REKONSTRUKTION NEUER *CHLOROFLEXI*-METAGENOME AUS KONTAMINIERTEM GRUNDWASSER

Bachelorarbeit

von

Dominik Hellmann

Matrikelnr.: 1959746

Fakultät für Chemie und Biowissenschaften

Institut für Biologische Grenzflächen 5

Gutachter:	Prof. Dr. Anne-Kristin Kaster
Betreuender Mitarbeiter:	Dr. John Vollmer

30. September 2019

Inhaltsverzeichnis

Abkürzungsverzeichnis

16S rRNA	16S ribosomalen RNA
antiSMASH	*antibiotics & Secondary Metabolite Analysis Shell*
RAM	Arbeitsspeicher
BWA	*Burrow-Wheeler-Aligner*
COGs	*Cluster of Orthologous Groups*
CDS	*coding sequences*
Contigs	*Contiguous Sequences*
DIAMOND	*double index alignment of next-generation sequencing data*
Gb	Gigabasen
HCC	*Hirarchical Contig Classification*
KMGW	kanadisches Mülldeponie-Grundwassermetagenom
kb	Kilobasen
lsu	*large subunit*
LCA	*Lowest common Ancestor*
Mb	Megabasen
MetaBAT	*Metagenome Binning with Abundance and Tetra-nucleotide Frequencies*
MG-RAST	*Metagenomic Rapid Annotations using Subsystems Technology*
MIG	*Minimum information for genome data*
NCBI	*National Center for Biotechnology Information*
NGS	*Next Generation Sequencing*
NRPS/PKS	*Non-ribosomal peptide synthetase cluster/Polyketidesynthase*
ORFs	*Open Reading Frames*
CPUs	*central processing unit* (Rechenkerne)
SRA	*Sequence Read Archive*
SINA	*SILVA Incremental Aligner*
ssu	*small subunit*
TOGW	texanischen Öl-kontaminierten Grundwassermetagenom

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Ziel der vorliegenden Arbeit war die Rekonstruktion neuer, bislang nicht-kultivierter Bakterienvertreter aus Metagenomen. Hierbei wurde sich der stetige Anstieg an öffentlich verfügbaren Sequenzdaten zu Nutze gemacht, indem mehrere unabhängige und vergleichbare Metagenomdatensätze kombiniert wurden. Hierdurch wurde insgesamt der Informationsgehalt, insbesondere differentielle *Coverage*-Informationen für möglichst effiziente Genomrekonstruktionen, maximiert.

Der Fokus lag hier auf dem Phylum *Chloroflexi*, ein auch in extremen Habitaten ubiquitäres Phylum, dessen Diversität jedoch noch nicht ausreichend durch kultivierte Vertreter abgedeckt ist. Nach Durchsicht aller verfügbaren Metagenomdatensätze wurden exemplarisch Schadstoff-belastete Grundwassermetagenome ausgewählt. Einige dieser Metagenome wiesen erhöhte *Chloroflexi*-Anteile auf. Da *Chloroflexi* aus Grundwasserproben bisher wenig beschrieben wurden, deutete dies auf ein hohes Potential hin, bislang unbekannte Vertreter zu finden.

Insgesamt konnten 1222 *Bins*, also potentielle partielle bakterielle Genome, rekonstruiert werden. Hiervon konnten 32 dem Phylum *Chloroflexi* zugeordnet werden, darunter waren elf hochqualitative Genome mit einer geschätzten Genomvollständigkeit von 60-98 % und potentiellen Kontaminationswerten von maximal 10 %. Diese stellten allesamt Vertreter neuer Spezies, zum größten Teil aber neuer Genera, Familien und sogar potentiell neuer Klassen dar.
Durch funktionelle Analysen konnten erste Erkenntnisse über das metabolische Potential dieser Vertreter gewonnen werden, wodurch sich bereits verschiedene prägnante Eigenschaften herausstellten. Dadurch konnten einzelne Vertreter voneinander bzw. von den jeweiligen nächstverwandten beschriebenen Referenzen oder von der übrigen Bakteriengemeinschaft der Grundwasserproben abgegrenzt werden. Beispiele hierfür wären der auffallend hohe Anteil an Genen des Lipidstoffwechsels in einem der rekonstruierten Genome oder der signifikant erhöhte Anteil an Genen für den Transport und Metabolismus anorganischer Ionen in sämtlichen rekonstruierten *Chloroflexi*-Genomen. Interessant ist auch, dass ausnahmslos alle hier rekonstruierten *Chloroflexi*-Genome das Potential zur Sekundärmetabolit-Synthese in Form von PKS/NRPS-Genclustern aufweisen.

Die hier gewonnen Erkenntnisse ergänzen und vertiefen unser Wissen über die Diversität und das metabolische Potential des Phylums *Chloroflexi*. Es steht zu hoffen, dass diese Erkenntnisse auch dazu genutzt werden können, um speziell angepasste Kultivierungsbedingungen zu entwickeln, die es eventuell ermöglichen, diese Vertreter in Zukunft in Reinkultur zu bringen.

1 Einleitung

Bakterien leben in den unterschiedlichsten, oft auch extremsten, Habitaten [1] und stellen dort wesentliche Faktoren der Biomasseproduktion, Nährstoffumsetzung sowie globaler Stoffkreisläufe dar [2]. Um die Lebens- und Anpassungsbedingungen verschiedener Habitate bzw. deren Einflüsse auf die Ökosphäre verstehen zu können, ist es daher unerlässlich auch die entsprechenden mikrobiellen Gemeinschaften zu studieren. Über 99 % der Bakterien gelten jedoch unter den gegenwärtigen Laborbedingungen als nicht kultivierbar. Diese Vielzahl an bisher nicht erforschten Bakterien wird als *Microbial Dark Matter* bezeichnet [3–5]. Zudem sind auch viele existierende Isolate nur schwer anzuziehen, z. B. da diese eine sehr geringe Zellteilungsrate haben oder da die Umweltbedingungen, an die sie angepasst sind, sich nur schwer rekonstruieren lassen. Aus diesen Gründen sind kultivierungsunabhängige Verfahren zur Analyse solcher Bakterien erforderlich.

Eine Methode, um unkultivierte Bakterien zu erfassen, ist Metagenomik. Hierbei werden die gesamten Genomfragmente aus einer Umweltprobe extrahiert, sequenziert und analysiert [6, 7]. Auf diese Weise kann die phylogenetische Zusammensetzung und das metabolische Potential ganzer Organismengemeinschaften beschrieben, verglichen und durch anschließendes „*binning*“ möglicherweise sogar einzelne Genome rekonstruiert werden. Es wurden bereits viele unabhängige Metagenomstudien aus verschiedenen Habitaten durchgeführt, welche zu großen Teilen in öffentlichen Datenbanken verfügbar sind. Für eine möglichst zuverlässige Genomrekonstruktion kann es sinnvoll sein, möglichst viele vergleichbare Datensätze zu integrieren und zu koassemblieren, um so den Informationsgehalt zu maximieren. Ziel dieser Arbeit ist es daher durch „*Data Mining*“ öffentlich verfügbare Metagenome zu vergleichbaren Gruppen zusammenzufassen, um durch Koassemblierung und *Binning* neue und vollständigere bakterielle Genome zu rekonstruieren, als dies eventuell durch Analyse der jeweiligen Einzel-Datensätze möglich gewesen wäre. Aus den gewonnenen Genomdaten können dann eventuell die zugehörigen Stoffwechselwege hergeleitet werden, wodurch optimierte Kultivierungsbedingungen geschaffen werden können. Dadurch könnten die entsprechenden unkultivierten Organismen in Zukunft gezielt in Kultur gebracht werden. Außerdem lassen sich dadurch potentiell neue Erkenntnisse für das bessere Verständnis ökologischer Zusammenhänge oder für medizinische bzw. biotechnologische Anwendungen gewinnen [6, 8].

1.1 Biologische Grundlagen

1.1.1 METAGENOMIK

Als Metagenom wird die Gesamtheit der DNA aller Lebewesen in einer Probe bezeichnet. Diese kann z. B. kloniert oder direkt sequenziert werden [6]. Mit Hilfe metagenomischer Methoden wird z. B. das Problem der unkultivierbaren Bakterien umgangen [1].

Bevor *Next Generation Sequencing* (*NGS*) - Technologien [9] in der Metagenomik etabliert wurden, wurden Genfragmente in Wirtsorganismen, wie z. B. *E. coli* kloniert (sogenannte Cosmid- und Fosmid-Banken) [10]. Aufgrund gentechnischer Sicherheitsbestimmungen brachte diese Methode starke Einschränkungen mit sich. Darüber hinaus konnte unterschiedliche Klonierungs-Effizienz verschiedener gencodierender Fragmente, sogenanntes „*Cloning Bias*“, teilweise die Ergebnisse beeinflussen [11–13]. Moderne Sequenziertechnologien erlauben nun aber die mehr oder weniger direkte Sequenzierung von Genomfragmenten ohne vorherige Klonierung. Im Vergleich zur Sanger-Sequenzierung sind die heutigen *NGS*-Methoden kostengünstiger, durchsatzstärker und benötigen weniger Probenmaterial, wodurch die für repräsentative Analysen notwendigen Sequenziertiefen erreicht werden können [14].

Alternativ werden in Form sogenannter Amplikons häufig nur ausgewählte Markergensequenzen wie die der 16S ribosomalen RNA (16S rRNA) einer Umweltprobe amplifiziert und sequenziert. Solche rRNA-Gensequenzen besitzen stark konservierte und variable Bereiche und dienen als phylogenetische Marker. Dadurch kann eine zuverlässige phylogenetische Einordnung und taxonomische Klassifizierung der Mitglieder einer Bakteriengemeinschaft, jedoch keine Bewertung des metabolischen Potentials erfolgen [15].

1.1.2 *CHLOROFLEXI*

Vertreter des Phylums *Chloroflexi* treten nahezu ubiquitär weltweit auf und sind teilweise auch an extremen Bedingungen, wie z. B. an extrem heiße oder kalte Habitate angepasst. Jedoch sind *Chloroflexi* auch in Schadstoff-belasteten Umgebungen, wie z. B. Industrie, Dünnung, Pestiziden oder anderen Umweltverschmutzungen zu finden.

Eine erwähnenswerte Eigenschaft des Phylums *Chloroflexi* sind die offenbar teilweise unterschiedlich aufgebauten Zellhüllen (Zellwand und Membran), wodurch bei sogenannter Gram-Färbung [16] manche Vertreter positive (charakteristisch für *monoderme* Bakterien) und manche negative (charakteristisch für *diderme* Bakterien) Ergebnisse zeigen [17].

Das Phylum der *Chloroflexi* ist mannigfaltig. Ihm gehören sauerstoffarme photoautotrophe, aerobe chemoheterotrophe, thermophile sowie anaerobe Bakterien an [18]. *Chloroflexi* sind ubiquitär, jedoch noch wenig erforscht. Der heutige Stand der Forschung unterteilt *Chloroflexi* in acht Klassen: *Chloroflexia* [18,19], *Thermomicrobia* [20], *Dehalococcoidia* [21,22], *Ktedonobacteria* [23,24], *Ardenticatenia* [25], *Thermoflexia* [26], *Anaerolineae* [27] und

Caldilineae [28]. Eine neunte beschriebene Klasse, aus welcher jedoch bisher noch kein Vertreter in Reinkultur kultiviert wurde, ist *Candidatus Thermofonsia* [29, 30] (Abbildung 1).
Manche *Chloroflexi*, beispielsweise viele Vertreter der Klasse *Dehalococcoidia,* können ihre Energie durch reduktive Dehalogenierung organischer chlorierter Verbindungen gewinnen [18]. Neben *Dehalococcoidia* sind auch *Anaerolineae* eine in vielen verschiedenen Habitaten weit verbreitete Klasse mit relativ vielen kultivierten Vertretern der repräsentativen Familie *Anaerolineaceae* [31–33]. Nennenswerte Vertreter dieser Familie sind *Pelolinea submarina* und *Longilinea arvoryzae*. Beide Spezies sind Gram-negativ, filamentös, unbeweglich und nicht sporenbildend [34]. Allerdings wächst *L. arvoryzae* lediglich unter strikt anaeroben Bedingungen [28]. Eine weitere besonders interessante Klasse sind die *Ktedonobacteria,* welche häufig in thermophilen Habitaten anzufinden sind, auffällig große Genome und ein hohes Potential für Sekundärmetabolitbildung besitzen [24]. Der erste beschriebene Vertreter und somit Typstamm dieser Gruppe ist *Ktedonobacter racemifer* ein filamentöses, aerobes, nicht-motiles, mesophiles und gram-positives heterotrophes Bakterium [23].

Es werden noch weitere Gattungen, Familien, Ordnungen und Klassen dieses vielfältigen Phylums vermutet, welche jedoch noch nicht isoliert und daher nicht analysiert werden konnten. Diese Arbeit soll hier teilweise einen Ansatz liefern diese große Wissenslücke zu schließen und somit zur Vervollständigung des phylogenetischen Stammbaums beizutragen (Abbildung 1).

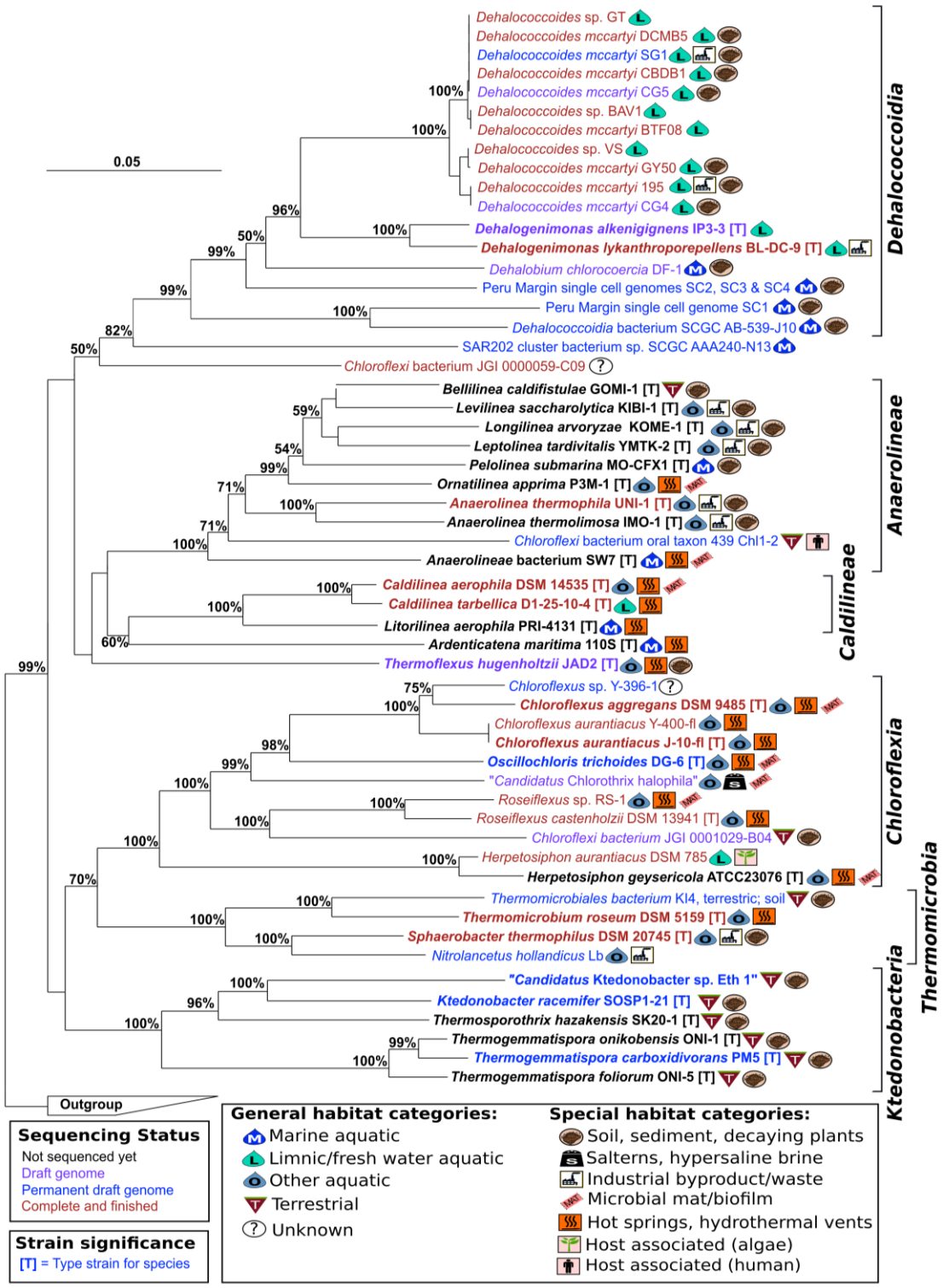

Abbildung 1. 16S rRNA Gensequenz-Stammbaum des Phylum *Chloroflexi*. *Chloroflexi* werden in folgende acht Klassen eingeteilt: *Chloroflexia* [18,19], *Thermomicrobia* [20], *Dehalococcoidia* [21,22] , *Ktedonobacteria* [23,24], *Ardenticatenia* [25], *Thermoflexia* (nur durch *Thermoflexus hugenholzii* vertreten) [26] , *Anaerolineae* (nur durch *Ardenticatena maritima* vertreten) [27] und *Caldilineae* [28] . Eine neunte Klasse, welche bisher nicht kultiviert wurde, ist *Candidatus Thermofonsia* [29,30]. Der Stammbaum wurde mittels *Neighbor-Joining* berechnet. Prozentzahlen und Knotenpunkte stellen *Bootstraps*-Konfidenzwerte basierend auf 1000 Permutationen dar. *Bootstraps*-Werte unter 50 % wurden für die Übersichtlichkeit entfernt. Als *Outgroup* diente *Escherichia coli* k12 und *Bacillus licheniformis* DSM 13. (DFG Antrag, Kaster)

1.2 Bioinformatische Grundlagen

1.2.1 DATENBANKEN

Mit fortschreitenden Sequenziertechniken werden immer mehr Daten gewonnen. Diese Datenmenge stellt in zunehmendem Maße einen Flaschenhals dar, der ohne spezialisierte informatische Techniken kaum zu bewältigen ist. Die gewonnen Daten werden oftmals in Datenbanken eingespeist, damit diese öffentlich zugänglich sind.

Vertreter öffentlich zugänglicher Datenbanken sind das *National Center for Biotechnology Information (NCBI)* [35] und *Metagenomic Rapid Annotations using Subsystems Technology (MG-RAST)* [36]. Bei *NCBI* ist ein Mindestmaß an Metainformationen zwingend dem Datensatz beizufügen [37]. Allerdings ändert sich die genaue Zusammensetzung der geforderten Metadaten stetig. *MG-RAST* hingegen lässt teilweise unzureichend oder fehlerhaft beschriebene Datensätze zu. Ein Vorteil ist hier, dass *MG-RAST* im Gegensatz zu *NCBI* nicht nur zur reinen Datenspeicherung dient, sondern auch integrierte Analysepipelines beinhaltet und somit bereits Schätzungen über die mögliche taxonomische Zusammensetzung der hier hochgeladenen Metagenome zur Verfügung stellt [35]. Es ist jedoch anzumerken, dass diese Information bei *MG-RAST* nicht auf Markergenen, sondern auf der Klassifikation aller Genomfragmente, teilweise auch unassemblierter *Reads*, basiert [36]. Das bedeutet, bei *MG-RAST* können taxonomische Klassifikationen auf schwach konservierten und daher wenig zuverlässigen Genbereichen beruhen. Gerade bei unassemblierten Datensätzen können die so geschätzten Anteile verschiedener Taxa durch Unterschiede in den Genomgrößen stark verzerrt sein. Große Genome liefern mehr *Reads*, werden also mit dieser Methode stärker repräsentiert als kleine Genome, selbst wenn diese mit gleicher Häufigkeit auftreten.
Für die Hinterlegung von Metainformationen eines Datensatzes gibt es mittlerweile eine allgemeine Richtlinie, die sogenannte „*Minimum information for genome data“* (MIG) [38]. Dennoch ist es notwendig alle Datensätze vor weiterführenden Analysen auf Konsistenz zu prüfen, da diese Richtlinie häufig nur unvollständig umgesetzt wird.

1.2.2 DATENPROZESSIERUNG

Ziel der hier angewandten Prozessierung von Metagenomdaten ist zum einen die Bestimmung der genauen phylogenetischen Zusammensetzung der zugehörigen Bakteriengemeinschaften, zum anderen die Bestimmung der genomischen Zusammensetzung einzelner Vertreter dieser Gemeinschaften. Dadurch werden spezifische Informationen über das jeweilige Bakterium, wie z. B. das Vorhandensein bestimmter Schlüssel-Enzyme und Kenntnisse über die jeweiligen metabolischen Eigenschaften erhalten. Nachdem die Datensätze ausgewählt wurden, beginnt die eigentlich Bioinformatik. Zuerst werden einzelne Sequenzen (engl. *Reads*) zu längeren

Genfragmenten (engl. *Contiguous Sequences* bzw. *Contigs*) assembliert. Durch sogenanntes *Mapping* wird die *Contig*-Abdeckung, durch die entsprechenden ursprünglichen *Reads*, also die „*Coverage“* bestimmt [39]. *Coverage*-Unterschiede können als Indikator der relativen *Contig*-Abundanzen dienen. *Contigs* können anschließend unter Einbezug von Sequenzsignaturen und der *Coverage*-Information evtl. sogar einzelnen Genomen bestimmter Spezies zugeordnet werden (*Binning*) [40]. Mit den erhaltenen *Bins* kann danach eine Analyse bislang unkultivierter Organismen ermöglicht werden. Alternativ oder parallel zum *Binning* kann eine Markergen-basierte (z. B. 16S rRNA) Genanalyse durchgeführt werden. Dies ermöglicht eine allgemeine Beschreibung der Zusammensetzung der Bakteriengemeinschaft, da sich vermutlich nicht jedes Genom aller Gemeinschaftsmitglieder vollständig rekonstruieren und sich aufgrund hochkonservierter Bereiche nicht jedes Markergen qualitativ hochwertig *binnen* lässt.
Um Proteine vorherzusagen wird die Assemblierung auf gültige *Open Reading Frames* (*ORFs*) untersucht und daraus tatsächlich Protein-kodierende Gensequenzen (engl. *coding sequences*, CDS) bestimmt.
Dies geschieht mit Zuordnung bekannter Markergene von entsprechend nah verwandten Vertretern aus einer Referenzdatenbank. Existiert kein Vertreter, sind lediglich schwache Ähnlichkeiten zu entfernt Verwandten vorhanden, wodurch dann anhand des sogenannten kleinsten gemeinsamen Vorfahren (engl. *Lowest Common Ancestor* bzw. LCA) klassifiziert wird. Verweisen die Markergene auf verschiedene Organismen, existiert dementsprechend kein nah verwandter Vertreter in der Datenbank. Es werden somit die nächstverwandten Vertreter angezeigt und aus diesen wird der kleinste gemeinsame Vorfahre bestimmt, wodurch eine relativ zuverlässige klassifizierte taxonomische Ebene erhalten wird.

