Analysis and interpretation of (meta-)genomic data from host-associated microorganisms

Analyse und Interpretation von (meta-)genomischen Daten aus Wirt-assoziierten Mikroorganismen

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Graduate School of Life Sciences,
Julius–Maximilians–Universität Würzburg
Klasse: Integrative Biology

vorgelegt von

Hannes Horn, M. Sc.

aus

Schwäbisch Gmünd

Würzburg, 2017

Eingereicht am: ______________________

Mitglieder des Promotionskomitees:

Vorsitzender:	Prof. Dr. Thomas Müller
1. Betreuer:	Prof. Dr. Ute Hentschel Humeida
2. Betreuer:	Prof. Dr. Markus Riederer
3. Betreuer:	Dr. Alexander Keller
4. Betreuer:	Dr. Mitja N. P. Remus–Emsermann

Tag des Promotionskolloquiums: ______________________

Promotionsurkunde ausgehändigt am: ______________________

To my family

"*It is the time you have wasted for your rose*
that makes your rose so important."
Antoine de Saint–Exupéry

Acknowledgements

Completing this work would have been impossible without the support of many people. I would like to express my gratitude to all of you.

First of all I wish to thank Prof. Dr. Ute Hentschel Humeida and Prof. Dr. Markus Riederer for the opportunity to work on this project and to open the door to many different and exciting topics.

I am very thankful to my thesis committee, consisting of Prof. Dr. Ute Hentschel Humeida, Prof. Dr. Markus Riederer, Dr. Alexander Keller and Dr. Mitja–Remus Emsermann. I appreciate their mentoring and they provided me advice, support, guidance, patience, ethusiasm, and constructive criticism.

Sincere thanks to my peers from the Department of Botany II, in particular Kristina, Usama, Beate, Martin, Lucas, Cheng, Lucia and Tine, Anni, Eli, Anton, Katja, Jana, Moni, Wilma, Andrea, Jutta, Ulrich, Markus and Michael. Thanks for helpful discussions, support in the lab, manuscript writing, administrative purposes, and amicable atmosphere! Special thanks to Usama who introduced me to many projects and guided me through the writing process of my PhD thesis.

Many thanks to my new colleages at the GEOMAR, Jutta, Tanja, Bettina, Alvaro, Ignacio and Prof. Dr. J. Imhoff, who gave us a warm welcome to Kiel and helped us to arrive in good spirits in our new environment.

I appreciate the help of all my collaboration partners, notably the people from Alexander Keller's Group, Wiebke Sickel and Markus Ankenbrand (and Alexander himself, of course!). Thanks to Dr. Stefanie Gläser and Prof. Dr. P. Kämpfer from the University of Gießen, Dr. Michael Richter from the MPI Bremen as well as Dr. Rodrigo Costa and Elham Karimi from the University of Algarve, and to the people of the Department of Electron Microscopy, Daniela Bunsen and Prof. Dr. G. Krohne. Special thanks to Frank Förster for his all–time support and organizing funny Kässpätzles–Events.

Coming to an end, I am being grateful to my friends for all your support and distractions, but also your understanding when there was limited time from my side.

Lastly, I want to express my gratitude to my family for your encouragement, your belief in me, never–ending support, and love.

Contents

Summary

Host–microbe interactions are the key to understand why and how microbes inhabit specific environments. With the scientific fields of microbial genomics and metagenomics, evolving on an unprecedented scale, one is able to gain insights in these interactions on a molecular and ecological level. The goal of this PhD thesis was to make (meta–)genomic data accessible, integrate it in a comparative manner and to gain comprehensive taxonomic and functional insights into bacterial strains and communities derived from two different environments: the phyllosphere of *Arabidopsis thaliana* and the mesohyl interior of marine sponges.

This thesis focused first on the *de novo* assembly of bacterial genomes. A 5–step protocol was developed, each step including a quality control. The examination of different assembly software in a comparative way identified SPAdes as most suitable. The protocol enables the user to chose the best tailored assembly. Contamination issues were solved by an initial filtering of the data and methods normally used for the binning of metagenomic datasets. This step is missed in many published assembly pipelines. The described protocol offers assemblies of high quality ready for downstream analysis.

Subsequently, assemblies generated with the developed protocol were annotated and explored in terms of their function. In a first study, the genome of a phyllosphere bacterium, *Williamsia* sp. ARP1, was analyzed, offering many adaptions to the leaf habitat: it can deal with temperature shifts, react to oxygen species, produces mycosporins as protection against UV–light, and is able to uptake photosynthates. Further, its taxonomic position within the *Actinomycetales* was infered from 16S rRNA and comparative genomics showing the close relation between the genera *Williamsia* and *Gordonia*.

In a second study, six sponge–derived actinomycete genomes were investigated for secondary metabolism. By use of state–of–the–art software, these strains exhibited numerous gene clusters, mostly linked to polykethide synthases, non–ribosomal peptide synthesis, terpenes, fatty acids and saccharides. Subsequent predictions on these clusters offered a great variety of possible produced compounds with antibiotic, antifungal or anti–cancer activity. These analysis highlight the potential for the synthesis of natural products and the use of genomic data as screening toolkit.

In a last study, three sponge–derived and one seawater metagenomes were functionally compared. Different signatures regarding the microbial composition and GC–distribution were observed between the two environments. With a focus on bacerial defense systems, the data indicates a pronounced repertoire of sponge associated bacteria for bacterial defense systems, in particular, Clustered Regularly Interspaced Short Palindromic Repeats, restriction modification system, DNA phosphorothioation and phage growth limitation. In addition, characterizing genes for secondary metabolite cluster differed between sponge and seawater microbiomes. Moreover, a variety of Type I polyketide synthases were only found within the sponge microbiomes. With that, metagenomics are shown to be a useful tool for the screening of secondary metabolite genes. Furthermore, enriched defense systems are highlighted as feature of sponge-associated microbes and marks them as a selective trait.

Zusammenfassung

Mikroben–Wirt Interaktionen sind der Schlüssel, um zu verstehen "Wie?" und "Warum?" Mikroben in bestimmten Umgebungen vorkommen. Mithilfe von Genomik und Metagenomik lassen sich Einblicke auf dem molekularen sowie ökolgischen Level gewinnen. Ziel dieser Arbeit war es, diese Daten zugänglich zu machen und zu vergleichen, um Erkenntnisse auf taxonomischer und funktionaler Ebene in bakterielle Isolate und bakterielle Konsortien zu erhalten. Dabei wurden Daten aus zwei verschiedenen Umgebungen erhoben: der Phyllosphäre von *Arabidopsis thaliana* und aus der Mesohyl–Matrix mariner Schwämme.

Das Ziel war zunächst, bakterieller Genome *de novo* zu assemblieren. Dazu wurde ein Protokoll, bestehend aus 5 Schritten, entwickelt. Durch Verwendung verschiedener Software zum Assemblieren konnte SPAdes als am besten geeignet für die gegebenen Daten herausgearbeitet werden. Durch anfängliches Filtern der Daten konnte erste Kontamination entfernt werden. Durch das Anwenden weiterer Methoden, welche ursprünglich für metagenomische Datensätze entwickelt wurden, konnten weitere Kontaminationen erkannt und von den "echten" Daten getrennt werden. Ein Schritt, welcher in den meisten publizierten Assembly–Pipelines fehlt. Das Protokoll ermöglicht das Erstellen hochqualitativer Assemblies, welche zur weiteren Analyse nicht weiter aufbereitet werden müssen.

Nachfolgend wurden die generierten Assemblies annotiert. Das Genom von *Williamsia* sp. ARP1 wurde untersucht und durch dessen Interpretation konnten viele Anpassungen an die Existenz in der Phyllosphäre gezeigt werden: Anpassung an Termperaturveränderungen, Produktion von Mycosporinen als Schutz vor UV–Strahlung und die Möglichkeit, von der Pflanze durch Photosynthese hergestellte Substanzen aufzunehmen. Seine taxonomische Position wurde aufgrund von 16S rRNA sowie vergleichende Genomik bestimmt. Dadurch konnte eine nahe Verwandtschaft zwischen den Gattungen *Williamsia* und *Gordonia* gezeigt werden.

In einer weiteren Studie wurden sechs Actinomyceten–Genome, isoliert aus Schwämmen, hinsichtlich ihres Sekundärmetabolismus untersucht. Mihilfe moderner Software konnten in zahlreiche Gen–Cluster identifiziert werden. Zumeist zeigten diese eine Zugehörigkeit zu Polyketidsynthasen, Nichtribosomalen Peptidsynthasen, Terpenen, Fettsäuren oder Sacchariden. Durch eine tiefere Analyse konnten die Cluster mit chemischen Verbindungen assoziiert werden, welche antibakterielle oder fungizide Eigenschaften besitzen.

In der letzten Untersuchung wurden Metagenome von drei Schwämmen sowie Meerwasser auf funktioneller Ebene verglichen. Beobachtet wurden Unterschiede in deren mikrobiellen

Konsortien und GC–Gehalt. Schwamm–assoziierte Bakterien zeigten ein ausgeprägtes Inventar an Verteidigungsmechanismen gegenüber deren Vertretern aus dem Meerwasser. Dies beinhaltete vor allem: Clustered Regularly Interspaced Short Palindromic Repeats, das Restriktions-Modifikationssystem, DNA Phosphorothioation, oder Gene, welche das Wachstum von Phagen hemmen können. Gene für Sekundärmetabolite waren zwischen Schwamm– und Meerwasser–Metagenomen unterschiedlich stark ausgeprägt. So konnten Typ I Polyketidsynthasen ausschließlich in den Schwamm–Metagenomen gefunden werden. Dies zeigt, dass metagenomische Daten ebenso wie genomische Daten zur Untersuchung des Sekundärmetabolismus genutzt werden können. Des Weiteren zeigt die Anhäufung an Verteidigungsmechanismen eine Anpassung von Schwamm–assoziierten Mikroben an ihre Umgebung und ist ein Hinweis auf deren mögliche selektive Eigenschaft.

Part I.

General introduction

The road to –omics

Historical background

> *“The order of nucleic acids in polynucleotide chains ultimately contains the information for the hereditary and biochemical properties of terrestrial life. Therefore the ability to measure or infer such sequences is imperative to biological research.”* (Heather and Chain 2016)

One could consider the year 1869 as the birth of sequencing, when Friedrich Miescher discovered and isolated “nuclein”, a phosphate–rich substance residing within the cell nuclei, better known today as desoxyribonucleic acid (DNA). For many years, proteins were believed to be the inheritable genetic molecules, as they varied more in a physical and chemical way compared to DNA. Surprisingly, in 1944, Oswald T. Avery, Colin MacLeod and Maclyn McCarty demonstrated, that DNA, serves as genetic carrier:

> *“If the results of the present study of the transforming principle are confirmed, then nucleic acids must be regarded as possessing biological specificity ...”* (Avery et al. 1944)

Based on the results obtained by Avery and his colleagues, Alfred Hershey and Martha Chase conducted experiments on bacteriophages infecting bacterial cells in 1952. Upon this process, they recognized only DNA entering the cell, whereas proteins did not, confirming DNA to carry genetic information and thus as responsible for inheratibility:

> *“This protein probably has no function in the growth of intracellular phage. The DNA has some function ...”* (Hershey and Chase 1952)

Only one year later, James Watson and Francis Crick were able to solve the three-dimensional structure of DNA supported through the crystallographic data of Rosalind Franklin and Maurice Wilkins (Watson and Crick 1953; Zallen 2003). They proposed a double–helical structure consisting of two chains running in opposite directions (i.e. complementary) with the sugar–phosphate backbones on the outside and the four (nucleo–) bases — adenine (A), cytosine (C), guanine (G) and thymine (T) — on the inside. Pairwise bonds between the nucleotides (i.e. nucleobase + sugar + phosphate) are formed to stabilize the helical structure: A with T, C with G (Watson and Crick 1953). Within the next decade, reading or even sequencing DNA was impossible, due to its length and double-stranded appearance (Heather and Chain 2016). However, sequencing of the first transfer RNA (tRNA) from *Saccharomyces cerevisiae* in 1965 (Holley et al. 1965), the coat protein of the ribonucleic acid (RNA) bacteriophage MS2 and subsequent its complete RNA sequence (Fiers et al. 1976; Jou et al. 1972) were published. Even the measurement of nucleotide composition became possible in 1961 (Holley et al. 1961) — but not the order of nucleotides.

Development and impact of sequencing technologies

Only in 1968, Ray Wu and colleagues employed DNA polymerase to sequence the "sticky ends" of phage λ (Wu and Kaiser 1968). With this, they were the first to publish a nucleotide sequence, consisting of 12 basepairs (Wu and Taylor 1971).

First–generation sequencing commenced in 1975, with the *chemical cleavage* technique of Allan Maxam and Walter Gilbert (Heather and Chain 2016). Independently, Frederick Sanger and Alan Coulson invented their *plus and minus* system and revolutionized the way of DNA sequencing presenting the famous *dideoxy* or *chain–termination* method (Sanger and Coulson 1975; Sanger et al. 1977a). Soon after, the 5.3 kbp genome of the bacteriophage ϕX174 was released (Sanger et al. 1977b). Technical improvements towards automation, capillary systems and fluorometric based detection of nucleotides (Heather and Chain 2016), the use of shotgun sequencing — i.e. DNA is randomly fragmented, cloned and these fragments/*reads* are sequenced — and *suitable computer programs* (Staden 1979) to stitch obtained reads together (= assembly) led to the first fully sequenced bacterial genome in 1995, *Haemophils influenza*, with a size of 1.8 Mbp (Fleischmann et al. 1995) and changed the science of bacteria dramatically (Land et al. 2015). A milestone in sequencing was set in 2001, when the sequence of the human genome with a length of 3 Gbp was presented (Lander et al. 2001; Venter et al. 2001) a few years ahead of schedule.

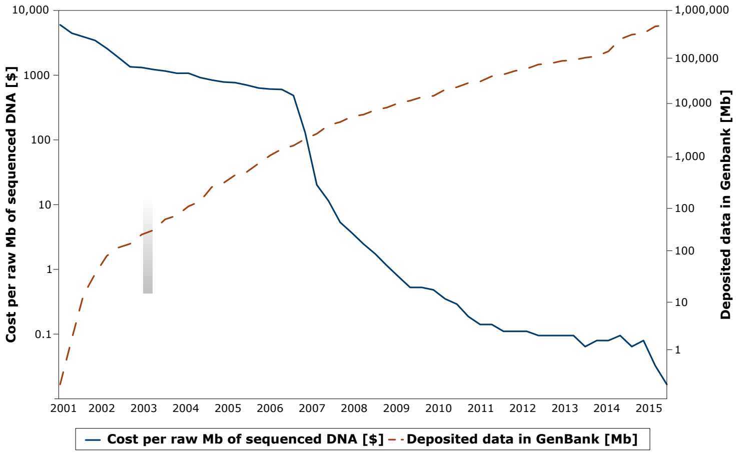


Figure 1. – Development of DNA sequencing costs according to the National Human Genome Research Institute (NHGRI) (Wetterstrand 2016) for one megabasepair and the amount of deposited sequencing data in GenBank (GenBank 2016)

Concurrent with further advances in large–scale *dideoxy* sequencing methods, a wave of next–generation sequencing (NGS) platforms were released (Heather and Chain 2016). In 2005, the first commercially sequencer, the GS20 by 454 Life Sciences became available. It heralded the era of NGS (Shendure and Ji 2008) as the mass of sequences produced in parallel within one run was considered a paradigm shift (Margulies et al. 2005). Since its release, a massive drop in sequencing costs concurrent with increasing throughput took place (Figure 1). In particular, Illumina has lowered sequencing costs drastically, allowing the human genome to be sequenced with ≤ 1 000 $ (Illumina 2016), which had costs of 1 000 000 $ back in 2001 and have brought the company to near monopoly (Greenleaf and Sidow 2014). Among the two described NGS technologies, others appeared (and disappeared) with variable impact, namely Sequencing by Oligonucleotide Ligation and Detection (SOLiD) (Applied Biosystems) and Ion Torrent (Life Technologies) in 2006 and 2010, respectively. More recently, the third generation of sequencing technologies emerged. That is, single molecule sequencing (SMS) in real–time without the need for amplification (Heather and Chain 2016; Liu et al. 2012a). Considered to be the most wideley used third–generation sequencing platform to date is PacBio (Pacific Biosciences) (Dijk et al. 2014) over nanopore sequencing (Oxford Nanopore Technologies).

Sequencing technologies have changed massively within the last three decades. The overall throughput has increased from 0.1 Mbp using Sanger–seqencing to over 1 000 Gbp with Illumina platforms. Further, there were trends towards smaller read–lenghts using NGS platforms compared to Sanger–sequencing, but this has changed again to lengths exceeding 1 000 bp using the Third–Generation sequencing technologies (Figure 2).

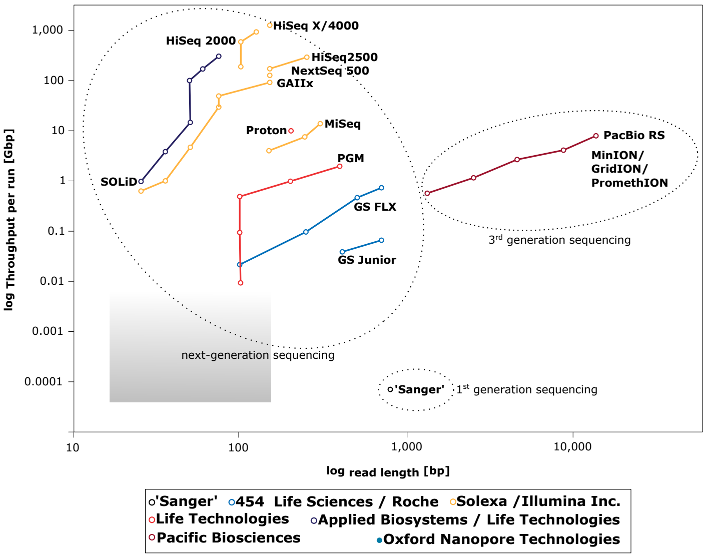


Figure 2. – Development in NGS with regards to throughput and read length. Adapted from Nederbragt (2015)

As a result of the technological improvements, it was possible to reduce the price for sequencing projects dramatically. Especially Illumina led to a massive drop to $\leq 0.1\,\$$ per megabase (Glenn 2011) (Sanger $\leq 400\,\$$ per megabase (Thomas et al. 2012)), and the availability of benchtop sequencers (e.g. Illumina MiSeq, Roche 454 Junior, Ion Torrent PGM), has made the sequencing of bacterial genomes (and others) affordable to many laboratories, but also dropped the impact of a bacterial genome projects from Journals as *Science* (Impact factor: 31.477) in 1995 to journals as *Standards in Genomic Sciences* (Impact factor: 3.167) nowadays. Whatsoever, the era of NGS and the vast data produced allows new questions to be asked. It is now possible to sequence multiple species at a time or even the complete DNA inventory of microbial communities. These endeavours and developments paved the way for a new scientific field: –omics.

Omics

In molecular biology, the term –omics refers to the suffix –ome to form nouns, which adresses to the objects of study (e.g. gen*ome* — gen*omics*). According to the Oxford English Dictionary, –omics describing "[a]ll constituents considered collectively" (Oxford English Dictionary 2016), meaning in science the use of "[l]arge-scale data/information to understand life summed up in omes" (Yadav 2007) .

Different –omics data types help to understand the inherent relationship between biological regimes, which can be described by the central dogma of molecular biology (Buescher and Driggers 2016; Crick 1970; Crick 1958): From DNA, i.e. the blueprint of an organism which is carried by RNA, describing what actually happens within an organism, to proteins, the functional products (Buescher and Driggers 2016) (Figure 3).

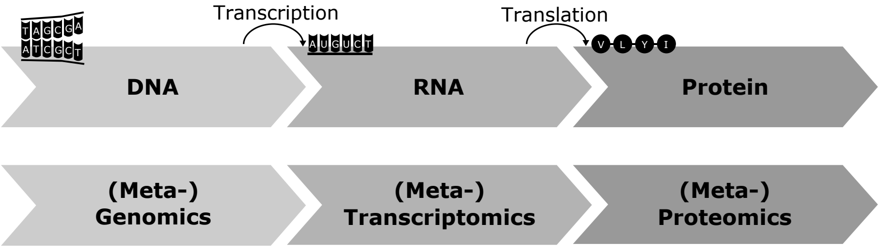


Figure 3. – The central dogma of molecular biology according to F. Crick (Crick 1970; Crick 1958) and the connection to –omics

Below, a brief overview of microbial genomics and metagenomics is given, as these are the main research fields used within this thesis. Further, approaches and challenges using that kind of data are described.

Genomics – the word was proposed by the geneticist Thomas H. Roderick in early 1986 as the name of the yet–to–be-published journal *Genomics* (Kuska 1998). Genomics refers to a field in genetics and concerns the analyis of an organisms genome within its cells (Lockhart and Winzeler 2000). In particular, it describes the determination of DNA sequences, and subsequent functional analysis. Two decades earlier, this was an extensive process requiring a lot of time and money, but became — at least for bacterial genomes — a standard procedure (Land et al. 2015) by use of NGS technologies. The ease of producing genomics data today is also reflected within the public databases such as The Genomes OnLine Database (GOLD) (Reddy et al. 2015) or GenBank. The number of registered genome projects increased dramatically since 2001, in particular for bacterial genomes, but also increased for eukaryotic ones within the last 5 years (Figure 4). Also GenBank has grown massively since its invention 1982 (Land et al. 2015) and now harbours amounts of 4 000 eukaryotic, 6 000 viral, and more than 80 000 available bacterial genomes comprising over 50 bacterial phyla (NCBI 2014). This is also due to the drop in sequencing costs (Figure 1) and initiatives like the 100K Pathogen Genome Project or the Genomic Encyclopedia of Bacteria and Archaea (GEBA) (Kyrpides et al. 2014) among others. However, the last decades have not only been about generating sequencing data, but also illuminated functionalities of genomic data and their coherence to the environment. Indeed, there have been many projects on eukaryotes like the human genome in 2001 (Lander et al. 2001; Venter et al. 2001), the nematode *Caenorhabditis elegans* 1998 (Sequencing Consortium 1998) or the first flowering plant *Arabidopsis thaliana* in 2000 (Initiative 2000). The next paragraph will focus on insights and knowledge gained through bacterial genome sequencing.

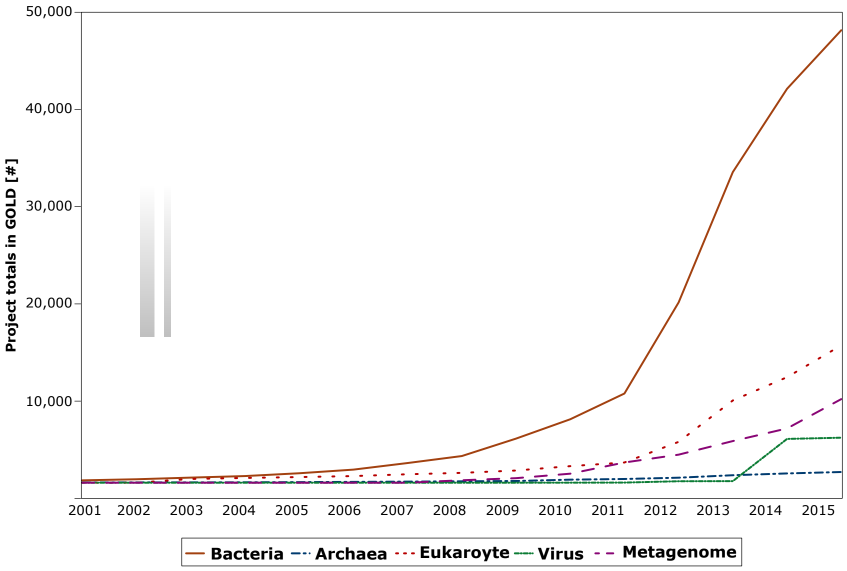


Figure 4. – Development of registered DNA sequencing projects in the Genomes Online Database (GOLD) since 2001 (Genomes Online Database 2016)

The sequencing of the first bacterial genomes in 1995 — *H. influenza* (Fleischmann et al. 1995) and *Mycoplasma genitalium* (Fraser et al. 1995) — were non–pathogenic strains. But, it started a race (Loman and Pallen 2015) to sequence genomes from pathogens as the white plague *Mycobacterium tuberculosis* (Cole et al. 1998), model organisms as *Bacil-*

lus subtilis and *Escherichia coli* K–12 (Blattner 1997; Kunst et al. 1997) and extremophiles as *Deinococcus radiodurans* (White 1999) or *Tropheryma whipplei.* Though many of these strains were hard to study *in vitro*, it allowed the design of a growth medium for e.g. *T. whipplei* through metabolic reconstruction and its subsequent cultivation (Renesto et al. 2003). The availability of genomes from the same genus or species opened the door for new scientific fields: comparative and evolutionary genomics. It involves the comparison of genomic features — i.e. DNA sequences and, genes and their order, regulatory elements or even Single nucleotide polymorphisms (SNPs) — of different organisms (Xia 2013). This information can be used to define genomic inventories identical/different across genomes, thus defining homologs/paralogs or origins of genes and species (Nature 2014). With that background, it was possible to define pan– (an entire gene set) and core–genomes (genes present in all strains of a taxononomic group) and classify strains using their whole genomic repertoire (Land et al. 2015) or nucleotide composition infered through average nucleotide identity (ANI) (Richter and Rosselló-Móra 2009) instead of single marker genes (e.g. 16S rRNA for bacteria). Using these toolkits, high strain diversity and horizontal gene transfer (HGT) within *E. coli* was detected (Welch et al. 2002). This knowledge of gene transfer and recombination changed the way of understanding genome evolution for microbes and tree–like structures were found to be inappropriate (Spratt and Maiden 1999) to show their phylogeny as bacterial cells were designated as individuals. Focussing on human pathogens and symbionts, a process called genome reduction was observed. Due to sexual isolation and adaptions to restricted niches as found in *Rickettsia* (Kurland et al. 1998) or *Mycobacterium* (Cole et al. 2001) strains, non–functional genes were created and lost as they were no longer needed (Darby et al. 2007) when inhabiting a host organism, e.g. genes responsible for flagellar motility (Maurelli 2007). In addition, within–host genome evolution was detected when analyzing host–microbe interactions, e.g. insertions and deletions within noncoding areas of different *Burkholderia* strains (Romero et al. 2006). This also opened the "eco–evo perspective" (ecology–evolution). In this context, microbes were recognized in combination with their lifestyle. They were shown to have the ability to shift from pathogenic to commensal states and vice versa (Loman and Pallen 2015) as non–pathogenic strains were found to encode virulence factors, e.g. antimicrobial peptide violacein in the free–living *Chromobacterium violaceum* ATCC 12472 (Holden et al. 2004). Beside, the usage of comparative genomics led to one of the very recent scientific breakthroughs: the detection of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) as a prokaryotic immune system detected in many bacterial genomes (Makarova et al. 2006) and its usage for genome engineering (Pennisi 2013).

Beside functional and evolutionary genomics, the field can be divided into other research areas. Functional genomics involves not only genomic data, but also transcriptomics (that is, the sequencing complete RNA), to describe gene functions, expression and is focussed on dynamic processes such as transcription, translation, regulation of genes, and their interactions. Following, forward–engineering of metabolic pathways led to the rerouting of biosynthetic pathways and production of biodiesel (Steen et al. 2010), gasoline (Choi and Lee 2013), or even the antimalarial drug artemisinin in 2013 (Paddon et al. 2013). A new and emerging field is single cell genomics, using isolated DNA from individual cells for genomic sequencing (e.g. Kamke et al. 2014; Siegl et al. 2011). It is a promising tool as it can be used on (bacterial) cells, which can not be cultured. However, it is technically challenging as cells need to kept viable during enzymatic/mechanical isolation. Due to whole genome amplifications, artifacts as genome loss, mutations and chimeras may be introduced (Gawad et al. 2016). But, this approach filled the phylogenetic tree with reference genomes previously unculutivated and only known by barcodes or even unknown phyla as TM7 from human mouth or soil (Marcy et al. 2007; Podar et al. 2007), or phyla as Poribacteria (Siegl et al. 2011), or Tectomicrobia from sponges with high potential for the synthesis of secondary metabolites (Wilson et al. 2014) as source for novel drugs. To sum up, genomics can provide insights into genetic mark–ups, functional and metabolic capabilities, evolution of organisms on a molecular level, resolve phylogenetic patterns, and can be used for bioprospecting approaches.

The term *metagenomics* was first coined by Jo Handelsman in 1998 and defined as "[cloning] and functional analysis of collective microbial genomes" (Handelsman et al. 1998). It is also refered to as environmental genomics, community genomics or population genomics (Handelsman 2004). Until today its defintion has modified as the direct sequencing — i.e. circumvents unculturability — of microbial genomes within an environmental sample.

Classical microbial genome sequencing, except single cell genomics, has the drawback to rely on culturing the organism of interest. As observed by microscopy, environments contain millions of microbial cells of different species, but only few of them grow on petri plates (Amann 1911). The number of culturable microorganisms was estimated $\leq 1\,\%$ (Hugenholtz 2002). This phenomenon was called *the great plate count anomaly* (Staley and Konopka 1985). First attempts by Norman R. Pace focused on the amplification without "[i]solation of the 16S rRNA or cloning of its gene" (Lane et al. 1985). This lead to the idea of cloning DNA directly from the environment (Pace et al. 1985). Thanks to the efforts of Pace and colleagues, extraction of phylogenetic marker genes from environmental samples became possible (Giovannoni et al. 1990) and opened a world far more complex than was seen based on morphological features. First limited to non–protein coding genes, subsequent studies by DeLong and Healy also reported the direct isolation of functional genes (Healy et al. 1995; Stein et al. 1996). These concepts are in use today to explore microbial communites through massive parallel sequencing of 16S ribosomal RNA (rRNA) marker genes (= amplicon sequencing) and might be considered as targeted metagenomics (Knief 2014). It can also be applied to explore the diversity of functional genes such as polyketide synthases (PKSs) using distinct primers (Della Sala et al. 2014) and subsequent identification of the underlying product.

In 2002, whole metagenome shotgun sequencing was invented. In short, DNA is extracted from all cells in a community sample. Instead of targeting genomic features for amplification, DNA is sheared into small fragments that are sequenced independently. These DNA sequences occur from different genomic locations for the various genomes present in the sample, which includes also non–microbial DNA. Some fragments relate to taxonomically informative loci (e.g. 16S rRNA), and others to coding sequences (Sharpton 2014). Hence, with shotgun metagenomics, one is able to not only answer the question “Who is there?”, but also “What are they doing?” (Handelsman 2004) as metagenomics give access to the functional repertoire of microbial communities (Thomas et al. 2012). Functional comparative metagenomics enabling the reconstruction of metabolic pathways (Escobar-Zepeda et al. 2015), or may be as simple as comparisons of GC–content (Foerstner et al. 2005) or genome sizes. Moreover, functions of microbial communities can be linked to geographic locations or explain interactions and adaptions between hosts and microbes. In addition, it is possible to reconstruct single genomes from metagenomic samples (Albertsen et al. 2013), which might reveal conditions of yet unculturable strains (Burgsdorf et al. 2015). Metagenomics is able to connect function to phylogenetic, chemical or other biological traits to characterize the environment (Thomas et al. 2012). This can be considered an advantage over genomics, to the expense, generated data is more complex and at higher costs.

Since the first well–analyzed shotgun metagenomic studies as the Acid Mine Drainage microbial biofilm (Tyson et al. 2004), the Sargasso Sea surface water (Venter et al. 2004), or the Whale Fall (whale bone) and Minnesota farm soil metagenomes (Tringe et al. 2005), the number of registered metagenomic projects in GOLD has increased to more than 10 000 by the end of 2015 (Figure 4). Moreover, consortia for the sequencing of metagenomes have been launched as the TerraGenome project (Vogel et al. 2009) for soil or the Human Microbiome Project (HMP) (Human Microbiome Project Consortium 2012). Endeavours in generating huge amounts of metagenomic data shows its importance to biological sciences, but also leads to challenges due to their dimensions, sequence diversity and their fragmentation.

With regards to other –omics, both genomics and metagenomics are limited to the description of functional potential within one or more organisms. To adress questions as “What genes are expressed?”, “What are the microbes doing at the moment?” or “Which functions are up– or down–regulated during a specific event?”, techniques as (meta–)transcriptomics (that is the total set of mRNA sequenced) or (meta–)proteomics (that is, detecting all proteins in given cells) are necessary. Despite its limitations, metagenomics is considered one of the “[m]ost remarkable events in the field of microbial ecology” (Thomas et al. 2012), as it offers functional insights into microbial communities remaining unknown by culture–depdendent methods based on the blueprint of life — DNA.

A (meta–)genomic workflow

Current genome and metagenome projects sequenced on NGS platforms produce high throughput with usually short reads. Data generation is estimated to be doubled every 7 months (Stephens et al. 2015). According to Moore’s Law, this can no longer be compensated by computer power, which doubles every 18 months. Consequently, the amount of data poses bioinformatic challenges when it comes to sequence quality assessments, alignment, assembly, storage and release (Shendure and Ji 2008) or integration and interpretation. Below described are the steps towards genomic and metagenomic data. A workflow from sampling to data storage is illustrated in Figure 5.

The adequate sampling and its procession is a first and most crucial step to assure high quality. Especially in the field of metagenomics the sample has to be representative for all cells (Thomas et al. 2012). For DNA extraction, suitable treatments, e.g. lysis, fractionation or disruption may be applied to the sample (Figure 5, DNA isolation and fragmentation). The obtained DNA can then be sequenced on an appropiate platform (see Figure 5 DNA sequencing). Next steps conduct an extensive quality control (Figure 5 Quality control). Whereas the quality (PHRED score, length, ambiguous basepairs, throughput) of sequenced reads can be measured with ease, i.e. using FastQC (Andrews 2016), the quality of a sequencing project becomes available not until initial processing. For example, within 202 metagenome studies, 145 were found with human contamination of up to 64 % (Schmieder and Edwards 2011), but also genomic studies are not free of contamination (Merchant et al. 2014). This is a serious concern often overlooked and may lead to erroneous downstream analysis (Schmieder and Edwards 2011). It also affects the assembly of genomes, i.e. put

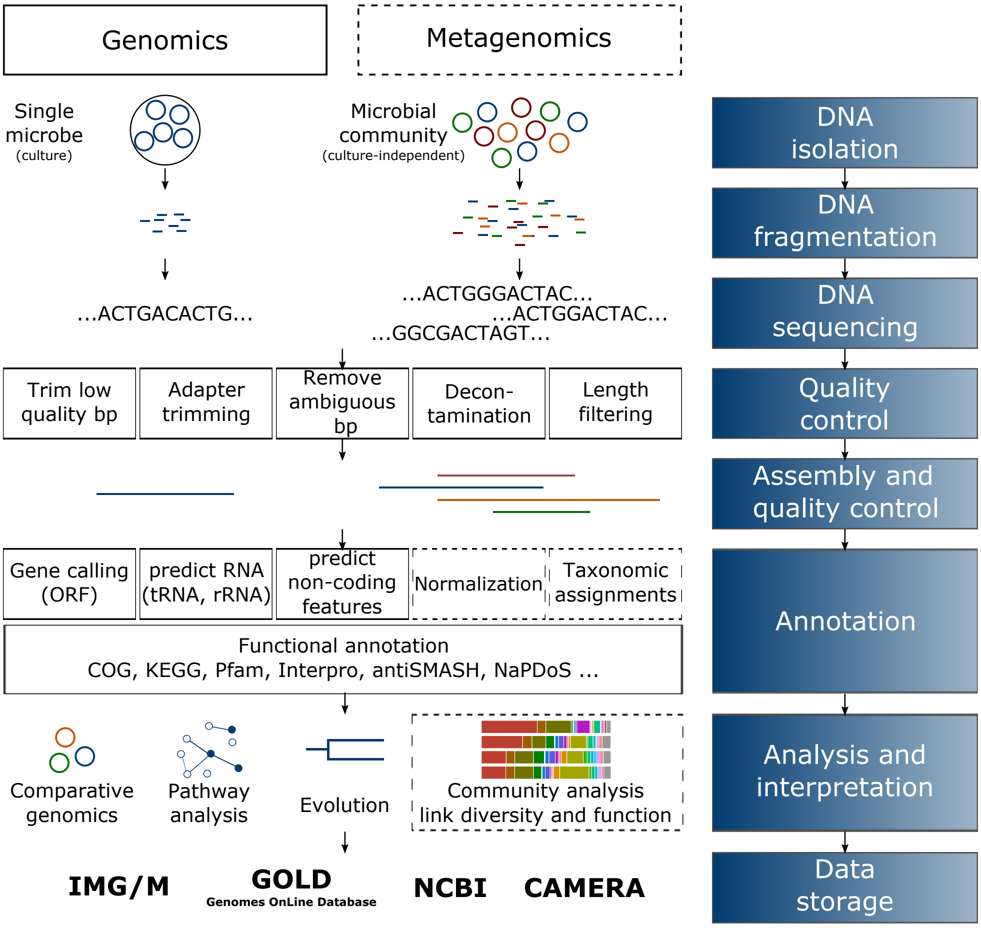


Figure 5. – Simplified flowchart of genomic (left) and metagenomic (right) projects from sampling to storage. Dashed lines indicate steps only conducted for metagenomics.

together reads of DNA into longer fragments, so called contigs (Figure 5 Assembly and qualtiy control). This is considered one of the most complex computational tasks in biology (Baker 2012), and is even more complicated for metagenomes (Sharpton 2014) due to chimeric sequences, uneven coverage, and the large amount of different species. For analysis of long genetic elements or genes in genomic proximity as secondary metabolite gene clusters or CRISPR, it is inevitable to work on assembled sequences rather than separated reads (Thomas et al. 2012). The assembly process is followed by the prediction of open reading frames (ORFs) and their functional annotation with available databases such as Clusters of Orthologous Groups (COG) (Galperin et al. 2015), TIGRFAM (Haft et al. 2013), or Protein Families (PFAM) (Finn et al. 2016). Several web servers are specialized ressources and merge the information of several databases for functional annotation and comparison of genomic and/or metagenomic datasets. Among these prominent ones are RAST and MG–RAST servers, IMG/M or EDGAR (Aziz et al. 2008; Blom et al. 2016; Markowitz et al. 2012; Meyer et al. 2008). If one is interested in more specialized analysis, e.g. secondary metabolism, web servers as antiSMASH (Weber et al. 2015) may be conducted. Also, the use of locally installed and manually set up databases for Basic Local Alignment and Search Tool (BLAST) or hidden–markov model (HMM) searches may be necessary for a more specific annotation, but requires great computational ressources (Figure 5 Annotation). Noteworthy, only around 50 % of all sequences may be successfully annotated (Thomas et al. 2012) as this process is limited to known genes in databases. Interpretation of the created data hardly depends on the project behind (Figure 5 Analysis and interpretation). For single genomes, one may conduct analysis on evolutionary genomics by searching for orthologous or paralogous genes. Also, the creation of core– or pan–genomes is a way to learn about evolution when looking for absent or present features between close bacterial strains. Using metagenomic approaches, it is possible to get an overview of a whole bacterial community and its diversity which can than be associated to their functional composition (Tringe 2005). Platforms such as NCBI, IMG, RAST/MG–RAST or CAMERA provide an integrated environment for the analysis, storage, management and allow sharing and visualization of genomic and metagenomic data (Figure 5 Data storage) (Pavlopoulos et al. 2015).

As biology has become a big data science, the challenges are multilayered, starting with sampling and data generation generation, over quality control and assembly, analysis, final interpretation, and finding appropriate software in the plethora of available ones.

A microbial world

With an estimated number of 10^{29} prokayrotic cells around the world (Lougheed 2012) exceed other lifeforms by orders of magnitudes. They can be found in the harshest environments (e.g. Huber et al. 2000; Tyson et al. 2004) probably due to their metabolic diversity (Kostadinov 2011). First observations by Robert Koch revealed connections between microbes and diseases in various hosts as humans, plants or other animals (De Kruif 2002) and launched a generation of "microbe hunters" (Choffnes et al. 2013) sytematically searching for pathogenic microbes. Around the same time, the question arose, if microorganisms do not only harm, but also support their respective host (Choffnes et al. 2013), and given rise to the concept of symbiosis. This can be separated in three relationships between hosts and microorganisms, which form a continuum: parasitic (=interaction, in which one partner benefits at the expense of the other) , commensal (=beneficial interaction for one partner without affecting the other) and mutualistic (=beneficial interaction for both partner) (Little et al. 2008). However these relationships can not be separated clearly, the definition by Anton de Bary is used here: "the living together of unlike organisms" (Bary 1878). Microbes were found to be essential for all life on earth. They produce important compounds as oxygen for the earths atmosphere through photosynthesis (Pedrós-Alió 2006), are responsible for catalyzing carbon (Falkowski et al. 1998), nitrogen, oxygen or phosphorus into accessible forms and make nutrients available to their hosts (Handelsman et al. 2007) — to only name a few examples among thousands. With that our view on microbies has changed within the last decades.

We are aware of their real impact, but understanding microbes and the combined activities within microbial communities is still in its infancy. Modern biology relies not only on microscopy and the description of microbial phenotypes. These techniques and culturing of microbes will still be important (Handelsman et al. 2007), but are limited to the visibility of features. With the discovery of DNA (Watson and Crick 1953), the door was opened for molecular biology and the research of intrinsic, genotypic features. Tradititional methods, have in the last few decades, supplemented by the sequencing of DNA, helping to understand the functional and metabolic repertoire of microbes and determine interactions between microorganisms and the environment or their hosts.

Below, examples of microbes and their impact — in particular bacteria — are provided for two different environments: the plant surface and within marine sponges as these were the main research fields in this thesis. Each starting with an overview of environmental variables, the microbial community inhabiting the habitat and insights from modern sequencing approaches.

Microbial life in the phyllosphere

The phyllosphere is known as the aerial surfaces of plants (Kembel et al. 2014) and is dominated by the leaves (Vorholt 2012). It is estimated, that the phyllosphere spans an area up to 6.4*10^8 km^2. The leaf surface is far from sterile and is considered as a hostile and short–lived environment for microbes, due to temperature shifts, UV–radiation, fluctuating humidity and desiccation, limited nutrients (oligotrophic) and their uneven distribution among leaves (Remus-Emsermann et al. 2011; Remus-Emsermann and Leveau 2010), and their heterogenity at the micro– and macroscale (Lindow and Brandl 2003; Vorholt 2012). Also surrounding conditions as the atmosphere, wind and gas exchange have been shown on lettuce leaves to have an effect on epiphytic bacterial communities (Medina-Martinez et al. 2015) as they might induce stomata closure and reduce nutrient availability (Leveau and Lindow 2001; Miller et al. 2001). On the one hand, inhabitants encounter antimicrobial compounds either produced by other microbes or the plant itself (Vorholt 2012) (Figure 6 A and B). On the other hand, phyllosphere microbes, such as *Pseudomonas sp*, were also shown to protect the plant against pathogen invasion (e.g. fungi) by producing plant hormones (Ritpitakphong et al. 2016) or even support their host plant through drought tolerance (Lau and Lennon 2012) and disease resistance (Santhanam et al. 2015). Interactions between plants and their microbiome seem to be mutualistic, thus it is not surprising, that the planetary population of bacteria – by far the most abundant inhabitants — is as large as 10^{26} cells in the phyllosphere (Lindow and Brandl 2003) and is assumed to be as high as 10^6 – 10^7 cells/cm^2 (Lindow and Brandl 2003). Maybe explained, as phyllosphere–bacteria represent an ancient symbiosis (Partida-Martinez and Heil 2011) in which the bacteria extend the host phenotype (Wagner et al. 2016).

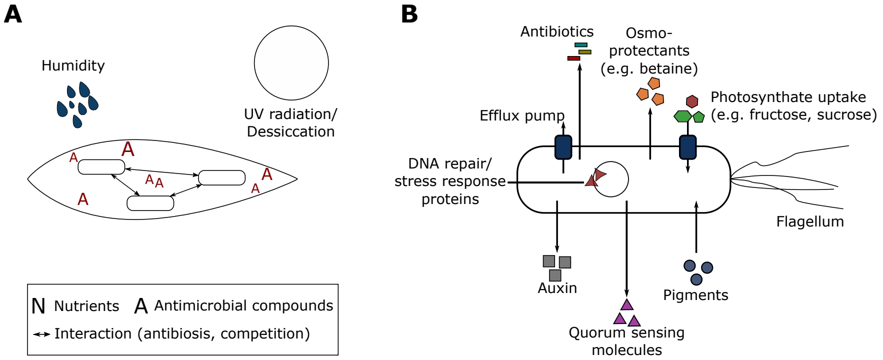


Figure 6. – (A) Microbe–microbe and plant–microbe interactions which may occur in the phyllosphere due to competition through antibiosis for nutrients and best environmental conditions. (B) Traits which allow and support microbes to establish and colonize the phyllosphere comprising the production of osmoprotectants, auxin, antibiotics or the uptake of photosynthates. Adapted from Vorholt (2012)

Insights into microbial phyllosphere communities have been obtained mostly through amplification and pyrosequencing of 16S rRNA marker genes (Vorholt 2012). Overall, the species richness was found to be lower compared to soil, the rhizosphere, costal seawater, or the human gut (Delmotte et al. 2009; Knief et al. 2012b). Identified dominant bacterial phyla on different plants are *Proteobacteria*, *Firmicutes*, and *Actinobacteria* (Williams and Marco 2014), but their proportions vary depending on several traits: Knief et al. (2010) have shown, that geographic location of a plant has more impact than the plant species, but only shown for *Methylobacterium* communites. Redford et al. (2010) concludes communities do not become more distinct with increasing geographic distance. Instead, they found "[i]nterspecies variability exceeds intraspecies variability" for 56 tree species. In another study, the wax composition of *A. thaliana* leaves was shown to have an effect on bacterial community composition (Reisberg et al. 2013). Further variations in microbial communities were linked to season: *Firmicutes* were dominant in planting of lettuce in june, in august and october it were *Proteobacteria* (Williams et al. 2013). Also the genotype of the plant seems to be an determining factor for communities (Hunter et al. 2010; Knief et al. 2010; Redford et al. 2010).

Genomic investigations revealed functional adaptions to the phyllospheric habitat as repair of UV–damaged DNA, uptake of photosynthates (sucrose, fructose) or production of osmoprotectants such as trehalose or betaine (**Horn** et al. 2016a; Remus-Emsermann et al. 2013) (Figure 6 B). Deeper insights in plant–microbe interactions on a functional scale and using cultivation–independet techniques are scarce (Berlec 2012). However, metaproteogenomic approaches to investigate bacteria on soya bean, *A. thaliana*, clover (Delmotte et al. 2009) and rice (Knief et al. 2012b) are outstanding examples. They were able to identify more than 4 600 proteins, a methanol–based methylotrophy within the genus *Methylobacterium* and specific functions as response to reactive oxygen species the invasion–associated locus B when compared to the rhizosphere. Atamna-Ismaeel et al. (2012) identified rhodopsins in the microbiome of the *Tamarix*, *A. thaliana* and rice phyllosphere and subsequent light sensing genes in *Tamarix* as adaption to the leaf habitat using a metagenomic approach (Finkel et al. 2016). In a recent resarch on 400 draft genomes from the phyllosphere and rhizosphere of *A. thaliana*, not only a huge phylogenetic overlap was detected, but also a large overlap of functionalities was found. An enriched function in phyllosphere bacteria was carbohydrate metabolism, possibly due to the fact, carbon is easier accessible in roots htan from the leaves (Bais et al. 2006). Vice versa, root microbial genomes are enriched in genes to process aromatic compounds. Overally, it was hypothesized, that functional and phylogenetic diversification is driven by taxonomic affiliation rather than ecological effects (Bai et al. 2015). Taxonomic identity as a major driver in microbial community structure have also been found before (Kembel et al. 2014; Redford et al. 2010). With the use of genome—wide association studys (GWASs), genetic loci varying in plants were identified leading to different bacterial communities. Based on 196 accessions of *A. thaliana*, gene loci responsible for cell wall integritiy and defense were suggested to alter the bacterial community. Further, genetic variation linked to morphogenesis and trichome branching was shown to shape the richness of the inhabiting community (Horton et al. 2014). Comprising, it is assumed that there is more than one single factor playing the key role in the process of microbial community assembly in the phyllosphere (Finkel et al. 2012; Hunter et al. 2010; Knief et al. 2010; Redford et al. 2010).

Microbial life within sponges

Sponges (phylum Poriferea) represent one of the oldest metazoans of the extant animal lineages (Li et al. 1998; Love et al. 2009) with a record dating back 600 million years (Erwin et al. 2011) reflecting their evolutionary success. Around 15 000 species are estimated to exist, belonging to three major classes (Hexactinellida, Calcarea, and Demospongiae). Sponges are widespread among aquatic habitats such as freshwater lakes, the deep sea, tropical and subtropical oceans as well as polar regions (Hooper and Van Soest 2002). They are sessile animals with a simple but optimized body–plan towards filter–feeding (Leys and Hill 2012) which enables them to pump $\leq 24\,000$ liters per day (Vogel 1977) to sequester nutrients. Due to their pumping performance, they contribute much to the function of benthic communities, e.g. through carbon or nitrogen cycling (Bell 2008).

Their immobile lifesytle resulted in a well adapted and complex innate immune system (Müller and Müller 2003) to defend against pathogens and invading parasites. Thus surprising, organisms of all domains of life, *Bacteria*, *Archaea*, *Eukarya* reside within sponges (Taylor et al. 2007; Wiens et al. 2007). Hence, recognition upon symbiotic or non–symbiotic microbes is evident (Hentschel et al. 2012). Even more interesting is the high abundance of microbes, which can make up to 35 % of the sponge biomass (Vacelet 1975) and may exceed 10^9 cells per sponge tissue (Webster and Hill 2001). Based on these numbers, sponges can be classified as either low microbial abundance (LMA) or high microbial abundance (HMA) sponges (Bayer et al. 2014; Gloeckner et al. 2014; Hentschel et al. 2003).

Initially, the diversity of sponge–associated microbes was observed based on microscopy (Vacelet 1975), but has changed from denaturating gradient gel electrophoresis (DGGE) and flueorescence in situ hybridization (FISH) to 16S rRNA clone libraries (Hentschel et al. 2006) and direct sequencing of 16S rRNA from environmental samples. Studies recovered 32 different bacterial phyla (Schmitt et al. 2012) in sponges with the most abundant ones appertaining to the *Proteobacteria*, *Chloroflexi*, *Actinobacteria*, *Nitrospirae* and the candidate phylum *Poribacteria* (Hentschel et al. 2012). Further phylogenetic analysis showed, that bacterial phyla fall into sponge–specific clusters, (Hentschel et al. 2002) but with low abundances in other environements such as sediment or seawater (Taylor et al. 2013; Webster et al. 2010). Species–specific microbes of sponges have been found to be stable across geographic distances (Taylor et al. 2005), over different time periods (Friedrich et al. 2001) starvation, antibiotic treatment and transplantation (Friedrich et al. 2001; Thoms et al. 2003). Latest studies revealed varying richness of sponge microbes ranging from 50 to 3 820 distinct symbionts but with a low variablity between the hosts of the same species. Further, 41 different phyla were identified with at least 13 of them residing in all 81 investigated sponge species (Thomas et al. 2016). Thus, sponges show a distinct community with low abundance in other environments, which is stable within the same host species.

Within the last years, not only the community composition, but also their functional repertoire attributing to commensal interactions with the host was investigated. Heterotrophic microorganisms have been reported many times to be associated with sponges (Hentschel et al. 2012; Taylor et al. 2007), not surprising as sponges provide a high amount of nutrients, and thus represeent an ideal environment. Organic carbon assimilation was found to be mediated by different sponge species and their microbial symbionts (Goeij et al. 2008; Yahel et al. 2003). The degradation of carbohydrates is performed by the sponge–specific candidate phylum *Poribacteria*, and may thus be considered as relevant for symbiosis (Kamke et al. 2013). One carbon metabolism has been reported for methanotrophic bacteria associated to deep–sea sponges (Vacelet et al. 1995) and *Arenosclera brasiliensis* (Trindade-Silva et al. 2012). Further examples are the photosynthetic carbon fixation through cynobacteria, which provide up to 50 % of energy for their respective host (Cheshire and Wilkinson 1991; Steindler et al. 2002; Wilkinson 1983) or methylotrophy for Gammaproteobacteria revealed through metatranscriptomics (Moitinho-Silva et al. 2014).

