads	DNA	1907.B	1	21,648,241 x 2	250	10,846,444,690
		Pe3	1907		Paired-end reads	DNA	1907.C	1	10,904,818 x 2	250	5,463,702,645
		Pe1	1906	HiSeq	Single reads	total RNA	1906.A 1906.B	1	12,998,858	150	1,904,775,898
			1906		Single reads	rRNA-depleted RNA	1906.C	1	57,878,109	150	8,482,580,003
			1843		Single reads	total RNA	1843.A	1	36,052,513	150	5,249,816,556
							1843.E	1	14,307,025	150	2,107,927,713
			1843		Single reads	rRNA-depleted RNA	1843.F	1	60,205,817	150	8,918,223,863
		Pe2	1906		Single reads	total RNA	1906.D	1	7,063,485	150	1,033,365,196
			1906		Single reads	rRNA-depleted RNA	1906.E	1	56,879,058	150	8,306,692,118
		Pe3	1906		Single reads	total RNA	1906.F 1906.G	1	10,025,317	150	1,500,837,967

Tab. S5. Overview on trimmed read data.

NGS technique	Library no.	Trimming treatment		Read type	Average read length (bp)	Read no.	Total bases no.
454 sequencing	385.A	mothur-1.29.1		se	418	426,697	178,148,218
MiSeq Illumina sequencing	1221.A	SolexaQA v.2.2./Khmer 1.0	workflow 1	pe/se combined	225	5,474,931	1,231,671,678
		Bbmap tools	workflow 2	pe/se combined	242	8,725,608	2,108,729,112
			workflow 3	pe/se combined	252	8,045,685	2,029,852,600
			workflow 4	pe/se combined	232	18,739,184	4,339,851,180
HiSeq Illumina sequencing	1907.A	SolexaQA v.3.1.4/Khmer 1.4.1 & Trimmomatic v.0.32		pe	193	8,222,350	1,586,898,275
				se	179	1,160,518	208,182,346
	1907.B			pe	188	3,674,326	690,559,509
				se	169	719,609	121,529,554
	1907.C			pe	187	1,610,392	301,535,200
				se	167	383,399	63,840,374
	1906.ABCDEFG	BBDuk/SolexaQA v.3.1.4 & Trimmomatic v.0.32		se	110	134,406,170	14,764,370,293
	1843.A	SolexaQA v.3.1.4		se	146	18,095,103	2,634,610,850
	1843.E			se	147	7,231,469	1,061,150,891
	1843.F			se	147	30,248,871	4,460,683,279

[1] Details to used tools and commands are shown in Fig. S 21.

[2] Merged file of reads of all performed libraries for project 1906.

Tab. S6. Statistics of *de novo* assemblies (SPAdes-3.8.0) obtained from each of the pellet fraction materials (Pe1, Pe2, Pe3) and from the combined meta-dataset (Pe123) using the QUAST web interface (57).

	Pe1	**Pe2**	**Pe3**	**Pe123**
Contigs	66,295	49,682	37,084	60,275
Total length (≥ 0 bp)	138,911,059	69,185,133	41,658,908	143,570,422
Total length (≥ 10,000 bp)	73,066,546	33,865,897	11,010,571	86,992,345
Largest contig	947,465	730,866	1,402,382	845,110
GC (%)	52.95	55.20	52.81	52.81
N50	14,596	23,779	2,784	24,276

Chapter 6

Discussion

After the discovery of candidate division OP3 in sediment of the Obsidian Pool (OP) based on 16S rRNA gene sequences, a diverse range of habitats, mostly anoxic environments, were discovered to contain 16S rRNA gene sequences of members of the c*andidate phylum* Omnitrophica. A broad phylogeny of this phylum was established, with 16S rRNA gene dissimilarities of up to 26%.

The phylotype OP3 LiM, one of the OP3 members, was detected in limonene-degrading, methanogenic enrichment cultures. The cells occurred either free-living or attached to larger cells of different morphology (Rotaru *et al.*, 2012). All attempts to isolate the bacterium in pure culture failed so far (Rotaru *et al.*, 2012). The characterization of OP3 LiM promised to gain insights into a so far undisclosed biology and this thesis describes my insights.

6.1 Culture stability

The methanogenic enrichment culture that utilized limonene as carbon and energy source originated from a wastewater sample which was taken in 1997 (Harder & Foss, 1999). A dilution-to-extinction series was prepared in 1999. A culture with one µl of inoculum grew and all cultures used in this study originated from this culture. The small inoculum was used to obtain a limited diversity in the methanogenic enrichment cultures.

Bacterial genomes vary extensively in terms of both gene content and gene sequence (Lees *et al.*, 2016). The initial bottleneck should enable a close-to-clonal population structure to obtain population genomes with little sequence variation. Twelve culture lines were established in 2005. Cultures produced methane for two to three years. These culture lines were maintained by one annual transfer of 10% v/v inoculum. Active cultures of each line were chosen as inoculum. First studies observed that 18% of all cells were OP3 LiM cells in the enrichment during the exponential growth as revealed by catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH) (Rotaru *et al.*, 2012).

Populations of the same species are likely to be maintained, when similar environmental conditions are available (Konstantinidis & Rosselló-Móra, 2015). This may also be expected for stable enrichment cultures. But the 12 lineages varied in their activity. The methane production profiles varied for the cultures of the same passages and demonstrated a range from inactive to highly active cultures. Thus, we observed an instability of the microbial community. In an early set of experiments of this thesis in 2014, OP3 LiM-specific PCR (colony PCR and PCR with genomic DNA) as well as CARD-FISH were performed to investigate the presence of OP3 LiM cells in the lineages. At that time, seven cultures were active and contained OP3 LiM. One culture was active, but OP3 LiM was apparently lost in the last transfers. Four cultures had no OP3 LiM cells. Such variations in the structure of microbial communities, including members of OP3, were previously described. Among lineages of benzene-degrading, methanogenic enrichment cultures, the OP3 population was absent in culture lineages, which were more frequently transferred suggesting very slow doubling times of OP3 (Luo *et al.*, 2016). Shifts of archaeal and bacterial communities, induced by changes in the dilution rate, were identified for propionate-degrading, methanogenic chemostat cultures (Shigematsu *et al.*, 2006) as well as for butyrate-degrading, methanogenic chemostats. In the latter, candidate division OP3 were predominant at high dilution rate (Tang *et al.*, 2007).

To understand the instable coexistence of microbial species within mixed cultures, all members of the community should be defined and their role and relationships among them should be investigated (Kato *et al.*, 2005). The microbial community compositions of the limonene degrading enrichments were revealed by the full cycle 16S rRNA approach. Clone libraries for bacterial and archaeal 16S rRNA genes were established and revealed a complex microbial community. Syntrophic interactions between these members were proposed (Rotaru, 2009). However, a balance of the various types of relationships (both positive and negative) is considered to be essential for the stability (Kato *et al.*, 2005). The examination of general characteristics of each of the members in pure culture and the co-cultivation of these isolates in order to investigate their functions and relationships would facilitate the successful construction of mixed cultures (Kato *et al.*, 2005). However, currently there is no isolate available for members of *candidate phylum* Omnitrophica.

OP3 LiM is metabolically active as attached organism and preys on *Methanosaeta.* Thus, it is not essential for the methanogenic process and may be lost in highly active cultures. Alternatively, it may severely impair and kill the methanogenic process by a very

successful predation. Thus, we have to consider the enrichment cultures as potentially instable. The maintenance of OP3 LiM in the cultures is the result of a sensitive equilibrium.

6.2 Challenges in genome closure and finishing

The genome of the bacterium OP3 LiM was obtained from metagenomic datasets of OP3-enriched cell fractions and of size-fractionated cells of the enrichment cultures. Metagenomics provides ways to characterize not yet cultured microbes as well as microbial communities and their interactions (Segata *et al.*, 2013). Population genomes are assembled from metagenomic data (Tsai *et al.*, 2016).

Complex microbial communities such as the limonene-degrading methanogenic enrichment cultures require a large sequencing effort. Physical cell separation was performed to reduce the complexity of the metagenome by the separation of the diversity into individual gradient fractions. Highly enriched OP3 LiM fractions were obtained. A higher representation of a population in a metagenomic dataset (greater dominance) will result in a greater likelihood of assembly and recovery of contigs (Kunin *et al.*, 2008; Sangwan *et al.*, 2016). Consequently, a significant part of the genome and, in some cases, a near-complete genome of the dominant species is likely to be obtained by assembly (Kunin *et al.*, 2008). OP3 LiM cells were enriched to over 80% of all cells in two consecutive density gradient centrifugations and allowed the extraction of genomic DNA and consequently the start of the metagenome project (Chapter 5). An alternative approach is fluorescent activated cell sorting (FACS) by flow-cytometry after FISH or CARD-FISH hybridization in combination with genome amplification. This application has the potential of direct extraction of specific subpopulations from environmental samples (Kalyuzhnaya *et al.*, 2006; Gawad *et al.*, 2016).