Abschließend können aus den ausgewerteten Daten Rückschlüsse auf vorhandene Enzyme, die daraus resultierenden Metabolismen und die phylogenetische Abstammung der Bakterien gezogen werden.

2 Methoden

Um an die gewünschten Daten zu gelangen wurden teilweise eigens geschriebene Python- und BASH-Skripte benutzt. Diese können unter ***https://github.com/DominikHe93/Thesis.git*** [41] abgerufen werden. Die Python-Skripte wurden unter Linux mit Python Version 2.7 geschrieben. Es ist zu empfehlen, die geschriebenen Python-Skripte mit dieser Version auszuführen. Nachfolgend wird die Datenprozessierung vom Auswählen der Datensätze bis hin zur Identifizierung von Sekundärmetaboliten beschrieben.

2.1 Metagenomdaten

Metagenomdaten wurden der *MG-RAST* und der *NCBI* Datenbank entnommen. Dabei handelt es sich um öffentlich zugängliche Datenbanken. Aus diesen wurden geeignete Datensätze aufgrund ihres Informationsgehalts ausgewählt. Datensätze, welche Sequenz-Informationen beinhalten, können in verschiedenen Sequenz-Dateiformaten, wie z. B. im *Fasta*-Format, abgespeichert werden.

2.1.1 SEQUENZ-DATEIFORMATE

2.1.1.1 *Fasta*-Format

Beim *Fasta*-Format [42] handelt es sich um ein reines Text-Format, welches nach folgendem Schema aufgebaut ist: Sequenz-Überschriften (mit einem Größer-als-Zeichen „>“ markiert), beinhalten eine für jede Sequenz spezifische Bezeichnung. Darauf folgen jeweils eine oder mehrere Zeilen der Sequenzdaten. Das *Fasta*-Format wurde für die Handhabung von Rohsequenzdaten mittlerweile vom *Fastq*-Format abgelöst, ist aber dennoch weit verbreitet.

2.1.1.2 *Fastq*-Format

Auch das *Fastq*-Format ist mit einer einheitlichen Struktur rein textbasiert. Der Aufbau besteht immer aus vier Zeilen pro Sequenz: Sequenz-Überschriften, beginnend mit einem „@“-Zeichen, gefolgt von der Sequenzbezeichnung. Darauf folgt jeweils die Sequenz in einer einzelnen Zeile. Die dritte Zeile beginnt stets mit einem „+“- Zeichen, gefolgt von optionalen Kommentaren oder Annotationen. In der vierten Zeile finden sich die korrespondierenden *Quality Scores* zu jeder der in Zeile zwei aufgeführten Basen. Der *Quality Score* gibt an, mit welcher Wahrscheinlichkeit die vorhergesagte korrespondierende Base fehlerhaft ist [43].

2.1.1.3 *BAM*-Format

Das *Bam*-Dateiformat wird speziell für ausgerichtete Sequenzen bis zu einer Größe von 128 Mb genutzt. Diese besteht aus einem *Header* mit allgemeinen Informationen wie Name und Länge, die

Alignments beinhaltet den Namen, die Sequenz, Qualität und *Alignment*-Information des *Reads* [44].

Es existieren noch weitere Formate, wie z. B. das GenBank-Format. Dieser und alle weiteren wurden in dieser Arbeit nicht behandelt.

2.1.2 *MG-RAST* METADATEN

Um die Übersichtsinformationen über die bei *MG-RAST* verfügbaren Metagenomdatensätze zu sammeln und herunterzuladen, wurde das Skript ***mg_rast_id.py*** erstellt. Das Skript greift auf die *Application Programming Interface (API)* [45] von *MG-RAST* zu und erstellt eine Tabelle ***mg_rast_data.tab*** [41] mit folgenden Einträgen:

id, name, biome, collection_date, env_package_type_mixs, feature ,material, country, location, latitude, longitude, prokaryote_count, perc_target_prokaryote, perc_target_total, bp_count_raw, project_metagenomes, sequence_type, seq_meth, seq_method, sequence_count_raw, target_count, total_count, type, project_name, project_id, public, project_description, study_abstract, study_description, study_title

Obwohl theoretisch möglich, wurde an dieser Stelle bewusst noch nicht nach bestimmten *Chloroflexi*-Anteilen gefiltert. Aus den Datensätzen wurden für diese Arbeit zunächst alle Amplikons und Transkriptome entfernt, dann grob in terrestrische und marine Habitate eingeteilt. Die terrestrischen Habitate wurden näher betrachtet, ausgewählt wurden Schadstoff-belastete Grundwassermetagenome mit einem durchschnittlichen *Chloroflexi*-Anteil von 3 % (Tabelle 1).

2.1.3 *NCBI* METADATEN

Nachdem die *MG-RAST*-Datensätze ausgewählt wurden, wurde festgestellt, dass alle Metagenome aus dem gleichen Habitat stammten, wodurch kein Vergleich mit ähnlichen Schadstoff-belasteten Habitaten stattgefunden hätte. Deshalb wurden anschließend komplementäre Metagenome manuell bei *NCBI* über die *NCBI Sequence Read Archive (SRA)* Webseite gesucht. Hierfür wurden die Suchbegriffe *„groundwater“* und *„contaminated“* genutzt, wodurch dann der Datensatz SRX3574179 erhalten wurde.

2.1.4 METAGENOMDATEN

Die Nukleotidsequenzen der *MG-RAST* Metagenome wurden im *Fastq*-Format heruntergeladen (siehe Sequenzformate 2.1.1.2). Der Datensatz der *NCBI*-Datenbank wurde mit dem Shell-Skript ***cmd_fastqdump.sh*** [41] heruntergeladen.

2.2 Assemblierung

Für die Assemblierung der heruntergeladenen Metagenom-Rohsequenzdaten (*Reads*) wurde das dedizierte Metagenomassemblierungsprogramm MEGAHIT [46] genutzt. MEGAHIT bedient sich der *de Brujin* Graph Methode. Bei dieser Methode werden *Reads* zunächst in noch kürzere, überlappende Abschnitte der Länge *k*, den sogenannten *k-meren*, eingeteilt. [47, 48].
Anschließend wird überprüft, welche dieser *k-mere*, sich in ihrer vollen Länge mit Ausnahme einer Base, also um genau *k-1*, überlappen. Da es im DNA-Code nur vier Basen gibt, sind nur vier unterschiedliche Kombinationsmöglichkeiten auf jeder Seite des *k-mer*s möglich, was die Suche nach überlappenden *k-meren* zu einer simplen Text-Suche vereinfacht. Die Überlappungen zwischen verschiedenen *k-meren* werden in Form eines *de Brujin* Graphen festgehalten. Hierbei werden die Überlappungen als Knoten (*Nodes*) im *de Brujin* Graph bezeichnet, die *k-mere* selbst stellen die Kanten (*Edges*) dar. Die Assemblierung besteht also darin, den längsten kontinuierlichen Pfad des Graphen zu ermitteln, der alle Knoten besucht und dabei jede Kante nur genau einmal durchwandert. Dieser Pfad beschreibt die Reihenfolge in der die *k-mere* angeordnet werden müssen, um die ursprüngliche Sequenz zu rekonstruieren. Resultat ist eine längeres zusammenhängendes Sequenzfragment, welches als *Contig* bezeichnet wird [4, 47].
Ein Vorteil von MEGAHIT liegt in der komprimierten Datenstruktur, welche durch den *succinct de Brujin* Graph [48, 49] erfolgt. Somit werden weniger Arbeitsspeicher und Rechenleistung verbraucht. Darüber hinaus ist MEGAHIT in der Lage sowohl *single*- als auch *paired-end Reads* zu akzeptieren [47].

Um MEGAHIT auszuführen wurde folgendes Befehls-Schema genutzt:

```
megahit -out-prefix <out-prefix> -t 8 -m 0.5 --k-min 21 --k-max 99 --k-step 10
-l <forward-reads_1> <reverse-reads_1>, <forward-reads_2> <reverse-reads_2>, […]
```

Mit „*-t* 8“ wurde die Anzahl der zu nutzenden Rechenkerne (CPUs) zugewiesen. Mit „*-m 0,5“* wurde der zu nutzende Arbeitsspeicher (RAM) auf die Hälfte des vorhandenen Arbeitsspeichers begrenzt. *„--k-min“* und *„--k-max“* legten die minimale bzw. maximale *k-mer* Größe von 21 bzw. 99 Basen fest. Mit „*--k-step“* 10 wurde MEGAHIT übergeben, iterativ die *k-mer* Größe in Zehnerschritten zu erhöhen.
Für den *NCBI* Datensatz wurde das Shell-Skript ***cmd_megahit_template_ncbi.sh*** [41] ausgeführt. Hier wurde jedoch die maximale *k-mer* Größe auf 141 erhöht, da die größten *k-mere* 150 Basen groß waren. Bei *MG-RAST* betrug die maximale *k-mer* Länge 100 Basen. Zusätzlich wurde für den *NCBI*-Datensatz die minimale *Contig* Größe auf 750 Basen festgelegt.
Die beiden erhalten Einzelassemblierungen wurden anschließend mit dem Programm Minimus2

koassembliert (*merging*). Dies erfolgte über das *Wrapper*-Skript ***run_minimus2.py*** [41], welches wiederum über das Shell-Skript ***start_minimus_script.sh*** aufgerufen wurde [41]**.**

```
run_minimus2.py -1 <Path_to_1st_Fasta/ 1st_fasta.fa> -2 <Path_to_2nd_Fasta /2nd_fasta.fa>
--minident 97 -o minimus_merged_1
```

Für diesen *merging*-Schritt müssen Referenz- und *Query*-Assemblierung festgelegt werden, welche gerichtet gegeneinander assembliert (ge*merged*) werden sollen (in diesem Fall mittels der Argumente „*-1*“ bzw. „*-2*“). Als Mindestkriterien für die Assemblierung überlappender Fragmente wurde mit „--minident“ eine Mindestübereinstimmung der Sequenz von 97 % und eine Mindestüberlappung von 500 bp (Standardeinstellung des *Wrapper*-Skripts) bestimmt.

2.3 *Mapping*

Anschließend wurde für das *Mapping*, welches die *Coverage* von *Contigs* einer Assemblierung bestimmt, *BamM* verwendet. *BamM* basiert auf dem *Burrow-Wheeler-Aligner (BWA)* [44] und richtet die ursprünglichen *Reads* der Sequenzierung auf der Assemblierung aus. Die entsprechenden *Alignment*-Resultate wurden teilweise mittels der Programmsuite *SAMtools* [50] weiter prozessiert. Der *Aligner* wird mit folgenden Skripten ausgeführt: ***cmd_bamm_make.sh***, ***cmd_samtools_merge.sh*** und ***cmd_bamm_parse.sh*** [41]**.**

```
bamm make -c <pass_1> <pass_2> -p <prefix> -o <out_folder> -k -K -t 4 -d <database>
```

Das erste Skript erstellt *bam* Dateien (siehe Sequenzformate 2.1.1.3), hierfür wurden mit dem Argument „*-c“* die gepaarten *Files* übergeben, mit „*-p“* wurde ein Präfix für die Ausgabedatei definiert, mit „*-o“* wurde ein Ausgabe-Verzeichnis festgelegt, mit „*-d“* wurde die Referenz-Datenbank zugeordnet und mit „*-k -K“* wurde festgelegt, dass Zwischenschritte nach Abschluss nichtgelöscht werden.

Das ***cmd_samtools_merge.sh*** [41] Skript führt die zusammengehörigen *bam*-Dateien zusammen, sortiert diese und indiziert die Dateien eindeutig, indem es nacheinander die *SAMtools* Befehle „*samtools merge*“, „*samtools sort*“ und „*samtools index*“ aufruft.

```
samtools merge -@ 4 -O BAM <output_name> <input_bam> && rm <orginale_datei>
```

Beim ersten Schritt „*samtools merge“* wird mit dem Argument*„-@“* die Anzahl der CPUs und mit „*-O“* das Ausgabeformat und der Name der ausgegebenen Datei festgelegt. Als *Input*-Datei diente die Ausgabe *Alignment*-Datei des vorherigen „*bamm make*“ Schritts, welche anschließend mittels des BASH-Befehls *rm* wieder gelöscht wurde.

Die Sortierung erfolgte mittels *„samtools sort“*.

```
samtools sort -@ 4 -O BAM -o <new_output_name> <new_input_bam> && rm <old_input_BAM>
```

Auch hier wurden mit *„-@“* und *„-O“* die Rechenkerne bzw. das Ausgabeformat festgelegt und nach Abschluss des Prozessierungschrittes, die Eingabedaten mittels des BASH-Befehls „rm“ wieder gelöscht.

Die zusammengeführten Ergebnisse wurden mit dem Befehl *samtools index* indiziert.

```
samtools index <new_input_bam>
```

Anschließend wurde mittels *„bamm parse“* anhand den *bam-Alignmentdateien* die *Coverage* aller *Contigs* berechnet.

```
bamm parse -b < new_input_bam > -c <coverage.tab> -m opmean --max_distance 50 -t 4
```

Hierbei wurden mittels des Arguments „*-t 4*“ wieder vier CPUS zugewiesen, mit „*-b“* die Eingabe-*bam-Alignmentdateien* festegelt, mit „*-c“* die Ergebnis-*Coveragetabelle* benannt und mit „*-m opmean“* festgelegt, dass Basen mit einer *Coverage* außerhalb einer Standardabweichung von ±1 ausgeschlossen werden. Die maximale erlaubte Sequenzunterschiede zwischen *Query* und Referenz wurde mit „- *-max_distance 50“* festgelegt.

2.4 *Binning*

Beim *Binning* wurde versucht, unter Einbezug der *Coverage,* Tetranukleotid-Frequenzen und eventuell vorhandene universelle Markergene, die *Contigs* einzelnen Genomen verschiedener Spezies zuzuordnen.
MetaBAT (*Metagenome Binning with Abundance and Tetra-nucleotide Frequencies*) [51] und *MaxBin* [52] werden als *Binning*-Programme bezeichnet. *MaxBin* nutzt relative *Contig*-Abundanzen (also die *Coverage*-Information des *Mapping*-Schritts) und Tetranukleotid-Frequenzen (also *k-mere* der Länge 4) sowie potentielles Vorhandensein universeller Markergene, um die *Contigs* eines Metagenoms zu „*Bins*“ zu gruppieren und einzuteilen. *MaxBin binnt* die *Contigs* unter Verwendung des sogenannten *Expectation-Maximization (EM)* Algorithmus [52]. Das Programm wurde mit dem Shell-Skript ***cmd_maxbin.sh*** [41] ausgeführt.

```
run_MaxBin.pl -thread 4 -abund_1 [...] -abund_x  -contig ${contig} -out ${output}
```

Es wurden mit „*-thread 4“* 4 Kerne zugewiesen, mit „*-abund“* wurden die jeweiligen Abundanz-Dateien aufgerufen, mit dem Argument „*-contig“* wurden die *Contig-Files* zugewiesen und mit „*-out“* wurde der *Output-File* festgelegt.

Auch *MetaBAT* nutzt Tetranukleotid-Frequenzen, erstellte sich jedoch eigene *Coverage*-Werte. Dabei nutzt *MetaBAT* einen modifizierten *k-medoid clustering* Algorithmus [51]. *MetaBAT* wurde mit dem Skript ***cmd_metabat_binning.sh*** [41] ausgeführt.

```
runMetaBat.sh -t 4 -m 1500 ${assembly} ${sample_1} [...] ${sample_x}
```

Dieses Skript hat einen ähnlichen Aufbau wie das von *MaxBin*. *MetaBAT* wurde mit *„-t 4“* 4 Rechenkernen zugewiesen, mit dem Argument „*-m 1500*“ wurde die Mindestgröße eines *Contigs*, welche für das *Binning* berücksichtigt werden soll, festgelegt. Nachfolgend wurde die assemblierte *Fasta*-Datei und die jeweilige *bam*-Datei an *MetaBAT* übergeben.

2.5 Genomqualität

Anhand von Markergenen bekannter Spezies wurde mithilfe des Programms c*heckM* die Genomqualität überprüft, wodurch die Kontamination eingeschätzt und die *Bins* auf ihre Vollständigkeit geprüft wurden. Darüber hinaus wurde durch das Programm c*heckM* eine taxonomische Zuordnung durchgeführt [53]. Der genaue Aufruf ist im Folgenden schematisch dargestellt und mittels der Skripte ***cmd_checkm_maxbin.sh*** bzw. ***cmd_checkm_metabat.sh*** [41] nachzuvollziehen. „*checkm taxonomy_wf"*, „*checkm lineage_wf" und "checkM tree_qa"* sind Unterprogramme von *CheckM*.

```
checkm taxonomy_wf -t 4 -x .fasta -f <name.tab>  --tab_table <{rank}{taxon}{input}{output}>

checkm lineage_wf -t 4 -x .fasta -<name.tab> --tab_table <{input}> <{output2}>
```

Mit *„taxonomy_wf"* wurde die Taxonomie *Bins* mit Taxonomie-spezifischen Markergenen analysiert.
Der „*lineage workflow*" („*lineage_wf")* analysierte *Bins* auf linienspezifische sowie universelle Markergene und verglich diese mit Referenz-Markergenen bekannter Spezies. Auf diesen Ergebnissen basierend schätzte c*heckM* anschließend die Vollständigkeit und Kontamination der *Bins* ein.

Eine grobe taxonomische Einordnung der *Bins* erfolgt mit dem *CheckM*-Befehl *tree_qa*.

```
checkm tree_qa -o 2 -f <name.tab> --tab_table ./lineage_wf/
```

Mit dem Argument *„-o 2"* wurde der Informationsgehalt, welche in der Ausgabedatei abgespeichert werden soll, festgelegt. Mit *„-f"* wurde die Ausgabedatei benannt und mit *„-tab_table"* wurden die Informationen Tabstopp-separiert abgespeichert.

2.6 Klassifikation

2.6.1 PROTEINVORHERSAGE

Für die Proteinvorhersage wurde *Prodigal* [54] verwendet. *Prodigal* sagt, unter Berücksichtigung der *Codon-Usage* und Sequenzsignaturen, aus allen möglichem offenen Leseraster (engl. *Open Reading Frames* bzw. ORFs) potentielle proteincodierende Sequenzen (engl. *coding sequence* bzw. CDS) voraus. *Prodigal* durchsucht die Assemblierung nach Start- und Stopp-Codons, wodurch mögliche ORFs eingegrenzt werden. Anhand der *Codon-Usage* werden unter den möglichen ORFs die vermutlich tatsächlichen proteincodierende Bereiche (engl. *coding sequence* bzw. CDS) herausgesucht. *Prodigal* wurde mit dem Skript ***cmd_prodigal.sh*** [41] ausgeführt. Dabei wurde die finale Koassemblierung als *Input* an *Prodigal* übergeben. *Prodigal* gab eine Protein-*Fasta*-Datei aus.

```
prodigal -q -a <output_protein.fasta> -i <assembly.fasta > -o /dev/null -p meta
```

Mit dem Argument „*-p meta*“ wurden das Metagenom-Verfahren ausgewählt.

2.6.2 DETEKTION KONSERVIERTER MARKERGENE

Die Protein-*Fasta*-Datei von *Prodigal* wurde mit dem Skript ***cmd_fetchmg_prodigal.sh*** [41] an *fetchMG* übergeben. Bei *fetchMG* handelt es sich um ein Programm, welches aus einer beliebigen Reihe von Eingabe-Proteinsequenzen bestimmte universelle Markergene identifiziert [55]. Aufgerufen wurde das Programm durch folgenden Befehl.

```
fetchMG.pl -p -t 4 -m extraction <protein.fasta> -o fetchmg_results
```

Mit *„-p“* wurden lediglich Proteinsequenzen übergeben, mit *„-t 4“* wurden 4 Rechenkerne zugewiesen. Das Argument *„-m extraction“* griff auf die Protein-*Fasta* zu und mit *„-o“* wurde der Name der Ausgabedatei festgelegt.

FetchMG nutzt hier 42 ausgewählte *Cluster of Orthologous Groups* (COGs) [56]. Dabei handelt es sich um eine Datenbank, die versucht funktionell verwandte Proteinsequenzen zu Gruppen (COGs) zusammenzufassen. *FetchMG* nutzt 42 dieser COGs, welche in der überwiegenden Mehrheit der bekannten Organismen in genau einfacher Kopienzahl vorkommen, als Markergene. Aus den Referenz-Proteinsequenzen der COGs wurden *Hidden-Markov* Modelle erstellt. Mit diesen Modellen ist es möglich neuen Eingabeproteinsequenzen entsprechend zueinander verwandte Proteine zuzuordnen und somit Rückschlüsse auf mögliche Funktionen und Stoffwechsel der Proteine zu ziehen.