Apart from carbon, nitrogen cycling has been well–investigated, as it is the base for the synthesis of amino acids and proteins (Taylor et al. 2007). Especially in oligotrophic environments, such as coral reefs, symbiotic microbes may contribute to the nitrogen level of sponges through fixation of atmospheric nitrogen (Wilkinson et al. 1999). Moreover, archaea and bacteria were found to be involved in nitrification processes and relevant functional genes, such as *amoA*, (Bayer et al. 2008; Radax et al. 2012) were detected using several meta–omic approaches (Fan et al. 2012; Liu et al. 2012b; Radax et al. 2012; Thomas et al. 2010). Specific transport functions between symbionts and their sponge host were revealed (Liu et al. 2012b), suggesting a close link for this relationship. Notably, not only the symbionts benefit from the sponge–excreted *waste*, but also sponges digest their symbionts (Vacelet et al. 1996) to gain energy and nutrients. Aside these major cycles, symbionts were shown to be able to degrade halogenates using a metageneomic approach in *Aplysina aerophoba* (Bayer et al. 2013), synthesizing vitamins and cofactors (Fan et al. 2012; Hentschel et al. 2012; Thomas et al. 2010) or protect the sponge and help to respond to variable environmental conditions by expressing stress protection proteins (Fan et al. 2012; Liu et al. 2012b). The high abundance of CRISPR in symbionts compared to seawater microbes (Fan et al. 2012) may further protect the host from phages.

Sponges are a well–known source for diverse secondary metabolites, which are considered a chemical defense system (Proksch 1994) helping them to defend against harmful microbes (Taylor et al. 2007) and regulate their numbers. The diverse organic compounds include terpenoids, peptides, alkaloids (Lejon et al. 2011), PKS, nonribosomal peptide synthetase (NRPS), showing anti–microbial, anti–fungal or even anti–cancer activity (Abdelmohsen et al. 2012, 2014c; Blunt et al. 2016). Many of the bioactive compounds have been identified to be produced by their symbionts rather than the sponge (Piel 2009). In particular, actinomycetes have been found to be a rich source of secondary metabolites among sponge–

associated bacteria (Abdelmohsen et al. 2014c; Cheng et al. 2015). As much as 50 % of all bioactive compounds can be traced back to actinomycetes (Abdelmohsen et al. 2014a). This has drawn much attention to these symbionts. Among the plethora of secondary metabolites, polyketides are often in focus "[f]rom a drug discovery perspective" (Della Sala et al. 2014) and due to their diversity and various pharmacological activities. Polyketides are catalyzed by PKS in a modular fashion. PKS were isolated from metagenomes or via PCR–approaches from symbionts of different marine sponges (*Theonella swinhoei*, *Plakortis simplex*, *A. aerophoba*) and shown to produce the antitumor polyketides psymberin (Fisch et al. 2009), onnamide (Piel et al. 2004) and swinholide A (Bewley et al. 1996). Among the different PKS classes (cis–AT, trans–AT, FAS–like), two were found to be sponge exclusive: sup–PKS (symbiont ubiquitous type I PKS) (Fieseler et al. 2007) and *Swf* (Della Sala et al. 2014). Both represent mono–modular PKS, and the sup–PKS is predicted to produce mid-chain-branched fatty acids (Fieseler et al. 2007). Further, the sup–PKS cluster may be linked to *Poribacteria* (Hochmuth et al. 2010; Siegl and Hentschel 2010).

The microbial ecology, specifically community structure and function has received much recent attention. This is driven by the recognition of the holobiont concept – describing "[a] network of interactions between a host and its symbiotic microbial consortia" (Webster and Thomas 2016) within an environment (Figure 7). Patterns are independent whether the host is a plant, an animal or human. This concept allows for overarching questions to be addressed to understand the function of this kind of association as a whole (Bosch and Miller 2016). With latest sequencing technologies, insights into functional aspects are

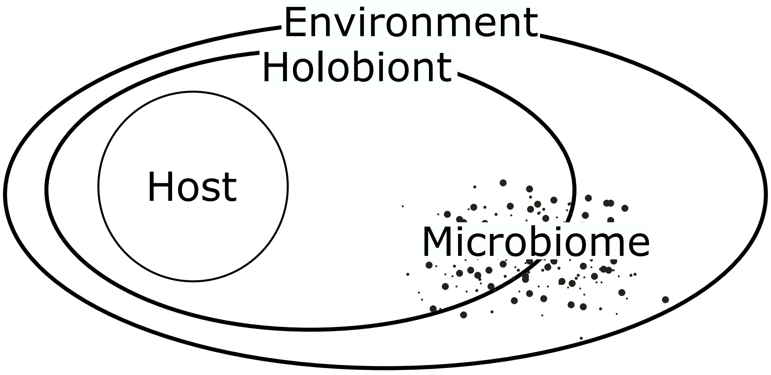


Figure 7. – The holobiont concept describing the asociation between a host and its microbiome in a defined environment.

now possible. Despite the shown efforts and findings, important questions as "How stable are microbial communities?", "What are effects on the host?" or "What are functional drivers of community structure?" remain unanswered. Future directions head towards the use of synthetic communities (e.g. (Bodenhausen et al. 2014)) to link observations with genomic information (Vorholt 2012) and the further integration of meta–omic studies in a combination with cultivation techniques (Müller and Ruppel 2014) to further clarify the functional capabilites of the microorganisms. These will help to understand the complex relationship between hosts and their associated microbes.

Scope and structure

The overall goal of this thesis was to explore bacterial genomic and metagenomic data of microbes and microbial comunities associated with different hosts – the leaf surface of *A. thaliana* and microbial communities within marine sponges. As metagenomic data is complex due to its dimensions and amounts of data, this states a challenge. The first task was to make that data accessible for downstream analysis. In particular, I analyzed functions contributing to host–microbe interactions, adaptions to the host/environment, and defense mechanisms. Further, genomes and metagenomes were screened for secondary metabolite genes. This thesis has four major goals and is structured as follows:

Part I, an overall introduction to sequencing, arised –omics and their terminology and applications are described accompanied by examples from previous studies on microbes the phyllsophere and sponge environments. Part II comprises material and methods used in this thesis. Part III describes the results as follows:

Chapter 1 Investigate, develop and compare methods for bacterial genome assembly. The main focus was on the decontamination and clean–up processes to generate high–quality draft genomes ready for downstream analysis.

Chapter 2 Generate a draft genome and study the phyllosphere bacterium, *Williamsia* sp. ARP1. The genomic sequence was analyzed in terms of adaptions to the phyllosphere environment. Further, the taxonomic relationship of the genera *Williamsia* and *Gordonia* was explored based on a taxonomic marker and on the genomic level.

Chapters 3–4 Investigate bacterial isolates from sponges for presence of secondary metabolite genes. The genomes of six bacterial isolates were analyzed in terms of secondary metabolism (=bioprospecting) and possible natural products using state–of–the–art software.

Chapter 5 Explore four microbial metagenomes from sponges and seawater towards bacterial defense systems. Metagenomes – originating from three mediterranean sponges and from seawater – were compared and differences regarding defense systems comprising CRISPR and restriction modification system (RMS). In addition, the genomic composition and features as the GC distribution were exposed.

Within Part IV the obtained results are discussed including the limits and potential of –omics data, and compared to similar studies. In Part V, the results are concluded and an outlook for future projects is given.

Part II.

Material and methods

Material

Samples and sequencing data

All samples were obtained from studies carried out and/or supervised by colleagues of the Department of Botany II, University of Würzburg, Germany. Described below are the sampling sites and necessary steps to obtain sequencing data, associated people and companies.

Samples for genomic sequencing

Bacterial strains were isolated from the phyllosphere of *Arabidopsis thaliana* in June 2012 in the Botanical Garden of the University of Würzburg, Germany (N 49.765°, E 9.932°) or from sponges derived from the Mediterranean Sea in Milos, Greece (N 36.767°, E 24.514°), Rovinj, Croatia (N 27.794°, E 34.215°) in May, 2013. Growth of the strains on substrate was carried out by Cheng Cheng, Usama Abdelmohsen (Department of Botany II, University of Würzburg, Germany) or by myself. Christine Gernert did the DNA extraction and PCR amplification of 16S rRNA for all strains in which I participated.

Library preparation and DNA sequencing on a MiSeq platform (Illumina, Inc, Department of Human Genetics, University of Würzburg, Germany) was carried out by Wiebke Sickel and Alexander Keller (both Department of Zoology III, University of Würzburg, Germany) with a 250 bp paired–end library design and targeted insert size of 400 bp. If necessary, two libraries were sequenced to reach enough throughput and coverage (Table 1).

Samples for metagenomic sequencing

Sponge samples of *Petrosia ficiformis*, *Sarcotragus foetidus* and one seawater sample were collected on May, 25th 2013 via SCUBA diving in Milos, Greece at a depth of 5–7 m (N 36.767°, E 24.514°). DNA extraction was carried out by Lucas Moitinho–Silva ands Kristina Bayer (both Department of Botany II, University of Würzburg, Germany). Library construction as well as DNA sequencing of associated microbes was performed at an external company (GATC Biotech AG, Konstanz, Germany). Libraries were designed using a paired–end configuration with 250 bp or 300 bp.

The samples for the sponge *Aplysina aerophoba* were collected on May, 7th 2013 via SCUBA diving from the Adriatic Sea (northernmost arm of the Mediterranean Sea) in the Gulf of Piran, Slovenia at a depth of 5 m (approx. N 45.530°, E 13.566°) . DNA extraction was carried out by Beate Slaby (Department of Botany II, University of Würzburg, Germany), library design by Tanja Woyke (JGI, Walnut Creek, CA, USA), library construction and

DNA sequencing of associated microbes was carried out by the Joint Genome Institute (JGI, Walnut Creek, CA, USA). A total of six libraries (3x pinacoderm, 3x mesohyl) were sequenced on an Illumina HiSeq 2000/2500 platform (Illumina, Inc., San Diego, CA, USA) in paired–end configuration and a length of 100 bp with a targeted insert size of 170 bp (Table 1).

Table 1. – Overview of paired–end sequencing data of used genomic and metagenomic libraries. Shown are the isolation source, total number of reads, library size, read length, targeted insert size and relative GC content (average of all libraries) per sample

Isolate	**ID**	**Isolation source**	**Sequencing platform**	**Raw reads [#]**	**Size [bp]**	**Read length [bp]**	**Insert size [bp]**	**GC–content [%]**
Streptomyces sp. SBT349[a]	HH1	Sponge host: *Sarcotragus spinosulus* Milos, Greece	MiSeq	4,094,430	1,023,607,500	2x250	400	71
Nonomuraea sp. SBT364[a]	HH2	Sponge host: *Sarcotragus foetidus* Milos, Greece	MiSeq	7,003,186	1,750,796,500	2x250	400	70
Nocardiopsis sp. SBT366[a]	HH3	Sponge host: *Chondrilla nucula* Milos, Greece	MiSeq	4,231,116	1,057,779,000	2x250	400	69
Williamsia sp. ARP1	HH6	Plant leaf: *Arabidopsis thaliana* Würzburg, Germany	MiSeq	2,608,588	652,147,000	2x250	400	59
Micromonospora sp. RV043	RV43	Sponge host: *Aplysina aerophoba* Rovinj, Croatia	MiSeq	5,900,702	1,475,175,500	2x250	400	71
Rubrobacter sp. RV113	RV113	Sponge host: *Aplysina aerophoba* Rovinj, Croatia	MiSeq	2,206,732	551,683,000	2x250	400	64
Nocardiopsis sp. RV163	RV163	Sponge host: *Dysidea avara* Rovinj, Croatia	MiSeq	4,851,800	1,212,950,000	2x250	400	71
Sample	**ID**	**Isolation source**	**Sequencing platform**	**Raw reads [#]**	**Size [bp]**	**Read length [bp]**	**Insert size [bp]**	**GC–content [%]**
Microbial community	-	Sponge host: *Petrosia ficiformis* Milos, Greece	MiSeq	41,383,600	10,345,900,000	2x250	-	59.5
Microbial community	-	Sponge host: *Sarcotragus foetidus* Milos, Greece	MiSeq	32,672,426	9,801,727,800	2x300	-	62
Microbial community	-	Seawater Milos, Greece	MiSeq	40,505,000	12,151,500,000	2x300	-	42.5
Microbial community[b]	-	Sponge host: *Aplysina aerophoba* Piran, Slovenia	HiSeq 2000/ 2500	945,906,728	283,772,018,400	2x150	170	-

[a] Sequencing data depends on two paired–end libraries

[b] Sequencing data depends on six paired–end libraries

Hardware ressources

Main parts of bioinformatic calculations were carried out on a Fujitsu R920 Workstation (Fujitsu, Germany) running on 64-bit Ubuntu 12.04 (Precise Pangolin) or 14.04 (Trusty Fahr) as operating systems, dual core Intel® Xeon® CPU E5-2690 @ 2.90GHz × 16 and 256GB of memory (DDR3) and 4 x 2TB local hard drives.

Additionally, minor calculations and writing of manuscripts were performed on a Fujitsu Esprimo P9900 (Fujitsu, Germany) running on Ubuntu 12.04.3 and Microsoft® Windows 7™ Professional as operating system, Intel® Core™ i5-650 CPU @ 3.20GHz, 4GB of memory (DDR3) and 2TB of local storage space.

Databases

In use were the primary major public databases Genbank (Benson et al. 2014) and the EMBL nucleotide sequence databases (Kulikova et al. 2007; Stoesser et al. 1997). Further databases were GOLD (Reddy et al. 2015), and the SILVA high quality ribosomal RNA database (Quast et al. 2013).

Secondary and locally installed databases including PFAM 27.0 (Finn et al. 2016) , TIGRFAM 12.0 (Haft et al. 2013), Protein ANalysis THrough Evolutionary Relationships (PANTHER) 9.0 (Thomas et al. 2003), Simple Modular Architectur Research Tool (SMART) (Letunic et al. 2012; Schultz et al. 1998), COG(Galperin et al. 2015; Marchler-Bauer et al. 2015; Tatusov et al. 1997), were used. Additional local database to perform annotations were the NCBI non-redundant protein sequences (NR) (as of September, 2014), the NCBI nucleotide collection (NT) (as of September, 2014), the 16S Ribosomal RNA sequences for bacteria and archaea (16S, as of September, 2014), and RMS genes types 1–3 (comprising restriction endonucleases, methyltransferases and specifity domains) downloaded from The Restriction Enzyme Database (REBASE) (as of October, 2015).

Methods

General data processing

The processing of data was performed with the command line or scripts written in perl 5.12 (The Perl Programming Language, https://www.perl.org/), python 2.7.3 (Python Programming Language, https://www.python.org/) and R 3.0.2 (R Development Core Team 2014). Writing of scripts was done with the GNOME editor (gedit, https://wiki.gnome.org/Apps/Gedit), manuscripts and this thesis with either LaTeX(Latex Project Team 2010, https://www.latex-project.org/) or Microsoft® Word™ 2007/2010. As citation manager Thomson Reuters EndNote™ X7.3 (http://endnote.com/), Mendeley 1.16.1 (https://www.mendeley.com/) and JabRef 3.3 (www.jabref.org) were used. The manipulation of images and graphics was done with Inkscape 0.91 (https://inkscape.org/de/) and the GNU Image Manipulation Program 2.8 (GIMP, https://www.gimp.org/).

Bioinformatics

Processing of Sanger–sequenced data

If not stated otherwise, each approach within this section was applied to both, 16S rRNA sequences and sequences derived from PKS types I–II and NRPS PCR reactions.

Initial quality control

ABI-output files/chromatograms were initially analyzed for suspicious and/or multiple peaks due to contamination with Sequencher 4.9 (Gene Codes Corporation, Ann Arbor, Michigan, USA http://www.genecodes.com). ABI files were translated to the fastq/fasta format (Cock et al. 2010) using *seqret* from the EMBOSS 6.6.0 package (Rice et al. 2000). Obtained sequences containing chimeras were removed using Pintail 1.1 (Ashelford et al. 2005). Remaining sequences were quality trimmed with BWA's dynamic trimming algorithm (Li and Durbin 2009), removing nucleotides with a Phred quality score (Ewing and Green 1998; Ewing et al. 1998) ≥ 20 (=accuracy $\geq 99\,\%$).

Generation of consensus sequences

If available, reads of foward and reverse strands were merged with either the Contig assembly programm 3 (CAP3, Huang and Madan 1999) or by using a local alignment calculated with MUSCLE 1.3.8.31 (Edgar 2004) including quality scores for the overlapping positions. Alignments were curated manually if necessary.

Taxonomic and functional assignments

Genus level assignments to 16S rRNA consensus sequences were assigned either using the RDP classifier (Wang et al. 2007) or the Silva incremental aligner (SINA, Pruesse et al. 2012) with enabled *search and classify* option. Additional search against the nt or 16S ribsomal databases using the BLAST 2.2.28+ (Altschul et al. 1990) revealed closest (type–) strains, possible species level assignments for 16S rRNA sequences and possible functional assignments for PKS type I and II or NRPSs sequences.

Multiple sequence alignment and phylogenetic tree construction

For 16S rDNA sequences and closest relatives, multiple alignments were calculated with SINA (Pruesse et al. 2012) against the expert–based SILVA database with the bacterial variability profile. Best fitting substitution models for phylogenetic trees were choosen upon the Akaike information criterion (AIC) implemented in ModelGenerator 0.85 (Keane et al. 2006) – in all cases the generalised time reversible (GTR) model was found. A tree was constructed with a maximum–likelihood algorithm using RaxML 7.28 (Stamatakis 2006) or PhyML 3.0 (Guindon et al. 2010). Usually, 1000 bootstrap replicates were performed and consensus trees generated. All trees were saved in the Newick format.

Tree drawing

Obtained trees were drawn and edited with either TreeGraph 2 (Stöver and Müller 2010) or the interactive tree of life v2 (iTOL, Letunic and Bork 2011) and saved as portable document format (PDF) or scalable vector graphic (SVG). If necessary, trees were refined using Inkscape 0.91.

Processing of Illumina–sequenced data

If not stated otherwise, each approach within this section was applied to both, genomic and metagenomic datasets.

Initial quality control

Sequencing quality, overall throughput and integrity of reads was verified using FastQC 0.11.2 (Andrews 2016). Computed statistics as kmer–content, duplication levels, overrepresented sequences (adapter sequences), N content, GC content, and length distribution were evaluated and used for subsequent trimming. FastQC was applied to the data before and after the trimming process for data verification.

Quality trimming of reads

During sample preparation, sequencing libraries are constructed by ligating adapter, important for flow cell binding and amplification, to fragmented sequences. These adapter sequences as well as low quality regions have to be trimmed in order enable high–quality downstream analysis.

Sequences were submitted to Trimmomatic 0.32 (Bolger et al. 2014). Only sequences were retained, if both of a pair passed the trimming process with (i) trimming basepairs within a 10 bp window requiring a minimum phred score ≥ 25, (ii) an average phred quality score (Ewing and Green 1998; Ewing et al. 1998) ≥ 30 (=accuracy 99.9 %), and (iii) a minimum length ≤ 50 bp of the initial length.

Taxonomic assignments of metagenomic reads

Trimmed reads were submitted to the Metagenomic Rapid Annotations using Subsystems Technology (MG–RAST) server 3.3.6 (Meyer et al. 2008). Uploading the reads was done with default parameters and enabled contamination filter. The taxonomic profiling of MG–RAST is based on the NCBI taxonomy and the implemented lowest common ancestor (LCA) method (Huson et al. 2011) was used to assign taxonomic levels to the reads.

Sequence assembly

To analyze genomes and metagenomes for functional properties, their reads have to be stitched together (=assembled) into contiguous sequences (=contigs) (Nagarajan and Pop 2013) to reconstruct the original genomic sequence/s.

A protocol for the assembly of all genomic samples used within this thesis is given in Section 1, page 35. The workflow includes all steps, from initial quality control to annotation. It is focussed on the comparison of different assembler and the decontamination of the samples.

All paired reads which passed the quality control of the samples *P. ficiformis*, *S. foetidus* and seawater were merged with bbmerge.sh from the bbnorm release (Bushnell 2014).

Merged and unmerged reads for each sample were then assembled using IDBA–UD 1.1.1 (Peng et al. 2012) with kmer lengths ranging from 10 to 100 and a step size of 10. Reads obtained from the Integrated Microbial Genomes (IMG) (Markowitz et al. 2012) for the *A. aerophoba* sample were normalized with bbnorm (Bushnell 2014), and assembled using SPAdes 3.5.0 (Bankevich et al. 2012) with kmer sizes of 21,33,55,77,99,127, disabled mismatch corrector, and enabled single cell option to account for the uneven coverage of the sample. *A. aerophoba* was processed by B. Slaby.

Functional annotation

Data normalization In general, the metagenomic datasets were not of the same size, depth and coverage. In addition, the genes and genomes within these datasets are of different lengths. To account for these differences, metagenomic samples were normalized right before annotation. To perform the normalization, reads were mapped against their respective assembly using bowtie 2.2.24 (Langmead et al. 2009) with enabled very–sensitive option. For each contig, mapped reads were counted and the coverage for each basepair was calculated with samtools depth 1.0. The average coverage of a contig was set as the mean coverage over each position:

$$average\,coverage_{contig} = \frac{mapped\,basepairs_{contig}}{length_{contig}}$$

This coverage was divided by all mapped basepairs and this number was multiplied by 10^6 to obtain copy numbers per megabase:

$$copy\,per\,megabase = \frac{average\,coverage_{contig}}{all\,mapped\,basepairs} * 10^6$$

Each annotated feature was assigned the normalized copy per megabase from its respective contig.

RNA prediction To annotate tRNA features, contigs were submitted to tRNAscan–SE 1.3.1 (Lowe and Eddy 1997) using default parameters. rRNA was annotated in the same ways using RNAmmer (Lagesen et al. 2007) searching for small subunit rRNA (16S rRNA and 18S rRNA) within the kingdom bacteria. Non–coding RNAs (ncRNAs) were searched by submitting contigs to INFERNAL 1.1 (Nawrocki and Eddy 2013) using prokaryote covariance models obtained from Prokka/RFAM 11 (Burge et al. 2013; Seemann 2014) and an e–value threshold of 10^{-6}.

ORF prediction Coordinates of RNA features were masked (replacing nucleotides with X) within the contig sequences using bedtools (Quinlan 2014) to prevent Prodigal 2.6.1 (Hyatt et al. 2010) from predicting ORFs within non–coding regions. Prodigal was run with the following parameters: (i) genes were not allowed to predicted over/within gaps, (ii) genes were not allowed at the edge of sequences, (iii) using translation table 11 (prokaryotic),

(iv) in *normal* mode for single isolates or in *anon* mode for metagenomic datasets and (v) ORFs were outputted in nucleotide and protein format.

COG annotation Assignments of COGs was done via Reversed Position Specific BLAST (RPS BLAST) 2.2.28+ (Altschul et al. 1990). The COG position–specific scoring matrices were downloaded from the NCBI Conserved Domains Database (CDD) (as of September, 2014, `ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/little_endian/Cog_LE.tar.gz`) (Marchler-Bauer et al. 2015). RPS BLAST was run with an e–value of 10^{-6} and only top hits were retained for analysis.

Searching multiple protein databases To annotate protein functions of predicted ORFs, the tool Interproscan 5.18 (Jones et al. 2014) was conducted. It comprises several databases of which four were used for functional assignments: (i) PFAM 27.0 (Finn et al. 2016), (ii) PANTHER 9.0 (Thomas et al. 2003), (iii) TIGRFAM 12.0 (Haft et al. 2013) and (iv) SMART 6.0 (Letunic et al. 2012; Schultz et al. 1998). Only hits with at least 25 % identity to the target sequences, 70 % alignment length and an e–value $\leq 10^{-2}$ were retained for bacterial genomes. For metagenomic samples, searches were performed with an e–value of 10^{-6}.

CRISPR and *cas*–gene annotation For the bacterial genomes, CRISPR arrays were predicted on contigs with PILER–CR 1.06 (Edgar 2007) and default parameters.

For the metagenomic samples, a multiple–step approach was performed as proposed by Gogleva et al. (2014). In a first step, CRISPR arrays were predicted with PILER–CR 1.06 and CRT 1.1 (Bland et al. 2007) with direct repeat sizes ranging from 24 bp to 48 bp according to Haft et al. (2005). In addition, the Interposcan annotation of each contig was searched for *cas*–genes. Contigs with a found CRISPR array or *cas*–gene were validated with CRISPRFinder (Grissa et al. 2007a). Of these, only arrays with at least two repeats were retained as true hits.

To explore targets of CRISPR spacer, their sequences were submitted to CRISPRTarget and compared to four different databases: (i) ACLAME (as of August, 2009), (ii) RefSeq–Viral, (iii) GenBank–Plasmid and (iv) GenBank–Phage (all as of September, 2015), with an e–value of 0.1 and gap open costs of -5 (Biswas et al. 2013).

Direct repeats derived from CRISPR arrays were analyzed with CRISPRmap to obtain their structure and superclasses (Lange et al. 2013). In additon, they were compared to the CRISPRdb (Grissa et al. 2007b) with an e–value of 0.01 to validate their occurence in earlier studies.

Detecting genes of the RMS Reference sequences of type I–III RMS (each containing restriction endonucleases, methyltransferases, and type I additional specifity domains) were obtained form REBASE. A BLAST database was built for each type of restriction endonucleases, methyltransferases and specifiy domains. ORFs of the metagenomes were queried against all databases using blastp 2.2.28 with an evalue of 10^{-6} and a query coverage of $\geq 70\,\%$. If restriction endonuclease and methyltransferase (for type I also specifiy domain) were in genomic promximity (≤ 4 ORF apart from each other), a RMS was considered complete (Oliveira et al. 2014). To avoid double counts, overlapping regions of restriction endonucleases and methyltransferases of the same type within four genes were combined.

Secondary metabolite genes and possible products To screen for genes, which were characterizing for secondary metabolite clusters (PKS types I-IV PKS, NRPS, terpenes, lanthipeptides, bacteriocins, thiopeptides, linaridins, lassopeptides, microcins, proteusins, beta–lactams, siderophores, ectoines, butyrolactones, indoles, nucleosides, furans, homoserine lactones, phenazines, phosphonates, saccharides and fatty acids) ORFs were submitted to HMMER3 (Eddy 2009) using e–value cutoofs and HMMs obtained from the antiSMASH pipeline 3.0 (Weber et al. 2015). To avoid multiple counts, genes belonging to the same cluster type were joined, if they were found within a range of six ORFs as proposed by Doroghazi and Metcalf (2013).

For a more detailed view in the metagenomics datasets, contigs containing PKS type I genes were submitted to the antiSMASH web server to view complete secondary metabolite clusters and genes in near promximity. Possible products of PKS type I clusters were revealed by submitting their protein sequences to Natural Product Domain Seeker (NaPDoS) (Ziemert et al. 2012) using an e–value $leq\,10^{-5}$ for pathway assignments, condensation– and ketosynthase–domains.

Comparing the functional inventory To compare functional annotations of genes, either their relative abundance (genomes) or copies per megabase (metagenomes) were used. Heatmaps were drawn with the heatmap.2 function , Bray-Curtis dissimilarities were caluclated using the vegdist function from the *vegan* package (Dixon 2003) and clustered with the complete linkage method and euclidian distance. Both functions were implemented in R (R Development Core Team 2014).

ANI and ortholog detection

To compare relatedness among prokaryotic strains, the ANI was calculated with BLAST as proposed by Goris et al. (2007) and Mummer as implemented in JSpecies (Richter and Rosselló-Móra 2009). As this measure depends only on contigs split into fragments of 1020 nucleotides — and therewidth includes also non–coding regions — and two–sided BLAST searches between bacterial strains, an additional method was used.

Orthologs (=genes that have evolved from common ancestry by speciation) were searched between strains, based on genes with a valid COG annotation. This procedure was performed using InParanoid 8 (Sonnhammer and Östlund 2015). A two–sided BLAST was conducted for each pair of orthologous genes. Their summed up similarities divided by the total number of orthologs was used as a comparative value to ANI.

Part III.

Results

Chapter 1.

De novo assembly protocol for Illumina–sequenced bacterial genomes

Author: **Horn, H.**

This chapter describes an unpublished protocol to assemble sequenced bacterial genomes. The protocol was essential for the results shown in Chapters 2–5.

Abstract

Sequencing bacterial genomes has become a standard procedure nowadays. Existing tools focus on automated and fast assembly algorithms, but lack steps of curation. As a consequence, many of these pipelines may lead to assemblies including errorneous and contaminated contigs. To enable also biologists untrained in bioinformatics, I developed a protocol to perform high–quality bacterial assemblies based on Illumina–sequenced data.

With this 5–step–protocol one is able to produce high–quality assemblies of bacterial isolates. The included decontamination step based on single copy genes, coverage and GC–content was shown to be able to delineate between real data and contamination in a non–isolate sample. The usage of multiple assemblers in a competetive way enables the user to (i) assign the *best* algorithm to each dataset and (ii) validate the robustness of the assemblies. Performing all steps of this protocol offers assemblies of high quality and a starting point for downstream analysis.

Introduction

Since the invention of second-generation sequencing, the costs for sequencing a bacterial genome has decreased by orders of magnitudes. Contrary, the numbers of sequenced genomes has increased dramatically and with that the demand for bioinformatic tools and protocols to analyse them (Howison et al. 2013). Many different sequencing platforms have been developed so far, comprising 454 pyrosequencing (454 Life Sciences), SOLiD (Applied Biosystems), Ion PGM (Thermofisher), and PacBio RS II (Pacific Biosciences). The most prevalent technology is Illumina (Illumina, Inc.) with its HiSeq and MiSeq platforms generating the short reads for most projects and also the ones described in this thesis.

The assembly of short reads is a complex problem (Koren et al. 2014). Assembly tools use algorithms (de Bruijn graphs or Overlap Layout Consensus) for generation and varius heuristics for assembly optimization (Miller et al. 2010). Many *de novo* assemblers have been developed so far, comprising SPAdes (Bankevich et al. 2012), IDBA–UD (Peng et al. 2012), MaSuRCA (Zimin et al. 2013), Velvet (Zerbino 2010), SOAPdenovo (Luo et al. 2012), ABySS (Simpson et al. 2009), and more. Many of them were tested and evaluated on different bacterial datasets within the Assemblathlon (Earl et al. 2011) or the Genome Assembly Gold-standard Evaluation for Bacteria (GAGE–B) (Magoc et al. 2013) and are implemented in pipelines such as Computational Genomics–pipeline (CG–pipeline) (Kislyuk et al. 2010), A5 (Coil et al. 2015), MyPro (Liao et al. 2015), or iMetAMOS (Koren et al. 2014). However, these pipelines focus on fully automated assembly of sequencing data and, with exception of iMetAMOS, lack validation as likelihood scores (i.e. Clark et al. 2013) or substantial decontamination of assemblies.

For this purpose, an easy–to–use protocol for the *de novo* assembly of Illumina–sequenced prokaryotic genomes is introduced. Other than comparable protocols or pipelines, it makes use of comprehensive decontamination steps including tools used for metagenomic binning, incorporating single copy genes, GC–content and coverage information of contigs. Described here is the entire process from raw reads processing, quality control, comparison and competition of different assemblers, refinement of the best assembly and final scaffolding. This process ends up with a high–quality assembly and basic annotation making it accessible for downstream analysis.

Material

Isolates

Seven bacterial isolates were assembled to test this protocol. The samples originated from sponge tissues and the *Arabidopsis thaliana* phyllosphere and belonged to the phylum *Actinobacteria* (Table 2). For collection and pre–processing of samples, see Section II. All of the isolates are discussed in the following chapters and are published.

Table 2. – Overview of the sequencing data for the used bacterial isolates before and after application of quality trimming. Read numbers are given as the sum of forward and reverse reads.

Isolate	**ID**	**Sequencing platform**	**Raw reads [#]**	**Trimmed reads [#]**	**DeconSeq reads [#]**	**Final reads [#]**	**Reference**
Streptomyces sp. SBT349[a]	HH1	MiSeq (2x250bp)	4,094,430	4,044,464	4,003,548	4,003,134	**Horn** et al. 2015a
Nonomuraea sp. SBT364[a]	HH2	MiSeq (2x250bp)	7,003,186	6,871,588	6,814,609	6,813,724	**Horn** et al. 2015a
Nocardiopsis sp. SBT366[a]	HH3	MiSeq (2x250bp)	4,231,116	4,151,036	4,148,532	4,148,422	**Horn** et al. 2015a
Williamsia sp. ARP1	HH6	MiSeq (2x250bp)	2,608,588	2,574,764	2,574,529	2,574,494	**Horn** et al. 2016a
Micromonospora sp. RV043	RV043	MiSeq (2x250bp)	5,900,702	3,522,538	-	3,522,538	**Horn** et al. 2015b
Rubrobacter sp. RV113	RV113	MiSeq (2x250bp)	2,206,732	1,683,986	-	1,683,986	**Horn** et al. 2015b
Nocardiopsis sp. RV163	RV163	MiSeq (2x250bp)	4,851,980	3,944,354	-	3,944,354	**Horn** et al. 2015b

[a] Sequencing data depends on two paired–end libraries

- This step was not performed for this isolate

Data deposition

Bacterial isolated of *Streptomyces* sp. SBT349, *Nonomuraea* sp. SBT364 and *Nocardiopsis* sp. SBT366 were registered under the BioBroject PRJNA280805 and were deposited in GenBank under the accession numbers LAVK00000000, LAVL00000000, and LAVM00000000

The sequencing project of *Williamsia* sp. ARP1 was deposited in Genbank under the BioProject PRJNA272726 and the accession number JXYP00000000.

Bacterial isolated of *Micromonospora* sp. RV043, *Rubrobacter* sp. RV113 and *Nocardiopsis* sp. RV163 were registered under the BioBroject PRJNA280805 and were deposited in GenBank under the accession numbers LEKG00000000, LEKH00000000, and

LEKI00000000.

Assembly workflow

Figure 8 details the 5–step–protocol, with the central steps (ii) and (iii), describing an assembly as a competition between mutliple assemblers and the evaluation of their outcomes.

The initial step was the **(i) pre–processing of the raw reads** by submitting them to FastQC for initial screening. Based on these results, reads were subjected to Trimmomatic 0.32 (Bolger et al. 2014) to remove low–quality regions (sliding window size 4, average quality ≥ 30) , sequences containing "Ns", adapters, and deleting subsequent short sequences (read length$_{\text{start}}$ - 50 nt $\geq$ read length$_{\text{end}}$). Next, viral (all NCBI Viral genomes, ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/) and human (Humane Genome Reference GRch37, ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/ARCHIVE/BUILD.37.3) contamination within the quality controlled reads were detected and removed using DeconSeq 0.43 (Schmieder and Edwards 2011). Remaining reads were subjected to KmerGenie 1.6213 (Chikhi and Medvedev 2014) to estimate the best k–mer length for single–kmer genome assemblers (i.e. Velvet and MaSuRCA) and to estimate an expected genome size. Within step **(ii), assembly** of quality controlled sequences using five different tools (A5–miseq, IDBA–UD, MaSuRCA, SPAdes and Velvet (Bankevich et al. 2012; Coil et al. 2015; Peng et al. 2012; Zerbino 2010; Zimin et al. 2013)) was performed.

Step **(iii)** comprised the **assembly evaluation**. First, contigs ≥ 1000 nt were discarded. Remaining contigs were subjected to several software packages listed below. Their results upon each assembly can be found in Table 4 and Figure 9. With Quality Assessment Tool for Genome Assemblies (QUAST) 2.3 assembly statistics as N_{50}, number of contigs, genome size, and the largest contig (Gurevich et al. 2013) were collected. It was run with default parameters. All reads were mapped back to their rexpective assembly using bowtie 2.2.24 (Langmead et al. 2009). Mapping statistics such as overall coverage, duplication of reads and quality of mapped reads were evaluated with Qualimap 2 (Okonechnikov et al. 2016) using the *Bam QC mode*. An likelihood score for each assembly was calculated with Assembly likelihood estimator (ALE) (Clark et al. 2013) and the use of the bowtie 2.2.24 output. This measure was based on distance of paired–end reads (insert size), read quality, sequencing coverage, read orientation, and the frequency of kmers. To calculate single copy genes per assembly, contigs were subjected to Phylosift 1.1 (Darling et al. 2014) using the *align mode* and enabled option for bacterial isolates. The number of assignments to each single copy gene were calculated, including multiple counts. Blobology was conducted to explore the assembly data for contamination by using GC content vs coverage plots including taxon annotations (Kumar et al. 2013) calculated with BLAST against the NCBI NT database with an evalue $\leq 10^{-6}$ only retaining top hits

The *best* assembly is selected upon the ranking of the calculated assembly statistics. For each metric, a number from best (1) to worst (5) is assigned according to Table 3 to each of the five assembly algorithms. The assembly with the smallest score after summarizing the

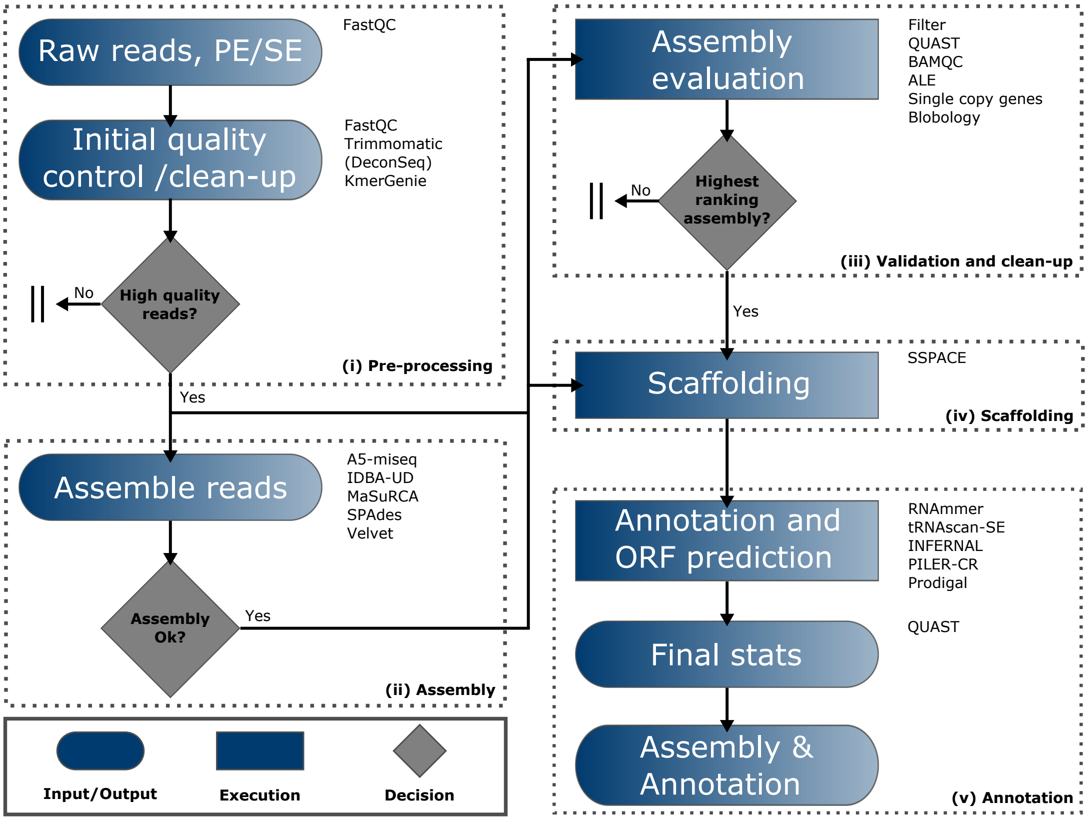


Figure 8. – Assembly flowchart for prokaryotic genomes. The first stage (i) of the pipeline checks the quality of incoming reads, performs quality trimming, removes contaminant reads and calculates a expected genome size. Then the pipeline assembles (ii) contigs with five different assemlber using the qualitry controlled reads. These contigs are then evaluated (iii) for length, contiguity and cleaned up, if necessary. Cleaned contigs are scaffolded (iv) with reads from stage (i). Finally, the scaffolds are annotated (v).

ranks is used for further refinement as clean–up through the GC–coverage plots (Figure 10) and **(iv) scaffolding** the contigs using all trimmed reads and SSPACE 3.0 (Boetzer et al. 2011) with default parameters. In the last step **(v), the annotation** of features as tRNA, rRNA, ncRNA and CRISPR was conducted using tRNAscan–SE 1.3 (Lowe and Eddy 1997), RNAmmer (Lagesen et al. 2007), INFERNAL 1.1 (Nawrocki and Eddy 2013) with the covariance models obtained form prokka (Seemann 2014), and PILER–CR 1.03 (Edgar 2007), respectively.

Table 3. – Metrics used for comparison and ranking assemblies produced by different assembler

Metric	Ranking decision	Tool
N_{50}	greater is better	QUAST
Largest contig	greater is better	QUAST
Coverage across contigs	more even is better	Qualimap
Assembly likelihood score	greater is better	ALE
Place score	greater is better	ALE
Insert score	greater is better	ALE
Depth score	greater is better	ALE
Kmer score	greater is better	ALE
Single copy genes	closest to 100% (=37/37 genes)	Phylosift
Single copy genes - duplication level	less is better	Phylosift

Results and discussion

The application of quality trimming to the reads is mandatory as the existence of low quality reads or basepairs may lead to unreliable sequences (Del Fabbro et al. 2013) and influences the downstram analysis. After trimming the reads of the first four genomes, the use of DeconSeq was cancelled as it had not that great effect on their number (Table 2). As an assembly is known as a testing problem, several assemblers were used to (i) ensure robustness (Koren et al. 2014) and (ii) find the algorithm fitting best to the given data.

Table 4. – Assembly statistics for all used isolates over five different assembly tools, a5–miseq, IDBA–UD, MaSuRCA, SPAdes and velvet. All statistics are based on contigs ≥ 1000 nt. Explanations of the values are given below the table.

Isolate	**ID**	**Expected size [Mbp]**	**a5-miseq**				**IDBA-UD**				**MaSuRCA**				**SPAdes**				**velvet**			
			Size	**N50**	**ALE**	**SCG**	**Size**	**N50**	**ALE**	**SCG**	**Size**	**N50**	**ALE**	**SCG**	**Size**	**N50**	**ALE**	**SCG**	**Size**	**N50**	**ALE**	**SCG**
Streptomyces sp. SBT349	HH1	7.88	8.03	16,038 (794)	$-6.47 * 10^7$	37 (6)	7.91	13,832 (893)	$-7.11 * 10^7$	37 (5)	error	error	error	error	8.07	18,975 (707)	$-6.74 * 10^7$	37 (6)	7.77	6494 (1604)	$-7.56 * 10^7$	37 (5)
Nonomuraea sp. SBT364	HH2	10.01	10.00	244,851 (433)	$-11.03 * 10^7$	37 (4)	9.96	160,153 (585)	$-11.53 * 10^7$	37 (4)	11.10	16,781 (1184)	$-14.15 * 10^7$	37 (5)	10.02	231,607 (378)	$-11.01 * 10^7$	37 (3)	9.83	17,292 (933)	$-12.04 * 10^7$	37 (4)
Nocardiopsis sp. SBT366	HH3	6.02	5.99	52,282 (309)	$-5.97 * 10^7$	37 (3)	5.96	114,171 (324)	$-6.37 * 10^8$	37 (3)	6.38	23.584 (541)	$-8.66 * 10^7$	37 (6)	6.04	171,199 (187)	$-6.23 * 10^7$	37 (3)	5.94	17,188 (548)	$-6.39 * 10^7$	37 (3)
Williamsia sp. ARP1	HH6	4.77	4.74	384,340 (52)	$-3.08 * 10^7$	37 (1)	4.73	428,053 (60)	$-3.07 * 10^7$	37 (1)	5.07	168,369 (312)	$-4.57 * 10^7$	37 (1)	4.75	428,335 (50)	$-2.97 * 10^7$	37 (1)	4.75	51,738 (161)	error	37 (1)
Micromonospora sp. RV043	RV043	7.67	-	-	-	-	8.48	116,864 (846)	$-7.88 * 10^{-7}$	37 (36)	-	-	-	-	8.69	253,928 (829)	$-8.08 * 10^7$	37 (11)	-	-	-	-
Rubrobacter sp. RV113	RV113	3.21	-	-	-	-	3.19	340,901 (53)	$-1.80 * 10^7$	37 (2)	-	-	-	-	3.19	401,647 (40)	$-1.78 * 10^7$	37 (2)	-	-	-	-
Nocardiopsis sp. RV163	RV163	6.03	-	-	-	-	6.04	87,173 (417)	$-5.84 * 10^7$	37 (3)	-	-	-	-	6.12	123.006 (359)	$-5.77 * 10^7$	37 (3)	-	-	-	-

Expected size: Calculated size with KmerGenie using trimmed reads

Size: Assembly size in Mbp

N50: N50 value in bp, number of contigs in brackets

ALE: Alignment likelihood score

SCG: Single copy gene number, multiple gene number in brackets. Maximal number is 37

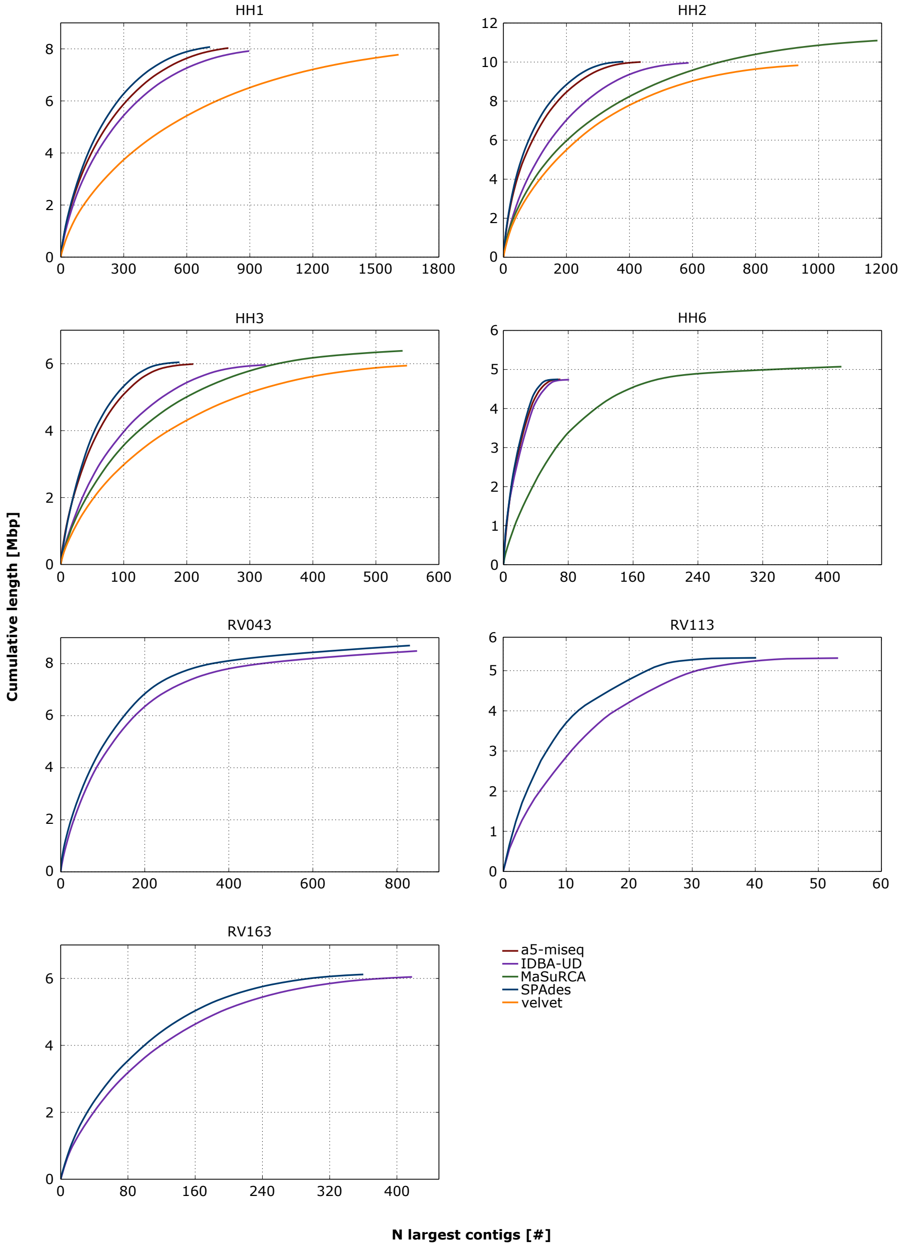


Figure 9. – Sequence length accumulation curve for all assemblies of the used isolates. Different assemblers are given in different colors. Steeper curves indicate longer contigs.

Assembly evaluation

For all genomes assembled, SPAdes worked best followed by IDBA–UD and the A5–pipeline in terms of contiguity as it produced the longest contigs with lowest number shown by the steepest curves in the QUAST output (Figure 9). Further, the SPAdes assemblies had the highest N_{50} values, except for *Nonomuraea* sp. SBT364 in which the A5–pipeline performed better (Table 4). Overall, the ALE scores were best (=smallest) for the A5–pipeline and SPAdes (Table 4), whereas MaSuRCA had the lowest scores for all datasets. With regards to the number of single copy genes, all assembler performed on the same level. Comparing the duplication levels of single copy genes, the MaSuRCA assemblies seemed to be more prone to duplications.

Due to the highly fragmented genome sequences calculated with Velvet, it was only applied to the first four datasets, and so was for MaSuRCA. In addition, MaSuRCA produced assemblies between 0.3 Mbp and 1 Mbp bigger compared to all other tools. IDBA–UD was selected over the A5–pipeline for the last three genomic datasets (IDs: RV043, RV113, RV163), as their performance was similar (Table 4). Mapping statistics between all assembler did not offer qualitative differences (data not shown).

Surprisingly, MaSuRCA did not work well on the given datasets. In contrast to the GAGE–B assembler evaluation, where best assemblies (with regards contiguity and assembly errors) were performed with SPAdes and MaSuRCA based on eight bacterial datasets (Magoc et al. 2013).

Assembly decontamination

During the assembly of bacterial genomes, the removal of contamination is often a forgotten task and leads to a reduced quality of public databases (Koren et al. 2014). To support the detection of contamination, the introduced protocol possesses various steps concerning the reads in the early stage as well as the contigs after the assembly. In the majority of the samples, the process of removing reads by mapping them against viral and the human genome had little effect on the number of reads. As a consequence, this step was not applied to all genomic datasets (Table 2). This process may be more suitable for sequencing projects, in which host and viral DNA may be extracted and sequenced as well. Thus, DeconSeq can better be used on metagenomic data, on which it was originally tested (Schmieder and Edwards 2011)

Two more decontamination steps were carried out on the *best* assembly. In all cases, SPAdes had the highest rank based on the chosen metrics (Table 4). In a first step, a scan for single copy genes of bacterial origin was carried out. Single copy genes occuring more than once were a sign for contamination. Within the most samples, between 1 and 6 genes were counted multiple times. With regards to sample *Micromonospora* sp. RV043, 11 and 36 double counts were found and assembly sizes significantly higher than expected (Table 4), indicating a contaminated sample. The dominant fraction of contigs was assigned to the phylum *Actinobacteria* (*Micromonospora*), but a large fraction of contigs (453 contigs, 0.62 Mbps) was assigned to *Rubrobacteriales* and had significantly less coverage

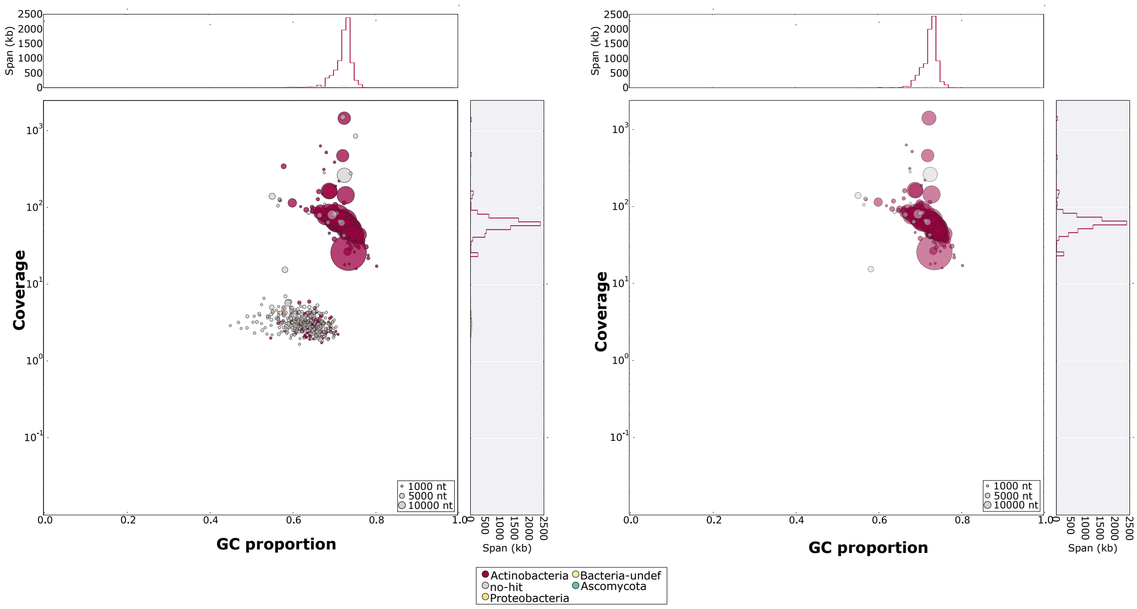


Figure 10. – Exemplary blobplot before (A) and after (B) clean–up of the strain *Micromonospora* sp. RV043 using SPAdes as assembler. The plot is showing contig coverage over GC content of each contig with assigned taxonomy based on BLAST analysis and contig length.