Currently, about 72,000 datasets of microbial genomes are available for studies. However, only a small part of about 12,000 genomes are complete, they have been finished. In contrast, draft genomes or permanent draft genomes are incomplete. The finishing process is mostly performed for sequences obtained from cultured organisms. Only 179 finished genomes have been extracted from metagenomic datasets. The majority of genomes derived from metagenomes are published as draft genomes or permanent draft genomes (June 2017, https://img.jgi.doe.gov). As complete genomes are not a requisite for many applications and the creation of the best possible assembly from the available data may already reveal biological insight, the effort to finish the genome is usually avoided (Wences & Schatz, 2015;

Mardis *et al.*, 2017). Consequently, large percentage of genomes only achieves high-quality draft status (Utturkar *et al.*, 2014; Land *et al.*, 2015). However, a complete genome is highly desired for a full understanding of the biology of the organism. A finished genome sequence is required to examine all rather than most of the genes in an organism (Mardis *et al.*, 2017) and allows the argumentation of the absence of functions.

The assembly of DNA reads to correctly reconstruct genomes is an essential task in bioinformatics (Utturkar *et al.*, 2014). Assembly of the complete genome sequence of each single microbial population is limited by biological sequence information, e.g. long direct or inverted repeats, longer than sequence reads and by problems of the sequencing technology to cover all regions equally (Klumpp *et al.*, 2012; Konstantinidis & Rosselló-Móra, 2015; Tsai *et al.*, 2016). Assembly algorithms usually finish the contig construction at these repeats resulting in a genome assembly with many contigs (Klumpp *et al.*, 2012; Tsai *et al.*, 2016). Finishing is the process of resolving the contig order and filling the sequence gaps. This process of finishing a genome requires the availability of template DNA, time, resources and expertise (Ekblom & Wolf, 2014; Mardis *et al.*, 2017). In my work, I first explored *in silico* methods to improve the assembly and then used combinational PCR and amplicon sequencing for finishing the OP3 LiM genome.

De novo draft assemblies of high quality but with many contigs can be obtained from Illumina short read datasets by the application of few bioinformatics tools and without the requirement of specific bioinformatics expertise. I explored variables in existing bioinformatics tools as well as the use of taxonomic information gained from the contigs. Iterative rounds of taxonomic evaluation of contigs, read mapping to taxonomically coherent contigs, and reassembly of the recruited reads efficiently improved draft assemblies significantly (Chapter 3). The usage of a taxonomic classifier resolved some gaps between contigs of single-genome draft assemblies. The differentiation and separation into taxon-related and taxon-unrelated groups of contigs by the Metawatt binner (Strous *et al.*, 2012) improved Illumina read assemblies, e.g. of different planctomycetes strains (SWK21, SM50, K833) (Chapter 3). These genomes demonstrated the usage of currently available bioinformatics programmes and tools in novel ways to refine the genome assembly. The visual inspection of mapping results in a bioinformatics software platform (e.g., Geneious) contributed to an understanding of the assembly problem, e.g. the read coverage reflected the numbers of repeats in the genome, and provided an *in silico* read walking with 454 reads for the OP3 LiM genome.

The small number of finished metagenome-derived genomes integrated into the Integrated Microbial Genomes & Microbiomes (IMG/M) system (https://img.jgi.doe.gov) is an indication for the complexity of finishing genomes from metagenomes. A variety of strategies and different tools were needed to close all gaps of the OP3 LiM draft assembly and to optimize the final sequence. An available finished reference genome is a distinct advantage for the challenge of genome finishing due to the performance of mapping and the discovery of genetic variations (Klumpp *et al.*, 2012). For OP3 LiM a reference genome was absent and members of the c*andidate phylum* Omnitrophica represent an extremely phylogenetically diverse bacteriological group in nature (Chapter 1, Table S1; Fig. 2). Consequently, *de novo* assembly was approached with read datasets of the 454 sequencing technology and the Illumina technology obtained from genomic DNA of OP3 LiM-enriched fractions. A second biological sample was size-fractionated and provided three Illumina datasets for confirmation of the OP3 LiM genome sequence. Each sequencing platform has its own unique features and advantages over other platforms as well as produces technology based sequencing errors (Klumpp *et al.*, 2012).

Contigs of a Newbler assembly of 454 reads was taxonomically characterized to identify 22 contigs associated to OP3 LiM. These served as starting point for the mapping and contig elongation using 454 and Illumina MiSeq reads. The combination of data of different read lengths and from several and different sequencing platforms can counterbalance drawbacks of each method (Ekblom & Wolf, 2014). Today, the 454 technology is replaced by the PacBio platform and actual projects use a combination of short reads from the second generation sequencing (SGS) platforms and long reads generated by third generation sequencing (TGS) technology, such as PacBio (Au *et al.*, 2012; Utturkar *et al.*, 2014; Tsai *et al.*, 2016).

Central challenge in genome assembly are repetitive sequences which can give rise to false overlaps or fragments (Wences & Schatz, 2015). Read mapping onto the most advanced *in silico* refined OP3 LiM draft assembly revealed the presence of repetitive sequences. The assembler had collapsed tandem repeats of two or more copies of sequences. *In silico* approaches, e.g. by read walking, failed to solve these suspicious regions because the repeats were longer than the length of 454 reads. PCR with primers located outside of the repetitive sequences yielded amplicons that filled the sequence gaps and indicated the physical map of the repetitive elements.

One repeat was detected in the largest ORF, a gene of 119,034 bp and required the design of 17 primers to integrate two additional Newbler contigs with similar repeat

sequences in the accurate order into the OP3 LiM genome. To ensure that unique primers correspond to the regions outside of the repeats, the design of a second, more specific reverse primer was required for one contig-end resulting in 9 reverse and 8 forward primers. Another repeat present within both Newbler contigs was the cause of false-positive assemblies. PCR reactions with primers, derived from unique contig regions located next to this repeat, revealed the misassembly. One contig had to be integrated into the other contig before the integration into the OP3 LiM genome. Sequence information obtained by combinatory PCR also revealed small sequence variations in the repeats. The physical sequence order around the repetitive elements was deduced by accurate mapping of acquired sequences and by identification of the approximate length of PCR products providing by the used molecular weight size marker. DNA fragments of 500 bp up to 3,000 bp size were generated allowing the implementation of the regular PCR method.

Combinatorial PCR reactions were also performed for a second questionable region of multiple insertions in the OP3 LiM genome. The genetic element contained duplicated genes of unknown function (CHP1 and CHP2), each with a highly-variable region of 83 bp, a poly-G region consisting of 21 guanine bases, a group II intron with a gene for a reverse transcriptase and three 23S rRNA intervening sequences (IVSs). PCR reactions with combinations of 7 forward and 7 reverse primers, designed for the detection of the overall suspicious region as well as small segments of the regions, confirmed the accuracy of the assembly for this region with multiple insertions. A special problem was the inability of dideoxy sequencing to confirm the polymeric G sequence in amplicons.

Small variations between the datasets of different sequencing platforms (454 and Illumina technology) were identified. The comparison of these data revealed similarity between Illumina datasets generated from two different biological samples and variations of these datasets to that generated from 454 sequencing. The manual inspection showed homopolymer-associated, single-base errors established as high error rate for Roche 454 (Luo *et al.*, 2012). Consequently, OP3 LiM genome was manually edited to the sequence information of matched reads from the Illumina metagenomes. The confirmation of the finished circular OP3 LiM genome by a second biological sample additionally indicated the constructed average “mosaic” genome (Konstantinidis & Rosselló-Móra, 2015) with the very low single-nucleotide polymorphisms of the population genome as nearly clonal consensus genome.

Overall, the assembly and finishing of a complete OP3 LiM genome required a variation of read processing methods, a variety of assembly and mapping software, the

introduction of taxonomic binning software in the quality evaluation and a large set of lab experiments required to resolve long repeated sequences in the genome. Software tools greatly vary on different datasets (Wences & Schatz, 2015) and in this study, visual inspection revealed their limitations. The adaption of the developed approach to the individual dataset and especially the manual inspection of all results were essential.

6.3 Visualization of OP3 LiM

Visualization experiments applying CARD-FISH indicated an epibiontic lifestyle for OP3 LiM. The signal intensity of attached cells indicated via a large number of ribosomes that bacterial as well as archaeal members of the methanogenic enrichment cultures provide nutrition to OP3 LiM. Transmission electron microscopy (TEM) also provided evidence of OP3 LiM attachment to larger cells. With a size of 200 nm to 300 nm, OP3 LiM cells have to be closely associated to larger cells via cell-to-cell contact to become pelleted in the centrifugation.