2.6.3 AUSRICHTUNG (*ALIGNMENT*) VON PROTEINSEQUENZEN

Mit dem Programm *double index alignment of next-generation sequencing data* (*DIAMOND*) [57] wurden die mit *FetchMG* ermittelte Marker-Proteinsequenzen durch die paarweise Sequenzausrichtung (*Alignment*) mit der *NCBI-nr*-Referenzdatenbank verglichen, um Funktionen bzw. Phylogenie der Proteinsequenzen vorherzusagen. *DIAMOND* ist bis zu 20000 mal schneller als andere *Aligner*, wie z. B. *BLASTX* [57], weshalb dieses Programm ausgewählt wurde. Hierfür wurde das Skript ***cmd_fetchmg_diamond.sh*** [41] geschrieben. Mit „- *-query"* wurde die erhaltene Protein-*Fasta* zugewiesen.

```
diamond blastp --query <Path_to_input_fasta> --db <Path_to_database> --threads 4 --outfmt 6 --out <output_name>
```

Sowohl die Gesamt-Proteine (hierfür wurde das Skript ***cmd_prodigal_diamond.sh*** [41] benötigt), als auch die gezielt daraus extrahierten „Einzelkopie-Markergene" wurden gegen die Referenzdatenbank ausgerichtet.

2.6.4 GENANALYSE UND STAMMBAUMBERECHNUNG

Zur Detektion von rRNA Genen wurde das Programm *RNAmmer* [58] ausgeführt. Dieses basiert ebenfalls auf dem *Hidden-Markov*-Modell. Es wurden potentielle 16S rRNA Gensequenzen auf den Metagenom-*Contigs* gesucht.
Hierbei wurde das Skript ***cmd_rnammer_16S_tempdirmultiproc.sh*** [41] verwendet, welches folgenden Befehl ausführt:

```
rnammer -S bac -m ssu -f <fasta_file>
```

Dem Programm *RNAmmer* wurde mit „*-S bac"* übergeben ausschließlich bakterielle rRNA Sequenzen zu suchen, mit „*-m ssu"* wurde die Suche auf 16S rRNA spezifiziert. Die 16s rRNA gehört der kleinen Untereinheit (*small subunit* bzw. *ssu*) des 70S Ribosoms an, wohingegen die 23S rRNA der großen Untereinheit (engl. *large subunit* bzw. *lsu*) des 70S Ribosoms angehört [59]. Anschließend wurde die Gensequenzen mit dem online basierten Programm „SILVA Incremental Aligner" (SINA) [60] gegen die SILVA-Datenbank [61] *aligned*, verglichen und klassifiziert. Analog wurden die 23S rRNA Gensequenzen mit dem Skript ***cmd_rnammer_23S_tempdirmultiproc.sh*** [41] prozessiert.

```
rnammer -S bac -m lsu -f <fasta_file>
```

Mit *Arb [62]*, einem Phylogenie Programm, wurde ein 16S rRNA phylogenetischer Stammbaum mithilfe der *Neighbour-Joining*-Methode ermittelt. Es wurden 1000 zufällige Permutationen durchgeführt, um für jede Verzweigung anzeigen zu können, wie viele der Permutationen die jeweilige Verzweigung unterstützten.
Darüber hinaus wurde mit RAxML ein weiterer Stammbaum berechnet, hier wurden aufgrund des hohen Rechenaufwands für die *rapid bootstrapping* Methode nur 200 zufällige Permutationen durchgeführt.

2.6.5 TAXONOMISCHE KLASSIFIZIERUNG

2.6.5.1 Hierarchische *Contig*-Klassifizierung

Um Sequenzen möglichst zuverlässig taxonomisch zu klassifizieren wurde nach dem Prinzip des kleinsten gemeinsamen Vorfahren (engl. *Lowest common ancestor* bzw. *LCA*) vorgegangen. Hierfür wurde die *Hirarchical Contig Classification* (HCC) Methode angewandt [63]. Mittels RNAmmer [58] extrahierte 16S und 23S rRNA Sequenzen wurden mithilfe von SINA [60] nach dem LCA-Prinzip klassifiziert. Ebenfalls wurden die Gesamtproteine sowie die Markergenprodukte jeden *Contigs* mithilfe von DIAMOND gegen eine Referenzdatenbank (*NCBI*-nr Datenbank, Stand von Januar 2019) verglichen und mittels KRONA-Tools [64] ebenfalls nach dem LCA-Prinzip klassifiziert.

Abschließend wurde in folgender hierarchischer Reihenfolge für jeden *Contig* überprüft, auf welchen der eben genannten Ebenen jeweils Ergebnisse bzw. taxonomische Klassifikationen vorlagen. Bevorzugt wurden 16s rRNA Klassifizierungen akzeptiert. Waren solche nicht vorhanden wurden als „nächst-höhere“ Ebene, 23S rRNA Gene, dann die universellen Markergene und zuletzt die Gesamtprotein-Klassifizierung zu Rate gezogen. Es wurde somit für jeden *Contig* immer nur die höchstrangige uns aussagekräftigste Analysen-Ebene zur genauen Klassifikation genutzt.

Ausgeführt wurde HCC mit folgenden Skripten: ***blast2kronaclass.py***, wodurch die erhaltenen *Blast*-Tabellen für die nachfolgende LCA-Klassifizierung mit KRONA angepasst wurden. Mit ***get_full_taxpath_from_krona_class.py*** wurden aus allen Einzel-Klassifizierungen von 16S rRNA, 23S rRNA, Gesamtproteine und vorhandene Markergene eine gemeinsame hierarchische Klassifizierung erstellt. Das letzte Skript ***taxids_2_kronainput.py*** erzeugte eine für KRONA passend formatierte Eingabe-Tabelle mit zugehörigen *Coverage*-Informationen, für die Erstellung von relativen Abundanz-Plots.

2.6.5.2 Klassifizierung mit der GTDB-Tk Methode

Das öffentlich zugängliche Klassifizierungsprogramm GTDB-Tk [65] wurde zum Vergleich der bisher erlangten Annotationen herangezogen. GTDB-Tk wurde für die finalen *Bins* aus dem Methodenteil 2.4 mit folgendem Befehl ausgeführt.

```
gtdbtk classify_wf --cpus 4 -x .fa --genome_dir <bin_directory> --out_dir <gtdbtk_output>
```

Es wurden 4 CPUS mit dem Argument *„--cpus 4“* zugewiesen. Mit dem Argument *„-x .fa“* wurden lediglich GTDB-Tk ausschließlich *Fasta*-Dateien übergeben, mit *„--genome_dir“* wurde das Verzeichnis der *Bins* übergeben und mit *„--out_dir“* wurde die Ausgabedatei benannt.

2.6.6 Funktionale Zuordnung von Proteinsequenzen zu *Clusters of Orthologeous Groups* (COGs)

Den extrahierten COG Kategorien wurden mit dem öffentlich zugänglichen Skript ***cdd2cog.pl*** [66] Proteinsequenzen zugewiesen. Hierbei wurden die Proteinsequenzen gegen die *NCBI Conserved Domain*-Datenbank *geblastet* [66], wodurch die Genprodukte den *Cluster of Orthologous Groups* (COG-)Kategorien zugeordnet wurden. Bei den COGs handelt es sich um Einträge einer (COG-)Datenbank, in der funktionell verwandte Proteinsequenzen zu Gruppen (COGs) zusammengefasst sind. Durch Zuordnung der Proteinsequenzen zu verschiedenen COGS konnten also erste Aussagen über mögliche Funktionen der zugehörigen Gene getroffen werden.

2.6.7 Sekundärmetabolit-Gencluster

Das webbasierte Programm *antibiotics & Secondary Metabolite Analysis Shell* (antiSMASH) identifizierte, annotierte und analysierte mögliche Sekundärmetabolit-Biosynthese Gencluster in Bakteriengenomen [67]. Sekundärmetabolite können verschiedene Stoffe, wie z. B. antimikrobielle Peptide oder andere bioaktive Verbindungen sein. Diese sind für verschiedene Bereiche der Forschung, wie z. B. in der Antibiotikaforschung interessant.

Das Grundprinzip von antiSMASH ist einfach: Hier werden Proteinsequenzen mit bekannten und gut charakterisierten Sekundärmetabolit-Clustern verglichen. Sind diese in ausreichendem Maße ähnlich, wurde die Sequenz diesem Sekundärmetaboliten zugeordnet. Nachteil ist, dass auf diese Weise nur bereits bekannte grundlegende Mechanismen der Sekundärmetabolit-Synthese erkannt werden können.

3 Ergebnisse

3.1 Metagenomdaten

Eine Übersicht der grob eingeteilten vorläufigen terrestrischen Habitate inklusive der jeweiligen Anzahl an vorhandenen Metagenomdatensätzen aus der *MG-RAST* Datenbank ist in Tabelle 1 gezeigt. Diese Metagenomdatensätzen wurde nach unabhängigen aber stark vergleichbaren Proben mit jeweils relativ hohem *Chloroflexi*-Anteil durchsucht. Für weiterführende Analysen wurden schließlich Datensätze der Kategorie „*groundwater*“ mit einem *Chloroflexi*-Anteil von durchschnittlich 3 % ausgewählt. Alle so ausgewählten Datensätze stammten von Proben, die als Schadstoff-belastet beschrieben waren, was dieser Kategorie zusätzliche ökologische bzw. biotechnologische Relevanz gab, beispielsweise im Zusammenhang mit sogenannter „Bio-Remediation“ [68].

Beim näheren Betrachten dieser Kategorie stellten sich fünf der 27 darin enthaltenen Metagenome jedoch als nicht öffentlich zugänglich heraus. Auf direkte Nachfrage bei den Datenbankbetreibern (siehe digitaler Anhang [41]) stellte sich heraus, dass diese Datensätze teilweise mit Humangenomfragmenten (eventuell der damit arbeitenden Labormitarbeiter) kontaminiert waren, weshalb die resultierenden Sequenzdaten aus datenschutzrechtlichen Gründen nicht mehr veröffentlicht werden konnten.
Bei den verbliebenden 22 Datensätzen war es nun von Nachteil, dass diese alle aus einem einzelnen Projekt und somit aus demselben Probenort stammten. In der *MG-RAST* Datenbank standen also keine weiteren unabhängigen Schadstoff-belasteten Grundwasserproben anderer Regionen für Vergleiche bzw. Ko-Varianzbinnig zur Verfügung. Es stellte sich weiterhin bei den 22 übrigen *MG-RAST* Datensätzen heraus, dass *forward*- und *reverse-Reads* desselben Datensatzes getrennt voneinander hochgeladen wurden. Immer zwei aufeinanderfolgende Datensätze bilden somit einen gesamten Metagenomdatensatz. Daraus resultierten also letztendlich elf *MG-RAST* Datensätze.
Bei der Schadstoffbelastung dieser elf Grundwassermetagenomproben handelte es sich um eine Ethanol-Kraftstoff-Mischung. Die Probenumgebung lag in Houston, Texas (USA).
Um einen vergleichbaren Datensatz aus einer ähnlich Schadstoff-belasteten Probenumgebung zu erhalten, wurde auf die *NCBI*-Datenbank zurückgegriffen. Nach kurzer Recherche wurde in der *NCBI*-Datenbank ein geeigneter Grundwassermetagenom-Datensatz, welcher nahe einer kanadischen Mülldeponie gelegenen und somit einer vermeintlich Schadstoff-belasten Umgebung entnommen wurde, gefunden. Die Mülldeponie lag in Ontario (Kanada). Der kanadische Mülldeponie-Metagenomdatensatz beinhaltet bereits *forward*- und *reverse-Reads*. Alle Datensätze wurden mittels Illumina-Technologien sequenziert. Der zusätzliche kanadische Mülldeponie-Datensatz wurde analog zu den bereits erwähnten *MG-RAST* Datensätzen behandelt.

Tabelle 1. Einteilung der *MG-RAST* Metagenomdaten aus verschiedenen terrestrischen Habitaten. Für jedes Habitat wurde die Anzahl der vorhandenen Metagenomdatensätze angegeben. Nachfolgend ist der durchschnittliche prozentuale Anteil an *Chloroflexi* bezüglich der vorhandenen Prokaryoten und des gesamten Metagenoms aufgelistet. Grün hinterlegt wurde das ausgewählte Schadstoff-belastete Habitat.

Habitate	Anzahl	Anteil der *Chloroflexi* in den Prokaryoten [%]
Antarctic	13	1,98
Mine	20	2,80
Industry	7	4,03
Groundwater	**27**	**3,74**
Grassland	120	1,57
Mountain	13	0,94
Forest	41	1,70
Desert	22	2,50
Tundra	40	2,41
Glacier	1	0,31
Agricultrue	197	2,42
Volcano	2	0,91
Microbial mat	16	3,34
Wetland	103	2,48

Um die Übersicht zu wahren, wurden die elf texanischen mit Ethanol-Kraftstoffgemisch belasteten Metagenome aus der *MG-RAST* Datenbank im Folgenden mit TOGW1-11 (texanisches Öl-kontaminiertes Grundwasser) und das kanadische Mülldeponie-Grundwassermetagenom aus der *NCBI*-Datenbank als KMGW12 bezeichnet (Tabelle 2). Diese zwölf Datensätze wurden somit für weitere Analysen ausgewählt.

Tabelle 2. Ausgewählte Metagenomdatensätze mit einem durchschnittlichen Anteil von 3 % *Chloroflexi*. TOGW1-11 (texanische Öl-kontaminierte Grundwassermetagenome) stellen die in *MG-RAST* gefundenen Datensätze dar. KMGW12 (kanadisches Mülldeponie-Grundwassermetagenom) ist der manuell gesuchte Datensatz der *NCBI* Datenbank.

Metagenomdatensätze					
	forward	***reverse***		***forward***	***reverse***
TOGW1	mgm4519753.3	mgm4519754.3	**TOGW7**	mgm4519765.3	mgm4519766.3
TOGW2	mgm4519755.3	mgm4519756.3	**TOGW8**	mgm4519767.3	mgm4519768.3
TOGW3	mgm4519757.3	mgm4519758.3	**TOGW9**	mgm4519769.3	mgm4519770.3
TOGW4	mgm4519759.3	mgm4519760.3	**TOGW10**	mgm4519771.3	mgm4519772.3
TOGW5	mgm4519761.3	mgm4519762.3	**TOGW11**	mgm4519773.3	mgm4519774.3
TOGW6	mgm4519763.3	mgm4519764.3	**KMGW12**	SRX3574179	

3.2 Finale Koassemblierung

Die Gesamtlänge der Koassemblierung aller ausgewählten Schadstoff-belasteten Grundwasser Metagenomdatensätze umfasst ca. 2 Gb, verteilt auf über 800.000 *Contigs* mit einer durchschnittlichen Länge von ca. 2,3 kb (Tabelle 3).
Die Durchschnittsgröße der in der *NCBI*-Datenbank [35] hinterlegten vollständigen bakteriellen Genome beträgt ca. 3,7 Mb. Wird nur die Gesamtlänge aller *Contigs* betrachtet, könnten folglich rein rechnerisch bis zu ca. 540 bakterielle Genome daraus zusammengesetzt werden.

Tabelle 3. Übersicht der vorhandenen *Contigs*. Es ist die Anzahl, die gesamte Länge aller *Contigs*, die Länge der kürzesten und längsten *Contigs* sowie die durchschnittliche Länge und der Median über alle *Contigs* zu sehen.

***Contig* Anzahl**	820.168
Gesamt-*Contig*-Länge	1.928.648.161 bp
Mindest-*Contig*-Länge	750 bp
Maximal-*Contig*-Länge	1.563.906 bp
Durchschnittliche *Contig*-Länge	2.351,5 bp
Median der *Contig*-Längen	1.177,5 bp

3.2.1 Vergleich der Metagenomanalyse

In den texanischen Öl-kontaminierten Grundwasser Metagenom-Datensätze TOGW1-11 lag der *Chloroflexi*-Anteil zwischen 7 bis 18 %. Der Durchschnittswert lag somit bei ca. 13 %. Im kanadischen Mülldeponie Grundwasser Metagenom-Datensatz KMGW12 liegt der Anteil jedoch lediglich bei 0,2 % (Abbildung 2). Trotz des geringen *Chloroflexi*-Anteils wurde KMGW12 beim *Binning* miteinbezogen, da diese Information für anschließendes Co-Varianz basiertes *Binning* genutzt werden kann.

Ein in beiden Beprobungsorten stark vertretenes Phylum sind *Proteobacteria*. Weiterhin sind in einigen der hier analysierten Metagenomen Vertreter der „*Planctomycetes, Verrucomicrobia* und *Chlamydiae"* (PVC)- und „*Fibrobacteres, Flavobacteria, Chlorobi* und *Bacteroidetes*" (FCB)-Gruppe, sowie *Nitrospira* in vergleichsweise hohen Anteilen zu finden (Abbildung 2).
Die PVC-Gruppe wird als Superphylum bezeichnet [69] und beinhaltet verschiedene Phyla, wie z. B. die namensgebenden *Planctomycetes*, *Verrucomicrobia* und *Chlamydiae. Planctomycetes* sind in der Umwelt weit verbreitet und kommen in Böden, Meer- und Süßwasser vor. Sie können sowohl aerob als auch anaerob sein. *Verrucomicrobia* kommen häufig im Boden vor, sind jedoch auch in marinen Lebensräumen anzutreffen [70]. *Chlamydiae* benötigen in der Regel eukaryotische

Wirtszellen zum Überleben [71].
Das Superphylum FCB-Gruppe hat eine hohe Bedeutung für die Umwelt- und Darm-Mikrobiologie erlangt [72]. *Fibrobacteres* sind beispielsweise im Pansen von Wiederkäuern zu finden und verdauen Cellulose. *Chlorobi* sind photolithotroph, können anaerob Photosynthese betreiben und kommen in anoxischen Gewässern vor. *Bacteroidetes* können in terrestrischen und marinen Habitaten gefunden werden. In jedem Phyla der FCB-Gruppe gibt es Vertreter, welche enge Assoziationen zu Menschen und Tieren zeigen [73].

Unter den texanischen Öl-kontaminierten Grundwassermetagenomen weisen TOGW1, TOGW6 und TOGW9, TOGW3, TOGW8 und TOGW11 sowie TOGW2, TOGW4, TOGW7 und TOGW10 eine sehr ähnliche Verteilung des *Chloroflexi*-Anteils auf. Die anderen vorkommenden Bakterien(super-)phyla, wie PVC- und FCB-Gruppe, *Nitrospira* und *Proteobacteria* besitzen teilweise gravierende Unterschiede in ihren Verteilungen.
Im Vergleich zu TOGW1-11 weist KMGW12 einen signifikant höheren prozentualen Anteil von 47 % an C*andidatus* (engl. *candiate*) *Phyla* auf. Hierbei handelt es sich um Phyla, aus welchen noch keine Vertreter kultiviert werden konnten. Mit 18 % ist auch die FCB-Gruppe im Vergleich zu den texanischen Metagenomen deutlich stärker vertreten. Eine genaue Auflistung aller *Bins* ist im digitalen Anhang [41] hinterlegt.

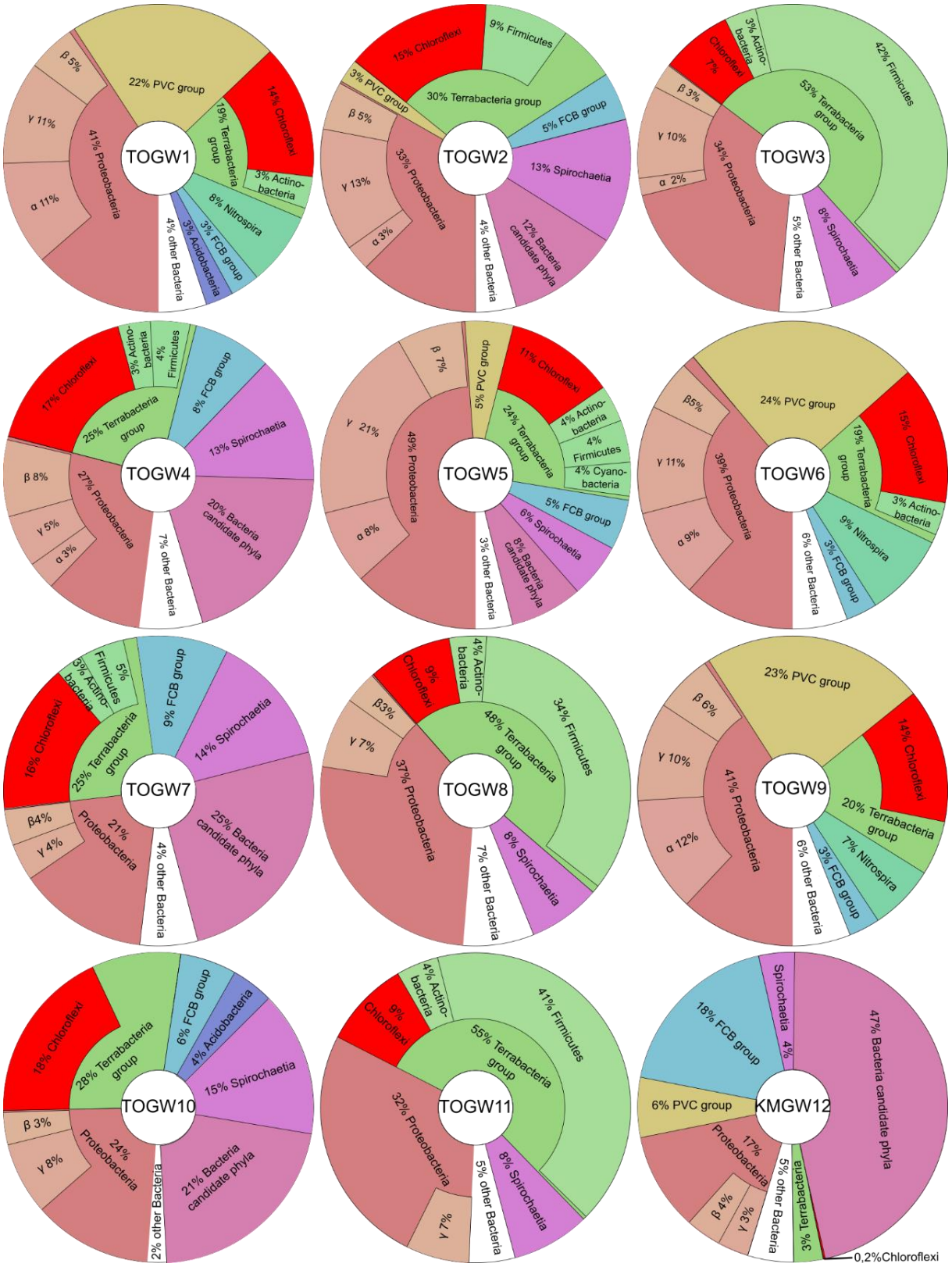

Abbildung 2. KRONA-Charts der ausgewählten Metagenome. TOGW1-11 zeigen texanische Öl-kontaminierte Grundwassermetagenome, KMGW12 das kanadische Mülldeponie-Grundwasser-Metagenom an. In Rot ist der jeweilige *Chloroflexi*-Anteil hervorgehoben. *Proteobacteria* (rosa) und C*andidate phyla* (violett) sind neben *Terrabacteria group* (grün) häufig vertreten. Ein signifikanter Unterschied ist jedoch bei KMGW12 zu sehen, hier stellen *Chloroflexi* lediglich 0,2% des bakteriellen Metagenoms.