(Figure 10 A). After manual curation and deletion of contigs with low coverage and assignment to *Rubrobacterales* (Figure 10), the assembly was close to its expected size (Table 4) and single copy genes with multiple hits were reduced to 2. All other assemblies were also analyzed using GC content vs coverage plots. This lead to drastically reduced numbers of duplicated single copy genes of 1 or 2. Together with with a final round of scaffolding, contig numbers were up to 353 smaller than before (compare Table 4, Table 5).

Table 5. – Final assembly statistics for all used isolates.

		Assembly statistics									Annotation statistics[a]				
Organism	**ID**	**Assembler**	**SCG**	**Size [bp]**	**Contigs [#]**	**%GC**	**N50 [bp]**	**Largest contig [bp]**	**Coverage [fold]**		**tRNA [#]**	**rRNA [#]**	**ncRNA [#]**	**CRISPR repeats [#]**	**ORF [#]**
Streptomyces sp. SBT349	HH1	SPAdes +SSPACE	2	8,064,387	691	71.67	19,800	85,158	100		54	7	136	5	6939
Nonomuraea sp. SBT364	HH2	SPAdes +SSPACE	2	9,986,141	361	70.74	50,206	308,714	115		57	7	53	0	9338
Nocardiopsis sp. SBT366	HH3	SPAdes +SSPACE	1	5,784,200	177	72.72	60,060	171,199	146		57	8	51	11	5123
Williamsia sp. ARP1	HH6	SPAdes + SSPACE	1	4,745,080	50	68.63	140,970	428,355	65		71	5	21	2	4509
Micromonospora sp. RV043	RV043	SPAdes +SSPACE	2	8,069,233	376	72.46	35,501	253,928	72		96	3	111	0	7334
Rubrobacter sp. RV113	RV113	SPAdes +SSPACE	2	3,187,461	33	65.05	210,704	458,670	146		48	3	32	0	3127
Nocardiopsis sp. RV163	RV163	SPAdes +SSPACE	1	6,110,531	348	72.45	30,291	123,006	117	1	58	3	43	2	5320

[a] Annotation statistics may differ from those calculated by the NCBI prokaryotic pipeline as the underlying software differs from those used in this study

Comparison to other tools

Published pipelines as the A5–pipeline (Coil et al. 2015), MyPro (Liao et al. 2015), the CG–pipeline (Kislyuk et al. 2010) and iMetAMOS (Koren et al. 2014) are designed for the automated assembly of bacterial genomes/metagenomes in contrast to the step–by–step protocol presented in this study. Even though this might be seen as a disadvantage, it offers the possibilty for monitoring and the curation of each intermediate result as well as adaption of parameters.

Whereas the CG–pipeline and the A5–pipeline are limited to one assembler, MyPro includes 5, iMetAMOS 13, and this protocol also 5. This protocol and iMetAMOS can be extended by more assemblers according to the users needs. While MyPro focussed on higher contiguity obtained by the integration of different assemblies by use of Contig Integrator for Sequence Assembly (CISA) (Lin and Liao 2013), this step was left out within this protocol. In GAGE–B, it was shown to cause misassemblies and inferior assemblies compared to individual ones. Moreover, finding assemblies which complement each other is an extensive process requiering many trials (Magoc et al. 2013). In addition, highest contiguity (i.e. highest N_{50} value) does not necessarely reflect the assembly of highest quality (Salzberg et al. 2012).

The A5–pipeline is targeted towards finding mis–assemblies and split contigs at these sides or correct them through mapping of reads. In this study, this lead to improved assemblies reagarding the number of contigs, in 3 of 4 cases better ALE scores compared to assemblies produced with IDBA–UD. In only two cases, the largest contig was bigger with A5, due to the misassembly correction. But, SPAdes overally outperformed the A5–pipeline with the used metrics. However, the use of mis–assembly correction may lead to assemblies with less errors and was also proposed for the iMetAMOS pipeline through an iterative process with the tool Recognition of Errors in Assemblies using Paired Reads (REAPR) (Hunt et al. 2013). Thus, an implementation would be meaningful. Of note, this tool and also the mis–assembly correction of the A5–pipeline is limited to data with paired–end information.

Of the named tools, only the here invented protocol and iMetAMOS make use of tools for comparing assemblies, decontamination and ranking of assemblies. As the protocol relies on ALE for computing assembly likelihood scores, in iMetAMOS two more tools, Computing Genome Assembly Likelihoods (CGAL) (Rahman and Pachter 2013) and Log Average Probability (LAP) (Ghodsi et al. 2013), are included for this purpose. Vice versa, iMetAMOS lacks analysis of coverage values. Within both approaches, QUAST is used to obtain assembly statistics.

Whereas the decontamination process in iMetAMOS relies on taxonomic classification through Kraken (Wood and Salzberg 2014), the approach used here includes taxonomic classification based on BLAST, but also GC–content and coverage information (Kumar et al. 2013). The use of such information is widely used in binning (i.e. separating) single genomes in metagenomic samples (e.g. Albertsen et al. 2013; Seah and Gruber-Vodicka 2015). As taxonomic assignments can not be done for unknown strains or reference genomes are may not be available, decontamination on taxonomic classification only is not possible, thus might not be detected by iMetAMOS. Including GC–content and coverage enables the identification of contaminant prokaryotic (low coverage) or eukaryotic (low coverage, low GC–content) contigs as shown in this study (Figure 10).

Conclusions

Overall, this study supports the use of state–of–the–art assemblers such as IDBA–UD and SPAdes as these performed well on the data produced on Illumina MiSeq platforms. The use of different assembly algorithms may also be useful for follow–up projects, in which different data (read length, coverage) is to be assembled. Also the ranking of different assemblies can be adapted to a users need from high contiguity (high N_{50}) to more errorless assemblies (low duplication levels of single copy genes) similar to iMetAMOS. The named pipelines may autmate many of the steps included in this protocol, but lack either the usage of several assembler or decontamination methods.

The implementation of mis–assembly correction might further improve the presented protocol. Extensive efforts were taken in the decontamination step by use of GC–content, coverage and singly copy genes and makes the presented protocol superior in cleaning–up assemblies compared to other tools.

Chapter 2.

Draft genome sequence of *Williamsia* sp. ARP1, a phyllosphere bacterium

Authors: **Horn, H.**, A. Keller, U. Hildebrandt, P. Kämpfer, M. Riederer, and U. Hentschel
Published: 2016
Journal: Standards in Genomic Sciences
Volume: 11,8
DOI: dx.doi.org/10.1186/s40793-015-0122-x

A reprint of this article is used here with the permission from the authors and BioMed Central. The documentation of the author's contributions is found in Part VII. Appendix. This article is open access and published under the terms of the Creative Commons Attribution International license 4.0 (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/)).

Horn *et al. Standards in Genomic Sciences* (2016) 11:8
DOI 10.1186/s40793-015-0122-x

Standards in
Genomic Sciences

SHORT GENOME REPORT Open Access

Draft genome of the *Arabidopsis thaliana* phyllosphere bacterium, *Williamsia* sp. ARP1

Hannes Horn[1,2], Alexander Keller[3], Ulrich Hildebrandt[1], Peter Kämpfer[4], Markus Riederer[1] and Ute Hentschel[1,2*]

Abstract

The Gram-positive actinomycete *Williamsia* sp. ARP1 was originally isolated from the *Arabidopsis thaliana* phyllosphere. Here we describe the general physiological features of this microorganism together with the draft genome sequence and annotation. The 4,745,080 bp long genome contains 4434 protein-coding genes and 70 RNA genes. To our knowledge, this is only the second reported genome from the genus *Williamsia* and the first sequenced strain from the phyllosphere. The presented genomic information is interpreted in the context of an adaptation to the phyllosphere habitat.

Keywords: Draft genome, Phyllosphere, *Williamsia* sp. ARP1, Adaptation, Whole genome sequencing, Next generation sequencing, Assembly, Annotation, *Arabidopsis thaliana*

Introduction

The genus *Williamsia* was originally proposed by Kämpfer et al. in 1999 [1] to accommodate an unusual mycolic-acid containing actinomycete. Members of the genus *Williamsia* are Gram-positive, non-spore forming, and form round, orange colonies. Their cell shape is coccoid- or rod-like [2]. The genus *Williamsia* forms a distinct group within actinomycetes of the suborder *Corynebacterineae* [3], which also comprises the genera *Corynebacterium*, *Dietzia*, *Gordonia*, *Mycobacterium*, *Nocardia*, *Rhodococcus*, *Skermania*, *Tsukamurella and Turicella*. Based on the mycolic-acid profile with carbon chain lengths ranging from 50 to 56, the genus *Williamsia* is likely to be placed between the genera *Gordonia* and *Rhodococcus* [1]. At the time of writing, only one other draft genome of *Williamsia* sp. D3 was publicly available [4] and nine species of this taxon were recognized with valid scientific names: *Williamsia deligens* [5], *Williamsia faeni* [6], *Williamsia limnetica* [7], *Williamsia marianensis* [8], *Williamsia maris* [9], *Williamsia muralis* [1], *Williamsia phyllosphaerae* [10], *Williamsia serinedens* [11] and *Williamsia sterculiae* [12]. Further this genus has been linked with the degradation of hexahydro-1,3,5-trinitro-1,3,5-triazine in soils as a sole nitrogen source [13], the degradation of carbonyl sulfide in soils [14] and polychlorinated biphenyls in tree habitats [15]. *Williamsia* was isolated from various sources, including indoor building material [1], human blood [5] and following pulmonary infections [16], oil-contaminated and Antarctic soils [4, 11], extreme environments as glacier ice [17], deep sea sediments of the Mariana Trench [8], hay meadows [6], and the rare soil biosphere [18]. Besides, *Williamsia* was also reported as an endophyte of grey box eucalyptus tree roots [19] and as an epiphytic bacterium residing in the phyllosphere of white clover [20].

The phyllosphere, known as the aerial surface of plant leaves, is a short-lived environment [21] to diverse microorganisms of various taxonomic groups comprising bacteria, filamentous fungi, yeasts, viruses and protists. The phyllosphere presents a challenging environment for microbial colonizers with respect to climatic conditions, UV radiation, desiccation, water availability, reactive oxygen species, and in terms of antimicrobial compounds produced by the plant or possibly also microbes [21–25].

* Correspondence: uhentschel@geomar.de
[1] Department of Botany II, Julius-von-Sachs Institute for Biological Sciences, University of Würzburg, Julius-von-Sachs-Platz 3, D-97082 Würzburg, Germany
[2] GEOMAR Helmholtz Centre for Ocean Research, RD3 Marine Microbiology and Christian-Albrechts University of Kiel, Düsternbrooker Weg 20, D-24105 Kiel, Germany
Full list of author information is available at the end of the article

Additionally, the wax composition of the cuticle, surface characteristics such as stomata and veins affect nutrient availability and leaching, as they are likely to retain more water [23, 26].

Here, we present a summary, classification and general physiological features of the strain *Williamsia* sp. ARP1 together with the genomic sequencing, assembly, annotation, and its putative adaptions to the phyllosphere.

Organism information

Classification and features

The genus *Williamsia* belongs to the suborder *Corynebacterineae* [3] of actinomycetes owing to the presence of mycolic acid in the cell wall [2]. Since 2009, it was assigned to the family *Nocardiaceae* [27, 28]. *Williamsia* and other genera of this family form a distinct clade in a 16S rRNA phylogenetic tree as well as by using a combination of phenotypic markers [29]. In order to resolve the taxonomic position of *Williamsia* sp. ARP1, a 16S rRNA sequence (length of 1504 bp) derived from the assembled genome was compared with the NCBI non-redundant and 16S microbial database using BLASTn [30]. The five nearest sequences with the highest identity (all <100 %), the nine validly described *Williamsia* species, as well as representative sequences of the suborder *Corynebacterineae* – *Gordonia*, *Rhodococcus*, *Dietzia*, *Mycobacterium*, *Tsukamurella* and *Turicella* - were used for phylogenetic analysis. A strain of the family *Frankineae* was chosen as the outgroup. All 16S rRNA sequences were aligned using the SINA web aligner (variability profile: Bacteria) [31] and the phylogenetic tree was assessed using PhyML [32] with a generalised time reversible (GTR) substitution model, gamma distribution and 1000 bootstrap replications. All genera formed distinct clades (except *Rhodococcus*) and were well supported by bootstrap values ≥50 %. *Williamsia* formed two well supported distinct clades consisting of five and nine sequences, respectively. Within these clades, however, bootstrap values were weaker, due to low variation between 16S sequences. Closest sequences to *Williamsia* sp. ARP1 were *Williamsia* sp. 7B-582, A2-614 and A2-437 (all three originating from sediment), and phylogeny in this subclade could not be resolved better due to a multifurcation (Fig. 1).

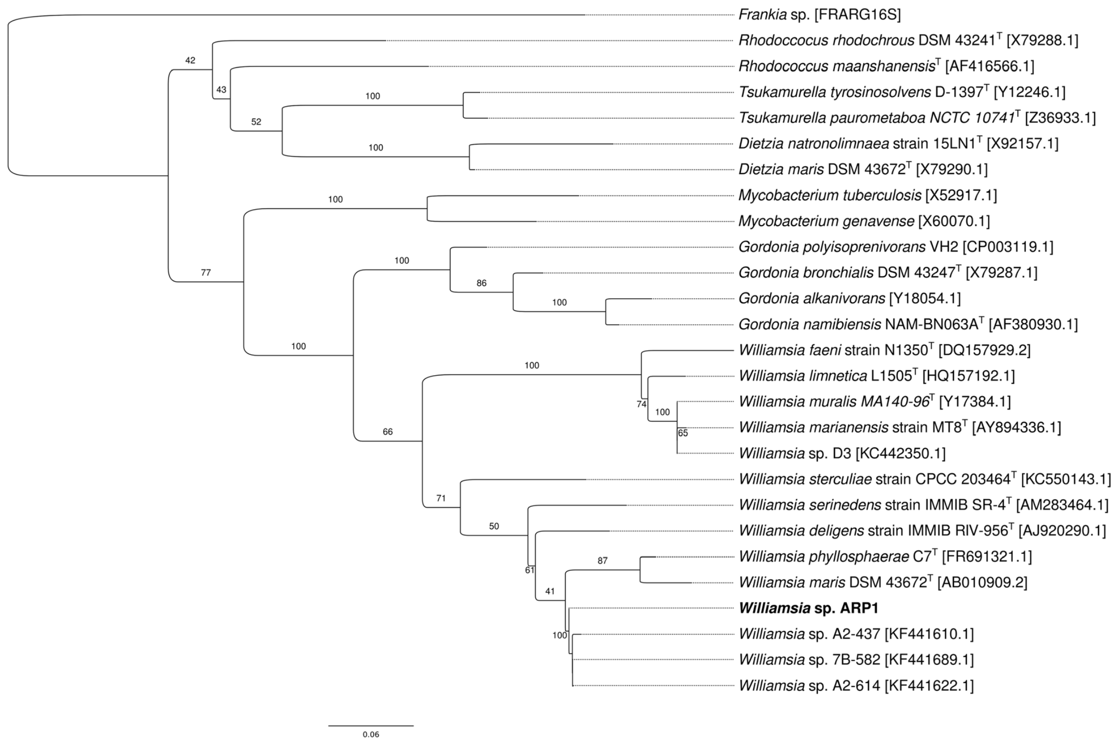

Frankia sp. [FRARG16S]
Rhodoccocus rhodochrous DSM 43241^T [X79288.1]
Rhodococcus maanshanensis^T [AF416566.1]
Tsukamurella tyrosinosolvens D-1397^T [Y12246.1]
Tsukamurella paurometaboa NCTC 10741^T [Z36933.1]
Dietzia natronolimnaea strain 15LN1^T [X92157.1]
Dietzia maris DSM 43672^T [X79290.1]
Mycobacterium tuberculosis [X52917.1]
Mycobacterium genavense [X60070.1]
Gordonia polyisoprenivorans VH2 [CP003119.1]
Gordonia bronchialis DSM 43247^T [X79287.1]
Gordonia alkanivorans [Y18054.1]
Gordonia namibiensis NAM-BN063A^T [AF380930.1]
Williamsia faeni strain N1350^T [DQ157929.2]
Williamsia limnetica L1505^T [HQ157192.1]
Williamsia muralis MA140-96^T [Y17384.1]
Williamsia marianensis strain MT8^T [AY894336.1]
Williamsia sp. D3 [KC442350.1]
Williamsia sterculiae strain CPCC 203464^T [KC550143.1]
Williamsia serinedens strain IMMIB SR-4^T [AM283464.1]
Williamsia deligens strain IMMIB RIV-956^T [AJ920290.1]
Williamsia phyllosphaerae C7^T [FR691321.1]
Williamsia maris DSM 43672^T [AB010909.2]
Williamsia sp. ARP1
Williamsia sp. A2-437 [KF441610.1]
Williamsia sp. 7B-582 [KF441689.1]
Williamsia sp. A2-614 [KF441622.1]
42
43
52
100
100
100
77
100
86
100
100
100
74
100
65
66
71
50
61
87
41
100
0.06

Fig. 1 16S rRNA gene based maximum likelihood phylogenetic tree highlighting the position of *Williamsia* sp. ARP1 within the suborder *Corynebacterineae*. The tree is based on 16 s rRNA sequences comprising the genera *Williamsia*, *Gordonia*, *Mycobacterium*, *Dietzia*, *Tsukamurella*, *Rhodococcus* and *Frankia* as an outgroup. The *Williamsia* sp. ARP1 is highlighted in *bold text* to show its position. The maximum-likelihood phylogenetic tree was generated using PhyML with the GTR substitution model. Numbers at the nodes are percentages of 1000 bootstrap replicates. Genbank accession numbers are indicated in *parentheses*; type strains are tagged with a *superscripted T*. The *scale bar* represents 0.06 substitutions per nucleotide position

All three 16S rRNA gene sequences showed a sequence identity of 99.93 % for strain 7B-582, 99.93 % for strain A2-614, 99.64 % for strain A2-437 to *Williamsia* ARP1. Minimum information about the genome sequence of *Williamsia* sp. ARP1 (MIGS) is provided in Table 1.

The colonies of *Williamsia* sp. ARP1 were orange to red in color on LB agar medium (Fig. 2a). Strain ARP1 was shown to be Gram-positive by Gram staining (data not shown). The cells of the strain were coccoid to rod-like with a diameter of about 1.0–1.5 μm (Fig. 2b). Further, the strain showed positive oxidase and catalase reaction and an aerobic respiratory metabolism. Cells were growing at a temperature range between 4 and 36 °C. Optimal growth was observed between 25 and 30 °C after 3 days on tryptic soy agar, Reasoner's 2A agar, and nutrient agar (all Oxoid). NaCl tolerance was investigated at different concentrations of NaCl (0.5–8.0 (*w*/*v*) %) in tryptic soy broth (TSB, Oxoid) with the cells growing in the presence of 1.0–6.0 % NaCl. The strain lacked motility after 3 days of growth in TSB at 30 °C, as observed under the light microscope. In agreement with this observation, a flagellum was not observed which is further backed up by the lack of flagellar genes (i.e., fliX, flgX and motX genes) on its genome. These findings were consistent with previous descriptions for this genus.

Genome sequencing information

Genome project history

The organism was selected for sequencing as part of ongoing *Arabidopsis* phyllosphere microbiology studies [33]. The sequencing project was completed in July 2014 and sequencing data was deposited as a Whole Genome Shotgun (WGS) project in Genbank under the BioProject PRJNA272726 and the accession number JXYP00000000

Table 1 Classification and general features of *Williamsia* sp. ARP1 [34]

MIGS ID	Property	Term	Evidence code[a]
	Classification	Domain *Bacteria*	TAS [73]
		Phylum *Actinobacteria*	TAS [74]
		Class *Actinobacteria*	TAS [3]
		Order *Actinomycetales*	TAS [3, 28, 75, 76]
		Family *Nocardiaceae*	TAS [3, 28, 75, 76]
		Genus *Williamsia*	TAS [1]
		Species *Williamsia* sp.	IDA
		(Type) strain: ARP1	IDA
	Gram stain	Positive	IDA
	Cell shape	Coccoid to rod-like	IDA
	Motility	Non-motile	IDA
	Sporulation	Non-sporulating	IDA
	Temperature range	4–36 °C	IDA
	Optimum temperature	25–30 °C	IDA
	pH range; Optimum	Not reported	NAS
	Carbon source	organic carbon	IDA
MIGS-6	Habitat	Phyllosphere	IDA
MIGS-6.3	Salinity	1.0–6.0 %	IDA
MIGS-22	Oxygen requirement	Aerobic	IDA
MIGS-15	Biotic relationship	Commensal	IDA
MIGS-14	Pathogenicity	Non-pathogenic	NAS
MIGS-4	Geographic location	Würzburg, Germany	IDA
MIGS-5	Sample collection	2012	IDA
MIGS-4.1	Latitude	49.766556	IDA
MIGS-4.2	Longitude	9.931768	IDA
MIGS-4.3	Depth	Plant surface	IDA
MIGS-4.4	Altitude	198 m above sea level	IDA

[a] Evidence codes - *IDA* Inferred from Direct Assay, *TAS* Traceable Author Statement (i.e., a direct report exists in the literature), *NAS* Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [77]

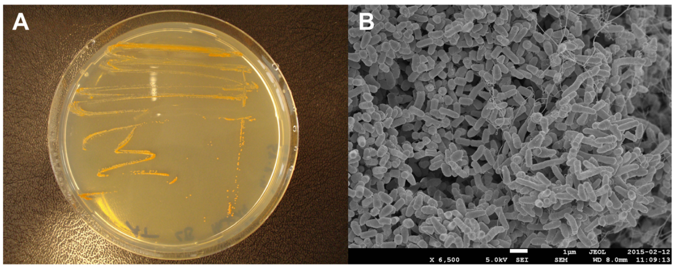


Fig. 2 General characteristics of *Williamsia* sp. ARP1. **a** The morphology of the colonies after three days of growth on LB-agar at 30 °C. **b** Image of *Williamsia* sp. ARP1 using scanning electron microscopy

consisting of 50 contigs (≥1000 bp). The genome sequencing was carried out with a MiSeq (Illumina Inc.) located in-house at our University. A summary of the project information according to the MIGS version 2.0 is shown in Table 2 [34].

Growth conditions and genomic DNA preparation

Several plants were collected from a Landsberg *erecta* (L*er*) population of *Arabidopsis thaliana* from the Botanical Garden (University of Würzburg, June 2012). Leaf washings [35] were used for inoculation of minimal media with C_{16} alkane (Sigma-Aldrich) as the sole carbon source in order to enrich for bacteria with the ability to degrade long-chain hydrocarbons. Aliquots were streaked (in duplicate) on agar plates prepared with minimal media and supplemented with C_{22} alkane (Sigma-Aldrich). This procedure provided a total of 17 isolates, of which most belonged to the genus *Rhodococcus* and two to genus *Williamsia* [33].

Williamsia sp. ARP1 was grown in 10 ml Luria-Bertani broth medium (10 g peptone, 5 g yeast extract, 5 g NaCl in 1000 ml demineralized water) for 24 h at 30 °C and rotary shaking at 180 rpm. For genomic DNA isolation, 2 ml of overnight culture were centrifuged at 8000 rpm for 5 min at room temperature. The pellet was rinsed in 1 ml TNE (1 ml 1 M Tris pH 8, 0.2 ml 5 M NaCl, 2 ml 0.5 M EDTA pH8, and 100 ml demineralized water) and resuspended in 270 μl TNEx (TNE, 1 % *v*/*v* TritonX-100) and 25 μl lysozyme (10 mg/ml). After a 30 min incubation at 37 °C, 50 μl of proteinase K (20 mg/ml) were added. After an incubation of 2 h and 55 °C, 15 μl of 5 M NaCl and 500 μl of 100 % EtOH were added. The mixture was then centrifuged at 13,000 rpm for 15 min at room temperature, rinsed with 70 % EtOH, air dried and resuspended in 150 μl TE buffer. The quality and quantity of the extracted DNA was evaluated by 0.8 % (*w*/*v*) agarose gel electrophoresis, by measuring absorption ratios 260/280 and 260/230 with a Nanodrop 2000c

Table 2 Project information

MIGS ID	Property	Term
MIGS 31	Finishing quality	Draft genome
MIGS-28	Libraries used	One Illumina paired-end library (400 bp insert size)
MIGS 29	Sequencing platforms	Illumina MiSeq
MIGS 31.2	Fold coverage	65×
MIGS 30	Assemblers	SPAdes 3.0, SSPACE 3.0
MIGS 32	Gene calling method	Prodigal 2.6.1
	Genbank ID	JXYP00000000
	Locus Tag	TU34
	GenBank Date of Release	July 1, 2015
	GOLD ID	Gp0118481
	BIOPROJECT	PRJNA272726
MIGS 13	Source Material Identifier	DSM 46827
	Project relevance	Phyllosphere, Environmental

Spectrophotometer (Thermo Fisher Scientific) and an additional Qubit dsDNA HS assay (Life Technologies).

Genome sequencing and assembly

High molecular weight DNA was cleaned with the DNA Clean & Concentrator kit (Zymo Research). The genomic DNA library for the Illumina platform was generated using Nextera XT (Illumina Inc.) according to the manufacturer's instructions. After tagmentation, size-selection was performed using NucleoMag NGS Clean-up and Size Select (Macherey-Nagel) to obtain a library with median insert-size around 400 bp. After PCR enrichment, the library was validated with a high-sensitivity DNA chip and Bioanalyzer 2100 (both Agilent Technologies, Inc.) and additionally quantified using the Qubit dsDNA HS assay (Life Technologies). Sequencing was performed on a MiSeq device using v2 2 × 250 bp chemistry, and the genome was multiplexed together with ten other bacterial genomes from other sources. Multiplexing was done via dual indexing, with the official Nextera indices N706 and S503 for *Williamsia* sp. ARP1.

In total, 1,304,294 (mean length 237.86 bp) raw paired-end sequences were subjected to the Trimmomatic software [36] for adapter and quality trimming (mean Phred quality score ≥30), filtering of sequences containing ambiguous bases and a minimum length of 200 bp. Subsequently, human and viral decontamination was excluded using DeconSeq [37]. The 1,287,247 (mean length 236.95 bp) remaining paired-end sequences were assembled with five different tools: a5-miseq [38], IDBA-UD [39], MaSuRCA [40], SPAdes [41] and Velvet [42]. In order to obtain the most reliable contigs, all assemblies were evaluated with QUAST [43], REAPR [44], ALE [45] and Feature Response Curves [46]. According to those evaluations, we have selected SPAdes assembler with enabled pre-correction and k-mer sizes ranging from 15 to 125 (step size of 10) as the best assembly. Obtained contigs were extended with remaining reads where possible. This led to 50 large contigs (≥1000 bp, N_{50}: 140,970 bp, longest contig: 428,355 bp) and an overall genome size of 4,745,080 bp (GC content: 68.63 %). As a final step, the contigs were ordered according to the nearest related complete genome by functional content using Mauve in 12 iterations [47]. As *Williamsia* sp. D3 was only available as a draft genome, *Gordonia bronchialis* was used for this step.

Genome annotation

Open reading frames were identified using Prodigal [48] followed by manual correction. The predicted coding sequences were translated into amino acid sequences and searched against COG position-specific scoring matrices obtained from the Conserved Domains Database [49] using RPS-BLAST [30]. Comparisons with TIGRFAM, Pfam, and PANTHER databases were performed with the InterProScan pipeline [50]. Only matches with an e-value ≤1 10^{-2}, ≥25 % identity and a minimum of 70 % alignment length to the target sequence were maintained. During this run, matches were also mapped to Gene Ontology terms. Additional gene prediction and functional annotation was performed with the Integrated Microbial-Genomes Expert Review [51] and the Rapid Annotation using Subsystem Technology webserver [52, 53]. Features as tRNA, rRNA, ncRNA, transmembrane helices, signal peptides, CRISPR elements and secondary metabolite gene clusters were predicted using tRNAscan-SE [54], RNAmmer [55], INFERNAL [56] and Prokka's prokaryotic RNA covariance models [57], TMHMM [58], SignalP [59] PILER-CR [60] and antiSMASH [61]. Searching for essential genes [62] was performed using HMMER3 [63]. Ortholog detection between *Williamsia* sp. ARP1 and three other genomes were carried out with InParanoid [64] whereas the mean percentage of nucleotide identity among the found orthologous genes was calculated using BLASTn. Average nucleotide identities between *Williamsia* sp. ARP1 and reference genomes were calculated with JSpecies [65].

Genome properties

The *Williamsia* sp. ARP1 draft genome sequence contained a total of 4,745,080 bp distributed over 50 large contigs (≥1000 bp) with an average GC content of 68.63 %. Of

Table 3 Genome statistics

Attribute	Value	% of total
Genome size (bp)	4,745,080	100.00
DNA coding (bp)	4,347,123	91.61
DNA G+C (bp)	3,256,678	68.63
DNA scaffolds	50	
Total genes	4509	100.00
Protein coding genes	4438	98.42
RNA genes	71	1.57
tRNA genes	45	1.00
rRNA genes	5	0.01
rRNA operons	1[a]	
Pseudo genes	0	0.00
Genes in internal clusters	NA	
Genes with function prediction	3505	77.73
Genes assigned to COGs	2207	48.95
Genes with Pfam domains	1330	29.50
Genes with TIGRFAM domains	793	17.59
Genes with signal peptides	334	7.41
Genes with transmembrane helices	1140	25.28
CRISPR repeats	2	0.04

[a]Only one RNA operon appears to be complete

the 4509 predicted genes, 4438 (98.42 %) were protein-coding, and 3505 (77.73 %) annotated with putative function. Pseudogenes were not detected. Genes not linked to a function were annotated as hypothetical or unknown function. Of these, 45 belonged to tRNA genes, 21 to ncRNA genes and five to rRNA genes (Table 3). One operon comprising a 16S rRNA, a 5S rRNA and a 23S rRNA gene was found. However two additional 5S rRNA genes suggest the presence of at least three rRNA operons. Functional assignments using COGs, a total of 2204 (59.59 %) of the coding sequences were classified into 23 different classes (Table 4, Fig. 3). Using TIGRFAM or Pfam, 793 (17.59 %) and 1330 (29.50 %) of the sequences could be classified (Table 3). For testing the genome completeness, a set of 111 essential gene markers was searched and 106 (=95.50 %) of them were present in *Williamsia* sp. ARP1. Except two marker genes (ribosomal proteins bS18 and bl28), all of them were found only once (Additional file 1). Within the RAST annotation, 1625 sequences were assigned to 402 metabolic subsystems. The highest ranking among the metabolic subsystems are linked to amino acids and derivatives (8.41 %), cofactors, vitamins and pigments (6.25 %), carbohydrates (5.77 %), protein metabolism (5.61 %), fatty acids, and lipids and isoprenoids (4.32 %) followed by stress response (2.86 %), (Fig. 4).

Insights from the genome sequence

The genome of *Williamsia* sp. ARP1 was smaller but displayed a higher CG content (68.63 %) than its nearest relative genomes (Table 5), thus rendering this genome more similar to the *G. bronchialis* and *G. polysoprenivorans* VH2 (67.00 and 66.96 %) than to *Williamsia* sp. D3

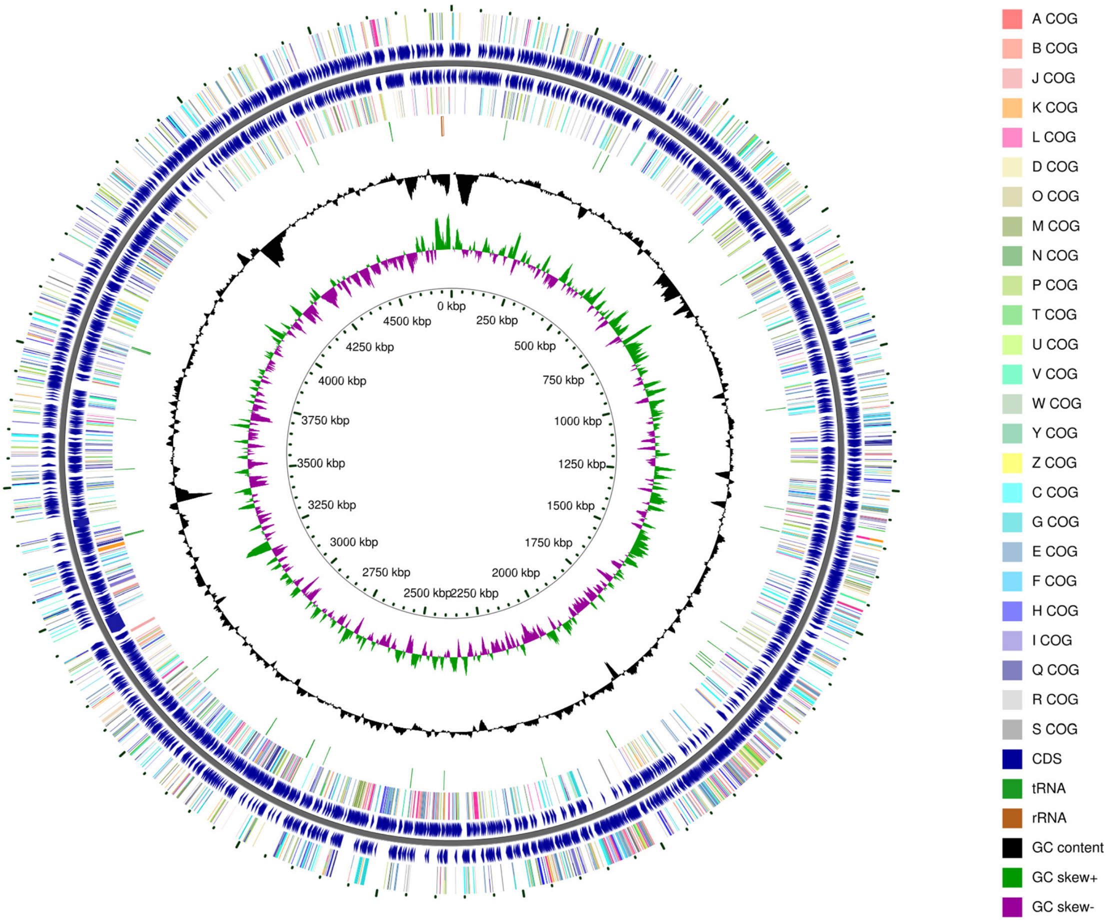


Fig. 3 Graphical circular map of the *Williamsia* sp. ARP1 genome. Starting from the outmost circle and moving inwards, each ring of the circle contains information of the genome: genes on the forward strand (*colored* according to their COG categories), CDS on the forward strand (*blue arrows*), CDS on the reverse strand (*blue arrows*), genes on the reverse strand (*colored* according to their COG categories), tRNA and rRNA genes on both strands (*green* and *orange*), GC content (*black*), GC skew (*green* and *purple*) and genome region by kbp

Table 4 Number of genes associated with general COG functional categories

Code	Value	% age	Description
J	143	3.17	Translation, ribosomal structure, and biogenesis
A	1	0.02	RNA processing and modification
K	183	4.06	Transcription
L	85	1.89	Replication, recombination, and repair
B	1	0.02	Chromatin structure and dynamics
D	0	0.00	Cell cycle control, Cell division, chromosome partitioning
V	31	0.69	Defense mechanisms
T	74	1.64	Signal transduction mechanisms
M	102	2.26	Cell wall/membrane biogenesis
N	11	0.24	Cell motility
U	18	0.40	Intracellular trafficking and secretion
O	79	1.75	Posttranslational modification, protein turnover, chaperones
C	184	4.08	Energy production and conversion
G	125	2.77	Carbohydrate transport and metabolism
E	226	5.01	Amino acid transport and metabolism
F	66	1.46	Nucleotide transport and metabolism
H	118	2.62	Coenzyme transport and metabolism
I	194	4.30	Lipid transport and metabolism
P	154	3.42	Inorganic ion transport and metabolism
Q	141	3.13	Secondary metabolites biosynthesis, transport and catabolism
R	346	7.67	General function prediction only
S	184	4.08	Function unknown
-	2231	49.48	Not in COGs

The total is based on the total number of protein coding genes in the genome

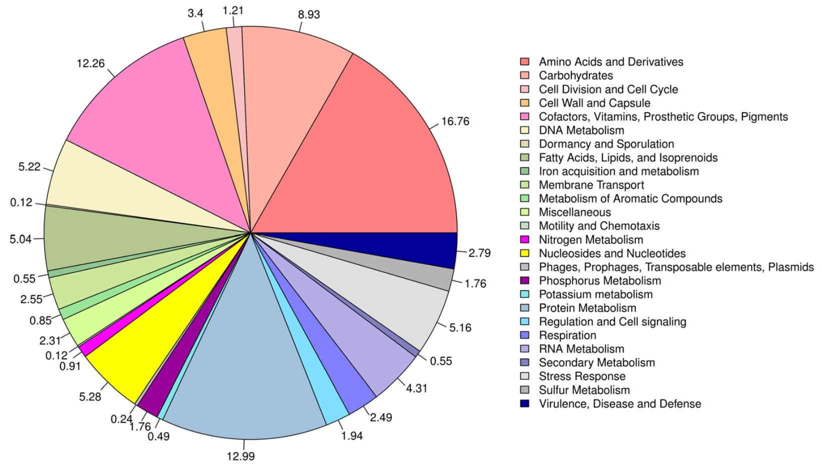


Fig. 4 Metabolic subsystems of *Williamsia* sp. ARP1 annotated through the RAST webserver

Table 5 Used actinomycete reference genomes in this study

Species	Strain	Accession number	Genome Size [Mbp]	G+C content
Williamsia sp.	D3	NZ_AYTE000000000.1	5.62	64.60
Gordonia bronchialis		CP001802.1	5.21	67.00
G. polysoprenivorans	VH2	NC_016906.1	5.67	66.96

(64.60 %) (Table 5). Considering the similarity between 16S rRNA sequences and its placement in the phylogenetic tree, strain ARP1 was however clearly assigned to the genus *Williamsia* (Fig. 1). With respect to orthologous genes, *Williamsia* sp. D3 was found to be the most similar strain to *Williamsia* sp. ARP1 with an average nucleotide identity of these orthologs of 75.53 %. Notably, the differences between *Williamsia* sp. ARP1 and the *Gordonia* strains and VH2 (75.17 and 74.84 % identity, respectively) is similar to the difference between the two *Williamsia* strains (75.53 %), (Additional file 2). Neither the clustering of COG classes nor the average nucleotide identities (ANI) were discriminative between the two genera (Fig. 5, Additional file 3). The ANI values are noticeably lower than the calculated cut-off values for species level identification (95) [66].

Extended insights

UV radiation UV radiation may impose stress on bacteria inhabiting plant leaves. In this context, a cluster of genes synthesizing mycosporins was found. These secondary metabolites are known to protect cells by absorbing UV light without generating reactive oxygen species (ROS) [67, 68]. Additionally, genes involved in the repair of UV-damaged DNA were found, which comprise DNA photolyases, the UvrABC endonuclease enzyme complex, and the DNA helicase II UvrD of the UvrABC system. The red color of *Williamsia* sp. ARP1 might protect it against photo-oxidative stress as pigmentation is known to be a common feature of phyllosphere colonizers [69]. All genes of the carotenoid biosynthetic pathway were found, consisting of a geranylgeranyl diphosphate synthase, a phytoene synthase, a phytoene desaturase, a carotene desaturase and a lycopene-β-cyclase. The products of this pathway are lycopene and β-carotene, both producing orange to red pigments.

Oxidative stress Further adaptions to an epiphytic lifestyle are encoded on genes responding to reactive oxygen species (ROS; e.g. hydrogen peroxide, superoxide, hydroperoxil radical), which are products of the plant defense [70, 71]. Here, two genes encoding for glutathione peroxidases, two superoxide dismutases with copper/zinc or manganese as active site, two glutaredoxins, three thioredoxins, and one catalase were found.

Temperature shifts Regarding temperature shifts, the heatshock chaperones DnaK, DnaJ and GrpE and the cold shock protein CspC were identified.

Uptake ABC transporters for the uptake of carbohydrates such as ribose, glycerol or maltose, amino acids such as methionine, known plant photosynthates such as fructose, and enzymes for fructose utilization were identified. Also, genes mediating the uptake of choline and subsequent

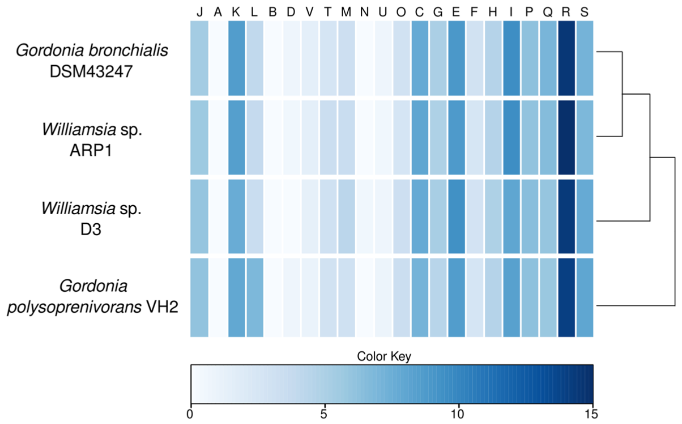


Fig. 5 Comparison of COG classes between strain ARP1 and reference genomes. The *color keys* provide the relative percentage of each COG class per genome. The dendrogram is based on correlation analysis

biosynthesis (choline dehydrogenase, betaine-aldehyde dehydrogenase) of the osmoprotectant betaine were found.

Desiccation Trehalose is a compatible solute and known to prevent cells from desiccation and water loss [72]. Eight genes encoding for the biosynthesis pathway (Malto-oligosyltrehalose synthase, 1,4-alpha-glucan (gly cogen) branching enzyme, GH-13-type trehalose-6-phosphate phosphatase, putative glucanase glgE, malto-oligosyltrehalose trehalohydrolase, glycogen debranching enzyme alpha, alpha-trehalose-phosphate synthase, glucoamylase) were identified.

Conclusions

The isolate ARP1 was isolated from the *Arabidopsis thaliana* phyllosphere. Phylogenetic analysis based on the 16S rRNA gene confirmed its affiliation to the genus *Williamsia*. However genomic properties also showed close similarities to *Gordonia*, as derived from GC content, COGs, and average nucleotide identities. Thus, an unequivocal delinearization based on the functional genomics level was not possible, which may be due to the underrepresentation of genomes from this genus. The genomic features of strain ARP1 would be consistent with a lifestyle within the phyllosphere, including putative adaptions to UV radiation, heat and cold shock, desiccation and oxidative stress. With this study, we provide novel genomic insights into the rarely sequenced genus *Williamsia* and discuss its putative adaptations to the phyllosphere habitat.

Additional files

Additional file 1: Identified essential genes in the *Williamsia* sp. ARP1 genome. (PDF 75 kb)

Additional file 2: Orthologous gene comparison of *Williamsia* sp. ARP1 and three other actinomycete genomes. (PDF 58 kb)

Additional file 3: Average nucleotide identities between *Williamsia* sp. ARP1 and nearest actinomycete genomes. (PDF 54 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HH designed the study, carried out the genome analysis, performed electron microscopy, phylogenetic analysis, and drafted the manuscript. AK carried out the sequencing and helped to draft the manuscript. UHi participated in the study design. PK performed laboratory experiments. MR conceived the study design and participated in its coordination. UHe conceived of the study, participated in its design, coordinated and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Eva E. Reisberg for original isolation of the strain, Christine Gernert and Srikkanth Balasubramanian for technical assistance, Wiebke Sickel for library preparation, the Department of Human Genetics for access to the MiSeq device, and Daniela Bunsen for help with the electron microscope (all University of Würzburg). This work was financially supported by the DFG Graduiertenkolleg GK1342 (TP A8).

Author details

[1]Department of Botany II, Julius-von-Sachs Institute for Biological Sciences, University of Würzburg, Julius-von-Sachs-Platz 3, D-97082 Würzburg, Germany. [2]GEOMAR Helmholtz Centre for Ocean Research, RD3 Marine Microbiology and Christian-Albrechts University of Kiel, Düsternbrooker Weg 20, D-24105 Kiel, Germany. [3]Department of Animal Ecology and Tropical Biology, Biocenter, University of Würzburg, Am Hubland D-97074, Germany. [4]Institut für Angewandte Mikrobiologie, Justus-Liebig-Universität Giessen, Heinrich-Buff-Ring 26, D-35392 Giessen, Germany.

Received: 16 April 2015 Accepted: 21 December 2015
Published online: 16 January 2016

References

1. Kämpfer P, Andersson MA, Rainey FA, Kroppenstedt RM, Salkinoja-Salonen M. *Williamsia muralis* gen. nov., sp. nov., isolated from the indoor environment of a children's day care centre. Int J Syst Evol Microbiol. 1999; 49:681–7. doi:10.1099/00207713-49-2-681.
2. Ludwig W, Euzéby J, Schumann P, Busse H-J, Trujillo M, Kämpfer P, et al. Bergey's manual of systematic bacteriology. Vol. 5, The actinobacteria. New York: Springer; 2012.
3. Stackebrandt E, Rainey FA, WardRainey NL. Proposal for a new hierarchic classification system, Actinobacteria classis nov. Int J Syst Bacteriol. 1997; 47(2):479–91. doi:10.1099/00207713-47-2-479.
4. Guerrero LD, Makhalanyane TP, Aislabie JM, Cowan DA. Draft genome sequence of *Williamsia* sp. strain D3, isolated from the Darwin Mountains, Antarctica. Genome Announc. 2014;2(1). doi:10.1128/genomeA.01230-13.
5. Yassin AF, Hupfer H. *Williamsia deligens* sp. nov., isolated from human blood. Int J Syst Evol Microbiol. 2006;56(Pt 1):193–7. doi:10.1099/ijs.0.63856-0.
6. Jones AL, Payne GD, Goodfellow M. *Williamsia faeni* sp. nov., an actinomycete isolated from a hay meadow. Int J Syst Evol Microbiol. 2010; 60(Pt 11):2548–51. doi:10.1099/ijs.0.015826-0.
7. Sazak A, Sahin N. *Williamsia limnetica* sp. nov., isolated from a limnetic lake sediment. Int J Syst Evol Microbiol. 2012;62(6):1414–8. doi:10.1099/ijs.0.032474-0.
8. Pathom-Aree W, Nogi Y, Sutcliffe IC, Ward AC, Horikoshi K, Bull AT, et al. *Williamsia marianensis* sp. nov., a novel actinomycete isolated from the Mariana Trench. Int J Syst Evol Microbiol. 2006;56(Pt 5):1123–6. doi:10.1099/ijs.0.64132-0.
9. Stach JE, Maldonado LA, Ward AC, Bull AT, Goodfellow M. *Williamsia maris* sp. nov., a novel actinomycete isolated from the Sea of Japan. Int J Syst Evol Microbiol. 2004;54(Pt 1):191–4. doi:10.1099/ijs.0.02767-0.
10. Kämpfer P, Wellner S, Lohse K, Lodders N, Martin K. *Williamsia phyllosphaerae* sp. nov., isolated from the surface of *Trifolium repens* leaves. Int J Syst Evol Microbiol. 2011;61(Pt 11):2702–5. doi:10.1099/ijs.0.029322-0.
11. Yassin AF, Young CC, Lai WA, Hupfer H, Arun AB, Shen FT, et al. *Williamsia serinedens* sp. nov., isolated from an oil-contaminated soil. Int J Syst Evol Microbiol. 2007;57(Pt 3):558–61. doi:10.1099/ijs.0.64691-0.
12. Fang XM, Su J, Wang H, Wei YZ, Zhang T, Zhao LL, et al. *Williamsia sterculiae* sp. nov., isolated from a Chinese medicinal plant. Int J Syst Evol Microbiol. 2013;63(Pt 11):4158–62. doi:10.1099/ijs.0.052688-0.
13. Andeer P, Stahl DA, Lillis L, Strand SE. Identification of microbial populations assimilating nitrogen from RDX in munitions contaminated military training range soils by high sensitivity stable isotope probing. Environ Sci Technol. 2013;47(18):10356–63. doi:10.1021/es401729c.
14. Kato H, Saito M, Nagahata Y, Katayama Y. Degradation of ambient carbonyl sulfide by *Mycobacterium* spp. in soil. Microbiology. 2008;154(Pt 1):249–55. doi:10.1099/mic.0.2007/011213-0.
15. Leigh MB, Prouzova P, Mackova M, Macek T, Nagle DP, Fletcher JS. Polychlorinated biphenyl (PCB)-degrading bacteria associated with trees in a PCB-contaminated site. Appl Environ Microbiol. 2006;72(4):2331–42. doi:10.1128/aem.72.4.2331-2342.2006.
16. del Mar Tomas M, Moure R, Saez Nieto JA, Fojon S, Fernandez A, Diaz M, et al. *Williamsia muralis* pulmonary infection. Emerg Infect Dis. 2005;11(8):1324–5. doi:10.3201/eid1108.050439.
17. Miteva VI, Sheridan PP, Brenchley JE. Phylogenetic and physiological diversity of microorganisms isolated from a deep greenland glacier ice core. Appl Environ Microbiol. 2004;70(1):202–13. doi:10.1128/AEM.70.1.202-213.2004.
18. Shade A, Hogan CS, Klimowicz AK, Linske M, McManus PS, Handelsman J. Culturing captures members of the soil rare biosphere. Environ Microbiol. 2012;14(9):2247–52. doi:10.1111/j.1462-2920.2012.02817.x.