CARD-FISH allowed the visualization of strong signals. Low cellular rRNA contents of the target organisms hamper *in situ* identification by FISH. The usage of CARD-FISH enhanced the fluorescent FISH signals (Molari & Manini, 2012). Nevertheless, this method has also limitations which must be overcome by adaption the application. A critical step of this approach is the penetration of the horseradish peroxidase (HRP)-labeled probes into fixed cells due to their large size. Because the cell-wall composition of *Bacteria* and *Archaea* is diverse, no standard protocol exists and an optimization of the permeabilization step is usually necessary which considers the specific cell wall composition of the target organisms (Amann & Fuchs, 2008). We observed an inadequate and heterogeneous detection of rRNA signals of *Methanosaeta* cells. Lysozyme, usually used for the permeabilization of prokaryotic cells, was not effective for the diffusion of HRP-labeled ARCH-915 into the *Methanosaeta* sheath. A harsh treatment with proteinase K was necessary for the detectability of homogeneous CARD-FISH signals of the filamentous cells. Nevertheless, some cells in filaments lacked DNA and RNA. Because of the visibility of these filaments in phase-contrast micrographs and due to similar results by the application of *in situ* DNA-hybridization chain reaction (HCR), an approach without permeabilization step (Yamaguchi *et al.*, 2015), a loss of their cellular content by predation is suggested (Chapter 4).

However, the visualization of pure-cultured *Methanosaeta* cells should also confirm our finding of biological reasons for the partial staining. All uninfected cells of the filaments should be stained by ARCH-915 and DAPI.

Metabolically active cells can be identified using FISH (Molari & Manini, 2012). A significant problem for FISH of 16S rRNA with oligonucleotide probe is also the target site inaccessibility. Unlabeled oligonucleotides (helpers) that bind adjacent to the probe target site have the potential to increase weak probe hybridization signals (Fuchs *et al.*, 2000). Introducing of four helpers improved the CARD-FISH signals of OP3 LiM cells. We detected weak signals for free-living OP3 LiM cells in comparison to strong signals for attached cells to archaeal and bacterial cells suggesting a higher metabolic activity of OP3 LiM during the attachment. Due to bright signals of OP3 LiM when attached to bacterial cells of different morphology, we suggested also the ability to prey on a wide range of bacterial cells supported by the visualization of the 23S rRNA intron of OP3 LiM within a bacterial cell aggregate (Chapter 5, Fig. S13).

6.4 OP3 LiM, a new predator

Association of OP3 LiM cells to *Methanosaeta* together with the absence of DNA and rRNA in filamentous cells and the presence of the bacterial intron RNA within this cell morphology indicated a predation of archaea. The OP3 LiM genome has the potential for a predatory life of the cell. The combination of both visualization and genome information suggests that OP3 LiM is a novel epibiotic predator. So far, few predatory prokaryotes with an epibiotic lifestyle are known: "*Bdellovibrio* and like organism" (BALO) strains, represented by *Bdellovibrio exovorus* sp. nov. (Jurkevitch & Davidov, 2006) and *Micavibrio aeruginoavorus* (Dashiff *et al.*, 2011), as well as the bacteria *Vampirococcus*, *Vampirovibrio* and *Ensifer* (Esteve & Gaju, 1999).

A range of epibiontic lifestyles are known including different preys, cell morphology, response to oxygen, and mode of feeding and reproduction. Still, common features are present (Pérez *et al.*, 2016) and we locked for characteristics of these in OP3 LiM.

The observation of different intensities of CARD-FISH signals as well as different cell sizes of small attached coccoid cells indicated a lifestyle similar to obligate bacterial predators. The observed biphasic lifecycle for OP3 LiM consisting of a free living attack phase and a growth and replication phase seem to be similar to that of epibiotic predators,

such as *Bdellovibrio exovorus* and *Vampirococcus*, which remain attached to the prey outer membrane while attacking (Mahmoud & Koval, 2010).

A small cell size might be a common feature of epibiotic bacteria. *Bdellovibrio exovorus* is a gram-negative, comma-shaped rod with 0.5 µm width and 0.5 µm to 1.4 µm length (Koval *et al.*, 2013) and *Vampirococcus* is gram-negative and ovoidal (0.6 µm wide) (Esteve & Gaju, 1999). With a cell size of about 0.2 µm to 0.3 µm OP3 LiM is smaller than the other members of the methanogenic community. "Small eats big" has been established as rule for bacterial predators (Pérez *et al.*, 2016).

A small genome length appears to be a characteristic signature of epibiotic predators. The complete genome of OP3 LiM consists of 1,974,201 bp and is predicted to encode for 2,015 open reading frames (ORFs). These genome features correspond to the genome lengths and number of predicted ORFs of *Bdellovibrio exovorus* strain JSS (2.66 Mb; 2,669 ORFs), *Micavibrio aeruginosavorus* EPB (2.46 Mb; 2,460 ORFs), and *Micavibrio aeruginosavorus* ARL13 (2.48 Mb; 2,432 ORFs) (Pasternak *et al.*, 2014). Periplamic predators have larger genomes (*Bdellovibrio bacteriovorus* HD 100: 3.78 Mb, 3,586 ORFs and *Bdellovibrio marinus* JS: 3.44 Mb, 3,231 ORFs) (Pasternak *et al.*, 2014). This significant difference between both main predatory strategies is a further indication for the epibiotic strategy of OP3 LiM.

We identified an active Sec pathway and a type II secretion / type IV pilus assembly system in the high organized OP3 LiM genome. Genes for a secretion pathway and a very large multi-enzyme surface protein with a size of 4,384 kDa were directly located after genes involved in DNA replication. In addition to the used annotation pipelines, NCBI Prokaryotic Genome Annotation (Angiuoli *et al.*, 2008), Rapid Annotations using Subsystems Technology (RAST) (Aziz *et al.*, 2008) and the in-house annotation based on GenDB (Meyer *et al.*, 2003), two other ORF prediction programs, the open reading frame finder by NCBI (https://www.ncbi.nlm.nih.gov/orffinder/) and the ORF-finder of the Sequence Manipulation Suite (Stothard, 2000), predicted the ORF of the very large multi-enzyme surface protein. This large protein involved enzymes that function abundantly in polymer degradation including a repertoire of ten carbohydrate-active enzymes (CAZymes) (Yin *et al.*, 2012) including domains of glucosidases and several glycosyltransferases. Bright detection signals for several OP3 LiM cells, especially when attached to individual bacterial cells (Chapter 5 Fig. 7 and 8), indicated a high metabolic activity of these cells. Due to these observations we suggest OP3 LiM may also nourish on lipopolysaccharides (LPSs) present in gram-negative bacteria. All metabolic enzymes involved in glycolysis were found indicating the bacterium is

capable of utilizing sugars. For bacteriophages, which infect gram-negative bacteria, enzyme activities of LPS hydrolysis was already identified (Yan *et al.*, 2014). Beside the degradation of macromolecules the multifunctional protein has receiver and transmitter domains for the reception of environmental signals. A diguanylate cyclase/phosphodiesterase (GGDEF & EAL domains) was identified, capable of the synthesis and degradation of the bacterial second messenger cyclic diguanylate (c-di-GMP) (Tamayo *et al.*, 2007).

When a predator finds a suitable prey cell, a cascade of quickly events for the growth phase has to start. To have a protein for complex cellular functions including domains of receiver and transmitter might influence their metabolic activity. Free-living OP3 LiM cells are in a state of low metabolic activity to survive starvation. Signaling molecules interact with the prey cell. Certain components of the outer membrane of the prey cell, e.g., the bacterial LPS, are postulated to serve as trigger for the activation of second messengers, followed by a cascade of effects for the infection. The perception of this extracellular stimulus is coupled to an adaptive response and all steps will be exerted, which are necessary for the growth phase of OP3 LiM. Consequently, OP3 LiM is metabolically active as attached organism.

OP3 LiM lacks genes for a flagellum, but genes for stator elements (proteins of Mot complexes) are present. These proteins are able to form a channel (Bardy *et al.*, 2003). The most abundant protein is a 18-stranded beta-sheet outer membrane pore protein. Beta barrels are described to be suitable objects for channel engineering (Schulz, 2002) and are involved in the transport of polymers across the outer membrane (Whitney *et al.*, 2011). The pore may also serve for pili. Several locations in the genome harbor *pil* genes encoding type IV pili and *pil*-related genes. Many of the proteins involved in type IV pilus biosynthesis and function share similarity with proteins in type II secretion systems and archaeal flagellum systems. Several proteins are phylogenetically related, suggesting that these three machineries may share functional characteristics. Indeed, the formation of pilin-like structures was observed in an overexpression of the pseudopilin (PulG) from the type II secretion system in *Klebsiella oxytoca* (Jakovljevic *et al.*, 2008). The pilus fibre is composed of a single pilin protein, PilA (Bardy *et al.*, 2003). The genome of OP3 LiM possesses a PilA-related pseudopilin PulG as most abundant pilin (peg.1182) suggesting PilA polymerization might be initiated as suggested for the *Klebsiella oxytoca* type II pullulanase secretion system (Peabody *et al.*, 2003; Köhler *et al.*, 2004; Seitz & Blokesch, 2013). So far, the visualizations have not resolved pili on the surface of OP3 LiM. Because of the small cell size of the bacterium, advanced tools are required. Pili of ultra-small bacteria have been visualized using

cryogenetic (cryo)-TEM (Fig. 1, Luef *et al.*, 2015). OP3 LiM may have similar functional pili-like structures.

