UNIVERSITY OF CALIFORNIA
SANTA CRUZ

DIVERSITY, FORM, AND FUNCTION WITHIN THE MARINE MAMMAL MICROBIOTA

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ECOLOGY AND EVOLUTIONARY BIOLOGY

by

Natasha K. Dudek

September 2018

The Dissertation of Natasha K. Dudek
is approved:

Professor Beth Shapiro, Chair

Professor David Relman

Professor Daniel Costa

Professor Richard Green

Lori Kletzer
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures **vii**

List of Tables **x**

Abstract **xi**

Acknowledgments **xiii**

1 Introduction **1**
1.1 Co-evolution of animals and their microbiota 1
1.2 The marine mammal microbiome 2
1.3 Outline of research presented in this thesis 4

2 Characterization of the gingival and gut microbiota of the Southern sea otter *(Enhydra lutris nereis)* **8**
2.1 Abstract 8
2.2 Introduction 9
2.3 Results 12
2.3.1 Sea otter sample collection 12
2.3.2 Taxonomic composition and alpha diversity of gingival and rectal sea otter samples 14
2.3.3 Sea otter distal gut microbiota is distinct from that of other otters 16
2.3.4 Sea otter body site communities are distinct from those of adjacent seawater and from one another 17
2.3.5 Wild sea otter gingival communities conform to one of three composition types 21
2.3.6 *Helicobacter* in sea otter samples 24
2.3.7 Sources of DNA and taxonomic composition of the sea otter fecal metagenome 24

2.3.8 Marine mammal gut microbiome may have increased potential for the degradation of arthropod-derived chitin compared to the terrestrial mammal gut microbiome 28
2.4 Discussion 30
2.5 Methods 36
2.5.1 Sea otter population 36
2.5.2 Sample collection for the 16S rRNA gene amplicon survey from sea otters 37
2.5.3 Sample collection for the 16S rRNA gene amplicon survey from other otters 37
2.5.4 DNA extraction, 16S rRNA gene amplification, and amplicon sequencing 38
2.5.5 Amplicon sequence variant inference, taxonomic assignment, and filtering 39
2.5.6 Determination of alpha and beta diversity 40
2.5.7 Clustering and comparison of sea otter gingival communities 41
2.5.8 Differential abundance testing 41
2.5.9 Sample collection and selection for shotgun sequencing . . 41
2.5.10 Shotgun sequencing and quality filtering 42
2.5.11 Metagenome assembly and annotation 43
2.5.12 Identification of genes of interest 44
2.5.13 Estimation of the number of bacterial reads per sample . . 45
2.5.14 Estimation of prey DNA present in shotgun sequencing reads 46
2.5.15 Chitinase diversity present in sea otters, other mammals, and seawater 46

3 Novel microbial diversity and functional potential in the marine mammal oral microbiome 48
3.1 Abstract 48
3.2 Introduction 49
3.3 Results 51
3.3.1 Dolphin oral microbiota composition and structure 51
3.3.2 Novel, deeply divergent phylum-level lineages 55
3.3.3 Functional profile of the Delphibacteria lineage 59
3.3.4 Large biosynthetic gene cluster in the dominant Actinobacteria genome 61
3.3.5 Novel Cas9 diversity 63
3.3.6 Saccharibacteria type II CRISPR-Cas systems and a Saccharibacteria-infecting phage 66
3.4 Discussion 68
3.5 Methods 70
3.5.1 Experimental model and subject details 70

3.5.2 DNA extraction, sequencing, and quality filtering 71
3.5.3 Metagenome assembly, annotation, and binning 72
3.5.4 Identification of phage scaffolds 73
3.5.5 Refining selected scaffolds 74
3.5.6 Bin completeness and characterization 74
3.5.7 Phylogenetic placement of genomes 76
3.5.8 Metabolic reconstruction of DolZOral124_Bacteria-_64_63 (*Candidatus* Delphibacteria) 78
3.5.9 Biosynthetic gene cluster structural predictions 79
3.5.10 Identification and classification of CRISPR-Cas systems and predicted Cas9 proteins 79
3.5.11 Identification and analysis of scaffolds targeted by CPR spacers 81
3.5.12 Data availability 82

4 Previously uncharacterized rectangular microbial units in the marine mammal oral cavity 83
4.1 Abstract 83
4.2 Introduction 84
4.3 Results 86
4.3.1 Light microscopy insights into the structure and spatial organization of rectangular cell-like units 86
4.3.2 Cryogenic electron transmission microscopy insights into the internal ultrastructure of rectangular cell-like units 88
4.3.3 Cryogenic electron transmission microscopy insights into surface structures of rectangular cell-like units 89
4.3.4 Potential contacts between rectangular cell-like units and other cells 91
4.3.5 Taxonomic identification of rectangular cell-like units . . . 91
4.4 Discussion 96
4.5 Methods 105
4.5.1 Sampling collection 105
4.5.2 Light microscopy data acquisition 106
4.5.3 Cryo-TEM data acquisition 106
4.5.4 Single-cell genomics experiment 107

5 Conclusion 111
5.1 Summary remarks 111
5.2 Open problems for future work 112
5.2.1 Co-evolution of marine mammals and their microbiota . . 112
5.2.2 Novel microbial diversity and functional potential associated with marine mammals 115

5.3 Coda 118

Appendix A Additional material for 'Characterization of the gingival and gut microbiota of the Southern sea otter *(Enhydra lutris nereis)*' 119
A.1 Additional figures and tables 119

Appendix B Additional material for 'Novel microbial diversity and functional potential in the marine mammal oral microbiome' 138
B.1 Additional figures and table 138
B.2 Additional Discussion 149
B.2.1 Naming novel lineages 149
B.2.2 Linking 16S rRNA genes to genomes from novel lineages . 149
B.2.3 High proportion of type II CRISPR-Cas systems in the dolphin oral microbiome 150
B.2.4 Additional information on the insertion in Cas9 from DolZOral124_scaffold_26_62 150
B.2.5 Identity of Saccharibacteria genomes with type II CRISPR-Cas systems 151
B.2.6 Analysis of spacer sequences from Saccharibacteria CRISPR arrays 151

Appendix C Additional material for 'Previously uncharacterized rectangular microbial units in the marine mammal oral cavity' 153
C.1 Additional figures 153

List of supplemental data files 156

Bibliography 158

List of Figures

2.1 Sampling locations of sea otters included in this study 13
2.2 Relative abundances of major taxonomic groups present in the sea otter gingival and rectal microbiota 16
2.3 Comparison of distal gut communities obtained from different otter species 19
2.4 Relationship between sea otter gingival, rectal, and adjacent seawater bacterial communities 20
2.5 Community profiles characteristic of gingival samples from wild sea otters 23
2.6 Relative abundance of *Helicobacter* ASVs in gingival, rectal, and seawater communities 25
2.7 Taxonomic composition of reads resulting from shotgun sequencing of sea otter feces 27
2.8 Distribution of marine-like vs gut-like chitin-degrading proteins in different environments 29

3.1 Phylogenetic relationships among genomes recovered from the dolphin mouth 53
3.2 Community structure of the dolphin oral microbiota 56
3.3 Functional profile of Delphibacteria 60
3.4 Novel non-ribosomal peptide synthesis BGC encoded by the dominant Actinobacteria genome 62
3.5 Unusual predicted Cas9 protein sequences in the dolphin oral samples 65

3.6 Genome organization of the Saccharibacteria phage 67

4.1 Light microscopy images of rectangular cell-like units 87
4.2 Cryo-TEM image with paired parallel segments in rectangular cell-like units 89
4.3 Cryo-TEM images reveal morphological features of rectangular cell-like units 90
4.4 Cryo-TEM image documenting rectangular cell-like unit in proximity to another cell 92
4.5 Cryo-TEM image documenting rectangular cell-like unit in proximity to other cells and probable extracellular vesicles 93

A.1 Comparison of distal gut communities obtained from different otter species using multiple distance metrics 120
A.2 Comparison of genera present in sea otter and North American river otter distal gut communities 121
A.3 Alpha diversity of gingival and rectal bacterial communities from wild sea otters, and adjacent seawater 122
A.4 Relative abundance and overlap of ASVs between gingival, rectal, and seawater communities 124
A.5 Composition of gingival community profiles 126
A.6 ASVs are differentially abundant between sea otter gingival CPs . 127
A.7 Variation between gingival community profiles is correlated with time of sampling 128
A.8 Variation between gingival community profiles by month of sampling 129
A.9 Variation between sampling location by month of sampling 130
A.10 Variation between gingival community profiles by sampling location 131
A.11 Approximate number of bacterial genomes assembled per sea otter fecal metagenome 132
A.12 Gap statistic for the sea otter gingival dataset 133

B.1 Comparison of the relative abundances of the most common 16S rRNA gene sequences identified with PCR and pyrosequencing (Bik *et al.*, 2016) with those assembled from Illumina paired-end reads using IDBA-UD 140
B.2 Maximum likelihood 16S rRNA gene phylogenies 142
B.3 Distribution of the taxonomic identity of top protein matches to the novel Fibrobacteres-Chlorobi- Bacteroidetes superphylum lineage and the CPR lineages 143
B.4 Characterization of the BGC-encoding scaffold 144
B.5 Distribution of CRISPR-Cas types across dolphin, harbour seal, and human oral environments 145
B.6 Phylogeny of Cas9 proteins 146

C.1 Cryo-TEM image documenting rectangular cell-like unit in association with other cells 154
C.2 Tetranucleotide ESOM of genomic material recovered the single-cell genomics experiment 155

List of Tables

2.1 Habitat and ecology of four otter species included in this study . . 18

4.1 Properties of genome bins recovered from the single-cell genomics experiment 95

A.1 Estimate of the number of bacterial reads using assembly-driven metagenomics vs read-based metagenomics 134

A.2 Number of marine-like vs gut-like glycoside hydrolase per mammalian species or seawater 135

A.3 gDNA sizes for final libraries 136

A.4 Amount of sequencing data generated per sample 137

B.1 Comparison of phyla detected in a pyrosequencing amplicon survey versus those identified after assembly of Illumina paired-end reads with IDBA-UD 148

Abstract

Diversity, form, and function within the marine mammal microbiota

by

Natasha K. Dudek

Animals can be viewed as complex, co-evolving networks of microbes and host cells. Understanding the diversity, form, and function of microbes associated with different animals is therefore essential to understanding the patterns and processes underlying evolution across all domains of life. Extant marine mammals present an interesting opportunity to study the microbiota of animals with an unusual lifestyle that has arisen independently six times since the time of their last common ancestor. The manner in which the marine mammal-associated microbiota has evolved in response to the host's marine lifestyle remains unclear. In this thesis, I describe three studies of the microbiota of marine mammals. In the first, I characterize bacterial community composition associated with sea otters, which are a keystone species that is listed as endangered by the IUCN. They are also the sole representatives of an entire lineage of marine mammal. The findings suggest that environment plays a major role in structuring sea otter-associated bacterial community composition and raises the question of whether sea otters may have a reduced bacterial biomass in their guts compared to other mammals. As seen in other marine mammal species, results show that sea otters host a diversity of 'microbial dark matter'. In chapter two of this thesis, I study such 'microbial dark matter' present in the dolphin mouth and propose two new bacterial phyla (*Candidatus* Delphibacteria and *Candidatus* Fertabacteria), the former of which our metabolic reconstruction suggests may have a direct effect on dolphin physiology and health. In the third chapter of my thesis, I operate under the assumption

that novel phylogenetic diversity is correlated with novel functional diversity, and thereby discover a previously uncharacterized rectangular microbe in dolphin oral samples with several unusual morphological features, such as pili-like appendages whose architecture differs substantially from known surface structures seen in bacteria and archaea. A single-cell genomics experiment suggested that this microbe was a type of bacteria from one of the following three groups: Bacteroidetes, TM7, or Epsilonproteobacteria. Collectively, these studies provide insight into diversity, form, and function within the marine mammal microbiota, and contribute towards our understanding of the microbial diversity, both phylogenetic and functional, which has evolved on Earth.

Acknowledgments

I am very grateful to my adoptive advisors, David Relman and Beth Shapiro, for welcoming me into their labs and creating an environment in which I was able to pursue my graduate research. I am lucky to have had them both as advisors, and greatly appreciate all the time, ideas, and funding that they contributed towards making my PhD pursuit a valuable and formative experience. To Beth: thank you for providing me with a home at UCSC, for giving me the opportunity to grow as a scientist, and for exposing me to some very cool fields of biology. To David: thank you for sharing your interest and enthusiasm for a diverse range of microbiology-related fields, for having been so encouraging and supportive, and for the many thought-provoking, helpful, and oftentimes fun conversations over the years.

In addition to my advisors, I would like to thank Dan Costa and Ed Green for serving on my dissertation reading committee and for their guidance and suggestions over the course of my graduate work. I would also like to thank the members of the Relman lab and Shapiro-Green lab for insightful discussion, feedback, and guidance, especially Christine Sun, Liz Costello, Alix Switzer, Rob Pesich, Les Dethlefsen, Elies Bik, Daniela Aliaga Goltsman, Ania Robaczewska, and Arati Patankar. An especially big thank you goes to Christine, who first got me started doing metagenomic analyses and was an amazing mentor over the course of my PhD work. I am further grateful to have received financial support for my PhD in part from a B1 fellowship from the Fonds de Recherche du Québec - Nature et Technologie.

The research presented in this thesis includes a wide range of techniques and approaches. To this end, I worked with many collaborators. In regards to Chapter two, a big thank you goes to Alix Switzer, who conceived of and initiated the

sea otter 16S rRNA gene survey branch of the study, and played a major role in it's development, including performing many DNA extractions. This project would not have been possible without the hard work of those who coordinated and executed the collection of sea otter microbial samples, from the United States Geological Survey, Monterey Bay Aquarium, and California Department of Fish and Wildlife. In particular, for the work presented in this thesis I would like to thank Tim Tinker, Joe Tomoleoni, Michelle Staedler, and Francesca Batac. I'd also like to thank Liz Costello for her assistance in learning how to analyze 16S rRNA gene amplicon sequencing data, as well as Ania Robaczewska and Arati Patankar for guidance in preparing 16S rRNA gene amplicons for sequencing.

The text in Chapter three is a reprint of previously published material from the journal *Current Biology* under the title 'Novel microbial diversity and functional potential in the marine mammal microbiome'. I would like to thank my co-authors Christine Sun, David Burstein, Rose Kantor, Daniela Aliaga Goltsman, Elisabeth Bik, Brian Thomas, and Jillian Banfield for their guidance and assistance. Author contributions are as follows: N.D, C.S., E.B., and D.R. designed the study. N.D., C.S., and B.T. processed sequence and assembly data. N.D., C.S., D.B., R.K., D.A.G., J.B., and D.R. conducted data analysis. N.D., C.S., and D.R. wrote the manuscript with input from all authors. I would additionally like to thank Mohamed Donia for guidance in the analysis of a biosynthetic gene cluster, Alix Switzer for the contribution of DNA extracts from a harbor seal, and Frances Gulland and coworkers at the Marine Mammal Center (Sausalito, CA, USA) for collection of the harbor seal oral sample.

For the investigation of the rectangular cell-like units presented in Chapter four, my sincere thanks goes to KC Huang, who helped guide this research, as well as to members of the Huang Lab who helped me learn microscopy techniques,

especially Handuo Shi. I'd also like to thank Wah Chiu and members of the Chiu lab who were instrumental in collecting and processing electron microscopy images, especially Megan Mayer, Gong-Her Wu, and Jesus Galaz-Montoya. I thank Barry Behr for assistance in capturing rectangular cells using a cell micromanipulator, as well as Kat Ng and Brian Yu for helping attempt to capture cells via laser capture microdissection and microfluidics, respectively.

I would also like to thank Celeste Parry and colleagues at the U.S. Navy Marine Mammal Program (MMP) Biosciences Division, Space and Naval Warfare Systems Center Pacific, San Diego, California. Their generous collection of dolphin-associated samples was invaluable to the work presented in this thesis, both Chapters three and four.

A huge thank you goes to my Mom, Dad, and brother, Nick. Thank you guys for always having encouraged me, ranging from providing moral support to helping me with computer programming to proofreading many drafts of my thesis and first paper. Thank you especially for the many phone calls, emails, texts, and visits.

Finally, my deep thanks to Dmitriy for having been so supportive of me and my endeavour to get a PhD. In the last few weeks before my thesis was due, thank you for reassuring me while I panickedly threw things together and for helping me with last minute R and python problems. Mostly though, thank you for having such a wonderful sense of humour and for always making me smile.

Chapter 1

Introduction

1.1 Co-evolution of animals and their microbiota

Bacteria predate the existence of metazoans on Earth by approximately 3 billion years (Knoll, 2003). From introducing oxygen into early Earth's atmosphere to serving as the basis for mitochondria, these single-cellular organisms have fundamentally shaped our origin and evolution (Andersson, 2003; McFall-Nagai *et al.*, 2003). In modern animals, there is ample evidence that bacterial symbionts supplement their host's genetic makeup with novel, relatively plastic functional diversity (Zilber-Rosenberg & Rosenberg, 2008), and in particular, have myriad effects on host development, immunity, physiology, and even behaviour (Mazmanian *et al.*, 2005; Turnbaugh *et al.*, 2006; Zheng *et al.*, 2008; Gaboriau-Routhiau *et al.*, 2009; Ichinohe *et al.*, 2010; Hooper *et al.*, 2012; Koropatkin *et al.*, 2012; Buffie & Pamer, 2013, and others). Thus animals can be regarded as units consisting of complex networks of interacting species upon which selection acts, rather than as autonomous entities (Zilber-Rosenberg & Rosenberg, 2008). A general synthesis on the nature of the interactions between hosts and their microbiota, as well as the patterns and processes that underlie their co-evolution, has applications in

fields such as medicine, conservation biology, and evolutionary theory.

The most well-developed characterization of the range of interactions between a host species and their microbiota comes from studies of humans. Recognition of the importance of the microbiome in human health led to the Human Microbiome Project (Relman & Falkow, 2002; Turnbaugh *et al.*, 2007), thereby galvanizing scientific inquiry into the diversity of the human microbiota, the factors driving the establishment and distribution of human-associated taxa, and the functional consequences of the microbial consortia that inhabit the human body (Mazmanian *et al.*, 2005; Turnbaugh *et al.*, 2006; Zheng *et al.*, 2008; Gaboriau-Routhiau *et al.*, 2009; Ichinohe *et al.*, 2010; Hooper *et al.*, 2012; Koropatkin *et al.*, 2012; Buffie & Pamer, 2013, and others). Similar research on other host species has also flourished (Cardoso *et al.*, 2012; Hooda *et al.*, 2012; Lavery *et al.*, 2012; Abdelrhman *et al.*, 2016; Bik *et al.*, 2016, Hammer *et al.*, 2017, and others), although oftentimes there is significant bias in terms of the host species that receive the greatest focus, such as those that serve as model organisms for human health (e.g., mice) or are economically valuable (e.g., cows). Incorporating knowledge from a phylogenetically diverse selection of host species with a wide variety of lifestyles has the potential to greatly expand our understanding of the full diversity, and therefore general nature, of host-microbiome interactions.

1.2 The marine mammal microbiome

Marine mammals consist of five distinct lineages that each independently re-invaded the marine environment. While clear adaptations such as streamlined bodies and increased thermal retention have repeatedly evolved in host species via modification of the slowly-evolving mammalian genome (reviewed in Berta *et al.*, 2005), many unanswered questions remain regarding how a marine lifestyle

has affected the relatively rapidly evolving microbiota of these animals and vice-versa. For example, what types of bacteria are present, how has this selection been shaped by life in the sea, and does the microbiota differ in composition and/or function from that of terrestrial mammals?

An important first step towards answering these questions is characterizing the types of microorganisms that are associated with marine mammals. Substantial progress has been made towards culture-independent surveys of the microbiota of four out of the six extant lineages of marine mammals, namely cetaceans (Lima *et al.*, 2012; Sanders *et al.*, 2015; Bik *et al.*, 2016; Soverini *et al.*, 2016, Godoy-Vitorino, *et al.*, 2017; Erwin *et al.*, 2017; Russo *et al.*, 2018; and others), polar bears (Glad *et al.*, 2010), sirenians (Eigeland *et al.*, 2012; Merson *et al.*, 2014), and pinnipeds (Nelson *et al.*, 2013; Bik *et al.*, 2016; Delport *et al.*, 2016; Lavery *et al.*, 2012, and others). (Notably lacking are the sea otter and marine otter.) One interesting finding that has resulted from such studies is that marine mammals, and especially dolphins, serve as hosts for a rich diversity of bacteria from many poorly characterized branches of the tree of life, including 'candidate phyla' (Bik *et al.*, 2016), which are entire phylum-level lineages for which no cultured representatives exist (i.e. 'microbial dark matter'). As such, their ecology and evolution, let alone the significance of their presence in the marine mammal microbiota, is essentially unknown.

It is also worth noting that marine mammals are an ecologically important group of animals, 25-37% of which are in danger of extinction (Schipper *et al.*, 2008; Davidson *et al.*, 2012), as is their microbiome. In addition to being of interest from theoretical perspective, a greater understanding of their microbiota may have practical applications in veterinary and conservation science.

1.3 Outline of research presented in this thesis

This thesis focuses on exploring the taxonomic and functional diversity of the marine mammal microbiota. The following overarching questions have guided the research herein:

1. What is the community composition of the marine mammal microbiota?

2. In the case of novel, previously undescribed lineages, what is their taxonomic affiliation and the nature of their lifestyle?

3. Novel phylogenetic diversity is correlated with novel functional potential (Wu *et al.*, 2009). Given that marine mammals, and especially dolphins, are host to a rich diversity of poorly characterized bacterial lineages (Bik *et al.*, 2016), do we observe novel functional or morphological characteristics in these communities?

To study such questions, a combination of the following complementary methodologies were employed: a 16S rRNA gene survey, assembly-driven metagenomics (including at the genome-resolved level), light microscopy, cryogenic electron transmission microscopy, and single-cell genomics.

In Chapter two of this thesis, I studied the community composition of the sea otter microbiota. Sea otters are an IUCN endangered, keystone species (Estes, 1990; Estes & Palmisano, 1974; Doroff & Burdin, 2015) and are the sole representative of one of the six extant lineages of mammals to have independently re-invaded the marine environment (reviewed in Berta *et al.*, 2005). I performed a 16S rRNA gene survey of the gingival (oral) and rectal microbiota of 151 wild sea otters, followed by shotgun metagenomic sequencing to characterize in more detail a set of twelve fecal samples. Bacterial communities in the sea otter gut

were distinct from those of three other semi-aquatic otters, suggesting that provenance (life in the sea) may shape the bacterial community of the sea otter gut. Sea otter gingival community composition tended towards one of three different profiles, which appeared to be strongly influenced by their environment. Shotgun sequencing suggested that little bacterial or host DNA was present in fecal samples from sea otters; rather, most DNA was likely from prey species. This may potentially reflect a reduced bacterial biomass in the sea otter gut, rapid transit time, and a relatively large dietary mass:body weight ratio when compared to other mammalian species. As seen in other species of marine mammals (Bik *et al.*, 2016), a diversity of bacteria from candidate phyla were present in sea otter-associated communities. This study establishes a baseline for understanding the community composition, structure, and function of the microbiota of healthy sea otters, which may assist in future management of sea otter populations and of sick and/or vulnerable animals.

The significance of members of candidate phyla in mammalian environments, including marine mammal species such as sea otters and dolphins (Bik *et al.*, 2016) is poorly understood. In the next Chapter of this thesis, I investigated the ecology and evolution of such lineages present in the dolphin mouth. This was done by recovering and analyzing genomes from uncultured dolphin oral bacteria directly from environmental samples using genome-resolved metagenomics. Three genomes from two lineages were discovered to be representative of deeply branching, previously undescribed bacterial phyla, for which the names '*Candidatus* Delphibacteria' and '*Candidatus* Fertabacteria' were proposed. The former was found in both wild and managed dolphins. Metabolic reconstruction suggested that this taxon likely has the capacity for denitrification and therefore may have an effect on dolphin physiology and health. Novel taxonomic diversity was

accompanied by novel functional diversity in the community as a whole, including unusual CRISPR-Cas9 systems that were hypothesized to have previously uncharacterized functional properties. While the biotechnological value of such systems remains to be seen, this work further establishes the potential of assembly-driven metagenomics in uncovering novel biochemical systems in previously unexplored environments.

While metagenomic analysis offers a powerful tool to learn about uncultured bacteria, it is limited in its ability to reveal functional/structural characteristics that have not previously been described and/or whose genetic basis is unknown. Operating under the assumption that novel phylogenetic diversity is correlated with novel functional diversity (Wu *et al.*, 2009), in Chapter four I used a microscopy-based approach to survey dolphin oral communities for interesting cell morphotypes or features. Since form typically follows function, we reasoned that this could provide insight into novel properties or lifestyles of 'microbial dark matter'. With this approach, unusual rectangular microbial morphotypes were discovered in the dolphin mouth, which were dubbed 'rectangular cell-like units'. Each rectangular cell-like unit was composed of numerous parallel, seemingly paired, membrane-bound segments. These segments are likely individual cells, whereas each rectangular cell-like unit is likely an aggregate, similar to what is seen with bacteria from the genus *Simonsiella*. Cryogenic transmission electron microscopy revealed that pili-like appendages with a complex architecture projected from segments. These appendages consisted of stalks of hair-like structures that splayed out at the tips, in contrast to known pili which consist of single hair-like appendages with little to no further morphological features (Fernandez & Berenguer, 2000; Hospenthal *et al.*, 2017). Each rectangular cell-like unit was encapsulated in an S-layer-like structure, suggesting that if segments are

individual cells, there is likely cooperation between cells in secreting the proteins or glycoproteins involved in assembly of the S-layer-like structure. A single-cell genomic experiment suggested that the cells are bacterial and affiliated with the Bacteroidetes, Saccharibacteria (TM7), or Epsilonproteobacteria taxa.

Finally, in the thesis conclusion I discuss the significance of these results with respect to the field of marine mammal microbiome research and with respect to the exploration of the diversity of microbial life which has evolved on Earth.

Chapter 2

Characterization of the gingival and gut microbiota of the Southern sea otter *(Enhydra lutris nereis)*

2.1 Abstract

The microbiome plays an important role in mammalian health. Sea otters are an IUCN endangered, keystone species in coastal ecosystems for which relatively little is known about their indigenous microbiota. To characterize the bacterial community composition of Southern sea otters, we used 16S rRNA gene amplicon sequencing to study 144 gingival (oral), 82 rectal, and 75 adjacent seawater samples from 151 wild individuals living off the coast of California, USA, and shotgun metagenomic sequencing to characterize twelve fecal samples in more detail. Bacterial communities in the sea otter gut were distinct from those of semi-aquatic

otters, such as the North American river otter. Sea otter gingival community composition tended towards one of three different types of profiles, which most likely occur along a gradient rather than being distinct entities. To learn about the functional potential of the sea otter gut microbiome, we performed shotgun sequencing on 12 fecal samples, with particular interest in whether chitinolytic gene diversity might show evidence of adaptation for the purpose of degrading the chitin-rich prey that are known to be preferred as food by this species. Our results suggest that there may be a shift towards increased prey-derived chitin-degradation potential in the gut microbiome of sea otters and other marine mammals compared to terrestrial mammals. However, metagenomic analysis was hindered by the finding that the majority of DNA in samples appeared to have been derived from prey (up to 63% in one case) rather than from indigenous bacteria or host cells. We hypothesize that this may reflect a reduced bacterial biomass in the sea otter gut compared to other mammalian species. This study establishes a baseline for understanding the community composition and structure of the microbiota of healthy sea otters, which may assist in future management of sea otter populations and of sick and/or vulnerable animals.

2.2 Introduction

Sea otters are an IUCN endangered, keystone species that was nearly hunted to extinction in the 1800's (Estes, 1990; Estes & Palmisano, 1974; Doroff & Burdin, 2015). Removal of sea otters from their native coastal ecosystems leads to increased herbivory by invertebrates and destruction of plant communities (ex: kelp forests) that provide a habitat for other species (Estes & Palmisano, 1974; Estes *et al.*, 1998). Following the introduction of US federal protections in 1911, the Southern sea otter *(Enhydra lutris nereis)*, which is listed as threatened by

the US Endangered Species Act, has shown promise of recovery. They have a long way to go, however, before their stock is considered to be a 'significant functioning element in the ecosystem' under the Marine Mammal Act, which would require an excess of 8,400 individuals in contrast to the 3,272 individuals estimated to exist as of 2016 (US Fish and Wildlife Service, 2015; Tinker and Hatfield, 2016).