19. Kaewkla O, Franco CM. Rational approaches to improving the isolation of endophytic actinobacteria from Australian native trees. Microb Ecol. 2013; 65(2):384–93. doi:10.1007/s00248-012-0113-z.
20. Stiefel P, Zambelli T, Vorholt JA. Isolation of optically targeted single bacteria by application of fluidic force microscopy to aerobic anoxygenic phototrophs from the phyllosphere. Appl Environ Microbiol. 2013;79(16):4895–905. doi:10.1128/AEM.01087-13.
21. Vorholt JA. Microbial life in the phyllosphere. Nat Rev Microbiol. 2012; 10(12):828–40. doi:10.1038/nrmicro2910.
22. Knief C, Delmotte N, Vorholt JA. Bacterial adaptation to life in association with plants–A proteomic perspective from culture to in situ conditions. Proteomics. 2011;11(15):3086–105. doi:10.1002/pmic.201000818.
23. Leveau JH, Lindow SE. Appetite of an epiphyte: quantitative monitoring of bacterial sugar consumption in the phyllosphere. Proc Natl Acad Sci U S A. 2001;98(6):3446–53. doi:10.1073/pnas.061629598.
24. Lindow SE, Brandl MT. Microbiology of the phyllosphere. Appl Env Microbiol. 2003;69(4):1875–83. doi:10.1128/AEM.69.4.1875-1883.2003.
25. Newton AC, Gravouil C, Fountaine JM. Managing the ecology of foliar pathogens: ecological tolerance in crops. Ann Appl Biol. 2010;157(3):343–59. doi:10.1111/j.1744-7348.2010.00437.x.
26. Marcell LM, Beattie GA. Effect of leaf surface waxes on leaf colonization by *Pantoea agglomerans* and *Clavibacter michiganensis*. Mol Plant Microbe Interact. 2002;15(12):1236–44. doi:10.1094/MPMI.2002.15.12.1236.
27. Castellani A, Chalmers AJ. Manual of tropical medicine. 1919.
28. Zhi XY, Li WJ, Stackebrandt E. An update of the structure and 16S rRNA gene sequence-based definition of higher ranks of the class Actinobacteria, with the proposal of two new suborders and four new families and emended descriptions of the existing higher taxa. Int J Syst Evol Microbiol. 2009;59(3):589–608. doi:10.1099/ijs.0.65780-0.
29. Goodfellow M, Isik K, Yates E. Actinomycete systematics: an unfinished synthesis. Nova Acta Leopold. 1999.
30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. doi:10.1016/S0022-2836(05)80360-2.
31. Pruesse E, Peplies J, Glockner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28(14):1823–9. doi:10.1093/bioinformatics/bts252.
32. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–21. doi:10.1093/sysbio/syq010.
33. Reisberg EE. Der Einfluss von Trichomen und kutikulären Lipiden auf die bakterielle Besiedelung von *Arabidopsis thaliana* Blättern. PhD thesis: University of Wuerzburg. 2013.
34. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008;26(5):541–7. doi:10.1038/nbt1360.
35. Reisberg EE, Hildebrandt U, Riederer M, Hentschel U. Phyllosphere bacterial communities of trichome-bearing and trichomeless *Arabidopsis thaliana* leaves. Antonie Van Leeuwenhoek. 2012;101(3):551–60. doi:10.1007/s10482-011-9669-8.
36. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. doi:10.1093/bioinformatics/btu170.
37. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011;6(3), e17288. doi:10.1371/journal.pone.0017288.
38. Coil D, Jospin G, Darling AE. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. Bioinformatics. 2014. doi:10.1093/bioinformatics/btu661.
39. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28(11):1420–8. doi:10.1093/bioinformatics/bts174.
40. Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29(21):2669–77. doi:10.1093/bioinformatics/btt476.
41. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comp Biol. 2012;19(5):455–77. doi:10.1089/cmb.2012.0021.
42. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics. 2010;Chapter 11:Unit 11 5. doi:10.1002/0471250953.bi1105s31.
43. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5. doi:10.1093/bioinformatics/btt086.
44. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013;14(5): R47. doi:10.1186/gb-2013-14-5-r47.
45. Clark SC, Egan R, Frazier PI, Wang Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics. 2013;29(4):435–43. doi:10.1093/bioinformatics/bts723.
46. Vezzi F, Narzisi G, Mishra B. Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. PLoS One. 2012;7(12), e52210. doi:10.1371/journal.pone.0052210.
47. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6): e11147. doi:10.1371/journal.pone.0011147.
48. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119. doi:10.1186/1471-2105-11-119.
49. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, et al. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 2007;35(Database issue):D237–40. doi:10.1093/nar/gkl951.
50. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9): 1236–40. doi:10.1093/bioinformatics/btu031.
51. Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics. 2009;25(17):2271–8. doi:10.1093/bioinformatics/btp393.
52. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75. doi:10.1186/1471-2164-9-75.
53. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014;42(Database issue):D206–14. doi:10.1093/nar/gkt1226.
54. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5): 955–64. doi:10.1093/nar/25.5.0955.
55. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100–8. doi:10.1093/nar/gkm160.
56. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5. doi:10.1093/bioinformatics/btt509.
57. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9. doi:10.1093/bioinformatics/btu153.
58. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998;6:175–82.
59. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785–6. doi:10.1038/nmeth.1701.
60. Edgar RC. PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics. 2007;8:18. doi:10.1186/1471-2105-8-18.
61. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, et al. antiSMASH 2.0–a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 2013;41(Web Server issue):W204–12. doi:10.1093/nar/gkt449.
62. 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. doi:10.1038/nbt.2579.
63. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755–63. doi:10.1093/bioinformatics/14.9.755.
64. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 2001;314(5): 1041–52. doi:10.1006/jmbi.2000.5197.

65. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106(45):19126–31. doi:10.1073/pnas.0906412106.
66. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007;57(Pt 1):81–91. doi:10.1099/ijs.0.64483-0.
67. Gao Q, Garcia-Pichel F. An ATP-grasp ligase involved in the last biosynthetic step of the iminomycosporine shinorine in Nostoc punctiforme ATCC 29133. J Bacteriol. 2011;193(21):5923–8. doi:10.1128/JB.05730-11.
68. Gao Q, Garcia-Pichel F. Microbial ultraviolet sunscreens. Nat Rev Microbiol. 2011;9(11):791–802. doi:10.1038/nrmicro2649.
69. Jacobs JL, Carroll TL, Sundin GW. The role of pigmentation, ultraviolet radiation tolerance, and leaf colonization strategies in the epiphytic survival of phyllosphere bacteria. Microb Ecol. 2005;49(1):104–13. doi:10.1007/s00248-003-1061-4.
70. Liu X, Williams CE, Nemacheck JA, Wang H, Subramanyam S, Zheng C, et al. Reactive oxygen species are involved in plant defense against a gall midge. Plant Physiol. 2010;152(2):985–99. doi:10.1104/pp.109.150656.
71. Hammond-Kosack KE, Jones JD. Resistance gene-dependent plant defense responses. Plant Cell. 1996;8(10):1773–91. doi:10.1105/tpc.8.10.1773.
72. Brown AD. Microbial water stress. Bacteriol Rev. 1976;40(4):803–46.
73. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990;87(12):4576–9. doi:10.1073/pnas.87.12.4576.
74. Garrity GM, Holt JG. The road map to the manual. In: Boone D, Castenholz R, Garrity G, editors. Bergey's manual® of systematic bacteriology. New York: Springer; 2001. p. 119–66.
75. Buchanan RE. Studies in the nomenclature and classification of the bacteria: II. The primary subdivisions of the schizomycetes. J Bacteriol. 1917;2(2):155–64.
76. Skerman VBD, Mcgowan V, Sneath PHA. Approved lists of bacterial names. Int J Syst Bacteriol. 1980;30(1):225–420.
77. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9. doi:10.1038/75556.

Chapter 3.

Mining Genomes of Three Marine Sponge-Associated Actinobacterial Isolates for Secondary metabolism

Authors: **Horn, H.**, Hentschel, U. and U.R. Abdelmohsen
Published: 2015
Journal: Genome Announcements
Volume: 3,5
DOI: dx.doi.org/10.1128/genomeA.01106-15

A reprint of this article is used here with the permission from the authors and American Society for Microbiology (ASM). The documentation of the author's contributions is found in Part VII. Appendix. This article is open access and published under the terms of the Creative Commons Attribution Unported license 3.0 (CC BY 3.0, http://http://creativecommons.org/licenses/by/3.0/).

genomeA
Journals.ASM.org

Mining Genomes of Three Marine Sponge-Associated Actinobacterial Isolates for Secondary Metabolism

Hannes Horn,[a,b] Ute Hentschel,[a,b] Usama Ramadan Abdelmohsen[a*]

Department of Botany II, Julius-von-Sachs Institute for Biological Sciences, University of Würzburg, Würzburg, Germany[a]; Department of Marine Microbiology, GEOMAR Helmholtz Centre for Ocean Research, RD3 Marine Microbiology and Christian-Albrechts University of Kiel, Kiel, Germany[b]

* Permanent address: Usama Ramadan Abdelmohsen, Department of Pharmacognosy, Faculty of Pharmacy, Minia University, Minia, Egypt.

Here, we report the draft genome sequences of three actinobacterial isolates, *Micromonospora* sp. RV43, *Rubrobacter* sp. RV113, and *Nocardiopsis* sp. RV163 that had previously been isolated from Mediterranean sponges. The draft genomes were analyzed for the presence of gene clusters indicative of secondary metabolism using antiSMASH 3.0 and NapDos pipelines. Our findings demonstrated the chemical richness of sponge-associated actinomycetes and the efficacy of genome mining in exploring the genomic potential of sponge-derived actinomycetes.

Received 14 August 2015 **Accepted** 17 August 2015 **Published** 1 October 2015

Citation Horn H, Hentschel U, Abdelmohsen UR. 2015. Mining genomes of three marine sponge-associated actinobacterial isolates for secondary metabolism. Genome Announc 3(5):e01106-15. doi:10.1128/genomeA.01106-15.

Address correspondence to Usama Ramadan Abdelmohsen, usama.ramadan@uni-wuerzburg.de.

Actinomycetes are known for their unprecedented ability to produce novel lead compounds of clinical and pharmaceutical importance (1–4). Among the many actinobacterial genera, *Streptomyces*, *Micromonospora*, *Nocardiopsis*, and *Rhodococcus* are the most prolific producers of secondary metabolites, which display broad chemical diversity and diverse pharmaceutically and medically relevant bioactivities (5–8). Recent genomic sequencing data have revealed the presence of a plethora of putative biosynthetic gene clusters on the genomes of actinomycetes encoding for secondary metabolites that are not observed under standard fermentation conditions (9–13). In the present study, draft genomes of three actinobacterial isolates, *Micromonospora* sp. RV43, *Rubrobacter* sp. RV113, and *Nocardiopsis* sp. RV163 that had previously been cultivated from the Mediterranean sponges *Aplysina aerophoba* (RV43 and RV113) and *Dysidea avara* (RV163) (14), were established.

The genomic DNA of the isolates was extracted from 5-day-old ISP2 cultures. Paired-end, 2 × 250-bp libraries were prepared with the Nextera XT kit (Illumina, Inc.). Sequencing was performed on an Illumina MiSeq device. A total of 5,900,702 raw reads were produced for *Micromonospora* sp. RV43, 2,206,732 raw reads for *Rubrobacter* sp. RV113 and 4,851,980 raw reads were delivered for *Nocardiopsis* sp. RV163. Reads were adapter clipped, quality trimmed and length filtered (15). Initial contigs were generated using SPAdes (16) and only contigs ≥1000 bp were maintained. A further clean-up of contigs was performed using G+C-content, coverage, and taxonomic assignments (17). For *ab initio* gene prediction, prodigal was applied (18) and functional annotation of the predicted protein sequences was performed with the RAST webserver (19). Secondary metabolite gene clusters and possible encoded compounds were predicted with antiSMASH (20) and NapDos (21).

A number of 101 (RV43), 33 (RV113), and 82 (RV163) secondary metabolite gene clusters were detected with antiSMASH. For strain RV43, 5 terpene clusters, 4 type 1 PKS clusters, 2 lantipeptides, 1 type 2 PKS cluster, 1 siderophore, 1 NRPS cluster, and 1 bacteriocin were found. For strain RV113, 3 terpene clusters, 1 fatty acid, and 1 mixed type 3 PKS-fatty acid cluster were found. The draft genome sequence of strain RV163 showed homologies to 7 NRPS clusters, 4 terpene gene clusters, 2 type 1 PKS clusters, 2 ectoines, 2 bacteriocins, 1 phenanzine, 1 butyrolacetone, 1 type 2 PKS, and 1 siderophore.

For *Micromonospora* sp. RV43, NaPDoS predicted the presence of gene clusters encoding for compounds such as leinamycin, kirromycin, aclacinomycin, and tetronomycin. For *Nocardiopsis* sp. RV163, compounds such as alnumycin, avermectin, and neocarzinostatin were predicted. For *Rubrobacter* sp. RV113, only gene clusters encoding for fatty acids synthesis were found. These results highlight the genomic potential of at least two of three inspected isolates for natural products discovery.

Nucleotide sequence accession numbers. This whole-genome shotgun project was deposited in DDBJ/ENA/GenBank under the accession numbers LEKG00000000, LEKH00000000, and LEKI00000000. The versions described in this paper are the first versions LEKG01000000, LEKG01000000, and LEKH01000000.

ACKNOWLEDGMENTS

We gratefully acknowledge C. Gernert for technical assistance in the laboratory. We thank the Group of Molecular Biodiversity/Zoology 3 for library preparation and the Department of Human Biology for access to the sequencing device (both University of Würzburg).

Financial support was provided by the DFG (SFB 630 TP A5) to U.H.

REFERENCES

1. **Abdelmohsen UR, Bayer K, Hentschel U.** 2014. Diversity, abundance and natural products of marine sponge-associated actinomycetes. Nat Prod Rep **31:**381–399. http://dx.doi.org/10.1039/C3NP70111E.
2. **Li JW, Vederas JC.** 2009. Drug discovery and natural products: end of an era or an endless frontier? Science **325:**161–165. http://dx.doi.org/10.1126/science.1168243.

3. **Eltamany EE, Abdelmohsen UR, Ibrahim AK, Hassanean HA, Hentschel U, Ahmed SA.** 2014. New antibacterial xanthone from the marine sponge-derived *Micrococcus* sp. EG45. Bioorg Med Chem Lett **24:** 4939–4942. http://dx.doi.org/10.1016/j.bmcl.2014.09.040.
4. **Grkovic T, Abdelmohsen UR, Othman EM, Stopper H, Edrada-Ebel R, Hentschel U, Quinn RJ.** 2014. Two new antioxidant actinosporin analogues from the calcium alginate beads culture of sponge-associated *Actinokineospora* sp. strain EG49. Bioorg Med Chem Lett **24:**5089–5092. http://dx.doi.org/10.1016/j.bmcl.2014.08.068.
5. **Reimer A, Blohm A, Quack T, Grevelding CG, Kozjak-Pavlovic V, Rudel T, Hentschel U, Abdelmohsen UR.** 20 May 2015. Inhibitory activities of the marine streptomycete-derived compound SF2446A2 against *Chlamydia trachomatis* and *Schistosoma mansoni*. J Antibiot. http://dx.doi.org/10.1038/ja.2015.54.
6. **Dashti Y, Grkovic T, Abdelmohsen UR, Hentschel U, Quinn RJ.** 2014. Production of induced secondary metabolites by a co-culture of sponge-associated actinomycetes, *Actinokineospora* sp. EG49 and *Nocardiopsis* sp. RV163. Mar Drugs **12:**3046–3059. http://dx.doi.org/10.3390/md12053046.
7. **Abdelmohsen UR, Yang C, Horn H, Hajjar D, Ravasi T, Hentschel U.** 2014. Actinomycetes from Red Sea sponges: sources for chemical and phylogenetic diversity. Mar Drugs **12:**2771–2789. http://dx.doi.org/10.3390/md12052771.
8. **Abdelmohsen UR, Szesny M, Othman EM, Schirmeister T, Grond S, Stopper H, Hentschel U.** 2012. Antioxidant and anti-protease activities of diazepinomicin from the sponge-associated *Micromonospora* strain RV115. Mar Drugs **10:**2208–2221. http://dx.doi.org/10.3390/md10102208.
9. **Harjes J, Ryu T, Abdelmohsen UR, Moitinho-Silva L, Horn H, Ravasi T, Hentschel U.** 2014. Draft genome sequence of the antitrypanosomally active sponge-associated bacterium *Actinokineospora* sp. strain EG49. Genome Announc 2(2):e00160-14. http://dx.doi.org/10.1128/genomeA.00160-14.
10. **Abdelmohsen UR, Grkovic T, Balasubramanian S, Kamel MS, Quinn RJ, Hentschel U.** 2015. Elicitation of secondary metabolism in actinomycetes. Biotechnol Adv **33:**798–811. http://dx.doi.org/10.1016/j.biotechadv.2015.06.003.
11. **Challis GL.** 2008. Mining microbial genomes for new natural products and biosynthetic pathways. Microbiology **154:**1555–1569. http://dx.doi.org/10.1099/mic.0.2008/018523-0.
12. **Jensen PR, Chavarria KL, Fenical W, Moore BS, Ziemert N.** 2014. Challenges and triumphs to genomics-based natural product discovery. J Ind Microbiol Biotechnol **41:**203–209. http://dx.doi.org/10.1007/s10295-013-1353-8.
13. **Nett M, Ikeda H, Moore BS.** 2009. Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat Prod Rep **26:**1362–1384. http://dx.doi.org/10.1039/b817069j.
14. **Abdelmohsen UR, Pimentel-Elardo SM, Hanora A, Radwan M, Abou-El-Ela SH, Ahmed S, Hentschel U.** 2010. Isolation, phylogenetic analysis and anti-infective activity screening of marine sponge-associated actinomycetes. Mar Drugs **8:**399–412. http://dx.doi.org/10.3390/md8030399.
15. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:**2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.
16. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA.** 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19:**455–477. http://dx.doi.org/10.1089/cmb.2012.0021.
17. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.** 2009. BLAST+: architecture and applications. BMC Bioinformatics **10:**421. http://dx.doi.org/10.1186/1471-2105-10-421.
18. **Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11:**119. http://dx.doi.org/10.1186/1471-2105-11-119.
19. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O.** 2008. The RAST server: Rapid Annotations using Subsystems Technology. BMC Genomics **9:**75. http://dx.doi.org/10.1186/1471-2164-9-75.
20. **Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH.** 2015. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res **43:** W237–W243. http://dx.doi.org/10.1093/nar/gkv437.
21. **Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR.** 2012. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One **7:**e34064. http://dx.doi.org/10.1371/journal.pone.0034064.

Chapter 4.

Draft genome sequences of three chemically rich actinomycetes isolated from Mediterranean sponges

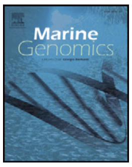
Authors: **Horn, H.**, C. Cheng, Edrada–Ebal, R., Hentschel, U. and U.R. Abdelmohsen
Published: 2015
Journal: Marine Genomics
Volume: 24,3
DOI: dx.doi.org/10.1016/j.margen.2015.10.003

Marine Genomics 24 (2015) 285–287

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen



Genomics/technical resources

Draft genome sequences of three chemically rich actinomycetes isolated from Mediterranean sponges

Hannes Horn [a,1], Cheng Cheng [a], RuAngelie Edrada-Ebel [b], Ute Hentschel [a,1], Usama Ramadan Abdelmohsen [a,*,2]

[a] Department of Marine Microbiology, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany
[b] Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, The John Arbuthnott Building, Glasgow, UK

ARTICLE INFO

Article history:
Received 10 July 2015
Received in revised form 8 October 2015
Accepted 8 October 2015
Available online 20 October 2015

ABSTRACT

Metabolomic analysis has shown the chemical richness of the sponge-associated actinomycetes *Streptomyces* sp. SBT349, *Nonomureae* sp. SBT364, and *Nocardiopsis* sp. SBT366. The genomes of these actinomycetes were sequenced and the genomic potential for secondary metabolism was evaluated. Their draft genomes have sizes of 8.0, 10, and 5.8 Mb having 687, 367, and 179 contigs with a GC content of 71.6, 70.7, and 72.7%, respectively. Moreover, antiSMASH 3.0 predicted 108, 149, and 75 secondary metabolite gene clusters, respectively which highlight the metabolic capacity of the three actinomycete species to produce diverse classes of natural products.

1. Short introduction

Actinomycetes harbor a wealth of natural products with structural complexity and diverse biological activities (Abdelmohsen et al., 2014a; Nett et al., 2009; Li and Vederas, 2009; Abdelmohsen et al., 2015). Genomic sequence data have revealed the presence of putatively silent biosynthetic gene clusters in the genomes of actinomycetes that encode for secondary metabolites, which are not seen under standard fermentation conditions (Cimermancic et al., 2014). The actinomycete isolates *Streptomyces* sp. SBT349, *Nonomureae* sp. SBT364, and *Norcardiopsis* sp. SBT366 were cultivated from marine sponges *Sarcotragus spinosulus*, *Sarcotragus foetidus*, and *Chondrilla nucula*, respectively. The strains have been deposited in the German Collection of Microorganisms and Cell Cultures (DSMZ) with accession numbers DSM 100667 (SBT349), DSM 100666 (SBT364), and DSM 100668 (SBT366). The sponges were collected by SCUBA diving at 5–7 m depth from offshore Pollonia, Milos, Greece (N36.76612°; E24.51530°) in May 2013 under the umbrella of the EU-FP7 project entitled "SeaBioTech: From sea-bed to test-bed: harvesting the potential of marine biodiversity for industrial biotechnology" that aims to create innovative marine biodiscovery pipelines. Members of the genera *Streptomyces and Nocardiopsis* are widespread in terrestrial environments, including soil and plants and have also been isolated from the marine environment, i.e., from marine sponges (Abdelmohsen et al., 2010; Vicente et al., 2013; Abdelmohsen et al., 2014b; Eltamany et al., 2014). We report here, to our knowledge for the first time, the isolation of members of the genus *Nonomureae* from marine environment. Among the 50 actinomycetes cultivated from the Milos collection, the organic extracts of isolates SBT349, SBT364, and SBT366 exhibited rich HPLC-peak profiles as well as diverse bioactivities including antioxidant, antitrypanosomal and anticancer, respectively (Cheng et al., 2015). These isolates were selected based on their HPLC-peak richness and bioactivity profile for further genomic sequencing.

2. Data description

Genomic DNA of the actinomycetes was extracted and prepared as described (Harjes et al., 2014). 250 bp paired-end sequencing was performed on a MiSeq benchtop sequencer (Illumina). Obtained reads were adapter trimmed as well as quality and length filtered using Trimmomatic 0.32 (Bolger et al., 2014). Assembly was performed using SPAdes 3.1.1 (Bankevich et al., 2012) and calculated contigs were manually filtered due to low coverage. Remaining contigs were extended and merged wherever possible using SSPACE 3.0. (Boetzer et al., 2011). The RAST webserver was used for annotation (Aziz et al., 2008) (Table 1).

The draft genomes were mined using antiSMASH 3.0 ("Antibiotic and Secondary Metabolites Analysis Shell") (Weber et al., 2015) and NapDos ("The natural product domain seeker") (Ziemert et al., 2012). Among the three genomes sequenced, *Streptomyces* sp. SBT349 displayed the most diverse antiSMASH read-out. A total of 108 potential secondary metabolite gene clusters were predicted, encoding for 23 type I polyketide synthases (PKS), 11 non-ribosomal peptide synthetases (NRPSs), 2 terpenes, 21 saccharides, 3 siderophores, 3 lantipeptides, 1 butyrolactone, 1 bacteriocin, 1 phenazine, 1 ladderane,

* Corresponding author.
E-mail address: usama.ramadan@uni-wuerzburg.de (U.R. Abdelmohsen).
[1] Department of Botany II, Julius-von-Sachs Institute for Biological Sciences, University of Würzburg, Germany.
[2] Permanent address: Department of Pharmacognosy, Faculty of Pharmacy, Minia University, Minia 61,519, Egypt.

http://dx.doi.org/10.1016/j.margen.2015.10.003
1874-7787/

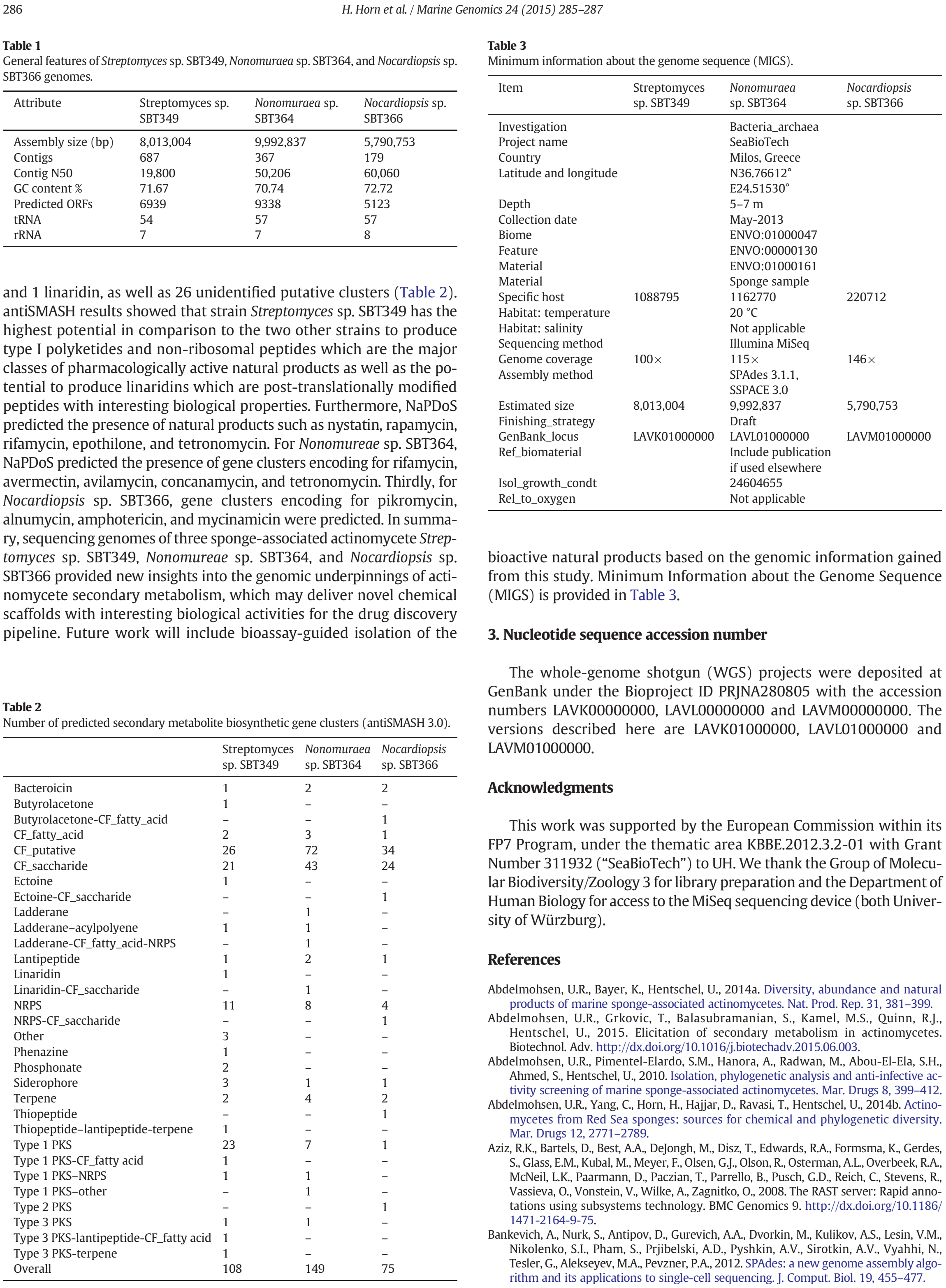


Table 1
General features of *Streptomyces* sp. SBT349, *Nonomuraea* sp. SBT364, and *Nocardiopsis* sp. SBT366 genomes.

Attribute	Streptomyces sp. SBT349	*Nonomuraea* sp. SBT364	*Nocardiopsis* sp. SBT366
Assembly size (bp)	8,013,004	9,992,837	5,790,753
Contigs	687	367	179
Contig N50	19,800	50,206	60,060
GC content %	71.67	70.74	72.72
Predicted ORFs	6939	9338	5123
tRNA	54	57	57
rRNA	7	7	8

Table 3
Minimum information about the genome sequence (MIGS).

Item	Streptomyces sp. SBT349	*Nonomuraea* sp. SBT364	*Nocardiopsis* sp. SBT366
Investigation		Bacteria_archaea	
Project name		SeaBioTech	
Country		Milos, Greece	
Latitude and longitude		N36.76612° E24.51530°	
Depth		5–7 m	
Collection date		May-2013	
Biome		ENVO:01000047	
Feature		ENVO:00000130	
Material		ENVO:01000161	
Material		Sponge sample	
Specific host	1088795	1162770	220712
Habitat: temperature		20 °C	
Habitat: salinity		Not applicable	
Sequencing method		Illumina MiSeq	
Genome coverage	100×	115×	146×
Assembly method		SPAdes 3.1.1, SSPACE 3.0	
Estimated size	8,013,004	9,992,837	5,790,753
Finishing_strategy		Draft	
GenBank_locus	LAVK01000000	LAVL01000000	LAVM01000000
Ref_biomaterial		Include publication if used elsewhere	
Isol_growth_condt		24604655	
Rel_to_oxygen		Not applicable	

and 1 linaridin, as well as 26 unidentified putative clusters (Table 2). antiSMASH results showed that strain *Streptomyces* sp. SBT349 has the highest potential in comparison to the two other strains to produce type I polyketides and non-ribosomal peptides which are the major classes of pharmacologically active natural products as well as the potential to produce linaridins which are post-translationally modified peptides with interesting biological properties. Furthermore, NaPDoS predicted the presence of natural products such as nystatin, rapamycin, rifamycin, epothilone, and tetronomycin. For *Nonomureae* sp. SBT364, NaPDoS predicted the presence of gene clusters encoding for rifamycin, avermectin, avilamycin, concanamycin, and tetronomycin. Thirdly, for *Nocardiopsis* sp. SBT366, gene clusters encoding for pikromycin, alnumycin, amphotericin, and mycinamicin were predicted. In summary, sequencing genomes of three sponge-associated actinomycete *Streptomyces* sp. SBT349, *Nonomureae* sp. SBT364, and *Nocardiopsis* sp. SBT366 provided new insights into the genomic underpinnings of actinomycete secondary metabolism, which may deliver novel chemical scaffolds with interesting biological activities for the drug discovery pipeline. Future work will include bioassay-guided isolation of the bioactive natural products based on the genomic information gained from this study. Minimum Information about the Genome Sequence (MIGS) is provided in Table 3.

Table 2
Number of predicted secondary metabolite biosynthetic gene clusters (antiSMASH 3.0).

	Streptomyces sp. SBT349	*Nonomuraea* sp. SBT364	*Nocardiopsis* sp. SBT366
Bacteroicin	1	2	2
Butyrolacetone	1	–	–
Butyrolacetone-CF_fatty_acid	–	–	1
CF_fatty_acid	2	3	1
CF_putative	26	72	34
CF_saccharide	21	43	24
Ectoine	1	–	–
Ectoine-CF_saccharide	–	–	1
Ladderane	–	1	–
Ladderane–acylpolyene	1	1	–
Ladderane-CF_fatty_acid-NRPS	–	1	–
Lantipeptide	1	2	1
Linaridin	1	–	–
Linaridin-CF_saccharide	–	1	–
NRPS	11	8	4
NRPS-CF_saccharide	–	–	1
Other	3	–	–
Phenazine	1	–	–
Phosphonate	2	–	–
Siderophore	3	1	1
Terpene	2	4	2
Thiopeptide	–	–	1
Thiopeptide–lantipeptide-terpene	1	–	–
Type 1 PKS	23	7	1
Type 1 PKS-CF_fatty acid	1	–	–
Type 1 PKS–NRPS	1	1	–
Type 1 PKS–other	–	1	–
Type 2 PKS	–	–	1
Type 3 PKS	1	1	–
Type 3 PKS-lantipeptide-CF_fatty acid	1	–	–
Type 3 PKS-terpene	1	–	–
Overall	108	149	75

3. Nucleotide sequence accession number

The whole-genome shotgun (WGS) projects were deposited at GenBank under the Bioproject ID PRJNA280805 with the accession numbers LAVK00000000, LAVL00000000 and LAVM00000000. The versions described here are LAVK01000000, LAVL01000000 and LAVM01000000.

Acknowledgments

This work was supported by the European Commission within its FP7 Program, under the thematic area KBBE.2012.3.2-01 with Grant Number 311932 (“SeaBioTech”) to UH. We thank the Group of Molecular Biodiversity/Zoology 3 for library preparation and the Department of Human Biology for access to the MiSeq sequencing device (both University of Würzburg).

References

Abdelmohsen, U.R., Bayer, K., Hentschel, U., 2014a. Diversity, abundance and natural products of marine sponge-associated actinomycetes. Nat. Prod. Rep. 31, 381–399.

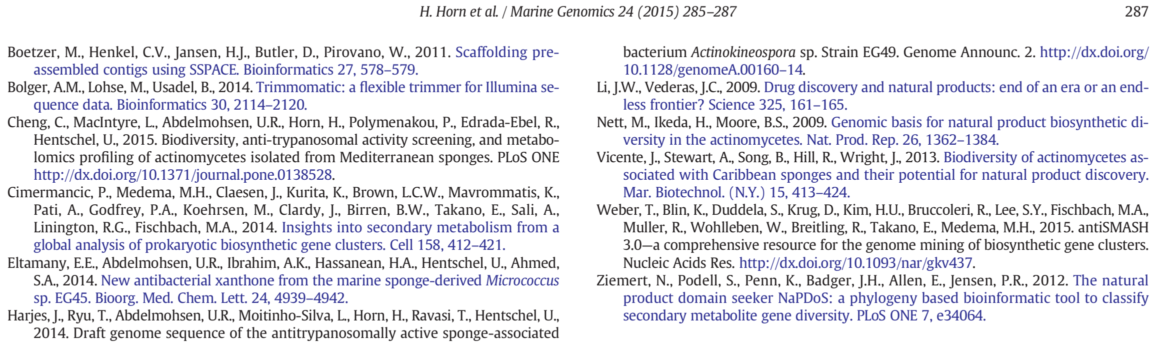
Abdelmohsen, U.R., Grkovic, T., Balasubramanian, S., Kamel, M.S., Quinn, R.J., Hentschel, U., 2015. Elicitation of secondary metabolism in actinomycetes. Biotechnol. Adv. http://dx.doi.org/10.1016/j.biotechadv.2015.06.003.

Abdelmohsen, U.R., Pimentel-Elardo, S.M., Hanora, A., Radwan, M., Abou-El-Ela, S.H., Ahmed, S., Hentschel, U., 2010. Isolation, phylogenetic analysis and anti-infective activity screening of marine sponge-associated actinomycetes. Mar. Drugs 8, 399–412.

Abdelmohsen, U.R., Yang, C., Horn, H., Hajjar, D., Ravasi, T., Hentschel, U., 2014b. Actinomycetes from Red Sea sponges: sources for chemical and phylogenetic diversity. Mar. Drugs 12, 2771–2789.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O., 2008. The RAST server: Rapid annotations using subsystems technology. BMC Genomics 9. http://dx.doi.org/10.1186/1471-2164-9-75.

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



Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., Pirovano, W., 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579.

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

Cheng, C., MacIntyre, L., Abdelmohsen, U.R., Horn, H., Polymenakou, P., Edrada-Ebel, R., Hentschel, U., 2015. Biodiversity, anti-trypanosomal activity screening, and metabolomics profiling of actinomycetes isolated from Mediterranean sponges. PLoS ONE http://dx.doi.org/10.1371/journal.pone.0138528.

Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Brown, L.C.W., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J., Birren, B.W., Takano, E., Sali, A., Linington, R.G., Fischbach, M.A., 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158, 412–421.

Eltamany, E.E., Abdelmohsen, U.R., Ibrahim, A.K., Hassanean, H.A., Hentschel, U., Ahmed, S.A., 2014. New antibacterial xanthone from the marine sponge-derived *Micrococcus* sp. EG45. Bioorg. Med. Chem. Lett. 24, 4939–4942.

Harjes, J., Ryu, T., Abdelmohsen, U.R., Moitinho-Silva, L., Horn, H., Ravasi, T., Hentschel, U., 2014. Draft genome sequence of the antitrypanosomally active sponge-associated bacterium *Actinokineospora* sp. Strain EG49. Genome Announc. 2. http://dx.doi.org/10.1128/genomeA.00160–14.

Li, J.W., Vederas, J.C., 2009. Drug discovery and natural products: end of an era or an endless frontier? Science 325, 161–165.

Nett, M., Ikeda, H., Moore, B.S., 2009. Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat. Prod. Rep. 26, 1362–1384.

Vicente, J., Stewart, A., Song, B., Hill, R., Wright, J., 2013. Biodiversity of actinomycetes associated with Caribbean sponges and their potential for natural product discovery. Mar. Biotechnol. (N.Y.) 15, 413–424.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., Lee, S.Y., Fischbach, M.A., Muller, R., Wohlleben, W., Breitling, R., Takano, E., Medema, M.H., 2015. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. http://dx.doi.org/10.1093/nar/gkv437.

Ziemert, N., Podell, S., Penn, K., Badger, J.H., Allen, E., Jensen, P.R., 2012. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS ONE 7, e34064.

Chapter 5.

An enrichment of CRISPR and other defense–related features in marine sponge–associated microbial metagenomes

Authors: **Horn, H.**, Slaby, B.M., Jahn, M.T., Bayer, K., Moitinho–Silva, L., Förster, F., Abdelmohsen, U.R., and U. Hentschel

Published: 2016
Journal: Frontiers in Microbiology
Volume: 7,1751
DOI: dx.doi.org/10.3389/fmicb.2016.01751

frontiers in Microbiology

ORIGINAL RESEARCH
published: 08 November 2016
doi: 10.3389/fmicb.2016.01751

An Enrichment of CRISPR and Other Defense-Related Features in Marine Sponge-Associated Microbial Metagenomes

Hannes Horn[1,2], Beate M. Slaby[1,2], Martin T. Jahn[1,2], Kristina Bayer[1], Lucas Moitinho-Silva[3], Frank Förster[4,5], Usama R. Abdelmohsen[2,6] and Ute Hentschel[1,7]*

[1] RD3 Marine Microbiology, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany, [2] Department of Botany II, Julius-von-Sachs Institute for Biological Sciences, University of Würzburg, Würzburg, Germany, [3] School of Biological, Earth and Environmental Sciences, Centre for Marine Bio-Innovation and School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia, [4] Department of Bioinformatics, University of Würzburg, Würzburg, Germany, [5] Center for Computational and Theoretical Biology, University of Würzburg, Würzburg, Germany, [6] Department of Pharmacognosy, Faculty of Pharmacy, Minia University, Minia, Egypt, [7] Christian-Albrechts-Universität zu Kiel, Kiel, Germany

OPEN ACCESS

Edited by:
*Joerg Graf,
University of Connecticut, USA*

Reviewed by:
*Mark J. Mandel,
Northwestern University, USA
Kimberly B. Ritchie,
Mote Marine Laboratory, USA*

***Correspondence:**
*Ute Hentschel
uhentschel@geomar.de*

Specialty section:
This article was submitted to Microbial Symbioses, a section of the journal Frontiers in Microbiology

Received: *28 July 2016*
Accepted: *19 October 2016*
Published: *08 November 2016*

Citation:
Horn H, Slaby BM, Jahn MT, Bayer K, Moitinho-Silva L, Förster F, Abdelmohsen UR and Hentschel U (2016) An Enrichment of CRISPR and Other Defense-Related Features in Marine Sponge-Associated Microbial Metagenomes. Front. Microbiol. 7:1751. doi: 10.3389/fmicb.2016.01751

Many marine sponges are populated by dense and taxonomically diverse microbial consortia. We employed a metagenomics approach to unravel the differences in the functional gene repertoire among three Mediterranean sponge species, *Petrosia ficiformis*, *Sarcotragus foetidus, Aplysina aerophoba* and seawater. Different signatures were observed between sponge and seawater metagenomes with regard to microbial community composition, GC content, and estimated bacterial genome size. Our analysis showed further a pronounced repertoire for defense systems in sponge metagenomes. Specifically, clustered regularly interspaced short palindromic repeats, restriction modification, DNA phosphorothioation and phage growth limitation systems were enriched in sponge metagenomes. These data suggest that defense is an important functional trait for an existence within sponges that requires mechanisms to defend against foreign DNA from microorganisms and viruses. This study contributes to an understanding of the evolutionary arms race between viruses/phages and bacterial genomes and it sheds light on the bacterial defenses that have evolved in the context of the sponge holobiont.

Keywords: metagenomes, defense, CRISPR, restriction modification, sponge microbiome, seawater

INTRODUCTION

Marine sponges (Porifera) represent the oldest metazoan phylum with a fossil record dating back 580 million years in time (Li et al., 1998). Many sponges host dense and diverse communities of unicellular microorganisms within their tissues (Taylor et al., 2007; Hentschel et al., 2012; Thomas et al., 2016). Based on 16S rRNA gene amplicon sequencing, a recent study observed 1000s of symbiont lineages [operational taxonomic units (OTUs)] within sponges, which are dominated by *Proteobacteria* (mostly *Alpha-* and *Gammaproteobacteria*), *Acidobacteria*, *Actinobacteria*,

Abbreviations: CRISPR, clustered regularly interspaced short palindromic repeats; MTase, methyltransferase; nt, NCBI nucleotide database; nr, NCBI non-redundant protein database; REase, restriction endonuclease; RMS, restriction modification system.

Chloroflexi, *Cyanobacteria*, *Crenarchaeota*, as well as symbionts of several candidate phyla. Representatives of 41 different phyla were thus far recovered from sponges with representatives of 13 phyla being shared among all sponge hosts (Thomas et al., 2016). Sponges are ecologically important in benthic environments (Bell, 2008). The sponge-associated microorganisms carry out functions related to nutrient cycling including carbon, nitrogen, and possibly sulfur and vitamin metabolism (Taylor et al., 2007; Bayer et al., 2008; Hentschel et al., 2012) as well as to secondary metabolism and chemical defense (Wilson et al., 2014). As sessile filter feeders, sponges are capable of pumping seawater at rates up to 1000s of liters per kilogram of sponge per day (Vogel, 1977; Weisz et al., 2008). Small particles are retained from the incoming seawater and transferred into the mesohyl interior where they are digested by phagocytosis (Bell, 2008; Southwell et al., 2008; Maldonado et al., 2012). Sponges and their microbial consortia (hereafter referred to as the sponge holobiont) are thus continuously exposed to incoming microorganisms, that serve as a food source, but that may also be harmful (Webster, 2007; Wehrl et al., 2007). Despite considerable research effort and several published sponge genomes (Srivastava et al., 2010; Ryu et al., 2016), little is known as to how the sponge holobiont protects itself against potentially harmful microorganisms, whether eukaryotic, prokaryotic, or viral in nature.

One major line of prokaryotic defense is based on the self – non-self-discrimination principle, which recognizes and targets foreign DNA (Makarova et al., 2013). It comprises various systems, among them the clustered regularly interspaced short palindromic repeats (CRISPR). CRISPRs are based on conserved repeats and variable spacer sequences which are incorporated into the host genomes upon encounters from viruses or phages and plasmids and are thus able to memorize the attack (Horvath and Barrangou, 2010). Hence, it is described as the adaptive immune system of prokaryotes (Makarova et al., 2013). Structurally, CRISPRs are associated with *cas* genes, which are essential for their function and which are also used for the CRISPR classification (Makarova et al., 2011). Additional defense systems are the RMS and the DNA phosphothiolation (DND) system (Makarova et al., 2013). The RMS is nearly ubiquitous among bacteria (Vasu and Nagaraja, 2013). RMS can be classified into types I–IV depending on their subunits, recognition sites, cleavage positions, and substrate specificities (Roberts et al., 2003). Both, the RMS and DMD systems, make use of labeling own DNA, either by methylation or by phosphorothioation, and recognize and destroy unmodified non-self DNA (Wang et al., 2007; Vasu and Nagaraja, 2013). The Phage growth limitation (Pgl) system is another line of defense that allows phage burst upon initial infection. In *Streptomyces coelicolor* A(3)2, PgI was shown to target phage Φ31 and its relatives. Here, the DNA of the phage progeny was methylated, which resulted in activation and consequently, in prevention of phage growth through presumed methyl-specific restriction endonuclease activity (Abedon, 2012; Hoskisson et al., 2015). The PglZ protein family is a central element of Pgl, however, the mechanisms of this complex system are poorly understood (Makarova et al., 2013). Another major line of defense is based on dormancy or programmed cell death (Makarova et al., 2013). These can be separated into toxin–antitoxin (T–A) systems and abortive infection (ABI). In the T–A system, the protein toxin kills cells above a certain expression level. The antitoxin component then regulates and/or inactivates toxin expression and prevents killing of the cell. The ABI system is also based on cell death or dormancy and it is also based on two modules (Fineran et al., 2009). The ABI system activates cell death to prevent viral replication and thereby protects the bacterial population.

In the present study we aimed to characterize defense systems of marine sponge-associated microbial consortia. The microbial metagenomes of three Mediterranean sponges (*Petrosia ficiformis*, *Sarcotragus foetidus*, *Aplysina aerophoba*) and seawater were compared toward this goal. Besides insights into the microbial community composition and overall GC content, we present defense-related features that consist of the CRISPR system, restriction modification, phage growth inhibition, and genes related to DNA phosphothiolation. The results of the present study are consistent with the concept of "functional convergence" (Fan et al., 2012) that shows similar functional profiles in the microbiomes of different sponge species and that are distinct from those of seawater.

MATERIALS AND METHODS

The sponges *P. ficiformis* (sample ID: 1Biotec2_S07) and *S. foetidus* (sample ID: 1Biotec2_S06) were collected on 25 May 2013, by SCUBA diving in Milos, Greece (N36.76759° E24.51422°), at 5–7 m depth. Sponge tissues (5 ml each) were washed with sterile-filtered seawater, passed through a 100 μm Nitex cloth (Hartenstein, Germany) and transported to the laboratory in glycerol solution (15% v/v) at −20°C until further processing. A total of 10 L seawater (sample ID: Biotec_SW) was collected from the vicinity of the sponges. Within 2–3 h after collection, seawater was filtered consecutively through 100 μm Nitex (Hartenstein), 5 μm durapore (Merck-Millipore), and finally through 0.22 μm durapore membrane filters, which were then frozen at −20°C.

Sponge samples of *A. aerophoba* were collected in the Mediterranean Sea from a depth of 5 m (Piran, Slovenia), on 07 May 2013. Upon transport back to the laboratory, samples of pinacoderm and mesohyl were separated using a sterile scalpel blade. One scalpel blade was used per each sample to prevent cross-contamination between samples. Microbial cells were enriched from the different sponge tissues by differential centrifugation (Fieseler et al., 2006). Microbes from *P. ficiformis* and *S. foetidus* samples were prepared using the same protocol. Fractions of sponge-associated prokaryotes (SAPs) were frozen at −80°C in 15% glycerin.

DNA Extraction and Sequencing

Genomic DNA was extracted from the sponge SAP preparations of *P. ficiformis* and *S. foetidus* and the seawater filters using the FastDNA Spin Kit for Soil (MP Biomedicals, USA). The quantity of metagenomic DNA was determined by spectrophotometry using a NanoDrop 2000c reader (PEQLAB Biotechnologie

GmbH, Germany). The quality and size were analyzed by visual inspection on 0.8% agarose gels following electrophoresis.

DNA of *A. aerophoba* was extracted in triplicates for each pinacoderm and mesohyl using the FastDNA SPIN Kit for Soil (MP Biomedicals). In order to maximize DNA yield from bacteria with different cell properties, the cell lysis step varied for the three replicates of each tissue type: (i) bead beating, following the manufacturer's protocol, (ii) freeze-thaw cycling (three cycles of 20 min at −80°C and 20 min at 42°C), (iii) proteinase K digestion (bacterial pellet re-suspended in 567 μl TE with SDS in a final concentration of 0.5% and proteinase K in 100 ng/ml final concentration) for 1 h at 37°C. After cell lysis, the manufacturer's protocol was followed for all six samples. Extracted metagenomic DNA from *A. aerophoba* samples was sequenced on an Illumina HiSeq2000 platform (150 bp paired-end reads) and quality filtered at the DOE Joint Genome Institute (Walnut Creek, CA, USA). Seawater, *P. ficiformis* and *S. foetidus* derived DNA was sequenced at GATC Biotech AG (Cologne, Germany) on an Illumina MiSeq Personal Sequencer (250 or 300 bp paired-end reads, respectively).

Raw Data Processing and Assembly

The raw reads obtained for the samples of *P. ficiformis*, *S. foetidus*, and seawater were initially analyzed with FastQC 0.11.2[1] for adapters, overall quality, length and ambiguous bases. In a first step, the reads were trimmed using Trimmomatic 0.31 (PE -phred 33 LEADING:3 ILLUMINACLIP:2:30:10) (Bolger et al., 2014) and then merged using bbmerge[2]. All reads, merged and unmerged, were again subjected to Trimmomatic for further quality trimming and length filtering (SE -phred 33 SLIDINGWINDOW:4:25 MINLEN:150 AVGQUAL:30). The remaining reads were assembled with IDBA-UD 1.1.1 (-mink 10 -maxk 100) (Peng et al., 2012). Contigs with a length ≤1000 nt were discarded. The reads obtained for the *A. aerophoba* dataset were processed via the IMG/ER webserver (Markowitz et al., 2012). Quality filtered reads were normalized using bbnorm and assembled with SPAdes 3.5.0 (-only-assembler, -k 21,33,55,77,99,127, -sc) (Bankevich et al., 2012). Only contigs ≥1000 nt were used for further analysis. To remove eukaryotic contamination, all contigs, that were further analyzed, were subjected to blastn 2.2.28 (e-value 10e-6 -task blastn) (Altschul et al., 1990) and searched against the NCBI nucleotide database (nt, as of September 29, 2015). The blast hits were analyzed with Krona 2.6 (Ondov et al., 2011). All reads of eukaryotic origin were removed. Information about the metagenomics datasets is presented in **Tables 1** and **2**.

Taxonomic Affiliation of Reads and Contigs

The processed reads were submitted to MG-RAST with enabled screening for human contamination and disabled dynamic trimming (Meyer et al., 2008). Contigs obtained from the metagenomic assemblies were assigned to taxonomy using blastx 2.2.28 (e-value 10e-6) and the NCBI non-redundant protein database (nr). All hits were submitted to *blast2lca* (default parameters), a last common ancestor algorithm implemented in MEGAN5 (Huson et al., 2011).

Comparison of GC Content and Average Genome Sizes

The GC content of all four metagenomes was calculated for all processed and filtered reads, using an in-house perl script. In addition, the average genome size per metagenome was computed with MicrobeCensus 1.0.7 (Nayfach and Pollard, 2015) using the same reads and their average calculated length (**Table 1**).

Data Normalization

Processed reads were mapped to their respective assembly using bowtie2 2.2.4 (very-sensitive) (Langmead and Salzberg, 2012). The coverage for each position on a contig was calculated with samtools depth 1.2 (Li et al., 2009). With this data, the coverage of each contig was set as the mean coverage over each position. To account for the different sequencing depths, the number of mapped reads and assembly size, the coverage for each contig was divided by the total number of mapped basepairs and multiplied by 10^6 to obtain copy numbers per megabase (cpm).

Functional Annotation

All contigs were subjected to Prodigal 2.6.0 (-p meta, -c, -g 11) (Hyatt et al., 2010) to predict open reading frames (ORFs). Clusters of Orthologous Groups (COGs) obtained from the Conserved Domains Database (CDD) (Marchler-Bauer et al., 2015) were annotated using rpsblast 2.2.28 (e-value 10e-6). Protein families (Pfam) and TIGRFAM were assigned with the InterProScan pipeline 5.17 (Jones et al., 2014) based on the best hit (e-value 10e-6).

Characterization of CRISPR Arrays, Repeats, and Spacers

The presence of CRISPR arrays was analyzed with a multiple tool approach similar as proposed by Gogleva et al. (2014) using CRT, PILER-CR, and CRISPRFinder (Bland et al., 2007; Edgar, 2007; Grissa et al., 2007b). *Cas* genes were identified by subjecting ORFs of CRISPR-containing contigs to TIGRFAM and Pfam databases using InterProScan. Assignment of CRISPR-Cas types was accomplished according to Makarova et al. (2011) using the TIGRFAM and Pfam annotations. Contigs containing CRISPR arrays found with CRT and PILER-CR or included *cas* genes

TABLE 1 | Samples analyzed in this study.