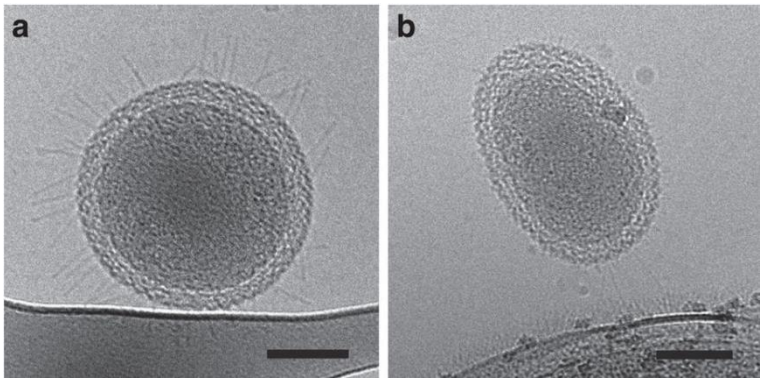


Fig. 1. Cryo-TEM images (2D) of an ultra-small bacterium with pili-like structures, (a) distributed across the cell surface, (b) connected to an adjacent bacterium. Scale bar, 100 nm (Luef *et al.*, 2015).

The ATPases PilB and PilT drive energetically the extrusion and retraction of pili (Bardy *et al.*, 2003). Cycles of pilus assembly and disassembly have been linked to "twitching" motility (Peabody *et al.*, 2003). Several expressed *pilT* and *pilB* genes are identified in the OP3 LiM proteome. This may provide a mode of motility for the flagellum-less OP3 LiM cells. In addition, PilT ATPases are involved in the uptake of foreign macromolecules which suggests together with many nucleic acid binding domains in expressed proteins that nucleic acids are a nutrient for OP3 LiM. A cell-to-cell contact via pili has been considered as essential element of predation (Wang *et al.*, 2011; Pasternak *et al.*, 2014; Pérez *et al.*, 2016). Genomes of epibiotic predators such as *Micavibrio* and *Bdellovibrio exovorus* JSS^{T} also possess genes encoding type IV pili suggesting a function of type IV pili for the initial attachment to prey cells (Chanyi & Koval, 2014).

OP3 LiM shares many characteristics with other epibionic predators, especially with the obligate predator *Vampirococcus*. Both are able to grow under anaerobic conditions, which differs they from the strict aerobe bacteria *Bdellovibrio*, *Ensifer* and *Vampirovibrio*. *Vampirococcus* cannot be sustained long enough outside its natural environment to isolate a pure culture according to Guerrero *et al.* (1986). The bacterium appears to absorb the contents of their prey *Chromatium* like a "vampire", which resulted in name of *Vampirococcus* (Guerrero *et al.*, 1986). We observed the absence of DNA and rRNA in *Methanosaeta.* Growth of OP3 LiM during attachment on the archaeal filaments can be imagined from different cell sizes of OP3 LiM cells in TEM and scanning electron microscopy (SEM)

images. Some of these filaments were deformed and seemed to have lost their cellular proteins responsible for the cell morphology. In conclusion, our description of OP3 LiM fulfills the requirements of a *Candidatus* proposal (Konstantinidis & Rosselló-Móra, 2015). We propose to name the OP3 LiM cells *"Candidatus* Vampirococcus archaeovorus".

6.5 Diversity of *candidate phylum* Omnitrophica

Based on 16S rRNA sequences, a diverse range of habitats was identified for the *candidate phylum* Omnitrophica, covering anoxic aquatic and terrestrial environments (Kumar & Saravanan, 2010; Morrison *et al.*, 2017). The ubiquity of the phylum may imply a prominent role in the functioning of the ecosystem (Pisapia *et al.*, 2017). Genomic insights of *candidate phylum* Omnitrophica has been obtained from draft genomes originating from metagenomes and single cell genomes. A draft genome with a size of 4.0 Mb and consisting of 181 contigs (Genbank acc. no. LMZT00000000) has been obtained from granulates of wastewater treatment plant and described as the most complete *Omnitrophica* genome (Speth *et al.*, 2016). Therefore, this study provides the first completely closed high-quality genome of this phylogenetic group. The OP3 LiM sequence was reconstructed from metagenomic data of methanogenic, limonene-degrading enrichment cultures. OP3 16S rRNA sequences were previously found in slow-growing methanogenic enrichment cultures thriving on propionate (Shigematsu *et al.*, 2006), butyrate (Tang *et al.*, 2007) and benzene (Luo *et al.,* 2016).

Omnitrophica is described as phylogenetical broad phylum (Glöckner *et al.*, 2010; Kolinko *et al.,* 2012; Ivanova & Dedysh, 2012). The 16S rRNA gene sequence of OP3 LiM also indicated a large phylogenetic distinction to the next described species. The sequence has an identity of 99% to 16S rRNA genes of LiM clones from our previous study (Rotaru *et al.*, 2012) and of 98% to the 16S rRNA gene of another clone obtained from benzene-degrading methanogenic enrichment cultures (Genbank acc. no. KT025835). Identities of lower than 95% to other published cloned-derived sequences (Fig. 2), of 77% to the 16S rRNA of the next-related single cell genome (Genbank acc. no. ASOC01000103) and of 94% to the next-related 16S rRNA of a metagenome-derived genome (Genbank acc. no. MNVR01000026) represented the phylogenetic diversity. Except for the closed relatedness of OP3 LiM to one clone found in a benzene-degrading methanogenic enrichment culture (Genbank acc. no. KT025835), the broad phylogeny of *candidate phylum* Omnitrophica is also confirmed by the low 16S rRNA gene similarity of OP3 LiM to the next described OP3-related species of

methanogenic enrichment cultures. The sequence has an identity of 82% to genes of clones found in propionate-degrading methanogenic enrichment cultures (Genbank acc. no. AB232821, AB232806, and AB232814) and of 82% to genes of clones found in butyrate-degrading methanogenic enrichment cultures (Genbank acc. no. AB248652, AB248653, and AB248654).

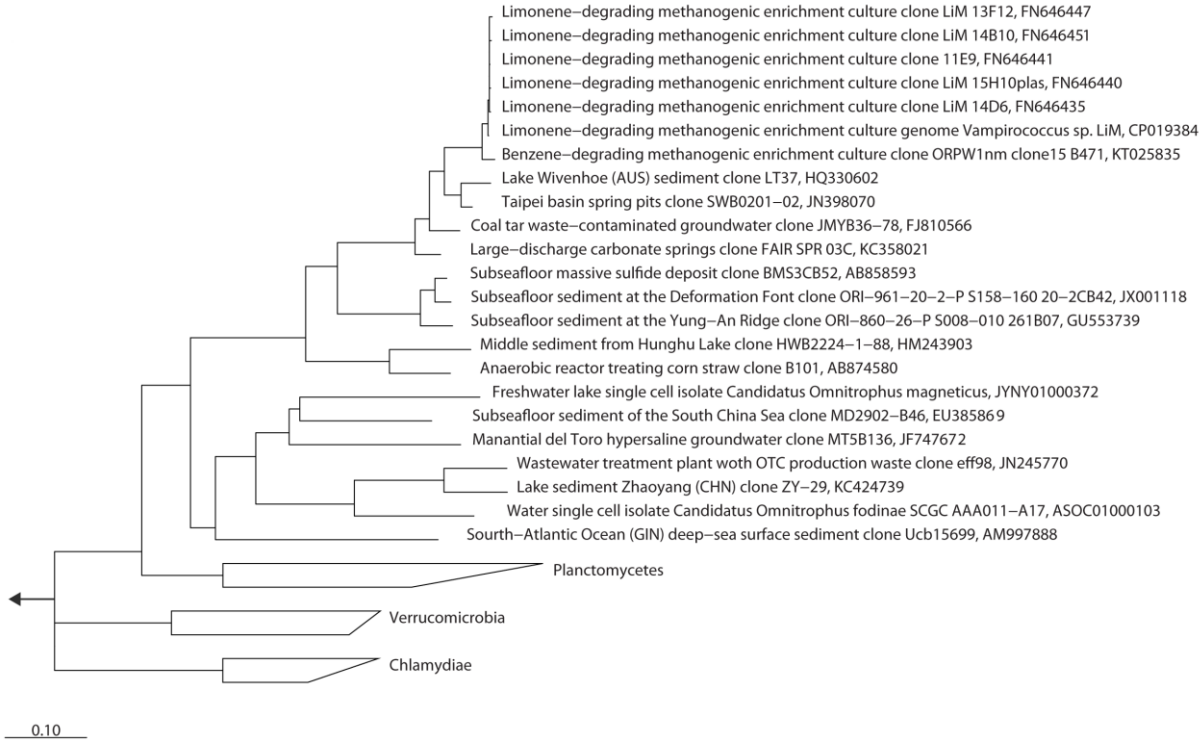


Fig. 2. Phylogenetic 16S rRNA-based tree of OP3 LiM and related organisms, calculated by randomized axelerated maximum-likelihood (RAxML) analysis without a filter. The phylogenetical distance of OP3 LiM (Genbank acc. no. CP019384) to described species of *candidate phylum* Omnitrophica and to closely relative groups was calculated including all sequences with 16S rRNA gene identity of over 90%, some representatives with identities < 90% to the 16S rRNA gene of OP3 LiM as well as the affiliation to *Planctomycetes*, *Verrucomicrobia*, and *Chlamydiae* (PVC superpylum). *Lentisphaerae* were used as outgroup (Tree reconstruction in cooperation with K. Knittel, 2017).