The primary obstacle hindering Southern sea otter recovery is thought to be high mortality rates (Kreuder *et al.*, 2003). One important cause of death in the population as a whole is infectious disease, especially encephalitis caused by *Toxoplasma gondii* (Kreuder *et al.*, 2003). By demographic group, the highest mortality rates occur in pups and juveniles, with prime-age females also disproportionately affected (Estes *et al.*, 2003; Tinker *et al.*, 2006). In part, this is driven by the difficulty of obtaining enough calories to fuel their astronomically high metabolic rates, which requires that sea otters consume the equivalent of 20-25% of their body mass per day (Costa and Kooyman, 1982; Costa and Kooyman, 1984). This can be particularly difficult for adult females with pups, in whom caloric insufficiency via end-lactation syndrome is a primary or major contributing cause of death in 56% of adult female mortalities (Chinn *et al.*, 2015), and who can experience increases in daily energetic costs of up to 96% as a result of having a dependent pup (Thometz *et al.*, 2014). Adult female mortality has a strong influence on population trajectory (Gerber *et al.*, 2004; Tinker *et al.*, 2006).

Extensive evidence from studies of other mammalian species supports a key role of the microbiome in mammalian health, including pathogens, immune system maturation, and energy acquisition from food in the gut (Mazmanian *et al.*, 2005; Turnbaugh *et al.*, 2006; Zheng *et al.*, 2008; Gaboriau-Routhiau *et al.*, 2009; Ichinohe *et al.*, 2010; Hooper *et al.*, 2012; Koropatkin *et al.*, 2012; Buffie & Pamer,

2013 and others). Thus characterizing the microbiome of healthy sea otters is of interest for two primary reasons. First, a baseline understanding of the sea otter microbiota and the factors that determine its composition may be of use to sea otter veterinary and conservation communities. To understand the role of the microbiota in health and disease, it is first necessary to have a baseline measurement of the state of the healthy microbiota (Turnbaugh *et al.*, 2007). Second, while extensive research has been conducted on the community composition and functional potential of the human microbiome, there remains much to be learned about evolutionary patterns underlying microbiota assembly and evolution in other species, of which marine mammals are particularly interesting.

Marine mammals represent a textbook case of convergent evolution of mammals to a new environment (reviewed in Berta *et al.*, 2005) and therefore offer an ideal opportunity to study the evolution of the host-associated microbiome. A major shift that occurred repeatedly in mammalian lineages (i.e. baleen whales, some seals, and sea otters) as they adapted to the marine environment was a move towards a diet rich in chitinous invertebrates (reviewed in Berta *et al.*, 2005). Correspondingly, a long-standing question is whether the gut microbiota of these mammals is enriched in bacteria that can degrade chitin, and if so, whether the host may benefit from the energy extracted by their gut bacteria from an otherwise indigestible polysaccharide (Martensson *et al.*, 1994; Olsen *et al.*, 1999; Simunek *et al.*, 2000; Sanders *et al.*, 2015). Sea otters are the sole representative of an entire mammalian lineage that independently re-invaded the marine environment and therefore provide an additional replicate in what can be viewed as a 'natural experiment' on how the functional potential of the gut microbiome adapts to a marine lifestyle.

In this study, we first sought to characterize the bacterial community composi-

tion of the Southern sea otter oral and gut microbiota. To this end, we performed a 16S rRNA gene amplicon survey on 144 gingival and 82 rectal samples from 151 wild sea otters, as well as 75 samples of seawater adjacent to these sea otters at the time of capture. The wild sea otter samples were collected by the United States Geological Survey (USGS), the Monterey Bay Aquarium (MBA), and the California Department of Fish and Wildlife (CDFW). To learn about the functional potential of the sea otter gut microbiome, and whether gut bacteria may be involved in prey-derived chitin-degradation, we employed an assembly-driven metagenomic approach with shotgun sequencing data from 12 sea otter fecal samples. Our findings establish a baseline understanding of the healthy sea otter microbiome and offer the first community-wide insights into the microbiota of one of the five lineages of mammals to independently re-invade the marine environment.

2.3 Results

2.3.1 Sea otter sample collection

We analyzed 144 gingival, 82 rectal, and 75 adjacent seawater samples from 151 wild Southern sea otters living off the coast of California, USA (Figure 2.1). Samples were collected over a six year period from 2011-2017. The 16S rRNA gene V3-V4 region was PCR amplified and amplicons were sequenced across a single Illumina HiSeq 2500 lane to produce 2 x 250 bp reads. After quality control and filtering this yielded a dataset consisting of 17,795 ASVs represented by 34,235,857 merged reads.

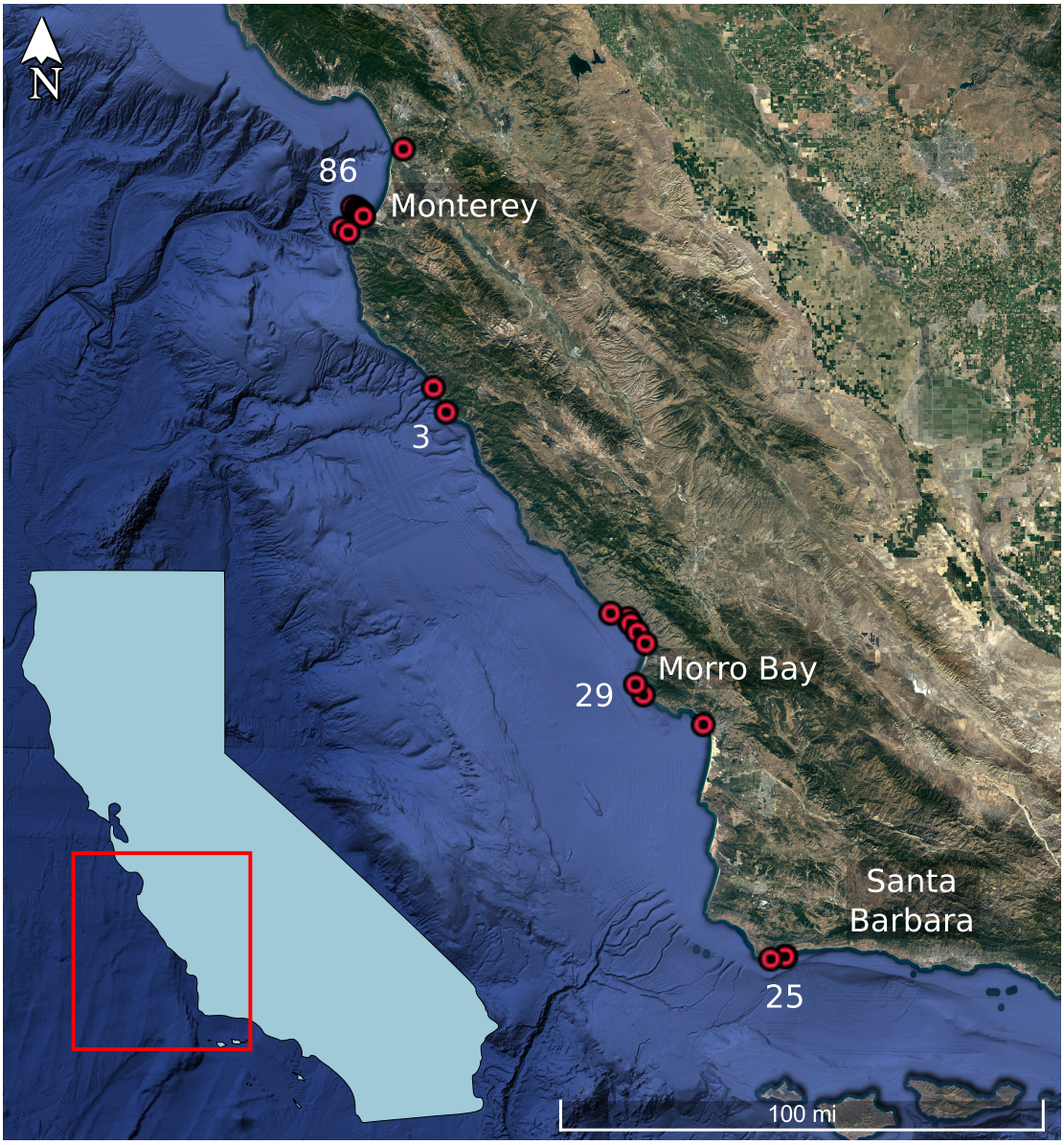


Figure 2.1: Sampling locations of sea otters included in this study. A satellite map of the coastline of central California, USA is shown. The state outline is shown in blue in the lower left corner, with a red box indicating the area shown in the main image. Sea otter sampling locations are indicated by red markers. The number of sea otters sampled per set of locations is denoted. Location data for eight sea otters was not collected - these sea otters are omitted from the map. Scale bar: 100 miles.

2.3.2 Taxonomic composition and alpha diversity of gingival and rectal sea otter samples

In total, 36 high-level bacterial lineages were detected (approximately phyla; the polyphyletic Proteobacteria 'phylum' was split into classes and phyla in the Parcubacteria (OD1) and Microgenomates (OP11) groups were collapsed at the superphylum level). The taxonomic groups with the highest prevalence in gingival samples were Gammaproteobacteria, Firmicutes, Betaproteobacteria, Bacteroidetes, and Actinobacteria, while in rectal samples they were Gammaproteobacteria, Firmicutes, Fusobacteria, Bacteroidetes, and Actinobacteria (Figure 2.2).

Taxa from several poorly understood phylum-level lineages with no cultured representatives were detected (i.e. candidate phyla). Of particular interest are those from the Candidate Phyla Radiation (CPR), which is a monophyletic radiation of candidate phyla with highly reduced genomes, missing core biosynthetic pathways, hypothesized symbiotic lifestyles, and other unusual properties (Wrighton *et al.*, 2012; Albertsen *et al.*, 2013; Kantor *et al.*, 2013; Brown *et al.*, 2015; Hug *et al.*, 2016). From the CPR, the following phyla and superphyla were identified: Absconditabacteria (SR1), Microgenomates (OP11), Parcubacteria (OD1), and Saccharibacteria (TM7). We also detected members of other candidate phyla or phyla for which the first isolates were only recently cultured (Tamaki *et al.*, 2011), and whose biology remains relatively poorly understood: Aminicenantes (OP8), Armatimonadetes (OP10), BRC1, Marinimicrobia (SAR406 / Marine Group A / MGA), and WPS-1.

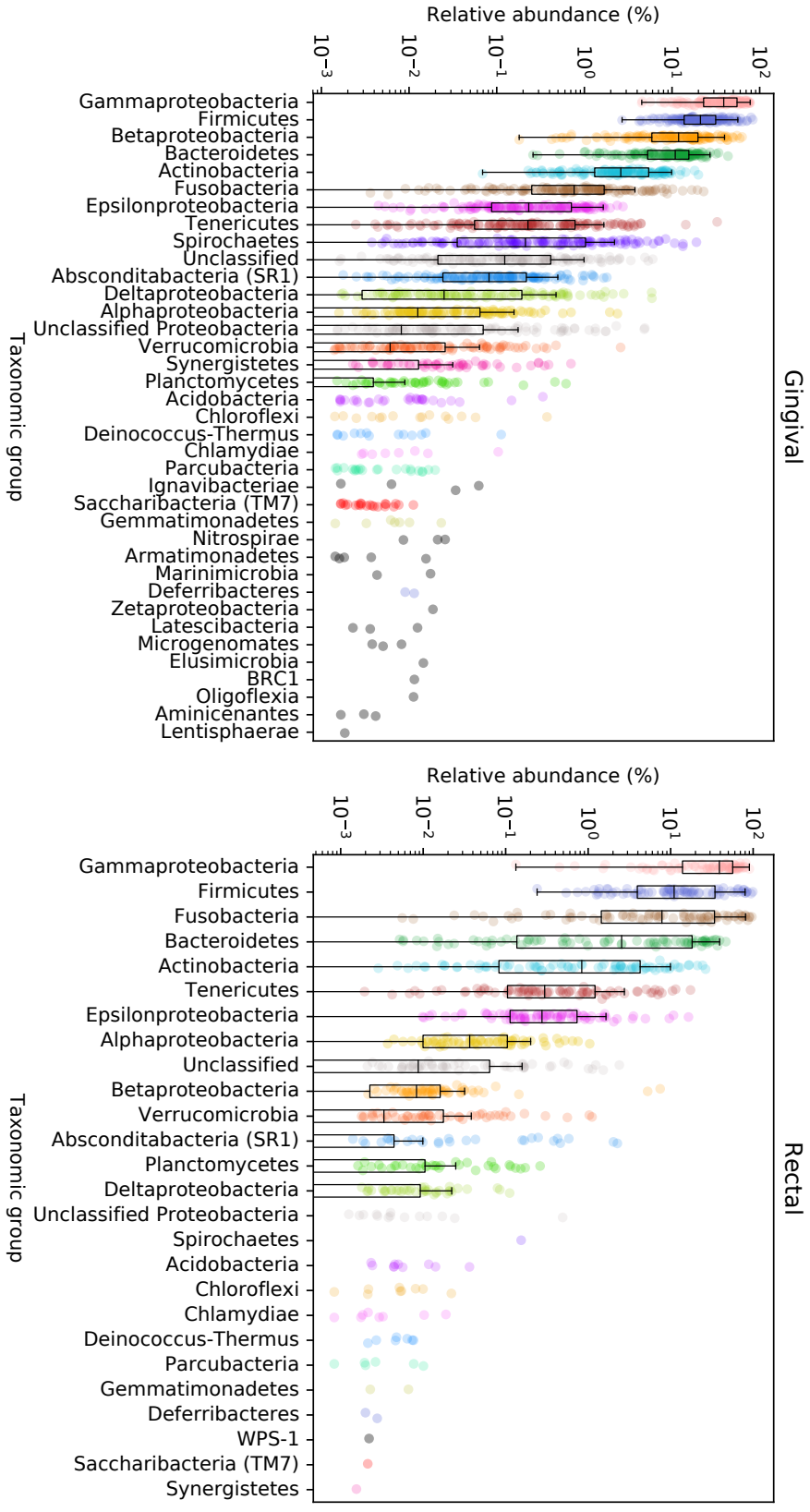

Relative abundance (%)
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Taxonomic group
Gingival
Gammaproteobacteria
Firmicutes
Betaproteobacteria
Bacteroidetes
Actinobacteria
Fusobacteria
Epsilonproteobacteria
Tenericutes
Spirochaetes
Unclassified
Absconditabacteria (SR1)
Deltaproteobacteria
Alphaproteobacteria
Unclassified Proteobacteria
Verrucomicrobia
Synergistetes
Planctomycetes
Acidobacteria
Chloroflexi
Deinococcus-Thermus
Chlamydiae
Parcubacteria
Ignavibacteriae
Saccharibacteria (TM7)
Gemmatimonadetes
Nitrospirae
Armatimonadetes
Marinimicrobia
Deferribacteres
Zetaproteobacteria
Latescibacteria
Microgenomates
Elusimicrobia
BRC1
Oligoflexia
Aminicenantes
Lentisphaerae
Relative abundance (%)
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Taxonomic group
Rectal
Gammaproteobacteria
Firmicutes
Fusobacteria
Bacteroidetes
Actinobacteria
Tenericutes
Epsilonproteobacteria
Alphaproteobacteria
Unclassified
Betaproteobacteria
Verrucomicrobia
Absconditabacteria (SR1)
Planctomycetes
Deltaproteobacteria
Unclassified Proteobacteria
Spirochaetes
Acidobacteria
Chloroflexi
Chlamydiae
Deinococcus-Thermus
Parcubacteria
Gemmatimonadetes
Deferribacteres
WPS-1
Saccharibacteria (TM7)
Synergistetes

Figure 2.2: Relative abundances of major taxonomic groups present in the sea otter gingival and rectal microbiota. High level taxonomic groups (approximately phyla) present in (A) 144 gingival and (B) 82 rectal samples are shown in order of decreasing median rank abundance per sample type. Each taxonomic group is colour-coded between panels. Light grey indicates artificial groupings ('Unclassified' and 'Unclassified Proteobacteria') while dark grey indicates a taxonomic group detected in only gingival or only rectal samples, but not both.

2.3.3 Sea otter distal gut microbiota is distinct from that of other otters

Major determinants of gut bacterial community composition include the host's diet (whether they are carnivorous, ominivorous, or herbivorous), phylogeny, and morphology (whether they have a simple gut or are fore- vs hind- gut fermenters) (Ley *et al.*, 2008). Provenance, such as whether a host species is marine or terrestrial, is also likely an important factor, but it's role has been less well characterized to date and is particularly unclear for carnivorous marine mammals (Bik *et al.*, 2016). To gain insight into whether and how the sea otter gut microbiota differs from those of other closely related, non-marine species, we compared the rectal microbiotas of sea otters to those of other otters: 13 North American river otters (*Lontra canadensis*), an Asian small clawed otter (*Aonyx cinerea*), and a giant otter (*Pteronura brasiliensis*). Notably, these four species have similar gut morphologies and diets (for information on host species biology, see Table 2.1).

The gut microbiota of sea otters was distinct from that of North American river otters (Unifrac distance metric, bootstrap test: $p < 0.001$) and also other otter species (Figure 2.3). The latter three species appeared to have communities that were more similar to one-another than to sea otters. This appears to be driven by differences in the relatedness of low abundance bacterial genera between host species, as the pattern was pronounced when using unweighted

Unifrac but less so when using the Bray-Curtis, Jaccard, and weighted Unifrac distance metrics (Appendix A Figure A.1). Differential abundance testing with DESeq2 (Love *et al.*, 2014) revealed that differences in the composition of the microbiota between sea otters and North American river otters was driven by 50 genera (Appendix A Figure A.2). The top ten genera most characteristic of sea otters were *Bisgaardia*, *Cloacibacterium*, *Cardiobacterium*, *Gromnita*, *Helcococcus*, *Peptoniphilus*, *Guggenheimella*, *Atopobacter*, *Ornithobacterium*, and *Otariodibacter*, whereas the top ten genera most characteristic of North American river otters were *Sporosarcina*, *Paenalcaligenes*, *Kurthia*, *Vagococcus*, *Lysinibacilus*, *Neorhizobium*, *Bacilus*, *Bhargavaea*, *Erysipelothrix*, and *Atopostipes*. A total of 56 ASVs were shared between rectal samples from a set of 13 randomly selected sea otters and 13 North American river otters, rarefied to 83,871 reads. This may be suggestive of otter-associated lineages of bacteria, although we cannot rule out contamination as an alternative possibility.

2.3.4 Sea otter body site communities are distinct from those of adjacent seawater and from one another

Marine mammals are in constant contact with seawater, raising the question of how similar their microbiota is to that of the water around them. Previous research on bottlenose dolphins, humpback whales, and California sea lions found a sharp distinction between the bacterial communities associated with hosts and adjacent seawater (Apprill *et al.*, 2010; Bik *et al.*, 2016). The community composition of wild sea otter gingival and rectal samples was also distinct from that of seawater and from each other (Adonis permanova: $p = 0.001$) (Figure 2.4a). Within-sample diversity (alpha diversity) differed at the ASV level between sea otter body sites (Kruskal-Wallis: $p < 0.001$, post-hoc Dunn test: all groups significantly

Species	Lifestyle	Habitat	Primary dietary components	IUCN status
Sea otter[1]	Marine	Nearshore marine environment	Marine invertebrates such as sea urchins, clams, abalone, crabs, and snails, as well as fish	Endangered
North American river otter[2]	Semi-aquatic	Lakes, streams, rivers, ponds, swamps, marshes	Mostly fish, as well as amphibians (mostly frogs) and crustaceans (mostly crayfish)	Least concern
Asian small-clawed otter[3]	Semi-aquatic	Streams, rivers, irrigated rice fields, mangroves, tidal pools, swamps	Invertebrates such as crabs, snails and other molluscs, and insects, as well as fish	Vulnerable
Giant otter[4]	Semi-aquatic	Large, slow-moving rivers, streams, and lakes	Mostly fish, some caiman and turtle	Endangered

Table 2.1: Habitat and ecology of four otter species included in this study. Lifestyle, habitat, primary dietary components, and IUCN status of sea otters, North American river otters, Asian small-clawed otters, and giant otters. Citations as follows, 1: Doroff & Burdin, 2015; 2: Serfass *et al.*, 2015; 3: Wright *et al.*, 2015; 4: Groenendijk *et al.*, 2015.

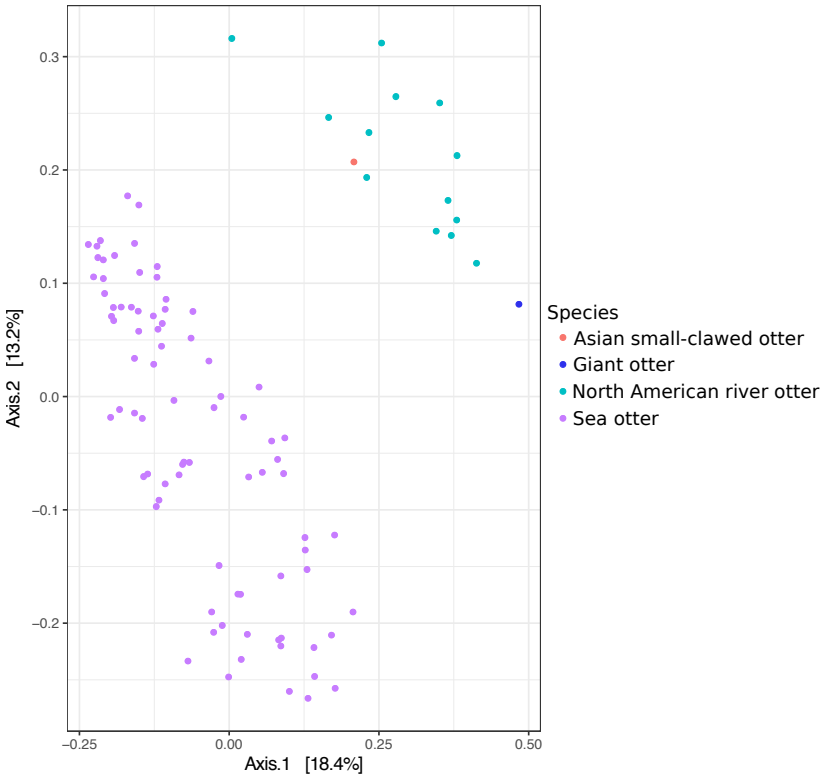


Figure 2.3: Comparison of distal gut communities obtained from different otter species. Principal coordinates analysis (PCoA) Unifrac ordination of sea otter (n = 82), North American river otter (n = 13), asian small clawed otter (n = 1), and giant otter (n = 1) rectal microbiota composition based on 16S rRNA gene amplicon sequences. Taxa were collapsed at the genus level prior to comparison.

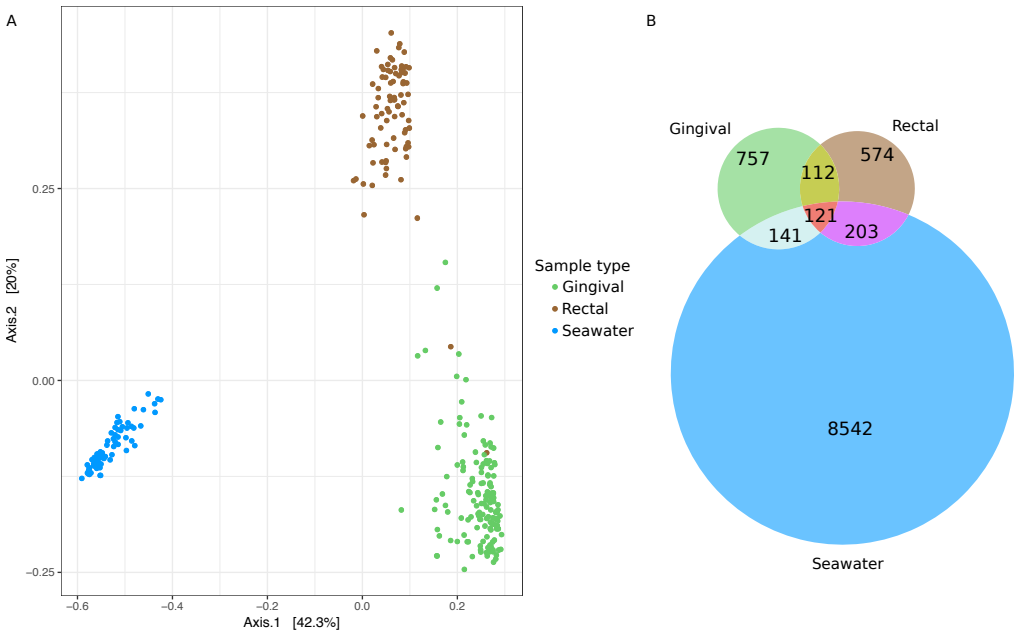


Figure 2.4: Relationship between sea otter gingival, rectal, and adjacent seawater bacterial communities. (A) Principal coordinates analysis (PCoA) of sea otter gingival samples, calculated with the unweighted Unifrac distance metric. Colours indicate sample type. (B) Venn diagram demonstrating overlap in ASVs between gingival, rectal, and seawater samples from a set of 38 sea otters from whom all three sample types were collected. Samples were rarefied to a depth of 83,871 reads prior to comparison.

different from one another with $p < 0.001$). Median alpha diversity was highest for seawater, intermediate for oral samples, and lowest for rectal samples (Appendix A Figure A.3), similar to what is seen in other mammals such as bottlenose dolphins (Bik *et al.*, 2016).

To explore the extent to which ASVs overlapped between environments, we studied the 38 sea otters from whom gingival, rectal, and adjacent seawater samples were collected. The majority of ASVs in each environment were unique to that environment (Figure 2.4b). Between gingival and rectal samples, 23% (233 out of 1,010) of rectal ASVs overlapped with gingival ASVs whereas 20% of gingival ASVs overlapped with rectal ASVs. 24% of sea otter-associated ASVs were shared with seawater (n = 465 out of 1908) whereas only 5% (n = 465 out of

9,007) of seawater ASVs were found in sea otter samples.

To better understand whether the ASVs shared between body sites and seawater were common and highly abundant, we compared the prevalence of ASVs present across the different sampling environments (Appendix A Figure A.4). Many of the most abundant ASVs from both the gingival sulcus and rectum were shared with seawater, but most often these ASVs were present in low abundance in seawater. For example, the most prevalent gingival ASV, from the genus *Mannheimia*, was found in relative abundances as high as 69%, 37% and 0.15% in gingival, rectal, and seawater samples, respectively, whereas the most prevalent ASV in rectal samples, from the species *Otariodobacter oris*, was found in maximum relative abundances of 72%, 1.32%, and 0.02% in rectal, gingival, and seawater samples, respectively. There are two possible interpretations. First, these results could be due to sea otter defecation following capture, a frequently observed event, contaminating sea water. Alternatively, the ASVs that are present in sea water may seed sea otter microbial communities. More specifically, ASVs that are not necessarily highly competitive in seawater may be competitive in the sea otter body (i.e. selected for in that environment), leading to the establishment of resident populations.

2.3.5 Wild sea otter gingival communities conform to one of three composition types

The dearth of gingival microbiota datasets available for mammals hindered our ability to make meaningful comparisons of sea otter gingival microbiotas to those of other host species. Instead, we investigated the landscape of variation in bacterial community composition and structure that exists within the gingival microbiota of sea otters. To do so, we performed clustering of gingival communities

using the k-medoids algorithm, which is a statistically robust variant of k-means that also assures that cluster centers are defined by pre-existing data points (Kaufman & Rousseeuw, 1987). The k-medoids algorithm determines a set of points that minimizes the mean distance between the cluster centers and all other points within the cluster. To better understand which taxa were related to how samples were assigned to clusters, we performed a redundancy analysis (RDA) (Ter Braak, 1986; Legendre & Legendre, 1998), which is a constrained form of principal component analysis (PCA). In this analysis, axes were constrained by CP assignment. Since differences between communities may be driven by changes in closely related taxa (for example see Ravel *et al.*, 2011), the Bray-Curtis compositional dissimilarity metric was selected since it does not account for phylogenetic relatedness.