	Seawater	*Petrosia ficiformis*	*Sarcotragus foetidus*	*Aplysina aerophoba*
Sample date	29.05.2013	29.05.2013	29.05.2013	07.05.2013
Location	Mediterranean Sea, Milos, Greece	Mediterranean Sea, Milos, Greece	Mediterranean Sea, Milos, Greece	Mediterranean Sea, Piran, Slovenia
Depth	5–7 m	5–7 m	5–7 m	5 m
Temperature	20°C	20°C	20°C	18°C

[1] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[2] https://sourceforge.net/projects/bbmap/

TABLE 2 | Statistics on the processing of the metagenomics samples from sequencing throughput to analysis.

	Seawater	*Petrosia ficiformis*	*Sarcotragus foetidus*	*Aplysina aerophoba*
Sequencing platform	Illumina MiSeq (2 × 300 bp)	Illumina MiSeq (2 × 250 bp)	Illumina MiSeq (2 × 300 bp)	Illumina HiSeq (2 × 150 bp)
Sequenced reads (#)	40,505,000	41,383,600	32,672,426	945,906,728
Sequenced bp	12,151,500,000	10,345,900,000	9,801,727,800	283,772,018,400
Reads after QC (#)	18,273,997	29,213,518	17,525,606	–
Bp after QC	6,240,860,642	7,655,556,186	4,909,483,386	–
Assembly algorithm	IDBA-UD	IDBA-UD	IDBA-UD	SPAdes
Assembly size (bp)	216,407,276	226,772,563	190,159,175	489,999,481
Contigs > 1000 bp	116,626	82,740	41,164	110,609
N_{50} contigs (bp)	1,853	3,381	9,706	8,958
Largest contig	58,177	342,148	369,775	1,056,271
Average %GC	41	63	63	58
Open reading frames	215,442	221,522	175,356	455,396
ORF with COG annotation	129,900	119,914	103,075	203,692
ORF with Pfam annotation	78,569	55,478	29,253	56,643
ORF with TIGRFAM annotation	28,017	17,214	10,675	17,267
Average genome size (bp)	1,347,075.38	3,034,048.52	3,744,502.76	5,165,191.54
Reads mapped to assembly	13,396,184	15,900,219	22,478,672	537,464,688
Bp mapped to assembly	3,642,606,507	3,539,701,337	5,378,691,619	80,619,703,200
Average coverage	16.83	15.61	28.29	164.53

were uploaded to CRISPRfinder and were validated as true hits. Of these, only confirmed CRISPR with at least two spacers were retained. Possible targets of spacers were identified by submitting their sequences to CRISPRtarget using the ACLAME (as of August, 2009), GenBank-Phage, GenBank-Plasmid and RefSeq-Viral databases (all as of September, 2015) (gap open -5, gap extend -2, nucleotide match +1, mismatch −1, e-value 0.1, word size 7) (Biswas et al., 2013). Direct repeat seqences were submitted to CRISPRdb (Grissa et al., 2007a) and blasted against the CRISPRfinder database (e-value 10e-2) and CRISPRmap (Lange et al., 2013) to examine their superclasses by sequence and structure and to determine if they were reported before. The origin of the CRISPR arrays was determined through their respective contigs as described in Section "Taxonomic Affiliation of Reads and Contigs."

Analysis of Restriction Modification Systems

Reference protein sequences of type I [restriction endonucleases (REase), methyltransferases (MTase), and specificity domains], type II (REases ant MTases), and type III (REases and MTases) RMSs) were downloaded from REBASE (as of October 15, 2015) (Roberts et al., 2015). For each type of REases, MTases and specificity domains, a blast database was built. Predicted ORFs from all metagenomes were queried against the databases using blastp 2.2.28 (e-value 10e-6) and only hits with a coverage ≥70% were kept. A RMS was considered as being complete, if its restriction endonuclease and methyltransferase were at least four genes apart from each other (Oliveira et al., 2014). Finally, overlapping regions of REases and MTases (and specificity domains for type I) of the same type within four genes were combined to one cluster to avoid double counts.

Deposition of Sequence Data

The sequencing projects were completed in 2013 and sequencing data was deposited in the Sequence Read Archive (SRA), metagenome assemblies as a Whole Metagenome Shotgun (WGS) projects in GenBank under the BioProject PRJNA318959 and the BioSample IDs SAMN04870510, SAMN04870527, SAMN04870528 and SAMN05860141 for *P. ficiformis* (SRA: SRP074318, WGS: LXNJ00000000), *S. foetidus* (SRA: SRP074318, WGS: LXNI00000000), seawater (SRA: SRP074318, WGS: LXNH00000000), and *A. aerophoba* (WGS: MKWU00000000). Raw sequencing data of *A. aerophoba* is available under the GOLD Study ID Gs0099546[3] with the GOLD Project IDs Gp005580–Gp005585 which can be downloaded via the JGI Genome Portal.

RESULTS

Sample Description

Three samples from the sponges *P. ficiformis*, *S. foetidus,* and *A. aerophoba* as well as one seawater sample from the Mediterranean Sea were investigated in this study for functional differences of their associated microbiomes (**Table 1**). Using Illumina MiSeq and HiSeq platforms, more than 1,064,000,000 high-quality sequences (∼310 Gbp) were generated. The metagenomes had assembly sizes ranging from 190 to 489 Mbp. The predicted ORFs ranged from 175,356 in *S. foetidus* up to 455,396 in *A. aerophoba.* A total of 44.73–60.29% could be annotated via COGs (**Table 2**). In order to compare the generated data, the metagenomes were normalized based on their coverage which ranged from 15.61- to 164.53-fold.

[3] https://gold.jgi.doe.gov/biosamples?Study.GOLD%20Study%20ID=Gs0099546

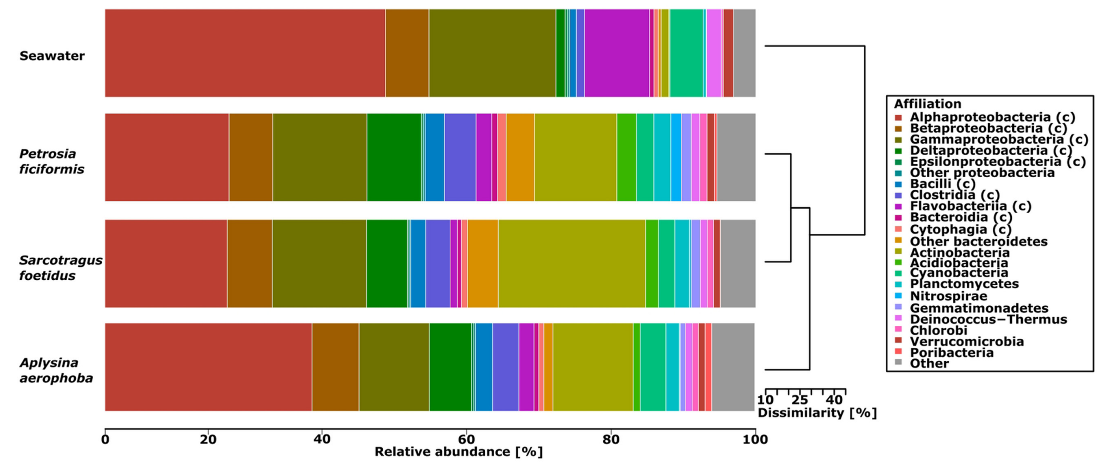


FIGURE 1 | Barplot showing the relative genomic diversity and associated hierarchical clustering of the metagenomic samples using the Bray–Curtis dissimilarity and complete linkage. The relative abundance is scaled on the x-axis. Taxonomic assignments are based on MG-RAST annotated phylum and class level (indicated by (c)). For the groups contributing ≥5% of relative abundance (*Proteobacteria, Bacteroidetes, Firmicutes*) class level assignments are given. The group 'other' comprises eukaryotes, viruses, archaea, fungi, bacteria at ≤1% abundance and unclassified sequences.

Genomic Composition

Based on phylogenetic affiliations using the lowest common ancestor algorithm (LCA) in MG-RAST, over 95% of the reads of all four metagenome samples were assigned to 38 bacterial phyla. Proportions of archaea, eukaryotes, viruses and unclassified sequences were consistently low (each ≤ 2.72%) in all metagenomes and were thus not further analyzed (Supplementary Figure S1). The genomic composition of the metagenomes was then analyzed on the phylum and class level. As indicated by Bray–Curtis dissimilarity, the metagenomes of *P. ficiformis and S. foetidus* were closest to each other (11.18% dissimilarity). Both showed dissimilarities of 17.01 and 19.46% to *A. aerophoba*. The seawater sample displayed dissimilarities of 41.04, 43.55, and 31.26% to *P. ficiformis*, *S. foetidus,* and *A. aerophoba,* respectively (**Figure 1**).

A limited number of sequences (0.15–0.21%) was not assigned to any known bacterial taxa. The *Proteobacteria*, *Firmicutes,* and *Bacteroidetes* were the most abundant phyla in all metagenomes. The *Actinobacteria* (12.3–22.63% vs. 1.18%) and the *Deltaproteobacteria* (6.28–7.24% vs. 1.38%) were more abundant in the sponge samples than in seawater. In contrast, the *Alphaproteobacteria* were less abundant in sponges than in seawater (18.80–31.81% vs. 43.11%), and so were the *Flavobacteria* (1.12–2.43% vs. 9.99%) and the *Cyanobacteria* (2.49–3.96% vs. 4.98%). Only minor differences between the sponge and seawater samples were found for the *Gammaproteobacteria* (10.77–14.48% vs. 19.51%), the *Clostridia* (3.75–4.86% vs. 1.25%) and other unclassified *Bacteroidetes* (1.38–4.74% vs. 0.43%) according to a principal component analysis (Supplementary Figure S3). Overall, the sponge metagenomes were taxonomically distinct from the seawater metagenome based on taxonomic read assignment using MG-RAST, Bray–Curtis dissimilarity, and principal component analysis.

GC Footprint

Higher average GC contents were detected for the assembled metagenomes of sponges (58–63%) than for seawater (41%) (**Figure 2A**; **Table 2**). The highest GC content was detected for the metagenome sample of *S. foetidus*, followed by *P. ficiformis, A. aerophoba*, and seawater. Interestingly, a second smaller seawater peak around 50–55% overlapped with the lower GC tail ends of the sponge metagenomes. To test whether there is a correlation between high GC content and genome size, we calculated the average genome sizes for a bacterial cell within each metagenome. The calculated average genome sizes in the sponge sample were considerably higher than those of the seawater sample (**Figure 2B**; **Table 2**).

General Functional Properties

Functional analysis was based on COG assignments. All hits were normalized to copy number per megabase based on their contig coverage. We identified 103,075–203,692 COG hits for the metagenomes which corresponds to 44.73–60.29% of annotated ORFs (**Table 2**). This number includes however the general function (G) and unknown function (S) categories (10.9–21.2 and 5.9–6.9%, respectively). The functional profiles of the sponge samples were more similar to each other than to seawater, as reflected by a Bray–Curtis dissimilarity of 10% between sponge and seawater metagenomes. Overall, many genes relate to the COG categories general function (G) or unknown function (S), and most of the COG categories were neither enriched for sponges nor the seawater metagenomes ("enriched" is defined as

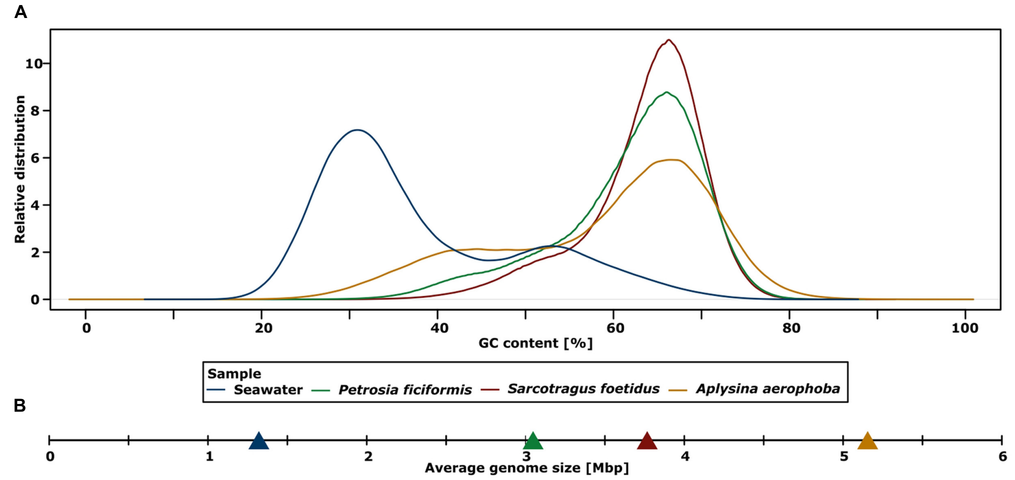


FIGURE 2 | (A) Plot of metagenomics samples showing the relative distribution of the GC content of filtered reads. **(B)** Calculated average genome sizes for bacteria of each metagenomic sample.

>1.5-fold more copies per megabase) (**Figure 3**). The category nucleotide transport and metabolism (F) was exceptionally high in *S. foetidus* (33.05 cpm), whereas the category of cell cycle control, cell division, chromosome partitioning (D) was exceptionally low in *P. ficiformis* (0.53 cpm) when compared to the other sponge metagenomes. Only few differences were identified between seawater and sponge metagenomes based on COG level assignments. The sponge metagenomes showed a higher number of genes assigned to functions related to defense mechanisms (V) and the cytoskeleton (Z), suggesting that these are important functional traits for sponge symbionts. On the other hand, fewer reads were assigned to translation, ribosomal structure and biogenesis (O), cell motility (N), and chromatin structure and dynamics (B) in the sponge sample, marking them as relevant functional features for free living bacteria.

Defense Mechanisms

COG and Pfam-Annotated Defense Mechanisms

With respect to defense mechanisms, the sponge datasets were more similar to each other than to seawater according to Bray–Curtis dissimilarity measure (**Figures 4A,B**). All comparisons are based on copies per megabase (cpm). Features were defined as "enriched" when being >1.5-fold abundant in either the sponge or seawater metagenome. Transport and efflux systems for drugs were found in all samples, and with the exception of a Na^+-efflux pump and ABC-type multidrug transporter, all related functions were enriched in the sponge samples over seawater. Furthermore, all annotations associated to CRISPR were enriched in the sponge microbiomes, as all (with the exception of one CRISPR-nuclease (COG3513) were absent from seawater. 11 features related to CRISPR in *S. foetidus* and one in *P. ficiformis* were missing from the sponge metagenomes, which were mostly related to the receptor activity-modifying proteins (RAMP) superfamily. Interestingly, the cas2-gene (COG1343) in *S. foetidus* may be substituted by a cas2-homolog (COG3512), which showed highest cpm within this metagenome. With respect to RMSs, all genes except one encoding for one endonuclease (COG1403) were enriched in the sponge datasets. However, one endonuclease (COG1787) copy was absent in *S. foetidus* and six were absent from seawater. The overall cpm's within the RMS were higher in sponge metagenomes than in the corresponding seawater metagenome. Classes A and C beta-lactamases (COG2367, COG1680) were further enriched in sponge metagenomes. On the other hand, seawater was enriched for the beta-lactamase class D (COG2602) and an inductive membrane protein (COG3725). A couple of genes related to resistance against colicin COG4452), a growth inhibiting toxin, bacteriophages (COG4823) and the antibiotic vancomycin (COG2720) were enriched in sponge metagenomes, whereas a cephalosporin hydroxylase (COG3510) was more abundant in the seawater metagenome (**Figure 4A**).

The overall gene copy numbers for DNA phosphorothioation (DND) and phage growth limitation (Pgl) were higher in the sponge than in the seawater metagenome (**Figure 4B**). The DndG (PF08747) was absent from seawater. With respect to the Pgl system, two core genes (COG1002, COG4930) and three additional genes (PF08849, PF10923, COG3472) were missing from seawater. The overall Pgl gene copy number in sponges was ~50% higher in *A. aerophoba* and *S. foetidus* than in *P. ficiformis* (**Figure 4B**).

Clustered Regularly Interspaced Short Palindromic Repeats

We analyzed CRISPR arrays and related components, i.e., direct repeat sequences separated by spacers and adjacent *cas* genes in the four metagenomes. The highest numbers of CRISPR array containing contigs were found by searching for *cas* genes.

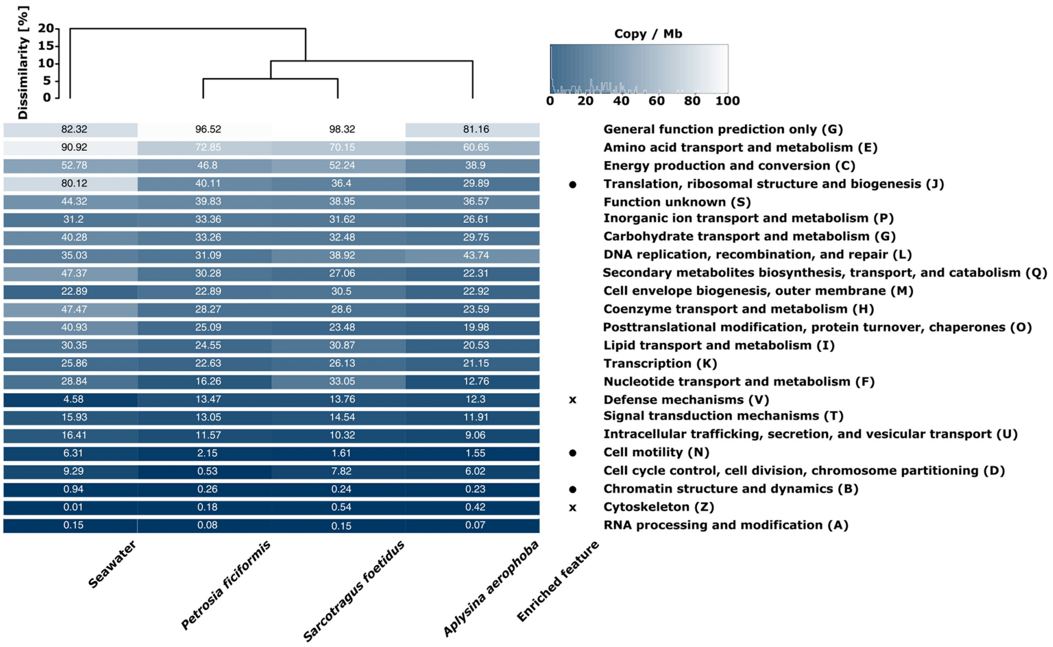


FIGURE 3 | Heatmap of COG functional categories for the four analyzed metagenomes. The color scale ranges from 0 (black) to 100 (white) and indicates copies per megabase metagenome (cpm). Functional dissimilarities (Bray–Curtis) are indicated by the dendrogram on top. The term "enriched feature" relates to COG classes which are on average at least 1.5-fold higher in seawater (circle) or in sponges (cross) over all sponge samples. COG classes are ordered from high to low copy numbers.

A. aerophoba showed the highest abundance of validated arrays, whereas none was identified in the seawater metagenome. The final number of identified CRISPR arrays was 77 (0.21 cpm), 47 (0.25 cpm), 283 (0.62 cpm), and 0 (0 cpm) for the metagenomes of *P. ficiformis*, *S. foetidus*, *A. aerophoba* and seawater, respectively (**Table 3**; **Figure 5**). On the domain level, taxonomy was assigned to 53 of 77 (68.83%) arrays in *P. ficiformis*, 40 of 47 (85.11%) arrays in *S. foetidus* and 240 of 283 (84.81%) arrays in *A. aerophoba*. Noteworthy, despite the differences in their geographic location and total number of identified arrays, a large overlap of taxonomic groups was found. The overall distribution of taxa containing CRISPR-contigs was similar in the three sponge datasets, with *Proteobacteria* as the most prevalent phylum followed by *Actinobacteria* and *Chloroflexi* in the sponges *P. ficiformis* and *S. foetidus* and *Firmicutes* in *A. aerophoba* (**Table 3**).

Different CRISPR-Cas types were categorized by their associated *cas* genes. In the metagenomic datasets, at least 26 (35.06%), 20 (42.55%), and 144 (46.78%) of the CRISPR arrays were adjacent to *cas* genes for *P. ficiformis*, *S. foetidus* and *A. aerophoba*, which indicates that these arrays might be complete (**Table 3**). The *cas* genes of all known CRISPR-Cas types were identified. CRISPR-Cas type I was the most abundant, with the subtypes I-E and I-C as the most prevalent for all metagenomes, followed by types II and III. Around 50% of *cas* genes could not be annotated in more detail (type unknown, **Table 3**). According to the number of *cas* genes per megabase, *cas1*, *cas2* were most abundant, followed by *cas3, cas4* and *cas7* in all three sponge metagenomes. Smallest proportions were detected for *cas8* and *cas9* (**Figure 5**). Even though CRISPR arrays were not identified in the seawater metagenome, two *cas6*-genes were detected.

Spacers are the functional part of the CRISPR defense that recognizes foreign DNA fragments. In the *P. ficiformis* metagenome, a set of 1,366 spacers was detected, of which 1,349 were unique (**Table 3**). The largest CRISPR array in *P. ficiformis* contained 112 spacers. For the *S. foetidus* metagenome, a total of 723 spacers was identified, of which 714 were unique. Here, the longest array contained 67 spacers. Thirdly, in the *A. aerophoba* metagenome, a total of 9,669 spacer sequences were detected with 125 of these occurring more than once and with 9,547 being unique. The longest array found in *A. aerophoba* comprised 169 spacers. None of the spacer sequences were shared between the metagenomic samples suggesting that the three sponge microbiomes have their own distinct CRISPR systems.

With respect to potential targets of the spacers, the number of hits decreased from unknown targets, to plasmids, phages and to viruses in all samples (**Table 3**). Combining these results with the spacer taxonomy, most spacers originated in *Alphaproteobacteria* and *Actinobacteria*. All taxonomic groups of spacers had hits

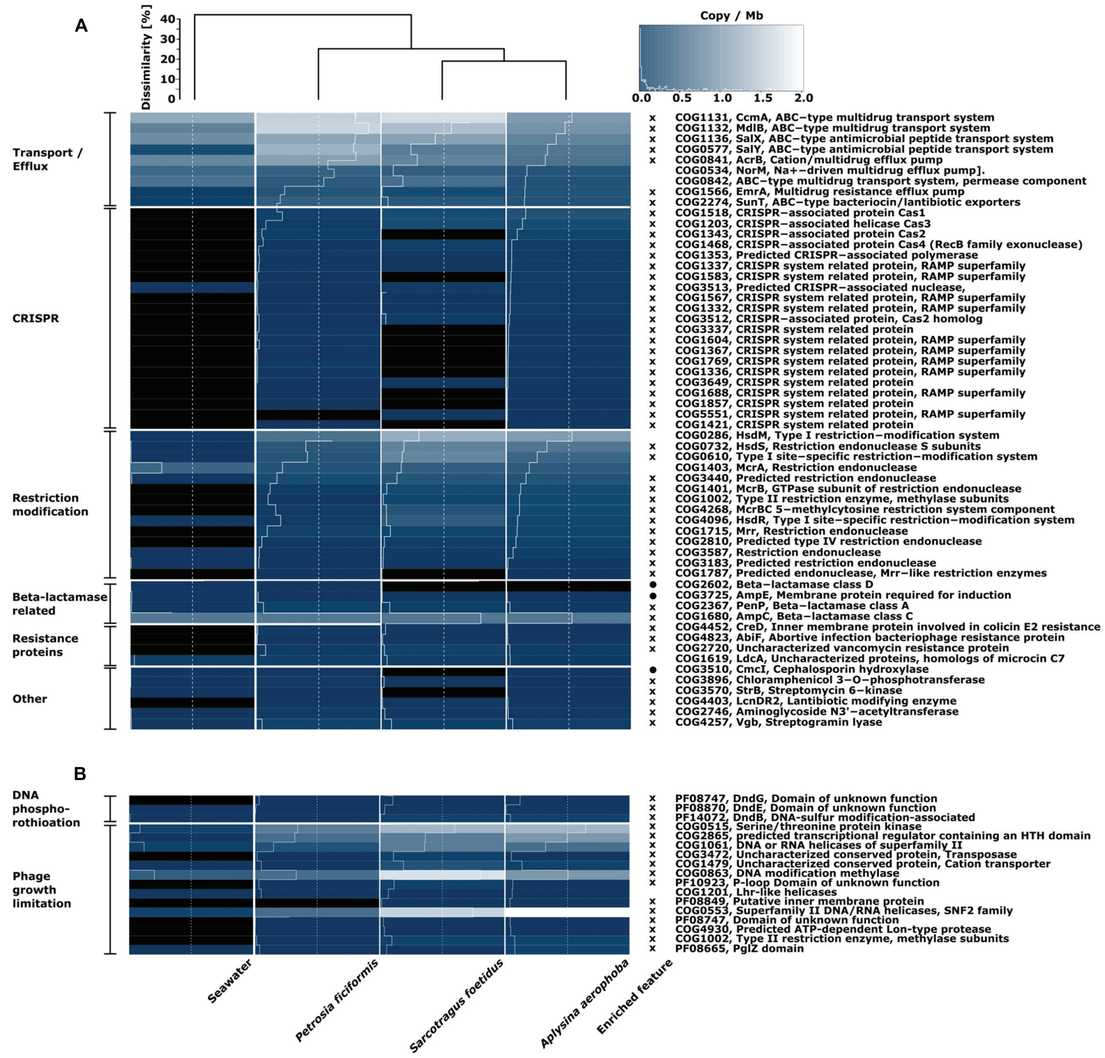


FIGURE 4 | Heatmap of defense mechanisms in (A) COG functional categories and (B) additional searches for the phage growth limitation and DNA phosphorothiotation in the COG and Pfam databases. The color scale ranges from 0 (black) to 2 (white) and indicates copies per megabase metagenome. Bray–Curtis dissimilarity is indicated by the dendrogram on top. Enriched feature relates to COG classes which are on average >1.5-fold higher in seawater (circle) or in sponges (cross) over all sponge samples. Similar COG annotations are labeled on the left side of the heatmap and ordered from high to low copy numbers.

in the four target groups, with the largest amount of hits found for unknown targets. The three sponge metagenomes were shaped similarly with respect to spacer origins and targets (**Figure 6**). The proportion of spacer sequences originating from *Betaproteobacteria* was highest in *S. foetidus*, whereas *Gamma-* and *Deltaproteobacteria* were highest in *A. aerophoba*. *S. foetidus* showed more spacers originating from *Firmicutes* than the other sponge samples. Spacers from *Spirochaetes* were only found in the *A. aerophoba* metagenome. Overall, the distribution of spacers with assigned taxonomy followed the genomic composition with correlation coefficients of 0.64, 0.59, and 0.88 (all p-values ≤ 0.05) for *P. ficiformis*, *S. foetidus,* and *A. aerophoba,* respectively (**Figures 1** and **6**). All spacers and direct repeat sequences are compiled in Supplementary Figures S2 and S3.

The number of unique direct repeats was 67, 40, and 218 for *P. ficiformis, S. foetidus,* and *A. aerophoba,* respectively (**Table 3**). In the datasets *A. aerophoba* and *S. foetidus*, three of the direct repeat sequences were shared, of which nine between

TABLE 3 | Raw counts of identified CRISPR arrays and their taxonomic assignments, *cas*-genes, spacers and direct repeats.

	Seawater	*Petrosia ficiformis*	*Sarcotragus foetidus*	*Aplysina aerophoba*
Pilerr-cr	5	90	76	384
CRT	9	108	83	529
Contigs with found Cas-genes	2	124	101	263
CRISPRFinder	0	77	47	290
CRISPR per megabase	0	0.21	0.25	0.62
CRISPR with assigned taxonomy	0	53	40	240
Proteobacteria	0	36	22	169
Actinobacteria	0	9	6	33
Chloroflexi	0	3	4	6
Cyanobacteria	0	1	1	6
Firmicutes	0	1	2	14
Acidobacteria	0	2	1	4
Verrucomicrobia	0	1	2	0
Bacteroidetes	0	0	1	7
Deinococcus–Thermus	0	0	1	3
CRISPR assigned to CAS-genes	0	26	20	144
Type I (A–F)	0	11	10	73
Type II (A–C)	0	1	2	10
Type III (A–B)	0	1	0	6
Type unknown	0	13	8	55
Largest array (# spacer)	0	112	67	169
Total number of spacer	0	1,366	723	9,669
Unique spacer	0	1,349	714	9,547
Spacer with found target	0	278	152	1,642
Phage	0	55	42	255
Virus	0	19	10	146
Plasmid	0	204	100	1,241
Target unknown	0	1,088	581	8,027
Total number of repeats	0	77	47	290
Number of unique repeats	0	67	40	218
Repeats with hits to CRISPRdb	0	55	25	144
CRISPRmap superclass A/B/C/D/E/F	0 – 0/0/0/0/0/0	47 – 1/7/19/2/6/2	21 – 0/6/8/0/5/2	88 – 3/30/36/0/23/7

P. ficiformis and *A. aerophoba*, suggesting a horizontal transfer of either CRISPR arrays or bacteria. An amount of 55 (81.09%), 25 (62.5%), and 144 (66.06%) of *P. ficiformis, S. foetidus,* and *A. aerophoba* derived repeats were assigned to known repeat sequences using CRISPRdb. With respect to the classification using CRISPRmap, 47 (70.15%), 21 (52.5%), and 88 (41.12%) direct repeats for *P. ficiformis, S. foetidus,* and *A. aerophoba* were assigned to known superclasses, with the most abundant classes C, E, and B (**Table 3**). Notably, the superclasses were ordered decreasing in their conservation (Lange et al., 2013), showing a mixture of repeats with a roughly corresponding structure (superclasses B and C) and little sequence conservation (superclass E). Overall, 81.09 and 70.15% of all direct repeat sequences could be classified using CRISPRdb and CRISPRmap, respectively.

Restriction Modification Systems

We identified a total of 3,057 RMSs in the metagenome datasets with 432 assigned to type I RMS, 2,379 to type II RMS, and 246 to type III RMS. A normalization of these raw counts to copies per megabase (cpm) resulted in a similar distribution of RMS types I-III in the metagenomes. The sponge metagenomes showed higher abundances of all RMS types than seawater (2.48–5.08 cpm in sponges vs. 0.18 cpm in seawater). Type II was the most prevalent RMS type in the inspected metagenomes (*S. foetidus* = 4.03 cpm, *A. aerophoba* = 3.16 cpm, *P. ficiformis* = 2.06 cpm, seawater = 0.17).

The majority of type I RMS genes were assigned to *Proteobacteria*, *Actinobacteria,* and *Deinococcus–Thermus* in the sponge metagenomes, while in seawater, type I RMS was assigned exclusively to the classes *Beta*- and *Gammaproteobacteria* (**Figure 7**). The majority of type II RMS in sponge metagenomes was assigned to *Proteobacteria (Alpha- and Gamma-)* and *Actinobacteria* as well as to a lesser extent, to *Bacteroidetes*, *Cyanobacteria,* and *Acidobacteria*. The *S. foetidus* metagenome contained an unusually high number of type II RMS affiliated to *Actinobacteria*. Type III RMS was the most underrepresented group. Type III RMS in sponge metagenomes was most represented by *Alpha*- and *Gamma-Proteobacteria* as well as *Bacteroidetes* and *Chloroflexi,* while type III in the seawater sample was only represented by the *Alpha*- and *Gammaproteobacteria*, *Bacteroidetes* and the *Clostridia*.

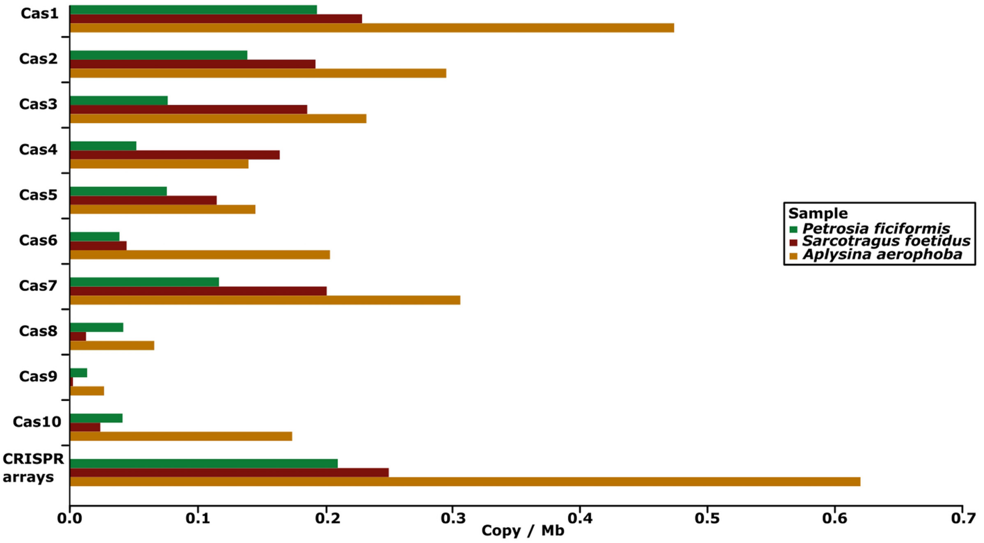


FIGURE 5 | Barplot showing the abundance of CRISPR arrays and *cas* genes in the four metagenomes. The x-axis shows their abundance in copy number per megabase.

DISCUSSION

General Features

We performed the taxonomic assignment of metagenomics reads by MG-RAST which has been previously attempted using microbial metatranscriptome data from a low microbial abundance sponge (Moitinho-Silva et al., 2014). While this approach offers the advantage of using the full metagenome dataset rather than a single gene marker (i.e., 16S rRNA gene), it may lose resolution for those phyla and candidate phyla where references genomes are not available. Our results confirm previous findings that sponges harbor a distinct microbiota which is different from that of the surrounding seawater. Principal component analysis of the relative abundance of reads revealed a clustering of the sponge samples (Supplementary Figure S3). The sponge metagenomes overlapped in their composition and they showed a higher proportion of *Actinobacteria* and *Deltaproteobacteria* than seawater based on assignment of complete metagenomic reads. In contrast, the seawater metagenome revealed higher abundances of *Alphaproteobacteria*, *Flavobacteria,* and *Cyanobacteria* compared to the sponge metagenomes. With increasing availability of sequence data and the completion of draft genomes by single cell genomics (Kamke et al., 2014) or binning approaches (Gao et al., 2014; Burgsdorf et al., 2015), the assignment of complete reads rather than single gene markers should become widely acceptable.

The sponge metagenomes displayed much higher GC contents (58–63%) than the seawater metagenome (41%) (**Figure 2A**). As has previously been recognized, the prokaryotic GC content can be highly variable between different environments (Foerstner et al., 2005; Reichenberger et al., 2015), ranging from 34% for Sargasso Sea surface water samples to 61% for terrestrial soils. The GC composition of the sponge metagenomes is much higher than most other metagenomes, only to be superseded by metagenomes from saline ponds and contaminated soils (Reichenberger et al., 2015). While an explanation for the variation in GC composition remains wanting, there is increasing evidence that both, the phylogenetic composition of the samples and the environment shape the GC composition of the resident microbiota. With respect to the sponge metagenomes, the GC contents are likely a result of bacterial community composition. *Actinobacteria*, which are known for their high GC content, are much more prevalent in the sponge metagenomes than in seawater. Accordingly, *S. foetidus* displayed the largest abundance of *Actinobacteria* (**Figure 1**) and the highest GC content (**Figure 2**). Nonetheless, this cannot be the only explanation, because in spite of variable abundances of *Actinobacteria* within the three sponges (**Figure 1**), the GC content is very narrow (**Figure 2A**). Therefore, we posit that the specific microenvironment within sponges has some yet to be characterized effect on the microbial GC composition of sponges.

The sponge metagenomes displayed larger calculated average genome sizes (3.0–5.1 Mb) than that calculated for seawater (1.35 Mb) (**Figure 2B**; **Table 2**). The estimates for sponge bacterial genomes are on the larger end of genome size estimates derived from diverse metagenomic data (Giovannoni et al., 2014). It should however be noted that the comparison of closely related *Synechococcus* genomes from sponge symbionts versus those from seawater did not reflect this pattern (Burgsdorf et al., 2015). Larger genomes of sponge-associated bacteria may be the

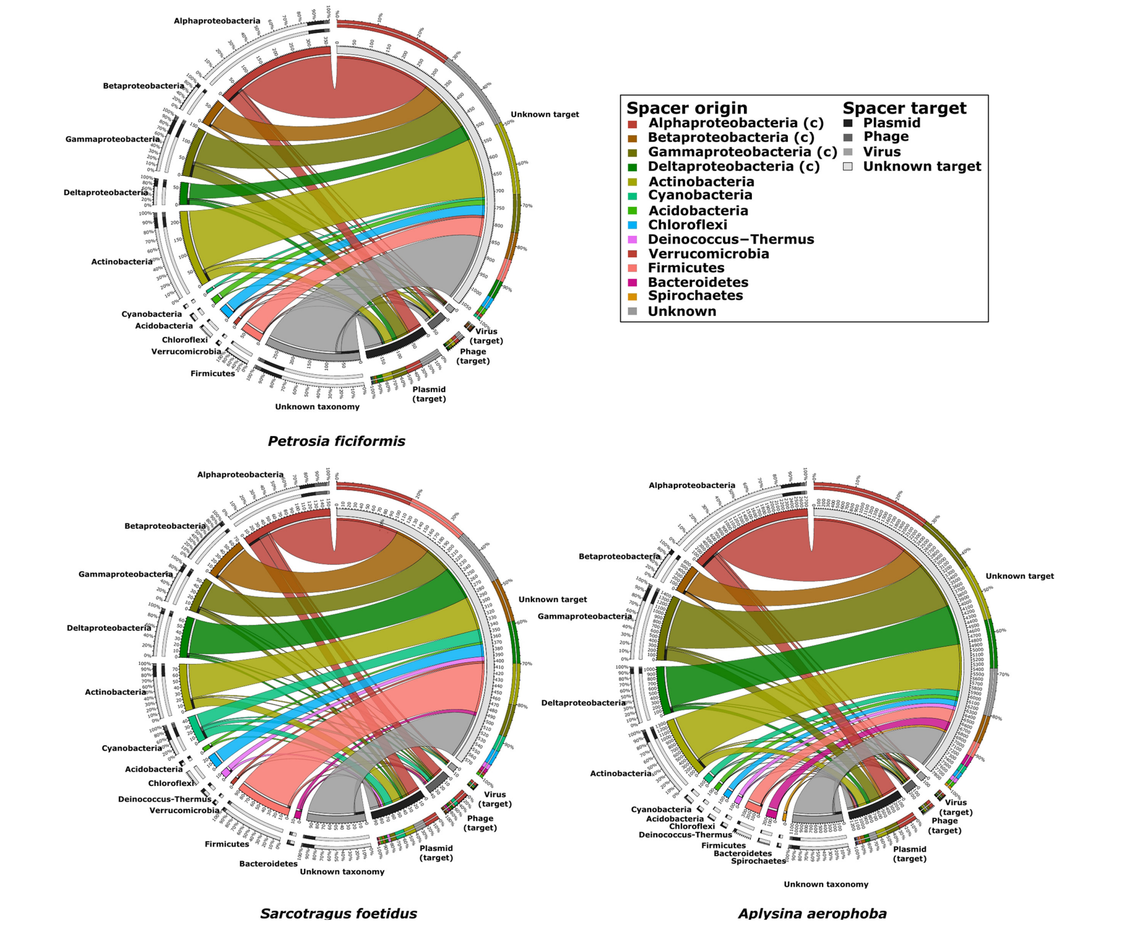


FIGURE 6 | Plots showing the origin (left side of circles) and targets (right side of circles) of spacer sequences for the three sponge datasets. The two outermost rings indicate the percentage of target found for each spacer and vice versa. The inner ring indicates the number of spacers connected to the origin and target, respectively.

evolutionary consequence of a more variable and nutrient-rich microenvironment within the sponge as opposed to the stable, nutrient poor seawater. Further the sponge-associated microbial consortia are constantly exposed to an ample source of free DNA resulting from the host's digestion of food bacteria. Whether and to what extent the mechanisms of horizontal gene transfer occur in sponges and whether this would then results in larger symbiont genomes remains to be investigated in future studies. The high prevalence of transposases and other mobile genetic elements within sponge microbiomes (Fan et al., 2012) does suggest that horizontal gene transfer is rampant in the sponge holobiont.

The overall functional annotation on the level of COG categories was more similar within the sponge samples than compared to the seawater sample (**Figure 3**). The functional profile of *A. aerophoba* was more distant to the other sponge samples, which may have been influenced by a higher functional diversity as shown in the rarefaction curve (Supplementary Figure S2). Overall, only two COG categories were enriched in sponge metagenomes (defense mechanisms; cytoskeleton), while three COG categories were depleted in sponge metagenome over seawater (translation, ribosomal structure, biogenesis; cell motility; chromatin structure and dynamics). The category cytoskeleton was not pursued further owing to low gene abundance (<0.6 cpm). These results are somewhat different from previous data (Thomas et al., 2010), where the metagenome of the Australian sponge *C. concentrica* was enriched in two COG categories (secondary metabolites biosynthesis, transport and catabolism; replication, recombination, and repair) while being depleted in three other categories (translation, ribosomal

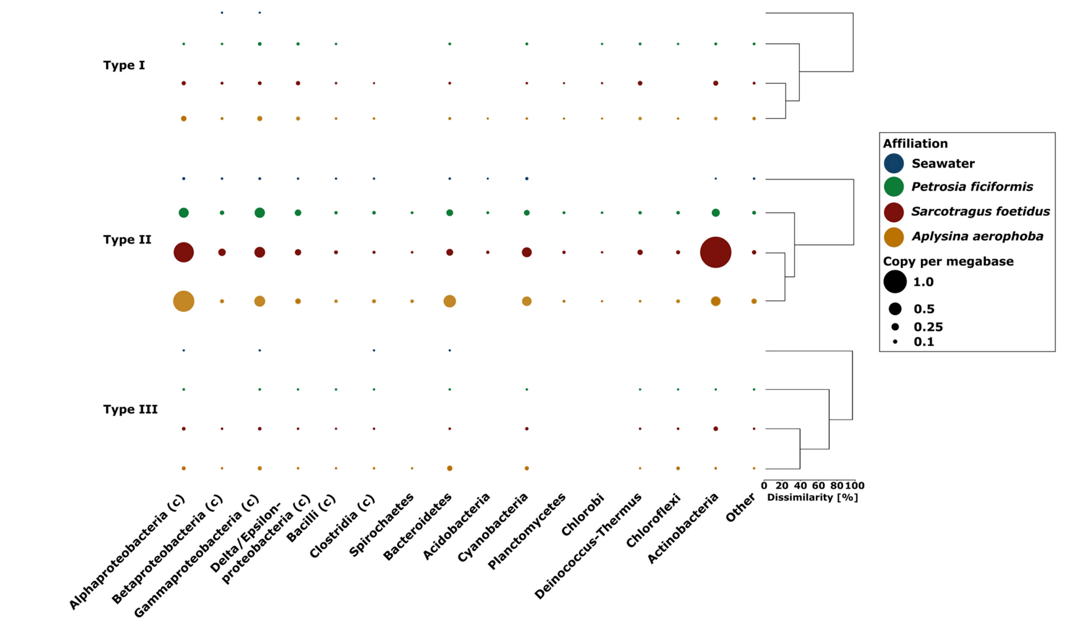


FIGURE 7 | Presence of types I–III restriction modification systems in the sponge and seawater metagenomes along with additional taxonomic assignments on the phylum and class (indicated through the c in brackets) level. The size of each bubble indicates the gene copy number per megabase. Bray–Curtis dissimilarity for each RMS type is indicated by the dendrograms.

structure and biogenesis; nucleotide transport and metabolism; energy production and conversion in comparison to seawater). The only shared feature between these analyses is the depletion of sponge metagenomes in the category: translation, ribosomal structure, biogenesis. The category defense mechanisms is discussed in detail below.

Defense Systems

The overall enrichment of the category defense mechanisms in sponge metagenomes over seawater metagenome is in agreement with earlier results, where functions related to viral defense were found to be enriched in sponges (CRISPR-Cas system, RMS) over surrounding bacterioplankton (Thomas et al., 2010; Fan et al., 2012) or in selected bacterial reference genomes (Burgsdorf et al., 2015). The defense system DNA phosphorothioation was, even though not functionally complete, also more prevalent in the sponge metagenomes. Further, genes associated with phage growth limitation (Pgl) were enriched in the sponge metagenomes. The microbial consortia within sponges may thus not only defend themselves against viruses and phages, but may also be capable of suppressing their growth. Since the Pgl system is only poorly characterized, further studies are needed to fully understand its potential impact on microbial communities.

Clustered regularly interspaced short palindromic repeats arrays were identified through a protocol using three different tools to avoid false positive hits. The total set of arrays (and direct repeats) was 77, 47, 283, and 0 for the metagenomes of *P. ficiformis*, *S. foetidus*, *A. aerophoba* and seawater, respectively. While the CRISPR arrays in sponge metagenomes (0.21–0.62 cpm) are below the values described for completely sequenced genomes (0.72 cpm), they are still an order of magnitude above the values for seawater metagenomes, such as derived from the Sorcerer II Global Ocean Sampling expedition (0.042 cpm) (Sorokin et al., 2010). This value suggests a low number of CRISPRs in seawater and indeed, we found 0 hits in our seawater metagenome. The variation in the number of observed CRISPR arrays between the sponge metagenomic datasets may be due the fragmentation of generated contigs and is based on the used sequencing technology and the assembly algorithms (Gogleva et al., 2014).

The overall taxonomic assignment (>68.83%) was comparable between the datasets with the largest fraction of CRISPR arrays affiliated to *Proteobacteria* followed by *Actinobacteria*, *Chloroflexi* and *Firmicutes* (**Table 3**). Similar results were observed for the origin of the spacer sequences (**Figure 6**). While this finding supports the presence of similar microbiomes within the different sponge species, only a small overlap of repeat sequences was identified. As ∼30% of direct repeats could not be assigned to a superclass or known repeats, they may represent novel direct repeats. The fact that we did not detect any shared spacers suggests that the acquisition of protospacers may vary between bacterial individuals (Gogleva et al., 2014). The sponge-associated bacteria may either be exposed to different types of viruses, phages or plasmids (Burgsdorf et al., 2015) or to distinct viral variations (Fan et al., 2012). With respect to the targets of the spacer sequences, their number decreased from the group "unknown" to plasmids, phages and viruses, and they were uniformly distributed among all identified phyla (**Figure 6**). The large fraction of hits to unknown and unique spacer sequences suggests that a large number of novel and diverse CRISPR targets and spacers can be expected in marine sponge metagenomes. The small overlap between spacers and direct repeats of the CRISPR-Cas systems likely reflects variations within each sponge metagenome as well as the specific acquisition of spacers from selected bacteria.

We found only ∼50% of all CRISPR arrays adjacent to *cas* genes, which is likely an effect of the fragmentation of the assemblies. The *cas* genes were used to classify CRISPR systems into types and subtypes according to Makarova et al. (2011). Overall and as was expected, *cas1* (an universal marker of all CRISPR-*cas* systems) and *cas2* were most prevalent. CRISPR-Cas type I, described via *cas3*, was the most prevalent in all three sponge metagenomes followed by types II and III, identified by *cas9* and *cas10*. The latter two types were only found in very low abundances, suggesting type I to be the most important CRISPR type in the sponge microbiome. Interestingly, type I was also most prevalent in other environments such as the human gut (Gogleva et al., 2014) or groundwater (Burstein et al., 2016). The most abundant subtypes I–E showed a strong link to *Actinobacteria* (Makarova et al., 2015). In ecological terms, the high prevalence of CRISPR-Cas systems in sponge microbiomes may be necessary to defend the sponge-associated bacteria against viral particles that are drawn into the sponge holobiont by filtration. It has previously been estimated that the sponge-associated bacteria may be exposed to as many as 1000 viral particles per day (Thomas et al., 2010), thus an efficient defense against viral onslaught could be essential.

Restriction modification system have previously been shown to be more abundant in metagenomes from Australian sponges than in seawater (Fan et al., 2012). We here confirm these results for the Mediterranean sponges (2.48–5.08 cpm vs. 0.18 for RMS in seawater metagenome). The difference might be explained by the observation that larger genomes tend to have more RMS than smaller genomes (Makarova et al., 2013), which is indeed the case for the sponge metagenomes over the seawater metagenomes (**Figure 2B**). Among the different types of RMS, type II was most abundant in the metagenomes (**Figure 7**) which is consistent with previous findings for bacterial isolates (Oliveira et al., 2014). Similar to CRISPR, the RMS are mostly affiliated with *Alphaproteobacteria*, *Gammaproteobacteria*, *Betaproteobacteria,* and *Actinobacteria*. Both CRISPR and RMS thus appear as the first line of defense against foreign DNA, in particular against attack by viruses or phages.

CONCLUSION

A comparison of microbial metagenomes from different Mediterranean sponge species versus seawater revealed bacterial defense systems as the consistently enriched feature in sponge metagenomes. These defenses include CRISPRs, RMSs, phage growth inhibition and DNA phosphorothioation as the main mechanisms to combat foreign DNA from viruses, phages or other sources. The expanded genomic repertoire for bacterial defenses is likely the result of an evolutionarily long-standing adaptation where the resident sponge microbiota is exposed to free DNA resulting from the immense filtration activities of the animal host. In support of this, higher GC contents and larger calculated genome sizes were identified in sponge metagenomes over seawater. Collectively, our results indicate that the genomes of sponge microorganisms are/have been subject to horizontal gene transfer and that defense against foreign DNA is one prerequisite for an existence within sponges.

AUTHOR CONTRIBUTIONS

HH designed the study, carried out metagenomics analysis, prepared figures and drafted the manuscript. BS collected the *A. aerophoba* sample and carried out its analysis. MJ participated in the analysis, preparation of figures and study design. KB sampled *P. ficiformis* and seawater and helped in the study design. LM-S sampled *S. foetidus* and participated in the study design. FF carried out analysis and wrote scripts. UA participated in the study coordination and drafting of the manuscript. UH coordinated the study and drafted/finalized the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was financially supported by the European Commission within its FP7 Program, under the thematic area KBBE.2012.3.2-01 with Grant Number 311932 ("SeaBioTech") to UH and the DOE Joint Genome Institute with Grant Number CSP1291 to UH. BS and MJ were supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg.