At present, the genus "Omnitrophus" has been published as candidatus genus with two species. A single cell draft genome, as representative of first substantive genomic data for candidate bacterial phyla OP3 (Rinke *et al.*, 2013), was tentatively named "*Candidatus* Omnitrophus fodinae" (Genbank acc. No. ASOC00000000). The draft genome of a magnetotactic representative from the *candidate phylum* Omnitrophica was tentatively named "*Candidatus* Omnitrophus magneticus" (Genbank acc. no. JYNY00000000) (Kolinko *et al.*,

2016). With 16S rRNA gene identities of 77% between OP3 LiM and both species (Fig. 2), the sequence divergence indicated that OP3 LiM belongs to another genus. We propose to name this taxonomic unit "*Candidatus* Vampirococcus".

The analysis of the genome of OP3 LiM revealed features unusual for *Bacteria*. We identified a transcription of 85% of the genes in direction of the DNA replication and only 15% of the genes along the opposite direction. In a large number of *Bacteria* from diverse lineages, strand-specific bias in gene distribution exists to reduce head-on collisions between DNA and RNA polymerases (Rocha, 2002; Mao *et al.*, 2012; Saha *et al.*, 2014) and to minimize transcriptional abortion, replication delay, and mutagenesis (Chen & Zhang, 2013). However, such strong gene strand bias of over 80% is much more important in low G + C *Firmicutes* than in any other taxonomic group (Saha *et al.*, 2014). Interestingly, members of *Firmicutes* were identified as closest neighbors for the OP3 LiM genome by RAST analysis. Highest similarity was predicted to the genome structure of *Thermoanaerobacter tengcongensis* MB4. Being a low G + C *Firmicute*, *Thermoanaerobacter tengcongensis* has almost 87% of genes were identified in the leading strand (Ochman & Santos, 2003). The strong bias was explained with the existence of two different isoforms of the DNA polymerase III alpha subunit, PolC and DnaE, responsible for synthesis of leading and lagging strand (Saha *et al.*, 2014). In other bacteria, DnaE is only responsible for the synthesis of both strands and an average of 58% of genes was identified in their leading strand (Rocha, 2002). The OP3 LiM genome has only an orthologous of DnaE (peg.1001), which identifies it as a unique bacterium with respect to the combination of genome organization and DNA polymerase. One explanation may be the viral-like metabolic activity shift from a low activity as free-living bacterium to the high activity as epibiont. For OP3 LiM, the high organized small genome and the efficient replication and transcription machineries are essential based on their dependence to prey cells and the necessity of a rapid growth phase. Some phages encode typically the majority of their genes on the leading strand (Merrikh *et al.*, 2012) or even all of their genes are located on the same strand and are transcripted in the same direction (Mrázek & Karlin, 1998; Petrovski *et al.*, 2012). The location of the gene for a large multi-enzyme surface protein (4,384 kDa) potentially acting in signal reception and transduction and the degradation of macromolecules, directly after genes involved in DNA replication and genes for a secretion pathway is an efficient genome organization for a predatory epibiont. OP3 draft genomes are fragmented and do not allow an identification of the genome organization. Therefore, the presence of the large multifunctional protein was investigated. Of the 43 conserved domains in the large protein, several other Omnitrophica

genomes have 13 of these domains. Four glycosyltransferases and one glucosidase domains represent the most widely present functions. These enzymes of the degradation of polysaccharides suggest that the ultramicrobacteria of the *candidate phylum* Omnitrophica graze on LPSs. Numerous OP3 draft genomes also code for both ssDNA binding domains as well as the diguanylate cyclase (GGDE) and signal receiver domain, which could indicate signaling as common function of the phylum. Several genomes have protein binding chaperaone, SecA protein, a peptidase as well as the ppGpp synthase/hydrolase domains of the large protein of OP3 LiM.

The genome includes several mobile elements. The 23S rRNA gene of OP3 LiM is interrupted by a group I intron at position 2,061 bp to 2,768 bp possesses a gene for a LAGLIDADG/HNH homing endonuclease. In this study, we visualized the intron in cells of different morphologies indicating the self-splicing character of this intron as well as an occurred intron RNA transfer. Group I introns are distributed widely in nature. In contrast to mitochondria and chloroplasts of lower eukaryotes, they are relatively rare in bacteria (Raghavan *et al.*, 2007). But most bacterial 23S rRNA genes contain conserved target sequences for intron-encoded homing endonucleases (Nesbø & Doolittle, 2003). The intron in the 23S rRNA gene of OP3 LiM is a special feature of the OP3 LiM population. None of published draft genomes of members of *candidate phylum* Omnitrophica contained such an intron in the 23S rRNA sequence. However, it has to be noted that the high degree of conservation among rRNAs usually causes an incomplete assembly of rRNAs in metagenomes. The intron showed similarities to mobile elements of species of other taxonomic groups. The first of three introns in the 23S rRNA gene of a clone of *Thiomargarita sp.* (Genbank acc. no. FR774200) showed highest similarity with an identity of 70%, followed by introns of a *Candidatus* Gloeomargarita lithophora strain (Genbank acc. no. CP017675), of an uncultured gamma proteobacterium (Genbank acc. no. HF954163), and of a *Synechoccus* strain (Genbank acc. no. DQ421380). A pathogenic lifestyle is suggested to contribute to the acquisition of mobile elements. *Simkania negevensis* and *Coxiella burnetii*, both obligate intracellular parasites, contain group I introns, which they might have acquired from their eukaryotic hosts (Nesbø & Doolittle, 2003).

A large number of various inserted genetic elements were identified for a small region of 6,505 bp in length, located from position 648,611 bp to 655,115 bp in the OP3 LiM genome. Both sites are flanked by genes for 23S rRNA intervening sequence (IVS) proteins (122 aa; peg.581 and 209 aa; peg.586) (Fig. 3). Unlike to group I introns, which are located in functionally vital loci, IVSs are found in highly variable regions and not in functionally

essential domains after they are excised from the rRNA by ribonucleases (Everett *et al.*, 1999). A group II intron was found in the genetic element, located between two copies of a gene of unknown function (peg.582 and peg.585) (Fig. 3). Both duplicated genes called CHP1 and CHP2 contained a highly variable region of 83 bp, potentially the insertion sites of the group II intron. An activity of the intron can attribute by the predicted reverse transcriptase gene, a RNA-dependent DNA polymerase (peg.584). A gene of a third 23S rRNA IVS protein (133 aa; peg.583) is also encoded by the group II intron and upstream (Fig. 3), a G homopolymer of 21 bases is additionally identified. Bacterial group II introns are often fragmented and have sizeable extraneous insertions (Dai & Zimmerly, 2002; Darmon & Leach, 2014). The deviation in GC content and GC skew from average of the OP3 LiM chromosome implied the presence of exogenous DNA (Chapter 5, Fig.9).

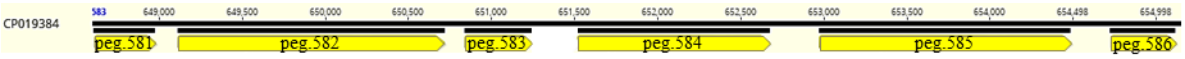


Fig. 3. Representation of the arrangement of the various genetic elements contained in the small region of 6,505 bp in the OP3 LiM genome.

In no other genome of currently described Omnitrophica species, the entire region with such variation of genomic elements was predicted, confirming the acquisition of foreign DNA for OP3 LiM. However, currently published Omnitrophica draft genomes possess a homologue of one copy of the duplicated genes with unknown function (CHP1 and CHP2) suggesting the duplication as result of an insertion. Still, it has to be noted that the absence of a repetitive element in a draft genome is expected as a result of the assembly process. For OP3 LiM, only a manual inspection of the read mapping frequency together with mismatches in the matched reads revealed the existence of the intron. Homologues for all three genes for 23S rRNA IVS proteins (peg.581, peg.583, peg.586) were present in draft genomes of other OP3. Interestingly, both genes located on the sides of the multiple insertion region of the OP3 LiM sequence (peg.581 and peg.586) were only detected in metagenomes obtained from groundwater of the same sample source (project acc. no. PRJNA288027) (Anantharaman *et al.*, 2016). Similarity to the third IVS gene (peg.583) encoded by the group II intron was only found to a gene of a groundwater metagenome obtained from another source of sample (Genbank acc. no. OIO35142). Homologues of the reverse transcriptase gene (peg.584) without the simultaneous detection of CHP1 were not present in genomes of other *candidate phylum* Omnitrophica but in genomes of species of other taxonomic groups. A gene of a

metagenome-derived draft genome of Candidate division WOR-1 had the highest similarity (Genbank acc. no. OGC23312).