We identified three different types of community profiles (CPs) within the sea otter gingival samples, which most likely occur along a gradient rather than being distinct entities. The gingival CPs displayed varying degrees of community evenness and were dominated by different taxa (Figure 2.5, Appendix A Figure A.5). CP1 was more diverse than the other two CPs (Kruskal-Wallis test: $p < 0.001$, post-hoc Dunn test: $p < 0.001$). The ASV that was most frequently the most abundant in CP1 gingival samples was from the species *Neisseria animaloris* (top most abundant ASV in 27 out of 50 CP1 samples, median relative abundance of 19.33%±8.20 median absolute deviation across all CP1 samples), while the most frequently most abundant ASV in CP2 samples was from the genus *Streptococcus* (top most abundant ASV in 21 out of 35 CP2 samples, median relative abundance of 26.46% ± 7.84 median absolute deviation across all CP2 samples), and from the genus *Mannheimia* in CP3 (top most abundant ASV in 58 out of 59 CP3 samples, median relative abundance of 69.11% ± 8.37 median absolute deviation

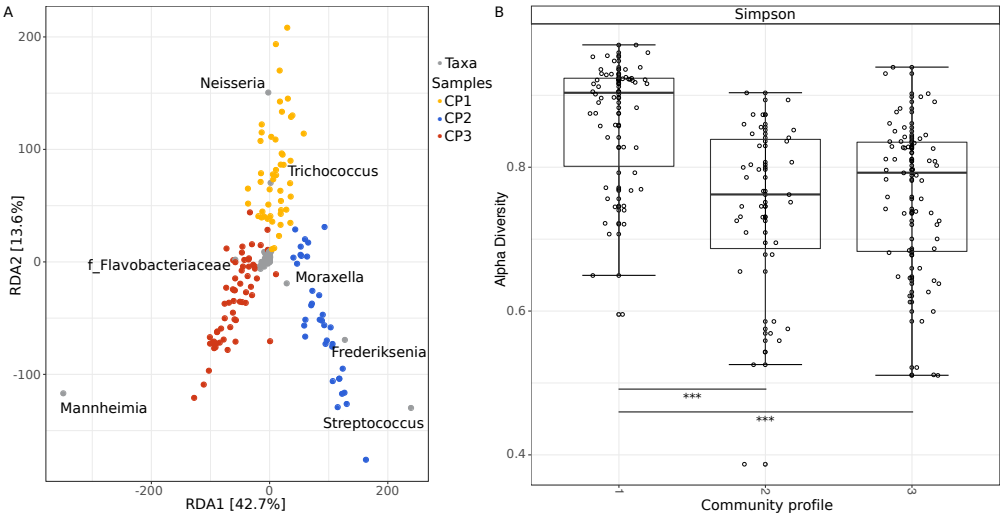


Figure 2.5: Community profiles characteristic of gingival samples from wild sea otters. Community profiles (CPs) vary in terms of composition and diversity as measured by the Simpson's diversity index. (A) Redundancy analysis (RDA) on Bray-Curtis distances with axes constrained by CP. The relationship between community profile and the ASVs that differ substantially between CPs is shown. Taxa are denoted by grey circles and are labeled by genus, or if no genus was assigned, by family. CPs are denoted by blue, yellow, or red circles. (B) Simpson's diversity for the three gingival CPs. Bars denote significance. Three stars represents a p-value < 0.001.

across all CP3 samples). Differential abundance testing with DESeq2 (Love *et al.*, 2014) revealed that 40 ASVs were differentially abundant between gingival CPs (Appendix A Figure A.6).

We next sought to understand the drivers of variation in sea otter-associated gingival bacterial community composition. Examination of PCoA ordinations led us to hypothesize that CP was correlated with the time at which sea otters were sampled, which was in fact statistically supported (Fisher's test, month captured: $p < 0.001$; year captured: $p < 0.001$) (Appendix A Figure A.7, A.8). However, sampling events occured in discrete boughts in which month, year, and location were intertwined (Appendix A Figure A.8, A.9, A.10). For example, while September samples were heavily biased towards CP3 (69%), 61% of September

samples were collected at Elkhorn Slough and Elkhorn Slough sea otters were only sampled one time during our study. As such, we cannot discriminate between the effects of time of sampling versus sea otter habitat on gingival bacterial community composition.

2.3.6 *Helicobacter* in sea otter samples

Members of the *Helicobacter* genus are potentially pathogenic in mammals, yet their prevalence in sea otters, let alone their effect on sea otter health, is unknown. In total, 11 *Helicobacter* ASVs were detected in sea otter-associated samples; five in gingival samples, seven in rectal samples, and two in seawater samples (Figure 2.6). Notably, one of the ASVs that was differentially abundant between gingival CPs was from the *Helicobacter* genus (ASV 1 in Figure 2.6). This ASV was relatively common in CP 2 samples (present in 40% (14 out of 35) samples, median relative abundance $0.04\% \pm 0.03$ M.A.D within sea otter with the ASV) but rare in others (CP1: present in 6% of samples (3 out of 50), median relative abundance $0.04\% \pm 0.02$ M.A.D.; CP3: present in <2% of samples (1 out of 59), relative abundance 0.009%).

2.3.7 Sources of DNA and taxonomic composition of the sea otter fecal metagenome

In an attempt to gain insight into the functional potential of the sea otter gut microbiome, we performed shotgun sequencing using the Illumina HiSeq platform on DNA extracted from one fecal sample from each of 12 sea otters, as described in the Methods section. We generated >879 Gbp data in total, with a minimum of 36 Gbp, median of 48.5 Gbp, and maximum of 208 Gbp per sample. Paired-end sequencing data for each sample was assembled into contigs (see Methods).

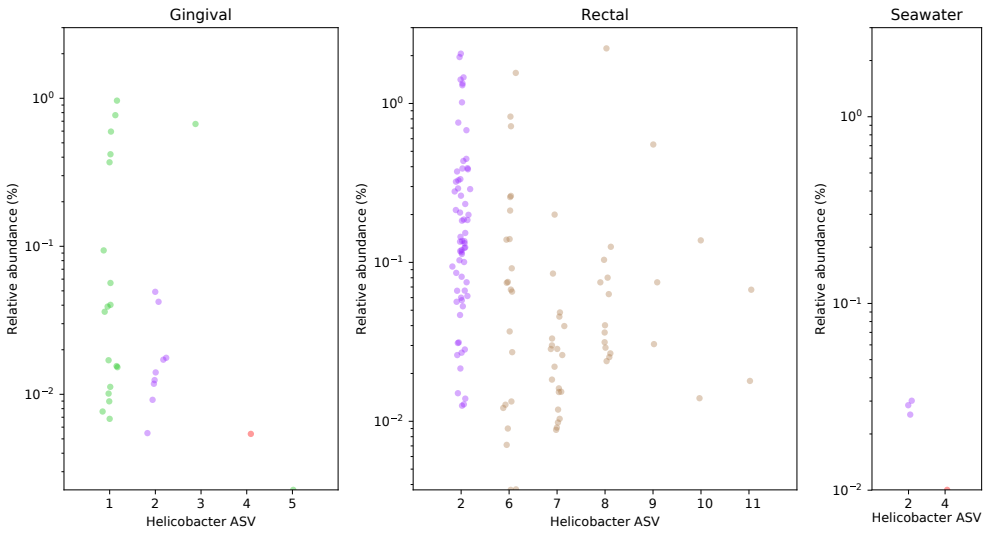


Figure 2.6: Relative abundance of *Helicobacter* ASVs in gingival, rectal, and seawater communities. The relative abundance of 11 *Helicobacter* ASVs present in sea otter gingival, rectal, and seawater communities is shown. No *Helicobacter* sample was present in every sample of a given type. Per panel, *Helicobacter* ASVs are in decreasing order of the number of samples in which they were present for the given sample type. ASVs that overlapped between sample types are indicated by unique colours, whereas those found only in gingival samples are shown in green, and those found only in rectal samples are shown in brown. No *Helicobacter* ASVs were found only in seawater.

Despite the relatively high depth of metagenomic sequencing, few bacterial genomes could be assembled (estimate of the median number of bacterial genomes assembled across samples: 9.5) (Appendix A Figure A.11). The number of bacterial genomes assembled per sample was determined by performing an HMM search for 139 bacterial single copy genes (bSCGs) (Campbell *et al.*, 2013) and using the median of the number of each of the 139 bSCGs identified as a proxy for the number of genomes assembled. This led us to ask: how much bacterial DNA was in the DNA extracts, and what were the other sources of DNA found in these samples?

To answer this question, we first estimated the number of bacterial reads by estimating the total coverage of all bacterial genomes combined and back-calculating from coverage to number of reads (see Methods). The estimated relative abundance of bacterial reads had a median relative abundance of 2.6% (median absolute deviation: 1.8%). To ensure this finding was not simply an artifact due to poor assemblies we also performed a read-based analysis with Centrifuge (Kim *et al.*, 2016), which yielded similar results (Appendix A Data A.1). We next mapped reads against publicly available sea otter and prey species reference genomes and estimated the percentage of reads attributable to each genome (see Methods). The primary prey items for the twelve sea otters studied here were sea urchin, crab, abalone, clam, snail, and mussel (see Methods). Few reference genomes from the same genus, let alone species, are available for sea otter prey species, hindering the analysis. Nonetheless, our results show that DNA extracts from sea otter fecal samples can contain large amounts of eukaryotic DNA from prey consumed by sea otters (Figure 2.7). For example, 63% of reads in sample C8 mapped to the purple sea urchin genome.

We did not pursue an analysis of the bacterial community functional potential

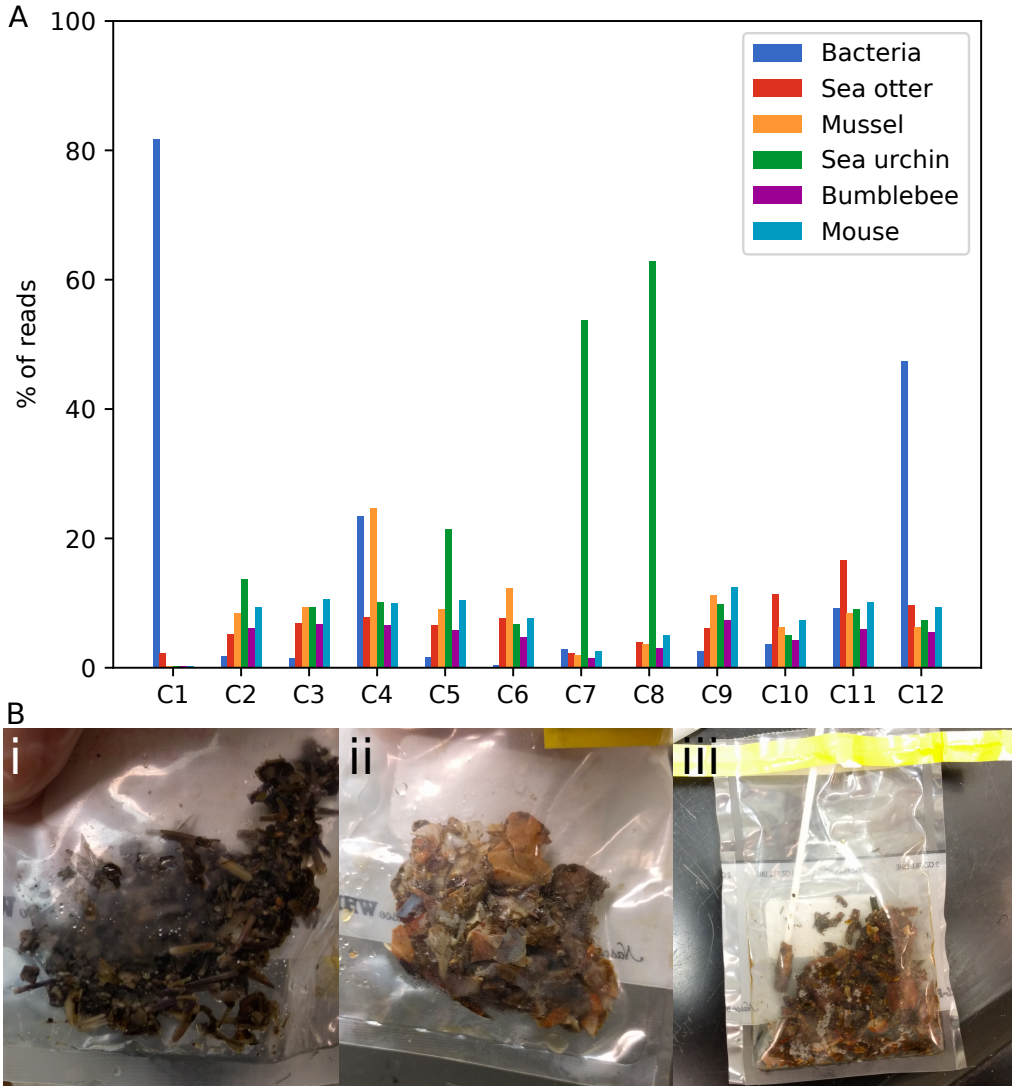


Figure 2.7: Taxonomic composition of reads resulting from shotgun sequencing of sea otter feces. (A) The percentage of reads assigned to each taxon is shown. Mouse and bumblebee genomes were used as control reference genomes to estimate the percentage of reads that map indiscriminately to eukaryotic genomes. The percentage of bacterial reads per sample was back calculated from the median coverage of 139 bacterial single copy genes identified in assemblies. (B) Fecal samples from sea otters are shown in sterile whirl-pak bags. (i) Feces from sea otter C8. Spines and shell of sea urchins are visually prominent components of the sample. (ii) Feces from sea otter C6. Crab shell is a visually prominent component of the sample. Panel (iii) shows a swab after sampling. Care was taken to ensure that no substantial amount of prey tissue (e.g. shell) was attached to the swab after sampling.

due to the poor quality of assemblies, inferred low depth of sequencing of bacterial communities, and low confidence in our ability to distinguish between eukaryotic vs bacterial/archaeal proteins that were predicted.

2.3.8 Marine mammal gut microbiome may have increased potential for the degradation of arthropod-derived chitin compared to the terrestrial mammal gut microbiome

To gain insight into whether marine mammal bacterial gut communities are involved in the degradation of chitin from ingested prey species, we asked whether the chitin-degrading genes present in marine mammal gut microbiomes were more similar to those from the seawater microbiome or the terrestrial mammal gut microbiome. The underlying logic was that a) different types of chitinous structures are produced by crustaceans vs fungi (reviewed in Tharanathan & Kittur, 2010); b) as a result, bacterially-encoded chitinolytic enzymes that degrade these two different types of chitinous structures likely have different modular structures and properties (Bai *et al.*, 2016); c) arthropods are major producers of chitin in marine environments (Cauchie, 2012) (i.e. in seawater and the gut of marine mammals eating chitin-rich marine invertebrate prey); and d) indigenous fungi are likely to be the major producers of chitin in the gut of terrestrial mammals that do not eat chitinous prey.

To test our hypothesis, we first constructed BLAST (Altschul *et al.*, 1990) databases of proteins from two major glycoside hydrolase families with activity for chitin: glycoside hydrolase family 18 (gh18) which are chitinases and family 20 (gh20) which are chitobiases (see Methods for details). Each database had an

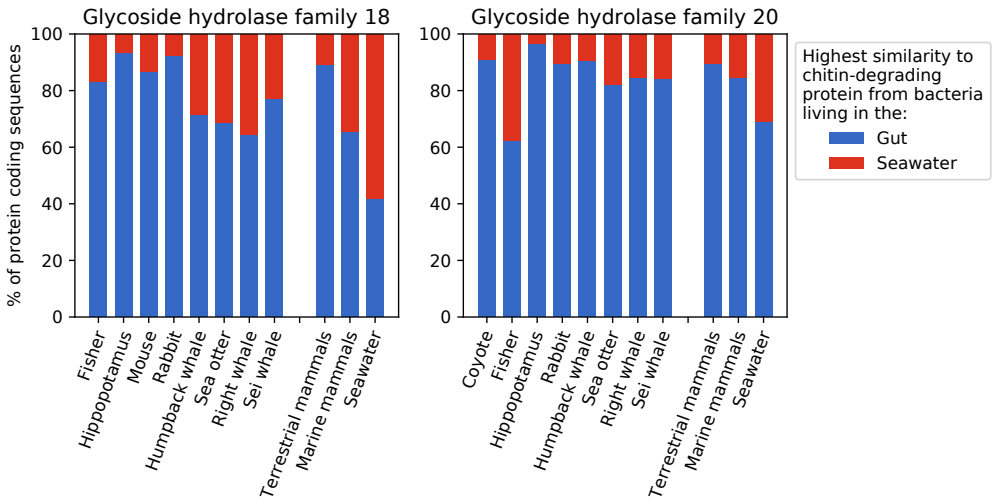


Figure 2.8: Distribution of marine-like vs gut-like chitin-degrading proteins in different environments. For each species, the corresponding bar shows the percentage of chitin-degrading proteins from that species' gut metagenome/s that had the most similarity to those from bacteria living in either the gut of terrestrial mammals or in seawater. Only showing species with sufficient depth of sequencing to have at least 10 glycoside hydrolase proteins detected. Note: dietary components of the fisher in includes insects, which contain chitin in their exoskeleton.

equal number of proteins from terrestrial gut bacteria and from seawater bacteria. Chitin-degrading protein-coding sequences from sea otters, other mammals (Sanders *et al.*, 2015), and seawater (Sungawa *et al.*, 2015) assemblies were then queried against these databases.

More specifically, we compared sea otter metagenomic assemblies to those from six right whales, a coyote, a fisher, a hippopotamus, a rabbit, a humpback whale, a sei whale (Sanders *et al.*, 2015) and five seawater samples (Sunagawa *et al.*, 2015). For both glycoside hydrolase families, marine mammals had a higher fraction of 'marine-like' chitin-degrading proteins than did terrestrial mammals (one-sided Fisher's exact test: $p < 0.001$ for both gh18 and gh20, post hoc pairwise independence test: marine mammals, terrestrial mammals, and seawater are all different from one another, $p < 0.001$ for both gh18 and gh20), as did sea otters

compared to terrestrial mammals (one-sided Fisher's exact test: p = 0.003 for gh18, p = 0.03 for gh20) (Figure 2.8, Appendix A Data A.2).

2.4 Discussion

Marine mammals are ecologically important in that they have the potential to affect structure, function, and nutrient cycling within ecosystems (Bowen *et al.*, 1997; Heithaus *et al.*, 2008; Roman & McCarthy, 2010; Kiszka *et al.*, 2015; and others). The evolution of mammals to a marine lifestyle has independently occurred in six extant, separate lineages (reviewed in Berta *et al.*, 2005), thereby presenting an opportune system in which to study the co-evolution of the microbiota and host in response to a new environment. Doing so, however, is challenging given the difficulties associated with obtaining samples from wild animals who live in the water. Here, we present the first large-scale, culture-independent survey of the bacterial communities associated with the one of these six lineages whose sole representative is the sea otter. First we characterized the gingival and rectal bacterial community composition of 151 wild, healthy sea otters by performing a 16S rRNA gene amplicon survey. Next we attempted to learn more about functional potential of the sea otter gut microbiome by performing assembly driven metagenomic analyses on fecal samples from twelve wild sea otters.

It is well established that variation in the composition of the vertebrate gut microbiota is driven by host diet, morphology, and phylogeny (Ley *et al.*, 2008). Provenance, such as whether an animal is marine or terrestrial, also is likely an important determinant (Bik *et al.*, 2016). To better understand the effect of these determinants on structuring bacterial gut communities in otters, we compared samples from sea otters, North American river otters, an Asian small-clawed otter, and a giant otter. All four species are carnivores with simple gut morphologies and

are part of the *Lutrinae* (otter) subfamily, which diverged from the *Mustelinae* (weasels, ferrets, etc.) subfamily 8.7-9.0 million years ago (mya) (Koepfli *et al.*, 2008). Within *Lutrinae*, giant otters diverged from all other otters 7.4-7.7 mya, new world river otters (including North American river otters) diverged from the remaining otters 6.4-6.6 mya, and the lineages giving rise to sea otters and Asian small-clawed otters diverged from one another 4.8-5.0 mya (Koepfli *et al.*, 2008). In contrast to the phylogenetic relatedness of otter species, we found that sea otter distal gut bacterial communities were distinct from that of the three semi-aquatic otter species. This suggests that a marine versus semi-aquatic lifestyle may be a significant determinant of bacterial community composition in otters. Interestingly, one of the strongest determinants of community composition in free-living communities is the environment's salinity (Ley *et al.*, 2008).

When comparing the genera that were differentially abundant between sea otters and North American river otters, two in particular stand out as potentially being associated with mammals living in a marine environment. The first is the genus *Bisgaardia*, which was first proposed in 2011 after it was recovered from ringed seals (Canada), grey seals (Scotland), and a harbour seal (Scotland) (Foster *et al.*, 2011). Another is the genus *Otariodibacter*, which was highly abundant in sea otter rectal samples and was first proposed as a genus in 2012 after the recovery of isolates from California sea lions and a walrus (Hansen *et al.*, 2012).

Microbiome studies of mammals are currently heavily biased towards the gut as opposed to other body sites. In addition to samples of the distal gut, in this study we obtained gingival swab samples. Within sea otter gingival bacterial communities, composition tended towards certain types of community profiles. This phenomenon superficially appeared to be correlated with the time of year in which sampling occurred, although time of year was inextricably linked with

sampling year and location/habitat. While the underlying cause of variation in the sea otter gingival microbiota is unknown, the results point towards environment as playing a prominent role in structuring these communities, rather than this effect primarily being driven by genetic or biological factors such as sex or age. It is also important to note that samples for this study were conducted over the course of six years and therefore it is possible that changes in personnel, etc., may have resulted in changes to the exact way in which samples were collected, although the official sampling protocol remained constant. Future work will be needed to disentangle the effects of time and sampling location, for example by performing an experiment where sea otters in a fixed location are sampled at regular intervals over the course of several years. Regardless of what drives differences in sea otter gingival communities, differences in community profile may be of significance to sea otter health. For example, we found that one strain of *Helicobacter* was more frequently present in gingival CP 2 samples than in the others, with 40% of sea otters with this type of community profile carrying this *Helicobacter* ASV while only 6% and 2% of sea otters with other community profiles did. While the clinical significance of this type of *Helicobacter* is unknown, this finding merits future investigation.

Gastric ulcers are a significant contributing cause of death in sea otters (Kreuder *et al.*, 2003). While the cause of gastric ulcers is unknown, infection by *Helicobacter* may play a role (Shen *et al.*, 2017). In the human stomach, *H. pylori* is associated with diseases such as peptic ulcer disease, gastric adenocarcinoma, and mucosa associated lymphoid tissue lymphoma, and in the human mouth, *H. pylori* infection has been associated with gastric infection and a variety of mouth-related pathologies, although the link is more controversial (reviewed in Adler *et al.*, 2014). In total, eleven *Helicobacter* ASVs were identified in the sea otter 16S rRNA gene

survey, including *Helicobacter cetorum*, which has long been suspected of being a potential etiological agent of disease in a variety of cetaceans and pinnipeds (Harper *et al.*, 2003; Goldman *et al.*, 2009; Goldman *et al.*, 2011; McLaughlin *et al.*, 2011; Davison *et al.*, 2014). While the most common *Helicobacter* ASV in this study was not identified at the species level and is of unknown clinical significance, it's presence in 79% of sea otter rectal samples is noteworthy.

Representatives from several poorly understood, uncultured bacterial candidate phyla, such as Absconditabacteria (SR1) and Saccharibacteria (TM7), were detected in sea otter samples. No cultured representatives exist for these phyla and few genomic studies have been conducted on host-associated representatives (in fact, only five Absconditabacteria genomes from any environment are publicly available through NCBI databases; accessed May 2018). Thus their role in the mammalian microbiome is of interest. Research from human studies suggests that some members of Saccharibacteria and Absconditabacteria may be associated with mucosal diseases (Fredricks *et al.*, 2005; Brinig *et al.*, 2007; Kuehbacher *et al.*, 2008; Griffen *et al.*, 2012), but the underlying mechanisms and the extent to which this finding carries to other mammalian species are unknown. Nonetheless, the presence of representatives from such phyla in sea otters offers an exciting opportunity to study the evolutionary history of such bacterial lineages and their adaptations that facilitate life in a host-associated, or even marine mammal-associated, environment (for example, by comparing against genomes from those associated with dolphins (Dudek *et al.*, 2017)).

In addition to describing the types of bacteria associated with sea otters, we sought to gain insight into the functional potential of the gut microbiome of sea otters. Attempts to do so were hindered by a low amount of bacterial DNA present in fecal samples. While this is not surprising, it is unusual that appre-

ciable amounts of DNA from food items was detected or found to be of concern (for counter examples from a variety of species, see Zhu *et al.*, 2010; Zhu *et al.*, 2011; Swanson *et al.*, 2011; Lavery *et al.*, 2012, Xu *et al.*, 2013; Xu *et al.*, 2015; Lloyd-Price *et al.*, 2017 and others). Hammer *et al.* (2017) suggested that high proportions of DNA from food may be indicative of a low biomass microbiota and that a reduced dependence on the gut microbiota may be widespread, especially in species with short guts and rapid gut transit times. We hypothesize that the sea otter gut may harbour relatively low bacterial biomass compared to other mammals, in line with the following observations: a) sea otters have an extremely rapid gut transit time on the order of 3 hours (Kirkpatrick *et al.*, 1955; Kenyon, 1969; Costa & Kooyman, 1984), which may make it difficult for bacterial populations to become established let alone contribute to digestion. Notably, despite the fast gut transit time sea otters assimilate a relatively high proportion of the energy ingested (Costa, 1982); b) sea otters are carnivores and therefore consume easily digestible animal-derived polysaccharides (Ley *et al.*, 2008; Doroff & Burdin, 2015), which may obviate the need to obtain breakdown products of complex polysaccharides from bacteria; and c) sea otters have elevated metabolisms that require high energy consumption rates (Costa and Kooyman, 1982), and therefore energy-consuming bacteria compete for a precious resource. These factors may reduce the degree to which a gut bacteria may be beneficial to the host species and are therefore 'tolerated' and/or able to establish resident populations in the gut (Dethlefsen *et al.*, 2007). An interesting follow-up experiment could be to measure changes in the assimilation efficiency and weight of sea otters given antibiotics, and thereby infer the extent to which the microbiota is involved in energy-acquisition of food from the gut (see Fadley *et al.*, 1994; Cho *et al.*, 2012).

Exploitation of new resources is a known driver of the co-evolution between

mammalian hosts and their gut bacteria (Ley *et al.*, 2008). For example, evolution to herbivory has occurred multiple times, and gut bacteria capable of degrading complex polysaccharides found in plant matter, which are otherwise not accessible to hosts, have repeatedly been acquired (Russell *et al.*, 2001; Stevens *et al.*, 2004). Sea otters (mustelids), baleen whales (cetaceans), and certain types of seals (pinnipeds) all independently moved towards a diet rich in chitinous marine invertebrates upon adaptation to life in the sea (review in Berta *et al.*, 2005). An interesting question is whether or not the bacteria in their gut are capable of utilizing chitin from prey, and if so, whether the host species may benefit.

In this study, we performed a pilot analysis and found moderate support for the idea that sea otters and other marine mammals have a greater percentage of bacterial chitin-degrading genes that may hydrolyze marine-type chitin than do terrestrial mammals. Such a phenomenon could potentially contribute to the relatively high assimilation efficiency of marine mammals with chitin-rich diets (reviewed in Costa, 1999). However, this result needs to be interpreted with caution for several reasons. First, our study was not able to distinguish between the genomes of resident gut bacteria vs those that may have been ingested along with prey items (for example, that live on crab shells, etc). Second, the analysis did not take into account the depth of sequencing within a community (i.e. potential resource partitioning). This means that if bacteria in a community use chitin differently (ex: if the most abundant bacteria do not utilize prey-derived chitin but less abundant ones do), the results presented here may not be representative of their respective whole communities. Third, this preliminary analysis was hindered by small sample size of comparable host metagenomes, which amongst other issues means that it does not account for phylogenetic relatedness amongst terrestrial versus marine mammals included in the analysis. Fourth, it is important to

note that the differences between arthropod versus fungi specialized chitinases are still poorly understood, and thus the simple BLAST comparisons performed here are in line with a hypothesis-generating rather than hypothesis-proving analysis. Future studies making use of multi-omic, whole-community sequencing of a large and phylogenetically balanced set of host species approaches will be required to gain maximal insight into this issue.

Sea otters are a charismatic, endangered, keystone species (Estes, 1990; Estes & Palmisano, 1974; Doroff & Burdin, 2015). Characterizing the baseline composition and function of bacterial communities associated with sea otters and other mammals is important for understanding mammalian health and the co-evolution of mammalian hosts with their microbiota. Such insights may ultimately be of use in the management of sick animals and at risk populations. This is especially salient in the face of ongoing changes in ocean ecosystems due to anthropogenically caused disturbances such as pollution and global warming.

2.5 Methods

2.5.1 Sea otter population

This study was conducted on wild, healthy Southern sea otters (*Enhydra lutris nereis*) living offshore of California, USA. Samples were collected from animals under anesthesia during routine population assessments carried out by the United States Geological Survey (USGS), the Monterey Bay Aquarium (MBA), and the California Department of Fish and Wildlife (CDFW) between 2011 and 2017. Wild sea otters were captured using a net and brought onto a boat prior to sedation. Details regarding animal capture and handling are described elsewhere (ex: Monson *et al.*, 2001). Sea otter samples were collected under permit number

MA672624-20.

2.5.2 Sample collection for the 16S rRNA gene amplicon survey from sea otters

The 16S rRNA gene survey consisted of samples from 151 sea otters. Gingival swabs were obtained from sea otters by brushing the lower left gingival sulcus eight to 10 times with a sterile foam Catch-All sample collection swab (Epicenter, WI, Cat. No. QEC091H). Rectal swabs were obtained by inserting the tip of a sterile swab (same brand) about an inch into the rectum and rotating the swab five to six times against the rectal wall. Accompanying seawater samples were collected once a sea otter was captured and divers were en route back to the boat. Seawater was collected in sterile 50 ml tubes by scooping water from the surface down to about one foot in depth and back up to the surface again. Samples were chilled on ice in coolers until return to land, at which point they were transferred to a -80°C freezer for long-term storage.