In der Vergangenheit basierte die taxonomische Klassifikation bei *MG-RAST* nur auf der *best BlAST hit*-Methode. Seit dem 4.0 Update wurden jedoch auch LCA-Methoden implementiert. Es ist nicht zweifelsfrei sicher, ob die über das Webinterface erhaltenen Taxonomieprofile auf der *best BLAST hit*- oder der LCA-Methode basieren.
Die *best BLAST hit*-Methode basiert auf dem nächstähnlichen Datenbankeintrag für den jeweiligen Sequenzread bzw. das jeweilige Genomfragment. Dadurch können die Sequenzreads bzw. Genomfragmente falsch zugeordnet werden.
Des Weiteren werden kurze *Reads* möglicherweise zu (irreleitenden) kurzen *Alignments* gegen die Referenzen ausgerichtet. Ein weiteres Problem ist, für mehrere gleich gute Treffer können widersprüchliche Referenzeinträge gefunden werden. In diesem Fall ist der „beste“ Treffer nicht repräsentativ.
Diese Problematik wird durch die Assemblierung der *Reads* mit anschließender HCC-Methode, welche mithilfe kurierter Referenzdatensätze klassifiziert, umgangen. Wodurch eine genauere Aussage über die tatsächlichen taxonomischen Anteile in TOGW1-11 getroffen werden können.
In der Gegenüberstellung der beiden Methoden ist dieser Unterschied deutlich zu sehen (Abbildung 3). Der *MG-RAST*-Algorithmus hat, basierend auf den unassemblierten *Reads* von TOGW1-11, deutlich geringere *Chloroflexi*-Anteile, als in dieser Arbeit nach Assemblierung identifiziert werden konnten, festgestellt. Die entsprechenden *MG-RAST*-Ergebnisse wurden manuell von der *MG-RAST*-Webseite [36] entnommen und sind im digitalen Anhang hinterlegt [41].

NCBI berechnet keine mögliche taxonomische Zusammensetzung der dort hochgeladenen Datensätze, weshalb das kanadische Mülldeponie-Grundwasser-Metagenom KMGW12 hier nicht direkt verglichen werden kann.

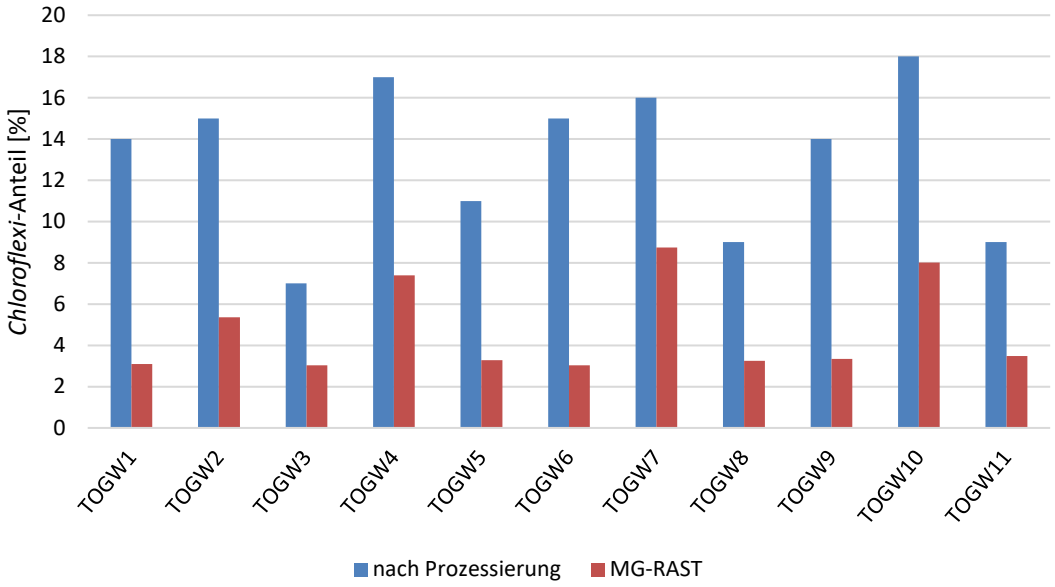

Abbildung 3. Vergleich der *Chloroflexi*-Anteile in den texanischen Metagenomen TOGW1-11. Basierend auf 16S Genen, welche in fast voller Länge assembliert und nach der *Lowest common Ancester* Methode mit kurierten 16S Referenzdatenbanken (blau) klassifiziert wurden. Die unassemblierten *Reads* wurden von *MG RAST* nach der *best BLAST hit*- bzw. LCA-Methode gegen nicht kurierte Referenzdatenbanken (rot) klassifiziert.

3.2.2 16S RIBOSOMALE RNA GENSEQUENZ BASIERTE PHYLOGENIE

Insgesamt wurden mittels *RNAmmer* (siehe Methodenteil 2.6.4) aus der finalen Metagenom-Koassemblierung 246 16S rRNA Gensequenzen extrahiert (siehe digitaler Anhang [41]). Sechs davon, welche im nachfolgenden als Metagenom-rRNA-Sequenz 1-6 durchnummeriert wurden, ließen sich mittels des phylogenetischen Klassifizierungprogramms SINA (siehe Methodenteil 2.6.4) als *Chloroflexi* identifizieren, wovon aufgrund ihrer Identitäten zum nächst verwandten Vertreter einer dem Phylum *Chloroflexi*, drei der Klasse *Anaerolineae*, einer der Ordnung *Anaerolineales* und einer dem Genus *Bellilinea* zugeordnet werden konnten (Abbildung 4).

Metagenom-rRNA-Sequenz 1 clustert mit einem *bootstrap-Konfidenzwert* von 75 % zusammen mit *Bellilinea caldifistulae*. Die Sequenzidentität zu *B. caldifistulae* beträgt 96 %, dies entspricht einer taxonomischen Übereinstimmung auf Genus-Ebene.
Metagenom-rRNA-Sequenz 2 mit einem *bootstrap*-Wert von 84 % bei *Longilinea arvoryzae*. Die Sequenzidentität zu *L. arvoryzae* beträgt 92 %, was mit der taxonomischen Ordnungs-Ebene übereinstimmt.
Die vier weiteren 16S Sequenzen clustern nicht direkt bzw. nur mit Konfidenzwerten unter 50 % bei bereits beschriebenen Spezies.
Der nächste Verwandte für Metagenom-rRNA-Sequenz 3 wäre *Pelolinea submarina* und für

Metagenom-rRNA-Sequenz 4 *Leptolinea tardivitalis*. Die Identitäten zu den jeweiligen nächst-verwandten Spezies betragen 88 % und 91 %, wodurch eine taxonomische Übereinstimmung auf der Klassen-Ebene stattfindet.

Der laut phylogenetischer Clusterung nächstverwandte kultivierte Verwandte zu Metagenom-rRNA-Sequenzen 5 und 6 wäre, allerdings mit erheblicher Distanz, *Dehalacoccoides mccartiy*. Die Sequenzidentitäten betrugen jeweils 80 % bzw. 85 %, das entspricht lediglich einer taxonomischen Übereinstimmung auf Phylum- bzw. Klassen-Ebene.

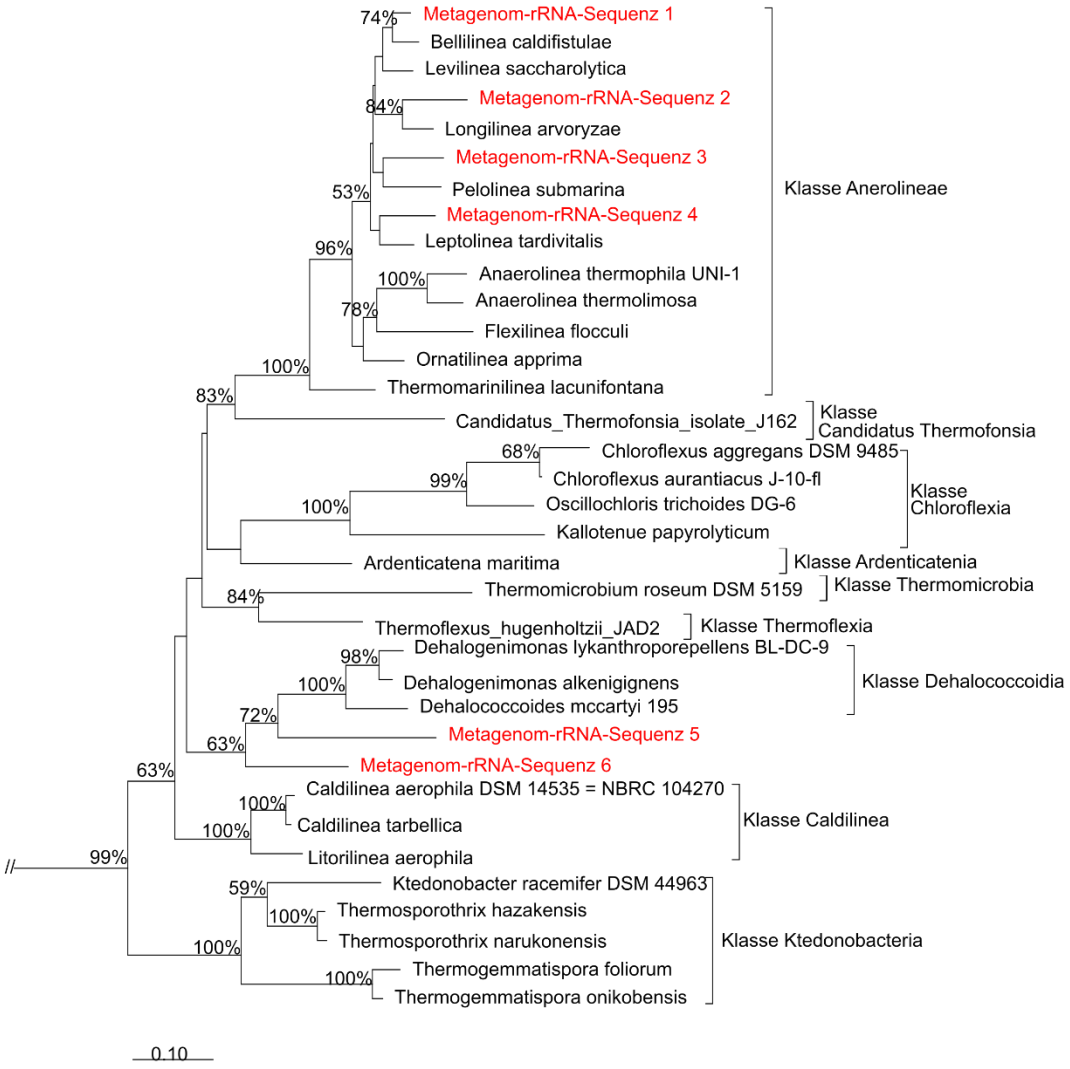

Abbildung 4. Phylogenetische Diversität der *Chloroflexi* 16S Sequenzen der finalen Metagenom-Koassemblierung. Dargestellt ist ein RaxML Stammbaum der *Chloroflexi* 16S Sequenzen der Metagenom-Koassemblierung in Relation zu Referenzsequenzen bekannter *Chloroflexi* Typstämme. Berechnet wurde der Baum mit 200 Permutationen nach der *rapid bootstrapping* Methode. Metagenom-rRNA-Sequenzen sind in Rot hervorgehoben, Klassen sind an der rechten Seite erkennbar. Prozentzahlen an den Knoten geben *bootsrtap*-Konfidenzwerte über 50 % an.

3.2.3 Relative Abundanzen

Im Vergleich zum kanadischen Mülldeponie-Metagenom KMGW12 sind in den texanischen Öl-kontaminierten Grundwassermetagenomen (TOGW1-11) deutlich weniger *Dehalococcoidia* (TOGW 31,5 %, KMGW12 68,0 %) vorhanden, dafür jedoch mehr *Anaerolineaceae* (TOGW 63,3 %, KMGW12 28,0 %). Weitere unbestimmte *Chloroflexi* Klassen sind in den Schadstoff-belasteten Grundwassermetagenomen mit einer ähnlichen Häufigkeit anzutreffen (TOGW 5,1 %, KMGW12 4,0 %).

Bei den texanischen Öl-kontaminierten Proben (TOGW1-11) schwankten die relativen Anteile an *Dehalococcoidia* um 1,24 %. Die relativen Anteile bei den *Anaerolineaceae* weisen eine deutlich höhere Schwankung von 25,68 % auf (Abbildung 5, digitaler Anhang [41]).

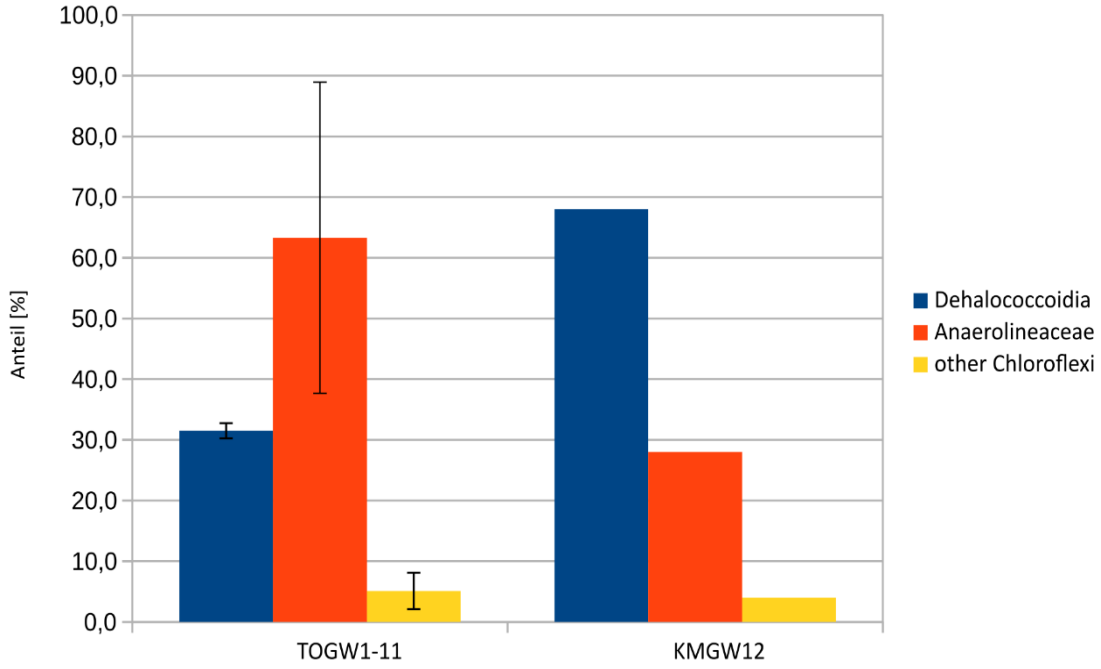

Abbildung 5. Relative Abundanzen verschiedener *Chloroflexi*-Klassen in den Metagenomdatensätzen. TOGW1-11 zeigen die texanischen Metagenome, KMGW12 ist das kanadische Metagenom. Es wurden lediglich *Dehalococcoidia* (blau) und *Anaerolineaceae* (orange) identifiziert, weitere *Chloroflexi*-Klassen sind unter „*other Chloroflexi*“ (gelb) zusammengefasst. Bei TOGW1-11 wurde zudem die Standardabweichung berechnet (Fehlerbalken).

3.3 Genomanalyse der rekonstruierten *Chloroflexi* Metagenome (*Bins*)

3.3.1 Genomqualität und taxonomische Klassifizierung

Um Aussagen über die Genomqualität der *Chloroflexi-Bins* zu treffen wurden diese auf ihre Reinheit und Vollständigkeit untersucht. Neben der „Genomvollständigkeit“ und potentiellen „Kontamination“ wurde auch eine theoretische „Stammheterogenität“ mithilfe des Programms *checkM* geschätzt und daraus schließlich eine „angepasste Kontamination“ berechnet (siehe Methodenteil 2.5) [53].

Bei einer „Genomvollständigkeit“ von beispielsweise 100 % ist jedes der Markergene auf dem untersuchten Genom vorhanden, wodurch dieses demnach als vollständig angesehen wird. Als Kontamination wurden mehrfach vorkommende Markergene desselben Typs angesehen. Da diese mehrfach pauschal als Verunreinigung angesehen wurden, in der Realität jedoch tatsächlich gelegentlich in bakteriellen Genomen auftreten, ist diese Kontaminations-Schätzung nicht zwangsläufig repräsentativ.

„Stammheterogenität“ gibt an, wie viele der mehrfach vorkommenden Markergene eine Sequenzidentität von mehr als 90 % aufweisen. Diese können zum Beispiel durch die Anwesenheit mehrerer sehr nah verwandter Stämme derselben Spezies oder aber lediglich durch eine mehrfache Kopienzahl einzelner Markergene innerhalb desselben Genoms hervorgerufen werden. Bei der „angepassten Kontamination“ wurde daher entsprechend die „Stammheterogenität“ aus der „Kontamination“ herausgerechnet.

Für tiefergehende Analysen wurden nur die vielversprechendsten und als „hoch-qualitativ“ angesehenen *Chloroflexi-Bins* mit einer angepassten Kontamination unter 10 % und einer Genom-Vollständigkeit über 50 % ausgewählt. Bei einander entsprechenden *Bins*, welche sowohl in den *MaxBin*- als auch in *MetaBAT*-Ergebnissen auftraten, wurde der jeweils qualitativ höherwertige ausgewählt (Tabelle 4). Dies führte zu elf finalen *Chloroflexi-Bins*, im Folgenden als *Bins* 01-11 bezeichnet.

Aus Assemblierungsgröße und Genomvollständigkeit lassen sich die theoretischen Genomgrößen bestimmen, wodurch überprüft werden kann, ob diese mit möglichen nah verwandten Spezies übereinstimmen. Neun *Bins* hatten eine Assemblierungsgröße zwischen 2,59 MB bis 4,09 Mb, bin04 lag bei 1,85 Mb und bin11 bei 5,28 Mb. Unter Berücksichtigung der jeweiligen Genomvollständigkeit lag die theoretische Genomgröße aller *Bins* somit zwischen 2,68 Mb und 5,67 Mb.

Auch anhand des GC-Gehalts können Aussagen über Verwandtschaftsbeziehungen getroffen werden. Wird der GC-Gehalt der *Chloroflexi-Bins* mit nächstverwandten Referenzspezies, welche

mit hoher Konfidenz beim jeweiligen *Bin* in den phylogenetischen Stammbäumen clustert, verglichen und weicht dieser stark ab, handelt es sich vermutlich um kontaminierte *Bins* bzw. um Artefakte. Existiert keine nächstverwandte Spezies mit hohen Konfidenzwerten, kann anhand des GC-Gehalt keine qualitative Aussage über die Nähe einer Verwandtschaft getroffen werden. Möglicherweise könnten jedoch *Bins* mit ähnlichem GC-Gehalt eine nahe Verwandtschaft aufweisen.

Der GC-Gehalt des Großteils der ausgewählten *Chloroflexi-Bins* lag im Bereich von ca. 49 % bis 67 %, bin07 hingegen wies einen auffallend hohen GC-Gehalt von 71,59 % auf.

Tabelle 4. Genomqualität ermittelt durch *checkM*. Für jeden *Bin* wurde der GC-Gehalt, die Genomgröße, Genomvollständigkeit, Kontamination und Stammheterogenität bestimmt. Die angepasste Kontamination wurde mit Hilfe der erhaltenen Werte berechnet. Grün hinterlegt sind die elf ausgewählten *Chloroflexi-Bins*. Grau hinterlegt sind die *Bins*, welche doppelt vorkamen, jedoch eine schlechtere Qualität aufwiesen. Die von MetaBAT und MaxBin benannten Urpsrungsbins sind ebenfalls dargestellt.

Finaler *Bin*	Ursprungs-bin*	GC-Gehalt [%]	Assemblier-ungsgröße [Mb]	Genom-vollständigkeit [%][A]	Konta-mination [%][B]	Stamm-heterogenität [%][C]	angepasste Kontamination [%][D]
bin01	max067	58,08	4,09	97,26	1,88	0,00	1,88
	met40	57,89	4,08	97,26	8,62	0,00	8,62
bin02	max199	67,27	3,05	60,17	6,35	0,00	6,35
bin03	met119	55,58	3,04	94,83	1,72	100,00	0,00
	max017	55,65	3,32	96,55	3,45	50,00	1,73
bin04	met137	49,56	1,85	58,90	0,63	0,00	0,63
bin05	met19	49,84	3,28	82,01	0,00	0,00	0,00
bin06	met190	54,10	3,30	98,28	0,16	100,00	0,00
	max038	54,26	3,49	98,28	11,68	7,69	10,78
bin07	met32	71,59	2,92	93,73	0,86	100,00	0,00
	max126	71,44	3,09	96,55	8,70	8,33	7,98
bin08	met4	53,29	2,59	96,55	0,86	100,00	0,00
bin09	met487	50,01	3,06	82,01	0,00	0,00	0,00
	max077	49,33	3,64	96,55	20,34	5,66	19,19
bin10	met54	56,74	3,90	70,92	1,88	100,00	0,00
	max161	56,70	4,78	87,88	9,33	30,00	6,53
bin11	met75	58,54	5,28	93,15	7,52	50,00	3,76
	max037	59,08	6,07	94,00	28,76	50,00	14,38

[A] siehe digitaler Anhang [41]
[B] Abschätzung, wie viele Markergene das gesamte Genom abdeckten
[C] Aussage über Anzahl der gefundenen Markergene, welche mehr als einmal vorliegen
[D] Aussage über mehrfacht vorkommende Markergene, welche eine Identität von mehr als 90 % aufweisen
[E] Stammheterogenität wurde aus der Kontamination herausgerechnet

Mittels der bereits erwähnten *Hirarchical Contig Classification* (HCC) (Methodenteil 2.6.5.1) wurde der kleinste gemeinsame Vorfahre der untersuchten *Chloroflexi-Bins* bestimmt (Tabelle 5). Hierbei wurde bin01 mit hoher Konfidenz die Spezies *Longilinea arvoryzae* als nächsten beschriebenen Verwandten identifiziert. Unterstützt wurde diese Vermutung durch die vorherige phylogenetische Analyse der 16S rRNA-Sequenzen, bei der bin01 ebenfalls mit hoher Konfidenz nahe bei *Longilinea arvoryzae* clusterte (Abbildung 6).