ACKNOWLEDGMENTS

We acknowledge the divers Panagiotis Vavilis and Thanos Dailianis (HCMR, Crete) for their diving support during the Milos sponge collection. We further thank the staff of the Marine Biology Station at Piran for their excellent support during our field work. We are grateful to Dr. Michael Richter (MPI Bremen) for help in first *P. ficiformis* metagenome assembly, and Dr. Rodrigo Costa (University of Algarve) as well as Dr. Alexander Keller (University of Würzburg) for helpful suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.01751/full#supplementary-material

REFERENCES

Abedon, S. T. (2012). Bacterial 'immunity' against bacteriophages. *Bacteriophage* 2, 50–54. doi: 10.4161/bact.18609

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comp. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bayer, K., Schmitt, S., and Hentschel, U. (2008). Physiology, phylogeny and in situ evidence for bacterial and archaeal nitrifiers in the marine sponge *Aplysina aerophoba*. *Environ. Microbiol.* 10, 2942–2955. doi: 10.1111/j.1462-2920.2008.01582.x

Bell, J. J. (2008). The functional roles of marine sponges. *Estuar. Coast. Shelf Sci.* 79, 341–353. doi: 10.1016/j.ecss.2008.05.002

Biswas, A., Gagnon, J. N., Brouns, S. J., Fineran, P. C., and Brown, C. M. (2013). CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol.* 10, 817–827. doi: 10.4161/rna.24046

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., et al. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8:209. doi: 10.1186/1471-2105-8-209

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Burgsdorf, I., Slaby, B. M., Handley, K. M., Haber, M., Blom, J., Marshall, C. W., et al. (2015). Lifestyle evolution in cyanobacterial symbionts of sponges. *MBio* 6, e391-15. doi: 10.1128/mBio.00391-15

Burstein, D., Sun, C. L., Brown, C. T., Sharon, I., Anantharaman, K., Probst, A. J., et al. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat. Commun.* 7:10613. doi: 10.1038/ncomms10613

Edgar, R. C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 8:18. doi: 10.1186/1471-2105-8-18

Fan, L., Reynolds, D., Liu, M., Stark, M., Kjelleberg, S., Webster, N. S., et al. (2012). Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. *Proc. Natl. Acad. Sci. U.S.A.* 109, E1878–E1887. doi: 10.1073/pnas.1203287109

Fieseler, L., Quaiser, A., Schleper, C., and Hentschel, U. (2006). Analysis of the first genome fragment from the marine sponge-associated, novel candidate phylum Poribacteria by environmental genomics. *Environ. Microbiol.* 8, 612–624. doi: 10.1111/j.1462-2920.2005.00937.x

Fineran, P. C., Blower, T. R., Foulds, I. J., Humphreys, D. P., Lilley, K. S., and Salmond, G. P. (2009). The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proc. Natl. Acad. Sci. U.S.A.* 106, 894–899. doi: 10.1073/pnas.0808832106

Foerstner, K. U., Von Mering, C., Hooper, S. D., and Bork, P. (2005). Environments shape the nucleotide composition of genomes. *EMBO Rep.* 6, 1208–1213. doi: 10.1038/sj.embor.7400538

Gao, Z. M., Wang, Y., Tian, R. M., Wong, Y. H., Batang, Z. B., Al-Suwailem, A. M., et al. (2014). Symbiotic adaptation drives genome streamlining of the cyanobacterial sponge symbiont "Candidatus *Synechococcus spongiarum*". *MBio* 5:e79-14. doi: 10.1128/mBio.00079-14

Giovannoni, S. J., Thrash, J. C., and Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *ISME J.* 8, 1553–1565. doi: 10.1038/ismej.2014.60

Gogleva, A. A., Gelfand, M. S., and Artamonova, I. (2014). Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs. *BMC Genomics* 15:202. doi: 10.1186/1471-2164-15-202

Grissa, I., Vergnaud, G., and Pourcel, C. (2007a). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8:172. doi: 10.1186/1471-2105-8-172

Grissa, I., Vergnaud, G., and Pourcel, C. (2007b). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* 35, W52–W57. doi: 10.1093/nar/gkm360

Hentschel, U., Piel, J., Degnan, S. M., and Taylor, M. W. (2012). Genomic insights into the marine sponge microbiome. *Nat. Rev. Microbiol.* 10, 641–654. doi: 10.1038/nrmicro2839

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. *Science* 327, 167–170. doi: 10.1126/science.1179555

Hoskisson, P. A., Sumby, P., and Smith, M. C. M. (2015). The phage growth limitation system in Streptomyces coelicolor A(3)2 is a toxin/antitoxin system, comprising enzymes with DNA methyltransferase, protein kinase and ATPase activity. *Virology* 477, 100–109. doi: 10.1016/j.virol.2014.12.036

Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560. doi: 10.1101/gr.120618.111

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., Mcanulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kamke, J., Rinke, C., Schwientek, P., Mavromatis, K., Ivanova, N., Sczyrba, A., et al. (2014). The candidate phylum Poribacteria by single-cell genomics: new insights into phylogeny, cell-compartmentation, eukaryote-like repeat proteins, and other genomic features. *PLoS ONE* 9:e87353. doi: 10.1371/journal.pone.0087353

Lange, S. J., Alkhnbashi, O. S., Rose, D., Will, S., and Backofen, R. (2013). CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res.* 41, 8034–8044. doi: 10.1093/nar/gkt606

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, C. W., Chen, J. Y., and Hua, T. E. (1998). Precambrian sponges with cellular structures. *Science* 279, 879–882. doi: 10.1126/science.279.5352.879

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J., Charpentier, E., Horvath, P., et al. (2011). Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* 9, 467–477. doi: 10.1038/nrmicro2577

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* 13, 722–736. doi: 10.1038/nrmicro3569

Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2013). Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* 41, 4360–4377. doi: 10.1093/nar/gkt157

Maldonado, M., Ribes, M., and Van Duyl, F. C. (2012). Nutrient fluxes through sponges: biology, budgets, and ecological implications. *Adv. Mar. Biol.* 62, 113–182. doi: 10.1016/B978-0-12-394283-8.00003-5

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., et al. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 43, D222–D226. doi: 10.1093/nar/gku1221

Markowitz, V. M., Chen, I. M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., et al. (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* 40, D115–D122. doi: 10.1093/nar/gkr1044

Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Moitinho-Silva, L., Seridi, L., Ryu, T., Voolstra, C. R., Ravasi, T., and Hentschel, U. (2014). Revealing microbial functional activities in the Red Sea sponge Stylissa carteri by metatranscriptomics. *Environ. Microbiol.* 16, 3683–3698. doi: 10.1111/1462-2920.12533

Nayfach, S., and Pollard, K. S. (2015). Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.* 16:51. doi: 10.1186/s13059-015-0611-7

Oliveira, P. H., Touchon, M., and Rocha, E. P. (2014). The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* 42, 10618–10631. doi: 10.1093/nar/gku734

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 12:385. doi: 10.1186/1471-2105-12-385

Peng, Y., Leung, H. C., Yiu, S. M., and Chin, F. Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174

Reichenberger, E. R., Rosen, G., Hershberg, U., and Hershberg, R. (2015). Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. *Genome Biol. Evol.* 7, 1380–1389. doi: 10.1093/gbe/evv063

Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J., et al. (2003). A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* 31, 1805–1812. doi: 10.1093/nar/gkg274

Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2015). REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* 43, D298–D299. doi: 10.1093/nar/gku1046

Ryu, T., Seridi, L., Moitinho-Silva, L., Oates, M., Liew, Y. J., Mavromatis, C., et al. (2016). Hologenome analysis of two marine sponges with different microbiomes. *BMC Genomics* 17:158. doi: 10.1186/s12864-016-2501-0

Sorokin, V. A., Gelfand, M. S., and Artamonova, I. (2010). Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. *Appl. Environ. Microbiol.* 76, 2136–2144. doi: 10.1128/AEM.01985-09

Southwell, M. W., Weisz, J. B., Martens, C. S., and Lindquist, N. (2008). In situ fluxes of dissolved inorganic nitrogen from the sponge community on Conch Reef, Key Largo, Florida. *Limnol. Oceanogr.* 53, 986–996. doi: 10.4319/lo.2008.53.3.0986

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E. A., Mitros, T., et al. (2010). The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* 466, 720–726. doi: 10.1038/nature09201

Taylor, M. W., Radax, R., Steger, D., and Wagner, M. (2007). Sponge-associated microorganisms: evolution, ecology, and biotechnological potential. *Microbiol. Mol. Biol. Rev.* 71, 295–347. doi: 10.1128/MMBR.00040-06

Thomas, T., Moitinho-Silva, L., Lurgi, M., Bjork, J. R., Easson, C., Astudillo-Garcia, C., et al. (2016). Diversity, structure and convergent evolution of the global sponge microbiome. *Nat. Commun.* 7, 11870. doi: 10.1038/ncomms11870

Thomas, T., Rusch, D., Demaere, M. Z., Yung, P. Y., Lewis, M., Halpern, A., et al. (2010). Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J.* 4, 1557–1567. doi: 10.1038/ismej.2010.74

Vasu, K., and Nagaraja, V. (2013). Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.* 77, 53–72. doi: 10.1128/MMBR.00044-12

Vogel, S. (1977). Current-induced flow through living sponges in nature. *Proc. Natl. Acad. Sci. U.S.A.* 74, 2069–2071. doi: 10.1073/pnas.74.5.2069

Wang, L., Chen, S., Xu, T., Taghizadeh, K., Wishnok, J. S., Zhou, X., et al. (2007). Phosphorothioation of DNA in bacteria by dnd genes. *Nat. Chem. Biol.* 3, 709–710. doi: 10.1038/nchembio.2007.39

Webster, N. S. (2007). Sponge disease: a global threat? *Environ. Microbiol.* 9, 1363–1375. doi: 10.1111/j.1462-2920.2007.01303.x

Wehrl, M., Steinert, M., and Hentschel, U. (2007). Bacterial uptake by the marine sponge *Aplysina aerophoba*. *Microb. Ecol.* 53, 355–365. doi: 10.1007/s00248-006-9090-4

Weisz, J. B., Lindquist, N., and Martens, C. S. (2008). Do associated microbial abundances impact marine demosponge pumping rates and tissue densities? *Oecologia* 155, 367–376. doi: 10.1007/s00442-007-0910-0

Wilson, M. C., Mori, T., Ruckert, C., Uria, A. R., Helf, M. J., Takada, K., et al. (2014). An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* 506, 58–62. doi: 10.1038/nature12959

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Part IV.

General discussion

Bioinformatic considerations

Genomic and metagenomic data was the backbone to most of the studies included in this thesis. In this section, the potential, application as well as advantages and disadvantages of both methods will be discussed.

Sequencing technologies have evolved fastly in the last decades from time–consuming methodologies to nearly real–time approaches. Genomic and metagenomic seqencing has somehow become a standard procedure and has revolutionized the microbiological landscape. In their simplest form, that is millions of reads, (meta–)genomic data contains a limited signal for homology searches (Wommack et al. 2008) and annotation is challenged by the availability of reference genomes. Consequently, this data is assembled into longer sequences, so called contigs. The assembly of genomic data can be monitored based on several features as genome size, GC content, single copy genes, orientation of read pairs. Including decontamination steps, genome assemblies end up in high–quality drafts as it has been shown here (Chapter 1). The presented pipeline uses all of these quality control steps. With that, it does not rely on a single metric as the N50 value, which would be a naive approach, as a single best metric does not exist (Ekblom and Wolf 2014). In addition, the described workflow can be used with multiple different assemblers to find the best fitting algorithm for each project or genome. This is an often underestimated issue, as some assemblers e.g. discard repetitive sequences, misjoin paired–end reads or "[p]erform beautifully on simulated data but fall down on actual data" (Baker 2012). With that, the workflow is also prepared for long–read sequencing projects as the assembly software can be easily exchanged. That is of high importance, as the algorithms are fundamentally different for short and long reads (Ekblom and Wolf 2014). Hence, genomes of high contiguity as well as low error rates can be guaranteed with the approach given here, also using latest technologies from Pacific Biosciences or Oxford Nanopore. However, the assembly of metagenomic data is much more complex due to the high intrinsic variation (Thomas et al. 2012). But once accomplished, it leads to full–length coding sequences (or ORF) and improves functional and taxonomic annotations (Huson et al. 2007). This was especially useful for the detection of long and complex genetic elemens (Thomas et al. 2012) as CRISPR (Chapter 5). There are also tools to predict CRISPR from unassembled reads (Skennerton et al. 2013), but they have been shown to highly underestimate their numbers (Gogleva et al. 2014).

Another challenge in metagenomic data is the enumeration of annotated coding sequences. In a single genome, one annotated coding sequence relates to a count of one. For metagenomics, there exist several methods to do a normalization of the count data. The easiest approach is dividing the counts of a specific gene X by all reads used, thus obtaining a relative abundance. Further, the counts of gene X can be divided by the number of counts for all 16S rRNA sequences, which is hardly the copy per genome. The last approach is biased, as 16S rRNA sequences exist in different copy numbers in bacteria (Větrovský and Baldrian 2013). For six sponge metagenome samples Fan et al. (2012) calculated average genome sizes based on multiple single copy genes, using this to infer copies per genome to compare different metagenomic datasets. In this thesis, I normalized the metagenomic data to copies per megabase. I propose this to be more robust than the method employed by Fan et al. (2012) as it does not rely on single copy genes which may (or may not) be uniformly sequenced among different metagenomes. There are also methods to circumvent normalization. Binning is a process, in which DNA is sorted into groups representing individual genomes (Thomas et al. 2012). This approach has successfully used before (Burgsdorf et al. 2015), but also has some drawbacks. Mostly, only the highest abundant organisms can be detected by binning, thus it is accompanied by loss of information towards the usage of full metagenomic data. Moreover, this process may not end in single genomes, but rather a consensus of similar ones. Thus events as HGT may hardly detected in that kind of data. The use of standard protocols — e.g. as used in the Earth Microbiome Project for 16S rRNA analysis — from sampling to sequencing would also help comparing different metagenomic datasets and reduce the need for different normalization steps.

By application of assembly and normalization, four metagenomic datasets were compared in this thesis for secondary metabolite genes. Further, a genomic–based approach and an approach based on the amplification with degenerated primer was conducted to detect and sequence biosynthetic genes. All of the presented methods are homology–based and circumvent the need for the expression of a biosnythetic gene cluster. The most common technique is the amplification of genes of interest using degenerated primers. This method is easy in use, fast in identifying, cataloging and quantifying. This method can be used on cultivated strains as shown here, but can also be applied to DNA directly isolated from the environment (Wilson and Piel 2013). Both, the genomic and metagenomic method rely on *in silico* predictions of biosynthetic genes and are — as the amplification approach — limited to the detection of previously known biosynthetic genes based on sequence similarities (Wilson and Piel 2013). However, genomic and metagenomic mining offer the ability to investigate the genomic proximity of identified genes and the identification of complete biosynthetic gene clusters. In addition, an abundance of PKS genes was detected with all thress methods, with some of them not recognized before. Furthermore, the investigation of the cultured isolates has shown their biosynthetic potential. In many cases, the isolation of secondary metabolites is lower compared to the genomic potential. This has been shown here, but also for well–studied strains as *Streptomyces coelicolor* A3 (Bentley et al. 2002). Despite the emergence of many bioinformatic tools and databases, it remains still a challenge to link biosynthetic genes or even clusters to chemical products. Thus, the complementation of sequencing methods to obtain genes or gene clusters with their heterologous expression using vectors can lead to the identification of natural products. Thus, the predictions obtained from all three approaches here are a good starting point towards this goal. Further, the use of transcriptomics aside to genomics in bioactivity screenings may help to not only identify possible biosynthetic gene clusters, but also which components are active.

The fragmentation of the metagenomic contigs was high within this study due to the use of Illumina short read technology and the complexity of the sponge microbiome. With the upcoming technologies by Pacific Biosciences and Oxford Nanopore Technologies, metagenomic as well as other omic–studies will benefit from longer reads. Biosnythetic gene clusters may be found on a single reads and single genomes will be closed. In summary, all analysis demonstrated the high value of genomic and metagenomic data to investigate the functional roles of microorganisms in the phyllosphere and in sponges.

Construction of a draft genome for the phyllosphere bacterium *Williamsia* sp. ARP1

Phyllosphere microbial communities are diverse and include many different fungi, yeasts, protozoa, and different genera of bacteria (Lindow and Brandl 2003). Main research was initially driven by plant–pathogens as *Pseudomonas syringae* or *Erwinia* spp. (Lindow and Brandl 2003), but also strains as *Panotoea agglomerans* (Remus-Emsermann et al. 2013) or different *Methylobacteria* (Knief et al. 2012a) were investigated and many colonizing microbes were found to be commensal symbionts (Müller and Ruppel 2014).

Here, strain ARP1 of the genus *Williamsia* was analysed. Most information about this genus was based on morphological and chemical features, and only 9 strains were described with valid names. Since the initial discovery in 1999 (Kämpfer et al. 1999), only two genomic sequencing studies were published, including the strains D3 and ARP1 (Guerrero et al. 2014; **Horn** et al. 2016a). To our knowledge, *Williamsia* sp. ARP1 is the first sequenced strain from the phyllosphere. To analyse the genomic potential of this phyllosphere bacterium, a sequencing approach of a culture isolated from an *Arabidopsis thaliana* leaf was conducted in this thesis. The presented study is the first to compare and characterize members of the genus *Williamsia* on a genomic level (Chapter 2). Below, I will discuss how the strain was chosen and the experiments behind, the genomic repertoire, adaption to its phyllospheric habitat, phylogeny and the contribution of this study to the knowledge about this genus.

Abundance, diversity and sources

Williamsia species have been isolated from various sources, ranging from human blood (Yassin and Hupfer 2006), glacier ice, deep sea (Pathom-Aree et al. 2006a), hay meadows (Jones et al. 2010), antarctic soil (Guerrero et al. 2014), building material, the leaf surface of white clover, and the phyllosphere of *Arabidopsis thaliana* (**Horn** et al. 2016a). Despite, only few cells and sequences were observed using 16S rRNA, metagenomic or culturing studies, e.g. in a foaming activated sludge community (Guo et al. 2015), in the roots of wheat (Conn and Franco 2004), the Mariana Trench (Pathom-Aree et al. 2006b), the rare soil biosphere (Shade et al. 2012), or more recently as “keystone genera” in deep subsurface fracture fluids (Purkamo et al. 2016) and from a high radiation environment in Chernobyl (Ruiz-González et al. 2016). Remarkably, the genus seem to be mostly observed in extreme environments and may thus be considered versatile in its adaptions to different lifestyles. These can be described as extremely cold (antarctic and glacier ice), under high pressure and dark (deep sea, Mariana Trench), under high radation (Chernobyl) or short–lived (leaf surfaces, phyllosphere). Interestingly, none of the publications describes the genus as highly abundant, and also within the major databases as NCBI or SILVA, only 453 and 390 nucleotide sequences are deposited related to *Williamsia* (as of July 11, 2016). On the data available, the genus can be assumed as low abundant in microbial populations, called the rare biosphere. Based on 16S rRNA the described strain ARP1 was clearly assigned to the genus *Williamsia*. Despite the low amount of sequences in the databases, similarities of $\geq 99\,\%$ to three *Williamsia* sequences derived from sediment were calculated. But, closest type strains, *Williamsia maris* DSM 43672 and *Williamsia phyllosphaerae* C7 exhibited similarities of only 98.3 % and 98.5 %, indicating a novel species.

Further investigation are needed to reveal abundance patterns of *Williamsia* species in different environments, either using deep sequencing strategies or designed primers to target this genus. This would also be helpful to explore the diversity of this genus along with its eveness and richness in different habitats. Moreover, marker sequences can be deposited in the databases. These, in turn, would help to identify *Williamsia* species in other studies.

Phylogeny and assignment

Since the recognition of microorganisms, scientists tried to classify them in a evolutionary and phylogenetic context (Clarke 1985). This has proved challenging as a species is defined as a group of organisms which can interbreed (Mayr 1940), a concept that can not be applied to asexual bacteria (Varghese et al. 2015). Hence, prokaryotic species have been defined based on multiple features as morphology, (gram–)staining, and metabolism (Kitahara and Miyazaki 2013). The first method to evaluate relationships was DNA–DNA hybridization (DDH) (Richter and Rosselló-Móra 2009) and became the gold standard using. It was replaced by sequence–based methods using marker genes such as the 16S rRNA (Stackebrandt and Goebel 1994) and cut–off values of 97 % to define a species. Today, the genome sequence is seen as the ultimate taxonomic information about a microbial strain (Kim et al. 2014). Thus, there is potential to replace the aforementioned methods by using whole–genome similarities such as ANI to circumscribe prokaryotic species. A cut–off in the range of 95–96 % was proposed to define a species (Richter and Rosselló-Móra 2009).

The obtained strain ARP1 was classified based on 16S rRNA gene phylogeny, ANI and pairwise best hit similarities using only coding genes with COG annotation. Using the 16S rRNA and the construction of a phylogenetic tree (Chapter 2, Figure 1), the strain was assigned clearly to the genus *Williamsia*. The assignment of phylogeny with genome–based methods led to a mixture with the genus *Gordonia*. The ANI method revealed a simlarity of 72.37 % between strain D3 and ARP1, but slightly higher similarities between strain ARP1 and the *Gordonia* species with 72.95 % and 72.71 % (Chapter 2, Table S1). A pairwise comparison of orthologous genes offered a similarity of 75.53 % between the two *Williamsia* strains. The same analysis comparing them to the *Gordonia* strains revealed similarities between 71.50 % and 75.15 % (Chapter 2, Figure 5). Thus, a discrepancy between marker gene and whole–genome analysis was observed. As the 16S RNA gene is not amenable to HGT (Kitahara and Miyazaki 2013) due to the *complexity hypothesis* (that is, "[m]aking horizontal transfer of informational gene products less probable") (Jain et al. 1999), it is unlikely to be transfered between species. Consequently, it can be considered to reflect true phylogeny and origin of a strain or cell. This is also supported by results from a recent project, in which I participated: strain ARP1 was investigated for its chemotaxonomix profile (Kämpfer et al. 2016, submitted), including mycolic acids, quinones, and polar lipids. These were in agreement with the description for the genus *Williamsia* (Kämpfer et al. 1999). One could think of artifacts due to the small sample size (2 *Williamsia* genomes) or low quality of sequencing data and assembly, which led to the low ANI scores and identities based on orthologs. But, latest studies revealed similarities based on ANI in a range of 75.4 % to 90.6 % for six available type strains of the genus *Williamsia* (Kämpfer et al. 2016, submitted). All of them clearly were below the proposed species boundary. Calculating this score between strain ARP1 and the recently available genome of *Williamsia* sp. Leaf354, an identity of 94.1 % is reached. This score and the overall range of ANI scores mostly below 80 % are indicative for a genus, which is diverse on the genomic level, potentially based on HGT. Moreover, isolates from the same source are more similar (ARP1 and Leaf354 were both isolated from *A. thaliana*) than compared to all other isolates. From this, one could hypothesize a specific gene gain or loss depending on the habitat of a strain in an adaptive manner.

Genomic functions and adaptations to the phyllosphere

As described, the phyllosphere have to adapt to a harsh environment (Rastogi et al. 2013). Among different studies, several traits have been proposed that enable bacteria to colonize this environment. One of the major determinants for leaf colonization is the availability of nutrients. In particular, carbon compounds are believed to be limited (Lindow and Brandl 2003). In the presented genome sequence, ABC transporters for the uptake of ribose, glycerol and fructose were identified. Specifically the photosynthate fructose is a known compound which is leached from the interior of the plant (Lindow and Brandl 2003). In addition, transporter mediating the uptake of amino acids and maltose were found. Whereas amino acids suggest the usage of nitrogen from plant–derived compounds, the latter indicates the utilization of disaccharides and was also detected in *Pseudomonas* species from the phyllosphere (Delmotte et al. 2009). The leaves of plants are exposed to fluctuating temperatures and relative humidity (Lindow and Brandl 2003). Genes enabling strain ARP1 to react on these shifts were heatshock chaperones, cold shock proteins as well as genes encoding for the osmoprotectants trehalose and betaine. Pathways for the biosynthesis of trehalose were also found in *Pseudomomnas* and *Methylobacterium* strains in the phyllosphere of clover and soybean, but not in *A. thaliana* (Delmotte et al. 2009). The synhetesis of betaine was also associated with phyllospheric fitness in bacteria as *Pseudomonas syringae* and *Pantoea agglomerans* (Farrer et al. 2009; Remus-

Emsermann et al. 2013). Other genes reported to be abundant in the bacteria of the phyllosphere (in comparison to the rhizosphere) (Knief et al. 2012b) were also detected in *Williamsia* sp. ARP1. These include genes involved in response to the plant defense as reactive oxygen species. Phyllosphere bacteria are exposed to sunlight. Strain ARP1 offered phentoypical adaptions through its red pigmentation (Chapter 2, Figure 1), genomically supported by detection of all genes in the carotenoid pathway. Moreover, genes for the synthesis of mycosporins to absorb UV light and DNA repair mechanisms as the UvrABC system were revealed. Several other adaptions as the production of biosurfactants or flagella were not identified, but might not be necessary for the investigated strain to survive in the phyllosphere.

Noteworthy, other environments offer similar conditions in terms of temperature and dessiccation (Lindow and Brandl 2003), UV radiation or nutrient avaliability. Thus, many of the found adaptions may also be transferred to other habitats. But, the number of found traits and the overlap to other studies implies them to be important for the leaf colonization and subsequent survival. If the genomic inventory of *Williamsia* sp. ARP1 was adapted over the years towards the phyllospheric lifestyle or if it can be considered the core genome of *Williamsia* remains open. Sequencing more genomes of this genus would help to investigate their genomic content and shed light on shared genes between strains from the same or different habitats, identifying such specific adaptions. Also, a core genome may be determined and the phylogenetic position of the genera *Williamsia* and *Gordonia* will be clarified. Following, the discrepancy between single marker genes and the whole genome can be explored and the phylogenomic position of genus *Williamsia* can be revisited.

Genomic mining of sponge-associated bacterial isolates

The search for new secondary metabolites and natural products is in focus for medical treatment as pathogenic microorganisms are considered a *major global challenge* for public health (Machado et al. 2015; Weber and Kim 2016). Historically, they have been the main source for antibiotics, but are also of interest in anti–cancer research, for insecticides, or crop protection among others (Weber and Kim 2016). So far, ≥ 80 000 natural products have been isolated from microorganisms (Bérdy 2012) and screening was mostly focused on soil microbes, in particular *Streptomyces* species (e.g. Li and Walsh 2010; Yu et al. 2013). In recent years, other environments were investigated and the marine habitat was found to be an underexplored ressource for novel compounds and chemical classes (Machado et al. 2015; Xiong et al. 2013) produced by microbes from different habitats as molluscs, algae and sponges. Mainly, the new identified natural products are mostly isolated from sponge–associated actinomycetes (Abdelmohsen et al. 2014a, 2015) and are produced either by PKS or NRPS (Xiong et al. 2013). Further, secondary metabolites from sponges or their associated microorganisms are known to mediate microbe-host as well as microbe-microbe interactions (Cimermancic et al., 2014;Hardoim and Costa, 2014). But, despite developments in combinatorial chemistry, novel drugs were not provided in expected percentage (Newman and Cragg 2007). In recent years, studies on the discovery of new molecules have benefited from the development in DNA sequencing technologies and led to publicly available genomic and metagenomic data and their subsequent screening using ever new bioinformatic tools (Weber and Kim 2016 and cited references).

Using culture–dependent sequencing, the whole genomic repertoire of single strains is available, thus making the identification of genes related to secondary metabolism a straightforward process. Presented in this thesis are six genomes belonging to the phylum Actinobacteria, isolated from different sponge species (Chapter 3 and 4). They were screened using antibiotics & Secondary Metabolite Analysis SHell (antiSMASH) (Weber et al. 2015) and NaPDoS (Ziemert et al. 2012) in a complementary way. The analysis revealed numbers of 4 to 42 diverse biosynthetic gene clusters for the six genomes. The results are in accordance with earlier studies, which revealed 2 to 20 for marine gram–negative bacteria (Machado et al. 2015) and between 1 and 70 cluster in actinomycetes from different habitats and are mostly distributed among Type I PKS, NRPS and terpenes (Doroghazi and Metcalf 2013). Isolate *Micromonospora* sp. RV43 exposed with 20 different gene cluster the highest diversity, followed by *Streptomyces* sp. SBT349 with 14, *Nonomuraea* sp. SBT364 with 11, the two *Nocardiopsis* strain with each 10 and *Rubrobacter* sp. RV113 with 2. Complemented with the abundance of gene clusters among the genomes, *Micromonospora* and *Streptomyces* offered again highest numbers. This is consistent with earlier studies (Doroghazi and Metcalf 2013) and as both genera were massively sampled and investigated, a discovery of biosynthetic gene cluster is more likely compared to all others. Noteworthy, in all six isolates two to five clusters associated to terpenes were detected. It was hypothesized, that terpenes may act as communication and defense mechanism or "safeguard for organsims in the marine world" (Paul et al. 2011). Similarly, Type I PKS, NRPS, lantipeptide and siderophore cluster were abundant in all but one genome: *Rubrobacter* sp, RV113. Siderophore–dependent iron transport is a feature linked to sponge–associated bacteria (Burgsdorf et al. 2015) and can be considered an adaption to their lifestyle.

Possible products for the gene cluster were predicted. Identified compounds were nystatin, rapamycin, epothilone, tetronomycin and rifamycin among others, but overall with similarities ranging between 26 % to 85 % and only based on single ketosynthase domains rather than complete gene cluster. From another genome mining study, it has been concluded, that results infered from homology searches like BLAST can be misleading, as parts of a gene cluster (e.g. domains) can be shared between different clusters. Thus, it is necessary to know investigate the complete architecture of a biosynthetic gene cluster (Aleti et al. 2015). But, using a genomic approach offers the possibility to scan genomes for their capabilities towards secondary metabolism and can highlight strains, which can be of biotechnological or pharmaceutical interest. Overall, this method can pave the way for future, bioassay–guided methods to isolate natural products.

Metagenomic mining of sponge-associated bacterial consortia

The composition of microbes is defined as their identity and relative abundance in an ecosystem (Reed and Martiny 2013). The knowledge of their distribution among different environments will help to understand these organisms and underlying processes shaping their community structure. Further, it helps to explain the symbiosis between a host and its associated microbes. Marine sponges are known to be inhabited by dense and diverse microorganisms (Hentschel et al. 2012) such as viruses, protozoa, archaea, fungi, and bacteria (Webster and Taylor 2012). Bacterial phyla in HMA sponges have been found to be the most prevalent organisms with the most common of these being *Proteobacteria*, *Cyanobacteria*, *Nitrospirae*, *Actinobacteria* and *Chloroflexi* (Hentschel et al. 2012). Usually, amplicon sequencing of a hypervariable region of the 16S rRNA gene is the applied approach to infer community composition, its organisation or spatiotemporal patterns (Sinclair et al. 2015). In this thesis, metagenomic sequencing and subsequent analysis of all obtained reads via the MG–RAST platform was performed. Therefore, this analyis is refered to as genomic composition rather than community composition.

All three investigated sponge samples, *Petrosia ficiformis*, *Saroctragus foetidus* and *Aplysina aerophoba*, are examples of HMA sponges. Their genomic composition is, with some exceptions, comparable to taxonomic profiles of microbial sponge–communities (see Thomas et al. 2016). Chloroflexi and Cyanobacteria are found to be more abundant, Firmicutes and Actinobacteria to be less abundant in community analysis compared to our metagenomes. In both studies, seawater samples exhibited highest abundances for Alpha– and Gammaproteobacteria. Also, the other phyla seem to be in congruence, but, Bacteroidetes had higher abundances in our study. Despite differences in the used approaches, there are overlaps in the composition. Variations may be explained by the fact, that the metagenomic analysis incorporates all reads and thus the genome sizes of the underlying organisms, thus highlighting the Actinobacteria in the *S. foetidus* sample with an relative abundance $\geq 20\,\%$. Both, the 16S rRNA and the metagenomic approach are biased by several sources comprising introduced by sequencing errors (regardless of technology), the error rates of polymerases or the formation of chimeras in heterogenous samples (Schloss et al. 2011). In a recent study, several tools, including MG–RAST, were tested for their capability to infer taxonomic composition by use of metagenomic reads, but all of them differed significantly from the original data (Lindgreen et al. 2016). Also, the assignment of taxonomy to reads is influenced by the tool used (Peabody et al. 2015) and so is for the analysis of 16S rRNA sequences (Schmidt et al. 2015).

The presented data may not reflect the true taxonomic composition. However, metagenomics is not limited to a single marker gene. Highly diverged microbes and also non–microbes may be missed by 16S rRNA studies (Sharpton 2014). Metagenomic analysis has the power to show the genomic contribution of different taxa in a microbial community as a function of their genome sizes. As a consequence, their real impact is depicted as genome sizes directly infer the number of genes, an thus functionality.

Screening for secondary metabolites

The number of genes characterizing secondary metabolite gene cluster was surprisingly high in seawater and renders them not as an adaptive or selective trait or necessary defense system for bacteria to inhabit a sponge. The overall composition was similar between both environments (Figure 11), but offered differences in their taxonomic assignment. Similar to the defense systems, most genes in the sponge datasets were affiliated to Proteobacteria and Actinobacteria, whereas the seawater metagenome was dominated by the Proteobacteria and Cyanobacteria, reflecting the genomic composition. Overall, the high number of detected genes might reflect the true number of secondary metabolites in both environments, but may also be a product of short sequences/contigs, especially in the seawater dataset. This could lead to a high fragmentation of cohesive gene clusters and thus to an overestimation. As described, the composition of indicator genes for secondary metabolites was similar, with the exception of Type I PKS (Figure 11). PKS are phylogenetically classified into two groups, which differ mainly by the presence (cis) or absence (trans) of the acyl–transferase (AT) domain in the extender modules (Helfrich and Piel 2016). Here, none of the Type I ketosynthases was affiliated to trans–AT PKS, hybrid NRPS–PKS or cis–AT PKS families, which are usually involved in the biosynthesis of natural products, which is in concordance of the data published by Fieseler et al. (2007), in which only 8 % of ketosynthases were assigned to one of the named families. 111 of the 120 validated Type I ketosynthase sequences were grouped into the sponge symbiont ubiquitous PKS (supA) group which reflect the high abundance of the microbial community containing the supA genes. Previous studies have reported the supA group as specific and exclusively found in sponges (Hochmuth and Piel 2009). This group describes small monomodular and unusual polyketide synthases, which might be responsible for the biosynthesis of mid–chain–branched fatty acids for which HMA sponges are a rich source (Della Sala et al. 2014; Fieseler et al. 2007; Hochmuth et al. 2010). Closest hits to our found PKS sequences were found in symbionts of *A. aerophoba* and *T. swinhoei* in earlier studies (Fieseler et al. 2007; Piel et al. 2004), thus suggesting the distribution of either similar Type I PKS systems or bacterial symbionts in different sponge species in different geographic locations. All found ketosynthases were assigned to symbionts with unknown taxonomic origin. Results with a lower e–value offered hits to the genus *Mycobacterium*, which belongs to the Actinobacteria. In a previous study focusing on the sponge *Arenosclera brasiliensis* (Trindade-Silva et al. 2013), 16 % of 235 PKS sequences were found to be of actinobacterial origin, with some of them assigned as *Mycobacterium marinum*. The genus *Mycobacterium* was also found to be a rich source of highly conserved PKS gene clusters (Doroghazi and Metcalf 2013)). Thus, it could be hypothesized, that supA are hosted by Actinobacteria in sponges, which are a well–known source for bioactive compounds (Subramani and Aalbersberg, 2012;Abdelmohsen et al., 2015;Cheng et al., 2015) and found in high abundance in our samples. Supporting this, only one ketosynthase was found in the seawater sample assigned to *Achromombacter xylosoxidans*, a Betaproteobacterium, and only a small abundance (1.06 %) of Actinobacteria. This ketosynthase was classified in the cis–AT PKS, and therewith confirms, that most sponge–associated PKS are distinct from PKS in free–living or non–sponge–associated microbes (Hochmuth and Piel 2009) (Chapter **??**, Figure 12). Moreover, the large variety of found type I PKS enzymes reflects the potential of metagenomics to be used as a backbone for novel polyketide natural product research.

Secondary metabolite genes

The most abundant secondary metabolite marker genes belonged to the groups of saccharides, bacteriocins, terpenes and fatty acids (≥ 0.1 cpm). Other indicator genes of secondary metabolism – linaridin, lantipeptides, ectoines, phosphonates, proteusin, polyketide synthases, nucleosides, microcins, siderophore or homoserine lactones - were only found in low copy numbers (≤ 0.02 cpm). Interestingly, while siderophores and homoserine lactone hits were only identified in seawater, lantipeptides, linaridines, and Type I Polyketide synthases – with the exception of one Type I PKS in seawater - were only found in the sponge metagenomes. In agreement with the calculated Bray–Curtis Dissimilarity based on genomic content (Chapter 5, Figure 1), the seawater sample is most dissimilar to all sponge samples with at least 21.03 % (Figure 11).

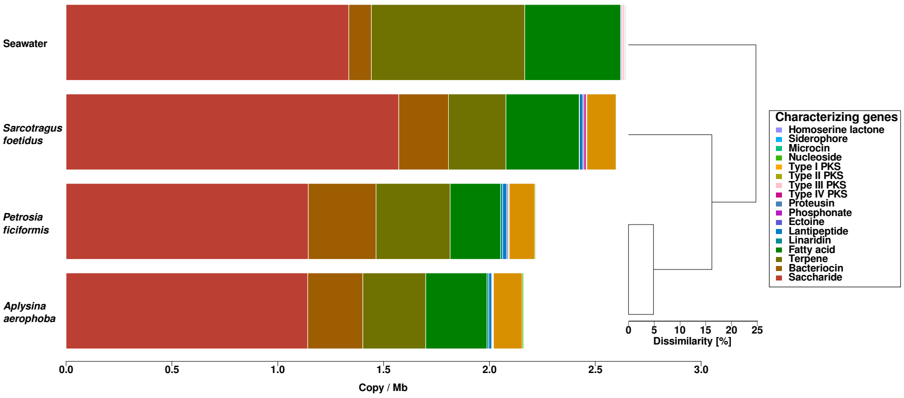


Figure 11. – Barplot of characterizing secondary metabolite genes for the four analyzed metagenomes. The x–axis indicates the copy per megabase metagenome for all characterizing genes. Functional dissimilarities (Bray–Curtis) are indicated by the dendrogram

Type I PKS phylogenetic tree

A total of 120 Type I PKS genes were identified in the three sponge metagenomes, 23 derived from P. ficiformis, 36 from S. foetidus, 1 from seawater and 60 from A. aerophoba. The sequence found in seawater was assigned to the cis-AT type I PKS system and the bacterial strain Achromobacter xylosoxidans, a betaproteobacterium. Phylogenetic analysis assigned the majority (109/120) to the symbiont ubiquitous supA-type PKS group. Most similar sequences from the sponge metagenomes according to the phylogenetic tree (Figure 12 A) were previously discovered and found to be bacterial symbionts associated to the sponges *Theonella swinhoei*, *Aplysina aerophoba* and *Discodermia dissoluta*(*Theonella swinhoei* bacterial symbiont clone pSW1H8 (ABE03935), *Aplysina aerophoba* bacterial symbiont clone pAPKS18 (ABE3915) and pAE27P20 (ABE3895), symbiont of *Discodermia dissoluta* (AAY0025-0027)). Further taxonomic analysis using blastp affiliated the supA-type PKS mostly to the genus *Mycobacterium* (Supplementary file 7). Only 11 non-supA-type PKS sequences were identified which fell into a FAS-like PKS cluster and were most similar to symbionts from the sponge *Plakortis simplex* (bacterium symbiont of *P. simplex* pPSA11D7 and pPS11G3 (aGH13590, aGH13577)). Nearest similar classifications indicated the Actinobacterium *Sciscionella marina* (WP020497474) as a possible host for the FAS-like PKS. Adding functions and possible products to the found PKS, we submitted them to NaPDoS. Most of the polyketide synthases in the supA clade of the tree resulted in a hit to epothilone, but with sequence identities ranging from only 38 % to 62 %. Despite the variance of possible products in the FAS-like PKS clade, the order of the genes surrounding the polyketide synthase was highly conserved (Figure 12 B).

Defense systems

Phages have a far–reaching impact, ranging from food industries to human health and nutrient cycling. Wherever we can find bacteria, there will also be phages (Seed 2015). Considered the most powerful driving force in evolution is the arms race between phages and their respective host (Stern and Sorek 2011), also considered a “struggle for existence” (tenOever 2016). Thus not surprising, many different defense systems have evolved in bacteria requiring large parts of their genomes (Makarova et al. 2011). A review by Makarova et al. (2013) has expanded our view and knowledge of these defense systems which are considered analogs to the adaptive and innate immunity in eukaryotes. In Chapter 5, the metagenomes of three sponges — *P. ficiformis*, *S. foetidus*, *A. aerophoba* — and from seawater were investigated towards the genomic inventory of bacterial defense systems. A broad range of genes mediating immunity were detected and found to be higher abundant in sponge microbiomes than in the seawater microbiome. Below, different immune systems will be shortly explained and discussed.

The CRISPR system is build of an array comprising spacer sequences (25 bp – 70 bp) alternating with direct repeats (24 bp – 48 bp). Spacer sequences are integrated sequences

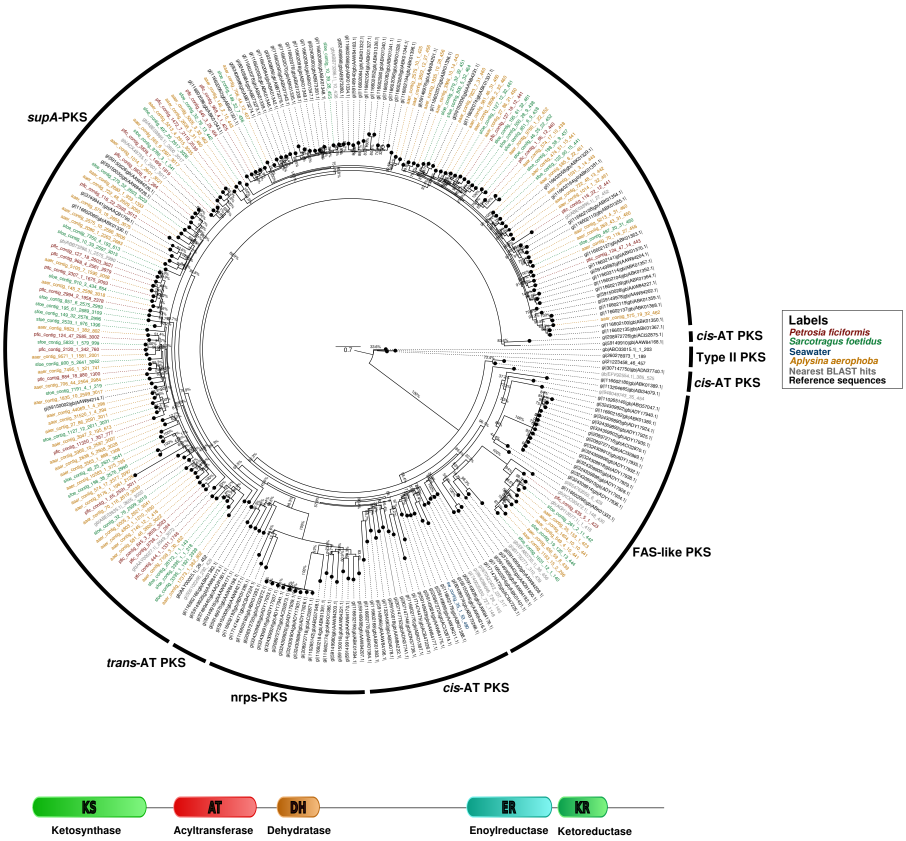


Figure 12. – (A) Phylogentic tree of extracted Type I polyketide ketosynthases from the analysed metagenomes and added reference sequences obtained from NCBI. Labels are colored according to their origin, possible products are marked as colored nodes. (B) Conserved structure of the FAS-like PKS cluster found in the three sponge samples

of foreign DNA. Further, it includes *cas*–genes (Gogleva et al. 2014), which encode proteins (helicases, polymerases, nucleases) important for the activity and upon which different CRISPR types can be distinguished (Makarova et al. 2015). A CRISPR array is transcribed completely as a precursor RNA and subsequent cleaved into fragments (cnRNA), each containing a spacer sequence. These fragments are used as guide RNA to recognize and cleave incoming viruses (Koonin and Makarova 2009).

The final sets of CRISPR arrays was 77 (0.21 copies per megabase (cpm)), 47 (0.25 cpm), 283 (0.62 cpm) and 0 (0 cpm) for the metagenomes of *P. ficiformis*, *S. foetidus*, *A. aerophoba* and seawater, respectively (Chapter 5, Table 3, Figure 5). A considerable fraction was found adjacent to *cas* genes and indicates these arrays as complete with counts of at least 26 (35.06 %), 20 (42.55 %) and 138 (46.78 %) for *P. ficiformis*, *S. foetidus*, *A. aerophoba*. Interestingly, not a single array was validated for the seawater dataset which is in accordance to the metagenomic study by Fan et al. (2012). Based on the *cas* genes, Type I was identified as the most abundant, followed by Type II and III for all sponge metagenomes. Similar results were found in three human gut metagenomes (Gogleva et al. 2014) and for most bacterial and archaeal genomes deposited in NCBI (Burstein et al. 2016). Thus, the distribution of CRISPR types seems to follow a similar pattern among different bacterial organisms and environments. Despite the high number of associated spacer sequences, none of them was shared between the microbiomes. This effect was also observed in two geographically distant sludge bioreactors for which no spacer sequences overlapped (Kunin et al. 2008). They concluded, that the investigated bacterial strains are able to disperse globally, but have to adapt to local phage pressure. For the sponge microbiome, this suggests an individual acquisition of spacer between bacterial cells (Gogleva et al. 2014) and/or massive differences in the local virome of the sponges (Fan et al. 2012). This concept would also apply to the distribution of spacer origins as they were found to be similar for all three metagenomes, decreasing from Proteobacteria to Actinobacteria, Chloroflexi and Firmicutes. These are the major lineages within the metagenomes suggesting an even distribution of CRISPR in the different phyla, but again with variations based on the local virome. Targets of spacer sequences were explored and the majority was assigned to unknown targets followed by plasmids and phages/viruses, implying that spacer sequences were mostly novel or unseen. This can be explained, as little to no studies focus on viral communities or even metagenomes. Comparable to the spacer sequences, only a few (up to nine) direct repeat sequences were shared between all three sponge microbiomes, despite the close geographic proximity at least for the samples of *P. ficiformis* and *S. foetidus*.

Due to the scarce overlap of spacer and repeat sequences, CRISPR arrays can be considered to be fast evolving and highly adaptive. Moreover, the absence of CRISPR arrays in seawater marks this system as an inevitable adaption to the host–sponge and demonstrate the potential of the CRISPR system to observe the co–evolution between bacteria and their attacking phages.

The RMS is an important component of the prokaryotic defense mechanimsms. Its activity is based on two enzymes: a restriction endonuclease and a methyltransferase. Whereas the methyltransferase is responsible for discrimination of self and nonself DNA by methylation of the own DNA, the restriction endonuclease cleaves foreign – non–methylated – DNA (Vasu and Nagaraja 2013). The RMS can be classified based on cleavage position, cofactor requirements, recognition sides and subunit composition (Roberts et al. 2003). Investigated here (Chapter 5) were RMS Type I (protein complex with methylation and restriction activity), Type II (separated methylation and restriction subunits), and Type III (heterotrimer or heterotetramers of methylation and restriction subunits).

In accordance to previous studies comparing sponge–associated to seawater or free–living relatives (Fan et al. 2012; Tian et al. 2016), the RMS was higher abundant in the sponge datasets compared to seawater. Type II was found to be most abundant. This is not surprising, as this type is the most widely studied (Vasu and Nagaraja 2013) and most common in prokaryotes (Oliveira et al. 2014). Whereas the RMS in seawater seems to be restricted to a few phyla, nearly all phyla contain Types I–III RMS in the sponge microbiota. Their distribution is in concordance with the genomic composition and, as for the CRISPR system, suggesting an even distribution of this system among sponge–associated bacteria. Burstein et al. (2016) reported the presence of the RMS in phyla lacking the CRISPR system due to little viral predation. Vice versa, Dupuis et al. (2013) showed the co–occurence of both systems as our study did, but may be due to high viral abundances in sponges. Thus, the two defense systems seem to work in a synergistic way complementing each other, hypothesizing, both are adaptive traits towards the sponge as a habitat. Further, the RMS has evidence to be involved in genome rearrangements and evolution of endosymbionts (Rocha et al. 2001) and might play a role in the co–evolution of phages and bacteria.

In this study, two more defense mechanisms were identified: the phage growth limitation (PGL) and the DNA phosphorothioation (DND) system. The DND works similar to the RMS, whereas PGL may be capable of inhibiting phage growth and avoid their proliferation (Makarova et al. 2013). Phages and bacteria controlling each others numbers: phages by constantly attacking, and bacteria by strategies to resist and thus maintain a dynamic equilibrium (Seed 2015) in their ongoing warfare. The diversity of at least four different defense mechanisms in sponge–associated microbiota suggests them an important trait. Possibly, they are functionally coupled making them more effective against phage attacks or work as a "backup solution" if one system fails. Even more defense systems are known for prokaryotes, including the toxin–antitoxin system or abortive infection and subsequent cell death. Also, many of the genes mediating defense are found in genomic proximity, so called genomic islands (Makarova et al. 2013). A property, which was not investigated due to the fragmentation of contigs in the underlying metagenomes. Whatsoever, this study gives deep insights on the interplay between baceria and phages and the possible obligate adaptions of bacteria to the sponge host. It is also a starting point for studies including the viral metagenome and the sponge to explore the whole system from a holobiontic side.

Part V.

Conclusion and perspectives

Overall, genomics and metagenomics are useful in detecting a high diversity of biosynthetic genes which can be used in expression experiments. Genomics may not always answer all questions, but may also raise new issues, for example, the phylogenetic discrepancy between 16S rRNA genes and whole genome sequences in *Williamsia* strains. But, genomics helped to understand adaptions to the phyllosphere environment and gave insights into the rare genus *Williamsia*. Metagenomics have been shown to be a powerful tool to detect adaptions to the sponge and surrounding seawater, thus giving a direction for subsequent studies. The complementation of –omics studies with an experimental framework, e.g. modulating the microbial community, even more insights into general properties of host–microbe and microbe–microbe interactions can be gleaned. Further studies on –omics data may be supported by these concepts. Perspective from my thesis are as follows:

State–of-the–art sequencing technologies and comparable protocols The use of Third–Generation sequencing technologies as established by Pacific Biosciences or Oxford Nanopore Technologies an its very long reads combined with Second–Generation sequencing of high quality may circumvent the need for assembly or at least simplify this process. With this, closed genomes will be available. Also, the binning of complete single genomes from metagenomic data are possible. The usage of these technologies in combination with sampling, library preparation and sequencing based on identical protocols will help to make the comparison of such data easier in the future.

The analysis of more *Williamsia* genomes In addition to the 2 published *Williamsia* genomes (strains ARP1 and D3), the genome of *Williamsia phyllosphaerae* C7 was sequenced and analyzed, in which I participated. Even more are registered within the databases (e.g strain Leaf354). A comparative study would help to define the core genome of this genus and determine factors important for specific habitats as they were isolated from different sources. Further, the phylogenetic position of the genus *Williamsia* may be better resolved.

Multi–omics approaches for secondary metabolite discovery in combination with experiments Genomic approaches can solve many biological problems and are the base layer of –omic research. But, they depend on static data and thus can not represent the actual metabolism of an organism. For this reason I propose the use of multi–omics studies to cover not only the "building plan" of organisms, but also their transcriptome along with their proteome and metabolome, to investigate, what really happens. This can also help to better explore the primary and secondary metabolism and support the hunt for bioactive compounds and drug discovery. Also many of the found gene cluster within genomes and metagenomes relate to *putative* ones, saccharides or fatty acids and possibly are no secondary metabolites. Tools for finding these gene clusters are permanently updated and help finding interesting organisms. But, a cluster of genes may not be representative for specific compounds, as the synthesis of secondary metabolites is not straightforward but iterative. Thus, I suggest the use of –omic data alongside to elicitation or heterologous expression experiments to create a link between gene clusters and produced compounds.

Further investigation of the *P. ficiformis*, *S. foetidus* and *A. aerophoba* microbiomes The named microbiomes have been found to be rich in defense mechanisms as CRISPR, restriction modification and phage growth limitation among others. A sampling and extraction approach with adapted protocols may unveil also the virome within these sponge hosts. Subsequent, targets of the defense system can be revealed, which would help understand the microbe–microbe interactions. Further, these results can be used in experiments to uncover the adaption of the CRISPR system towards viral cells. Finally, it is estimated, only 1x10^{-22} % of the total DNA on earth is sequenced so far (Microbiol 2011):

$$0.00000000000000000001\,\%$$

A number, which can be considered close to zero, thus leaving still more room for upcoming sequencing projects.

Part VI.

Bibliography

Abdelmohsen, U. R., K. Bayer, and U. Hentschel (2014a). “Diversity, abundance and natural products of marine sponge-associated actinomycetes.” In: *Nat Prod Rep* 31.3, pp. 381–399. DOI: 10.1039/c3np70111e (cit. on pp. 18, 92).

Abdelmohsen, U. R., T. Grkovic, S. Balasubramanian, M. S. Kamel, R. J. Quinn, and U. Hentschel (2015). “Elicitation of secondary metabolism in actinomycetes.” In: *Biotechnol Adv* 33.6 Pt 1, pp. 798–811. DOI: 10.1016/j.biotechadv.2015.06.003 (cit. on p. 92).

Abdelmohsen, U. R., S. M. Pimentel-Elardo, A. Hanora, M. Radwan, S. H. Abou-El-Ela, S. Ahmed, and U. Hentschel (2010). “Isolation, phylogenetic analysis and anti-infective activity screening of marine sponge-associated actinomycetes.” In: *Mar Drugs* 8.3, pp. 399–412. DOI: 10.3390/md8030399.