2.5.3 Sample collection for the 16S rRNA gene amplicon survey from other otters

Samples from North American river otters were collected from captive otters at the following zoos: the Hogle Zoo (UT, USA), the Cincinnati Zoo and Botanical Garden (OH, USA), the San Francisco Zoo (CA, USA), and the Potawatomi Zoo (IN, USA). The giant otter sample was collected from an individual at the Cincinnati Zoo and Botanical Garden (OH, USA) and the Asian small-clawed otter sample was collected from an individual at Six Flags Discovery Kingdom (CA, USA). The sampling protocol was the same as for sea otters.

2.5.4 DNA extraction, 16S rRNA gene amplification, and amplicon sequencing

Genomic DNA was extracted using the QiaAMP DNA Mini Kit (Qiagen, Valencia, CA, Cat No. 51304) as described in Bik *et al.* (2016). A total of 29 negative DNA extraction controls were included in the study. The V4 region of the 16S rRNA gene was PCR amplified in triplicate using barcoded 515F forward primers (5'-GTGYCAGCMGCCGCGGTAA-3') and the 806rB reverse primer (5'-GGACTACNVGGGTWTCTAAT-3'). PCRs were performed using the 5 Prime Hot Master Mix (Quantabio, MA, Cat No. 2200410) as follows: 3 minutes at 94°C, followed by 30 cycles of 45 seconds at 94°C, 1 minute at 50°C, and 1.5 minutes at 72°C, followed by 10 minutes at 72°C. Per 96-well reaction plate, we included one negative PCR control in which sterile molecular biology grade water (Sigma-Aldrich, MO, Cat No. W4502-1L) was added in the place of DNA extract. Triplicate reactions were pooled and PCR cleanup was performed using the UltraClean 96 PCR Cleanup Kit (Qiagen, CA, Cat No. 12596-4), after which DNA was quantified using the Quant-iT dsDNA assay kit (Thermo Fisher Scientific, MA, Cat No. Q33120) and pooled in equimolar ratio using an epMotion 5075 liquid handler (Eppendorf, Germany). Pooled DNA was run through a Zymo Clean and Concentrate Spin Column (Zymo Research Corporation, Irvine, CA, Cat No. D4013) and further purified using the QIAquick gel extraction kit (Qiagen, Hilden, Germany, Cat No. 28704). Amplicons were sequenced across a single 2 x 250nt Illumina HiSeq 2500 lane at the W.M. Keck Center for Comparative Functional Genomics at the University of Illinois, Urbana-Champaign (USA).

2.5.5 Amplicon sequence variant inference, taxonomic assignment, and filtering

Demultiplexing was performed using QIIME version 1.9.1 (Caporaso *et al.*, 2010). Amplicon sequencing variants (ASVs) were inferred using DADA2 version 1.6.0 (Callahan *et al.*, 2016), following guidelines provided in the 'Big Data Workflow' (https://benjjneb.github.io/dada2/bigdata_paired.html). In brief, forward and reverse reads were trimmed to lengths of 245 nt and 200 nt, respectively. ASVs were inferred separately for forward and reverse reads using lane-specific error rate profiles, and paired reads were merged. DADA2's 'removeBimeraDenovo' function was used to identify and remove chimeras from sample datasets and taxonomic assignments were created using the DADA2 'assignTaxonomy' and 'assignSpecies' functions, using RDP training set 16 as a reference database (Cole *et al.*, 2014). This yielded 29,625 ASVs represented by 48,972,130 reads.

Additional stringent filtering of ASVs was accomplished as follows. We performed a second round of chimera screening and removal using VSEARCH version 2.8.0 (Rognes *et al.*, 2016). From this set, ASVs that were not assigned to the bacterial domain, as well as mitochondrial or chloroplast ASVs, were removed based on taxonomic assignments from DADA2 (Callahan *et al.*, 2016). Subsequently, ASVs with low sequence similarity to known bacterial 16S rRNA gene sequences were removed. This was achieved by using BLAST version 2.7.1 (Altschul *et al.*, 1990) to query ASVs against the Schulz *et al.* (2017) 16S rRNA gene set, which is a high-quality set of 16S rRNA genes that are representative of the phylogenetic diversity across the bacterial tree of life. More specifically, we built a BLAST (Altschul *et al.*, 1990) database consisting of 16S rRNA gene sequences that Schulz *et al.* (2017) recovered from genomes in the IMG database (Markowitz *et al.*, 2011) genomes (IMGG_SSU1200.fasta) and metagenomes in IMG (bac-

SSU_prefiltering.fna), and removed sequences that were <1200 bp or contained N's or X's. ASVs with with <50% identity over <50% length were discarded. We also removed ASVs with anomalous lengths (>1.25% expected length of 235 nt). Finally, contaminant ASVs were filtered from the dataset with Decontam version 0.99.3 (Davis *et al.*, 2017), such that contaminant ASVs identified with either the frequency (threshold 0.1) or prevalence method (threshold 0.5) were removed. In total, 3,274 ASVs were removed during this filtering pipeline.

Only samples with greater than 83,871 reads were retained for analysis. This cut-off was selected as it represented the 1st quartile of sampling depth (including control samples, those from other species, sea otter fecal samples not included in analysis due to low sample size, etc) and was substantially higher than the maximum number of reads produced from any DNA extraction control or PCR negative control (n_{max} = 2,480). The final dataset of wild sea otter samples consisted of 17,795 ASVs represented by 34,235,857 reads across 301 samples.

2.5.6 Determination of alpha and beta diversity

Alpha and beta diversity was calculated using the R package phyloseq (McMurdie & Holmes, 2013) without rarefaction of data beforehand (note: these alpha diversity estimators automatically deal with differences in library size). To calculate diversity metrics that consider the phylogenetic relatedness of ASVs, we used fragment insertion to insert ASVs into a reference phylogeny. This was achieved using the QIIME2 version 2018.4.0 fragment insertion module (a wrapper for SEPP) (Caporoso *et al.*, 2010; Warnow, 2015; Janssen *et al.*, 2018) and the Greengenes 13_8 99% reference phylogeny (DeSantis *et al.*, 2006).

2.5.7 Clustering and comparison of sea otter gingival communities

Clustering analysis was based on that performed by DiGiulio *et al.* (2015). Briefly, a Bray-Curtis dissimilarity matrix was calculated for all sea otter gingival samples. Denoising of the matrix was performed by selecting eigenvectors with a significance $\geq$0.05. Partitioning (clustering) of the data was performed using pam in R (Reynolds *et al.*, 1992) after determining the number of clusters (k = 3) from the gap statistic (Tibshirani *et al.*, 2000; Tibshirani *et al.*, 2001; Broberg, 2006) (Appendix A Figure A.12).

Redundancy analysis (RDA) (Braak & Caro, 1986; Legendre & Legendre, 1998) was performed on the relationships between the gingival community profiles and bacterial communities using a Bray-Curtis dissimilarity matrix.

2.5.8 Differential abundance testing

Differential abundance testing revealed ASVs that differ between between distal gut communities of sea otters and North American river otters and between gingival community profiles. In both cases, Bray-Curtis dissimilarity matrices were calculated from raw ASV counts using phyloseq and differential abundance testing was performed using the phlyoseq wrapper for DESeq2 (McMurdie & Holmes, 2013; McMurdie & Holmes, 2014; Love *et al.*, 2014).

2.5.9 Sample collection and selection for shotgun sequencing

DNA extractions from sea otter rectal swabs tended to have insufficient amounts of DNA for shotgun sequencing. Therefore we obtained sea otter fecal samples

from a BioBank at the California Department of Fish and Wildlife in Santa Cruz, California, USA. Sample collection during live captures occurred when a sea otter defecated after being captured in a box off the side of a boat. Due to the nature of working with live, wild animals, potential contamination sources include seawater, sea otter feet/fur, the box, and kelp. Feces were transferred into a sterile Whirl-Pak bag and transferred to a -80°C freezer for long term storage upon return to land.

We selected twelve samples for shotgun sequencing. Selection was based on which sea otters were observed to exhibit the most extreme diet specialization at the prey phylum level (see Watt *et al.* (2000) for details on observational estimation of diet composition). Sub-samples of feces were collected for DNA extraction by inserting a sterile foam Catch-All sample collection swab (Epicenter, WI, Cat. No. QEC091H) into feces and maneuvering the swab such that it came in contact with as much of the fecal material as possible.

2.5.10 Shotgun sequencing and quality filtering

We performed shotgun sequencing on fecal samples from twelve wild sea otters with known diets. The same DNA extracts for these samples were used for both 16S rRNA gene amplification and shotgun sequencing, except for in the case of sample C10 which required two separate DNA extractions due to low DNA yields. Library preparation and sequencing was performed at the Keck Center at the University of Chicago at Urbana-Champaign. DNA was purified using a Zymo Clean and Concentrate Spin Column (Zymo Research Corporation, Irvine, CA, Cat No. D4013), after which libraries were constructed using the Kapa Hyper Prep Kit (Kapa Biosystems, Wilmington, MA, Cat No. KK8504) and quantitated by qPCR. The average length of gDNA in the resulting libraries ranged from 236-

644 bp (for more detail, see Appendix A Data A.3). For nine samples a single library was prepared, and for three samples (C4, C8, and C12) two libraries were prepared.

Libraries were sequenced across three Illumina HiSeq 2500 lanes. Each lane was run for 251 cycles using a HiSeq SBS sequencing kit version 4. Lane one consisted of libraries from all twelve samples and produced a total of 151,600,775 pairs of reads (2 x 250 bp). Lane two consisted of libraries from all twelve samples and produced a total of 133,615,084 pairs of reads (2 x 250 bp). Lane three consisted of libraries from the high-interest samples C3, C4, C10, and C12 and produced a total of 154,332,962 pairs of reads (2 x 100 bp). High-interest samples were defined as the two producing the best bacterial assemblies (greatest number of bacterial single copy genes) per depth of sequencing for both arthropod eaters and echinoderm eaters, as this study original intended to compare the functional potential of sea otters with different diets. The total number of reads generated per sample ranged from 17,968,495 - 103,830,968 pairs (Appendix A Data A.4). MultiQC indicated that read files received from the sequencing facility showed adapter levels below 0.1% (Ewels *et al.*, 2016) and thus no additional adaptor removal was performed.

2.5.11 Metagenome assembly and annotation

Reads were assembled into contigs with Megahit version 1.1.1 (Li *et al.*, 2015, Li *et al.*, 2016), using a minimum kmer size of 31, a maximum kmer size of 249, and a kmer step of 10. These assembly parameters were selected as they optimized the total length of contigs assembled that were greater than or equal to one kilobase pair (kbp) long (Supplemental Data File A.1). Optimized assemblies were still highly fragmented, with a median of 1.5% of contigs >1 kbp in length (min:

0.52%, max: 7.1%). To minimize the effect of discrepancies in overall assembly fragmentation when comparing features of different samples, contigs >600 bp were split into pieces that were 300-600 bp long. The average coverage of each contig was determined by using bowtie2 version 2.2.4 (Langmead & Salzberg, 2012) to map reads against fragmented contigs, using the samtools version 1.6 (Li *et al.*, 2009) depth function to compute depth at each position in a contig, and then by calculating the average depth per base across the entire contig (only including non-zero coverage bases). Genes were predicted using the metagenome implementation of Prodigal version 2.6.2 (Hyatt *et al.*, 2010).

2.5.12 Identification of genes of interest

Bacterial single copy genes and chitin-degrading genes were identified with HMMER suite version 3.1b2 (Finn *et al.*, 2011), using each HMM profile's gathering cutoffs to set significance thresholding (–cut_ga). Pre-compiled HMMs for the Campbell *et al.* (2013) set of 139 bacterial single copy genes (bSCG) were obtained from the Anvi'o (Eren *et al.*, 2015) Github repository. HMM alignments for the following chitin-degrading genes were obtained from the Pfam database (Finn *et al.*, 2015): glycoside hydrolase family 18 (chitinases, PF00704), glycoside hydrolase family 19 (chitinases, PF00182), and glycoside hydrolase family 20 (chitobioses, PF00728). Candidate chitinases were filtered by querying them against the NCBI non-redundant nucleotide database (downloaded March 2018) using BLAST version 2.7.1 (Altschul *et al.*, 1990) and an e-value threshold of 1e-10, and discarding sequences with no significant similarity to known chitinases or with a top hit to a eukaryote.

2.5.13 Estimation of the number of bacterial reads per sample

To better understand the composition of our samples we estimated the number of bacterial reads that were sequenced per sample. For each bSCG we computed the total coverage of all scaffolds on which a given bSCG was encoded, then calculated the median total coverage of all bSCGs as a measure of the of depth of sequencing across all bacterial genomes combined. We back-calculated the number of bacterial reads using the formula:

$$\#\ bacterial\ reads = \frac{coverage * number\ of\ bp\ per\ genome}{length\ of\ reads}$$

Since this formula is very sensitive to read length, coverage was estimated using only reads of a set length, which was the mode read length for each sample. This was accomplished by using a modified version of samtools version 1.7 (Li *et al.*, 2009) to compute depth using only reads of a set length. Specifically, the -l function, which tells the program to ignore reads under a user-defined length, was modified to only consider reads of a user-defined length via modification to one line as follows: 'if (aux->min_len && bam_cigar2qlen(b->core.n_cigar, bam_get_cigar(b)) != aux->min_len) continue;'. In the formula for the number of bacterial reads, we estimated that the average bacterial genome was 3 million bp (Land *et al.*, 2015) and that the length of reads was equal to the median length of reads sequenced for each sample. Importantly, error in average bacterial genome size on the order of ones of millions is insignificant given that the difference in genome size between bacterial and eukaryotic genomes tends to be on the order of thousands of millions.

2.5.14 Estimation of prey DNA present in shotgun sequencing reads

Reads generated from sea otter fecal samples were mapped against reference genomes using bowtie2 version 2.2.4 (Langmead & Salzberg, 2012). The following genomes were used as references: mouse assembly GRCm38.p6 (GenBank: GCA_000001635.8) (Church *et al.*, 2009), bumblebee assembly Bter_1.0 (GenBank: GCA_000214255.1) (Sadd *et al.*, 2015), purple sea urchin assembly Spur_4.2 (GenBank: GCA_000002235.3) (Sodergren *et al.*, 2006), Mediterranean mussel (Genbank: ASM167691v1) (Murgarella *et al.*, 2016), and Northern sea otter assembly ASM228890v2 (GenBank: GCA_002288905.2) (Jones *et al.*, 2017). The bowtie2 (Langmead & Salzberg, 2012) overall alignment rate, which represents the total percentage of mates aligned in pairs and mates aligned in singles, was reported when considering the percentage of reads mapping to a given genome.

2.5.15 Chitinase diversity present in sea otters, other mammals, and seawater

Custom databases were constructed that contained protein sequences from bacteria that lived in either seawater or the gut of a terrestrial mammal. One database was constructed for each of the glycoside hydrolase families 18, 19, and 20 as follows. First, the NCBI proteins database was searched using the term 'glycoside hydrolase X', where X was the relevant family number, and all candidate glycoside hydrolase protein sequences were downloaded. Candidates were screened by performing an HMM search for the given glycoside hydrolase using HMMER suite version 3.1b2 (Finn *et al.*, 2011), using each HMM profile's gathering cutoffs to set significance thresholding (–cut_ga). Metadata for chitin-

degrading protein sequences that passed this filter was obtained using the entrez programming utilities (eUtils). Proteins with 'marine', 'sea', or 'ocean' as the isolation source and 'Bacteria' as the organism were flagged as marine bacterial chitin-degrading genes, while proteins with 'gut', 'stool', 'feces', or 'faeces' as the isolation source and 'Bacteria' as the organism were flagged as gut bacterial chitin-degrading genes. To ensure that chitin-degrading protein sequences from the gut and marine ecosystems were equally represented in the custom database for each glycoside hydrolase family, we randomly subsampled to the smallest number of proteins present in the marine or gut group. For glycoside hydrolase family 18, this consisted of 63 gut proteins and 63 marine proteins, while for glycoside hydrolase family 19 and 20, it consisted of 170 and 50, respectively.

Non-eukaryotic chitin-degrading protein-coding sequences from sea otter and other assembled metagenomes used in this study were queried against the glycoside hydrolase databases using BLAST version 2.7.1 (Altschul *et al.*, 1990) and an e-value threshold of 1e-05. The number of best hits to a gut vs marine chitin-degrading protein from each glycoside hydrolase family was recorded. Results for glycoside hydrolase family 19 are not shown because so few glycoside hydrolase family 19 proteins were identified from terrestrial mammals (16 across two out of seven terrestrial mammal samples). Only samples from which at least 10 chitin-degrading proteins per family had a BLAST (Altschul *et al.*, 1990) hit are shown in the final barplot figure.

Chapter 3

Novel microbial diversity and functional potential in the marine mammal oral microbiome

3.1 Abstract

The vast majority of bacterial diversity lies within phylum-level lineages called 'candidate phyla', which lack isolated representatives and are poorly understood. These bacteria are surprisingly abundant in the oral cavity of marine mammals. We employed a genome-resolved metagenomic approach to recover and characterize genomes and functional potential from microbes in the oral gingival sulcus of two bottlenose dolphins (*Tursiops truncatus*). We detected organisms from 24 known bacterial phyla and one archaeal phylum. We also recovered genomes from two deep-branching, previously uncharacterized phylum-level lineages (here named '*Candidatus* Delphibacteria' and '*Candidatus Fertabacteria*'). The Delphibacteria lineage is found in both managed and wild dolphins; its metabolic

profile suggests a capacity for denitrification and a possible role in dolphin health. We uncovered a rich diversity of predicted Cas9 proteins, including the two longest predicted Cas9 proteins to date. Notably, we identified the first type II CRISPR-Cas systems encoded by members of the Candidate Phyla Radiation. Using their spacer sequences, we subsequently identified and assembled a complete Saccharibacteria phage genome. These findings underscore the immense microbial diversity and functional potential that await discovery in previously unexplored environments.

3.2 Introduction

The vast majority of bacterial diversity is found within phylum-level lineages that lack isolated representatives (Hug *et al.*, 2016), commonly referred to as 'candidate phyla'. Candidate phyla constitute at least 103 out of approximately 142 widely recognized bacterial phyla for which there is genomic representation (Anantharaman *et al.*, 2016; Eloe-Fadrosh *et al.*, 2016; Hug *et al.*, 2016); 46% of known bacterial phyla are clustered in the Candidate Phyla Radiation (CPR). However, there remain many phylum-level bacterial lineages that have no genomic representation and are not yet formally recognized (Brown *et al.*, 2015). Genome-resolved metagenomic studies offer unique and unprecedented insights into the biology of these uncultured, poorly understood lineages and their biochemical diversity (Wrighton *et al.*, 2012; Kantor *et al.*, 2013; Brown *et al.*, 2015; Sekiguchi *et al.*, 2015; Eloe-Fadrosh *et al.*, 2016; Hug *et al.*, 2016). In addition to revealing the environmentally and economically important roles played by such bacteria, these studies contribute greatly to our understanding of the distribution of lifestyles across the tree of life. For example, genomes from members of the CPR suggest that they are metabolically sparse and lack many biosynthetic pathways typically

required for life, presumably because these organisms are dependent on other microbes for survival (Kantor *et al.*, 2013; He *et al.*, 2015). Candidate phyla genomes may also reveal novel functional diversity, as phylogenetic diversity is correlated with novel proteomic diversity and biological properties (Wu *et al.*, 2009; Burstein *et al.*, 2017).

Marine mammals are an ecologically important group of animals harboring little-explored microbial communities. Previous research has shown that bottlenose dolphins, in particular, host a rich diversity of novel bacteria (Bik *et al.*, 2016). Nearly 70% of near full-length 16S rRNA genes from the dolphin microbiota were novel in 2015 at the species level, and representatives from 25 bacterial phyla were present in the mouth alone. Furthermore, a surprising number of candidate phyla such as Gracilibacteria (BD1-5/GN02), Modulibacteria (KSB3), and the Parcubacteria (OD1) supergroup, which are unusual in mammal-associated environments, were found in the dolphin mouth (Bik *et al.*, 2016). Genomes from such candidate phyla have nearly exclusively been retrieved from non-host-associated environments, and thus it is unknown how these bacteria adapt to a mammalian environment. Interestingly, despite evidence that the marine mammal microbiota is shaped by the sea, these bacteria were not detected in the adjacent seawater (Bik *et al.*, 2016).

On the basis of these prior observations, we concluded that marine mammals afford an unusual opportunity for studying bacterial diversity. Working under the hypothesis that novel phylogenetic diversity correlates with novel functional diversity, in this study we applied genome-resolved metagenomics to investigate the diversity and functional potential of the dolphin oral microbiome. The results hint at the wealth of evolutionary and biochemical diversity that remains uncharted within previously unexplored environments, including mammalian mi-

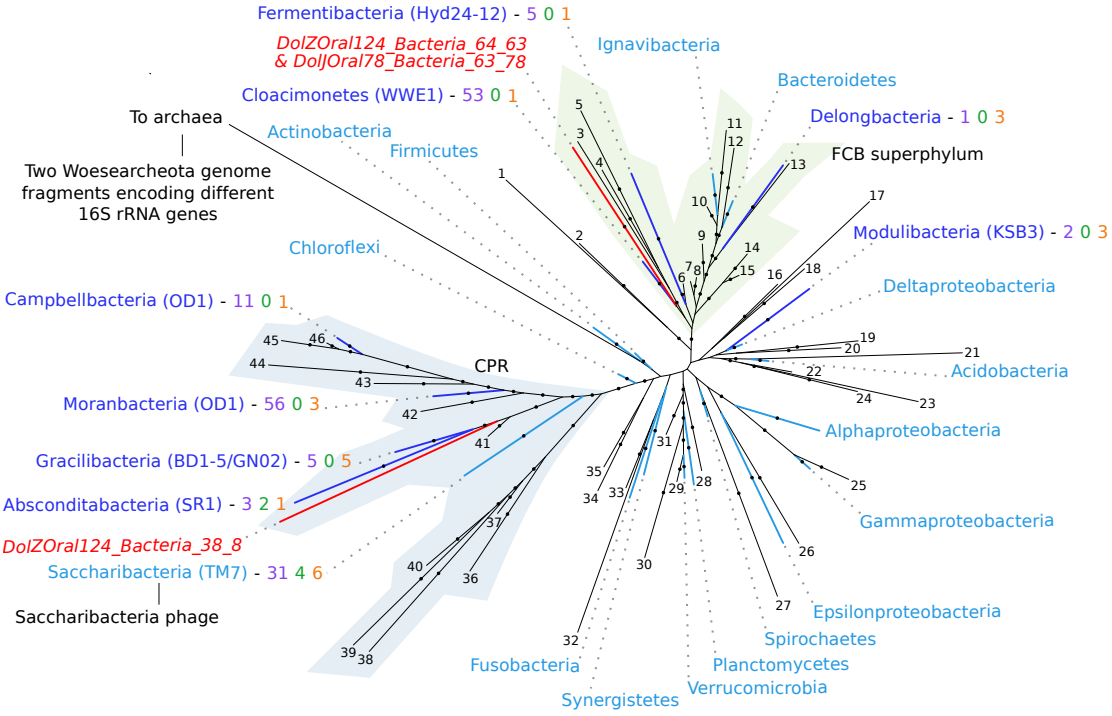
crobiomes, and will contribute to future comparative studies of host-associated versus non-host-associated candidate phyla bacteria.

3.3 Results

3.3.1 Dolphin oral microbiota composition and structure

Swab samples were collected from the gingival sulcus of healthy bottlenose dolphins (*Tursiops truncatus*) under the purview of the U.S. Navy's Marine Mammal Program in San Diego Bay, California. Samples from two dolphins were selected for shotgun sequencing based on the findings of Bik *et al.* (2016), which indicated that these two samples (DolJOral78 and DolZOral124) contained representatives from nine candidate phyla at relative abundances of $\geq$0.05% (Appendix B Table B.1). Paired-end Illumina HiSeq reads were generated, filtered, assembled, and used to recover microbial genomes, as described in the Methods.

From >63 Gbp of filtered paired-end sequences, we recovered 107 draft-quality genomes from 24 previously described bacterial phyla and one circular genome from a candidate Saccharibacteria (TM7) phage (presented below). These genomes derived from 22 organisms affiliated with the candidate phyla Absconditabacteria (SR1), Campbellbacteria (OD1), Cloacimonetes (WWE1), Delongbacteria, Fermentibacteria (Hyd24-12), Gracilibacteria (BD1-5/GN02), Modulibacteria (KSB3), and Moranbacteria (OD1), and the Saccharibacteria (TM7) phylum. Phylum-level assignments (or lack thereof, as was the case for three of our genomes) were determined by constructing a phylogeny based on an alignment of 15 concatenated ribosomal proteins (Figure 3.1, Supplemental Data File B.1; see Methods). Of note, we were able to link a 16S rRNA gene sequence to a member of the Delongbacteria phylum, which previously consisted of a single genome for which no



Phyla from which genomes were recovered from the dolphin mouth:
Candidate phylum Other phylum Novel lineage

Numbers next to candidate phyla:
genomes in NCBI databases # animal-associated genomes # dolphin-associated genomes

Reference phyla:

1 - Poribacteria
2 - Coatesbacteria
3 - *GUT_77*
4 - Fibrobacteres
5 - Raymondbacteria
6 - Latescibacteria (WS3) & Handelsmanbacteria
7 - Edwardsbacteria (TA06) & WOR-3
8 - Zixibacteria
9 - Marinimicrobia (SAR406)
10 - Ignavibacteria
11 - Kryptonia
12 - Chlorobi
13 - *Caldithrix abyssi*
14 - Gemmatimonadetes
15 - Glassbacteria
16 - Atribacteria
17 - BRC1
18 - Schekmanbacteria
19 - Rokubacteria
20 - Chrysiogenetes
21 - Fischerbacteria
22 - Nitrospirae
23 - *Nitrospira sp OLB3*
24 - Dadabacteria
25 - Betaproteobacteria
26 - Dependentiae (TM6)
27 - Hydrogenedentes (NKB19)
28 - Elusimicrobia
29 - Lentisphaerae
30 - Chlamydiae
31 - Omnitrophica
32 - *Caldiscericum exile*
33 - Thermotogae
34 - Cyanobacteria
35 - Margulisbacteria
36 - Curtisbacteria
37 - Gottesmanbacteria (OP11)
38 - Beckwithbacteria (OP11)
39 - Woesebacteria (OP11)
40 - Amesbacteria (OP11)
41 - Peregrinibacteria (PER)
42 - Falkowbacteria (OD1)
43 - Azambacteria (OD1)
44 - Wolfebacteria (OD1)
45 - Nomurabacteria (OD1)
46 - Taylorbacteria & Zambryskibacteria (OD1)

16S rRNA gene had been recovered (Anantharaman *et al.*, 2016). Additionally, low coverage ($\leq$3%) archaeal genome fragments were recovered from two members of the Woesearcheota phylum. Similar sequences have been recovered from host-associated environments (see SILVA database (Pruesse *et al.*, 2007; Quast *et al.*, 2013; Yilmaz *et al.*, 2014)), such as coral heads (Sato *et al.*, 2013) and human skin (Probst *et al.*, 2013), but were not originally recognized as affiliated with the Woesearcheota phylum or placed within a comprehensive phylogeny.

Figure 3.1: Phylogenetic relationships among genomes recovered from the dolphin mouth. The maximum-likelihood tree includes representation from all genomes that contained ≥8 of 15 ribosomal proteins used to infer the phylogeny (with the exception of one Delongbacteria genome with 7 ribosomal proteins) as well as from published genomes. Bootstrap support values ≥50% are denoted with a closed circle on the branches. Branches of phyla with genomic representation in the dolphin mouth are color coded such that dark blue indicates candidate phylum, light blue indicates other phylum, and red indicates novel, deep-branching lineage. Labels for these phyla appear around the tree, with dotted lines indicating the corresponding branch. Numbers next to candidate phyla names indicate the number of genomes from each phylum that are publicly available in NCBI databases prior to this study (purple), the number of those that come from an animal-associated environment (green), and the number that were recovered in this study (orange). Branches of the remaining phyla are included in the tree as references, are colored black, and can be identified using the legend at the bottom of the figure. The CPR is denoted with blue shadowing, and the FCB superphylum is denoted with green shadowing. The topology of the tree with respect to the position of the CPR does not recapitulate that of Hug *et al.* (2016), presumably due to lower sampling depth reducing the ability to resolve the branching order of the deepest lineages. See also Appendix B Figures B.1–B.3, Table B.1, Supplemental Data File B.1.