Bin02 konnte mit geringer Konfidenz lediglich bis zum Phylum *Chloroflexi* klassifiziert werden. Dies spricht wiederum dafür, dass es sich dabei um eine neue, bislang unbeschriebene Klasse handeln könnte. Als nächstähnlicher Vertreter in der Referenzdatenbank wurden, allerdings mit sehr geringer Konfidenz, Vertreter der Klasse *Ktedonobacteria* identifiziert.

Für *Bins* 03, 06 und 08 wurde mit ähnlichen sehr schwachen Konfidenzen als nächstmöglicher Vertreter die Familie *Anaerolineaceae* identifiziert, weshalb diese *Bins* lediglich die Phylum-Ebene, welche durchweg höhere Konfidenzwerte aufweist, repräsentieren. Bin09 repräsentiert mit ähnlich schwachen Konfidenzen ebenso lediglich das Phylum *Chloroflexi*. Hier wurde als nächstmöglicher Vertreter mit sehr schwacher Konfidenz die Klasse *Anaerolineae* identifiziert.

Bins 04 und 10 wurde als nächstmöglicher Vertreter auf Klassen-Ebene *Anaerolineae* identifiziert. Hier lagen die Konfidenzwerde bei 1,08 % bzw. 1,99 %.

Die *Bins* 05 und 07 konnten mit einer Konfidenz von jeweils 31,63 % bzw. 11,69 % lediglich auf Phylum-Ebene identifiziert werden. Aufgrund der geringen Wahrscheinlichkeit, welche die HCC-Methode bin07 den *Chloroflexi* zugeordnet hat, ist zu vermuten, dass es sich hierbei um eine neue Klasse der *Chloroflexi* bzw. ein nah verwandtes unbekanntes Phylum handelt. Bei bin05 handelt es sich vermutlich um eine neue *Chloroflexi*-Klasse.

Für bin11 wurde mit einer höheren Konfidenz von 4,82 % die Familie *Anaerolineaceae* identifiziert. Dies spricht dafür, dass es sich bei diesem *Bin* um eine neue, bislang unbeschriebene Gattung handeln könnte.

Tabelle 5. Taxonomische Klassifizierung der *Chloroflexi-Bins* nach der *Hirarchical Contig Classification* Methode. Es werden Phylum, Klasse, Ordnung, Familie und Gattung/Spezies gezeigt. Eine häufig vorkommende Klasse ist *Anaerolineae,* auch *Ktedonobacteria* ist vertreten. Die Prozentzahl gibt jeweils den Anteil der jeweiligen Genomsequenz an, welche die angegebene Klassifizierung unterstützt und stellt somit einen Konfidenzwert der Klassifikation dar.

bin	Phylum	Klasse	Ordnung	Familie	Gattung/Spezies
bin01	*Chloroflexi* [w: 63.95%]	*Anaerolineae* [w: 12.96%]	*Anaerolineales* [w: 12.05%]	*Anaerolineaceae* [w: 11.95%]	*Longilinea arvoryzae* [w: 3.87%]
bin02	*Chloroflexi* [w: 15.11%]	*Ktedonobacteria* [w: 0.24%]	*Ktedonobacterales* [w: 0.20%]	*Ktedonobacteraceae* [w: 0.10%]	*Ktedonobacter racemifer* [w: 0.10%]
bin03	*Chloroflexi* [w: 66.43%]	*Anaerolineae* [w: 0.93%]	*Anaerolineales* [w: 0.93%]	*Anaerolineaceae* [w: 0.93%]	None [w: 0.00%]
bin04	*Chloroflexi* [w: 59.01%]	*Anaerolineae* [w: 1.08%]	*Anaerolineales* [w: 0.85%]	*Anaerolineaceae* [w: 0.85%]	None [w: 0.00%]
bin05	*Chloroflexi* [w: 31.64%]	None [w: 0.00%]	None [w: 0.00%]	None [w: 0.00%]	None [w: 0.00%]
bin06	*Chloroflexi* [w: 55.41%]	*Anaerolineae* [w: 0.91%]	*Anaerolineales* [w: 0.91%]	*Anaerolineaceae* [w: 0.91%]	None [w: 0.00%]
bin07	*Chloroflexi* [w: 11.69%]	None [w: 0.00%]	None [w: 0.00%]	None [w: 0.00%]	None [w: 0.00%]
bin08	*Chloroflexi* [w: 56.44%]	*Anaerolineae* [w: 0.46%]	*Anaerolineales* [w: 0.46%]	*Anaerolineaceae* [w: 0.35%]	None [w: 0.00%]
bin09	*Chloroflexi* [w: 33.33%]	*Anaerolineae* [w: 0.26%]	None [w: 0.00%]	None [w: 0.00%]	None [w: 0.00%]
bin10	*Chloroflexi* [w: 67.86%]	*Anaerolineae* [w: 1.99%]	*Anaerolineales* [w: 0.21%]	None [w: 0.00%]	None [w: 0.00%]
bin11	*Chloroflexi* [w: 28.43%]	*Anaerolineae* [w: 4.82%]	*Anaerolineales* [w: 4.82%]	*Anaerolineaceae* [w: 4.82%]	*Anaerolinea* [w: 0.07%]

3.3.2 PHYLOGENETISCHE VERWANDTSCHAFTSBEZIEHUNGEN

In den elf hochqualitativen *Chloroflexi-Bins* wurden 42 universelle Einzelkopie-Markergene identifiziert (Methodenteil 2.6.2). Hiervon wurden anhand der entsprechenden Sequenzlängen und ihrer Verteilung in den elf *Bins* sieben repräsentative Marker für phylogenetische Analysen ausgewählt, welche den COG-Kategorien COG0016, COG0018, COG0085, COG0172, COG0495, COG0533 und COG0541 zugehörig sind [56].
Bei COG0016 handelt es sich um eine Phenylalanyl-tRNA Synthetase (α-Untereinheit), bei COG0018 um eine Arginyl-tRNA Synthetase.
COG0085 ist eine DNA-gesteuerte RNA Polymerase (β-Untereinheit). Bei COG0172 handelt es sich um eine Seryl-tRNA Synthetase. COG0495 ist eine Leucyl-tRNA Synthetase. Das Markergen von COG0533 ist der tRNA A37 Threonylcarbamoyltransferase TsaD zugehörig. COG0541 gehört einem Markergen für die Signalrezeptor GTPase an.

Für jedes dieser Markergene wurde ein phylogenetischer Stammbaum der 11 *Bins* sowie ausgewählter Referenz-Typstämme berechnet (digitaler Anhang [41]). Die zwei aussagekräftigsten Stammbäume (COG495 bzw. „Leucyl-tRNA Synthetase“ und COG541 bzw. „Signalrezeptor GTPase“) sind in vereinfachter Form im Folgenden dargestellt (Abbildung 6 und Abbildung 7).

Bins 01, 03 und 06 clustert, basierend auf den Leucyl-tRNA Synthetase- und Signalrezeptor GTPase-basierten Markern mit hoher Konfidenz innerhalb der *Anaerolineae*, also gehören diese wohl der Klasse *Anaerolineae* an. Bin01 lässt sich zudem mit hoher Konfidenz als naher Verwandter der Spezies *Longilinea arvoryzae* beschreiben. Die *Bins* 03 und 06 clustern in beiden Stammbäumen weiter entfernt bei der Spezies *Bellilinea caldifistulae*. Weitere *Chloroflexi-Bins*, welche den *Anaerolineae* zugeordnet wurden, jedoch schlechte Konfidenzwerte zum nächsten Verwandten aufweisen oder in den Bäumen nicht eindeutig clustern, sind *Bin* 04, 05, 08 bis 10.
Bin11 clustert im Leucyl-tRNA Synthetase-basierten Stammbaum mit niedriger Konfidenz bei der Klasse *Anaerolineae*. Im Signalrezeptor GTPase-basierten Stammbaum clustert dieser *Bin* bei unklassifizierten *Chloroflexi*.
Bins 02 und 07 clustern beide nur mit bislang unklassifizierten *Chloroflexi*-Vertretern und können somit keiner bekannten Klasse zugeordnet werden.

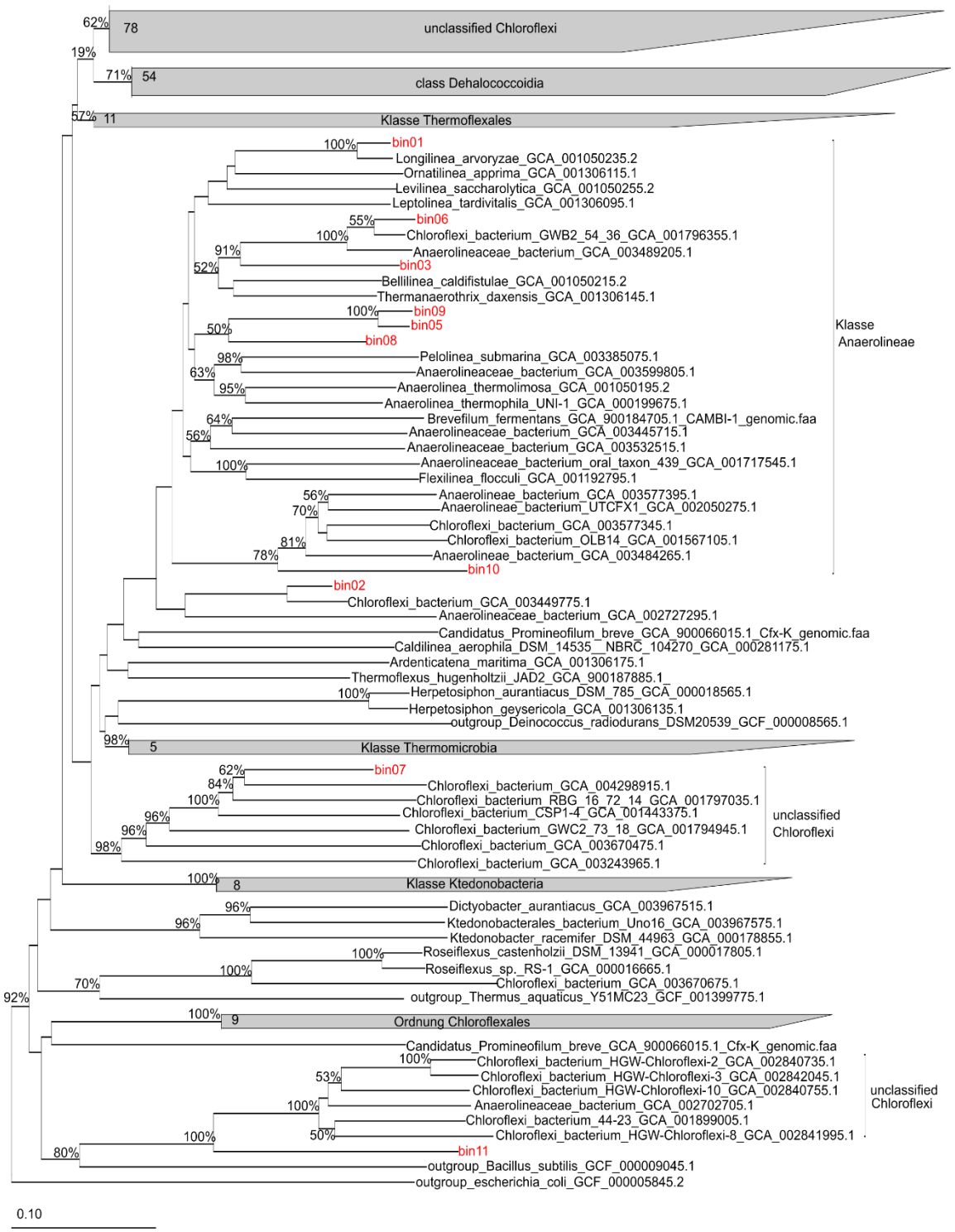

Abbildung 6. Vereinfachter phylogenetischer Stammbaum mit *Bins*, welchen ein Markergen des COG495 zugeordnet wurde. Der Stammbaum wurde mit allen zur Verfügung stehenden *Chloroflexi*-Referenzen berechnet. Es wurde die *Neighbour-Joining* Methode verwendet und mit 1000 Permutationen berechnet. Prozentzahlen an den Knoten geben *bootstrap*-Konfidenzwerte über 50 % an. *Chloroflexi-Bins* sind rot markiert. Manche Gruppen sind zur einfacheren Darstellung kollapsiert in Form grauer Kästen dargestellt. Die jeweilige taxonomische Ebene ist jeweils an der rechten Seite angegeben. Als *Outgroup* dienten *Bacillus subtilis, Deinococcus radiodurans, Thermus aquaticus* und *E. coli*.

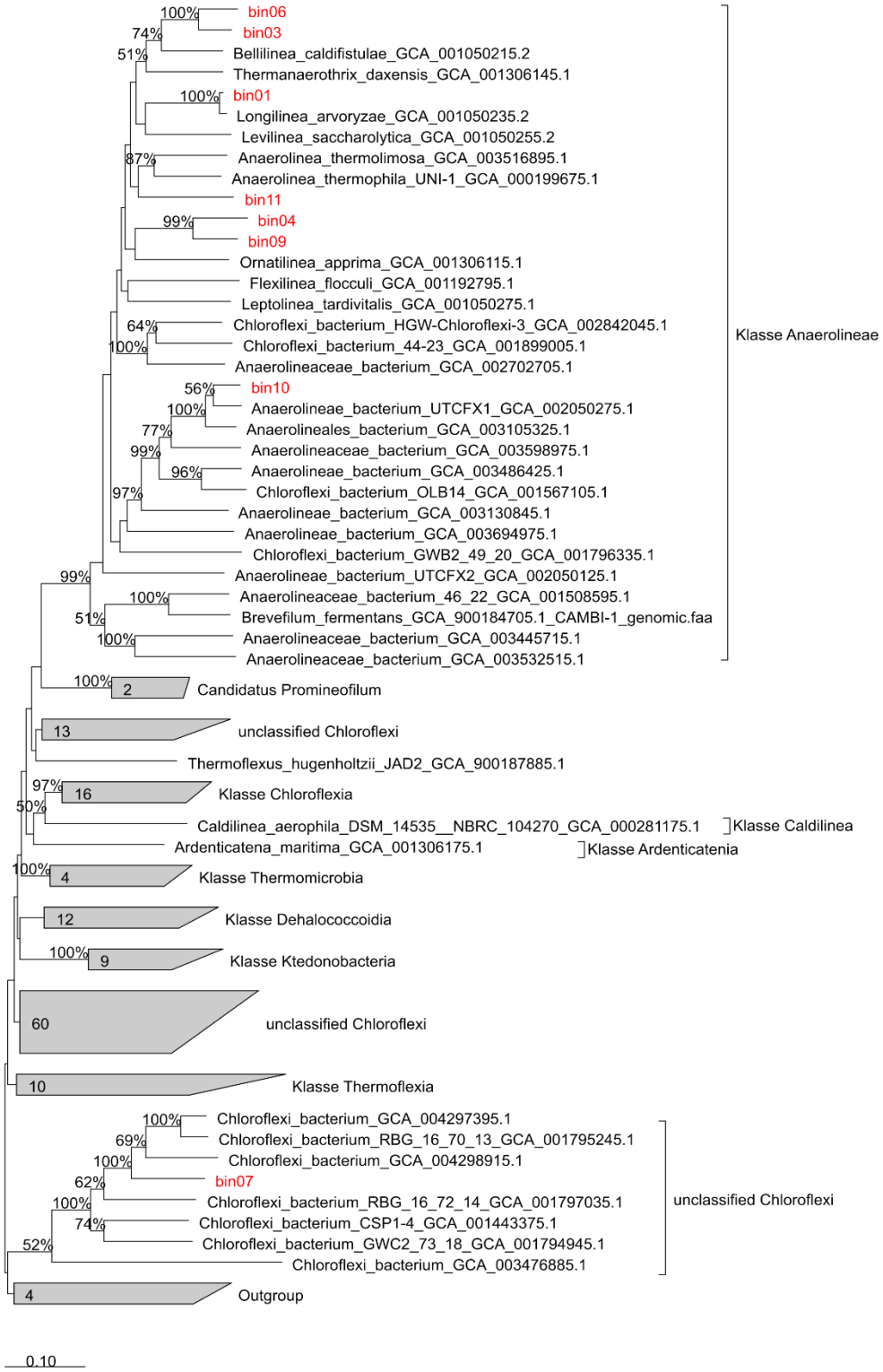

Abbildung 7. Vereinfachter phylogenetischer Stammbaum mit *Bins*, welchen ein Markergen des COG541 zugeordnet wurde. Der Stammbaum wurde mit allen zur Verfügung stehenden *Chloroflexi*-Referenzen berechnet. Es wurde die *Neighbour-Joining* Methode verwendet und mit 1000 Permutationen berechnet. Prozentzahlen an den Knoten geben *bootstrap*-Konfidenzwerte über 50 % an. *Chloroflexi-Bins* sind rot markiert. Manche Gruppen sind zur einfacheren Darstellung kollapsiert in Form grauer Kästen dargestellt. Die jeweilige taxonomische Ebene ist jeweils an der rechten Seite angegeben. Als *Outgroup* dienten *Bacillus subtilis, Deinococcus radiodurans, Thermus aquaticus* und *E. coli*.

3.3.3 Vergleich der taxonomischen Klassifizierungen

Es wurde eine unterschiedliche Anzahl an *Chloroflexi-Bins* durch die drei Klassifizierungsmethoden ermittelt. Bei der *Hirarchical Contig Classification* Methode wurden 32 *Chloroflexi-Bins* klassifiziert. Durch die GTDB-Tk-Methode wurden 30 *Bins* und durch *checkM* 236 *Bins* als *Chloroflexi* klassifiziert. Die genauen Zuordnungen der *Bins* sind dem digitalen Anhang zu entnehmen [41].

Die unterschiedlichen Häufigkeiten der identifizierten *Chloroflexi-Bins* sind anhand der verwendeten Algorithmen und der damit verbundenen Klassifizierung zu erklären. Einige *Chloroflexi-Bins* wurden beispielsweise bei der GTDB-Tk-Methode den Archaea zugeordnet [74]. Der *checkM*-Algorithmus klassifizierte viele *Candidatus Saccharibacteria* fälschlicherweise als *Dhecalococcoidia* (siehe digitaler Anhang [41]).

3.3.4 Funktionale Zuordnung von Proteinsequenzen zu *Clusters of Orthologeous Groups*

Die Proteinsequenzen der Gesamtassemblierung und der einzelnen *Chloroflexi-Bins* wurden „Clustern aus orthologen Gruppen“ (engl. *Clusters of Orthologous Groups* bzw. COGs) zugeordnet (Abbildung 8).

Einzelne *Bins* weisen im Vergleich zum Durchschnitt der *Chloroflexi-Bins* deutliche Unterschiede im relativen Anteil bestimmter funktioneller Kategorien auf. Bin02 beispielsweise weist einen deutlich erhöhten Anteil an Genen des Lipid- sowie einen schwach erhöhten Anteil an Genen des Koenzym- und Sekundärmetabolismus auf. Außerdem wurden in diesem *Bin* vergleichsweise mehr Gene der Zellfunktionen „Motilität“ sowie „intrazellulärer Transport“ gefunden. Im Gegenzug sind dort etwas niedrigere Anteile an Genen der funktionellen Kategorien „posttranslationale Modifikationen“, „Transkription“, „Verteidigungsmechanismen“ und „Metabolismus anorganischer Ionen“ aufzufinden.

Bei bin04 und bin10 sind Gene der Signaltransduktion häufiger als in den anderen *Chloroflexi-Bins* anzutreffen. Bei bin04 sind im Gegenzug Proteine der Funktionen für „posttranslationale Modifikationen“ und des „Metabolismus anorganischer Ionen“ im Vergleich zu den übrigen *Chloroflexi-Bins* in geringerer Anzahl anzutreffen.

Alle *Bins* weisen erhöhte Anteile der Kategorien „Metabolismus anorganischer Ionen“ und „Aminosäuremetabolismus“ auf.

Bei „Motilität“, „intrazellulärer Transport“, „Translation“, „Replikation“ und Proteine der Zellwand/Membran liegen die durchschnittlichen Anteile der *Chloroflexi-Bins* deutlich unter dem der Gesamtassemblierung.

Mit diesen Erkenntnissen könnten bereits optimierte Kultivierungsbedingungen geschaffen werden, wodurch zuvor unkultivierbare *Chloroflexi* kultiviert werden könnten.