Abdelmohsen, U. R., M. Szesny, E. M. Othman, T. Schirmeister, S. Grond, H. Stopper, and U. Hentschel (2012). “Antioxidant and anti-protease activities of diazepinomicin from the sponge-associated Micromonospora strain RV115.” In: *Mar Drugs* 10.10, pp. 2208–2221. DOI: 10.3390/md10102208 (cit. on p. 17).

Abdelmohsen, U. R., C. Yang, H. **Horn**, D. Hajjar, T. Ravasi, and U. Hentschel (2014b). “Actinomycetes from Red Sea sponges: sources for chemical and phylogenetic diversity.” In: *Mar Drugs* 12.5, pp. 2771–2789. DOI: 10.3390/md12052771.

Abdelmohsen, U. R. et al. (2014c). “Dereplication strategies for targeted isolation of new antitrypanosomal actinosporins A and B from a marine sponge associated-Actinokineospora sp. EG49.” In: *Mar Drugs* 12.3, pp. 1220–1244. DOI: 10.3390/md12031220 (cit. on pp. 17, 18).

Albertsen, M., P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen (2013). “Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes.” In: *Nat Biotechnol* 31.6, pp. 533–538. DOI: 10.1038/nbt.2579 (cit. on pp. 10, 45).

Aleti, G., A. Sessitsch, and G. Brader (2015). “Genome mining: Prediction of lipopeptides and polyketides from Bacillus and related Firmicutes.” In: *Comput Struct Biotechnol J* 13, pp. 192–203. DOI: 10.1016/j.csbj.2015.03.003 (cit. on p. 93).

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). “Basic local alignment search tool.” In: *J Mol Biol* 215.3, pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2 (cit. on pp. 27, 30).

Amann, J. (1911). “Die direkte Zählung der Wasserbakterien mittels des Ultramikroskops”. In: *Centralbl. Bakteriol* 29, pp. 381–384 (cit. on p. 9).

Andrews, S. (2016). *FastQC A Quality Control tool for High Throughput Sequence Data*. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (visited on Aug. 15, 2015) (cit. on pp. 11, 28).

Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman (2005). “At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies.” In: *Appl Environ Microbiol* 71.12, pp. 7724–7736. DOI: 10.1128/AEM.71.12.7724-7736.2005 (cit. on p. 26).

Atamna-Ismaeel, N. et al. (2012). “Microbial rhodopsins on leaf surfaces of terrestrial plants.” In: *Environ Microbiol* 14.1, pp. 140–146. DOI: 10.1111/j.1462-2920.2011.02554.x (cit. on p. 15).

Avery, O. T., C. M. Macleod, and M. McCarty (1944). “Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III”. In: *J Exp Med* 79.2, pp. 137–158. DOI: 10.1097/00003086-200010001-00002 (cit. on p. 3).

Aziz, R. K. et al. (2008). “The RAST Server: rapid annotations using subsystems technology.” In: *BMC Genomics* 9, p. 75. DOI: 10.1186/1471-2164-9-75 (cit. on p. 12).

Bai, Y. et al. (2015). “Functional overlap of the Arabidopsis leaf and root microbiota.” In: *Nature* 528.7582, pp. 364–369. DOI: 10.1038/nature16192 (cit. on p. 15).

Bais, H. P., T. L. Weir, L. G. Perry, S. Gilroy, and J. M. Vivanco (2006). “The role of root exudates in rhizosphere interactions with plants and other organisms.” In: *Annu Rev Plant Biol* 57.1, pp. 233–266. DOI: 10.1146/annurev.arplant.57.032905.105159 (cit. on p. 15).

Baker, M. (2012). “De novo genome assembly: what every biologist should know”. In: *Nat Methods* 9.4,

p. 333. DOI: 10.1038/nmeth.1935 (cit. on pp. 12, 85).

Bankevich, A. et al. (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." In: *J Comput Biol* 19.5, pp. 455–477. DOI: 10.1089/cmb.2012.0021 (cit. on pp. 29, 36, 38).

Bary, H. d. (1878). "Ueber Symbiose". In: *Tageblatt für die Versammlung deutscher Naturforscher und Aerzte Cassel*, pp. 121–126 (cit. on p. 13).

Bayer, K., J. Kamke, and U. Hentschel (2014). "Quantification of bacterial and archaeal symbionts in high and low microbial abundance sponges using real-time PCR." In: *FEMS Microbiol Ecol* 89.3, pp. 679–690. DOI: 10.1111/1574-6941.12369 (cit. on p. 16).

Bayer, K., M. Scheuermayer, L. Fieseler, and U. Hentschel (2013). "Genomic mining for novel FADH-dependent halogenases in marine sponge-associated microbial consortia." In: *Mar Biotechnol (NY)* 15.1, pp. 63–72. DOI: 10.1007/s10126-012-9455-2 (cit. on p. 17).

Bayer, K., S. Schmitt, and U. Hentschel (2008). "Physiology, phylogeny and in situ evidence for bacterial and archaeal nitrifiers in the marine sponge Aplysina aerophoba." In: *Environ Microbiol* 10.11, pp. 2942–2955. DOI: 10.1111/j.1462-2920.2008.01582.x (cit. on p. 17).

Bell, J. J. (2008). "The functional roles of marine sponges". In: *Estuar Coast Shelf Sci* 79.3, pp. 341–353. DOI: 10.1016/j.ecss.2008.05.002 (cit. on p. 16).

Benson, D. A., K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers (2014). "GenBank." In: *Nucleic Acids Res* 42.Database issue, pp. D32–D37. DOI: 10.1093/nar/gkt1030 (cit. on p. 25).

Bentley, S. D. et al. (2002). "Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2)." In: *Nature* 417.6885, pp. 141–147. DOI: 10.1038/417141a (cit. on p. 86).

Bérdy, J. (2012). "Thoughts and facts about antibiotics: where we are now and where we are heading." In: *J Antibiot* 65.8, pp. 385–395. DOI: 10.1038/ja.2012.27 (cit. on p. 92).

Berlec, A. (2012). "Novel techniques and findings in the study of plant microbiota: search for plant probiotics." In: *Plant Sci* 193-194, pp. 96–102. DOI: 10.1016/j.plantsci.2012.05.010 (cit. on p. 15).

Bewley, C. A., N. D. Holland, and D. J. Faulkner (1996). "Two classes of metabolites from Theonella swinhoei are localized in distinct populations of bacterial symbionts." In: *Experientia* 52.7, pp. 716–722. DOI: 10.1007/bf01925581 (cit. on p. 18).

Biswas, A., J. N. Gagnon, S. J. J. Brouns, P. C. Fineran, and C. M. Brown (2013). "CRISPRTarget: bioinformatic prediction and analysis of crRNA targets." In: *RNA Biol* 10.5, pp. 817–827. DOI: 10.4161/rna.24046 (cit. on p. 30).

Bland, C., T. L. Ramsey, F. Sabree, M. Lowe, K. Brown, N. C. Kyrpides, and P. Hugenholtz (2007). "CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats." In: *BMC Bioinformatics* 8, p. 209. DOI: 10.1186/1471-2105-8-209 (cit. on p. 30).

Blattner, F. R. (1997). "The Complete Genome Sequence of Escherichia coli K-12". In: *Science* 277.5331, pp. 1453–1462. DOI: 10.1126/science.277.5331.1453 (cit. on p. 8).

Blom, J., J. Kreis, S. Spänig, T. Juhre, C. Bertelli, C. Ernst, and A. Goesmann (2016). "EDGAR 2.0: an enhanced software platform for comparative gene content analyses." In: *Nucleic Acids Res.* DOI: 10.1093/nar/gkw255 (cit. on p. 12).

Blunt, J. W., B. R. Copp, R. A. Keyzers, M. H. G. Munro, and M. R. Prinsep (2016). "Marine natural products." In: *Nat Prod Rep* 33.3, pp. 382–431. DOI: 10.1039/c5np00156k (cit. on p. 17).

Bodenhausen, N., M. Bortfeld-Miller, M. Ackermann, and J. A. Vorholt (2014). "A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota." In: *PLoS Genet* 10.4, e1004283. DOI: 10.1371/journal.pgen.1004283 (cit. on p. 19).

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano (2011). "Scaffolding pre-assembled contigs using SSPACE." In: *Bioinformatics* 27.4, pp. 578–579. DOI: 10.1093/bioinformatics/btq683 (cit. on p. 39).

Bolger, A. M., M. Lohse, and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." In: *Bioinformatics* 30.15, pp. 2114–2120. DOI: 10.1093/bioinformatics/btu170 (cit. on pp. 28, 38).

Bosch, T. C. G. and D. J. Miller (2016). "Introduction: The Holobiont Imperative". In: *The Holobiont Imperative*. DOI: 10.1007/978-3-7091-1896-2_1 (cit. on p. 18).

Buescher, J. M. and E. M. Driggers (2016). "Integration of omics: more than the sum of its parts." In: *Cancer Metab* 4, p. 4. DOI: 10.1186/s40170-016-0143-y (cit. on p. 6).

Burge, S. W. et al. (2013). "Rfam 11.0: 10 years of RNA families." In: *Nucleic Acids Res* 41.Database issue, pp. D226–D232. DOI: 10.1093/nar/gks1005 (cit. on p. 29).

Burgsdorf, I. et al. (2015). "Lifestyle evolution in cyanobacterial symbionts of sponges." In: *MBio* 6.3, e00391–e00315. DOI: 10.1128/mBio.00391-15 (cit. on pp. 10, 86, 93).

Burstein, D. et al. (2016). "Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems." In: *Nat Commun* 7, p. 10613. DOI: 10.1038/ncomms10613 (cit. on pp. 99, 100).

Bushnell, B. (2014). *BBmap short read aligner, and other bioinformatic tools*. URL: http://sourceforge.net/projects/bbmap/ (visited on June 1, 2016) (cit. on pp. 28, 29).

Cheng, C., L. MacIntyre, U. R. Abdelmohsen, H. **Horn**, P. N. Polymenakou, R. Edrada-Ebel, and U. Hentschel (2015). "Biodiversity, anti-trypanosomal activity screening, and metabolomic profiling of actinomycetes isolated from Mediterranean sponges." In: *PLoS One* 10.9, e0138528. DOI: 10.1371/journal.pone.0138528 (cit. on p. 18).

Cheshire, A. and C. Wilkinson (1991). "Modelling the photosynthetic production by sponges on Davies Reef, Great Barrier Reef". In: *Mar Biol* 109.1, pp. 13–18. DOI: 10.1007/bf01320226 (cit. on p. 17).

Chikhi, R. and P. Medvedev (2014). "Informed and automated k-mer size selection for genome assembly." In: *Bioinformatics* 30.1, pp. 31–37. DOI: 10.1093/bioinformatics/btt310 (cit. on p. 38).

Choffnes, E. R., L. Olsen, T. Wizemann, et al. (2013). *The Science and Applications of Microbial Genomics: Workshop Summary*. National Academies Press. DOI: 10.17226/18261 (cit. on p. 13).

Choi, Y. J. and S. Y. Lee (2013). "Microbial production of short-chain alkanes". In: *Nature* 502.7472, pp. 571–574. DOI: 10.1038/nature12536 (cit. on p. 8).

Clark, S. C., R. Egan, P. I. Frazier, and Z. Wang (2013). "ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies." In: *Bioinformatics* 29.4, pp. 435–443. DOI: 10.1093/bioinformatics/bts723 (cit. on pp. 36, 38).

Clarke, P. H. (1985). *The scientific study of bacteria, 1780–1980*. Springer, pp. 1–37. DOI: 10.1007/978-1-4615-6511-6_1 (cit. on p. 89).

Cock, P. J. A., C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice (2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." In: *Nucleic Acids Res* 38.6, pp. 1767–1771. DOI: 10.1093/nar/gkp1137 (cit. on p. 26).

Coil, D., G. Jospin, and A. E. Darling (2015). "A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data." In: *Bioinformatics* 31.4, pp. 587–589. DOI: 10.1093/bioinformatics/btu661 (cit. on pp. 36, 38, 45).

Cole, S. T. et al. (1998). "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence". In: *Nature* 393.6685, pp. 537–544. DOI: 10.1038/31159 (cit. on p. 7).

Cole, S. T. et al. (2001). "Massive gene decay in the leprosy bacillus". In: *Nature* 409.6823, pp. 1007–1011. DOI: 10.1038/35059006 (cit. on p. 8).

Conn, V. M. and C. M. Franco (2004). "Analysis of the endophytic actinobacterial population in the roots of wheat (Triticum aestivum L.) by terminal restriction fragment length polymorphism and sequencing of 16S rRNA clones". In: *Appl Environ Microbiol* 70.3, pp. 1787–1794. DOI: 10.1128/aem.70.3.1787-1794.2004 (cit. on p. 88).

Crick, F. (1970). "Central dogma of molecular biology." In: *Nature* 227.5258, pp. 561–563. DOI: 10.1038/227561a0 (cit. on p. 6).

Crick, F. H. (1958). "On protein synthesis." In: *Symp Soc Exp Biol* 12, pp. 138–163. DOI: 10.1007/bf00930942 (cit. on p. 6).

Darby, A. C., N.-H. Cho, H.-H. Fuxelius, J. Westberg, and S. G. Andersson (2007). "Intracellular pathogens go extreme: genome evolution in the Rickettsiales". In: *Trends in genetics* 23.10, pp. 511–520. DOI: 10.1016/j.tig.2007.08.002 (cit. on p. 8).

Darling, A. E., G. Jospin, E. Lowe, F. A. Matsen 4th, H. M. Bik, and J. A. Eisen (2014). "PhyloSift: phylogenetic analysis of genomes and metagenomes." In: *PeerJ* 2, e243. DOI: 10.7717/peerj.243 (cit. on p. 38).

De Kruif, P. (2002). *Microbe hunters*. Houghton Mifflin Harcourt. DOI: 10.2307/3901290 (cit. on p. 13).

Del Fabbro, C., S. Scalabrin, M. Morgante, and F. M. Giorgi (2013). "An extensive evaluation of read trimming effects on Illumina NGS data analysis." In: *PLoS One* 8.12, e85024. DOI: 10.1371/journal.pone.0085024 (cit. on p. 40).

Della Sala, G., T. Hochmuth, R. Teta, V. Costantino, and A. Mangoni (2014). "Polyketide synthases in the microbiome of the marine sponge Plakortis halichondrioides: a metagenomic update." In: *Mar Drugs* 12.11, pp. 5425–5440. DOI: 10.3390/md12115425 (cit. on pp. 9, 18, 95).

Delmotte, N. et al. (2009). "Community proteogenomics reveals insights into the physiology of phyllosphere bacteria." In: *Proc Natl Acad Sci U S A* 106.38, pp. 16428–16433. DOI: 10.1073/pnas.0905240106 (cit. on pp. 15, 90).

Dijk, E. L. van, H. Auger, Y. Jaszczyszyn, and C. Thermes (2014). "Ten years of next-generation sequencing technology". In: *Trends Genet* 30.9, pp. 418–426. DOI: 10.1016/j.tig.2014.07.001 (cit. on p. 5).

Dixon, P. (2003). "VEGAN, a package of R functions for community ecology". In: *J Veg Sci* 14.6, pp. 927–930. ISSN: 1654-1103. DOI: 10.1111/j.1654-1103.2003.tb02228.x (cit. on p. 31).

Doroghazi, J. R. and W. W. Metcalf (2013). "Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes." In: *BMC Genomics* 14, p. 611. DOI: 10.1186/1471-2164-14-611 (cit. on pp. 31, 92, 93, 96).

Dupuis, M.-È., M. Villion, A. H. Magadán, and S. Moineau (2013). "CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance." In: *Nat Commun* 4, p. 2087. DOI: 10.1038/ncomms3087 (cit. on p. 100).

Earl, D. et al. (2011). "Assemblathon 1: a competitive assessment of de novo short read assembly methods." In: *Genome Res* 21.12, pp. 2224–2241. DOI: 10.1101/gr.126599.111 (cit. on p. 36).

Eddy, S. R. (2009). "A new generation of homology search tools based on probabilistic inference." In: *Genome Inform* 23.1, pp. 205–211. DOI: 10.1142/9781848165632_0019 (cit. on p. 31).

Edgar, R. C. (2004). "MUSCLE: a multiple sequence alignment method with reduced time and space complexity." In: *BMC Bioinformatics* 5, p. 113. DOI: 10.1186/1471-2105-5-113 (cit. on p. 27).

– (2007). "PILER-CR: fast and accurate identification of CRISPR repeats." In: *BMC Bioinformatics* 8, p. 18. DOI: 10.1186/1471-2105-8-18 (cit. on pp. 30, 39).

Eisen, J. A. (2016). *Sequencing Costs*. URL: https://phylogenomics.wordpress.com/sequencing-costs/ (visited on June 16, 2016).

Ekblom, R. and J. B. W. Wolf (2014). "A field guide to whole-genome sequencing, assembly and annotation". In: *Evol Appl* 7.9, pp. 1026–1042. DOI: 10.1111/eva.12178 (cit. on p. 85).

Erwin, D. H., M. Laflamme, S. M. Tweedt, E. A. Sperling, D. Pisani, and K. J. Peterson (2011). "The Cambrian conundrum: early divergence and later ecological success in the early history of animals." In: *Science* 334.6059, pp. 1091–1097. DOI: 10.1126/science.1206375 (cit. on p. 16).

Escobar-Zepeda, A., A. Vera-Ponce de León, and A. Sanchez-Flores (2015). "The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics." In: *Front Genet* 6, p. 348. DOI: 10.3389/fgene.2015.00348 (cit. on p. 10).

Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." In: *Genome Res* 8.3, pp. 186–194. DOI: 10.1101/gr.8.3.186 (cit. on pp. 26, 28).

Ewing, B., L. Hillier, M. C. Wendl, and P. Green (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." In: *Genome Res* 8.3, pp. 175–185. DOI: 10.1101/gr.8.3.175 (cit. on pp. 26, 28).

Falkowski, P. G., R. T. Barber, and V. Smetacek (1998). "Biogeochemical controls and feedbacks on ocean primary production". In: *Science* 281.5374, pp. 200–206. DOI: 10.1126/science.281.5374.200 (cit. on p. 13).

Fan, L., D. Reynolds, M. Liu, M. Stark, S. Kjelleberg, N. S. Webster, and T. Thomas (2012). "Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts." In: *Proc Natl Acad Sci U S A* 109.27, E1878–E1887.

DOI: 10.1073/pnas.1203287109 (cit. on pp. 17, 86, 98, 99).

Farrer, R. A., E. Kemen, J. D. G. Jones, and D. J. Studholme (2009). "De novo assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads." In: *FEMS Microbiol Lett* 291.1, pp. 103–111. DOI: 10.1111/j.1574-6968.2008.01441.x (cit. on p. 90).

Fiers, W. et al. (1976). "Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene". In: *Nature* 260.5551, pp. 500–507. DOI: 10.1038/260500a0 (cit. on p. 4).

Fieseler, L. et al. (2007). "Widespread occurrence and genomic context of unusually small polyketide synthase genes in microbial consortia associated with marine sponges." In: *Appl Environ Microbiol* 73.7, pp. 2144–2155. DOI: 10.1128/AEM.02260-06 (cit. on pp. 18, 95).

Finkel, O. M., A. Y. Burch, T. Elad, S. M. Huse, S. E. Lindow, A. F. Post, and S. Belkin (2012). "Distance-decay relationships partially determine diversity patterns of phyllosphere bacteria on Tamarix trees across the Sonoran Desert [corrected]." In: *Appl Environ Microbiol* 78.17, pp. 6187–6193. DOI: 10.1128/AEM.00888-12 (cit. on p. 16).

Finkel, O. M., T. O. Delmont, A. F. Post, and S. Belkin (2016). "Metagenomic Signatures of Bacterial Adaptation to Life in the Phyllosphere of a Salt-Secreting Desert Tree." In: *Appl Environ Microbiol* 82.9, pp. 2854–2861. DOI: 10.1128/AEM.00483-16 (cit. on p. 15).

Finn, R. D. et al. (2016). "The Pfam protein families database: towards a more sustainable future." In: *Nucleic Acids Res* 44.D1, pp. D279–D285. DOI: 10.1093/nar/gkv1344 (cit. on pp. 12, 25, 30).

Fisch, K. M. et al. (2009). "Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting." In: *Nat Chem Biol* 5.7, pp. 494–501. DOI: 10.1038/nchembio.176 (cit. on p. 18).

Fleischmann, R. D. et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." In: *Science* 269.5223, pp. 496–512. DOI: 10.1126/science.7542800 (cit. on pp. 4, 7).

Foerstner, K. U., C. von Mering, S. D. Hooper, and P. Bork (2005). "Environments shape the nucleotide composition of genomes." In: *EMBO Rep* 6.12, pp. 1208–1213. DOI: 10.1038/sj.embor.7400538 (cit. on p. 10).

Fraser, C. M. et al. (1995). "The Minimal Gene Complement of Mycoplasma genitalium". In: *Science* 270.5235, pp. 397–404. DOI: 10.1126/science.270.5235.397 (cit. on p. 7).

Friedrich, A. B., I. Fischer, P. Proksch, J. Hacker, and U. Hentschel (2001). "Temporal variation of the microbial community associated with the Mediterranean sponge Aplysina aerophoba". In: *FEMS Microbiol Ecol* 38.2-3, pp. 105–113. DOI: 10.1111/j.1574-6941.2001.tb00888.x (cit. on p. 16).

Galperin, M. Y., K. S. Makarova, Y. I. Wolf, and E. V. Koonin (2015). "Expanded microbial genome coverage and improved protein family annotation in the COG database." In: *Nucleic Acids Res* 43.Database issue, pp. D261–D269. DOI: 10.1093/nar/gku1223 (cit. on pp. 12, 25).

Gawad, C., W. Koh, and S. R. Quake (2016). "Single-cell genome sequencing: current state of the science." In: *Nat Rev Genet* 17.3, pp. 175–188. DOI: 10.1038/nrg.2015.16 (cit. on p. 9).

GenBank (2016). *GenBank and WGS Statistics*. URL: http://www.ncbi.nlm.nih.gov/genbank/statistics/ (visited on May 31, 2016) (cit. on p. 4).

Genomes Online Database, T. (2016). *Statistics*. URL: https://gold.jgi.doe.gov/statistics (visited on May 31, 2016) (cit. on p. 7).

Ghodsi, M., C. M. Hill, I. Astrovskaya, H. Lin, D. D. Sommer, S. Koren, and M. Pop (2013). "De novo likelihood-based measures for comparing genome assemblies." In: *BMC Res Notes* 6, p. 334. DOI: 10.1186/1756-0500-6-334 (cit. on p. 45).

Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field (1990). "Genetic diversity in Sargasso Sea bacterioplankton." In: *Nature* 345.6270, pp. 60–63. DOI: 10.1038/345060a0 (cit. on p. 9).

Glenn, T. C. (2011). "Field guide to next-generation DNA sequencers". In: *Mol Ecol Resour* 11.5, pp. 759–769. DOI: 10.1111/j.1755-0998.2011.03024.x (cit. on p. 6).

Gloeckner, V. et al. (2014). "The HMA-LMA dichotomy revisited: an electron microscopical survey of 56 sponge species". In: *The Biological Bul-*

letin 227.1, pp. 78–88. DOI: 10.1038/ncomms11870 (cit. on p. 16).

Goeij, J. M. de, H. van den Berg, M. M. van Oostveen, E. H. Epping, and F. C. Van Duyl (2008). "Major bulk dissolved organic carbon (DOC) removal by encrusting coral reef cavity sponges". In: *Mar Ecol Prog Ser* 357, pp. 139–151. DOI: 10.3354/meps07403 (cit. on p. 17).

Gogleva, A. A., M. S. Gelfand, and I. I. Artamonova (2014). "Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs." In: *BMC Genomics* 15, p. 202. DOI: 10.1186/1471-2164-15-202 (cit. on pp. 30, 85, 98, 99).

Goris, J., K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme, and J. M. Tiedje (2007). "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities." In: *Int J Syst Evol Microbiol* 57.Pt 1, pp. 81–91. DOI: 10.1099/ijs.0.64483-0 (cit. on p. 31).

Greenleaf, W. J. and A. Sidow (2014). "The future of sequencing: convergence of intelligent design and market Darwinism." In: *Genome Biol* 15.3, p. 303. DOI: 10.1186/gb4168 (cit. on p. 5).

Grissa, I., G. Vergnaud, and C. Pourcel (2007a). "CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats." In: *Nucleic Acids Res* 35.Web Server issue, W52–W57. DOI: 10.1093/nar/gkm360 (cit. on p. 30).

– (2007b). "The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats." In: *BMC Bioinformatics* 8, p. 172. DOI: 10.1186/1471-2105-8-172 (cit. on p. 30).

Guerrero, L. D., T. P. Makhalanyane, J. M. Aislabie, and D. A. Cowan (2014). "Draft Genome Sequence of Williamsia sp. Strain D3, Isolated From the Darwin Mountains, Antarctica." In: *Genome Announc* 2.1. DOI: 10.1128/genomeA.01230-13 (cit. on p. 88).

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel (2010). "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0." In: *Syst Biol* 59.3, pp. 307–321. DOI: 10.1093/sysbio/syq010 (cit. on p. 27).

Guo, F., Z.-P. Wang, K. Yu, and T. Zhang (2015). "Detailed investigation of the microbial community in foaming activated sludge reveals novel foam formers". In: *Scientific reports* 5, p. 7637. DOI: 10.1038/srep07637 (cit. on p. 88).

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler (2013). "QUAST: quality assessment tool for genome assemblies." In: *Bioinformatics* 29.8, pp. 1072–1075. DOI: 10.1093/bioinformatics/btt086 (cit. on p. 38).

Haft, D. H., J. D. Selengut, R. A. Richter, D. Harkins, M. K. Basu, and E. Beck (2013). "TIGRFAMs and Genome Properties in 2013." In: *Nucleic Acids Res* 41.Database issue, pp. D387–D395. DOI: 10.1093/nar/gks1234 (cit. on pp. 12, 25, 30).

Haft, D. H., J. Selengut, E. F. Mongodin, and K. E. Nelson (2005). "A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes." In: *PLoS Comput Biol* 1.6, e60. DOI: 10.1371/journal.pcbi.0010060 (cit. on p. 30).

Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman (1998). "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products." In: *Chem Biol* 5.10, R245–R249. DOI: 10.1016/s1074-5521(98)90108-9 (cit. on p. 9).

Handelsman, J. (2004). "Metagenomics: application of genomics to uncultured microorganisms." In: *Microbiol Mol Biol Rev* 68.4, pp. 669–685. DOI: 10.1128/MMBR.68.4.669-685.2004 (cit. on pp. 9, 10).

Handelsman, J. et al. (2007). "Committee on metagenomics: challenges and functional applications". In: *Washington: National Academy of Sciences*, pp. 1–158. DOI: 10.1002/9780470015902.a0020367 (cit. on p. 13).

Hansjakob, A., K. U. Foerstner, H. **Horn**, M. Riederer, and U. Hildebrand (2017). "Surface dependent gene expression of *Blumeria graminis* f. sp. hordei during the prepenetration processes". In: *New Phytol.* In preparation.

Harjes, J., T. Ryu, U. R. Abdelmohsen, L. Moitinho-Silva, H. **Horn**, T. Ravasi, and U. Hentschel (2014). "Draft genome sequence of the antitrypanosomally active sponge-associated bacterium *Actinokineospora* sp. strain EG49." In: *Genome Announc* 2.2. DOI: 10.1128/genomeA.00160-14.

Healy, F., R. Ray, H. Aldrich, A. Wilkie, L. Ingram, and K. Shanmugam (1995). "Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose". In: *Appl Microbiol Biotechnol* 43.4, pp. 667–674. DOI: 10.1007/bf00164771 (cit. on p. 9).

Heather, J. M. and B. Chain (2016). "The sequence of sequencers: The history of sequencing DNA". In: *Genomics* 107.1, pp. 1–8. DOI: 10.1016/j.ygeno.2015.11.003 (cit. on pp. 3–5).

Helfrich, E. J. N. and J. Piel (2016). "Biosynthesis of polyketides by trans-AT polyketide synthases." In: *Nat Prod Rep* 33.2, pp. 231–316. DOI: 10.1039/c5np00125k (cit. on p. 95).

Hentschel, U., L. Fieseler, M. Wehrl, C. Gernert, M. Steinert, J. Hacker, and M. Horn (2003). "Microbial diversity of marine sponges." In: *Prog Mol Subcell Biol* 37, pp. 59–88. DOI: 10.1007/978-3-642-55519-0_3 (cit. on p. 16).

Hentschel, U., J. Hopke, M. Horn, A. B. Friedrich, M. Wagner, J. Hacker, and B. S. Moore (2002). "Molecular evidence for a uniform microbial community in sponges from different oceans." In: *Appl Environ Microbiol* 68.9, pp. 4431–4440. DOI: 10.1128/aem.68.9.4431-4440.2002 (cit. on p. 16).

Hentschel, U., J. Piel, S. M. Degnan, and M. W. Taylor (2012). "Genomic insights into the marine sponge microbiome." In: *Nat Rev Microbiol* 10.9, pp. 641–654. DOI: 10.1038/nrmicro2839 (cit. on pp. 16, 17, 94).

Hentschel, U., K. M. Usher, and M. W. Taylor (2006). "Marine sponges as microbial fermenters." In: *FEMS Microbiol Ecol* 55.2, pp. 167–177. DOI: 10.1111/j.1574-6941.2005.00046.x (cit. on p. 16).

Hershey, A. D. and M. Chase (1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage". In: *J Gen Phyiol* 36.1, pp. 39–56. DOI: 10.1085/jgp.36.1.39 (cit. on p. 3).

Hochmuth, T. and J. Piel (2009). "Polyketide synthases of bacterial symbionts in sponges–evolution-based applications in natural products research." In: *Phytochemistry* 70.15-16, pp. 1841–1849. DOI: 10.1016/j.phytochem.2009.04.010 (cit. on pp. 95, 96).

Hochmuth, T. et al. (2010). "Linking chemical and microbial diversity in marine sponges: possible role for poribacteria as producers of methyl-branched fatty acids." In: *Chembiochem* 11.18, pp. 2572–2578. DOI: 10.1002/cbic.201000510 (cit. on pp. 18, 95).

Holden, M., L. Crossman, A. Cerdeño-Tárraga, and J. Parkhill (2004). "Genome watch: Pathogenomics of non-pathogens". In: *Nat Rev Microbiol* 2.2, pp. 91–91. DOI: 10.1038/nrmicro825 (cit. on p. 8).

Holley, R. W., J. Apgar, S. H. Merrill, and P. L. Zubkoff (1961). "Nuncleotide and oligonucleotide composition of the alanine-, valine-, and tyrosin-acceptor soluble ribonucleic acids of yeast". In: *J Am Chem Soc* 83.23, pp. 4861–4862. DOI: 10.1021/ja01484a040 (cit. on p. 4).

Holley, R. W. et al. (1965). "Structure of a ribonucleic acid". In: *Science* 147.3664, pp. 1462–1465. DOI: 10.1126/science.147.3664.1462 (cit. on p. 4).

Hooper, J. N. and R. W. Van Soest (2002). *Systema Porifera. A guide to the classification of sponges.* Springer, pp. 1–7 (cit. on p. 16).

Horn, H., C. Cheng, R. Edrada-Ebel, U. Hentschel, and U. R. Abdelmohsen (2015a). "Draft genome sequences of three chemically rich actinomycetes isolated from Mediterranean sponges." In: *Mar Genomics* 24 Pt 3, pp. 285–287. DOI: 10.1016/j.margen.2015.10.003 (cit. on p. 37).

Horn, H., U. Hentschel, and U. R. Abdelmohsen (2015b). "Mining genomes of three marine sponge-associated actinobacterial isolates for secondary metabolism." In: *Genome Announc* 3.5. DOI: 10.1128/genomeA.01106-15 (cit. on p. 37).

Horn, H., A. Keller, U. Hildebrandt, P. Kämpfer, M. Riederer, and U. Hentschel (2016a). "Draft genome of the *Arabidopsis thaliana* phyllosphere bacterium, *Williamsia* sp. ARP1." In: *Stand Genomic Sci* 11, p. 8. DOI: 10.1186/s40793-015-0122-x (cit. on pp. 15, 37, 88).

Horn, H., B. M. Slaby, M. T. Jahn, K. Bayer, F. Foerster, U. R. Abdelmohsen, and U. Hentschel (2016b). "An enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes." In: *Front Microbiol* 7, p. 1751. DOI: 10.3389/fmicb.2016.01751.

Horton, M. W. et al. (2014). "Genome-wide association study of Arabidopsis thaliana leaf microbial community." In: *Nat Commun* 5, p. 5320. DOI: 10.1038/ncomms6320 (cit. on p. 16).

Howison, M., F. Zapata, and C. W. Dunn (2013). "Toward a statistically explicit understanding of de novo sequence assembly". In: *Bioinformatics* 29.23, pp. 2959–2963. DOI: 10.1093/bioinformatics/btt525 (cit. on p. 36).

Huang, X. and A. Madan (1999). "CAP3: A DNA sequence assembly program." In: *Genome Res* 9.9, pp. 868–877. DOI: 10.1101/gr.9.9.868 (cit. on p. 26).

Huber, R., H. Huber, and K. O. Stetter (2000). "Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties." In: *FEMS Microbiol Rev* 24.5, pp. 615–623. DOI: 10.1111/j.1574-6976.2000.tb00562.x (cit. on p. 13).

Hugenholtz, P. (2002). "Exploring prokaryotic diversity in the genomic era." In: *Genome Biol* 3.2, REVIEWS0003. DOI: 10.1186/gb-2002-3-2-reviews0003 (cit. on p. 9).

Human Microbiome Project Consortium, T. (2012). "A framework for human microbiome research." In: *Nature* 486.7402, pp. 215–221. DOI: 10.1038/nature11209 (cit. on p. 10).

Hunt, M., T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T. D. Otto (2013). "REAPR: a universal tool for genome assembly evaluation." In: *Genome Biol* 14.5, R47. DOI: 10.1186/gb-2013-14-5-r47 (cit. on p. 45).

Hunter, P. J., P. Hand, D. Pink, J. M. Whipps, and G. D. Bending (2010). "Both leaf properties and microbe-microbe interactions influence within-species variation in bacterial population diversity and structure in the lettuce (Lactuca Species) phyllosphere." In: *Appl Environ Microbiol* 76.24, pp. 8117–8125. DOI: 10.1128/AEM.01321-10 (cit. on pp. 15, 16).

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster (2007). "MEGAN analysis of metagenomic data." In: *Genome Res* 17.3, pp. 377–386. DOI: 10.1101/gr.5969107 (cit. on p. 85).

Huson, D. H., S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster (2011). "Integrative analysis of environmental sequences using MEGAN4." In: *Genome Res* 21.9, pp. 1552–1560. DOI: 10.1101/gr.120618.111 (cit. on p. 28).

Hyatt, D., G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification." In: *BMC Bioinformatics* 11, p. 119. DOI: 10.1186/1471-2105-11-119 (cit. on p. 29).

Illumina (2016). *HiSeq X Ten System*. URL: http://www.illumina.com/systems/hiseq-x-sequencing-system/system.html (visited on June 17, 2016) (cit. on p. 5).

Initiative, T. A. G. (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana". In: *Nature* 408.6814, pp. 796–815. DOI: 10.1038/35048692 (cit. on p. 7).

Jain, R., M. C. Rivera, and J. A. Lake (1999). "Horizontal gene transfer among genomes: the complexity hypothesis." In: *Proc Natl Acad Sci U S A* 96.7, pp. 3801–3806. DOI: 10.1073/pnas.96.7.3801 (cit. on p. 90).

Jones, A. L., G. D. Payne, and M. Goodfellow (2010). "Williamsia faeni sp. nov., an actinomycete isolated from a hay meadow." In: *Int J Syst Evol Microbiol* 60.Pt 11, pp. 2548–2551. DOI: 10.1099/ijs.0.015826-0 (cit. on p. 88).

Jones, P. et al. (2014). "InterProScan 5: genome-scale protein function classification." In: *Bioinformatics* 30.9, pp. 1236–1240. DOI: 10.1093/bioinformatics/btu031 (cit. on p. 30).

Jou, W. M., G. Haegeman, M. Ysebaert, and W. Fiers (1972). "Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein". In: *Nature* 237, pp. 82–88. DOI: 10.1038/237082a0 (cit. on p. 4).

Kamke, J. et al. (2013). "Single-cell genomics reveals complex carbohydrate degradation patterns in poribacterial symbionts of marine sponges." In: *ISME J* 7.12, pp. 2287–2300. DOI: 10.1038/ismej.2013.111 (cit. on p. 17).

Kamke, J. et al. (2014). "The candidate phylum Poribacteria by single-cell genomics: new insights into phylogeny, cell-compartmentation, eukaryote-like repeat proteins, and other genomic features." In: *PLoS One* 9.1, e87353. DOI: 10.1371/journal.pone.0087353 (cit. on p. 9).

Kämpfer, P., M. A. Andersson, F. A. Rainey, R. M. Kroppenstedt, and M. Salkinoja-Salonen (1999). "Williamsia muralis gen. nov., sp. nov., isolated from the indoor environment of a children's day care centre." In: *Int J Syst Bacteriol* 49 Pt 2, pp. 681–687. DOI: 10.1099/00207713-49-2-681 (cit. on pp. 88, 90).

Kämpfer, P., H.-J. Busse, H. **Horn**, U. R. Abdelmohsen, U. H. Hentschel, and S. P. Glaeser (2016). "*Williamsia herbipolensis* sp. nov., isolated from the phyllosphere of *Arabidopsis thaliana*". In: *Int J Syst Evol Microbiol* 66.11, pp. 4609–4613. DOI: 10.1099/ijsem.0.001398 (cit. on p. 90).

Keane, T. M., C. J. Creevey, M. M. Pentony, T. J. Naughton, and J. O. Mclnerney (2006). "Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified." In: *BMC Evol Biol* 6, p. 29. DOI: 10.1186/1471-2148-6-29 (cit. on p. 27).

Keller, A., H. **Horn**, F. Förster, and J. Schultz (2014). "Computational integration of genomic traits into 16S rDNA microbiota sequencing studies." In: *Gene* 549.1, pp. 186–191. DOI: 10.1016/j.gene.2014.07.066.

Kembel, S. W., T. K. O'Connor, H. K. Arnold, S. P. Hubbell, S. J. Wright, and J. L. Green (2014). "Relationships between phyllosphere bacterial communities and plant functional traits in a neotropical forest." In: *Proc Natl Acad Sci U S A* 111.38, pp. 13715–13720. DOI: 10.1073/pnas.1216057111 (cit. on pp. 14, 15).

Kim, M., H.-S. Oh, S.-C. Park, and J. Chun (2014). "Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes". In: *Int J Syst Evol Microbiol* 64.2, pp. 346–351. DOI: 10.1099/ijs.0.059774-0 (cit. on p. 89).

Kislyuk, A. O. et al. (2010). "A computational genomics pipeline for prokaryotic sequencing projects." In: *Bioinformatics* 26.15, pp. 1819–1826. DOI: 10.1093/bioinformatics/btq284 (cit. on pp. 36, 45).

Kitahara, K. and K. Miyazaki (2013). "Revisiting bacterial phylogeny: Natural and experimental evidence for horizontal gene transfer of 16S rRNA." In: *Mob Genet Elements* 3.1, e24210. DOI: 10.4161/mge.24210 (cit. on pp. 89, 90).

Knief, C. (2014). "Analysis of plant microbe interactions in the era of next generation sequencing technologies." In: *Front Plant Sci* 5, p. 216. DOI: 10.3389/fpls.2014.00216 (cit. on p. 9).

Knief, C., V. Dengler, P. L. E. Bodelier, and J. A. Vorholt (2012a). "Characterization of Methylobacterium strains isolated from the phyllosphere and description of Methylobacterium longum sp. nov." In: *Antonie Van Leeuwenhoek* 101.1, pp. 169–183. DOI: 10.1007/s10482-011-9650-6 (cit. on p. 88).

Knief, C., A. Ramette, L. Frances, C. Alonso-Blanco, and J. A. Vorholt (2010). "Site and plant species are important determinants of the Methylobacterium community composition in the plant phyllosphere." In: *ISME J* 4.6, pp. 719–728. DOI: 10.1038/ismej.2010.9 (cit. on pp. 15, 16).

Knief, C. et al. (2012b). "Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice." In: *ISME J* 6.7, pp. 1378–1390. DOI: 10.1038/ismej.2011.192 (cit. on pp. 15, 91).

Koonin, E. V. and K. S. Makarova (2009). "CRISPR-Cas: an adaptive immunity system in prokaryotes." In: *F1000 Biol Rep* 1, p. 95. DOI: 10.3410/B1-95 (cit. on p. 98).

Koren, S., T. J. Treangen, C. M. Hill, M. Pop, and A. M. Phillippy (2014). "Automated ensemble assembly and validation of microbial genomes." In: *BMC Bioinformatics* 15, p. 126. DOI: 10.1186/1471-2105-15-126 (cit. on pp. 36, 40, 43, 45).

Kostadinov, I. (2011). "Marine Metagenomics. From high-throughput data to ecogenomic interpretation". PhD thesis. Max Planck Institute for Marine Microbiology, p. 139 (cit. on p. 13).

Kulikova, T. et al. (2007). "EMBL Nucleotide Sequence Database in 2006." In: *Nucleic Acids Res* 35.Database issue, pp. D16–D20. DOI: 10.1093/nar/gkl913 (cit. on p. 25).

Kumar, S., M. Jones, G. Koutsovoulos, M. Clarke, and M. Blaxter (2013). "Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots." In: *Front Genet* 4, p. 237. DOI: 10.3389/fgene.2013.00237 (cit. on pp. 38, 45).

Kunin, V. et al. (2008). "A bacterial metapopulation adapts locally to phage predation despite global dispersal." In: *Genome Res* 18.2, pp. 293–297. DOI: 10.1101/gr.6835308 (cit. on p. 99).

Kunst, F. et al. (1997). "The complete genome sequence of the gram-positive bacterium Bacillus subtilis." In: *Nature* 390.6657, pp. 249–256. DOI: 10.1038/36786 (cit. on p. 8).

Kurland, C. G. et al. (1998). "The genome sequence of Rickettsia prowazekii and the origin of mitochondria." In: *Nature* 396.6707, pp. 133–140. DOI: 10.1038/24094 (cit. on p. 8).

Kuska, B. (1998). “Beer, Bethesda, and biology: how "genomics" came into being.” In: *J Natl Cancer Inst* 90.2, p. 93. DOI: 10.1093/jnci/90.2.93 (cit. on p. 6).

Kyrpides, N. C. et al. (2014). “Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains.” In: *PLoS Biol* 12.8, e1001920. DOI: 10.1371/journal.pbio.1001920 (cit. on p. 7).

Lagesen, K., P. Hallin, E. A. Rødland, H.-H. Staerfeldt, T. Rognes, and D. W. Ussery (2007). “RNAmmer: consistent and rapid annotation of ribosomal RNA genes.” In: *Nucleic Acids Res* 35.9, pp. 3100–3108. DOI: 10.1093/nar/gkm160 (cit. on pp. 29, 39).

Land, M. et al. (2015). “Insights from 20 years of bacterial genome sequencing.” In: *Funct Integr Genomics* 15.2, pp. 141–161. DOI: 10.1007/s10142-015-0433-4 (cit. on pp. 4, 7, 8).

Lander, E. S. et al. (2001). “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062 (cit. on pp. 4, 7).

Lane, D. J., B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace (1985). “Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses.” In: *Proc Natl Acad Sci U S A* 82.20, pp. 6955–6959. DOI: 10.1073/pnas.82.20.6955 (cit. on p. 9).

Lange, S. J., O. S. Alkhnbashi, D. Rose, S. Will, and R. Backofen (2013). “CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems.” In: *Nucleic Acids Res* 41.17, pp. 8034–8044. DOI: 10.1093/nar/gkt606 (cit. on p. 30).

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg (2009). “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.” In: *Genome Biol* 10.3, R25. DOI: 10.1186/gb-2009-10-3-r25 (cit. on pp. 29, 38).

Lau, J. A. and J. T. Lennon (2012). “Rapid responses of soil microorganisms improve plant fitness in novel environments”. In: *Proc Natl Acad Sci U S A* 109.35, pp. 14058–14062. DOI: 10.1073/pnas.1202319109 (cit. on p. 14).

Lejon, D. P., J. Kennedy, and A. D. Dobson (2011). “Identification of novel bioactive compounds from the metagenome of the marine sponge Haliclona simulans”. In: *Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats*, pp. 553–562. DOI: 10.1002/9781118010549.ch52 (cit. on p. 17).

Letunic, I. and P. Bork (2011). “Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.” In: *Nucleic Acids Res* 39.Web Server issue, W475–W478. DOI: 10.1093/nar/gkr201 (cit. on p. 27).

Letunic, I., T. Doerks, and P. Bork (2012). “SMART 7: recent updates to the protein domain annotation resource.” In: *Nucleic Acids Res* 40.Database issue, pp. D302–D305. DOI: 10.1093/nar/gkr931 (cit. on pp. 25, 30).

Leveau, J. H. J. and S. E. Lindow (2001). “Appetite of an epiphyte: Quantitative monitoring of bacterial sugar consumption in the phyllosphere”. In: *Proc Natl Acad Sci U S A* 98.6, pp. 3446–3453. DOI: 10.1073/pnas.061629598 (cit. on p. 14).

Leys, S. P. and A. Hill (2012). “The physiology and molecular biology of sponge tissues.” In: *Adv Mar Biol* 62, pp. 1–56. DOI: 10.1016/B978-0-12-394283-8.00001-1 (cit. on p. 16).

Li, B. and C. T. Walsh (2010). “Identification of the gene cluster for the dithiolopyrrolone antibiotic holomycin in Streptomyces clavuligerus.” In: *Proc Natl Acad Sci U S A* 107.46, pp. 19731–19735. DOI: 10.1073/pnas.1014140107 (cit. on p. 92).

Li, H. and R. Durbin (2009). “Fast and accurate short read alignment with Burrows-Wheeler transform.” In: *Bioinformatics* 25.14, pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324 (cit. on p. 26).

Li, Chen, and Hua (1998). “Precambrian sponges with cellular structures”. In: *Science* 279.5352, pp. 879–882. DOI: 10.1126/science.279.5352.879 (cit. on p. 16).

Liao, Y.-C., H.-H. Lin, A. Sabharwal, E. M. Haase, and F. A. Scannapieco (2015). “MyPro: A seamless pipeline for automated prokaryotic genome assembly and annotation.” In: *J Microbiol Methods* 113, pp. 72–74. DOI: 10.1016/j.mimet.2015.04.006 (cit. on pp. 36, 45).

Lin, S.-H. and Y.-C. Liao (2013). “CISA: contig integrator for sequence assembly of bacterial genomes.” In: *PLoS One* 8.3, e60843. DOI: 10.1371/journal.pone.0060843 (cit. on p. 45).

Lindgreen, S., K. L. Adair, and P. P. Gardner (2016). “An evaluation of the accuracy and speed of metagenome analysis tools." In: *Sci Rep* 6, p. 19233. DOI: 10.1038/srep19233 (cit. on p. 94).

Lindow, S. E. and M. T. Brandl (2003). "Microbiology of the phyllosphere." In: *Appl Environ Microbiol* 69.4, pp. 1875–1883. DOI: 10.1128/aem.69.4.1875-1883.2003 (cit. on pp. 14, 88, 90, 91).

Little, A. E. F., C. J. Robinson, S. B. Peterson, K. F. Raffa, and J. Handelsman (2008). "Rules of engagement: interspecies interactions that regulate microbial communities." In: *Annu Rev Microbiol* 62, pp. 375–401. DOI: 10.1146/annurev.micro.030608.101423 (cit. on p. 13).

Liu, L. et al. (2012a). "Comparison of next-generation sequencing systems". In: *J Biomed Biotechnol* 2012. DOI: 10.1201/b16568-3 (cit. on p. 5).

Liu, M., L. Fan, L. Zhong, S. Kjelleberg, and T. Thomas (2012b). "Metaproteogenomic analysis of a community of sponge symbionts." In: *ISME J* 6.8, pp. 1515–1525. DOI: 10.1038/ismej.2012.1 (cit. on p. 17).

Lockhart, D. J. and E. A. Winzeler (2000). "Genomics, gene expression and DNA arrays." In: *Nature* 405.6788, pp. 827–836. DOI: 10.1038/35015701 (cit. on p. 6).

Loman, N. J. and M. J. Pallen (2015). "Twenty years of bacterial genome sequencing". In: *Nat Rev Microbiol* 13.12, pp. 787–794. DOI: 10.1038/nrmicro3565 (cit. on pp. 7, 8).

Lougheed, K. (2012). "There are fewer microbes out there than you think". In: *Nature*. DOI: 10.1038/nature.2012.11275 (cit. on p. 13).

Love, G. D. et al. (2009). "Fossil steroids record the appearance of Demospongiae during the Cryogenian period." In: *Nature* 457.7230, pp. 718–721. DOI: 10.1038/nature07673 (cit. on p. 16).

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." In: *Nucleic Acids Res* 25.5, pp. 955–964. DOI: 10.1093/nar/25.5.955 (cit. on pp. 29, 39).

Luo, R. et al. (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." In: *Gigascience* 1.1, p. 18. DOI: 10.1186/2047-217X-1-18 (cit. on p. 36).

Machado, H., E. C. Sonnenschein, J. Melchiorsen, and L. Gram (2015). "Genome mining reveals unlocked bioactive potential of marine Gram-negative bacteria." In: *BMC Genomics* 16, p. 158. DOI: 10.1186/s12864-015-1365-z (cit. on p. 92).

Magoc, T. et al. (2013). "GAGE-B: an evaluation of genome assemblers for bacterial organisms." In: *Bioinformatics* 29.14, pp. 1718–1725. DOI: 10.1093/bioinformatics/btt273 (cit. on pp. 36, 43, 45).

Makarova, K. S., N. V. Grishin, S. A. Shabalina, Y. I. Wolf, and E. V. Koonin (2006). "A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action". In: *Biol. Direct* 1.1, p. 7. DOI: 10.1186/1745-6150-1-7 (cit. on p. 8).

Makarova, K. S., Y. I. Wolf, and E. V. Koonin (2013). "Comparative genomics of defense systems in archaea and bacteria." In: *Nucleic Acids Res* 41.8, pp. 4360–4377. DOI: 10.1093/nar/gkt157 (cit. on pp. 97, 100).

Makarova, K. S., Y. I. Wolf, S. Snir, and E. V. Koonin (2011). "Defense islands in bacterial and archaeal genomes and prediction of novel defense systems." In: *J Bacteriol* 193.21, pp. 6039–6056. DOI: 10.1128/JB.05535-11 (cit. on p. 97).

Makarova, K. S. et al. (2015). "An updated evolutionary classification of CRISPR-Cas systems." In: *Nat Rev Microbiol* 13.11, pp. 722–736. DOI: 10.1038/nrmicro3569 (cit. on p. 98).

Marchler-Bauer, A. et al. (2015). "CDD: NCBI's conserved domain database." In: *Nucleic Acids Res* 43.Database issue, pp. D222–D226. DOI: 10.1093/nar/gku1221 (cit. on pp. 25, 30).

Marcy, Y. et al. (2007). "Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth". In: *Proc Natl Acad Sci U S A* 104.29, pp. 11889–11894. DOI: 10.1073/pnas.0704662104 (cit. on p. 9).

Margulies, M. et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." In: *Nature* 437.7057, pp. 376–380. DOI: 10.1038/nature03959 (cit. on p. 5).

Markowitz, V. M. et al. (2012). "IMG/M: the integrated metagenome data management and comparative analysis system." In: *Nucleic Acids Res* 40.Database issue, pp. D123–D129. DOI: 10.1093/nar/gkr975 (cit. on pp. 12, 29).

Maurelli, A. T. (2007). "Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens". In: *FEMS Microbiology Letters* 267.1, pp. 1–8. DOI: 10.1111/j.1574-6968.2006.00526.x (cit. on p. 8).

Mayr, E. (1940). "Speciation phenomena in birds". In: *The American Naturalist* 74.752, pp. 249–278. DOI: 10.1086/280892 (cit. on p. 89).