Bacterial community composition and structure inferred from the same DNA preparations differed depending on the survey method: genome-resolved metagenomics (this study) versus 16S rRNA gene amplification (Bik *et al.*, 2016) (Figure 3.2, Appendix B Figure B.1, Table B.1). Notably, the 16S rRNA gene that was associated with the highest-coverage genome in both samples (17% and 4% relative abundance in DolJOral78 and DolZOral124, respectively; Figure 3.2) was barely detected in the amplicon-based survey (not detected in DolJOral78; 0.04% relative abundance in DolZOral124). This is surprising because the PCR primers match the assembled sequence perfectly, the GC content of the gene is 58%, and it contains no unusual insertions. The two genomes are from the same species of Actinobacteria (order Micrococcales), and the GC content of the genome is 68%. Furthermore, members of the CPR were greatly under-detected using the amplicon-based approach. From the metagenomic assemblies, we detected 16 unique CPR species-level genomes, some of which ranked among the highest-coverage genomes recovered (Figure 3.2). For example, the fourth most abundant bacterial organism in the DolJOral78 sample was a member of the Saccharibacteria phylum (4% relative abundance), although no Saccharibacteria representatives were detected in the DolJOral78 sample in the previous 16S rRNA gene amplicon survey. In the amplicon-based study (Bik *et al.*, 2016), only nine unique operational taxonomic units (OTUs) from the CPR were identified from both samples combined, with a maximum relative abundance of 0.24%. This discrepancy can be explained at least partially by primer mismatches, consistent with previous reports on the CPR (Brown *et al.*, 2015). Of the 21 unique CPR 16S rRNA genes assembled and identified in the metagenomic data, nine span the region between the commonly used 338F and 906R bacterial primers (also used in Bik *et al.* (2016)) and have sufficient read coverage to validate the assembly. Eight

of these have 1–3 mismatches in at least one primer site. In the amplicon study, eight of the nine OTUs were detected among all samples, although only the one OTU with no primer site mismatches was detected in the two samples studied here.

Given the breadth of novel bacterial diversity in the dolphin oral samples, we next searched for novel phage diversity. Using a stringent set of criteria (see Methods), we identified a set of 33 and 55 sequences from DolJOral78 and DolZOral124, respectively, for which we had high confidence in their derivation from phage genomes. These sequences range in length from 1,583 to 119,885 bp (average 19,363 and 21,462; SD ± 13,243 and 19,615 bp). To assess overlap between samples, we performed a reciprocal best-hit BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009) search between phage sequences from the two samples. We identified 14 phage genome fragments that were present (or had close relatives present) in both samples. To evaluate the degree of phage genome novelty, we BLASTed (Altschul *et al.*, 1990; Camacho *et al.*, 2009) phage sequences against the NCBI non-redundant nucleotide database (https://ncbi.nlm.nih.gov/nucleotide). Only three alignments were longer than 1,000 bp, the longest of which was only 2,919 bp. These alignments corresponded to 2.3%, 3.8%, and 8.2% of the lengths of the respective phage scaffolds. This suggests that phages in the dolphin mouth are only distantly related to phages for which genomic fragments have previously been recovered, as one would expect under the hypothesis that novel bacterial diversity begets novel phage diversity.

3.3.2 Novel, deeply divergent phylum-level lineages

The concatenated ribosomal protein tree enabled determination of the phylum-level identity of recovered genomes (Figure 3.1). Within this tree, three genomes

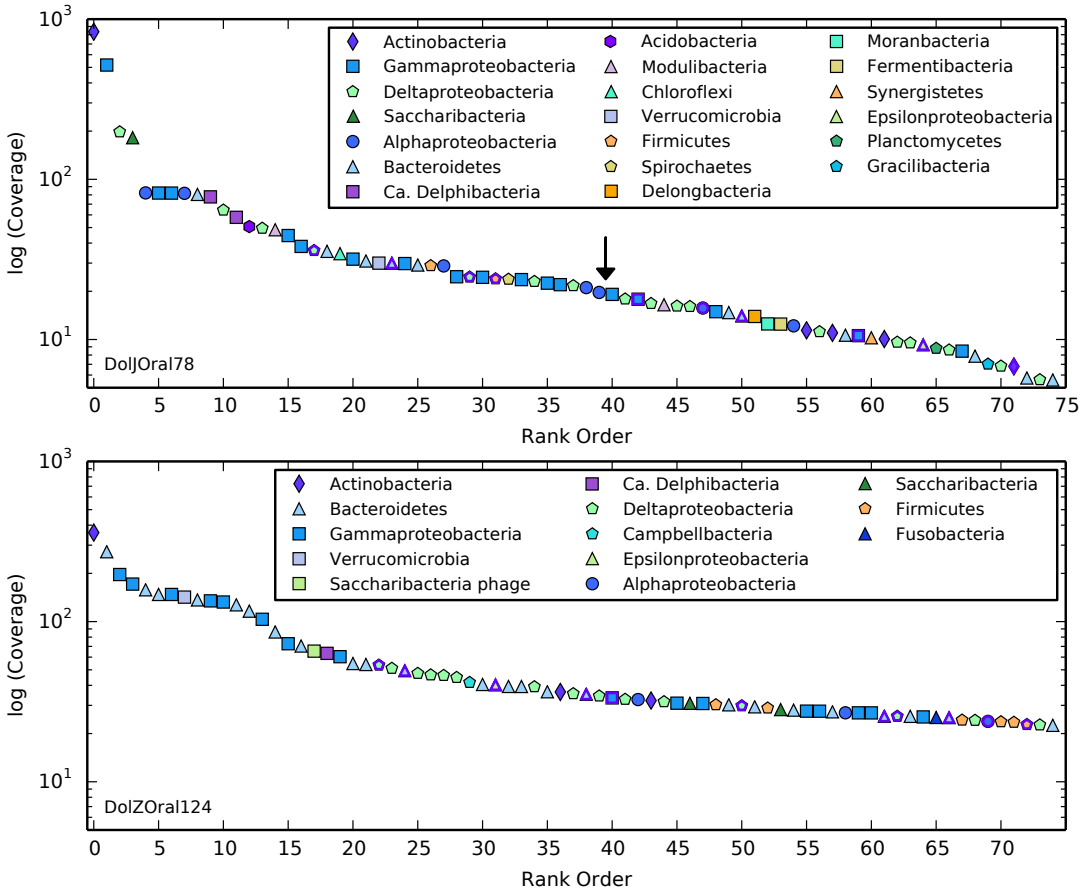


Figure 3.2: Community structure of the dolphin oral microbiota. The top panel presents the community structure of the DolJOral78 sample, and the bottom panel presents that of the DolZOral124 sample. Each symbol represents a bin, which is a set of scaffolds that share similar genomic signatures. In most cases, bins represent of a genome (or fragments of a genome) from a single organism. Bins that contain multiple genomes from organisms with similar genomic signatures are denoted by a purple outline around the symbol. The average coverage of all scaffolds in a bin is represented on the y axis, and bins are ranked in order of decreasing average coverage on the x axis. Due to the complexity of the samples, not all low-coverage genomes could be binned. This point, after which only a portion of genomes could be binned, is denoted by an arrow for DolJOral78 and is not reached in the top 75 bins for DolZOral124. See also Appendix B Figure B.1 and Table B.1.

belonging to two deep-branching lineages eluded identification. To evaluate whether these two lineages were representative of previously undescribed phyla, we examined whether (1) they formed monophyletic lineages in both the concatenated ribosomal protein phylogeny and the 16S rRNA gene phylogeny, and (2) the 16S rRNA gene sequences of such lineages were at least 25% divergent from those of known phyla (i.e., the threshold used by Yarza *et al.* (2014)).

One lineage, for which we propose the name 'Delphibacteria' (rationale in Appendix B Additional Discussion), is affiliated with the Fibrobacteres-Chlorobi-Bacteroidetes (FCB) superphylum and is represented by genomes DolJOral78_Bacteria_63_78 and DolZOral124_Bacteria_64_63. The names refer, for example, to sample DolZOral124, lowest taxonomic resolution Bacteria, GC content of 64%, coverage of 633). The 16S rRNA gene sequence from the Delphibacteria lineage clusters with sequences from what is currently recognized as the Latescibacteria phylum in the SILVA database (Pruesse *et al.*, 2007; Quast *et al.*, 2013; Yilmaz *et al.*, 2014) (see Appendix B Additional Discussion, Figure B.2, Supplemental Data File B.1). The diversity encompassed by this 'phylum' was recently found to be an assemblage of at least two phylum-level lineages: Latescibacteria and the newly proposed Eisenbacteria (Anantharaman *et al.*, 2016). Nearly all members of the Delphibacteria lineage share <75% sequence identity across the 16S rRNA gene with members of the Eisenbacteria phylum (Appendix B Figure B.2A) and <78.5% sequence identity with members of the Latescibacteria phylum (Appendix B Figure B.2B). Predicted proteins in the near-complete genome from this lineage were most similar to those from the Deltaproteobacteria phylum (Appendix B Figure B.3A). Notably, the Delphibacteria lineage was detected in 41 oral samples from 15 of 33 U.S. Navy dolphins and one of ten wild dolphins surveyed with 16S rRNA gene amplicon pyrosequencing in Bik *et al.* (2016),

although it was classified as a member of the Latescibacteria phylum. In the DolJOral78 sample, two Delphibacteria genomes were detected at relative abundances of 1.6% and 1.2%, while in the DolZOral124 sample one Delphibacteria genome was detected at a relative abundance of 0.7%.

The second previously uncharacterized lineage, for which we propose the name 'Fertabacteria' (rationale in Appendix B Discussion), is affiliated with the CPR and is represented by the genome DolZOral124_Bacteria_38_8. The 16S rRNA gene sequence from Fertabacteria clusters with sequences from what is currently recognized as the Peregrinibacteria (PER) phylum in the SILVA database (see Appendix B Additional Discussion, Supplemental Data File B.1). It is part of a well-supported clade with <75% sequence identity to the rest of the Peregrinibacteria phylum, including PER-ii (Appendix B Figure B.2C). Predicted proteins from this lineage are most similar to those from the Peregrinibacteria phylum (Appendix B Figure B.3B), yet the 16S rRNA gene sequence identity argues against its inclusion in this group. Out of all samples surveyed with 16S rRNA gene pyrosequencing in Bik *et al.* (2016), only a single Fertabacteria amplicon was detected. The amplicon was generated from a sample of forcefully expired air ('chuff') from the dolphin respiratory tract collected on sterile filter paper, and was originally classified as a member of the Gracilibacteria phylum. The 906R primer used in Bik *et al.* (2016) had two mismatches to the corresponding priming site, and therefore this organism may have been widely under-detected in the amplicon-based survey. The Fertabacteria genome is one of the lowest-coverage genomes (83X) in this study, with a relative abundance of 0.09% in the DolZOral124 sample.

3.3.3 Functional profile of the Delphibacteria lineage

Due to the abundance and prevalence of Delphibacteria organisms in the dolphin oral samples, we investigated the metabolic potential of the near-complete DolZOral124_Bacteria_64_63 genome. The genome contained 49 of 51 universal bacterial single-copy genes used to assess completeness (Raes *et al.*, 2007), was comprised of 3,362,850 bp, and was predicted to contain 3,011 protein-coding genes. The corresponding organism appears to utilize a variety of compounds as carbon and energy sources, including polysaccharides such as starch/glycogen, acetate, acetaldehyde, ethanol, and butyrate (Figure 3.3, Supplemental Data File B.2). DolZOral124_Bacteria_64_63 carries the potential to ferment to acetate, with ethanol and acetaldehyde being produced during regeneration of NAD+ required for glycolysis. Two of the three genes specific to gluconeogenesis are also present, as are those involved in the non-oxidative pentose phosphate pathway. The genome includes the capacity for amylose synthesis and possibly GDP-L-rhamnose synthesis.

The complete gene complement required for running the forward tricarboxylic acid (TCA) cycle is present. Accordingly, the DolZOral124_Bacteria_64_63 genome is predicted to support aerobic respiration and possibly also anaerobic respiration using nitrogen compounds as terminal electron acceptors. The catalytic subunit of a periplasmic nitrate reductase was detected (*napA*), as were accessory periplasmic nitrate reductase subunits. The catalytic subunit of a nitric oxide reductase (*norB*) and the terminal nitrous oxide reductase (*nosZ*) were also detected. Nitrite reductase genes (*nirK* or *nirS*) were not identified, nor were many of the subunits typically associated with the above reductases. Nonetheless, the presence of catalytic subunits for three out of the four steps involved in converting nitrate to dinitrogen suggests that this Delphibacteria representative is

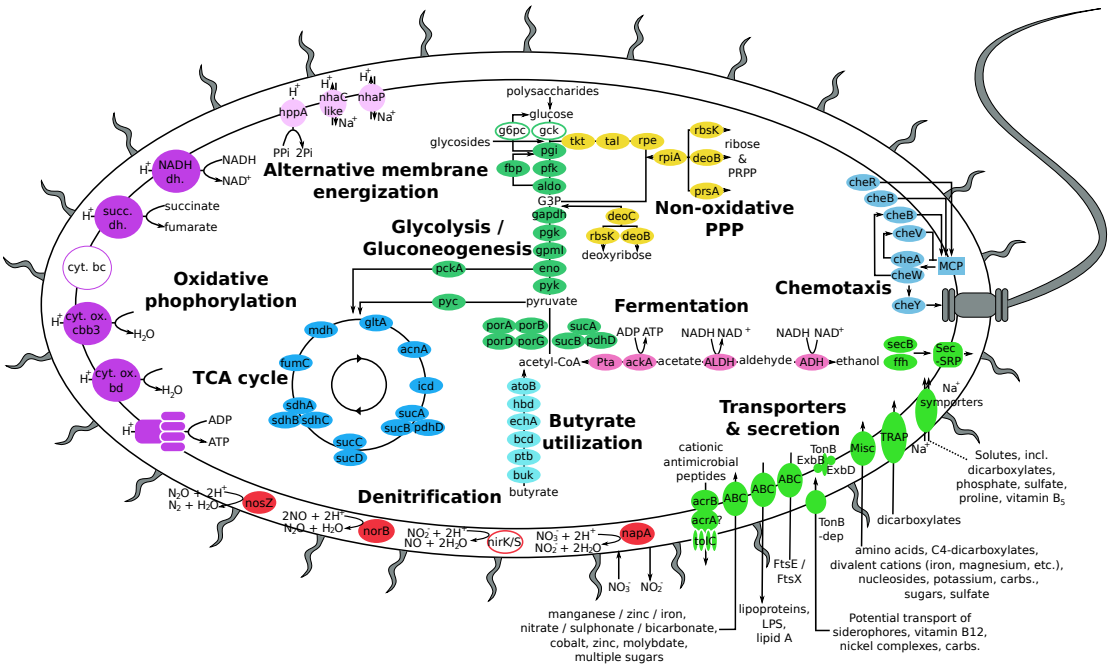


Figure 3.3: Functional profile of Delphibacteria. Key predicted metabolic and functional features are depicted. Genes of interest are denoted by abbreviations in the colored shapes. Filled shapes represent genes predicted to be present or likely to be present, whereas unfilled shapes represent genes that were not identified. See also Appendix B Figure B.3, Supplemental Data File B.2

capable of denitrification. We detected another mechanism for generating proton motive force in the form of a pumping pyrophosphatase, indicating that DolZOral124_Bacteria_64_63 may be able to utilize pyrophosphate as an alternative chemical energy carrier to ATP.

DolZOral124_Bacteria_64_63 is most likely a lipopolysaccharide-producing bacterial species with flagella and type IV pili and capable of chemotaxis. We identified ten acriflavin resistance proteins, which are typically involved in efflux of cationic antimicrobial peptides. Overall, we infer that this is a heterotrophic organism that has the genomic potential for oxygen and most likely nitrate reduction.

3.3.4 Large biosynthetic gene cluster in the dominant Actinobacteria genome

One of the two highest-coverage bins in both samples contained scaffolds that nearly exclusively encoded genes that were part of a small-molecule biosynthetic gene cluster (BGC). The products of BGCs are diverse and often act as mediators in bacteria-host or bacteria-bacteria interactions (Kadioglu *et al.*, 2008; Donia *et al.*, 2014). On first inspection, the BGC was not assigned to any draft-quality genomes from these samples. Extension of the BGC-associated scaffold revealed that it is part of the genome of the most abundant species in both samples (Actinobacteria phylum). The BGC is located within an 80,484 bp-long region of the genome flanked by mobile elements and has a relatively high GC content (74% versus 68% for the rest of the genome) (Appendix B Figure B.4A) and a distinct tetranucleotide composition (Appendix B Figure B.4B). Its read coverage is consistent with the rest of the genome (Appendix B Figure B.4C). These findings suggest that the BGC was acquired through a relatively recent horizontal gene

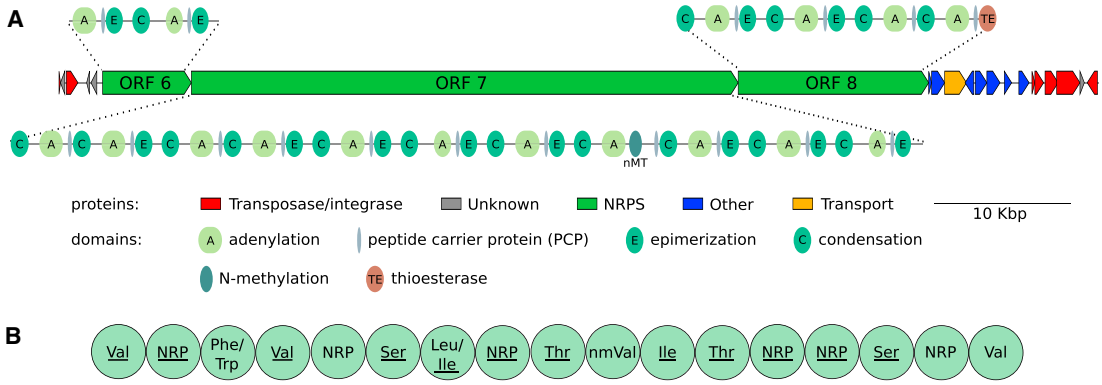


Figure 3.4: Novel non-ribosomal peptide synthesis BGC encoded by the dominant Actinobacteria genome. (A) Predicted protein and biosynthetic domain structure in the 80.5 kbp genomic region comprising the BGC. Open reading frames along the 80.5 kbp genomic region are color coded by function: red, transposase or integrase; gray, unknown function; green, non-ribosomal peptide synthesis (NRPS); blue, other; and yellow, transport-related. Biosynthetic domains of genes involved in NRPS are indicated: A, adenylation domain; E, epimerization domain; C, condensation domain; PCP, peptide carrier protein domain; nMT, N-methylation domain; and TE, thioesterase domain. Each of the 17 adenylation domains encoded by NRP synthesis genes is responsible for the recognition and activation of amino acids that will be incorporated into the peptide product. The cumulative length of these three genes is 69,771 bp. (B) Predicted structure of the peptide product. The amino acid sequence of the predicted peptide was established based on three A domain substrate specificity algorithms incorporated in antiSMASH (Medema *et al.*, 2011; Blin *et al.*, 2013; Weber *et al.*, 2015). Non-ribosomal peptide (NRP) was designated when no consensus was reached. Underlined amino acids are predicted to be in the D configuration, due to the presence of a dedicated epimerization domain in their modules. We cannot distinguish between the possibilities of a circular or linear product. See also Appendix B Figure B.4.

transfer event. Notably, the BGC is predicted to produce a relatively long non-ribosomal peptide (NRP) of 17 amino acids (Figure 3.4). NRPs are synthesized by NRP synthetase enzyme complexes, independent of the ribosome. In the MIBiG database (Medema *et al.*, 2015), the average size of NRPs synthesized by BGCs is only 6 amino acids long (SD ±4.5) (Appendix B Figure B.4D). Because the BGC does not have significant similarity to known BGCs and its predicted product does not resemble any known peptide, elucidation of the function of this BGC product will require heterologous expression—a daunting challenge given the large size of the BGC. Based on the prominence of this Actinobacterium in both dolphin oral microbiotas and the size of this genomic region (3% of the genome), the peptide product is likely to be advantageous to the organism, and may facilitate interactions within the community and/or with the host.

3.3.5 Novel Cas9 diversity

Given the wealth of both novel bacterial and phage genomes, we attempted to link phage sequences to bacterial hosts. We first identified CRISPR-Cas systems and, in doing so, discovered surprising CRISPR-Cas9 diversity (see Appendix B Additional Discussion, Figure B.5, Supplemental Data File B.3, B.4). We identified a total of 67 unique predicted Cas9 proteins (see Methods). Interestingly, two are longer than all Cas9 protein sequences in the RefSeq database (O'Leary *et al.*, 2016) (accessed December 2016) (Figure 3.5A) (DolZOral124_scaffold_19676_2: 1,895 amino acids; DolZOral124_scaffold_953_34: 1,794 amino acids). Neither was assigned to any of the recovered genomes. Another Cas9 contains a large insertion in the RuvC-III domain (DolZOral124_scaffold_26_62, also unassigned). We aligned all three novel Cas9 amino acid sequences against AnaCas9 from *Actinomyces naeslundii* (Figure 3.5B). AnaCas9 was selected as a reference because it

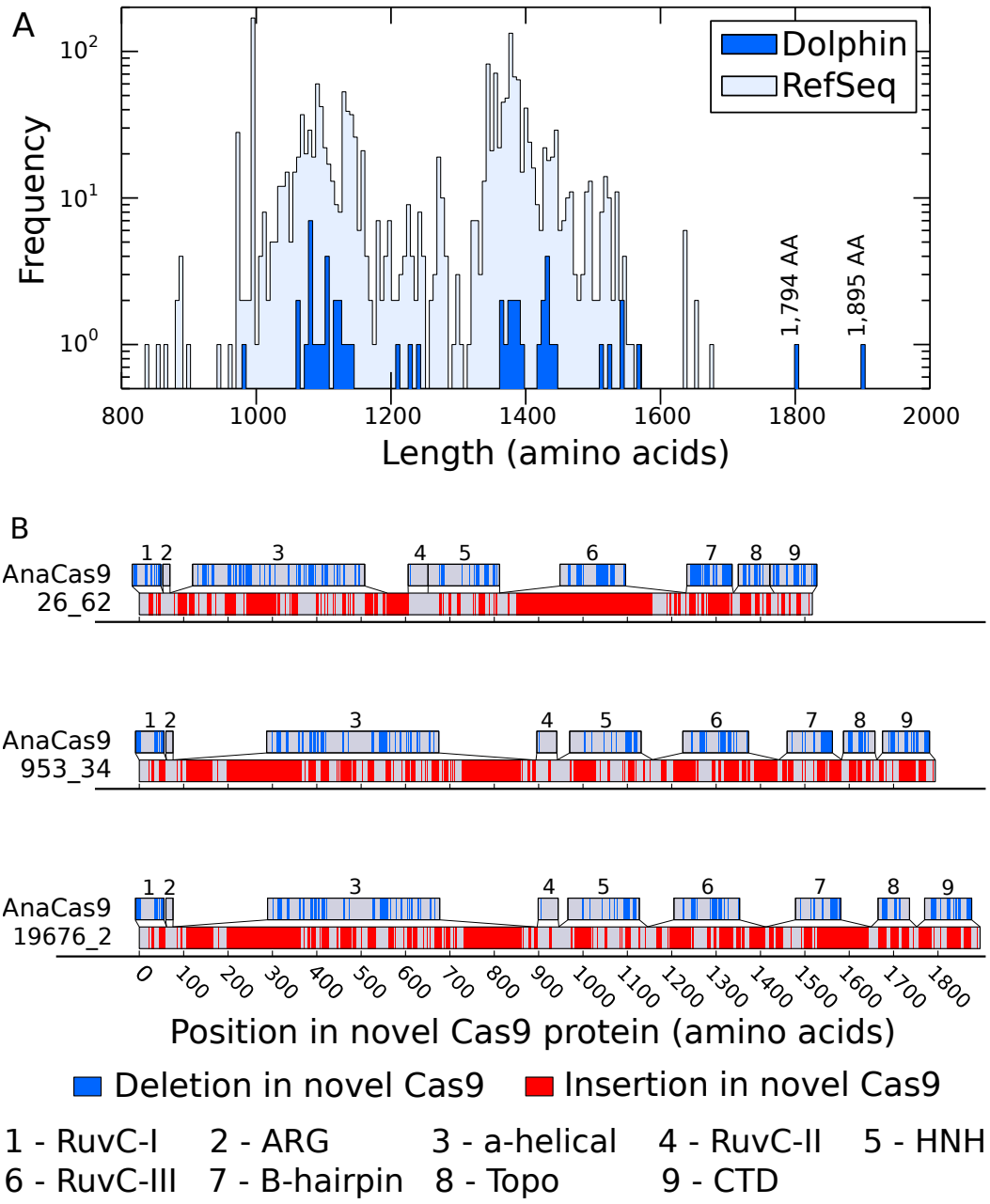

A
Dolphin
RefSeq
Frequency
10^2
10^1
10^0
1,794 AA
1,895 AA
800
1000
1200
1400
1600
1800
2000
Length (amino acids)
B
AnaCas9
26_62
AnaCas9
953_34
AnaCas9
19676_2
1 2
3
4
5
6
7
8
9
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
Position in novel Cas9 protein (amino acids)
Deletion in novel Cas9
Insertion in novel Cas9
1 - RuvC-I
2 - ARG
3 - a-helical
4 - RuvC-II
5 - HNH
6 - RuvC-III
7 - B-hairpin
8 - Topo
9 - CTD

Figure 3.5: Unusual predicted Cas9 protein sequences in the dolphin oral samples. (A) Length distribution of 1,799 complete Cas9 proteins from the RefSeq database (O'Leary *et al.*, 2016) (light blue) and 53 complete Cas9 proteins from the dolphin datasets (dark blue). The longest Cas9 protein in the RefSeq database (O'Leary *et al.*, 2016) is 1,669 amino acids long, whereas the longest Cas9 proteins in the dolphin datasets are 1,794 and 1,895 amino acids long. (B) Insertions and deletions in the three dolphin-associated Cas9 proteins, DolZOral124_953_34, DolZOral124_19676_2, and DolZOral124_26_62, compared to the reference Cas9 protein, AnaCas9. The x axis represents the position with respect to the novel Cas9 protein sequence, in amino acids. The AnaCas9 protein is split into each of its nine functional domains. Regions where both proteins have a residue (although not necessarily the same one) are shown in gray, regions where the dolphin Cas9 has an insertion are shown in red, and regions where the dolphin Cas9 has a deletion are shown in blue. ARG, arginine-rich; CTD, C-terminal domain; HNH, histidine-asparagine-histidine nuclease. See also Appendix B Figures B.5, B.6, Supplemental Data File B.3.

has a resolved crystal structure and it is a type II-C Cas9, as are the three novel predicted proteins in the present study (Appendix B Figure B.6, Supplemental Data File B.1). We found that the largest insertions in the two long Cas9 proteins are concentrated in regions that align with the α-helical, β-hairpin, and RuvC-III domains of AnaCas9. The DolZOral124_scaffold_26_62 Cas9 has a 304 amino acid insertion in the RuvC-III domain when compared with AnaCas9. This insertion has significant homology ($\geq$30% identity over 100% sequence length; e value $<$ 1e-10) to seven other Cas9 proteins in the NCBI non-redundant protein database (https://www.ncbi.nlm.nih.gov/protein/). Attempts to infer the function of the insertion were inconclusive (see Appendix B Additional Discussion) (Soding *et al.*, 2005; Kelley *et al.*, 2015).

3.3.6 Saccharibacteria type II CRISPR-Cas systems and a Saccharibacteria-infecting phage

CRISPR-Cas systems are exceedingly rare within the CPR. In a survey of 354 high-quality draft genomes from the CPR, Burstein *et al.* (2016) found that only five genomes (1.4%) contained a CRISPR-Cas system, and none contained a type II system. We found complete type II CRISPR-Cas systems in two out of five Saccharibacteria (CPR) genomes (see Appendix B Additional Discussion). The Saccharibacteria genomes are not closely related to each other; the ribosomal protein S3 sequences share 67% amino acid identity, which is less than expected for genomes in the same family (Sharon *et al.*, 2015). Although the two complete Saccharibacteria Cas9 proteins are affiliated with a single clade of type II-C Cas9 proteins (Appendix B Figure B.6), neither of the CRISPR-Cas loci encodes a Cas4 protein, as would be expected for a type II-C system.