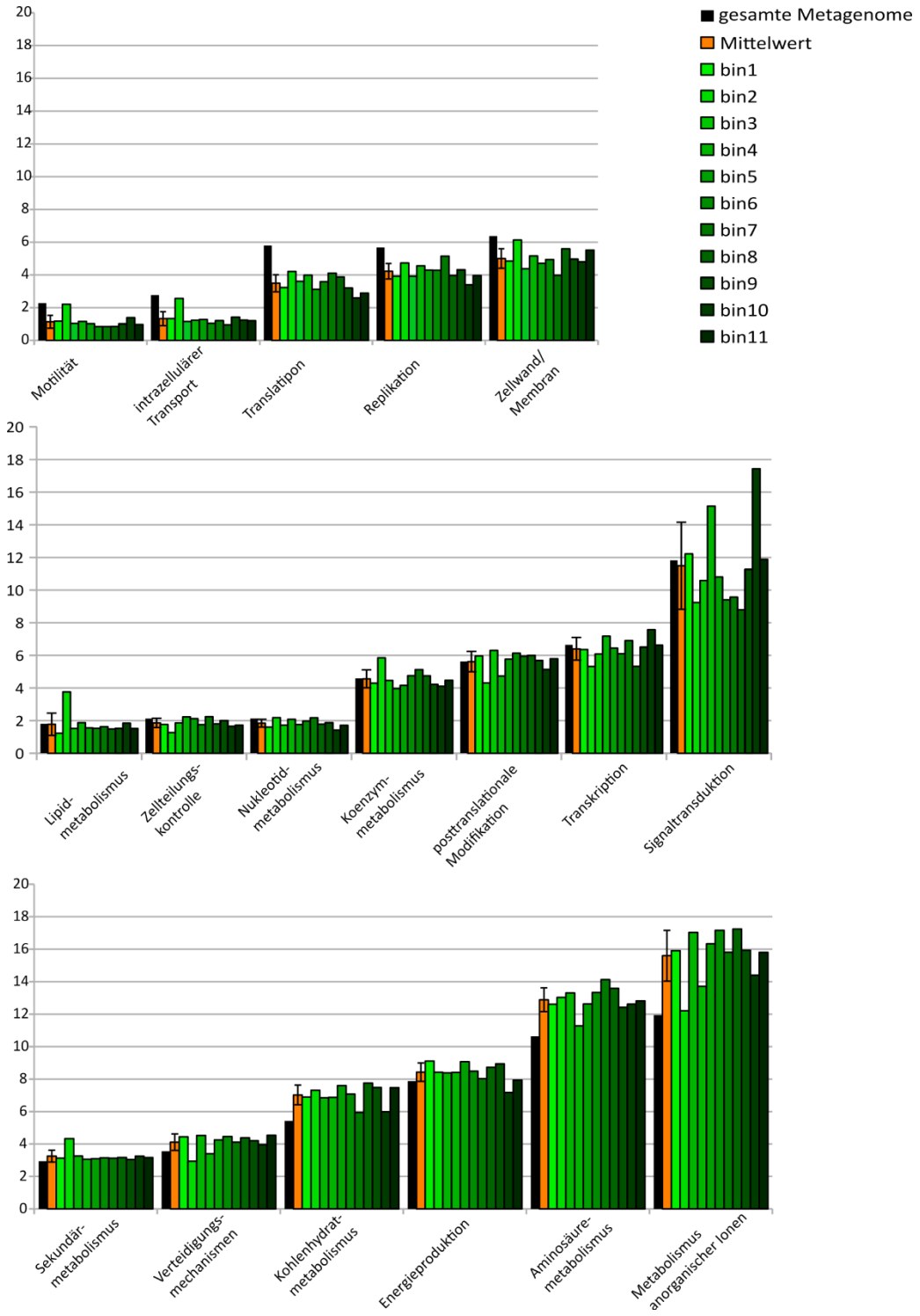

Abbildung 8. Relative Anteile von Proteinen verschiedener Zell-/Stoffwechselfunktionen in der Gesamtassemblierung bzw. den *Chloroflexi-Bins*. Aussagelose COG-Kategorien wie z. B. unklassifizierte Funktionen oder relative Anteile unter 1 % wurden zur Vereinfachung entfernt. Es sind das gesamte Metagenom, der Mittelwert der *Chloroflexi-Bins* und die einzelnen *Chloroflexi-Bins* dargestellt. Die erste Abbildung zeigt die Kategorien, welche in den *Chloroflexi-Bins* durchschnittlich weniger anzutreffen sind als im gesamten Metagenom. Das zweite Diagramm zeigt eine ähnliche Verteilung der COG-Kategorien zwischen den *Chloroflexi-Bins* und des gesamten Metagenoms und das letzte Diagramm zeigt die Kategorien, welche vermehrt in den *Chloroflexi-Bins* vorkommen.

3.3.5 SEKUNDÄRMETABOLIT-GENCLUSTER

Um vorhandene Sekundärmetabolite der *Chloroflexi-Bins* zu bestimmen wurde das webbasierte Programm antiSMASH ausgeführt (siehe Methodenteil 2.6.7). Hierbei konnten in den hochqualitativen *Chloroflexi-Bins* vier verschiedene Kategorien potentieller Sekundärmetabolit-Gencluster identifiziert und deren Vorkommen und relative Anteile mit denen ausgewählter *Chloroflexi*-Referenzgenome verglichen werden.

Eine für potentielle Antibiotika besonders relevante Kategorie sind *Non-ribosomal peptide synthetase cluster/Polyketidesynthase* (NRPS/PKS). In Bakterien werden viele Sekundärmetabolite von solchen Enzym-Clustern synthetisiert, welche von modularen biosynthetischen Genclustern codiert werden [69].

Unter die Gruppe der Terpene fallen unterschiedliche Synthasen, wie z. B. Mono-, Sesqui- und Diterpensynthasen [75]. Diese Enzyme können verschiedene Cyclisierungsreaktionen durchführen.

Ein weiterer identifizierter Sekundärmetabolit ist Bacteriocin. Bacteriocine sind antimikrobielle Peptide, welche von einigen Bakterien produziert werden können. Diese Peptide sind potentiell auch gegen antibiotikaresistente Stämme wirksam [76].

In sämtlichen *Bins*, sowie beiden *Chloroflexi*-Referenzgenomen wurde mindestens ein NRPS/PKS Sekundärmetabolit-Gencluster gefunden (Abbildung 9). Bin11 sticht hier mit drei NRPS/PKS clustern deutlich hervor.

Je eine Terpenesequenz wurde in bin02 und bin11 gefunden, auch in bin10 mit drei Sequenzen und *Bellilinea caldifistulae* mit zwei gefundenen Proteinsequenzen sind Terpene vorhanden. Bacteriocin-Gencluster wurden lediglich in bin11 detektiert.

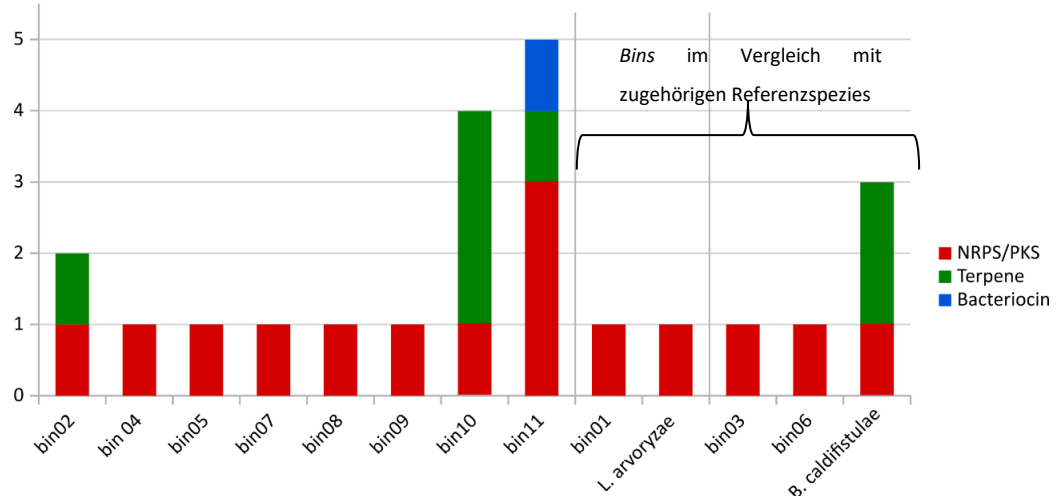

Abbildung 9. Anzahl der detektierten Sekundärmetabolite mittels antiSMASH. Gefunden wurden Gencluster der Kategorien Bacteriocin, Terpene und NRPS/PKS. *Bins* sind anhand ihrer Ähnlichkeit zu potentiellen nächstverwandten Referenzen (wenn vorhanden) gruppiert. Für bin01 wurde *Longilinea arvoryzae*, für bin03 und bin06 *Bellilinea caldifistulae* als Vergleichspezies gewählt.

4 Diskussion

4.1 Metagenomdaten

Nicht alle in öffentlichen Datenbanken hinterlegten Genome und Metagenome werden mit derselben Qualität und/oder Quantität an Metadaten versehen. Grundlegende Informationen, welche in Form von Metadaten idealerweise beigefügt sein sollten, sind u.a.: Name und Verantwortlicher des Projekts, Datum und Ort (mit Längen- und Breitegrad) der Proben-Entnahme, Art des Datensatzes (z. B. Metagenom, Transkriptom, Amplikons), Beschreibung der Probenumgebung (z. B. terrestrisch, maritim, human), verwendete Sequenziertechnologie und eine detaillierte Projektbeschreibung. Ohne ausreichende Metadaten lässt sich die Herkunft eines Datensatzes nicht genau bestimmen, was die Vergleichbarkeit mit anderen Datensätzen und die Nutzbarkeit für weiterführende Analysen stark einschränkt. Aufgrund nicht ausreichender Metadaten können viele der bei *MG-RAST* und *NCBI* hinterlegten Datensätze nicht für weiterführende vergleichende Analysen verwendet werden.
Zudem waren einige vermeintlich öffentlich verfügbare terrestrische Datensätze tatsächlich nicht zugänglich, da diese mit potentiell menschlichen DNA-Fragmenten verunreinigt waren. *MG-RAST* verfolgt hier eine strikte Politik, um keine Sequenzinformationen der entsprechenden menschlichen Ursprungs-Genome preiszugeben und so den Datenschutz der betreffenden Person(en) zu wahren.

Im Rahmen dieser Bachelorarbeit waren letztendlich elf Ethanol-Kraftstoff-kontaminierte Grundwasser-Metagenomdatensätze der *MG-RAST*-Datenbank nutzbar. Diese wiesen der *MG-RAST* Analyse-*Pipeline* zufolge relativ hohe Anteile an *Chloroflexi* auf. Es ist bekannt, dass einige *Chloroflexi* bestimmte Schadstoffe abbauen können und somit potential für sogenannte *Bio-Remediation* [68] bergen. Spezielle *Chloroflexi* bzw. hohe *Chloroflexi*-Vorkommen in Grundwasser-Habitaten wurden bislang noch nicht explizit beschrieben. Dies weckte den Verdacht eventuell potentiell neue und ökologisch sowie industriell relevante *Chloroflexi*-Vertreter vorfinden zu können. Jedoch stammten alle elf verbleibenden *MG-RAST* Datensätze aus demselben Projekt sowie demselben Beprobungsgebiet (einem Öl-kontaminiertem Habitat in Texas) und sind somit nicht unabhängig voneinander. Diese konnten mit einem unabhängigen vergleichbaren Grundwasser-Metagenomdatensatz aus der *NCBI*-Datenbank komplementiert werden, welches aus der Umgebung einer kanadischen Mülldeponie stammt. Aufgrund der Assoziation mit einer Mülldeponie wurde dort eine Schadstoffbelastung vermutet, jedoch lagen keine Metadaten vor, die diese Annahme belegen oder widerlegen konnten.
In verschiedenen Publikationen, welche Grundwasser in der Nähe von Mülldeponien untersuchten, wurden jedoch vor allem verschiedene Schwermetalle und erhöhte Chlorid-, Sulfat- und Eisen-Konzentrationen beschrieben [77,78]. Es wurde daher erhofft, sowohl in den texanischen

Öl-kontaminierten Grundwasser-Datensätzen (TOGW) aus *MG-RAST* als auch in dem kanadischen Mülldeponie-Grundwassermetagenom (KMGW) verschiedene Anteile an Genomen ähnlicher, aber teilweiser auch unterschiedlicher, Schadstoff-abbauender *Chloroflexi* zu finden. Wodurch sich diese dann anhand differentieller *Coverage-Binning* Methoden (siehe Methodenteil 2.4) rekonstruieren und eventuell den unterschiedlichen Schadstoffbelastungen zuordnen lassen konnten.

4.2 Analyse der gesamten bakteriellen Gemeinschaften

Mit Blick auf die hier erlangten *Chloroflexi*-Anteile (Abbildung 2) lässt sich sagen, dass die texanischen Öl-kontaminierten Grundwassermetagenomdatensätze (TOGW) deutlich mehr *Chloroflexi*-Anteile als das kanadische Mülldeponie-Grundwassermetagenom (KMGW) aufwiesen. Vergleicht man die *Chloroflexi*-Anteile mit weiteren *MG-RAST*-Datensätzen, welche aus nicht belasteten Grundwasserhabitaten stammten, liegt der *Chloroflexi*-Anteil in den Schadstoff-belasteten Grundwassermetagenomen TOGW1-11 höher (Abbildung 10). Daher lässt sich vermuten, dass die Ethanol-Kraftstoff Belastung ausschlagebend für den erhöhten *Chloroflexi*-Anteil ist.

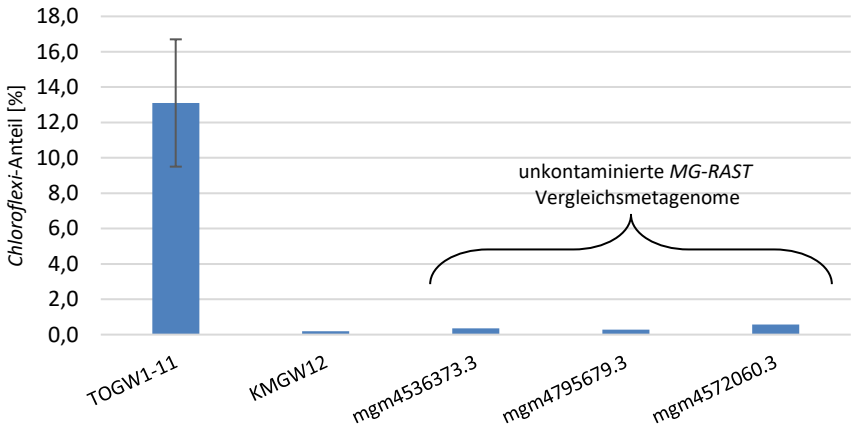

Abbildung 10. *Chloroflexi*-Anteile der prozessierten Schadstoff-belasteten Grundwassermetagenome TOGW1-11 und KMGW12 im Vergleich zu nicht Schadstoff-belasteten Grundwassermetagenomen aus der *MG-RAST* Datenbank. Als Bezeichnung wurden für diese Datensätze die *MG-RAST* IDs verwendet.

Um hier validierte Vergleichsdaten zu erhalten müssten die nicht belasteten Vergleichsmetagenome analog zu TOGW1-11 und KMGW12 behandelt werden, da der *MG-RAST*-Algorithmus den Anteil der *Chloroflexi* anhand der jeweils unassemblierten gesamt-*Reads* bestimmt. Kurze *Reads* besitzen im Vergleich zu langen *Reads* weniger Sequenzkontext und eignen sich nicht für ein wirklich zuverlässiges *Alignment*. Da alle *Reads* und nicht nur konservierte Markergene, klassifiziert werden, ist das Potential hoch, durch niedrig-konservierte oder in den

Datenbanken kaum repräsentierte Gene, falsche taxonomische Zuweisungen zu erhalten. Darüber hinaus werden die relativen Anteile der Organismen durch unterschiedliche Genomgrößen verzerrt, da ein großes Genom mehr *Reads* erzeugt als ein kleines, selbst wenn beide Genome in derselben Kopienzahl vorliegen. Aus diesem Grund sind Analysen von konservierten Markergenen, insbesondere wenn sie von assemblierten Daten stammen, zuverlässiger. Durch die Prozessierung und Assemblierung der ausgewählten Metagenomdatensätze vor anschließender taxonomischer Einteilung wird das Problem der kurzen *Reads* umgangen. Beispielsweise ist durch die Assemblierung ein größerer Sequenzkontext vorhanden, der bessere und aussagekräftigere *Alignements* ermöglicht.

Der *Chloroflexi*-Anteil von KMGW12 ähnelt mehr nicht belasteten Vergleichsmetagenomen (Abbildung 10). Die Probenumgebung von KMGW12 ist zwar in der Nähe einer Mülldeponie, doch ist nicht klar, ob diese genau unterhalb der Deponie liegt oder eventuell ein Stück davon entfernt. Somit liegt der zugehörige Probenort möglicherweise nicht im Einflussbereich des Schadstoffaustrags der Mülldeponie, worauf der geringe *Chloroflexi*-Anteil in KMGW12 zurückzuführen sein könnte. Außerdem ist es denkbar, dass ausreichende Versiegelungs- bzw. Schutzmaßnahmen getroffen wurden, um eine Schadstoffbelastung des Grundwassers durch diese Mülldeponie zu vermeiden. Wären die Proben auf eventuelle Schadstoff-Konzentrationen untersucht und den Metadaten beigefügt worden, könnte man validierte Aussagen treffen. Mit den vorhandenen Metadaten lassen sich lediglich Vermutungen äußern.
Obwohl sich die texanischen Öl-kontaminierten Grundwassermetagenome im relativen hohen Anteil von durchschnittlich ca. 13 % *Chloroflexi* sehr ähneln, weisen sie teilweise gravierende Unterschiede in der Verteilung anderer bakteriellen Taxa, insbesondere PVC-, *Firmicutes*- und „*Candidatus*“-Phyla auf (siehe Ergebnissteil 3.2.1 und Abbildung 2). Ob die Unterschiede in der Verteilung mit räumlichen oder zeitlichen Unterschieden der entsprechenden Probennahmen zusammenhängen, kann aus den angegebenen Metadaten nicht zweifelsfrei bestimmt werden. Um dies zu überprüfen, müssten aus denselben Standorten erneut Proben entnommen, sequenziert und analysiert werden.

Die in der Metagenom-Koassemblierung identifizierten *Chloroflexi* 16S rRNA Gensequenzen clustern vor allem innerhalb der Klassen *Anaerolineae* bzw. in der Nähe der *Dehalococcoidia* (Abbildung 4). Dies deckt sich mit Publikationen welche ähnliche Schadstoff-belasteten Habitate untersuchten und in denen *Anaerolineae* eine häufig vorkommende Klasse sind [31–33], aber sich auch *Dehalococcoidia* vermehrt wiederfinden [79,80].
Um zu überprüfen ob die vorgefundenen *Chloroflexi* potentiell neue, bislang nicht beschriebene, Spezies, Genera oder sogar teilweise neue Ordnungen und Klassen repräsentieren, wurden direkte 16S rRNA Gensequenzvergleiche mit beschriebenen Referenzen durchgeführt (Abbildung 6 und Abbildung 7). Ab einer Identität von 99 % wird von der gleichen Spezies ausgegangen. Bei einer

Identität von 96% bis 99% wird vom gleichen Genus ausgegangen. Bei 93 bis 96% wird von derselben Familie, ab 92% von derselben Ordnung und ab 82% von derselben Klasse ausgegangen [81,82]. Die genauen *Cutoff*-Werte dieser phylogenetischen Einordnung sind jedoch teilweise umstritten [82].

Metagenom-rRNA-Sequenz 1, welche mit hoher Bootstrap-Konfidenz und einer Genomidentität von 96% bei *Bellilinea caldifistulae* clustert, repräsentiert somit offenbar eine neue Spezies innerhalb des Genus *Bellilinea*.
Metagenom-rRNA-Sequenz 2 clustert mit hoher Konfidenz und einer Genomidentität von 92% bei *Longilinea arvoryzae*, wodurch Metagenom-rRNA-Sequenz 2 vermutlich eine neue Spezies innerhalb der Ordnung *Anaerolineales* repräsentiert.
Metagenom-rRNA-Sequenzen 3 und 4 clustern jeweils bei *Pelolinea submarina* bzw. *Leptolinea tardivitalis* und haben jeweils eine Identität von 88% bzw. 91% und repräsentieren somit neue Spezies der Klasse *Anaerolinea*.
Metagenom-rRNA-Sequenzen 5 und 6 clustern bei *Dehalococcoidia*. Die Identitäten betragen 80% bzw. 85%. Metagenom-rRNA-Sequenzen 5 konnte somit lediglich dem Phylum *Chloroflexi* zugeordnet werden. Metagenom-rRNA-Sequenzen 6 repräsentiert vermutlich eine neue Spezies der *Anaerolinea*.

Bei der Berechnung der 16S rRNA Phylogenien müssen die stark konservierten 16S Gensequenzen beachtet werden. Diese Bereiche können beim *Binning* oft nicht unterschieden werden und gehen dabei möglicherweise verloren oder können bei der taxonomischen Klassifizierung falsch klassifiziert werden, weshalb später nicht jede Metagenom-rRNA-Sequenz einem *Bin* zugeordnet werden konnte.

4.3 Analyse der *Binning*-Ergebnisse

Ein Vergleich der gesamt-*Binning*-Ergebnisse der verwendeten *Binning*-Methoden *MaxBin* und *MetaBAT* (Abbildung 11 bzw. digitaler Anhang [41]) zeigt teilweise Unterschiede in der jeweiligen Effizienz auf. Die *Binning*-Ergebnisse sind im digitalen Anhang zu finden [41].

MetaBAT hat einen höheren *Cutoff*-Wert der *Contig*-Größe als *MaxBin*, weshalb hier weniger *Contigs* vorzufinden sind. Da lediglich kleine *Contigs* entfernt wurden hat dies kaum Einfluss auf die Gesamtlänge der *Contigs*, weshalb diese bei beiden Methoden in einem ähnlichen Bereich liegen. Insgesamt konnten mehr als die Hälfte der *Contigs* und sogar mehr als 80 % des Gesamt-Nukleotidsequenz *gebinnt* werden. Nicht *gebinnt* wurden vor allem sehr kleine *Contigs*, welche jedoch die spätere Gesamt-*Contig*-Länge und somit die erhaltene Sequenzinformation nicht beeinflussten. Mit Hilfe von *MaxBin* wurden zwar deutlich mehr *Contigs* als durch *MetaBAT gebinnt*, daraus resultierte aber nicht deutlich mehr Sequenzinformationen, wodurch sich lässt der höhere *Contig-Cutoff*-Wert von *MetaBAT* erklären lässt [51].