6.6 Methanogenic community

6.6.1 Composition of the microbial community

Different groups of anaerobic microorganisms decompose organic matter in a series of steps in sulfate-reducing or methanogenic habitats. Methanogenesis is typically for anoxic freshwater environments producing methane and carbon dioxide as terminal products (Ito *et al.*, 2011; Pasche *et al.*, 2011). The biological methane production is usually dependent on syntrophic interactions (Stams & Plugge, 2009; Ito *et al.*, 2011). Methanogenic limonene degradation was also considered to depend on syntrophic interactions between the members of the microbial community (Rotaru, 2009). OP3 LiM was abundantly detected in these enrichment cultures. The instability of cultures indicated disturbances in the microbial relationships. However, syntrophy is a tightly regulated interaction between partners. Therefore, we propose that the lack of growth was caused by overwhelming predation by OP3 LiM. Because species of *candidate phylum* Omnitrophica were previously detected in other methanogenic enrichment cultures thriving on different carbon sources (Shigematsu *et al.*, 2006; Tang *et al.*, 2007) and these cultures showed instability (Luo *et al.*, 2016), we characterized the member populations to gain a better understanding of the community and consequently of the role of OP3 LiM as one of them. The mechanism that initiated the limonene degradation in methanogenic consortia was investigated.

The composition of the microbial community was previously determined by 16S rRNA gene clone libraries (Rotaru, 2009). OTUs were affiliated to the bacterial lineages of *Bacteroidetes*, *Deltaproteobacteria*, Candidate Division OP3, and *Firmicutes*. Archaeal 16S rRNA gene sequences were represented by OTUs related to *Methanomicrobiales* and *Methanosarcinalles* (Rotaru *et al.*, 2012). A proportion of 40% *Bacteria* and 33% *Archaea* was identified by CARD-FISH. OP3 LiM cells were detected abundantly (18%) in limonene-degrading enrichments (Rotaru *et al.*, 2012).

Tab. 1. Overview on obtained population genomes of members of the limonene-degrading methanogenic enrichment cultures.

Population genome	Contig no.	Total size (bp)	Largest contig (bp)	N50
Aminiphilus	107	2,509,332	426,441	249,750
Anaerolinea	729	8,190,433	675,216	148,540
Coriobacterineae	34	2,683,747	340,261	122,465
Desulfovibrionaceae[1]	13	3,064,128	825,615	459,888
Marinilabiliaceae	313	2,113,112	256,750	39,198
Mesotoga	187	2,953,278	321,343	62,338
Methanoculleus[1]	79	2,758,179	266,907	56,802
Methanoregula[1]	26	2,827,082	1,289,313	854,002
Methanosaeta[1]	92	2,562,790	130,105	47,079
Methanospirillum[1]	153	3,387,623	221,581	53,058
***Omnitrophica* (OP3 LiM)**[2]	1	1,974,201	1,974,201	1,974,201
Prolixibacteraceae	675	11,653,373	271,462	52,779
Synergistaceae[1]	451	997,826	12,213	2,567
Syntrophaceae[1]	141	3,766,219	359,483	137,984
Syntrophobacteraceae[1]	198	3,704,154	149,536	66,508
Thermoanaerobaculum	125	4,511,195	279,659	67,650
Treponema	38	2,768,450	522,577	257,485

[1] Improved bins produced in cooperation with Gerrit A. Martens as part of his student project.

[2] Closed population genome of OP3 LiM, reconstructed in this thesis.

The metagenomes from size-fractionated cells yielded assembled contigs of 139 Mb, 69 Mb and 42 Mb for Pe1, Pe2 and Pe3, respectively. This corresponded to 8–30 microbial species considering an average microbial genome. Binning of contigs into species-level groups by Metawatt-3.5.2 (Strous *et al.*, 2012) revealed 16 population genomes besides OP3

LiM (Tab. 1). For a comprehensive analysis of the complex syntrophic community, five metagenomic assemblies obtained from 454, MiSeq, Pe1, Pe2 and Pe3 datasets and a metagenome assembled from the combined reads of Pe1, Pe2 and Pe3 were combined to obtain population genomes.

Draft genomes from metagenomic data are incomplete (Sharon & Banfield, 2013). CheckM 0.9.7 (Parks *et al.*, 2015) was used to estimate the completeness and contamination. Selected population genomes were reassembled with taxonomic profiling by Metawatt, read mapping and *de novo* reassembly by SPAdes 3.3.0 (Bankevich *et al.*, 2012) (Tab. 1; Appendix Table S1). Refined population genomes were related to type strains using JSpecies (http://jspecies.ribohost.com, Richter & Rosselló, 2009) (Appendix Table S1).

The enrichment culture included populations of the genera *Methanoregula*, *Methanoculleus*, *Methanosaeta*, *Methanospirillum*, *Syntrophus*, *Syntrophobacter*, *Desulfovibrio*, and *Synergistaceae*. High similarities to published genomes were identified (Appendix Table S1). The population genome classified as *Methanosaeta* represents a species of *Methanosaeta concilii* that was presented by the genome of strain GP6. *Methanospirillum hungatei* JF-1 is the only genome of the family *Methanospirillaceae* (Gunsalus *et al.*, 2016). A population genome affiliated with the genus *Methanospirillum*, but represented a new species. Average nucleotide identity (ANI) values of over 80% indicated the presence of novel species of *Desulfovibrio* and of *Methanoculleus*. *Methanosaeta* was absent in the OP3 LiM enriched metagenome, in contrast to all other population genomes.

These population genomes reflect a syntrophy between limonene degradation and methane generation. The composition of the limonene-degrading, methanogenic enrichment cultures showed high similarities to a long-chain alkane-degrading methanogenic community (Embree *et al.*, 2014; Embree *et al.*, 2015). The conversion of hexadecane required the interaction of syntrophic bacteria with methanogenic archaea. Embree *et al.* (2014) identified methanogens related to *Methanosaeta concilli, Methanoculleus marisnigri* and *Methanocalculus* corresponding to the same species (*Methanosaeta concilli*), genus (*Methanoculleus*) and order of *Methanomicrobiales* (*Methanocalculus, Methanospirillum* and *Methanoregula*) among the population genomes. *Methanosaeta* is the specialist for acetate, whereas the other methanogens are specialists for hydrogen and formate as substrate. *Desulfovibrio sp.* and *Smithella*, a member of the family *Synthrophaceae*, are the syntrophs in the hexadecane enrichment, with *Smithella* as key organisms for the initial steps of alkane activation (Embree *et al.*, 2014). On limonene, members of *Syntrophus sp.*, *Syntrophobacter sp.*, *Desulfovibrio sp.* and *Synergistaceae* are involved in syntrophic processes.

6.6.2 Outlook

This study focused initially on an assembly and annotation of the OP3 LiM genome and the visualizations of interactions between OP3 LiM and community members. On the basis of these results, a complete study with metaproteomes, metagenomes and metatranscriptomes of size-fractionated cells of the enrichment was performed. The results of this study were so far only used to identify the proteome of OP3 LiM and to verify the OP3 LiM genome. The syntrophic relations within the limonene-degrading methanogenic enrichment can be investigated in detail. The study of the so-far unknown limonene degradation pathway is promising for the discovery of novel metabolisms. Metagenomes and metaproteomes will provide insight into the importance of the syntrophic intermediates acetate, hydrogen and formate. The metatranscriptome can complement the metaproteome, as it likely detects more transcribed genes than the metaproteome analysis identifies proteins. For these studies, the population genomes will be starting point and the identification of proteins and transcripts should be genome specific.

References

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

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun 7:13219.

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

Au KF, Underwood JG, Lee L, Wong WH. 2012. Improving PacBio long read accuracy by short read alignment. PLoS One 7:e46679.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.

Bardy SL, Ng SY, Jarrell KF. 2003. Prokaryotic motility structures. Microbiology 149:295–304.

Chanyi RM, Koval SF. 2014. Role of type IV pili in predation by *Bdellovibrio bacteriovorus*. PLoS One 9:e113404.

Chen X, Zhang J. 2013. Why are genes encoded on the lagging strand of the bacterial genome? Genome Biol Evol 5:2436–2439.

Dai L, Zimmerlya S. 2002. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. Nucleic Acids Res 30:1091–1102.

Darmon E, Leach DRF. 2014. Bacterial genome instability. Microbiol Mol Biol Rev 78:1–39.

Dashiff A, Junka RA, Libera M, Kadouri DE. 2011. Predation of human pathogens by the predatory bacteria *Micavibrio aeruginosavorus* and *Bdellovibrio bacteriovorus*. J Appl Microbiol 110:431–444.

Ekblom R, Wolf, JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042.

Embree M, Liu JK, Al-Bassam MM, Zengler K. 2015. Networks of energetic and metabolic interactions define dynamics in microbial communities. Proc Natl Acad Sci U S A 112:15450–15455.