The ability to identify phages that infect CPR bacteria is important to understanding CPR bacterial evolution and the constraints that they face in their natural settings. However, it is rare to identify phages that infect the CPR (Burstein *et al.*, 2016; Paez-Espino *et al.*, 2016; Paez-Espino *et al.*, 2017). Using CRISPRFinder (Grissa *et al.*, 2007) and Crass (Skennerton *et al.*, 2013), we identified a total of 42 unique spacers from Saccharibacteria CRISPR arrays (see Appendix B Additional Discussion, Supplemental Data File B.4). Of the Saccharibacteria spacers, only one (from the sole CRISPR array associated with DolZOral124_Saccharibacteria_55_12_B) matched a genomic fragment that was identifiable as a phage genome (DolZOral124_Phage_53_65). The phage and Saccharibacteria genomes were originally binned together based on tetranucleotide frequency. Convergence of tetranucleotide frequency is suggestive of a history of co-evolution between a phage and its bacterial host (Pride *et al.*, 2006). The

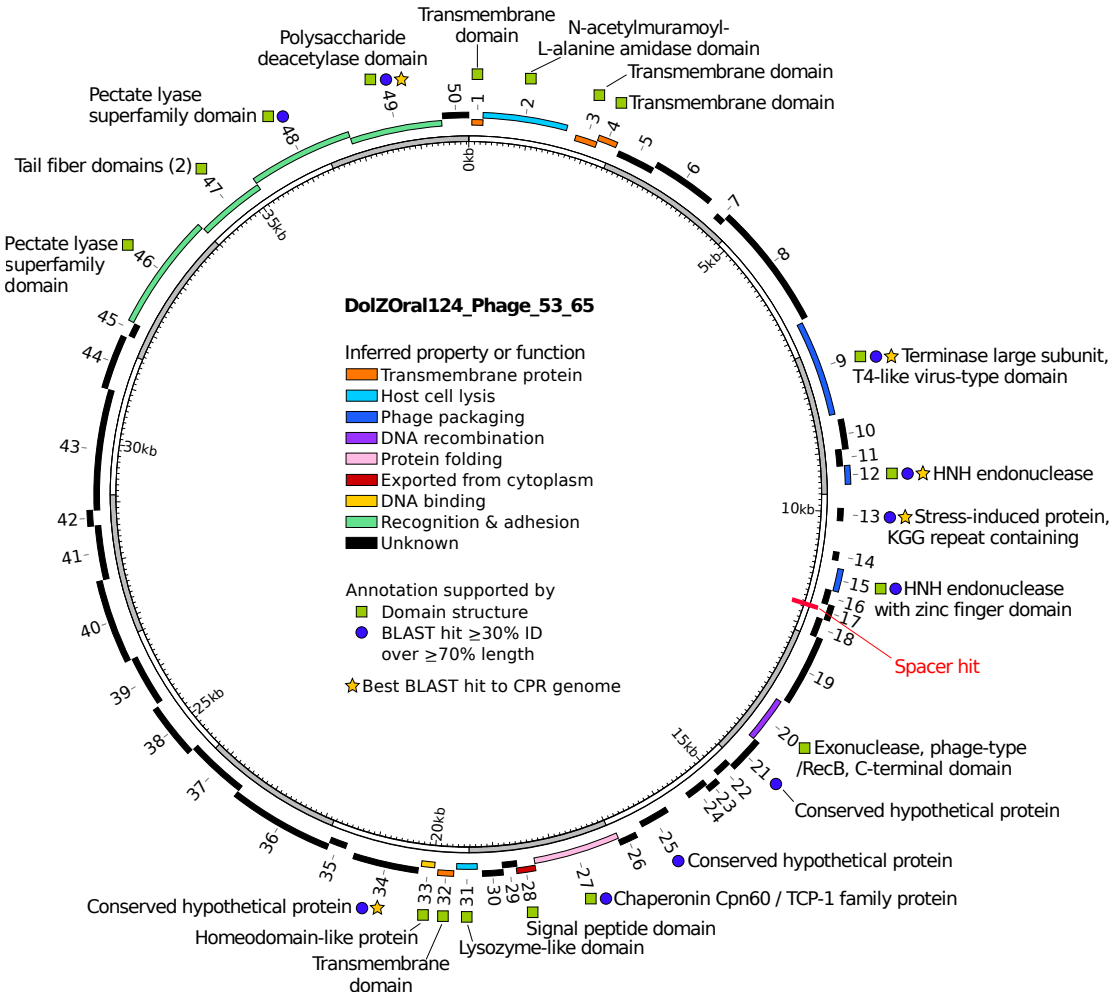


Figure 3.6: Genome organization of the Saccharibacteria phage. The inner ring represents the phage genome (total length 38.8 kbp; positions are indicated inside the ring). The outer ring shows the position of open reading frames (ORFs) around the genome, numbered from 1 to 50. ORFs are color coded based on inferred property or function. For those ORFs that have an inferred property or function, green squares denote annotations supported by domain structure, blue circles denote annotations supported by a BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009) hit of ≥30% identity over ≥70% length of the ORF with an e value of 1e-05, and yellow stars denote annotations whose top BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009) hit was to a genome in the CPR. The position of the spacer match from DolZOral124_Saccharibacteria_55_12_B is represented by a red slash perpendicular to the phage genome. See also Supplemental Data File B.3, B.4, B.5.

phage genome is circular and 38,841 bp long, with a GC content of 52%. No read pairs mapped to both the phage and Saccharibacteria genomes. Consequently, we infer that the phage was not integrated into the host genome at the time of sampling. The phage genome contains 50 predicted open reading frames (ORFs) and no tRNAs (Figure 3.6, Supplemental Data File B.5). Predicted functions of these ORFs include host cell lysis, phage packaging, and DNA recombination.

3.4 Discussion

We used genome-resolved metagenomics to study the microbial communities of two dolphin oral samples in order to explore the unusual evolutionary and functional diversity predicted by a previous 16S rRNA gene-based survey (Bik *et al.*, 2016). Of note, we detected and characterized novel lineages distantly related to and reproducibly unaffiliated with known phyla. We propose that they represent phylum-level lineages for which we put forth the names *Candidatus* Delphibacteria and *Candidatus* Fertabacteria. The Delphibacteria representative characterized here is predicted to denitrify, which is a process that may impact dolphin health and physiology. For example, in humans, denitrification by oral bacteria can affect oral and gastric blood flow, signaling in bacteria-bacteria and bacteria-host interactions, and mucus thickness in the stomach (Lundberg *et al.*, 2008; Schreiber *et al.*, 2010). It is unclear whether bacteria in the Delphibacteria candidate phylum remain uncultured due to intrinsic biological factors or due to the absence of a systematic effort to culture and identify them using traditional methods. Regardless, our genomic analysis may provide insights into the conditions required for successful cultivation of these and closely related bacteria, especially with regard to oxygen conditions and potential energy and carbon sources.

In addition, we recovered genomes from candidate phyla whose members are

seldom associated with animals. These genomes will be a valuable resource for future comparative studies aimed at understanding how such bacteria adapt to a mammalian environment. Interestingly, we detected members of the Saccharibacteria phylum. Members of this phylum have been associated with human oral disease (Brinig *et al.*, 2003). At least one Saccharibacteria strain, an obligate endobiont of an Actinobacterium, has the ability to modify human immune responses in vitro (He *et al.*, 2015). This may aid Saccharibacteria and potentially also their microbial host/s in avoiding clearance by the human immune system. It remains unclear whether oral Saccharibacteria are detrimental to dolphin health, and whether they may be associated with Actinobacteria in this setting.

An interesting aspect of our community composition analysis was that the highest-coverage genome was from an Actinobacterium that went virtually undetected in the previously published 16S rRNA amplicon survey. The underlying reasons for this discrepancy remain unknown. This finding highlights the fact that even among relatively well characterized phyla there exist unexplored branches represented by organisms with unusual predicted properties that are inherently distinct from the bacteria we are accustomed to studying.

By exploring the microbiology of the dolphin mouth, we uncovered an unexpected diversity of CRISPR systems that are related to those used in recently developed CRISPR-Cas9-based genome editing methods (Jinek *et al.*, 2012). At this time, the potential technological value of divergent proteins from class 2 CRISPR-Cas systems (those with single-subunit CRISPR RNA (crRNA)-effector molecules) remains relatively unexplored and so the significance of the findings remains unclear. However, the findings further establish the potential importance of genes discovered in the genomes of bacteria newly characterized by cultivation-independent metagenomics (Burstein *et al.*, 2017).

Previously unexplored environments, such as the marine mammal oral cavity, contain a wealth of phylogenetic and functional novelty of which we have only just scratched the surface. Populating the tree of life with genomes from poorly understood or previously unsampled microbial lineages from diverse environments, and characterizing the phages that infect them, is an important step toward creating a comprehensive picture of the evolutionary history of life on Earth.

3.5 Methods

3.5.1 Experimental model and subject details

Oral samples were obtained from the left gingival sulcus of dolphins managed by the U.S. Navy Marine Mammal Program (MMP) in San Diego, California. The swabbing protocol adhered to the guidelines described in the CRC handbook of Marine Mammal Medicine. From the 22 dolphin oral specimens included in Bik *et al.* (2016), two were selected for metagenomic analysis. Sample DolJOral78 originated from a healthy 5-year-old male and sample DolZOral124 originated from a healthy 29-year-old lactating female. The MMP is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) International and adheres to the national standards of the United States Public Health Service Policy on the Humane Care and Use of Laboratory Animals and the Animal Welfare Act. As required by the U.S. Department of Defense, the MMP's animal care and use program is routinely reviewed by an Institutional Animal Care and Use Committee (IACUC) and by the U.S. Navy Bureau of Medicine and Surgery. The animal use and care protocol for MMP dolphins in support of this study was approved by the MMP's IACUC and the Navy's Bureau of Medicine and Surgery (IACUC #92-2010, BUMED NRD-681).

To compare the proportion of CRISPR-Cas types across oral environments from different mammals (see Appendix B Additional Discussion and Figure B.5), we additionally analyzed data from two humans and a harbor seal. Saliva samples were obtained from two healthy, pregnant women who presented at Lucille Packard Children's Hospital in Stanford, California. These samples were collected from subjects who signed a written consent, and following procedures described in an IRB protocol (21956) that was approved by an Administrative Panel for the Protection of Human Subjects at Stanford University. Swab samples from the left gingival sulcus of a harbor seal were obtained from an animal originally admitted to the Marine Mammal Center in Sausalito, California, USA with pneumonia, malnutrition, and a left hind flipper injury. The animal was treated with Clavamox from July 5-18, 2012, recovered, and was released back into the wild. The sample used here was the last collected prior to release at a time of health, and was taken on August 22, 2012 during a routine clinical exam.

3.5.2 DNA extraction, sequencing, and quality filtering

We used the same DNA preparations from MMP dolphin gingival sulcus samples as used by Bik *et al.* (2016). These samples were processed using the QIAamp Mini Kit (QIAGEN, Valencia, CA). Library preparation and shotgun sequencing were performed by the Keck Center at the University of Illinois at Urbana-Champaign. Briefly, short read Illumina libraries (2 x 250 bp) were constructed using the Kapa Hyper Prep Kit (Kapa Biosystems, Wilmington, MA) and the two libraries were sequenced on a single Illumina HiSeq 2500 lane. The average gDNA fragment length was 580 bp (range: 350-800 bp). 93,369,641 raw read-pairs for sample DolJOral78 and 76,479,271 raw read-pairs for sample DolZOral124 were quality-filtered using Sickle (Joshi & Fass, 2011) with the '-q 28' flag specified

to increase the minimum threshold of acceptable quality scores. Adapters were removed and anomalously short reads (<100 bp) were discarded in one step using SeqPrep (https://github.com/jstjohn/seqprep). Reads that mapped to the dolphin genome (turTru2) (Linblad-Toh *et al.*, 2011) were considered to be host contamination and were removed from the dataset using bowtie2 version 2.2.4 (Langmead *et al.*, 2012). Six percent and two percent of reads from the DolJOral78 and DolZOral124 samples mapped to the dolphin genome, respectively. After host sequence removal, 58,250,929 and 82,272,429 read pairs were available for metagenome assembly.

3.5.3 Metagenome assembly, annotation, and binning

Assembly of read-pairs from each sample was performed using IDBA-UD version 1.1.1 (Peng *et al.*, 2012). IDBA-UD was patched to increase the maximum permissible length of paired end reads from 128 bp to 250 bp (via the kMaxShortSequence constant), thereby allowing for the use of 250 bp reads with the '-r' option. The DolJOral78 and DolZOral124 reads were assembled into 306,641 and 149,038 scaffolds greater than one kb in length, respectively. Genes were predicted using the metagenome implementation of Prodigal version 2.6.0 (Hyatt *et al.*, 2010). USEARCH version 7.0.1 (Edgar *et al.*, 2010) was used to compare protein sequences from all predicted ORFs against the UniRef 90 (Suzek *et al.*, 2015) and KEGG (Kanehisa *et al.*, 2000; Kanehisa *et al.*, 2016; Kanehisa *et al.*, 2017) databases, as well as an in-house database of predicted ORFs from candidate phyla genomes. 16S and 23S rRNA genes were predicted using in-house HMM-based rRNA gene identification scripts (Brown *et al.*, 2015) and tRNA genes were predicted using tRNAs can version 1.23 (Lowe *et al.*, 1997).

A bin is a set of scaffolds that share similar genomic features, and is typi-

cally representative of a genome. Binning of scaffolds was performed using ggKbase, based on %GC content, read coverage, and inferred taxonomy of scaffolds by best-hit annotations of predicted proteins. Bins were refined on the basis of tetranucleotide frequency using emergent self-organizing maps (ESOM). To do so, tetranucleotide frequency was calculated for all scaffolds greater than or equal to five kb in length over window sizes of five kb (as described in Dick *et al.* (2009)), and ESOMs were computed and visualized with the Databionics ESOM Tools software (Ultsch *et al.*, 2005).

3.5.4 Identification of phage scaffolds

To identify candidate phage sequences, we required that scaffolds have two or more gene annotations containing virus-specific keywords from the list: 'capsid, phage, terminase, base plate, baseplate, prohead, virion, virus, viral, tape measure, tapemeasure neck, tail, head, bacteriophage, prophage, portal, DNA packaging, T4, p22, holin' (excepting annotations with following terms: 'abortive, shock, forkhead, T7 exclusion, macrophage, hth-like transcriptional regulator, peptidase family t4, lamin a/c globular'). Candidate phage scaffolds were eliminated if any gene annotations contained prokaryote-specific terms from the list 'tRNA synthetase, tRNA synthase, ribosomal protein, preprotein translocase, DNA gyrase subunit A.' This yielded 322 and 708 candidate sequences for DolJOral78 and DolZOral124, respectively. To minimize the occurrence of false positives, we additionally required that at least one spacer from either dolphin oral metagenome match the candidate phage scaffold. Finally, we manually removed scaffolds which likely encoded prophage inserted into a bacterial genome (one scaffold was removed from each sample set).

3.5.5 Refining selected scaffolds

The PRICE assembly algorithm (Ruby *et al.*, 2013) was used to extend scaffolds of interest, such as those containing unbinned 16S rRNA genes of interest (in an attempt to associate them with binned scaffolds), the DolZOral124_Bacteria_38_8 genome, and the Saccharibacteria phage. For selected sets of scaffolds, such as those binned into one of the genomes from the two novel, phylum-level lineages, we attempted to resolve assembly errors using ra2 (Brown *et al.*, 2015). We visually confirmed that the scaffolds containing genes used for phylogenetic analysis of DolZOral124_Bacteria_64_63, DolJOral78_Bacteria_63_78, and DolZOral124_Bacteria_38_8 contained no assembly errors. This was done by mapping reads against scaffolds and using mapped.py (part of the ra2 suite) (Brown *et al.*, 2015) to filter out mate pairs where there was more than one mismatch to the assembled scaffold across both reads combined, and then confirming that there were no regions in the scaffolds whose assembly was not supported by the stringently mapped reads. Ra2 (Brown *et al.*, 2015) was also implemented on all scaffolds containing a *cas* gene prior to analysis, although deposited *cas*-containing scaffolds are the original versions assembled by IDBA-UD (Peng *et al.*, 2012).

3.5.6 Bin completeness and characterization

From sample DolJOral78, we recovered 34 near complete bacterial genomes ($\geq$80% complete), 16 draft-quality partial bacterial genomes ($\geq$50% complete), and 45 other bins. From DolZOral124, we recovered 31 near complete bacterial genomes, 1 complete (circular) phage genome, 25 draft-quality partial bacterial genomes, and 88 other bins. Bins that did not qualify as draft-quality genomes had $\geq$10 and $<$25 bacterial single copy genes present and/or, in some cases, contained multiple genomes from closely related bacteria. We calculated genome relative

abundance as follows: For every genome bin (plus an artificial bin consisting of all unbinned scaffolds) we calculated the cumulative length of all scaffolds in the bin (i.e., genome length), as well as the average coverage of all the scaffolds in the bin (i.e., genome coverage). To correct for genome size bias, we standardized genome coverage by genome length such that:

$$standardized\ binA\ coverage = \frac{fraction\ of\ reads\ that\ map\ to\ binA}{length\ binA}$$

Where:

$$fraction\ of\ reads\ that\ map\ to\ binA = \frac{\#\ reads\ that\ map\ to\ binA}{\#\ reads\ that\ map\ to\ the\ metagenome}$$

After performing this calculation for every bin, we calculated relative abundance as follows:

$$binA\ relative\ abundance = \frac{standardized\ binA\ coverage}{total\ standardized\ community\ composition} \times 100$$

Where:

$$total\ standardized\ community\ coverage =$$

$$standardized\ binA\ coverage\ + ... +\ standardized\ binN\ coverage$$

and N was the total number of bins recovered (including the artificial 'unbinned' scaffold 'bin').

Taxonomic assignment of 16S rRNA genes was performed using the RDP classifier with 16S rRNA gene training set 16 (Wang *et al.*, 2007). For 16S rRNA genes that could not be classified by RDP classifier, we attempted to identify them by a) determining whether the 16S rRNA gene was binned with a genome of

known taxonomic identity, or b) by using BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009) with OTUs from the previous 16S rRNA gene survey (Bik *et al.*, 2016) and determining whether close relatives (≥95% identity) had been detected and identified.

3.5.7 Phylogenetic placement of genomes

The concatenated ribosomal protein tree was created using a set of 15 ribosomal proteins (L2p, L3p, L4p, L5p, L14p, L15p, L16p, L18p, L22p, L24p, S3p, S8p, S10p, S17p, and S19p in bacteria and the homologous archaeal proteins L8e, L3e, L1e, L11e, L23e, L23Ae, L10e, L5e, L17e, L26e, S3e, S15Ae, S20e, S11e, and S15e) (Hug *et al.*, 2013). Ribosomal protein L6p was not included in the phylogenetic reconstruction because, later on, we ascertained that the alignment did not fit the same evolutionary model as the other 15 ribosomal proteins. Reference sets were obtained from PATRIC (Wattam *et al.*, 2014), ggKbase, and NCBI databases. Ribosomal protein sets from the dolphin samples were obtained from all genomes for which at least eight of the ribosomal proteins were present (with the exception of the DolJOral78_Delongbacteria_30_2 genome, which had seven ribosomal proteins present), and sets from candidate phyla genomes were curated and confirmed to have no assembly errors prior to analysis. Each individual protein set was created and refined using MUSCLE (Edgar, 2004) and then manually curated. Manual curation consisted of re-aligning misaligned C- or N- termini and removing protein sequences containing suspected frameshift mutations or assembly errors. Columns containing at least 5% gaps were removed using Geneious version 7.1.9 (Kearse *et al.*, 2012). Evolutionary model selection for each of the ribosomal protein sets was performed using ProtTest3 (Darriba *et al.*, 2011; Guindon & Gascuel, 2003). Protein sets were concatenated using Geneious version 7.1.9

(Kearse *et al.*, 2012). A phylogenetic tree was created using RAxML (Stamatakis *et al.*, 2014) under the LG+G (PROTGAMMALG) evolutionary model with 100 bootstrap replicates. The tree was visualized using iTOL (Letunic & Bork, 2011) and 'beautified' using Inkscape (https://inkscape.org/en/).

Phylogenetic analysis of 16S rRNA genes was primarily based on sequences in the SILVA NR Ref 99 database (Pruesse *et al.*, 2007; Quast *et al.*, 2013; Yilmaz *et al.*, 2014). For the Latescibacteria-Delphibacteria-Eisenbacteria phylogeny, we obtained all 16S rRNA genes present in what is currently labeled as the Latescibacteria phylum in the SILVA NR Ref 99 database (Pruesse *et al.*, 2007; Quast *et al.*, 2013; Yilmaz *et al.*, 2014), sequences from all genome assemblies from the Latescibacteria, Delphibacteria, and Eisenbacteria phyla with a 16S rRNA gene, and the top 20 BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009) hits from the NCBI non-redundant nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/) to the dolphin-associated sequence. For the Peregrinibacteria-Fertabacteria phylogeny, we used all 16S rRNA genes present in what is currently labeled as the Peregrinibacteria phylum in the SILVA NR Ref 99 database (Pruesse *et al.*, 2007; Quast *et al.*, 2013; Yilmaz *et al.*, 2014) and the PER 16S rRNA genes used by [1], which are approximately representative of each genus for which genomes have been sequenced. Sequences were aligned using the SINA aligner v1.2.11 (Pruesse *et al.*, 2012)) with the SILVA SSU Ref NR 99 database release 128 (Quast *et al.*, 2013; Yilmaz *et al.*, 2014; Pruesse *et al.*, 2012) as a reference. Columns containing at least 3% gaps were removed using Geneious version 7.1.9 (Kearse *et al.*, 2012) and a phylogenetic tree was run under the GTR+G (PROTGAMMAGTR) evolutionary model in RAxML (Stamatakis *et al.*, 2014) with 1000 bootstrap replicates. Estimation of the percent identity between different clades within 16S rRNA trees was based on the methods proposed by Yarza *et al.* (2014). We

used the 16S rRNA gene alignment created by SINA (before stripping columns) and removed insertions ≥10 bp long. Insertions were defined as any sequence shared by <5% of all aligned sequences. Sequences were sorted by length and clustered with a 75% identity threshold using USEARCH version 9.2.64 (Edgar, 2010) (-cluster_smallmem -query_cov 0.50 -target_cov 0.50 -id 0.75). Maximum likelihood trees overlayed with USEARCH clustering results were visualized using iTOL (Letunic & Bork, 2011).

3.5.8 Metabolic reconstruction of DolZOral124_Bacteria-_64_63 (*Candidatus* Delphibacteria)

Metabolic pathways were identified using KAAS (Moriya *et al.*, 2007). Amino acid sequences were queried against the KAAS database using the bi-directional best hits mode, using the following organism IDs to construct a reference set: eco, son, cje, gme, sme, rsp, mtu, bsu, cac, ctr, bfr, fjo, emi, cau, tma, mja, afu, pho, tac, ape, sso, pai, tne, tko, pab, pfu, mma, aae, dra, det, cte, pma, syw, fnu, fsu, cao, sru, lil, fra, and gau. Annotations for the genome from KAAS or the ggKbase pipeline were confirmed using a combination of BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009) searches against the NCBI non-redundant protein database (https://www.ncbi.nlm.nih.gov/protein/), pHMMER (Finn *et al.*, 2015), and/or InterProScan (Jones *et al.*, 2014). Searches for specific proteins of interest that were not identified by KAAS Moriya *et al.*, 2007) or our annotation pipeline (for example, proteins we wished to confirm as absent from the genome) were conducted by either obtaining the corresponding hidden Markov Models (HMMs) profile from the Pfam database (Finn *et al.*, 2016) and searching for it using the HMMER suite version 3.1b2 (Eddy, 2011), or by obtaining the corresponding protein sequence from the NCBI database and querying it against

our genome with BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009), and then confirming the identity of hits as described above. Potential ABC transporters were identified using an HMM search for the ATP-binding domain of ABC transporters (PF00005). Matches were then annotated using pHMMER (Finn *et al.*, 2015) and by performing BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009) searches of candidates against the ABCdb CleanDB (Fichant *et al.*, 2006), which is a specialized ABC transporter database containing only manually curated ABC transporter entries. The cell metabolism diagram was created using Inkscape (https://inkscape.org/en/).

3.5.9 Biosynthetic gene cluster structural predictions

The structure of the dolphin Actinobacteria BGC was characterized using antiSMASH version 3.0 (Medema *et al.*, 2011; Blin *et al.*, 2013; Weber *et al.*, 2015). Figure 3.4 was based on output from antiSMASH, which was modified using Inkscape (https://inkscape.org/en/).

3.5.10 Identification and classification of CRISPR-Cas systems and predicted Cas9 proteins

To search for Cas9 protein sequences, we performed an HMM search with HMMER suite version 3.1b2 (Eddy, 2011), using the Cas9 HMMs from Makarova *et al.* (2015) and a threshold e-value of 1e-10. To determine the number of unique proteins present in the two datasets combined, we used cd-hit (Li *et al.*, 2006; Fu *et al.*, 2012) to cluster together similar protein sequences $\geq$800 amino acids, using cutoffs of $\geq$90% identity over a maximum of 80% length difference. This cutoff length was selected since the shortest known functional Cas9 protein is 950 amino acids long (Shmakov *et al.*, 2015). To compare the dolphin Cas9 protein sequences against

previously sequenced Cas9 proteins, we downloaded all Cas9 proteins from the RefSeq database (O'Leary *et al.*, 2016) and confirmed whether they were genuine Cas9 proteins using the same HMM search pipeline. Only confirmed Cas9 proteins were used in downstream analysis. We then aligned all dolphin metagenome Cas9 proteins, Cas9 proteins classified into subtypes by Makarova *et al.* (2015), and the AnaCas9 protein using MUSCLE (Edgar, 2004). This alignment was used to determine the position of insertion sequences in the DolZOral124_953_34, DolZOral124_19676_2, and DolZOral124_26_62 proteins relative to AnaCas9. To create a Cas9 phylogeny, we removed all columns containing at least 5% gaps and used ProtTest3 (Guidon *et al.*, 2003; Darriba *et al.*, 2011) to determine the best fitting evolutionary model. A phylogenetic tree was constructed using RAxML (Stamatakis, 2014), applying the VT + G + F (PROTGAMMAVTF) evolutionary model. The tree was visualized using iTOL (Letunic & Bork, 2011). To evaluate the distribution of Cas9 protein lengths, we aligned all RefSeq and dolphin metagenome Cas9 proteins with the well-characterized AnaCas9 and SpyCas9 proteins using MUSCLE (Edgar, 2004), and removed partial sequences that did not span the domains present in AnaCas9 and SpyCas9. We then analyzed the length distribution of the remaining protein sequences.

To compare the proportion of CRISPR-Cas systems present in the dolphin, harbor seal, and human microbiomes (see Appendix B Additional Discussion and Figure B.5), the criteria used for identifying a CRISPR-Cas system required that a scaffold must contain a *cas* operon and a CRISPR array. Valid *cas* operons were considered as those that had at least one signature *cas* gene (*cas3*, *cas9*, *cas10*, *csf1*, or *cpf1*) and were composed of two or more *cas* genes. Operons were defined as sets of *cas* genes separated by four or fewer open reading frames of each other. To search for Cas proteins, we used the HMMER suite version 3.1b2 (Eddy, 2011)

to search for Cas protein HMMs constructed based on alignments from (Makarova *et al.*, 2015). We applied a cutoff e-value of 0.01 in order to identify Cas proteins with low sequence similarity to previously identified Cas proteins. CRISPR arrays were identified from assembled scaffolds using CRISPRFinder (Grissa *et al.*, 2007) and false positives were removed manually. These results were used to calculate the proportion of CRISPR-Cas types in mammalian oral microbiomes.

3.5.11 Identification and analysis of scaffolds targeted by CPR spacers

We identified spacers in assembled scaffolds using CRISPRFinder (Grissa *et al.*, 2007) and CRASS (Skennerton *et al.*, 2013). For the CRASS spacers, we identified which arrays matched Saccharibacteria CRISPR-Cas systems based on their having an identical direct repeat sequence to those identified by CRISPRFinder. We searched spacers (using BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009)) against both full metagenomic assemblies, the NCBI non-redundant nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/), and the NCBI virus database (Brister *et al.*, 2015) to identify scaffolds targeted by spacers. Spacers were required to have a match of $\geq$95% sequence identity over 100% of the spacer or 100% identity over $\geq$95% of the spacer to qualify as a match. The spacer sequence from DolZOral124_Saccharibacteria_55_12_B that matched the Saccharibacteria phage genome was 30 bp long (CGGCCTGAAAAGCTCGAGCCGGCCATTCAA) and had a match of 96.67% identity over 100% of the spacer. The Saccharibacteria phage genome was annotated using BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2009) searches against the NCBI non-redundant protein database (https://www.ncbi.nlm.nih.gov/protein/) and by submitting protein sequences to pHMMER (Finn *et al.*, 2015) and InterProScan (Jones *et al.*,

2014). Figure 3.6 was created using Circos (Krzywinski *et al.*, 2009) and Inkscape (https://inkscape.org/en/).

3.5.12 Data availability

Raw sequence reads, genomes, and assembled scaffolds from the dolphin oral datasets are available through NCBI BioProject database: PRJNA174530 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA174530/) with BioSample identifiers SAMN01162460 and SAMN01162508 for DolJOral78 and DolZOral124, respectively. Scaffolds and genome bins can be viewed through the online database ggKbase at http://ggkbase.berkeley.edu/DOLJORAL78/organisms and http://ggkbase.berkeley.edu/DOLZORAL124/ organisms. Raw sequence reads from the harbor seal oral dataset are available through NCBI under the BioProject identifier PRJNA412531 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA412531) with BioSample identifier SAMN07716580. Sequence data from the human oral metagenomes has been deposited under the BioProject identifier PRJNA288562 (https://www.ncbi. nlm.nih.gov/bioproject/PRJNA288562) with BioSample identifiers SAMN03845088, SAMN03845091, SAMN03845094, SAMN03845097, SAMN03845100, SAMN03845103, SAMN03845106, SAMN03845108, SAMN03845111, SAMN03845111, SAMN03845114, and SAMN03845224 for human A and SAMN03845448, SAMN03845451, SAMN03845454, SAMN03845458, SAMN03845460, SAMN03845463, SAMN03845466, SAMN03845469, SAMN03845472, SAMN03845475, and SAMN03845503 for human B.