Für weitere Analysen erscheint es sinnvoll, mehrere *Binning*-Methoden zu kombinieren, um so den maximalen Informationsgehalt aus den Metagenomdatensätzen zu erhalten.

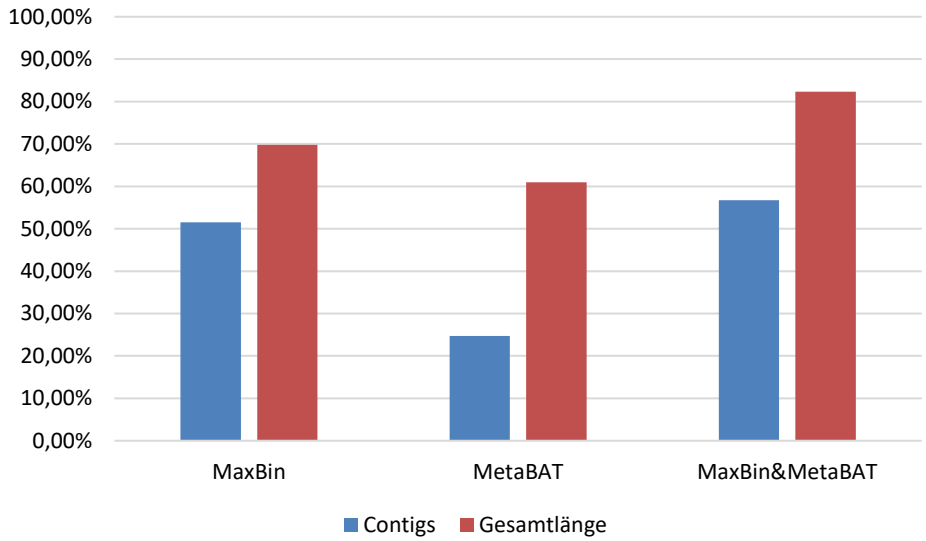

Abbildung 11. Effizienz des Gesamtmetagenom-*Binnings*. Dargestellt ist jeweils der Anteil des Gesamt-metagenoms, welches sich mittels verschiedener Methoden *Bins* zuordnen ließ. In blau dargestellt der Anteil *gebinnter Contigs*, in Rot der Anteil der summierten Gesamt-*Contig*-Länge.

Werden die Ergebnisse der taxonomischen Klassifizierung betrachtet (digitaler Anhang [41]) werden Unterschiede in der taxonomischen Zuordnung deutlich. *CheckM* klassifizierte deutlich mehr *Bins* als die beiden anderen Methoden. Jedoch nutzt *checkM* [53] zur Klassifizierung eine veraltete (letztes Update 2015) Referenzdatenbank. Diese beinhaltet u.a. unkultivierte „*Candidatus*“ Vertreter und ist somit unvollständig. Dies führt dazu, dass viele Vertreter der „*Candidate Phyla radiation*“ [83] falsch zugeordnet werden (in diesem Fall z. B. *Candidatus Saccharibacteria*, welche von *checkM* als *Dehalococcoidia* klassifiziert wurden).

Die HCC und die GTDB-TK Methoden weisen ähnliche Ergebnisse auf, wobei aber HCC 2 *Bins* mehr als *Chloroflexi* klassifizierte als GTDB-Tk. Eine zeitgleiche Bachelorarbeit wies problematische GTDB-Tk Klassifikationen auf, bei denen eindeutig *Chloroflexi-Bins* fälschlicherweise als *Archaea* interpretiert wurden [74].

Aus diesem Grund beruhen die weitere Identifikation und Auswahl von *Chloroflexi-Bins* daher ausschließlich auf der HCC Methode.

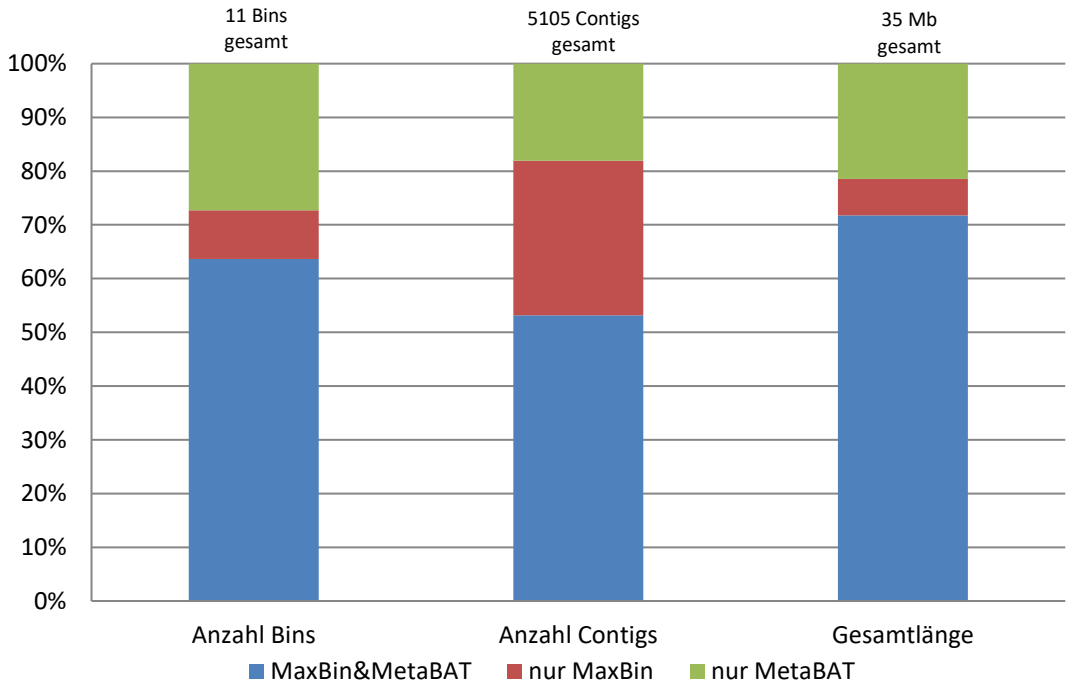

Abbildung 12. Relative Anteile der *Bins*, der *Contigs* und der Gesamtlänge bezüglich der angewandten *Binning*-Methoden.

4.4 Analyse der rekonstruierten *Chloroflexi*-Genome (*Bins*)

4.4.1 Allgemeine Genomeigenschaften

Um die korrekte Zuordnung der hier rekonstruierten *Chloroflexi*-Genome zu überprüfen, wurden diese mit den jeweils nächstverwandten Referenzen verglichen. Bin01 wurde mittels HCC (und vergleichsweise hohem Konfidenzwerten) der Spezies *Longilinea arvoryzae* zugeordnet (Tabelle 5), was sich auch mit den entsprechenden Markergenprodukt-Phylogenien deckt (Abbildung 6, Abbildung 7). Anhand der entsprechenden *Bin*-Größe und geschätzten Genomvollständigkeit ist von einer theoretischen Gesamthenomgröße von 4,2 Mb zu rechnen, was mit der Genomgröße des *L. arvoryzae* Typstamms von 4,4 Mb nahezu übereinstimmt [84]. Auch der GC-Gehalt von bin01 ähnelt mit 58,08 % den bei verschiedenen *L. arvoryzae* Stämmen beobachteten Werten von 56,84 % [84] und 57,6 % [34]. Es lässt sich somit vermuten, dass es sich bei bin01 eventuell um einen neuen Stamm der Spezies *L. arvoryzae*, zumindest aber um eine nahverwandte Spezies derselben Gattung handelt. Metagenom-rRNA-Sequenz 2 weist eine ähnliche phylogenetische Clusterung wie die Markergenprodukt-Phylogenien von bin01 auf, wodurch von einer neuen Spezies der Gattung *Longilinea* auszugehen ist, da die 16S rRNA-Genidentität mit 96 % unter den gängigen Spezies-*Cutoffs* liegt.

Bei der HCC-Methode wurde für bin02, mit sehr niedriger Konfidenz, als nächstverwandte Spezies *Ktedonobacter racemifer* identifiziert. Mit hoher Konfidenz konnte dieser lediglich auf Phylum-Ebene als *Chloroflexi* klassifiziert werden, was dafür spricht, dass dieser *Bin* eine neue Klasse repräsentiert. Diese Annahme wird ebenfalls durch die Markergen-Phylogenie bestärkt, da bin02 ausschließlich bei bislang unbeschriebenen *Chloroflexi*-Vertretern clustert. Darüber hinaus wird diese Vermutung durch die theoretische Genomgröße von 4,06 Mb und den GC-Gehalt von 67,27 % unterstützt, welche stark von den Werten des Typstamms der *Ktedonobacteria Ktedonobacter racemifer* mit einer Genomgröße von 13 Mb und einem GC-Gehalt von 54 % abweichen [85]. Interessant ist ebenfalls, dass sich keine 16S rRNA Gensequenz des Gesamtmetagenoms den *Ktedonobacteria* zuordnen ließ. Werden die vergleichsweise hohen Werte für „angepasste Kontamination“ und „Stammheterogenität“ dieses *Bins* miteinbezogen, lässt sich wiederum vermuten, dass bin02 eventuell ein Konsensusgenom mehrerer nahverwandte Stämme derselben Spezies oder zumindest mehrere sehr nah verwandte Spezies einer neuen *Chloroflexi*-Klasse darstellt.

Bins 03, *06*, *08* und *09* wurden in der taxonomischen Klassifizierung nach HCC mit hoher Konfidenz der Phylum-Ebene zuordnen. Diese *Bins* clusterten in den Markergen-Phylogenien mit teilweise niedrigen Konfidenzwerten bei der Klasse *Anaerolineae*, wodurch zumindest die Zugehörigkeit zum Phylum *Chloroflexi* bestärkt wird. Die theoretische Genomgröße betrug zwischen 2,7 Mb und 3,7 Mb. Der GC-Gehalt lag hier zwischen 50,01 % und 55,58 %. Zum

direkten Vergleich wurden hier die Typstämme der *Anaerolineae*-Vertreter *Anaerolinea thermophila* [27], *Bellilinea caldifistulae* [28] und *L arvoryzae* [27, 28, 84] herangezogen, welche Genomgrößen von 1,3 Mb bis 3,8 Mb und GC-Gehalt zwischen 53,3 % und 57,6 % aufwiesen. Sowohl die Genomgröße als auch der GC-Gehalt der *Bins* 03, 05-09 liegen also im Bereich der Literaturwerte, wodurch vermutet werden kann, dass diese *Bins* sogar der Klassen-Ebene *Anaerolineae* zugeordnet werden könnten. Auch in der 16S rRNA Gensequenz basierte Phylogenie clustern Metagenom-rRNA-Sequenzen 1, 3 und 4 ebenfalls bei *Anaerolineae*. Obwohl auf der hier vorliegenden Datengrundlage keine genaue Zuordnung dieser drei Metagenom-rRNA-Sequenzen zu bestimmten den *Bins* 03, 05-09 möglich ist, könnte dies ein weiterer Hinweis sein, dass es sich bei den genannten *Bins* um neue Vertreter der Klasse *Anaerolineae* handeln könnten, zumindest aber um neue Vertreter des Phylum *Chloroflexi*.

Bins 04 und 10 wurden nach der HCC-Methode mit niedrigen Konfidenzen der Klassen-Ebene *Anaerolineae* zugeordnet, weshalb auch hier die oben genannten Spezies als Vergleichsliteratur dienten. Die rechnerische Genomgröße betrug 3,1 Mb bzw. 5,5 Mb und der geschätzte GC-Gehalt betrug 49,56 % bzw. 56,74 %. Sowohl die theoretische Genomgröße als auch der GC-Gehalt liegen mit Abweichungen im Rahmen der Literaturwerte. Deutlich außerhalb liegt die Genomgröße von bin10. Dies lässt vermuten, dass bin04 den Vertretern der Klasse *Anaerolineae* ähnelt. Bei bin10 könnte aufgrund der hohen Genomgröße vermutlich einer neuen *Chloroflexi*-Klasse angehören.

Bins 05 und 07 konnten mittels HCC-Methode mit hoher bzw. niedriger Konfidenz lediglich bis auf Phylum-Ebene *Chloroflexi* zugeordnet werden. Den Marker-Genproduktphylogenien zufolge clustert bin05 innerhalb der Klasse *Anaerolineae* und bin07 bei unklassifizierten *Chloroflexi*. Die theoretischen Genomgrößen lagen hier bei 4,0 Mb bzw.3,1 Mb. Der GC-Gehalt lag bei 49,8 % bzw. 71,59 %. Werden wieder die oben genannten Typstämme für den Vergleich herangezogen, sticht der GC-Gehalt von bin07 deutlich hervor. Dies lässt vermuten, dass es sich bei bin07 um eine neue *Chloroflexi*-Klasse handelt. Bin05 wiederum repräsentiert vermutlich ein neuer Vertreter der Klasse *Anaerolineae.*

Für bin11 wurde mit sehr niedriger Konfidenz die Gattung *Anaerolinea* identifiziert. Mit höherem Konfidenzwert wurde die Familie *Anaerolineaceae* identifiziert, weshalb angenommen wurde, dass bin11 diese repräsentiert. In den phylogenetischen Markergen-Stammbäumen clustert bin11 mit niedrigen *bootstrap*-Werten innerhalb der Klasse *Anaerolineae* bzw. bei unklassifizierten *Chloroflexi*, was bereits darauf hindeutet, dass es sich um eine neue *Chloroflexi*-Klasse handeln könnte. Als Referenzspezies wurden die Typstämme der Spezies *A. thermophila* [27] und *Anaerolinea thermolimosa* [84] herangezogen. Im Vergleich zu den Referenzspezies weist bin11 eine deutlich höhere theoretische Genomgröße von 5,7 Mb und einen höheren GC-Gehalt 58,54 % auf, wodurch die Vermutung bestärkt wird, dass bin11 eine unbekannte *Chloroflexi*-Klasse repräsentieren könnte.

Auffällig erscheint die Abwesenheit von *Dehalococcoidia-Bins*, welche eigentlich gerade in Schadstoff-belasteten Habitaten zu vermuten gewesen wären [86].

4.4.2 FUNKTIONELLE ANALYSEN

Proteinsequenzen der *Bins* wurden verschiedenen COG-Kategorien zugeordnet. Diese Zuordnung ist eine erste Möglichkeit, eine grobe Übersicht über vorhandene Metabolismen und andere Eigenschaften, wie z. B. Motilität und Signaltransduktion, zu erhalten. Es steht zu hoffen, dass diese Informationen genutzt werden können, um die Kultivierungsbedingungen für zukünftige Isolierungsversuche zu optimieren.

Wird lediglich der Mittelwert der relativen Anteile von Proteinen verschiedener Zell- und Stoffwechselfunktionen der *Bins* (Abbildung 8) im Vergleich zum gesamten Metagenom betrachtet werden teilweise verschiedene Tendenzen der relativen Anteilen deutlich.

Höhere, zum Teil signifikante relative Anteile des Mittelwerts der *Chloroflexi-Bins* im Vergleich zum gesamten Metagenom sind in den folgenden Zell- und Stoffwechselfunktionen-Kategorien vorzufinden: „Sekundärmetabolismus“, „Verteidigungsmechanismen“, „Kohlenhydrat-metabolismus“, „Energieproduktion“, „Aminosäuremetabolismus“ und „Metabolismus anorganischer Ionen“. Hier könnte möglicherweise die genaue Ionenzusammensetzung eines möglichen Nährmediums von Bedeutung sein.

Bei „Motilität“, „intrazellulärer Transport“, „Translation“, „Replikation“ und „Proteine der Zellwand/Membran“ liegen die durchschnittlichen Anteile der *Chloroflexi-Bins* deutlich unter dem der Gesamtassemblierung. Geringere Anteile an Proteinen der Replikation und Translation lassen einen langsameren Stoffwechsel und somit eine geringere Zellteilungsrate vermuten.

Bei den übrigen Stoffwechselfunktionen und Eigenschaften sind keine allgemeinen signifikanten Unterschiede festzustellen, dennoch gibt es innerhalb der *Bins* Schwankungen, insbesondere bin02 ist zu erwähnen.

Bin02 hat einen deutlich erhöhten Lipidmetabolismus, einen erhöhten Koenzym- und Sekundärmetabolismus und zeigt erhöhte relative Anteile bezüglich der Motilität im Vergleich zu den gesamten Metagenomen, aber auch zu den anderen *Bins* (Abbildung 8). Es ist jedoch fraglich, ob dahinter der Abbau oder die Synthese der jeweiligen Stoffwechsel steht. Jedoch können diese Erkenntnisse nützlich sein, um entsprechende Nährmedien zur Kultivierung dieses Organismus zu entwickeln.

Für bin01 wurde die nah verwandte Spezies *Longilinea arvoryzae* als Referenzspezies bezüglich ihrer Zell- und Stoffwechselfunktionen analysiert (Abbildung 13). Hier wird lediglich ein erhörter Anteil der Signaltransduktion Proteine in bin01 deutlich. Es könnte hierbei eine erhöhte Zell-Zell-Kommunikation vermutet werden, was evtl. ein Hinweis auf mögliche Symbiosen ist oder auf

einen erhöhten Austausch mit anderen Organismen hindeutet. Alle weiteren Kategorien weisen eine ähnliche Verteilung auf.

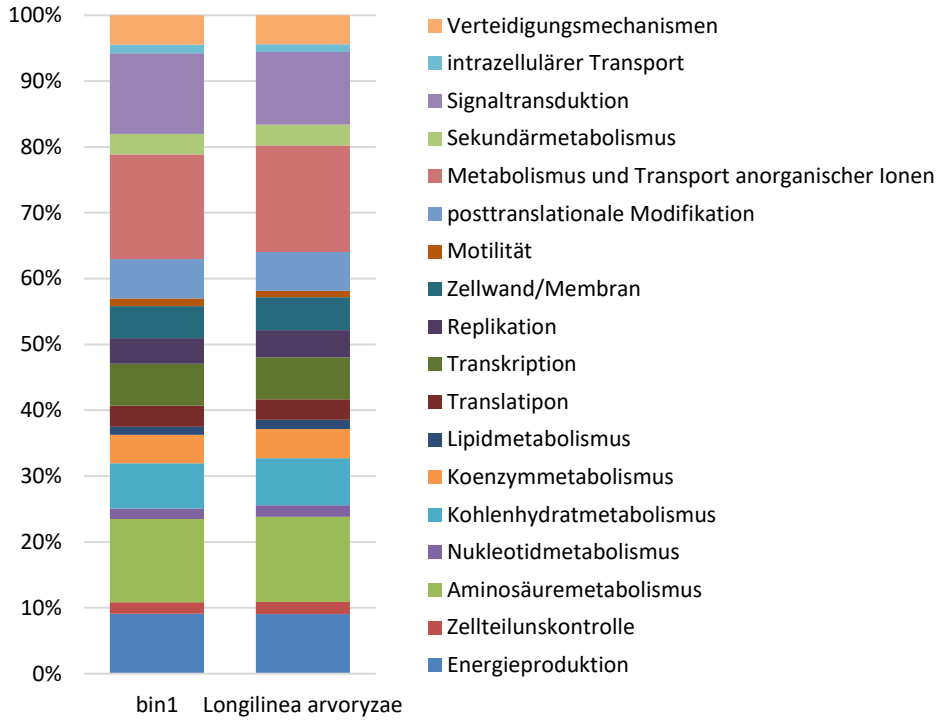

Abbildung 13. Relativer Anteil der COG-Kategorien für bin01 im Vergleich zur Referenzspezies *Longilinea arvoryzae*.

In allen hier beschriebenen *Chloroflexi-Bins* wurden NRPS/PKS-Cluster ermittelt (Abbildung 9). Es ist anzunehmen, dass NRPS/PKS-Cluster ein charakteristisches Merkmal für *Chloroflexi* sind, da diese in jedem einzelnen ausgewählten *Chloroflexi-Bin* sowie in beiden analysierten Referenzstämmen vorzufinden waren.

Solche Gencluster sind für die biotechnologische Forschung von großem Interesse, da diese die Synthese komplexer bioaktiver Peptidverbindungen katalysieren, welche oft antimikrobielle Wirkungen haben und eventuell als Antibiotika dienen können. Darüber hinaus können diese Erkenntnisse bei der Isolierung des jeweiligen Organismus helfen, da dieser selbst dagegen resistent ist.

4.5 Fazit

In der vorliegenden Arbeit wurden neue, bisher nicht beschriebene *Chloroflexi*-Genome identifiziert. Diese repräsentieren vermutlich vollkommen neue Klassen, Familien und Genera. Darüber hinaus wurden erste Einblicke über interessante Eigenschaften und Stoffwechsel dieser

Chloroflexi-Genome erhalten, welche vermutlich zur Kultivierung dieser *Chloroflexi* beitragen könnten und somit Grundlage für die weitere Erforschung von *Chloroflexi* ist.

5 Ausblick

Durch kostengünstige und schnelle Sequenziermethoden werden stetig neue Metagenomdaten produziert und in öffentliche Datenbanken hochgeladen. In dieser Arbeit wurden lediglich zwei Datenbanken genutzt, welche in nachfolgenden Projekten durch weitere Datenbanken, wie z. B. EBI Metagenomics [87] und gcMeta [88], ergänzt werden könnten. Außerdem wurde im Rahmen dieser Arbeit lediglich ein spezifisches Habitat analysiert. Auch hier sollten in nachfolgenden Projekten weitere Habitate analysiert werden.
Da lediglich zwei verschiedene Probenumgebungen analysiert wurden, ist es möglich, dass es sich bei den *Chloroflexi*-Anteilen um lokale Schwankungen handeln könnte. Um dies zu bestätigen müssten umfassende Probenentnahmen in unmittelbarer Nähe der beiden Entnahmestellen stattfinden und analog zu den bereits erhaltenen Proben sequenziert und bioanalytisch analysiert werden. Interessant wären hier weitere Grundwasser-Proben von Mülldeponien, sowohl kanadisch als auch von anderen Standorten, zu entnehmen und explizit auf ihren Schadstoffgehalt zu untersuchen, um so einen möglichen Zusammenhang zwischen Kontamination und *Chloroflexi*-Anteil bzw. *Chloroflexi*-Diversität zu bestätigen. Außerdem müssten zu weiteren vergleichbaren Schadstoff-belasteten Habitate auch mehr unbelastete Grundwasser-Habitate analysiert werden.