Embree M, Nagarajan H, Movahedi N, Chitsaz H, Zengler K. 2014. Single-cell genome and metatranscriptome sequencing reveal metabolic interactions of an alkane-degrading methanogenic community. ISME J 8:757–767.

Esteve I, Gaju N. 1999. Bacterial symbioses. Predation and mutually beneficial associations. Int Microbiol 2:81–86.

Everett KDE, Kahane S, Bush RB, FriedmanMG. 1999. An unspliced group I Intron in 23S rRNA links *Chlamydiales*, chloroplasts, and mitochondria. J Bacteriol 181:4734–4740.

Fuchs BM, Glöckner FO, Wulf J, Amann R. 2000. Unlabeled helper oligonucleotides increase the *in situ* accessibility to 16S rRNA of fluorescently labeled oligonucleotide probes. Appl Environ Microbiol 66:3603–3607.

Gawad C, Koh W, Quake SR. 2016. Single-cell genome sequencing: current state of the science. Nat Rev Genet 17:175–188.

Glöckner J, Kube M, Shrestha PM, Weber M, Glöckner FO, Reinhardt R, Liesack W. 2010. Phylogenetic diversity and metagenomics of candidate division OP3. Environ Microbiol 12:1218–1229.

Guerrero R, Pedros-Alio C, Esteve I, Mas J, Chase D, Margulis L. 1986. Predatory prokaryotes: predation and primary consumption evolved in bacteria. Proc Natl Acad Sci U S A 83:2138–2142.

Gunsalus RP, Cook LE, Crable B, Rohlin L, McDonald E, Mouttaki H, Sieber JR, Poweleit N, Zhou H, Lapidus AL, Daligault HE, Land M, Gilna P, Ivanova N, Kyrpides N, Culley DE, McInerney MJ. 2016. Complete genome sequence of *Methanospirillum hungatei* type strain JF1. Stand Genomic Sci 11:2.

Harder J, Foss S. 1999. Anaerobic formation of the aromatic hydrocarbon p-cymene from monoterpenes by methanogenic enrichment cultures. Geomicrobiol J 16:295305.

Ito T, Yoshiguchi K, Ariesyady HD, Okabe S. 2011. Identification of a novel acetate-utilizing bacterium belonging to *Synergistes* group 4 in anaerobic digester sludge. ISME J 5:1844–1856.

Ivanova AO, Dedysh SN. 2012. Abundance, diversity, and depth distribution of planctomycetes in acidic northern wetlands. Front Microbiol 3:5.

Jakovljevic V, Leonardy S, Hoppert M, Søgaard-Andersen L. 2008. PilB and PilT are ATPases acting antagonistically in type IV pilus function in *Myxococcus xanthus*. J Bacteriol 190:2411–2421.

Jurkevitch E, Davidov Y. 2006. Phylogenetic diversity and evolution of predatory prokaryotes. p 11–56. *In* Jurkevitch E (ed), Predatory prokaryotes. Springer, Berlin Heidelberg.

Kalyuzhnaya MG, Zabinsky R, Bowerman S, Baker DR, Lidstrom ME, Chistoserdova L. 2006. Fluorescence *in situ* hybridization-flow cytometry-cell sorting-based method for separation and enrichment of type I and type II methanotroph populations. Appl Environ Microbiol 72:4293–4301.

Kato S, Haruta S, Cui ZJ, Ishii M, Igarashi Y. 2005. Stable coexistence of five bacterial strains as a cellulose-degrading community. Appl Environ Microbiol 71:7099–7106.

Klumpp J, Fouts DE, Sozhamannan S. 2012. Next generation sequencing technologies and the changing landscape of phage genomics. Bacteriophage 2:190–199.

Köhler R, Schäfer K, Müller S, Vignon G, Diederichs K, Philippsen A, Ringler P, Pugsley AP, Engel A, Welte W. 2004. Structure and assembly of the pseudopilin PulG. Mol Microbiol 54:647–664.

Kolinko S, Richter M, Glöckner FO, Brachmann A, Schüler D. 2016. Single-cell genomics of uncultivated deep-branching magnetotactic bacteria reveals a conserved set of magnetosome genes. Environ Microbiol 18:21–37.

Kolinko S, Jogler C, Katzmann E, Wanner G, Peplies J, Schüler D. 2012. Single-cell analysis reveals a novel uncultivated magnetotactic bacterium within the candidate division OP3. Environ Microbiol 14:1709-1721

Konstantinidis KT, Rosselló-Móra R. 2015. Classifying the uncultivated microbial majority: A place for metagenomic data in the *Candidatus* proposal. Syst Appl Microbiol 38:223–230.

Koval SF, Hynes SH, Flannagan RS, Pasternak Z, Davidov Y, Jurkevitch E. 2013. *Bdellovibrio exovorus* sp. nov., a novel predator of *Caulobacter crescentus*. Int J Syst Evol Microbiol 63:146–151.

Kumar MR, Saravanan VS. 2010. Candidate OP phyla: importance, ecology and cultivation prospects. Indian J Microbiol 50:474–477.

Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. 2008. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev 72:557–578.

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

Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Davies MR, Steer AC, Tong SY, Honkela A, Parkhill J, Bentley SD, Corander J. 2016. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun 7:12797.

Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC, Singh A, Williams KH, Siegerist CE, Tringe SG, Downing KH, Comolli LR, Banfield JF. 2015. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun 6:6372.

Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. 2012. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One 7:e30087.

Luo F, Devine CE, Edwards EA. 2016. Cultivating microbial dark matter in benzene-degrading methanogenic consortia. Environ Microbiol 18:2923–2936.

Mahmoud KK, Koval SF. 2010. Characterization of type IV pili in the life cycle of the predator bacterium *Bdellovibrio*. Microbiology 156:1040–1051.

Mao X, Zhang H, Yin Y, Xu Y. 2012. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. Nucleic Acids Res 40:8210–8218.

Mardis E, McPherson J, Martienssen R, Wilson RK, McCombie WR. 2017. What is finished, and why does it matter. Genome Res 12:669–671.

Merrikh H, Zhang Y, Grossman AD, Wang JD. 2012. Replication-transcription conflicts in bacteria. Nat Rev Microbiol 10: 449–458.

Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Pühler A. 2003. GenDB–an open source genome annotation system for prokaryote genomes. Nucleic Acids Res 31:2187–2195.

Molari M, Manini E. 2012. Reliability of CARD-FISH procedure for enumeration of *Archaea* in deep-sea surficial sediments. Curr Microbiol 64:242–250.

Morrison JM, Baker KD, Zamor RM, Nikolai S, Elshahed MS, Youssef NH. 2017. Spatiotemporal analysis of microbial community dynamics during seasonal stratification events in a freshwater lake (Grand Lake, OK, USA). PLoS One 12:e0177488.

Mrázek J, Karlin S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci U S A 95:3720–3725.

Nesbø CL, Doolittle WF. 2003. Active self-splicing group I introns in 23S rRNA genes of hyperthermophilic bacteria, derived from introns in eukaryotic organelles. Proc Natl Acad Sci U S A 100:10806–10811.

Ochman H, Santos SR. 2003. Eyeing bacterial genomes. Curr Opin Microbiol 6:109–113.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

Pasche N, Schmid M, Vazquez F, Schubert CJ, Wuest A, Kessler JD et al. 2011. Methane sources and sinks in Lake Kivu. J Geophys Res Biogeosci 116:G03006.

Pasternak Z, Njagi M, Shani Y, Chanyi R, Rotem O, Lurie-Weinberger MN, Koval S, Pietrokovski S, Gophna U, Jurkevitch E. 2014. In and out: an analysis of epibiotic vs periplasmic bacterial predators. ISME J 8:625–635.

Peabody CR, Chung YJ, Yen MR, Vidal-Ingigliardi D, Pugsley AP, Saier MH Jr. 2003. Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. Microbiology 149:3051–3072.

Pérez J, Moraleda-Muñoz A, Marcos-Torres FJ, Muñoz-Dorado J. 2016. Bacterial predation: 75 years and counting! Environ Microbiol 18:766–779.

Petrovski S, Dyson ZA, Seviour RJ, Tillett D. 2012. Small but sufficient: the *Rhodococcus* Phage RRH1 has the smallest known siphoviridae genome at 14.2 kilobases. J Virol 86:358-363.

Pisapia C, Gérard E, Gérard M, Lecourt L, Lang SQ, Pelletier B, Payri CE, Monnin C, Guentas L, Postec A, Quéméneur M, Erauso G, Ménez B. 2017. Mineralizing filamentous bacteria from the Prony Bay Hydrothermal Field give new insights into the functioning of serpentinization-based subseafloor ecosystems. Front Microbiol 8:57.

Raghavan R, Miller SR, Hicks LD, Minnick MF. 2007. The unusual 23S rRNA gene of Coxiella burnetii: two self-splicing group I introns flank a 34-base-pair exon, and one element lacks the canonical omegaG. J Bacteriol 189:6572–6579.

Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci USA 106:19126–19131.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu WT, Eisen JA,. Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437.