Chapter 4

Previously uncharacterized rectangular microbial units in the marine mammal oral cavity

4.1 Abstract

Much remains to be explored regarding the diversity of morphologies that have evolved in uncharacterized environments, such as that represented by marine mammals. We discovered rectangular, likely bacterial, microbial units in the mouths of bottlenose dolphins (*Tursiops truncatus*), the identity of which remains uncertain. Cryogenic electron transmission microscopy revealed that each rectangular microbial unit is encapsulated in an S-layer-like structure and is composed of numerous parallel, seemingly paired, membrane-bound segments. These segments are likely individual cells, whereas each rectangular unit is likely an aggregate, similar to what is seen in bacteria from the genus *Simonsiella*. Pili-like appendages with an usual architecture project from segments. These consist of

stalks of hair-like structures that splayed out at the tips. Based on single-cell genomic analysis, we postulate that these microorganisms are bacteria from either the Saccharibacteria (TM7), Bacteroidetes, or (Epsilon-) Proteobacteria phyla. These observations highlight the diversity of novel microbial lifestyles that await discovery and characterization using microscopy and other traditional microbiological tools.

4.2 Introduction

The earliest descriptions of the microbial world centered around the morphology and motility patterns of 'animalcules' (Leeuwenhoek, 1677). Over 340 years have passed since Leeuwenhoek's revolutionary discovery, during which time a vast diversity of microbial forms have been unearthed. These range from peculiar star-shaped bacteria in the *Stella* genus (Nitkin *et al.*, 1966; Vasilyeva, 1985) to the striking multicellular fruiting bodies characteristic of *Myxobacteria* like *Stigmatella aurantiaca* (Voelz & Reichenbach, 1969; Dworkin, 2000). Morphology is a biologically important feature which is molded by selective pressures exerted as a result of lifestyle, and therefore offers an appealing route by which to glean insights into the diversity of novel microbial lifestyles that exist. For example, morphology plays an important role in nutrient acquisition, cell division, cell energetics, and interactions with other cells, all of which are strong determinants of survival (reviewed in Young, 2007). The diversity of microbial morphologies and lifestyles that exist in previously uncharted branches of the tree of life remains to be seen.

Genomics serves as a powerful lens through which to describe the microbial world. In recent years, metagenomic and single cell genomic analyses have substantially increased the number of known microbial phylum-level lineages. For

example, in the bacterial domain such techniques have increased this number by a factor of nearly four (Rinke *et al.*, 2013; Brown *et al.*, 2015; Castelle *et al.*, 2015; Anantharam *et al.*, 2016, and others). Novel phylogenetic diversity is correlated with novel functional potential (Wu *et al.*, 2009), and correspondingly genomes obtained from novel organisms have allowed for the discovery of new functional systems, types of protein variants, and lifestyles (Wrighton *et al.*, 2012; Brown *et al.*, 2015; Burstein *et al.*, 2017; Donia *et al.*, 2014; Dudek *et al.*, 2017). Such approaches, however, are limited to investigating proteins and systems that have homology or similar properties to those from well characterized organisms; they cannot be used to predict phenotypes and functions that are truly novel and/or whose genetic basis is unknown. Given the recalcitrance of the majority of the microbial diversity present on Earth to culture in the laboratory (Hug *et al.*, 2016), microscopy offers an appealing route by which to study novel morphological and functional properties of uncultured lineages.

Previous studies using 16S rRNA gene amplicon sequencing and genome-resolved metagenomics found that the mouths of bottlenose dolphins (*Tursiops truncatus*) host a rich diversity of novel microbial diversity (Bik *et al.*, 2016; Dudek *et al.*, 2017). This includes representatives from poorly characterized bacterial and archaeal phyla, some of which lack cultured representatives altogether and whose biology is therefore particularly poorly understood. To learn more about the morphology and lifestyle of novel microbial lineages, we surveyed dolphin oral microbial communities using microscopy.

In this study, we describe properties of unusual rectangular microbial units found in dolphin oral samples. These units are likely chains of individual cells and will therefore be referred to 'cell-like units' from here on. While the internal organization of segments is similar to that of the *Simonsiella* genus, there

are striking morphological differences. Single cell genomic analysis is suggestive of the rectangular cell-like units being bacteria from either the Saccharibacteria (TM7), Bacteroidetes, or Epsilonproteobacteria groups. Our findings raise questions about the function, genomic basis, and evolutionary history of unusual morphological features of rectangular cell-like units from the dolphin mouth.

4.3 Results

4.3.1 Light microscopy insights into the structure and spatial organization of rectangular cell-like units

Oral swab samples were collected from the mouths of eight bottlenose dolphins (*Tursiops truncatus*) under the purview of the Marine Mammal Program in San Diego Bay, California, USA (see Methods). Phase contrast microscopy images revealed the presence of rectangular cell-like units in dolphin oral samples (Figure 4.1). DAPI staining indicated that rectangular cell-like units contained multiple parallel, seemingly paired bands of what is likely DNA. Some cells have dark (dense) spots which may be storage granules. There appear to be two morphotypes of rectangular cell-like units, which have either 'short' or 'long' length DAPI-stained bands (highlighted in Figure 4.1A,E). The different morphotypes may represent a) different taxonomic groups (ex: strains or species), b) cells in different stages of development, or c) cells that have altered their shape in response to environmental conditions.

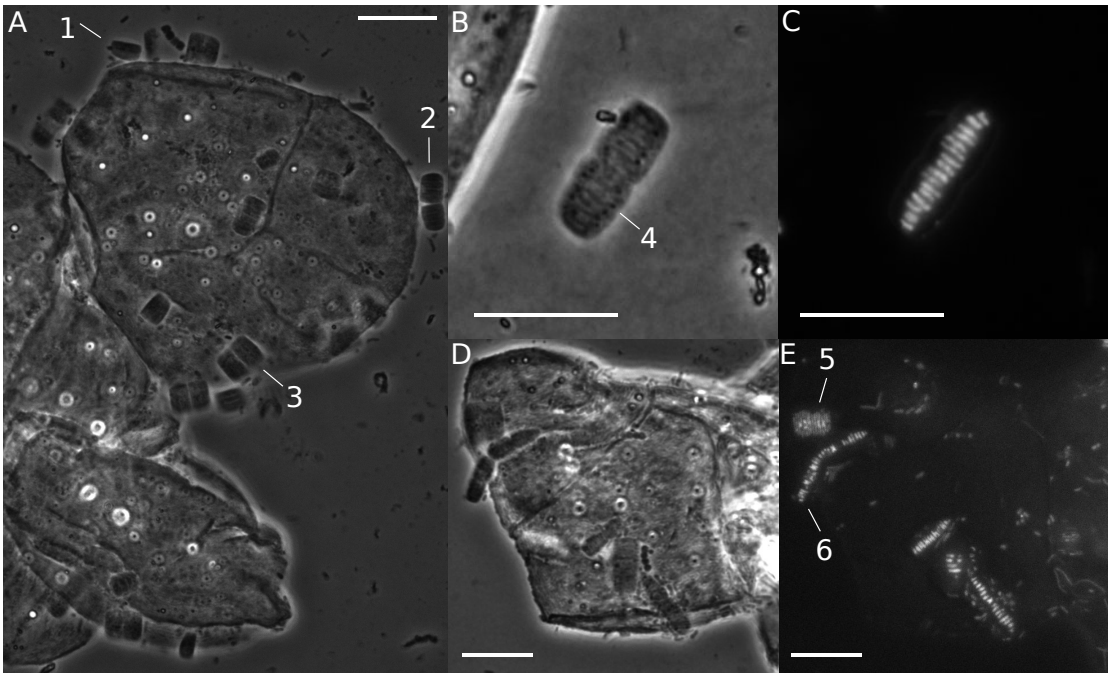


Figure 4.1: Light microscopy images of rectangular cell-like units. (A,B,D) Phase contrast microscopy, (C,E) Fluorescence microscopy, cells stained with DAPI, emission/excitation spectra of 340/488 nm. (A) Rectangular cell-like units on the surface of dolphin epithelial cell. Cell-like units are rectangular prisms; for side-view see arrow 1. Two morphotypes are present; see arrows 2 (short morphotype) and 3 (long morphotype) (B) Two rectangular cell-like units appear to be separating. Dark (dense) spots run in lines perpendicular to DAPI-stained bands; see arrow 4. (C) The same cell-like units and field of view as in (B). DAPI-stained bands are present. In this case the bands appear to be in pairs, with four pairs per rectangular cell-like unit, although other rectangular cell-like units had different numbers of bands. (D, E) Same field of view. Two morphotypes are present; see arrow 5 (long morphotype) and 6 (short morphotype). Scale bars: 10 um.

4.3.2 Cryogenic electron transmission microscopy insights into the internal ultrastructure of rectangular cell-like units

To learn more about the structure of the rectangular cell-like units, we used cryogenic electron transmission microscopy (cryo-TEM). Cryo-TEM images revealed that rectangular cell-like units consist of seemingly paired segments organized in parallel within each rectangular cell-like unit (Figure 4.2). These segments were oriented in the same manner as the DAPI-stained bands that were seen in light microscopy images. Groups of segments sometimes appeared to be separating from one another. As in the light microscopy images, dark (dense) structures are visible in the rectangular cell-like units. The dark spots are essentially spheroidal. The dimensions of two dark spots from one rectangular cell-like unit were approximately 192 nm x 200 nm x 192 nm (V $\approx$ 3.12 x 107 nm3) and 215 nm x 220 nm x 220 nm (V $\approx$ 4.36 x 107 nm3). The thickness at the edge of rectangular cell-like units was on the order of magnitude of 682 nm (average, n = 3). The middle of the cells was thicker than the edges to such an extent that it hindered analysis via cryo-TEM.

Individual segments were surrounded by a dark (dense) membrane-like layer and then a lighter (sparse) layer (Figure 4.3). Within segments there were often numerous bubble-like structures. In some cases, similar bubble-like structures were present on the exterior of the rectangular cell-like units. It is unclear if they are the same bubble-like structures that are seen inside, and if so whether this is a natural phenomenon or potentially due to artificially induced cellular disruptions that occurred after samples were collected.

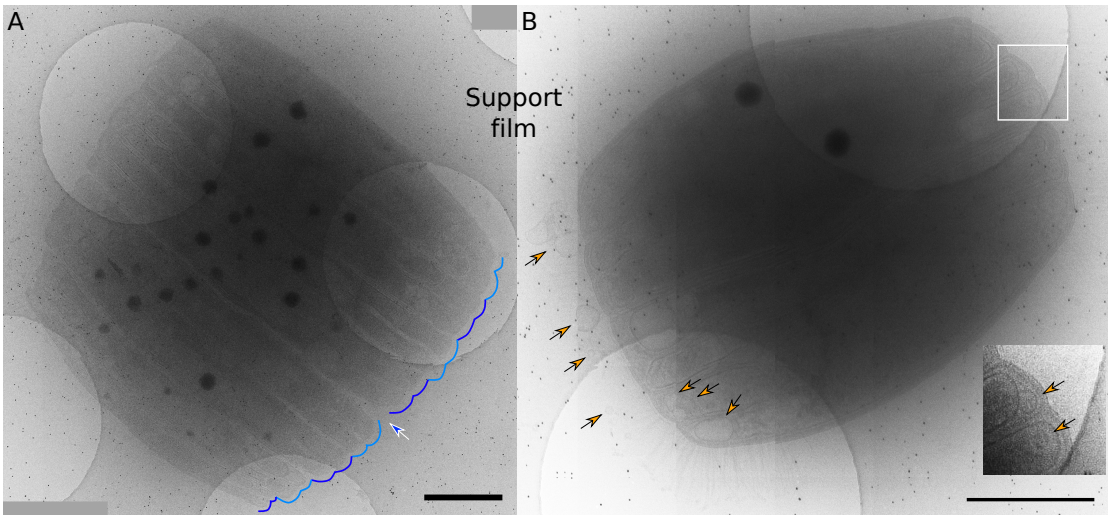


Figure 4.2: Cryo-TEM image with paired parallel segments in rectangular cell-like units. (A) Pairs of segments are denoted by lines in alternating shades of blue. Sharp indentations may occur between groups of paired segments; see blue arrow. (B) Bubble-like structures could be seen on the interior and exterior of the rectangular cell-like unit; see orange arrows for examples. Inset shows bubble-like structures near the edges of two segments. Scale bars: 1 um.

4.3.3 Cryogenic electron transmission microscopy insights into surface structures of rectangular cell-like units

A cell's surface is the frontline of its interaction with the environment. Surface proteins are involved in a wide range of biological functions, such as motility, adhesion, communication with other cells, chemosensing, and more (reviewed in Georgiou *et al.*, 1993). With regards to the surface of the rectangular cell-like units, pili-like appendages were prominent and could be observed to originate from individual segments within rectangular cell-like units (Figure 4.3). Such appendages were not always visible around rectangular cell-like units, though when present they tended to formed clusters with splayed tips.

The corrugated periodicity of the outermost layer of the rectangular cell-like units was suggestive of an S-layer-like structure. S-layers are composed of single protein or glycoprotein units that are secreted by cells. Upon exiting the cell, these

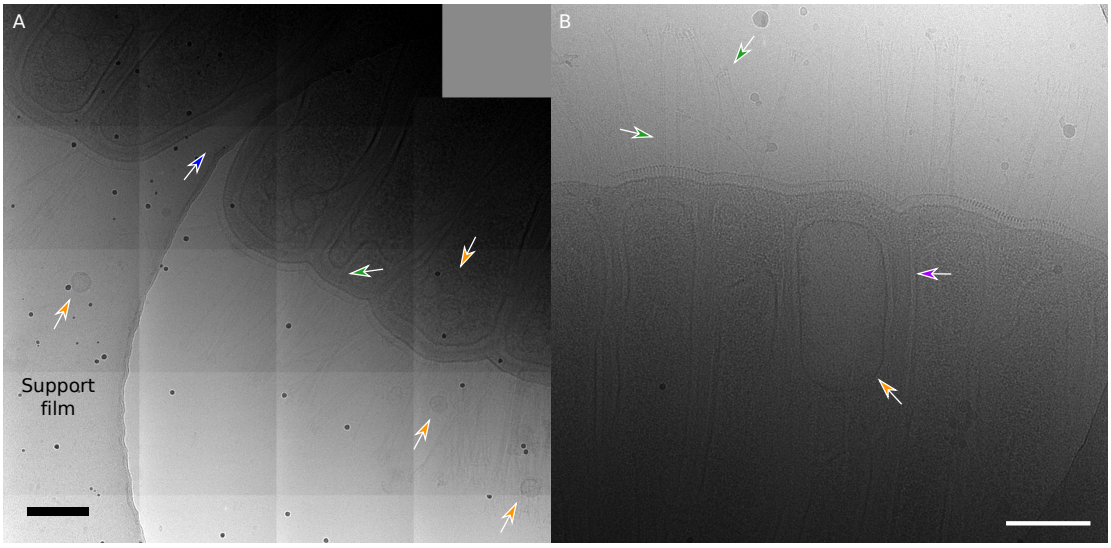


Figure 4.3: Cryo-TEM images reveal morphological features of rectangular cell-like units. Each panel shows a different rectangular cell-like unit. (A) A sharp indentation is present within a set of rectangular cell-like units; see blue arrow. An S-layer-like structure encircles each rectangular cell-like unit but is discontinuous at the indentation. Pili-like appendages originate from segments and protrude through the S-layer-like structure; see green arrow. Bubble-like structures can be seen within segments, with similar structures sometimes appearing on the outside of rectangular cell-like units; see orange arrows. (B) Segments are surrounded by a dark (electron-dense) layer followed by a light (less dense) layer; see purple arrow. Pili-like appendages often extended in a linearly arranged bunches and split out at the tips; see green arrows. A large bubble-like structure is present; see orange arrow. Scale bars: 200 nm.

units self-assemble into a porous closed lattice around the cell surface (reviewed in Sleytr *et al.*, 2006; Fagan & Fairweather, 2014). The S-layer-like structure encircled entire rectangular cell-likes units but discontinued at locations where groups of segments appeared to be splitting off from one another.

4.3.4 Potential contacts between rectangular cell-like units and other cells

Microbial communities are intertwined by complex networks of interspecies interactions (Kolenbrander *et al.*, 2000; Kuramitsu *et al.*, 2007; Anantharaman *et al.*, 2016). Such interactions have the potential to structure the spatial organization of cells within a community (Welch *et al.*, 2016). In dolphin oral samples, rectangular cell-like units were frequently observed in proximity to other unidentified bacterial or archaeal cells (Figure 4.4, 4.5, Appendix C Figure C.1). In Figure 4.5, the unidentified cell type in proximity to the rectangular cell-like unit at the bottom of the image (denoted by blue arrows) was repeatedly seen near rectangular cell-like units, and in some cases such cells appeared shriveled and possibly connected to rectangular cell-like units via pili-like appendages (ex: Appendix C Figure C.1B). In the same figure, insets A and B are reminiscent of nested vesicles (Dobro *et al.*, 2017) or possibly ultrasmall cells (Luef *et al.*, 2015). Interestingly, the S-layer-like structure of the rectangular cell-like unit appeared to be disrupted where it overlaps with the unidentified structure in Figure 4.5, inset B.

4.3.5 Taxonomic identification of rectangular cell-like units

The taxonomic identification of specific cell morphotypes from complex communities can be extremely difficult, to the point that it often remains unresolved (Alam *et al.* 1984; Wanger *et al.*, 2008; Luef *et al.*, 2015; and others). As a

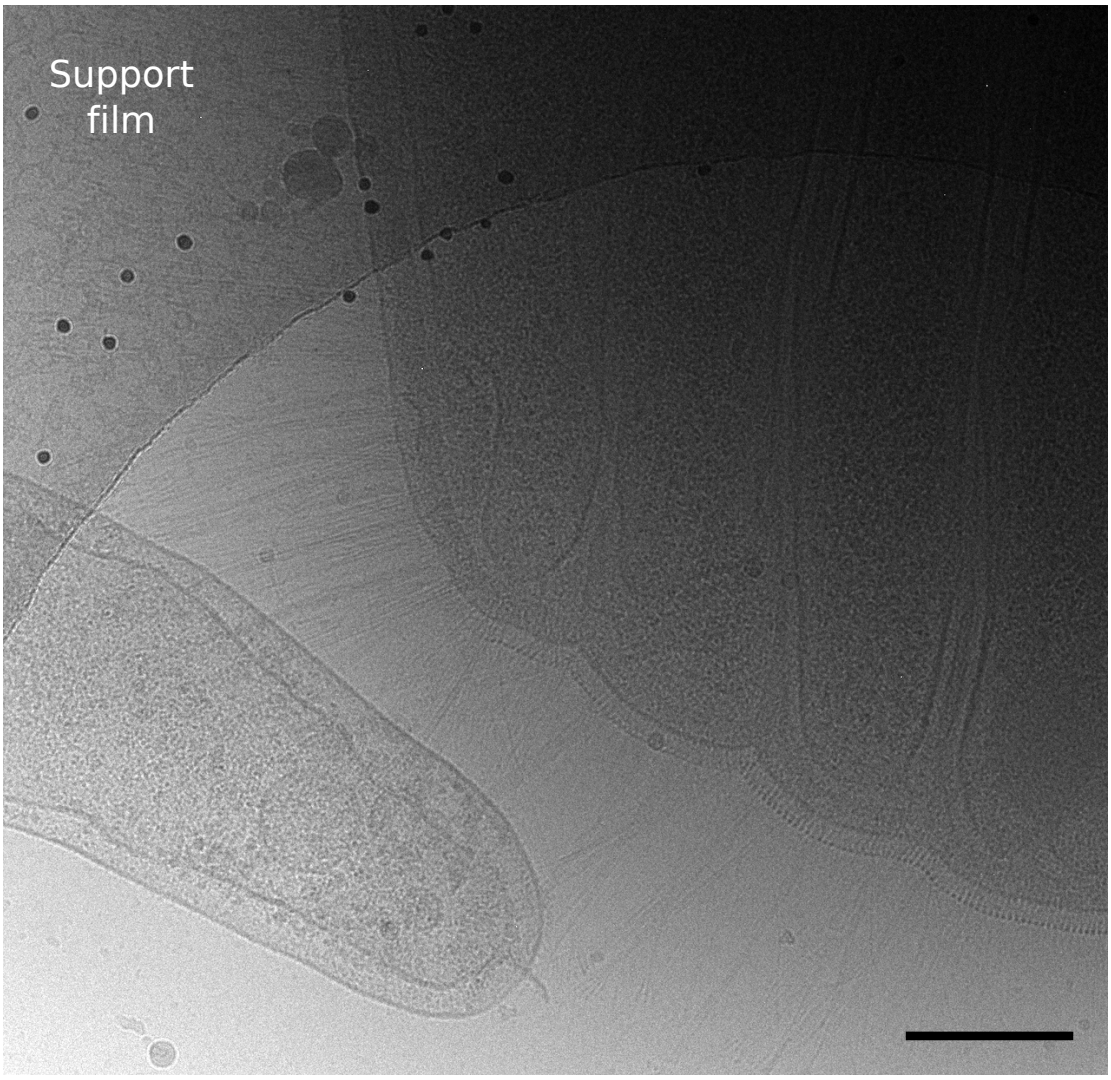


Figure 4.4: Cryo-TEM image documenting rectangular cell-like unit in proximity to another cell. A rectangular cell-like unit is present across the top right half of the image with pili-like appendages potentially projecting towards an unidentified bacteria-like cell. Scale bar: 200 nm.

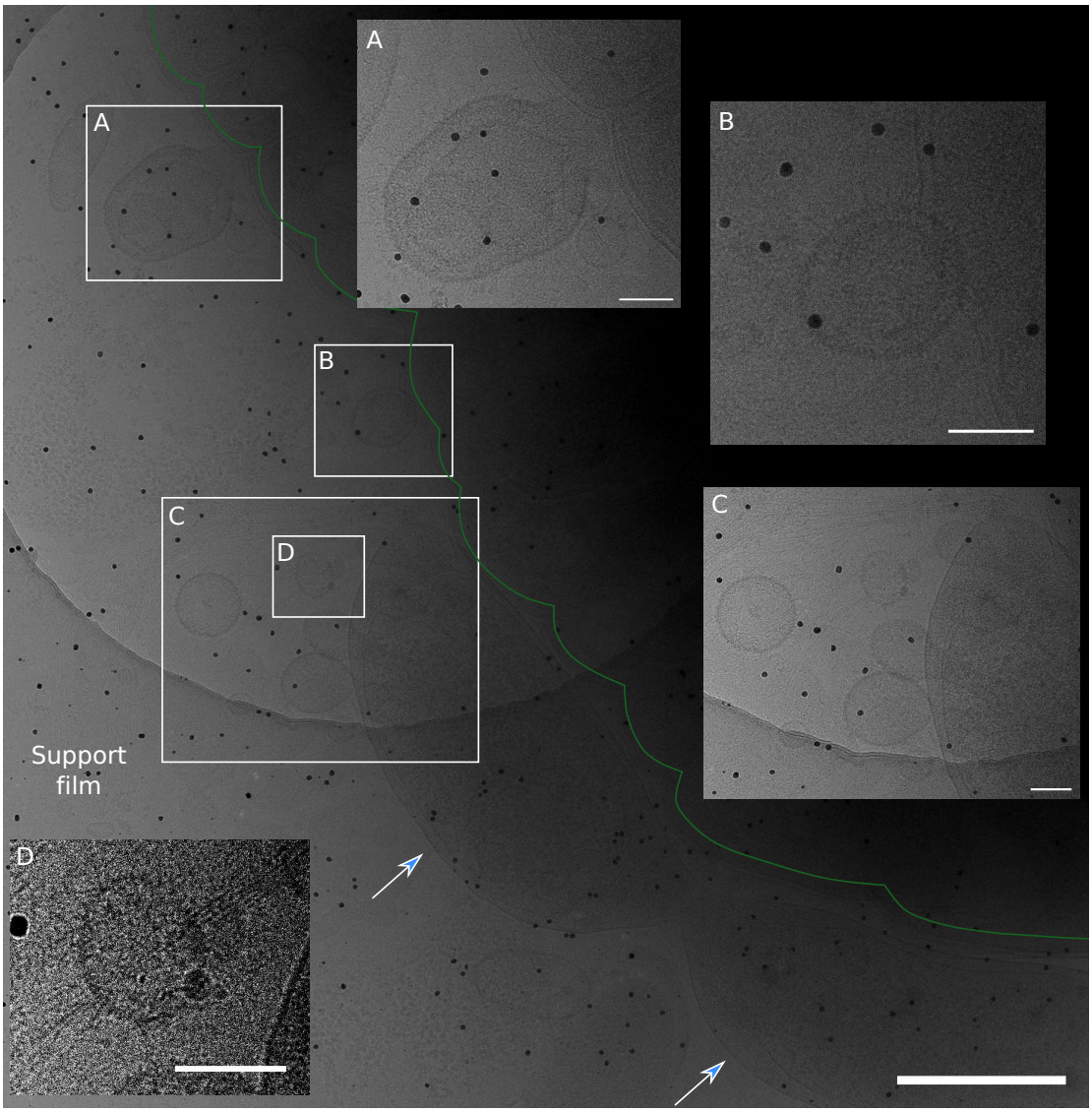


Figure 4.5: Cryo-TEM image documenting rectangular cell-like unit in proximity to other cells and probable extracellular vesicles. The rectangular cell-like unit can be seen across the top right half of the image and it is outlined by a green line. Blue arrows point towards a recurring, unidentified type of cell seen in close proximity to rectangular cell-like units. (A, B) Unidentified structures in proximity to the rectangular cell-like unit. In (B) a nick was present in the S-layer-like structure where the rectangular cell-like unit overlapped with the other structure. (C) Bubble-like structures of unknown identity and origin are featured. (D) Small spikes protruded from the surface of at least one bubble-like structure. Scale bar on main image: 500 nm, scale bars on insets: 100 nm.

first step towards identifying the rectangular cell-like units, we designed a single-cell genomics experiment. This experimental approach was selected to avoid bias towards any preconceptions about their possible identity.

We first tried three techniques to capture rectangular cell-like units for genomic sequencing: laser capture microdissection, microfluidics, and cell micromanipulation. Only the last allowed us to capture rectangular cell-like units in a collection tube. Non-target cells were also likely collected given the frequent close proximity between rectangular units and other smaller cells. Rectangular cell-like units would sometimes stick to the glass micropipette, making the exact number of rectangular units that were deposited into the collection tube unclear, although it was likely approximately five. While experimentally challenging, this observation offers insight into the biology of the rectangular cell-like units: they are sticky. DNA from captured cells and the surrounding aqueous environment was amplified via Multiple Displacement Amplification (MDA), as was the contents of a negative control in which water was added in the place of cells.

Paired-end Illumina MiSeq reads were generated, resulting in >323 Mbp and >190 Mbp of paired-end sequencing data for the sample and negative control, respectively, and these reads were used to generate assemblies. We first evaluated whether the negative control assembly (and therefore by proxy the sample) contained significant contamination resulting from the MDA reaction. It did not (see Methods) and was therefore excluded from downstream analysis. We next evaluated the degree to which the sample assembly was representative of the reads that were sequenced; a satisfactory 92% of reads mapped to the assembly. As such, we proceeded with the analysis by creating a tetranucleotide emergent self-organizing map (ESOM) of scaffolds for the purpose of binning (grouping together) scaffolds into microbial genome bins (Appendix C Figure C.2).

	Bin 1	**Bin 2**	**Bin 3**	**Bin 4**
Phylum	Bacteroidetes	Bacteroidetes	Sacchari-bacteria (TM7)	Proteobacteria
More specific taxonomy	Class Flavo-bacteriaceae & plausibly genus *Tenacibaculum*	Class Bacteroidia & plausibly order Marinilabiliales	-	Class Epsilon-Proteobacteria & Family *Campylo-bacteraceae*
# scaffolds	367	96	24	26
Length of assembly (bp)	4843190	1530584	619141	325155
N50 (bp)	16327	19832	55924	18717
Longest scaffold (bp)	108073	67276	87328	57664
%GC	33%	38%	50%	32%
Median read coverage	19.87	15.11	17.61	15.03
# protein coding genes	4662	1444	611	372
Bin completeness	94.51%	66.31%	47.41%	17.24%
% bSCG w/ >1 copy	57%	5%	0%	0%

Table 4.1: Properties of genome bins recovered from the single-cell genomics experiment. Four genomes were recovered from the single-cell genomics experiment. Scaffolds were assigned to genome bins on the basis of tetranucleotide frequency. In the "more specific taxonomy" row, the class and the rank of the lowest plausible taxonomic assignment is denoted (see Methods). Note: no widely accepted class designations exist for the phylum Saccharibacteria.