Auch die bioinformatische Prozessierung kann durch verschiedene weitere Methoden ergänzt und erweitert werden. Da, wie bereits erwähnt (Diskussionsteil 4.3), verschiedene *Binner* verschiedene *Bins* nicht erfassen können, könnte die parallele Nutzung weiterer *Binning*-Methoden, wie z. B. CONCOCT [89] zu höherer *Binning*-Effizienz führen. Besonders zu erwähnen ist in diesem Zusammenhang auch das Programm DASTool [90], welches Ergebnisse mehrerer *Binning*-Methoden vereinigt und so vermutlich eine höhere Vollständigkeit der *Bins* erzielt wird, was wiederum auch eine bessere Klassifizierung ermöglichen könnte.

Für die genaue phylogenetische und taxonomische Einordnung existieren ebenfalls weitere Klassifizierungsmethoden als die in dieser Arbeit genutzten (z. B. MLSA [91], Genecontent-Analysen [91] oder POCP [92]). Im Gegensatz zu rein auf 16S rRNA basierenden Phylogenie werden bei diesen Methoden möglichst viele verschiedene Marker bzw. Sequenzinformationen einbezogen, was die Auflösung sowie die Zuverlässigkeit der phylogenetischen Zuordnung erhöht. Die MLSA und POCP-Methoden kommen allerdings nur für möglichst vollständige *Bins* in Frage (idealerweise >80 %), Genecontent Analysen dagegen sind auch bei partiellen Genomen anwendbar.

Leider wird es weiterhin ein Problem sein, dass sich 16S rRNA Gensequenzen durch gängige *Binning*-Methoden nicht den effizient bestimmten Genomen bzw. *Bins* zuordnen lassen. Für die bessere nachträgliche Zuordnung von Metagenom 16S rRNA Gensequenzen zu bestimmten *Bins*

könnte jedoch auch eine detaillierte *Coverage*-Analyse herangezogen werden, in der geprüft wird, ob die Häufigkeitsverteilung bestimmter 16S rRNA Gensequenzen mit der Häufigkeitsverteilung bestimmter *Bins* auch über mehrere Probenorte bzw. -zeitpunkte übereinstimmt.

Bei der Berechnung der phylogenetischen Stammbäume wurde die *Neighbour-Joining* Methode verwendet. Als zuverlässigste Stammbaumberechnungsmethode gilt jedoch *Maximum-Likelihood* [93], wofür z. B. RaxML genutzt werden kann. Aus Zeitgründen wurde jedoch diese vergleichsweise rechenaufwendige Methode lediglich für die 16S rRNA Metagenomsequenzen angewandt.

Ebenfalls sind weitere funktionelle Analyse-Methoden, wie z. B. SEED System [94] und „GO::TermFinder“ [95] verfügbar. Mit diesen Methoden wäre es beispielsweise möglich, die hier erhaltenen hoch-qualitativen *Chloroflexi Bins* erneut auf NRPS/PKS-Cluster zu untersuchen, um so diese möglicherweise zu bestätigen. Hierdurch könnte die Vermutung bestärkt werden, dass die NRPS/PKS-Cluster möglicherweise in allen *Chloroflexi* vorkommen. Um diese Vermutung zu validieren wären jedoch deutlich mehr Analysen von *Chloroflexi*-Genomen nötig.

Die hier erhaltenen *Chloroflexi-Bins* sind basierend auf den vorliegenden Ergebnissen vielversprechend und sollten in weiteren Analysen näher betrachtet und analysiert werden, um so zur Vervollständigung des phylogenetischen Stammbaums beizutragen.

6 Erklärung

Ich versichere wahrheitsgemäß, die Arbeit selbstständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten Anderer unverändert oder mit Abänderungen entnommen wurde.

Karlsruhe, den 30.09.2019 _____

Dominik Hellmann

7 Literatur

[1] Madigan MT, Martinko JM, Stahl DA, Clark DP, Brock TD, Thomm M, Wirth R. Brock Mikrobiologie. 13th ed. München: Pearson; 2013.

[2] Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, Behrenfeld MJ, Boetius A, Boyd PW, Classen AT, Crowther TW, Danovaro R, Foreman CM, Huisman J, Hutchins DA, Jansson JK, Karl DM, Koskella B, Mark Welch DB, Martiny JBH, Moran MA, Orphan VJ, Reay DS, Remais JV, Rich VI, Singh BK, Stein LY, Stewart FJ, Sullivan MB, van Oppen, Madeleine J. H., Weaver SC, Webb EA, Webster NS. Scientists' warning to humanity: microorganisms and climate change. Nature Reviews Microbiology 2019;17(9):569–86.

[3] Solden L, Lloyd K, Wrighton K. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr Opin Microbiol 2016;31:217–26.

[4] Vollmers J, Wiegand S, Kaster A-K. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PLoS ONE 2017;12(1):e0169662.

[5] Daniel R. The metagenomics of soil. Nat Rev Microbiol 2005;3(6):470–8.

[6] Schmeisser C, Steele H, Streit WR. Metagenomics, biotechnology with non-culturable microbes. Appl Microbiol Biotechnol 2007;75(5):955–62.

[7] Simon C, Daniel R. Metagenomic analyses: past and future trends. Appl Environ Microbiol 2011;77(4):1153–61.

[8] Wilson MC, Piel J. Metagenomic approaches for exploiting uncultivated bacteria as a resource for novel biosynthetic enzymology. Chem Biol 2013;20(5):636–47.

[9] Metzker ML. Sequencing technologies — the next generation. Nature Reviews Genetics 2009;11:31 EP -.

[10] Hall BG. Predicting the evolution of antibiotic resistance genes. Nature Reviews Microbiology 2004;2(5):430–5.

[11] Frangeul L, Nelson KE, Buchrieser C, Danchin A, Glaser P, Kunst F. Cloning and assembly strategies in microbial genome projects. Microbiology (Reading, Engl) 1999;145 (Pt 10):2625–34.

[12] Forns X, Bukh J, Purcell RH, Emerson SU. How Escherichia coli can bias the results of molecular cloning: preferential selection of defective genomes of hepatitis C virus during the cloning procedure. Proc Natl Acad Sci U S A 1997;94(25):13909–14.

[13] Cardenas E, Tiedje JM. New tools for discovering and characterizing microbial diversity. Curr Opin Biotechnol 2008;19(6):544–9.

[14] Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods 2008;5(1):16–8.

[15] Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res 2011;166(2):99–110.

[16] Schrödel A. Darstellung von Bakterien mittels Gramfärbung. Biol. Unserer Zeit 2009;39(3):156.

[17] Sutcliffe IC. Cell envelope architecture in the Chloroflexi: a shifting frontline in a phylogenetic turf war. Environ Microbiol 2011;13(2):279–82.

[18] Gupta RS, Chander P, George S. Phylogenetic framework and molecular signatures for the class Chloroflexi and its different clades; proposal for division of the class Chloroflexia class. nov. corrected into the suborder Chloroflexineae subord. nov., consisting of the emended family Oscillochloridaceae and the family Chloroflexaceae fam. nov., and the suborder Roseiflexineae subord. nov., containing the family Roseiflexaceae fam. nov. Antonie Van Leeuwenhoek 2013;103(1):99–119.

[19] Garrity GM, Holt JG, Castenholz RW, Pierson BK, Keppen OI, Gorlenko VM. Phylum BVI. Chloroflexi phy. nov. In: Boone DR, Castenholz RW, Garrity GM (editors). Bergey’s Manual® of Systematic Bacteriology. New York, NY: Springer New York; 2001. p. 427–46.

[20] Hugenholtz P, Stackebrandt E. Reclassification of Sphaerobacter thermophilus from the subclass Sphaerobacteridae in the phylum Actinobacteria to the class Thermomicrobia (emended description) in the phylum Chloroflexi (emended description). Int J Syst Evol Microbiol 2004;54(Pt 6):2049–51.

[21] Moe WM, Yan J, Nobre MF, da Costa MS, Rainey FA. Dehalogenimonas lykanthroporepellens gen. nov., sp. nov., a reductively dehalogenating bacterium isolated from chlorinated solvent-contaminated groundwater. Int J Syst Evol Microbiol 2009;59(Pt 11):2692–7.

[22] Löffler FE, Yan J, Ritalahti KM, Adrian L, Edwards EA, Konstantinidis KT, Müller JA, Fullerton H, Zinder SH, Spormann AM (editors). Dehalococcoides mccartyi gen. nov., sp.

nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, Dehalococcoidia classis nov., order Dehalococcoidales ord. nov. and family Dehalococcoidaceae fam. nov., within the phylum Chloroflexi; 2013.

[23] Cavaletti L, Monciardini P, Bamonte R, Schumann P, Rohde M, Sosio M, Donadio S (editors). New lineage of filamentous, spore-forming, gram-positive bacteria from soil; 2006.

[24] Yabe S, Aiba Y, Sakai Y, Hazaka M, Yokota A. Thermosporothrix hazakensis gen. nov., sp. nov., isolated from compost, description of Thermosporotrichaceae fam. nov. within the class Ktedonobacteria Cavaletti et al. 2007 and emended description of the class Ktedonobacteria. Int J Syst Evol Microbiol 2010;60(Pt 8):1794–801.

[25] Kawaichi S, Ito N, Kamikawa R, Sugawara T, Yoshida T, Sako Y. Ardenticatena maritima gen. nov., sp. nov., a ferric iron- and nitrate-reducing bacterium of the phylum 'Chloroflexi' isolated from an iron-rich coastal hydrothermal field, and description of Ardenticatenia classis nov. Int J Syst Evol Microbiol 2013;63(Pt 8):2992–3002.

[26] Dodsworth JA, Gevorkian J, Despujos F, Cole JK, Murugapiran SK, Ming H, Li W-J, Zhang G, Dohnalkova A, Hedlund BP. Thermoflexus hugenholtzii gen. nov., sp. nov., a thermophilic, microaerophilic, filamentous bacterium representing a novel class in the Chloroflexi, Thermoflexia classis nov., and description of Thermoflexaceae fam. nov. and Thermoflexales ord. nov. Int J Syst Evol Microbiol 2014;64(Pt 6):2119–27.

[27] Yamada T, Sekiguchi Y, Hanada S, Imachi H, Ohashi A, Harada H, Kamagata Y. Anaerolinea thermolimosa sp. nov., Levilinea saccharolytica gen. nov., sp. nov. and Leptolinea tardivitalis gen. nov., sp. nov., novel filamentous anaerobes, and description of the new classes Anaerolineae classis nov. and Caldilineae classis nov. in the bacterial phylum Chloroflexi. Int J Syst Evol Microbiol 2006;56(Pt 6):1331–40.

[28] Yamada T, Imachi H, Ohashi A, Harada H, Hanada S, Kamagata Y, Sekiguchi Y. Bellilinea caldifistulae gen. nov., sp. nov. and Longilinea arvoryzae gen. nov., sp. nov., strictly anaerobic, filamentous bacteria of the phylum Chloroflexi isolated from methanogenic propionate-degrading consortia. Int J Syst Evol Microbiol 2007;57(Pt 10):2299–306.

[29] Vienne DM de. Lifemap: Exploring the Entire Tree of Life. PLoS Biol 2016;14(12):e2001624.

[30] Ward LM, Hemp J, Shih PM, McGlynn SE, Fischer WW. Evolution of Phototrophy in the Chloroflexi Phylum Driven by Horizontal Gene Transfer. Front Microbiol 2018;9:260.

[31] Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, Williams KH, Tringe SG, Banfield JF. Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome 2013;1(1):22.

[32] Tian M, Zhao F, Shen X, Chu K, Wang J, Chen S, Guo Y, Liu H. The first metagenome of activated sludge from full-scale anaerobic/anoxic/oxic (A2O) nitrogen and phosphorus removal reactor using Illumina sequencing. J Environ Sci (China) 2015;35:181–90.

[33] Fang H, Cai L, Yang Y, Ju F, Li X, Yu Y, Zhang T. Metagenomic analysis reveals potential biodegradation pathways of persistent pesticides in freshwater and marine sediments. Sci Total Environ 2014;470-471:983–92.

[34] Imachi H, Sakai S, Lipp JS, Miyazaki M, Saito Y, Yamanaka Y, Hinrichs K-U, Inagaki F, Takai K. Pelolinea submarina gen. nov., sp. nov., an anaerobic, filamentous bacterium of the phylum Chloroflexi isolated from subseafloor sediment. Int J Syst Evol Microbiol 2014;64(Pt 3):812–8.

[35] Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res 2017;45(D1):D12-D17.

[36] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 2008;9:386.

[37] Hong EL, Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, Rowe LD, Dreszer TR, Roe GR, Podduturi NR, Tanaka F, Hilton JA, Cherry JM. Principles of metadata organization at the ENCODE data coordination center. Database (Oxford) 2016;2016:baw001.

[38] Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, Vos P de, dePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone S-A, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, San Gil I, Wilson G, Wipat A. The minimum

information about a genome sequence (MIGS) specification. Nat Biotechnol 2008;26:541 EP -.

[39] Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 2013;31(6):533–8.

[40] Sharon I, Banfield JF. Genomes from Metagenomics. Science 2013;342(6162):1057.

[41] Hellmann D. Bachelorarbeit - digitaler Anhang: Rekonstruktion neuer Chloroflexi Genome aus kontaminiertem Grundwasser; Available from: https://github.com/DominikHe93/Thesis.git.

[42] Zhang Y. What is FASTA format; Available from: https://zhanglab.ccmb.med.umich.edu/FASTA/ (21 September 2019).

[43] Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 2010;38(6):1767–71.

[44] BamM - Working with the BAM; Available from: http://ecogenomics.github.io/BamM/ (20 August 2019).

[45] Wilke A, Bischof J, Harrison T, Brettin T, D'Souza M, Gerlach W, Matthews H, Paczian T, Wilkening J, Glass EM, Desai N, Meyer F. A RESTful API for accessing microbial community data for MG-RAST. PLoS Comput Biol 2015;11(1):e1004008.

[46] Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 2015;31(10):1674–6.

[47] Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 2016;102:3–11.

[48] Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Vardi MY, Weikum G, Raphael B, Tang J. Algorithms in Bioinformatics 2012;7534.

[49] El-Metwally S, Ouda OM, Helmy M. Next Generation Sequencing Technologies and Challenges in Sequence Assembly 2014;7.

[50] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078–9.

[51] Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 2015;3:e1165.

[52] Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 2016;32(4):605–7.

[53] Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 2015;25(7):1043–55.

[54] Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010;11:119.

[55] Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, Arumugam M, Pan Q, Liu B, Qin J, Wang J, Bork P. MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS ONE 2012;7(10):e47656.

[56] Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000;28(1):33–6.

[57] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;12(1):59–60.

[58] Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 2007;35(9):3100–8.

[59] Campbell MK, Farrell SO. Biochemistry. 6th ed. Belmont Calif.: Brooks/Cole; 2008.

[60] Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 2012;28(14):1823–9.

[61] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 2013;41(Database issue):D590-6.

[62] Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S,

Hermann S, Jost R, König A, Liss T, Lüssmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer K-H. ARB: a software environment for sequence data. Nucleic Acids Res 2004;32(4):1363–71.

[63] John Vollmers. Hirarchical Contig Classification (HCC) - noch nicht veröffentlicht.

[64] Ondov BD, Bergman NH, Phillippy AM. Krona: Interactive Metagenomic Visualization in a Web Browser. In: Nelson KE (editor). Genes, genomes and metagenomes: basics, methods, databases and tools: With 64 tables. New York, NY: Springer Reference; 2015. p. 339–46.

[65] GTDB-Tk; Available from: https://github.com/Ecogenomics/GTDBTk (19 September 2019).

[66] Andreas Leimbach. Bac-Genomics-Scripts: Bovine E. Coli Mastitis Comparative Genomics Edition. Zenodo; 2016.

[67] Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 2019;47(W1):W81-W87.

[68] Atlas RM. Bioremediation of petroleum pollutants. International Biodeterioration & Biodegradation 1995;35(1-3):317–27.

[69] Wagner M, Horn M. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr Opin Biotechnol 2006;17(3):241–9.

[70] Fuerst JA. The PVC superphylum: exceptions to the bacterial definition? Antonie Van Leeuwenhoek 2013;104(4):451–66.

[71] Taylor-Brown A, Vaughan L, Greub G, Timms P, Polkinghorne A. Twenty years of research into Chlamydia-like organisms: a revolution in our understanding of the biology and pathogenicity of members of the phylum Chlamydiae. Pathog Dis 2015;73(1):1–15.

[72] Gupta RS. The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes. Crit Rev Microbiol 2004;30(2):123–43.

[73] Munoz R, Rosselló-Móra R, Amann R. Revised phylogeny of Bacteroidetes and proposal of sixteen new taxa and two new combinations including Rhodothermaeota phyl. nov. Syst Appl Microbiol 2016;39(5):281–96.

[74] Jannick Fuchs. Genomrekonstruktion und Klassifizierung unkultivierter Chloroflexi aus Fumarolen von São Miguel (Azoren) 2019.

[75] Dickschat JS. Bacterial terpene cyclases. Nat Prod Rep 2016;33(1):87–110.

[76] Cotter PD, Ross RP, Hill C. Bacteriocins — a viable alternative to antibiotics? Nature Reviews Microbiology 2013;11(2):95–105.

[77] Nicholson RV, Cherry JA, Reardon EJ. Migration of contaminants in groundwater at a landfill: A case study 6. Hydrogeochemistry. Journal of Hydrology 1983;63(1-2):131–76.

[78] Mor S, Ravindra K, Dahiya RP, Chandra A. Leachate characterization and assessment of groundwater pollution near municipal solid waste landfill site. Environ Monit Assess 2006;118(1-3):435–56.

[79] Tas N, van Eekert MHA, Schraa G, Zhou J, Vos WM de, Smidt H. Tracking functional guilds: "Dehalococcoides" spp. in European river basins contaminated with hexachlorobenzene. Appl Environ Microbiol 2009;75(14):4696–704.

[80] van der Zaan B, Hannes F, Hoekstra N, Rijnaarts H, Vos WM de, Smidt H, Gerritse J. Correlation of Dehalococcoides 16S rRNA and chloroethene-reductive dehalogenase genes with geochemical conditions in chloroethene-contaminated groundwater. Appl Environ Microbiol 2010;76(3):843–50.

[81] Edgar RC. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ 2018;6.

[82] STACKEBRANDT E, GOEBEL BM. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int J Syst Evol Microbiol 1994;44(4):846–9.

[83] Danczak RE, Johnston MD, Kenah C, Slattery M, Wrighton KC, Wilkins MJ. Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. Microbiome 2017;5(1):112.

[84] Matsuura N, Tourlousse DM, Ohashi A, Hugenholtz P, Sekiguchi Y. Draft Genome Sequences of Anaerolinea thermolimosa IMO-1, Bellilinea caldifistulae GOMI-1, Leptolinea tardivitalis YMTK-2, Levilinea saccharolytica KIBI-1, Longilinea arvoryzae KOME-1, Previously Described as Members of the Class Anaerolineae (Chloroflexi). Genome Announc 2015;3(5).

[85] Chang Y-j, Land M, Hauser L, Chertkov O, Glavina Del Rio T, Nolan M, Copeland A, Tice H, Cheng J-F, Lucas S, Han C, Goodwin L, Pitluck S, Ivanova N, Ovchinikova G, Pati A, Chen A, Palaniappan K, Mavromatis K, Liolios K, Brettin T, Fiebig A, Rohde M, Abt B, Göker M, Detter JC, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk H-P, Lapidus A. Non-contiguous finished genome sequence and contextual data

of the filamentous soil bacterium Ktedonobacter racemifer type strain (SOSP1-21T). Standards in Genomic Sciences 2011;5(1):97–111.

[86] Kao C-M, Liao H-Y, Chien C-C, Tseng Y-K, Tang P, Lin C-E, Chen S-C. The change of microbial community from chlorinated solvent-contaminated groundwater after biostimulation using the metagenome analysis. J Hazard Mater 2016;302:144–50.

[87] Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, Salazar GA, Pesseat S, Boland MA, Hunter FMI, Hoopen P ten, Alako B, Amid C, Wilkinson DJ, Curtis TP, Cochrane G, Finn RD. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Res 2018;46(D1):D726-D735.

[88] Shi W, Qi H, Sun Q, Fan G, Liu S, Wang J, Zhu B, Liu H, Zhao F, Wang X, Hu X, Li W, Liu J, Tian Y, Wu L, Ma J. gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. Nucleic Acids Res 2019;47(D1):D637-D648.

[89] Alneberg J, Bjarnason BS, Bruijn Id, Schirmer M, Quick J, Ijaz UZ, Loman NJ, Andersson AF, Quince C. CONCOCT: Clustering cONtigs on COverage and ComposiTion; 2013.

[90] Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology 2018;3(7):836–43.

[91] Vollmers J. GitHub; Available from: https://github.com/jvollme (29 September 2019).

[92] Qin Q-L, Xie B-B, Zhang X-Y, Chen X-L, Zhou B-C, Zhou J, Oren A, Zhang Y-Z. A proposed genus boundary for the prokaryotes based on genomic insights. J Bacteriol 2014;196(12):2210–5.

[93] Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006;22(21):2688–90.

[94] Flemming U. Case-based design in the SEED system. Automation in Construction 1994;3(2-3):123–33.

[95] Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. GO:TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 2004;20(18):3710–5.

Anhang

Der digitale Anhang ist unter ***https://github.com/DominikHe93/Thesis.git*** zu finden. Hier sind alle verwendeten Abbildungen, Skripte, Stammbäume und Tabellen aufgelistet. Die Ordnerstruktur ist analog zu den hier aufgelisteten Kapiteln angelegt.