Medina-Martinez, M. S., A. Allende, G. G. Barbera, and M. I. Gil (2015). "Climatic variations influence the dynamic of epiphyte bacteria of baby lettuce". In: *Food Res Int* 68, pp. 54–61. DOI: 10.1016/j.foodres.2014.06.009 (cit. on p. 14).

Merchant, S., D. E. Wood, and S. L. Salzberg (2014). "Unexpected cross-species contamination in genome sequencing projects." In: *PeerJ* 2, e675. DOI: 10.7717/peerj.675 (cit. on p. 11).

Meyer, F. et al. (2008). "The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes." In: *BMC Bioinformatics* 9, p. 386. DOI: 10.1186/1471-2105-9-386 (cit. on pp. 12, 28).

Microbiol, N. R. (2011). "Microbiology by numbers". In: *Nat Rev Microbiol* 9 (9), p. 628. DOI: 10.1038/nrmicro2644 (cit. on p. 104).

Miller, J. R., S. Koren, and G. Sutton (2010). "Assembly algorithms for next-generation sequencing data." In: *Genomics* 95.6, pp. 315–327. DOI: 10.1016/j.ygeno.2010.03.001 (cit. on p. 36).

Miller, W. G., M. T. Brandl, B. Quinones, and S. E. Lindow (2001). "Biological Sensor for Sucrose Availability: Relative Sensitivities of Various Reporter Genes". In: *Appl Environ Microbiol* 67.3, pp. 1308–1317. DOI: 10.1128/aem.67.3.1308-1317.2001 (cit. on p. 14).

Moitinho-Silva, L. et al. (2014). "Specificity and transcriptional activity of microbiota associated with low and high microbial abundance sponges from the Red Sea." In: *Mol Ecol* 23.6, pp. 1348–1363. DOI: 10.1111/mec.12365 (cit. on p. 17).

Müller, T. and S. Ruppel (2014). "Progress in cultivation-independent phyllosphere microbiology." In: *FEMS Microbiol Ecol* 87.1, pp. 2–17. DOI: 10.1111/1574-6941.12198 (cit. on pp. 19, 88).

Müller, W. E. G. and I. M. Müller (2003). "Origin of the metazoan immune system: identification of the molecules and their functions in sponges." In: *Integr Comp Biol* 43.2, pp. 281–292. DOI: 10.1093/icb/43.2.281 (cit. on p. 16).

Nagarajan, N. and M. Pop (2013). "Sequence assembly demystified." In: *Nat Rev Genet* 14.3, pp. 157–167. DOI: 10.1038/nrg3367 (cit. on p. 28).

Nature (2014). *Definition of Genomics*. Ed. by S. by Nature Education. URL: http://www.nature.com/scitable/definition/genomics-126 (visited on June 13, 2016) (cit. on p. 8).

Nawrocki, E. P. and S. R. Eddy (2013). "Infernal 1.1: 100-fold faster RNA homology searches." In: *Bioinformatics* 29.22, pp. 2933–2935. DOI: 10.1093/bioinformatics/btt509 (cit. on pp. 29, 39).

NCBI (2014). *Genome Browser*. URL: https://www.ncbi.nlm.nih.gov/genome/browse/ (visited on Dec. 17, 2016) (cit. on p. 7).

Nederbragt, L. (2015). *Developments in NGS*. URL: https://github.com/lexnederbragt/developments-in-next-generation-sequencing (visited on May 10, 2016) (cit. on p. 5).

Newman, D. J. and G. M. Cragg (2007). "Natural Products as Sources of New Drugs over the Last 25 Years". In: *J Nat Prod* 70.3, pp. 461–477. DOI: 10.1021/np068054v (cit. on p. 92).

Okonechnikov, K., A. Conesa, and F. García-Alcalde (2016). "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data." In: *Bioinformatics* 32.2, pp. 292–294. DOI: 10.1093/bioinformatics/btv566 (cit. on p. 38).

Oliveira, P. H., M. Touchon, and E. P. C. Rocha (2014). "The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts." In: *Nucleic Acids Res* 42.16, pp. 10618–10631. DOI: 10.1093/nar/gku734 (cit. on pp. 31, 99).

Oxford English Dictionary, T. (2016). *Oxford English Dictionary Online*. URL: www.oed.com (visited on May 14, 2016) (cit. on p. 6).

Pace, N. R., D. A. Stahl, D. J. Lane, and G. J. Olsen (1985). "Analyzing natural microbial populations by rRNA sequences." In: *ASM News* 51.1, pp. 4–12. DOI: 10.1007/978-1-4757-0611-6_1 (cit. on p. 9).

Paddon, C. J. et al. (2013). "High-level semi-synthetic production of the potent antimalarial artemisinin".

In: *Nature* 496.7446, pp. 528–532. DOI: 10.1038/nature12051 (cit. on p. 8).

Partida-Martinez, L. P. and M. Heil (2011). "The microbe-free plant: fact or artifact?" In: *Front Plant Sci* 2, p. 100. DOI: 10.3389/fpls.2011.00100 (cit. on p. 14).

Pathom-Aree, W., Y. Nogi, I. C. Sutcliffe, A. C. Ward, K. Horikoshi, A. T. Bull, and M. Goodfellow (2006a). "Williamsia marianensis sp. nov., a novel actinomycete isolated from the Mariana Trench." In: *Int J Syst Evol Microbiol* 56.Pt 5, pp. 1123–1126. DOI: 10.1099/ijs.0.64132-0 (cit. on p. 88).

Pathom-Aree, W., J. E. Stach, A. C. Ward, K. Horikoshi, A. T. Bull, and M. Goodfellow (2006b). "Diversity of actinomycetes isolated from Challenger Deep sediment (10,898 m) from the Mariana Trench". In: *Extremophiles* 10.3, pp. 181–189. DOI: 10.1007/s00792-005-0482-z (cit. on p. 88).

Paul, V. J., R. Ritson-Williams, and K. Sharp (2011). "Marine chemical ecology in benthic environments." In: *Nat Prod Rep* 28.2, pp. 345–387. DOI: 10.1039/c0np00040j (cit. on p. 93).

Pavlopoulos et al. (2015). "Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies". In: *Bioinform Biol Insights*, p. 75. DOI: 10.4137/BBI.S12462 (cit. on p. 12).

Peabody, M. A., T. Van Rossum, R. Lo, and F. S. L. Brinkman (2015). "Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities." In: *BMC Bioinformatics* 16, p. 363. DOI: 10.1186/s12859-015-0788-5 (cit. on p. 94).

Pedrós-Alió, C. (2006). "Genomics and marine microbial ecology". In: *Int Microbiol* 9.3, pp. 191–198. DOI: 10.1007/s00248-002-1026-z (cit. on p. 13).

Peng, Y., H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin (2012). "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." In: *Bioinformatics* 28.11, pp. 1420–1428. DOI: 10.1093/bioinformatics/bts174 (cit. on pp. 29, 36, 38).

Pennisi, E. (2013). "The CRISPR Craze". In: *Science* 341.6148, pp. 833–836. DOI: 10.1126/science.341.6148.833 (cit. on p. 8).

Piel, J. (2009). "Metabolites from symbiotic bacteria." In: *Nat Prod Rep* 26.3, pp. 338–362. DOI: 10.1039/b703499g (cit. on p. 17).

Piel, J., D. Hui, G. Wen, D. Butzke, M. Platzer, N. Fusetani, and S. Matsunaga (2004). "Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge Theonella swinhoei." In: *Proc Natl Acad Sci U S A* 101.46, pp. 16222–16227. DOI: 10.1073/pnas.0405976101 (cit. on pp. 18, 95).

Podar, M. et al. (2007). "Targeted Access to the Genomes of Low-Abundance Organisms in Complex Microbial Communities". In: *Appl Environ Microbiol* 73.10, pp. 3205–3214. DOI: 10.1128/aem.02985-06 (cit. on p. 9).

Proksch, P. (1994). "Defensive roles for secondary metabolites from marine sponges and sponge-feeding nudibranchs". In: *Toxicon* 32.6, pp. 639–655. DOI: 10.1016/0041-0101(94)90334-4 (cit. on p. 17).

Pruesse, E., J. Peplies, and F. O. Glöckner (2012). "SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes." In: *Bioinformatics* 28.14, pp. 1823–1829. DOI: 10.1093/bioinformatics/bts252 (cit. on p. 27).

Purkamo, L. et al. (2016). "Microbial co-occurrence patterns in deep Precambrian bedrock fracture fluids". In: *Biogeosciences* 13.10, pp. 3091–3108. DOI: 10.5194/bg-13-3091-2016 (cit. on p. 88).

Quast, C. et al. (2013). "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools." In: *Nucleic Acids Res* 41.Database issue, pp. D590–D596. DOI: 10.1093/nar/gks1219 (cit. on p. 25).

Quinlan, A. R. (2014). "BEDTools: The Swiss-Army Tool for Genome Feature Analysis." In: *Curr Protoc Bioinformatics* 47, pp. 11.12.1–11.1234. DOI: 10.1002/0471250953.bi1112s47 (cit. on p. 29).

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria (cit. on pp. 26, 31).

Radax, R., F. Hoffmann, H. T. Rapp, S. Leininger, and C. Schleper (2012). "Ammonia-oxidizing archaea as main drivers of nitrification in cold-water sponges." In: *Environ Microbiol* 14.4, pp. 909–923. DOI: 10.1111/j.1462-2920.2011.02661.x (cit. on p. 17).

Rahman, A. and L. Pachter (2013). “CGAL: computing genome assembly likelihoods.” In: *Genome Biol* 14.1, R8. DOI: 10.1186/gb-2013-14-1-r8 (cit. on p. 45).

Rastogi, G., G. L. Coaker, and J. H. J. Leveau (2013). “New insights into the structure and function of phyllosphere microbiota through high-throughput molecular approaches.” In: *FEMS Microbiol Lett* 348.1, pp. 1–10. DOI: 10.1111/1574-6968.12225 (cit. on p. 90).

Reddy, T. B. K. et al. (2015). “The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification.” In: *Nucleic Acids Res* 43.Database issue, pp. D1099–D1106. DOI: 10.1093/nar/gku950 (cit. on pp. 7, 25).

Redford, A. J., R. M. Bowers, R. Knight, Y. Linhart, and N. Fierer (2010). “The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves.” In: *Environ Microbiol* 12.11, pp. 2885–2893. DOI: 10.1111/j.1462-2920.2010.02258.x (cit. on pp. 15, 16).

Reed, H. E. and J. B. H. Martiny (2013). “Microbial composition affects the functioning of estuarine sediments.” In: *ISME J* 7.4, pp. 868–879. DOI: 10.1038/ismej.2012.154 (cit. on p. 94).

Reisberg, E. E., U. Hildebrandt, M. Riederer, and U. Hentschel (2013). “Distinct phyllosphere bacterial communities on Arabidopsis wax mutant leaves.” In: *PLoS One* 8.11, e78613. DOI: 10.1371/journal.pone.0078613 (cit. on p. 15).

Remus-Emsermann, M. N. P., S. de Oliveira, L. Schreiber, and J. H. J. Leveau (2011). “Quantification of lateral heterogeneity in carbohydrate permeability of isolated plant leaf cuticles.” In: *Front Microbiol* 2, p. 197. DOI: 10.3389/fmicb.2011.00197 (cit. on p. 14).

Remus-Emsermann, M. N. P., E. B. Kim, M. L. Marco, R. Tecon, and J. H. J. Leveau (2013). “Draft Genome Sequence of the Phyllosphere Model Bacterium Pantoea agglomerans 299R.” In: *Genome Announc* 1.1. DOI: 10.1128/genomeA.00036-13 (cit. on pp. 15, 88, 90).

Remus-Emsermann, M. N. P. and J. H. J. Leveau (2010). “Linking environmental heterogeneity and reproductive success at single-cell resolution.” In: *ISME J* 4.2, pp. 215–222. DOI: 10.1038/ismej.2009.110 (cit. on p. 14).

Renesto, P., N. Crapoulet, H. Ogata, B. L. Scola, G. Vestris, J.-M. Claverie, and D. Raoult (2003). “Genome-based design of a cell-free culture medium for Tropheryma whipplei”. In: *The Lancet* 362.9382, pp. 447–449. DOI: 10.1016/s0140-6736(03)14071-8 (cit. on p. 8).

Rice, P., I. Longden, and A. Bleasby (2000). “EMBOSS: the European Molecular Biology Open Software Suite.” In: *Trends Genet* 16.6, pp. 276–277. DOI: 10.1016/s0168-9525(00)02024-2 (cit. on p. 26).

Richter, M. and R. Rosselló-Móra (2009). “Shifting the genomic gold standard for the prokaryotic species definition.” In: *Proc Natl Acad Sci U S A* 106.45, pp. 19126–19131. DOI: 10.1073/pnas.0906412106 (cit. on pp. 8, 31, 89).

Ritpitakphong, U., L. Falquet, A. Vimoltust, A. Berger, J.-P. Metraux, and F. LHaridon (2016). “The microbiome of the leaf surface of Arabidopsis protects against a fungal pathogen”. In: *New Phytol* 210.3, pp. 1033–1043. DOI: 10.1111/nph.13808 (cit. on p. 14).

Roberts, R. J. et al. (2003). “A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes.” In: *Nucleic Acids Res* 31.7, pp. 1805–1812. DOI: 10.1007/978-3-642-18851-0_1 (cit. on p. 99).

Rocha, E. P., A. Danchin, and A. Viari (2001). “Evolutionary role of restriction/modification systems as revealed by comparative genome analysis.” In: *Genome Res* 11.6, pp. 946–958. DOI: 10.1101/gr.153101 (cit. on p. 100).

Romero, C. M. et al. (2006). “Genome sequence alterations detected upon passage of Burkholderia mallei ATCC 23344 in culture and in mammalian hosts”. In: *BMC Genomics* 7.1, p. 228. DOI: 10.1186/1471-2164-7-228 (cit. on p. 8).

Ruiz-González, M. X., G. Á. Czirják, P. Genevaux, A. P. Møller, T. A. Mousseau, and P. Heeb (2016). “Resistance of Feather-Associated Bacteria to Intermediate Levels of Ionizing Radiation near Chernobyl.” In: *Sci Rep* 6, p. 22969. DOI: 10.1038/srep22969 (cit. on p. 88).

Salzberg, S. L. et al. (2012). “GAGE: A critical evaluation of genome assemblies and assembly algorithms.” In: *Genome Res* 22.3, pp. 557–567. DOI: 10.1101/gr.131383.111 (cit. on p. 45).

Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *J Mol Biol* 94.3, pp. 441–448. DOI: 10.1016/b978-0-12-131200-8.50040-x (cit. on p. 4).

Sanger, F., S. Nicklen, and A. R. Coulson (1977a). "DNA sequencing with chain-terminating inhibitors". In: *Proc Natl Acad Sci U S A* 74.12, pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463 (cit. on p. 4).

Sanger, F. et al. (1977b). "Nucleotide sequence of bacteriophage phi X174 DNA." In: *Nature* 265.5596, pp. 687–695. DOI: 10.1038/265687a0 (cit. on p. 4).

Santhanam, R., V. T. Luu, A. Weinhold, J. Goldberg, Y. Oh, and I. T. Baldwin (2015). "Native root-associated bacteria rescue a plant from a sudden-wilt disease that emerged during continuous cropping". In: *Proc Natl Acad Sci U S A* 112.36, E5013–E5020. DOI: 10.1073/pnas.1505765112 (cit. on p. 14).

Schloss, P. D., D. Gevers, and S. L. Westcott (2011). "Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies." In: *PLoS One* 6.12, e27310. DOI: 10.1371/journal.pone.0027310 (cit. on p. 94).

Schmidt, T. S. B., J. F. Matias Rodrigues, and C. von Mering (2015). "Limits to robustness and reproducibility in the demarcation of operational taxonomic units." In: *Environ Microbiol* 17.5, pp. 1689–1706. DOI: 10.1111/1462-2920.12610 (cit. on p. 94).

Schmieder, R. and R. Edwards (2011). "Fast identification and removal of sequence contamination from genomic and metagenomic datasets." In: *PLoS One* 6.3, e17288. DOI: 10.1371/journal.pone.0017288 (cit. on pp. 11, 38, 43).

Schmitt, S. et al. (2012). "Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges." In: *ISME J* 6.3, pp. 564–576. DOI: 10.1038/ismej.2011.116 (cit. on p. 16).

Schultz, J., F. Milpetz, P. Bork, and C. P. Ponting (1998). "SMART, a simple modular architecture research tool: identification of signaling domains." In: *Proc Natl Acad Sci U S A* 95.11, pp. 5857–5864. DOI: 10.1073/pnas.95.11.5857 (cit. on pp. 25, 30).

Seah, B. K. and H. R. Gruber-Vodicka (2015). "gbtools: Interactive Visualization of Metagenome Bins in R". In: *Front Microbiol* 6, p. 1451. DOI: 10.3389/fmicb.2015.01451 (cit. on p. 45).

Seed, K. D. (2015). "Battling Phages: How Bacteria Defend against Viral Attack." In: *PLoS Pathog* 11.6, e1004847. DOI: 10.1371/journal.ppat.1004847 (cit. on pp. 97, 100).

Seemann, T. (2014). "Prokka: rapid prokaryotic genome annotation." In: *Bioinformatics* 30.14, pp. 2068–2069. DOI: 10.1093/bioinformatics/btu153 (cit. on pp. 29, 39).

Sequencing Consortium, T. C. elegans (1998). "Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology". In: *Science* 282.5396, pp. 2012–2018. DOI: 10.1126/science.282.5396.2012 (cit. on p. 7).

Shade, A., C. S. Hogan, A. K. Klimowicz, M. Linske, P. S. McManus, and J. Handelsman (2012). "Culturing captures members of the soil rare biosphere". In: *Environ Microbiol* 14.9, pp. 2247–2252. DOI: 10.1111/j.1462-2920.2012.02817.x (cit. on p. 88).

Sharpton, T. J. (2014). "An introduction to the analysis of shotgun metagenomic data." In: *Front Plant Sci* 5, p. 209. DOI: 10.3389/fpls.2014.00209 (cit. on pp. 10, 12, 95).

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." In: *Nat Biotechnol* 26.10, pp. 1135–1145. DOI: 10.1038/nbt1486 (cit. on pp. 5, 10).

Siegl, A. and U. Hentschel (2010). "PKS and NRPS gene clusters from microbial symbiont cells of marine sponges by whole genome amplification." In: *Environ Microbiol Rep* 2.4, pp. 507–513. DOI: 10.1111/j.1758-2229.2009.00057.x (cit. on p. 18).

Siegl, A. et al. (2011). "Single-cell genomics reveals the lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges." In: *ISME J* 5.1, pp. 61–70. DOI: 10.1038/ismej.2010.95 (cit. on p. 9).

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol (2009). "ABySS: a parallel assembler for short read sequence data." In: *Genome Res* 19.6, pp. 1117–1123. DOI: 10.1101/gr.089532.108 (cit. on p. 36).

Sinclair, L., O. A. Osman, S. Bertilsson, and A. Eiler (2015). "Microbial community composition and diversity via 16S rRNA gene amplicons: evaluating the illumina platform." In: *PLoS One* 10.2, e0116955. DOI: 10.1371/journal.pone.0116955 (cit. on p. 94).

Skennerton, C. T., M. Imelfort, and G. W. Tyson (2013). "Crass: identification and reconstruction of CRISPR from unassembled metagenomic data." In: *Nucleic Acids Res* 41.10, e105. DOI: 10.1093/nar/gkt183 (cit. on p. 85).

Slaby, B. M., T. Hackl, H. **Horn**, K. Bayer, and U. Hentschel (2017). "Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization". In: *ISME J*. In review.

Sonnhammer, E. L. L. and G. Östlund (2015). "InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic." In: *Nucleic Acids Res* 43.Database issue, pp. D234–D239. DOI: 10.1093/nar/gku1203 (cit. on p. 31).

Spratt, B. G. and M. C. J. Maiden (1999). "Bacterial population genetics, evolution and epidemiology". In: *Philos Trans R Soc Lond B Biol Sci* 354.1384, pp. 701–710. DOI: 10.1098/rstb.1999.0423 (cit. on p. 8).

Stackebrandt, E. and B. Goebel (1994). "Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology". In: *Int J Syst Evol Microbiol* 44.4, pp. 846–849. DOI: 10.1099/00207713-44-4-846 (cit. on p. 89).

Staden, R. (1979). "A strategy of DNA sequencing employing computer programs." In: *Nucleic Acids Res* 6.7, pp. 2601–2610. DOI: 10.1093/nar/6.7.2601 (cit. on p. 4).

Staley, J. T. and A. Konopka (1985). "Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats." In: *Annu Rev Microbiol* 39, pp. 321–346. DOI: 10.1146/annurev.mi.39.100185.001541 (cit. on p. 9).

Stamatakis, A. (2006). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." In: *Bioinformatics* 22.21, pp. 2688–2690. DOI: 10.1093/bioinformatics/btl446 (cit. on p. 27).

Steen, E. J. et al. (2010). "Microbial production of fatty-acid-derived fuels and chemicals from plant biomass". In: *Nature* 463.7280, pp. 559–562. DOI: 10.1038/nature08721 (cit. on p. 8).

Stein, J. L., T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. DeLong (1996). "Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon." In: *J Bacteriol* 178.3, pp. 591–599. DOI: 10.1046/j.1462-2920.2001.00198.x (cit. on p. 9).

Steindler, L., S. Beer, and M. Ilan (2002). "Photosymbiosis in intertidal and subtidal tropical sponges". In: *Symbiosis* 33.3, pp. 263–274. DOI: 10.1007/s002270000482 (cit. on p. 17).

Stephens, Z. D. et al. (2015). "Big Data: Astronomical or Genomical?" In: *PLoS Biol* 13.7, e1002195. DOI: 10.1371/journal.pbio.1002195 (cit. on p. 10).

Stern, A. and R. Sorek (2011). "The phage-host arms race: shaping the evolution of microbes." In: *Bioessays* 33.1, pp. 43–51. DOI: 10.1002/bies.201000071 (cit. on p. 97).

Stoesser, G., P. Sterk, M. A. Tuli, P. J. Stoehr, and G. N. Cameron (1997). "The EMBL Nucleotide Sequence Database." In: *Nucleic Acids Res* 25.1, pp. 7–14. DOI: 10.1093/nar/25.1.7 (cit. on p. 25).

Stöver, B. C. and K. F. Müller (2010). "TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses." In: *BMC Bioinformatics* 11, p. 7. DOI: 10.1186/1471-2105-11-7 (cit. on p. 27).

Tatusov, R. L., E. V. Koonin, and D. J. Lipman (1997). "A genomic perspective on protein families." In: *Science* 278.5338, pp. 631–637. DOI: 10.1126/science.278.5338.631 (cit. on p. 25).

Taylor, M. W., R. Radax, D. Steger, and M. Wagner (2007). "Sponge-associated microorganisms: evolution, ecology, and biotechnological potential." In: *Microbiol Mol Biol Rev* 71.2, pp. 295–347. DOI: 10.1128/MMBR.00040-06 (cit. on pp. 16, 17).

Taylor, M. W., P. J. Schupp, R. De Nys, S. Kjelleberg, and P. D. Steinberg (2005). "Biogeography of bacteria associated with the marine sponge Cymbastela concentrica". In: *Environ Microbiol* 7.3, pp. 419–433. DOI: 10.1111/j.1462-2920.2004.00711.x (cit. on p. 16).

Taylor, M. W. et al. (2013). "'Sponge-specific' bacteria are widespread (but rare) in diverse marine environments.” In: *ISME J* 7.2, pp. 438–443. DOI: 10.1038/ismej.2012.111 (cit. on p. 16).

tenOever, B. R. (2016). “The Evolution of Antiviral Defense Systems.” In: *Cell Host Microbe* 19.2, pp. 142–149. DOI: 10.1016/j.chom.2016.01.006 (cit. on p. 97).

Thomas, P. D. et al. (2003). “PANTHER: a library of protein families and subfamilies indexed by function.” In: *Genome Res* 13.9, pp. 2129–2141. DOI: 10.1101/gr.772403 (cit. on pp. 25, 30).

Thomas, T., J. Gilbert, and F. Meyer (2012). “Metagenomics - a guide from sampling to data analysis.” In: *Microb Inform Exp* 2.1, p. 3. DOI: 10.1186/2042-5783-2-3 (cit. on pp. 6, 10–12, 85, 86).

Thomas, T. et al. (2010). “Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis.” In: *ISME J* 4.12, pp. 1557–1567. DOI: 10.1038/ismej.2010.74 (cit. on p. 17).

Thomas, T. et al. (2016). “Diversity, structure and convergent evolution of the global sponge microbiome.” In: *Nat Commun* 7, p. 11870. DOI: 10.1038/ncomms11870 (cit. on pp. 17, 94).

Thoms, C., M. Horn, M. Wagner, U. Hentschel, and P. Proksch (2003). “Monitoring microbial diversity and natural product profiles of the sponge Aplysina cavernicola following transplantation”. In: *Mar Biol* 142.4, pp. 685–692. DOI: 10.1007/s002270050562 (cit. on p. 16).

Tian, R.-M., J. Sun, L. Cai, W.-P. Zhang, G.-W. Zhou, J.-W. Qiu, and P.-Y. Qian (2016). “The deep-sea glass sponge Lophophysema eversa harbours potential symbionts responsible for the nutrient conversions of carbon, nitrogen and sulfur”. In: *Environ Microbiol.* DOI: 10.1111/1462-2920.13161 (cit. on p. 99).

Trindade-Silva, A. E., C. P. J. Rua, B. G. N. Andrade, A. C. P. Vicente, G. G. Z. Silva, R. G. S. Berlinck, and F. L. Thompson (2013). “Polyketide synthase gene diversity within the microbiome of the sponge Arenosclera brasiliensis, endemic to the Southern Atlantic Ocean.” In: *Appl Environ Microbiol* 79.5, pp. 1598–1605. DOI: 10.1128/AEM.03354-12 (cit. on p. 95).

Trindade-Silva, A. E. et al. (2012). “Taxonomic and functional microbial signatures of the endemic marine sponge Arenosclera brasiliensis.” In: *PLoS One* 7.7, e39905. DOI: 10.1371/journal.pone.0039905 (cit. on p. 17).

Tringe, S. G. (2005). “Comparative Metagenomics of Microbial Communities”. In: *Science* 308.5721, pp. 554–557. DOI: 10.1126/science.1107851 (cit. on p. 12).

Tringe, S. G. et al. (2005). “Comparative metagenomics of microbial communities.” In: *Science* 308.5721, pp. 554–557. DOI: 10.1126/science.1107851 (cit. on p. 10).

Tyson, G. W. et al. (2004). “Community structure and metabolism through reconstruction of microbial genomes from the environment.” In: *Nature* 428.6978, pp. 37–43. DOI: 10.1038/nature02340 (cit. on pp. 10, 13).

Vacelet, J. (1975). “Etude en microscopie electronique de l’association entre bacteries et spongiaires du genre Verongia (Dictyoceratida)”. In: *J Microsc Biol Cell* 23.3, pp. 271–288. DOI: 10.1016/0248-4900(92)90183-2 (cit. on p. 16).

Vacelet, J., A. Fiala-Medioni, C. Fisher, and N. Boury-Esnault (1996). “Symbiosis between methane-oxidizing bacteria and a deep-sea carnivorous cladorhizid sponge”. In: *Mar Ecol Prog Ser* 145, pp. 77–85. DOI: 10.3354/meps145077 (cit. on p. 17).

Vacelet, J., N. Boury-Esnault, A. Fiala-Medioni, and C. Fisher (1995). “A methanotrophic carnivorous sponge”. In: *Nature* 6547. DOI: 10.1038/377296a0 (cit. on p. 17).

Varghese, N. J., S. Mukherjee, N. Ivanova, K. T. Konstantinidis, K. Mavrommatis, N. C. Kyrpides, and A. Pati (2015). “Microbial species delineation using whole genome sequences.” In: *Nucleic Acids Res* 43.14, pp. 6761–6771. DOI: 10.1093/nar/gkv657 (cit. on p. 89).

Vasu, K. and V. Nagaraja (2013). “Diverse functions of restriction-modification systems in addition to cellular defense.” In: *Microbiol Mol Biol Rev* 77.1, pp. 53–72. DOI: 10.1128/MMBR.00044-12 (cit. on p. 99).

Venter, J. C. et al. (2004). “Environmental genome shotgun sequencing of the Sargasso Sea.” In: *Science* 304.5667, pp. 66–74. DOI: 10.1126/science.1093857 (cit. on p. 10).

Venter, J. C. et al. (2001). “The sequence of the human genome.” In: *Science* 291.5507, pp. 1304–

1351. DOI: 10.1126/science.1058040 (cit. on pp. 4, 7).

Větrovský, T. and P. Baldrian (2013). "The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses." In: *PLoS One* 8.2, e57923. DOI: 10.1371/journal.pone.0057923 (cit. on p. 86).

Vogel, S. (1977). "Current-induced flow through living sponges in nature." In: *Proc Natl Acad Sci U S A* 74.5, pp. 2069–2071. DOI: 10.1073/pnas.74.5.2069 (cit. on p. 16).

Vogel, T. M. et al. (2009). "TerraGenome: a consortium for the sequencing of a soil metagenome". In: *Nat Rev Microbiol* 7.4, p. 252. DOI: 10.1038/nrmicro2119 (cit. on p. 10).

Vorholt, J. A. (2012). "Microbial life in the phyllosphere." In: *Nat Rev Microbiol* 10.12, pp. 828–840. DOI: 10.1038/nrmicro2910 (cit. on pp. 14, 15, 19).

Wagner, M. R., D. S. Lundberg, T. G. del Rio, S. G. Tringe, J. L. Dangl, and T. Mitchell-Olds (2016). "Host genotype and age shape the leaf and root microbiomes of a wild perennial plant". In: *Nat Commun* 7, p. 12151. DOI: 10.1038/ncomms12151 (cit. on p. 14).

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." In: *Appl Environ Microbiol* 73.16, pp. 5261–5267. DOI: 10.1128/AEM.00062-07 (cit. on p. 27).

Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid". In: *Nature* 171.4356, pp. 737–738. DOI: 10.1097/blo.0b013e31814b9304 (cit. on pp. 3, 13).

Weber, T. and H. U. Kim (2016). "The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production". In: *Synth Syst Biotechnol* 1.2, pp. 69–79. DOI: 10.1016/j.synbio.2015.12.002 (cit. on p. 92).

Weber, T. et al. (2015). "antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters." In: *Nucleic Acids Res* 43.W1, W237–W243. DOI: 10.1093/nar/gkv437 (cit. on pp. 12, 31, 92).

Webster, N. S. and M. W. Taylor (2012). "Marine sponges and their microbial symbionts: love and other relationships." In: *Environ Microbiol* 14.2, pp. 335–346. DOI: 10.1111/j.1462-2920.2011.02460.x (cit. on p. 94).

Webster, N. S. and T. Thomas (2016). "The Sponge Hologenome". In: *mBio* 7.2, e00135–16. DOI: 10.1128/mBio.00135-16 (cit. on p. 18).

Webster, N. S. et al. (2010). "Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts." In: *Environ Microbiol* 12.8, pp. 2070–2082. DOI: 10.1111/j.1462-2920.2009.02065.x (cit. on p. 16).

Webster, N. and R. Hill (2001). "The culturable microbial community of the Great Barrier Reef sponge Rhopaloeides odorabile is dominated by an α-Proteobacterium". In: *Mar Biol* 138.4, pp. 843–851. DOI: 10.1007/s002270000503 (cit. on p. 16).

Welch, R. A. et al. (2002). "Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli". In: *Proc Natl Acad Sci U S A* 99.26, pp. 17020–17024. DOI: 10.1073/pnas.252529799 (cit. on p. 8).

Wetterstrand, K. A. (2016). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. URL: https://www.genome.gov/sequencingcostsdata/ (visited on May 41, 2016) (cit. on p. 4).

White, O. (1999). "Genome Sequence of the Radioresistant Bacterium Deinococcus radiodurans R1". In: *Science* 286.5444, pp. 1571–1577. DOI: 10.1126/science.286.5444.1571 (cit. on p. 8).

Wiens, M., M. Korzhev, S. Perovic-Ottstadt, B. Luthringer, D. Brandt, S. Klein, and W. E. G. Müller (2007). "Toll-like receptors are part of the innate immune defense system of sponges (demospongiae: Porifera)." In: *Mol Biol Evol* 24.3, pp. 792–804. DOI: 10.1093/molbev/msl208 (cit. on p. 16).

Wilkinson, C. C., R. R. Summons, E. Evans, et al. (1999). "Nitrogen fixation in symbiotic marine sponges: ecological significance and difficulties in detection". In: *Memoirs of the Queensland Museum-pages: 44: 667-673*. DOI: 10.1007/bf00403507 (cit. on p. 17).

Wilkinson, C. R. (1983). "Net primary productivity in coral reef sponges". In: *Science* 219.4583, pp. 410–412. DOI: 10.1126/science.219.4583.410 (cit. on p. 17).

Williams, T. R. and M. L. Marco (2014). "Phyllosphere microbiota composition and microbial community transplantation on lettuce plants grown indoors." In: *MBio* 5.4. DOI: 10.1128/mBio.01564-14 (cit. on p. 15).

Williams, T. R., A.-L. Moyne, L. J. Harris, and M. L. Marco (2013). "Season, irrigation, leaf age, and Escherichia coli inoculation influence the bacterial diversity in the lettuce phyllosphere." In: *PLoS One* 8.7, e68642. DOI: 10.1371/journal.pone.0068642 (cit. on p. 15).

Wilson, M. C. and J. Piel (2013). "Metagenomic approaches for exploiting uncultivated bacteria as a resource for novel biosynthetic enzymology." In: *Chem Biol* 20.5, pp. 636–647. DOI: 10.1016/j.chembiol.2013.04.011 (cit. on p. 86).

Wilson, M. C. et al. (2014). "An environmental bacterial taxon with a large and distinct metabolic repertoire". In: *Nature* 506.7486, pp. 58–62. DOI: 10.1038/nature12959 (cit. on p. 9).

Wommack, K. E., J. Bhavsar, and J. Ravel (2008). "Metagenomics: read length matters." In: *Appl Environ Microbiol* 74.5, pp. 1453–1463. DOI: 10.1128/AEM.02181-07 (cit. on p. 85).

Wood, D. E. and S. L. Salzberg (2014). "Kraken: ultrafast metagenomic sequence classification using exact alignments." In: *Genome Biol* 15.3, R46. DOI: 10.1186/gb-2014-15-3-r46 (cit. on p. 45).

Wu, R. and E. Taylor (1971). "Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA." In: *J Mol Biol* 57.3, pp. 491–511. DOI: 10.1016/0022-2836(71)90105-7 (cit. on p. 4).

Wu, R. and A. Kaiser (1968). "Structure and base sequence in the cohesive ends of bacteriophage lambda DNA". In: *J Mol Biol* 35.3, pp. 523–537. DOI: 10.1016/s0022-2836(68)80012-9 (cit. on p. 4).

Xia, X. (2013). "Comparative Genomics and the Comparative Methods". In: *Comp Genomics*. DOI: 10.1007/978-3-642-37146-2_2 (cit. on p. 8).

Xiong, Z.-Q., J.-F. Wang, Y.-Y. Hao, and Y. Wang (2013). "Recent advances in the discovery and development of marine microbial natural products." In: *Mar Drugs* 11.3, pp. 700–717. DOI: 10.3390/md11030700 (cit. on p. 92).

Yadav, S. P. (2007). "The wholeness in suffix -omics, -omes, and the word om." In: *J Biomol Tech* 18.5, p. 277. DOI: 10.1089/omi.2006.00e1 (cit. on p. 6).

Yahel, G., J. H. Sharp, D. Marie, C. Häse, and A. Genin (2003). "In situ feeding and element removal in the symbiont-bearing sponge Theonella swinhoei: Bulk DOC is the major source for carbon". In: *Limnol Oceanogr* 48.1, pp. 141–149. DOI: 10.4319/lo.2003.48.1.0141 (cit. on p. 17).

Yassin, A. F. and H. Hupfer (2006). "Williamsia deligens sp. nov., isolated from human blood." In: *Int J Syst Evol Microbiol* 56.Pt 1, pp. 193–197. DOI: 10.1099/ijs.0.63856-0 (cit. on p. 88).

Yu, D., F. Xu, J. Valiente, S. Wang, and J. Zhan (2013). "An indigoidine biosynthetic gene cluster from Streptomyces chromofuscus ATCC 49982 contains an unusual IndB homologue." In: *J Ind Microbiol Biotechnol* 40.1, pp. 159–168. DOI: 10.1007/s10295-012-1207-9 (cit. on p. 92).

Zallen, D. T. (2003). "Despite Franklin's work, Wilkins earned his Nobel". In: *Nature* 425.6953, pp. 15–15. DOI: 10.1038/425015b (cit. on p. 3).

Zerbino, D. R. (2010). "Using the Velvet de novo assembler for short-read sequencing technologies." In: *Curr Protoc Bioinformatics* Chapter 11, Unit 11.5. DOI: 10.1002/0471250953.bi1105s31 (cit. on pp. 36, 38).

Ziemert, N., S. Podell, K. Penn, J. H. Badger, E. Allen, and P. R. Jensen (2012). "The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity." In: *PLoS One* 7.3, e34064. DOI: 10.1371/journal.pone.0034064 (cit. on pp. 31, 92).

Zimin, A. V., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke (2013). "The MaSuRCA genome assembler." In: *Bioinformatics* 29.21, pp. 2669–2677. DOI: 10.1093/bioinformatics/btt476 (cit. on pp. 36, 38).

Part VII.

Appendix

Abbreviations

ALE	Assembly likelihood estimator 36, 43
ANI	average nucleotide identity 8, 29, 116, 117
BLAST	Basic Local Alignment and Search Tool 11, 25, 28, 29, 36, 43
CDD	Conserved Domains Database 28
CG–pipeline	Computational Genomics–pipeline 34, 43
CGAL	Computing Genome Assembly Likelihoods 43
CISA	Contig Integrator for Sequence Assembly 43
COG	Clusters of Orthologous Groups 10, 23, 28, 29, 116
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats 10, 16, 17, 28, 37, 113, 124
DDH	DNA–DNA hybridization 116
DGGE	denaturating gradient gel electrophoresis 14
DNA	desoxyribonucleic acid 3, 4, 6, 8, 9, 12, 113, 114
FISH	flueorescence in situ hybridization 14
GAGE–B	Genome Assembly Gold-standard Evaluation for Bacteria 34
GEBA	Genomic Encyclopedia of Bacteria and Archaea 7
GOLD	The Genomes OnLine Database 7, 9, 23
GWAS	genome—wide association study 14
HGT	horizontal gene transfer 114, 117
HMA	high microbial abundance 14
HMM	hidden–markov model 11, 29
HMP	Human Microbiome Project 9
IMG	Integrated Microbial Genomes 27
LAP	Log Average Probability 43
LCA	lowest common ancestor 26
LMA	low microbial abundance 14
MG–RAST	Metagenomic Rapid Annotations using Subsystems Technology 26
NaPDoS	Natural Product Domain Seeker 29
ncRNA	non–coding RNA 27, 37
NGS	next–generation sequencing 5–7, 9, 149
NHGRI	National Human Genome Research Institute 4
NR	NCBI non-redundant protein sequences 23
NRPS	nonribosomal peptide synthetase 16, 17, 21, 24, 25, 29
NT	NCBI nucleotide collection 23, 36
ORF	open reading frame 10, 27–29, 113
PANTHER	Protein ANalysis THrough Evolutionary Relationships 23, 28

PCR	Polymerase Chain Reaction 123
PFAM	Protein Families 10, 23, 28
PKS	polyketide synthase 8, 16, 17, 21, 24, 25, 29, 114
QUAST	Quality Assessment Tool for Genome Assemblies 36, 38, 41, 43
REAPR	Recognition of Errors in Assemblies using Paired Reads 43
REBASE	The Restriction Enzyme Database 23, 28
RMS	restriction modification system 17, 23, 28, 29
RNA	ribonucleic acid 4, 6, 7, 27, 117
RPS BLAST	Reversed Position Specific BLAST 28
rRNA	ribosomal RNA 8, 14, 27, 37, 114–116, 123
SMART	Simple Modular Architectur Research Tool 23, 28
SMS	single molecule sequencing 5
SNP	Single nucleotide polymorphism 7
SOLiD	Sequencing by Oligonucleotide Ligation and Detection 5
tRNA	transfer RNA 3, 27, 37

List of figures

List of tables

Statement of author contributions

Author contributions to the manuscripts

Publication (complete reference):

Horn, H., C. Cheng, R. Edrada-Ebel, U. Hentschel, and U. R. Abdelmohsen (2015a).
"Draft genome sequences of three chemically rich actinomycetes isolated from Mediterranean sponges." In: Mar Genomics 24 Pt 3, pp. 285–287. DOI: 10.1016/j.margen.2015.10.003.

Participated in	**Author initials:** Responsibility decreasing from left to right				
Study Design	URA	UHe	RAE	**HH**	CC
Methods Development	**HH**	CC	URA	UHe	RAE
Data Collection	CC	**HH**	URA	UHe	RAE
Data Analysis and Interpretation	**HH**	CC	URA	UHe	RAE
Manuscript Writing					
Writing of Introduction	**HH**	URA	CC	UHe	RAE
Writing of Materials and Methods	**HH**	URA	CC	UHe	RAE
Writing of Discussion	**HH**	URA	CC	UHe	RAE
Writing of Frist Draft	**HH**	URA	CC	UHe	RAE

Publication (complete reference):

Horn, H., U. Hentschel, and U. R. Abdelmohsen (2015b).
"Mining genomes of three marine sponge-associated actinobacerial isolates for secondary metabolism" In: Genome Announc 3.5,. DOI: 10.1128/genomeA.01106-15.

Participated in	**Author initials:** Responsibility decreasing from left to right		
Study Design	URA	**HH**	UHe
Methods Development	**HH**	URA	UHe
Data Collection	**HH**	URA	UHe
Data Analysis and Interpretation	**HH**	URA	UHe
Manuscript Writing			
Writing of Introduction	**HH**	URA	UHe
Writing of Materials & Methods	**HH**	URA	UHe
Writing of Discussion	**HH**	URA	UHe
Writing of Frist Draft	**HH**	URA	UHe

Publication (complete reference):

Horn, H., A. Keller, U. Hildebrandt, P. Kämpfer, M. Riederer, and U. Hentschel (2016a). “Draft genome of the *Arabidopsis thaliana* phyllosphere bacterium, *Williamsia* sp. ARP1.” In: Stand Genomic Sci 11, p. 8. DOI: 10.1186/s40793-015-0122-x.

Participated in	**Author initials:** Responsibility decreasing from left to right					
Study Design	**HH**	UHe	UHi			
Methods Development	**HH**	UHe	AK			
Data Collection	AK	**HH**	PK	UHi		
Data Analysis and Interpretation	**HH**	AK	PK	UHe	MR	UHi
Manuscript Writing						
Writing of Introduction	**HH**	UHe	AK	MR	UHi	PK
Writing of Materials & Methods	**HH**	UHe	AK	PK	MR	UHi
Writing of Discussion	**HH**	UHe	AK	MR	UHi	PK
Writing of Frist Draft	**HH**	Uhe	AK	MR	UHi	PK

Publication (complete reference):

Horn, H., B. M. Slaby, M. T. Jahn, K. Bayer, F. Foerster, U. R. Abdelmohsen, and U. Hentschel (2016b). An enrichment of CRISPR and other defense–related features in marine sponge-associated microbial metagenomes.” In: Front Microbiol 7, p. 1751. DOI 10.3389/fmicb.2016.01751.

Participated in	**Author initials:** Responsibility decreasing from left to right						
Study Design	UHe	**HH**	KB	LMS	URA		
Methods Development	**HH**	FF	MTJ				
Data Collection	KB	BMS	LMS	**HH**			
Data Analysis and Interpretation	**HH**	BMS	MTJ	KB	URA	FF	UHe
Manuscript Writing							
Writing of Introduction	**HH**	URA	KB				
Writing of Materials & Methods	**HH**	BMS	KB				
Writing of Discussion	**HH**	URA	KB				
Writing of Frist Draft	**HH**	URA	UHe				

Author contributions to figures and tables

Publication (complete reference):

Horn, H., C. Cheng, R. Edrada-Ebel, U. Hentschel, and U. R. Abdelmohsen (2015). "Draft genome sequences of three chemically rich actinomycetes isolated from Mediterranean sponges." In: Mar Genomics 24 Pt 3, pp. 285–287. DOI: 10.1016/j.margen.2015.10.003.

Participated in	**Author initials:** Responsibility decreasing from left to right	
Table 1	**HH**	URA
Table 2	**HH**	URA
Table 3	**HH**	URA

Publication (complete reference):

Horn, H., U. Hentschel, and U. R. Abdelmohsen (2015). "Mining genomes of three marine sponge-associated actinobacerial isolates for secondary metabolism" In: Genome Announc 3.5,. DOI: 10.1128/genomeA.01106-15.

Participated in	**Author initials:** Responsibility decreasing from left to right
Contains no figures or tables	

Publication (complete reference):

Horn, H., A. Keller, U. Hildebrandt, P. Kämpfer, M. Riederer, and U. Hentschel (2016a). "Draft genome of the *Arabidopsis thaliana* phyllosphere bacterium, *Williamsia* sp. ARP1." In: Stand Genomic Sci 11, p. 8. DOI: 10.1186/s40793-015-0122-x.

Participated in	**Author initials:** Responsibility decreasing from left to right		
Figure 1	**HH**	PK	UHe
Figure 2	**HH**	UHe	PK
Figure 3	**HH**		
Figure 4	**HH**		
Figure 5	**HH**		
Table 1	**HH**		
Table 2	**HH**	AK	
Table 3	**HH**		
Table 4	**HH**		
Table 5	**HH**		

Publication (complete reference):

Horn, H., B. M. Slaby, M. T. Jahn, K. Bayer, F. Foerster, U. R. Abdelmohsen, and U. Hentschel (2016b). An enrichment of CRISPR and other defense–related features in marine sponge-associated microbial metagenomes.” In: Front Microbiol 7, p. 1751. DOI: 10.3389/fmicb.2016.01751.

Participated in	**Author initials:** Responsibility decreasing from left to right		
Figure 1	**HH**		
Figure 2	**HH**	MTJ	KB
Figure 3	**HH**	MTJ	BMS
Figure 4	**HH**	MTJ	BMS
Figure 5	**HH**	FF	
Figure 6	**HH**	MTJ	
Figure 7	**HH**		
Figure 8	**HH**	FF	
Figure 9	**HH**	URA	
Table 1	**HH**		
Table 2	**HH**		
Table 3	**HH**	UHe	

The doctoral researcher confirms that he has obtained permission from both the publishers and the co-authors for legal second publication. The doctoral researcher and the primary supervisor confirm the correctness of the above mentioned assessment. I further confirm my primary supervisors acceptance.

Place, Date

Hannes Horn

Publications and presentations

Publications associated with this thesis

Horn, H., C. Cheng, R. Edrada-Ebel, U. Hentschel, and U. R. Abdelmohsen (2015a). "Draft genome sequences of three chemically rich actinomycetes isolated from Mediterranean sponges." In: *Mar Genomics* 24 Pt 3, pp. 285–287. DOI: 10.1016/j.margen.2015.10.003 (cit. on p. 37).

Horn, H., U. Hentschel, and U. R. Abdelmohsen (2015b). "Mining genomes of three marine sponge-associated actinobacterial isolates for secondary metabolism." In: *Genome Announc* 3.5. DOI: 10.1128/genomeA.01106-15 (cit. on p. 37).

Horn, H., A. Keller, U. Hildebrandt, P. Kämpfer, M. Riederer, and U. Hentschel (2016a). "Draft genome of the *Arabidopsis thaliana* phyllosphere bacterium, *Williamsia* sp. ARP1." In: *Stand Genomic Sci* 11, p. 8. DOI: 10.1186/s40793-015-0122-x (cit. on pp. 15, 37, 88).

Horn, H., B. M. Slaby, M. T. Jahn, K. Bayer, F. Foerster, U. R. Abdelmohsen, and U. Hentschel (2016b). "An enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes." In: *Front Microbiol* 7, p. 1751. DOI: 10.3389/fmicb.2016.01751.

Other publications

Abdelmohsen, U. R., C. Yang, H. **Horn**, D. Hajjar, T. Ravasi, and U. Hentschel (2014b). "Actinomycetes from Red Sea sponges: sources for chemical and phylogenetic diversity." In: *Mar Drugs* 12.5, pp. 2771–2789. DOI: 10.3390/md12052771.

Cheng, C., L. MacIntyre, U. R. Abdelmohsen, H. **Horn**, P. N. Polymenakou, R. Edrada-Ebel, and U. Hentschel (2015). "Biodiversity, anti-trypanosomal activity screening, and metabolomic profiling of actinomycetes isolated from Mediterranean sponges." In: *PLoS One* 10.9, e0138528. DOI: 10.1371/journal.pone.0138528 (cit. on p. 18).

Harjes, J., T. Ryu, U. R. Abdelmohsen, L. Moitinho-Silva, H. **Horn**, T. Ravasi, and U. Hentschel (2014). "Draft genome sequence of the antitrypanosomally active sponge-associated bacterium *Actinokineospora* sp. strain EG49." In: *Genome Announc* 2.2. DOI: 10.1128/genomeA.00160-14.

Kämpfer, P., H.-J. Busse, H. **Horn**, U. R. Abdelmohsen, U. H. Hentschel, and S. P. Glaeser (2016). "*Williamsia herbipolensis* sp. nov., isolated from the phyllosphere of *Arabidopsis thaliana*". In: *Int J Syst Evol Microbiol* 66.11, pp. 4609–4613. DOI: 10.1099/ijsem.0.001398 (cit. on p. 90).

Keller, A., H. **Horn**, F. Förster, and J. Schultz (2014). "Computational integration of genomic traits into 16S rDNA microbiota sequencing studies." In: *Gene* 549.1, pp. 186–191. DOI: 10.1016/j.gene.2014.07.066.

Manuscripts in preparation

Hansjakob, A., K. U. Foerstner, H. **Horn**, M. Riederer, and U. Hildebrand (2017). "Surface dependent gene expression of *Blumeria graminis* f. sp. hordei during the prepenetration processes". In: *New Phytol.* In preparation.

Slaby, B. M., T. Hackl, H. **Horn**, K. Bayer, and U. Hentschel (2017). "Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization". In: *ISME J.* In review.

Contributions to meetings and symposia

Horn, H. (2014). "Biodiversity and nutritional targets of phyllosphere bacteria (Talk)". In: Annual Retreat of the GK1342 - Lipid Signaling. Hausen, Germany.

– (2015). "The impact of phyllosheric bacteria on cuticular wax components (Talk)". In: Annual Retreat of the GK1342 - Lipid Signaling. Oberhof, Germany.

Horn, H., A. Keller, U. Hildebrandt, P. Kämpfer, M. Riederer, and U. Hentschel (2015c). "Draft genome of the *Arabidopsis thaliana* phyllosphere bacterium, *Williamsia* sp. ARP1. (Poster)". In: Poster6th European Conference on Prokaryotic and Fungal Genomics. Göttingen, Germany.

Horn, H., A. Trimbach, M. Remus-Emsermann, U. Hildebrandt, M. Riederer, and U. Hentschel (2015d). "Bacteria in the *Arabidopsis thaliana* phyllosphere: Genomic adaptions and the cuticular waxes as possible nutritional targets (Poster)". In: 10th International Symposium on Phyllosphere Microbiology. Ascona, Switzerland.

Reisberg, E. E., H. **Horn**, U. Hildebrandt, P. Kämpfer, M. Riederer, and U. Hentschel (2014). "*Arabidopsis thaliana* wax mutants influence the phyllosphere microbiome (Poster)". In: XVI International Congress on Molecular Plant-Microbe Interactions. Rhodes, Greece.

Curriculum vitae

Affidavit

I hereby confirm that my thesis entitled "Analysis and interpretation of (meta-)genomic data from host-associated microorganisms" is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Hiermit erkläre ich an Eides statt, die Dissertation "Analyse und Interpretation von (meta-)genomischen Daten aus Wirt-assoziierten Mikroorganismen" eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen, als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Place, Date

Hannes Horn