Rocha E. 2002. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol 10:393–395.

Rotaru AE. 2009. Anaerobic degradation of limonene and p-xylene in freshwater enrichment cultures. PhD thesis, Universität Bremen, Germany.

Rotaru AE, Schauer R, Probian C, Mussmann M, Harder J. 2012. Visualization of candidate division OP3 cocci in limonene-degrading methanogenic cultures. J Microbiol Biotechnol 22:457–461.

Saha SK, Goswami A, Dutta C. 2014. Association of purine asymmetry, strand-biased gene distribution and PolC within firmicutes and beyond: a new appraisal. BMC Genomics 15:430.

Sangwan N, Xia F, Gilbert JA. 2016. Recovering complete and draft population genomes from metagenome datasets. Microbiome 4:8.

Schulz GE. 2002. The structure of bacterial outer membrane proteins. Biochim Biophys Acta 1565:308–317.

Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. 2013. Computational meta'omics for microbial community studies. Mol Syst Biol 9:666.

Seitz P, Blokesch M. 2013. DNA-uptake machinery of naturally competent *Vibrio cholerae*. Proc Natl Acad Sci U S A 110:17987–17992.

Sharon I, Banfield JF. 2013. Microbiology. Genomes from metagenomics. Science 342:1057–1058.

Shigematsu T, Era S, Mizuno Y, Ninomiya K, Kamegawa Y, Morimura S, Kida K. 2006. Microbial community of a mesophilic propionate-degrading methanogenic consortium in chemostat cultivation analyzed based on 16S rRNA and acetate kinase genes. Appl Microbiol Biotechnol 72:401–415.

Speth DR, In ´t Zandt MH, Guerrero-Cruz S, Dutilh BE, Jetten MSM. 2016. Genome-based microbial ecology of anammox granules in a full-scale wastewater treatment system. Nat Commun 7:11172.

Stams AJM, Plugge CM. 2009. Electron transfer in syntrophic communities of anaerobic bacteria and archaea. Nat Rev Microbiol 7:568–577.

Stothard P. 2000. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28:1102–1104.

Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. 2012. The binning of metagenomic contigs for microbial physiology of mixed cultures. Front Microbiol 3:410.

Tamayo R, Pratt JT, Camilli A. 2007 Roles of cyclic diguanylate in the regulation of bacterial pathogenesis. Annu Rev Microbiol 61:131–148.

Tang YQ, Shigematsu T, Morimura S, Kida K. 2007. Effect of dilution rate on the microbial structure of a mesophilic butyrate-degrading methanogenic community during continuous cultivation. Appl Microbiol Biotechnol 75:451–465.

Tsai Y-C, Conlan S, Deming C, NISC Comparative Sequencing Program, Segreb JA, Kong HH, Korlach J, Oh J. 2016. Resolving the complexity of human skin metagenomes using single-molecule sequencing. MBio 7:e01948–15.

Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, Brown SD. 2014. Evaluation and validation of *de novo* and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics 30:2709–2716.

Wang Z, Kadouri DE, Wu M. 2011. Genomic insights into an obligate epibiotic bacterial predator: *Micavibrio aeruginosavorus* ARL-13. BMC Genomics 12:453.

Wences AH, Schatz MC. 2015. Metassembler: merging and optimizing *de novo* genome assemblies. Genome Biol 16:207.

Whitney JC, Hay ID, Li C, Eckford PD, Robinson H, Amaya MF, Wood LF, Ohman DE, Bear CE, Rehm BH, Howell PL. 2011. Structural basis for alginate secretion across the bacterial outer membrane. Proc Natl Acad Sci U S A 108:13083–13088.

Yamaguchi T, Kawakami S, Hatamoto M, Imachi H, Takahashi M, Araki N, Yamaguchi T, Kubota K. 2015. *In situ* DNA-hybridization chain reaction (HCR): a facilitated in situ HCR system for the detection of environmental microorganisms. Environ Microbiol 17:2532–2541.

Yan J, Mao J, Xie J. 2014. Bacteriophage polysaccharide depolymerases and biomedical applications. BioDrugs 28:265–274.

Yin Y, Mao X, Yang JC, Chen X, Mao F and Xu Y. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40:W445–451.

Appendix of Chapter 6:

Supplementary table

Table S1. Overview on obtained syntrophic-relevant population genomes (PGs) from the limonene-degrading, methanogenic enrichment cultures. Statistics of the highest-quality assembly of each of the PGs, quality parameters, closed related species of the PGs based on measured probabilities (ANI, TETRA) are shown.

Population genome	Contig no.	Total size (bp)	Completeness (%)	Contamination (%)	Closed related species[1]	ANIm (%)[2]	TETRA[2]
Desulfovibrionaceae	13	3,064,128	100.00	0.93	*Desulfovibrio aespoeensis* Aspo-2	82.53	0.8633
Methanoculleus	79	2,758,179	93.38	2.68	*Methanoculleus bourgensis* MS2T	81.94	0.97136
Methanoregula	26	2,827,082	98.14	3.04	*Methanoregula formicica* SMSP	89.76	0.98188
Methanosaeta	92	2,562,790	100.00	2.47	*Methanosaeta concilii* GP6	98.83	0.99808
Methanospirillum	153	3,387,623	99.44	1.53	*Methanospirillum hungatei* JF-1	85.36	0.97166
Synergistaceae	451	997,826	69.57	2.91	*Synergistaceae*[3]		
Syntrophaceae	141	3,766,219	98.41	3.97	*Syntrophus aciditrophicus* SB	86.86	0.80916
Syntrophobacteraceae	198	3,704,154	98.24	3.37	*Syntrophobacter fumaroxidans* MPOB	94.85	0.85987

[1] Reference genome with highest similarity to PGs.

[2] Calculations of average nucleotide identities (ANIs) and Tetranucleotide signatures (TETRA) were implemented in the JSpecies software (Richter and Rosselló, 2009). Based on these results, species of highest similarity to each PG was determined.

[3] Due to the low completeness, no conclusions to related species could be drawn.

Acknowledgements

There are a number of people who supported me during the time of my PhD and whom I am very grateful.

First, I would like to thank my supervisor Prof. Dr. Jens Harder to give me the opportunity to write this thesis. This fascinating project enabled me to learn new skills and techniques. I was able to broaden my horizons. Thank you for answer my questions as well as for giving me helpful suggestions exceptionally quick during my laboratory time if I needed advice.

I would like to express my gratitude to Prof. Dr. Rudolf Amann, for kindly reviewing my thesis.

Thank you to Prof. Dr. Ulrich Fischer and Prof. Dr. Karl-Heinz Blotevogel for kindly serving as members of my defense committee on quite short notice. Also, thank you Prof. Dr. Ulrich Fischer for all the great advice.

Nina Heinzmann and David Benito Merino, thanks for kindly serving as student representatives of the examining board.

Thank you to Dr. Katrin Knittel, Dr. Michael Richter and Prof. Dr. Rita Groß-Hardt for being supportive members of my thesis committee meetings and for your interest in my topic.

I want to thank all members and especially the technicians of the Department of Microbiology who create a pleasant working atmosphere. Thanks for all your support during my PhD life.

Special thanks go to Christina Probian for showing me the basic techniques in the lab as well as for your support in the lab and especially, during my final PhD phase. You gave me advices and motivation. I will never forget it. I also thank you for our nice conversations and walks.

Without the great support of many colleagues of other departments the thesis would not have been as successful!
I would like to thank Sten Littmann, Daniela Tienken and Swantje Lilienthal for your warm help to create wonderful scanning electron images.
I also thank Jörg Wulf for helpful suggestions to optimize my CARD-FISH images.
I want to express my special thanks to Dimitri Meier and Christian Quast for your open door to my bioinformatics questions.
Kin Ovanesov and Greta Reintjes, I thank you for answer the general questions that I had during my final PhD phase.
I thank the IT department for your excellent technical support.

During my years in the Department of Microbiology a lot of students started their internship by us. After few months, they left the institute. I would like to thank Martin Streuber for an unforgettable time. To this day we are friends and I have to thank you for always being there for me and that you bring back my smile.

I am deeply grateful of a lot of people for the moral support during my final PhD phase. Thank you for bringing back my optimism.

Mein größter Dank gilt meinen Eltern und meiner Schwester Anne. Ihr standet während dieser Zeit immer an meiner Seite und habt mir damit sehr viel Kraft geschenkt. Nur durch euren Zuspruch und eurer Unterstützung bin ich so weit gekommen. Lieben Dank!

Name: _____ Ort, Datum: _____

Anschrift: _____

ERKLÄRUNG

Hiermit erkläre ich, dass ich die Doktorarbeit mit dem Titel:

"Insights into the biology of Candidate Division OP3 LiM populations"

selbstständig verfasst und geschrieben habe und außer den angegebenen Quellen keine weiteren Hilfsmittel verwendet habe.

Ebenfalls erkläre ich hiermit, dass es sich bei den von mir angegebenen Arbeiten um drei identische Exemplare handelt.

_____
(Unterschrift)