Four genome bins were recovered: two from the Bacteroidetes phylum, one from the Saccharibacteria (TM7) phylum, and one from the (Epsilon-) Proteobacteria phylum. Genome bin properties are summarized in Table 1. Due to the dearth of phylogenetically informative marker genes assembled, namely 16S/18S rRNA genes and a commonly used set of 16 ribosomal proteins that can identify genomes from all domains of life (Hug *et al.*, 2013), taxonomic classification was based on protein similarity to those present in the NCBI non-redundant protein database (see Methods). In an attempt to gain further insight into the identity of recovered genomes, we queried them against two dolphin oral metagenomes, DolZOral124 and DolJOral78, that were assembled in Dudek *et al.* (2017). The Dudek *et al.* (2017) metagenomes were derived from oral samples from two dolphins from the US Navy MMP for the purpose of recovering genomes of microorganisms that inhabit this environment. Comparison against the two metagenomes revealed that the Bin 3 genome from this study had an average amino acid identity (AAI) of 99% and 91% to Saccharibacteria genomes from the DolJOral78 and DolZOral124 metagenomes, respectively. This is consistent with the three genomes originating from representatives of the same species (Rodriguez & Konstantinidis, 2014). None of the other genome bins recovered from the single-cell genomic analysis could be linked at the species level to genome bins previously recovered from the dolphin mouth (see Methods).

4.4 Discussion

The vast majority of microorganisms lack isolated representatives (Hug *et al.*, 2016). Sequencing-based analyses have proved invaluable in exploring and describing said diversity, yet cannot be used to explore all aspects of the biology of microorganisms. Notable blind spots in our understanding of uncultured or-

ganisms include the unique genes and corresponding features which have evolved within these lineages alone and the nature of their interactions with other members of their community. A shift towards a more multifaceted approach, drawing on a diversity of disciplines and techniques, will be required to create a comprehensive view of the biology of such lineages (Ponomarova & Patil, 2015; Xu *et al.*, 2017; Castelle & Banfield, 2018). The use of imaging techniques in describing the features and forms observed in microbes has great potential to provide insight into the biology of uncultured lineages of life (Baker *et al.*, 2010; Comolli & Banfield, 2014; Luef *et al.*, 2015).

In this study, microscopy revealed the presence of morphologically unusual rectangular cell-like units in the mouths of bottlenose dolphins. The rectangular cell-like units were present in at least one sample from all eight dolphins included in this study, with samples having been collected over the course of two years. This suggests that the rectangular cell-like units are endemic to the dolphin mouth.

Rectangular structures are a rarity in the microbial world and come in two flavours: cells that are rectangular and cell aggregates that are rectangular. To the best of our knowledge, the discovery of non-eukaryotic rectangular cells has thus far been restricted to the the family *Halobacteriaceae*, which consists of halophilic Archaea. Known rectangular cells from this family include *Haloquadratum walsbyi* (Walsby, 1980), *Haloarcula quadrata* (Oren *et al.*, 1999), and members of the pleomorphic genus *Natronrubrum* (Xu *et al.*, 1999). Additional occurrences of rectangular cells believed to be bacterial or archaeal have been discovered in high salinity environments but were not taxonomically identified (Alam *et al.* 1984; Oren *et al.*, 1996). Amongst eukaryotic microorganisms, species with cells that may have a rectangular appearance (at least in two dimensions) include diatoms such as those in the genera *Cerataulina*, *Lauderia*, and *Skeletonema*, although

these cells are cylindrical rather than being true rectangular prisms (Horner, 2002). In contrast, a variety of types of rectangular cellular aggregates are known to be formed by bacteria, such as a) rectangular sheets composed of coccoid bacteria (ex: *Thiopedia rosea* and those in the genus *Merismopedia*), b) cuboidal structures composed of coccoid bacteria (ex: those in the genera *Sarcina* and *Eucapsis*), c) rectangular chains of filamentous bacteria (ex: those in the genus *Simonsiella*), and d) rectangular trichomes formed by disc-shaped bacteria (ex: *Oscillatoria limosa* and other Cyanobacteria) (reviewed in Hedlund & Kuhn, 2006; Zwinder & Dworkin, 2006; and elsewhere).

The rectangular cell-like units observed here are mostly likely aggregates of cells, whereas segments are the units of individual cells. The following observations support this hypothesis: a) segments appeared to be surrounded by membrane-like structures, which are reminiscent of their having a plasma membrane and cell wall b) segments are arranged in the same manner as the bands that were stained with DAPI, which is suggestive of their having naked DNA within, c) pili, which are surface features of cells, projected out from individual segments, d) rectangular cell-like units often consisted of groups of variable numbers of segments that appeared to be separating from one another, suggesting that the rectangular structures are not the unit of an individual cell. The paired nature of segments can likely be explained by their undergoing longitudinal binary fission, as seen in members of the *Simonsiella* genus (Steed, 1963). Interestingly, the segments at the ends of rectangular cell-like units are often shorter in width than those nearer to the center. This suggests that there is a mechanism by which the length of segments is determined in a manner which is dependent on their spatial positioning within a unit.

If the segments really are the individual units of the cell, this raises interesting

questions about whether cells in rectangular cell-like units exhibit cooperation. Cryo-TEM imaging revealed that rectangular cell-like units were encapsulated by S-layer-like structures, which appeared to be discontinuous at points where groups of segments appeared to be separating from one another. S-layers are self-assembling, crystalline arrays of single proteins or glycoproteins that coat the exterior of the whole cell (reviewed in Sleytr *et al.*, 2006; Fagan & Fairweather, 2014). While their exact function varies widely (and is usually unknown), S-layers undoubtedly play an important role in cells, given their high metabolic cost (up to 20% of protein synthesis), their ubiquity across bacteria and archaea, and their multiple evolutionary origins (Sleytr *et al.*, 2006; Fagan & Fairweather, 2014). In the rectangular cell-like units, if the rectangular units are not individual cells, then the S-layer-like structure surrounding them is likely produced and secreted by the segments (cells). From a theoretical perspective, one can speculate that cooperation between cells could have evolved due to the fact that close kin (i.e. cells in a chain) have limited dispersal ability, and are therefore situated in close physical proximity. As such, the degree of relatedness between neighbouring cells in a unit may have been greater than the cost-to-benefit ratio of contributing to communal S-layer production, leading to the evolution of cooperation (see Nowak, 2006 and Bruger & Waters, 2015 for review on the evolution of cooperation).

From a functional perspective, if the S-layer-like structure of the rectangular cell-like units plays a critical role in cell biology such as protection from bacteriophage or bacteriocins (as reported for other microorganisms, see Fagan & Fairweather, 2014), then cells could benefit from cooperatively producing a single S-layer around the unit rather than each individual segment, as the surface area to cover per cell would decrease while the same function would be maintained. Such a phenomenon could potentially even have contributed to selection for the

aggregation of cells. An additional and not mutually-exclusive hypothesis it that the S-layer-like structure may be involved in maintenance of the shape and organization of the segments with respect to one another, similar to what is seen in archaea such as *Thermoproteus tenax* (Wildhaber & Baumeister, 1987). Questions for future study include: what is the function of the S-layer-like structure present around the rectangular cell-like units, is it the product of a cooperative task by multiple cells, and if so, how is cheating avoided?

Another interesting feature of the rectangular cell-like units is the large number of bubble-like structures contained within. The size of the bubble-like structures was highly variable. Some were <50 nm in diameter while others were the width of entire segments (≈200 nm). We hypothesize that they are likely vesicles, although they may also constitute ultrasmall bacterial or archaeal cells, bacteriophage, or a combination thereof. For perspective, vesicles have been reported with diameters in the range of 20 nm - 500 nm (Brown *et al.*, 2015), while the smallest cell diameter possible (not including the cell wall) is theorized to be in the range of 186 nm - 339 nm (de Duve *et al.*, 1999). The biogenesis of vesicles in bacteria is poorly understood, although they appear to be widespread in bacteria (Brown *et al.*, 2015; Dobro *et al.*, 2017). In *Gemmata obscuriglobus* (phylum Planctomycetes) vesicles mediate endocytosis-like protein uptake (Lonhienne *et al.*, 2010), while extracellular vesicles from a variety of bacteria and may contain nucleic acids, proteins, and polysaccharides and are likely important mediators of cell-cell interactions (Brown *et al.*, 2015). Notably, bubble-like structures were frequently observed on the exterior of rectangular cell-like units. We cannot determine from microscopy images whether these originated from the rectangular cell-like units, and if so whether this is a naturally occurring phenomenon. Nonetheless, it is interesting that bubble-like structures were often seen on the exterior of rectangular cell-like

units and were sometimes in close proximity to other types of cells, occasionally overlapping with or possibly inside of them (Figures 4.4, 4.5, Appendix C Figure C.1A).

On their surface, microbial cells often produce filamentous structures such as pili and flagella. These are involved in important functions such as motility, adhesion, biofilm formation, conjugative DNA transfer, and bacteriophage infection (Soto & Hultgren, 1999). There likely remains much to be explored regarding the diversity of such structures; as recently as 2016, a new type of bacterial pili was characterized from, and found to be ubiquitous in, the human gut microbiota (Xu *et al.*, 2016), while the discovery and ongoing characterization of hami in archaea provides a striking example of the novelty of features that have evolved in little-studied branches of the tree of life (Moissl *et al.*, 2005; Perras *et al.*, 2015). The pili-like appendages that protruded from segments within the rectangular cell-like units had an unusual architecture or organization in that they often formed stalks with splayed tips. This is in contrast to the comparatively simple bacterial and archaeal pili that typically exist as single hair-like appendages without further morphological features (Fernandez & Berenguer, 2000; Hospenthal *et al.*, 2017).

Microscopy-based observations have raised many questions about the biology of the rectangular cell-like units. Additional points not touched on above include: What are the dark (dense) essentially spheroidal spots (ex: storage granules)? Why do they sometimes appear in a disorganized fashion near the center of the cell (Figure 4.2) but in other rectangular cell-like units appear in lines that run perpendicular to segments (Figure 4.1B)? Do rectangular cell-like units interact with other cells in their environment, and if so, what is the nature of the interaction? The unique and complex structure of the rectangular cell-like units suggests there is much to be learned about a unique biology.

Given the intriguing morphology of the rectangular cell-like units, an important question is: what are they? As a first pass towards answering this question, we performed a single-cell genomics experiment. Importantly, this approach is unbiased towards preconceptions of what the cells may be and is essentially unbiased in terms of its ability to detect organisms across the tree of life. Four microbial genomes were recovered from the experiment, all of which were bacterial. This suggests that the rectangular cell-like units are bacteria, likely from either the Bacteroidetes, Saccharibacteria, or (Epsilon-) Proteobacteria phyla (note: due to the polyphyletic nature of the Proteobacteria phylum, we have included the class level designation).

It is unsurprising that genomes from multiple organisms were obtained. First, small non-target cells were likely captured alongside rectangular cell-like units, either because they were associated with the rectangular cell-like units (as cryo-TEM images reveal is often the case) or because they were accidentally collected alongside rectangular cell-like units. Second, cell-free DNA in the sample may have been captured. Third, it is possible that despite our best efforts to remove contaminant DNA via UV irradiation (see Methods), there was contamination associated with reagents and collection devices used for micromanipulation and cell capture. It is also important to recognize that the genomes of rectangular cell-like units may not be represented by those recovered from the single cell genomics experiment, as we cannot be absolutely certain that any target cells lysed during the MDA amplification process, although we have no reason to believe that they would not. The sequencing results provide important information for designing future experiments to conclusively determine the identity of the rectangular cell-like units. Ultimately any conclusion regarding their identity will require multiple lines of evidence all supporting the same taxonomic assignment.

In the interim, while may be tempting to jump to conclusions regarding the identity of the rectangular cell-like units based on morphology alone, such endeavours are futile at best and have the potential to be greatly misleading and bias-inducing at worst. For example, the rectangular nature of the units may lead one to conclude that the cells are diatoms. However, the segments (which are likely the individual units of cells) are not rectangular, the rectangular cell-like units show no evidence of having organelles that one would expect to find in diatoms (mitochondria, etc), and there is no evidence to suggest that the rectangular cell-like units have a frustule, which is a silica encasing that is a fundamental feature of diatoms (for review and comparison see Stoermer *et al.*, 1965; Hasle *et al.*, 1970; Ross & Simms, 1972; Falasco *et al.*, 2009). Furthermore, while diatoms may appear to form rectangular cells or chains when viewed in two dimensions, they tend to be cylindrical rather than being rectangular prisms, the latter of which corresponds with the shape of the rectangular cell-like units. Finally, diatoms are photosynthetic, and thus the ecology of said organisms makes them an unlikely candidate for cells that were consistently found in the dolphin mouth.

Another morphologically-driven hypothesis which is more in line with the ecology of the system in question is that the cells may be related to the genus *Simonsiella* (phylum Proteobacteria, class Betaproteobacteria, order Neisseriales, family *Neisseriaceae*) (reviewed in Hedlund & Kuhn, 2006). *Simonsiella* forms aggregates of longitudinally-dividing cells organized in chains and is a common commensal member of the vertebrate oral microbiota, including that of humans, dogs, and cows (reviewed in Hedlund & Kuhn, 2006). However, the unusual organization seen in *Simonsiella* is a characteristic known to have evolved at least twice within bacteria, as members of the genus *Moraxella* also display a chain like organization (Xie & Yokota, 2005), and were originally misclassified as *Simonsiella*

as a result (reviewed in Hedlund & Kuhn, 2006). Internal and surface features of the rectangular cell-like units appeared to be quite distinct from *Simonsiella*. For example, the rectangular cell-like units had an S-layer-like structure encapsulating each unit, a matrix-like substance in between segments, and numerous bubble-like structures, while they lacked any indication of the stark dorsal-ventral differentiation characteristic of *Simonsiella* members (for comparison see Pangborn *et al.*, 1977; McCowan *et al.*, 1979; Pankhurst *et al.*, 1988). Additionally, previous characterization of the microbiota of dolphins via a 16S rRNA gene amplicon survey (Bik *et al.*, 2016) did not detect *Simonsiella* in the mouths of dolphins under the purview of the Marine Mammal Program in San Diego Bay, California, USA, from which swab samples in this study were collected. Overall, these observations, combined with evidence from the single-cell genomic experiment, suggest that the rectangular cell-like units studied here are likely not members of *Simonsiella*, but rather have likely independently evolved to have similar organization of cell aggregates (if segments are individual cells). Morphology-based observations can lead to a variety of radically different hypotheses regarding the taxonomic of the rectangular cell-like units. The only way to satisfactorily address what the rectangular cell-like units are is to gather multiple lines of complementary evidence through a succession of carefully crafted experiments. The single-cell genomics experiment performed here was the first such experiment.

The rectangular cell-like units presented in this study highlight the incredible diversity of lifeforms that have evolved on Earth and remain to be discovered and understood. Characterization of the their structure has opened the door to many questions about their biology, not least of which is: what are they? Future fluorescence in situ hybridization (FISH) experiments will be used to evaluate the likelihood of the candidate identities generated through the single-cell genomics

experiment. Additionally, repeating the single-cell genomic experiment performed here numerous times will help bolster confidence in a taxonomic assignment for the rectangular cell-like units, and may also provide insight into whether they repeatedly are associated with specific other types of microorganisms. Future experiments may also provide greater insight into the general biology of the rectangular cell-like units. For example, genome-based reconstruction of its metabolic profile should offer insight into the ecology of this taxon, while microscopy-based investigations may shed light on aspects of cell biology such as the contents of the dense spots (via staining experiments) and characteristics of the S-layer-like structure (via single particle cryo-TEM). Ultimately, the rectangular cell-like units illustrate the fact that in regards to characterizing and understanding the diversity of lifeforms that exist on Earth, much remains to be discovered.

4.5 Methods

4.5.1 Sampling collection

Oral swab samples were obtained from bottlenose dolphins (*Tursiops truncatus*) managed by the U.S. Navy Marine Mammal Program (MMP) Biosciences Division, Space and Naval Warfare Systems Center Pacific, San Diego, California. Samples were obtained by swabbing the left gingival sulcus or a combination of three spots: the roof of the mouth, the tongue, and in between the tongue and mandible. The swabbing protocol adhered to the guidelines described in the CRC handbook of Marine Mammal Medicine. The MMP is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) International and adheres to the national standards of the United States Public Health Service Policy on the Humane Care and Use of Laboratory Animals and

the Animal Welfare Act. As required by the U.S. Department of Defense, the MMP's animal care and use program is routinely reviewed by an Institutional Animal Care and Use Committee (IACUC) and by the U.S. Navy Bureau of Medicine and Surgery. The animal use and care protocol for MMP dolphins in support of this study was approved by the MMP's IACUC and the Navy's Bureau of Medicine and Surgery (IACUC #92-2010, BUMED NRD-681). Samples were collected during the years 2017-2018. Rectangular cell-like units were identified in at least one sample from all eight dolphins from whom swabs were analyzed in this study.

To remove cells from swabs, swabs were immersed in 1X PBS (usually 50-100 ul, depending on cell density) in microcentrifuge tubes. Tubes were vortexed vigorously (>10 seconds) and lightly centrifuged to remove liquid from tube caps. The resulting PBS-cell solution was used for subsequent analyses.

4.5.2 Light microscopy data acquisition

Cells from PBS-cell solution were stained with DAPI for a final concentration of 0.5 ug DAPI / ml. The solution was applied to an agarose pad on a cover slip and imaged with a Nikon Eclipse Ti microscope at 100X magnification. For fluorescence microscopy of DAPI-stained cells, an emission/excitation spectrum of 340/488 nm was used.

4.5.3 Cryo-TEM data acquisition

PBS-cell solution was applied to glow-discharged 200-mesh copper Quantifoil grids with holey carbon or gold GridFinder Quantifoil grids. 2 ul of 15nm gold fiducial beads were applied to both sides of each grid. Grids were blotted for five seconds and plunge frozen in liquid ethane cooled by liquid nitrogen using

a Leica EM GP. Samples were loaded into one of two cryo-transmission electron microscopes: 1) TEM2 - a Titan Krios G3 operated at 300 kV with an energy filter or 2) TEM4 - a Titan Krios G4 operated at 300 kV with no energy filter. Both microscopes were outfitted with a K2 summit direct electron device (Gatan) to record micrographs. Montage images were acquired semi-automatically using the program SerialEM (Mastronarde *et al.*, 2003) and imaging data was processed using iMOD (Kremer *et al.*, 1996).

4.5.4 Single-cell genomics experiment

To obtain candidate identities for the rectangular cell-like units, we employed a single cell genomic approach. Since contamination by foreign DNA was of concern, reagents and tubes were UV treated such that they received a UV dose of 11.4 J/cm^2, following the guidelines proposed by Woyke *et al.* (2011). Individual rectangular cell-like units were captured using a cell micromanipulator. Captured units were deposited in a UV treated collection tube containing UV treated 1X PBS. As rectangular cell-like units sometimes adhered to the glass of the micropipette, it is unclear how many captured units were deposited in the UV treated collection tube, although the number was greater than three and less than 10. No dolphin cells were captured, although small non-target cells or free floating DNA from the sample may have been acquired as contaminants along with rectangular cell-like units.

DNA from cells was amplified via multiple displacement amplification (MDA) using the Repli-g single cell genomic kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. A negative control consisting of UV treated 1X PBS but no cells was included in the experiment and processed alongside the sample. Resulting amplified DNA was high molecular weight (>1 kb) for both

the sample and control. DNA from the sample and control was purified using a Zymo Clean and Concentrate Spin Column (Zymo Research Corporation, Irvine, CA, Cat No. D4013) and libraries were prepared using the Kapa Hyper Prep Kit (Kapa Biosystems, Wilmington, MA, Cat No. KK8504) at the W.M. Keck Center for Comparative Functional Genomics at the University of Illinois, Urbana-Champaign (USA). The two libraries were sequenced across a single MiSeq Nano PE v2 lane, resulting in 649,607 pairs of reads (2 x 250bp) for the rectangular cell-like units sample and 439,339 pairs of reads (2 x 250bp) for the negative control sample. Sequencing adaptors were removed at the Keck Center.

Reads were assembled using SPAdes (Bankevich *et al.*, 2012) with the single cell (–sc) and careful (–careful) modes specified and protein coding genes were identified using Prodigal version 2.6.2 (Hyatt *et al.*, 2010). To assess the degree of contamination resulting from the MDA reaction, we searched for 139 bacterial single copy genes (Campbell *et al.*, 2013) in the negative control assembly. This was done using HMMER suite version 3.1b2 (Finn *et al.*, 2011), using each HMM profile's gathering cutoffs to set significance thresholding (–cut_ga). No bacterial single copy genes were identified. Combined with the lack of long contigs ($\geq$1kb), this suggests that there was no appreciable amount of MDA-specific contaminant bacterial DNA (i.e. from reagents) amplified and subsequently sequenced and assembled. Therefore the negative control was excluded from downstream analysis. To determine the extent to which the sample assembly was representative of the reads sequenced, we mapped reads against the assembly using bowtie2 version 2.2.4 (Langmead & Salzberg, 2012). 92.43% of reads mapped to the assembly. Per scaffold average coverage was calculated using the samtools version 1.6 (Li *et al.*, 2009) depth function to calculate per base read coverage and a custom script was used to calculate average read coverage per scaffold.

To determine the taxonomic identity of sequenced cells, we employed a genome-resolved approach. Assignment of scaffolds to genome bins was performed using the tetranucleotide frequencies of all scaffolds ≥5000 bp long over windows of 5000 bp, as described in Dick et al (2009). Results were computed and visualized using the Databionics ESOM Tools software (Ultsch & Moren, 2005), leading to the reconstruction of four genome bins. To refine bins, we removed scaffolds for which fewer than 50% of keys were assigned to the bin. Scaffolds <5000 bp long were not binned. In total, 483 scaffolds with a cumulative length of 5,148,301 bp remained unassigned. The completeness and contamination per bin was assessed using CheckM version 1.0.7 (Parks *et al.*, 2014).

Taxonomic identification of bins posed a challenge since a) no 16S/18S rRNA gene assembled in the dataset, and b) genomes were partial and few phylogenetically informative bacterial single copy genes (Hug *et al.*, 2013) were present. As such, to obtain a phylogenetic signal we used BLAST version 2.2.30 (Altschul *et al.*, 1990) to query all protein coding genes from each genome against the NCBI non-redundant protein database using an e-value of 1e-10 and noted the taxonomic affiliation of the closest protein match. Taxonomic assignments were made based on the taxa with the highest level of similarity. Genome bin taxonomic assignments were considered to be highly likely if ≥50% of the top blast hits originated from a single taxon and were considered to be plausible if <50% but ≥33% of the top blast hits originated from a single taxon. To determine whether genomes from closely related taxa were recovered from the Dudek *et al.* (2017) genome-resolved metagenomic study of the dolphin mouth, we queried scaffolds from bins recovered here against dolphin metagenome assemblies using BLAST version 2.2.30 (Altschul *et al.*, 1990) with an e-value of 1e-10. We investigated whether scaffolds from metagenomes with a match of ≥90% identity over ≥5 kb were binned into

genomes in the Dudek *et al.* (2017) study. If so, we computed the average amino acid identity of the given genomes from the 2017 study and the present study using the Kostas Lab AAI calculator (http://enve-omics.ce.gatech.edu/aai/) with default parameters.

Chapter 5

Conclusion

5.1 Summary remarks

In this dissertation I presented three projects aimed at exploring the diversity, form, and function within the marine mammal microbiota. In Chapter two, I performed the first broad-scale, culture-independent survey of the bacterial communities associated with sea otters, which are an IUCN endangered, keystone species (Estes & Palmisano, 1974; Estes *et al.*, 1998; Doroff & Burdin, 2015). This provided the first insight into the bacterial communities associated with this species, which are the sole representatives of one of the six extant lineages of marine mammals. The results suggest that environment is a prominent determinant of bacterial community structure in both gingival and rectal environments in sea otters, and raises the question of whether sea otters may harbour a reduced gut microbiota compared to other mammals. As seen in other marine mammals (Bik *et al.*, 2016), sea otters were found to host a diversity of bacterial candidate phyla representatives. In Chapter three, I focused on learning about the ecology and evolution of specific, previously unstudied lineages of candidate phyla bacteria associated with dolphins, using genome-resolved metagenomics. This led

to our proposing two new bacterial phyla. Novel taxonomic diversity in these communities was accompanied by novel functional diversity, including unusual CRISPR-Cas9 systems. In Chapter four, I discovered and described rectangular cell-like units present in the dolphin mouth. Atypical morphological properties of this taxon were observed by microscopy such as architecturally complex pili-like appendages and an S-layer-like structure potentially encapsulating multiple individual cells. Evidence from a single-cell genomics experiment, combined with morphological features (or a lack thereof, in the case or membrane-bound organelles characteristic of eukaryotes) strongly suggested that the rectangular cell-like units are bacterial.

5.2 Open problems for future work

5.2.1 Co-evolution of marine mammals and their microbiota

An ever increasing number of studies have begun to shed light on the composition of the microbiota of marine mammals (Glad *et al.*, 2010; Eigeland *et al.*, 2012; Lavery *et al.*, 2012; Lima *et al.*, 2012; Nelson *et al.*, 2013; Merson *et al.*, 2014; Sanders *et al.*, 2015; Bik *et al.*, 2016; Delport *et al.*, 2016; Soverini *et al.*, 2016, Godoy-Vitorino, *et al.*, 2017; Erwin *et al.*, 2017; Russo *et al.*, 2018; and others). With the addition of the work presented in Chapter two of this thesis, culture-independent microbiota surveys have now been conducted on five of the extant lineages of marine mammals. This helps lay the groundwork for future studies aimed at evaluating the effect of a marine lifestyle on the microbiota of mammals. For example, the sea otter study presented here adds to a growing body of evidence that provenance, and more specifically the dichotomy between a

marine vs terrestrial lifestyle, is an important determinant of microbial community composition (Bik *et al.*, 2016; Nelson *et al.*, 2013). Whether there are reproducible trends driving such differences in the five lineages of marine mammal, as well as what the underlying mechanisms may be, remains open to question.

Notably, we found that the sea otter gut microbiota is distinct from that of three other closely related otter species, all of which share similar gut morphology and diet, and are semi-aquatic. This suggests that neither spending substantial amounts of time in an aquatic environment nor eating a diet heavy in fish and/or aquatic invertebrates like crab is sufficient to account for this difference. One possibility is that marine mammal-associated communities may be partially seeded by seawater and/or their prey. Previous research has shown that in free-living (i.e. non-host-associated) bacterial communities, salinity is a major determinant of community composition (Ley *et al.*, 2008), which could potentially be a contributor to the divergence of marine vs terrestrial mammalian microbial communities.

The following two experiments could provide much needed insight on this question. First, one could sample seawater microbial communities in tandem with those associated with specific marine mammal individuals over multiple seasons and years, taking care to obtain seawater samples that are not contaminated by marine mammal excretions. A comparison of the degree to which communities change over time (e.g., between seasons), and whether the presence/absence or abundance of any taxa change in concert in seawater and marine mammals, would be highly informative. A second experiment could consist of tracking the microbiota of prey fed to a set of marine mammal individuals, as well as the microbiota of those marine mammals themselves, and comparing the extent to which the two display similarity in terms of the taxa present. Artificially introducing a diet shift could provide additional insight into how closely the microbiota of a given marine

mammal species mirrors that of their prey.

One unexpected finding presented in Chapter two of this thesis was the low percentage of bacterial DNA and high percentage of prey DNA present in sea otter fecal samples. It has been suggested that such findings may potentially be indicative of a low biomass gut microbiota, and that such a phenomenon may be widespread in species with short guts and rapid transit times (Hammer *et al.*, 2017). This led to the hypothesis that sea otters may have a reduced gut microbiota compared to other mammalian species, in line with their extremely rapid gut transit time on the order of 3 hours (Kirkpatrick *et al.*, 1955; Kenyon, 1969; Costa & Kooyman, 1984) and other biological features of this species that may weaken the strength of the symbiosis between them and their gut microbiota. Future studies aimed at quantifying the biomass of the sea otter gut microbiota and comparing it against that of other mammalian species will be required to test this hypothesis. Supporting experiments aimed at evaluating the effect of the gut microbiota on host fitness may also be informative, such as measuring physiological factors such as host weight before and after the introduction of antibiotics. In upcoming years, research comparing bacterial biomass, as well other features of communities that go beyond sequencing-based characterizations, will aid in generating a more robust synthesis regarding the general nature of host-microbiome co-evolution. This in turn will have important implications for the extent to which findings regarding the interaction between humans and our microbiota can be applied to other mammalian species.

5.2.2 Novel microbial diversity and functional potential associated with marine mammals

An interesting feature of marine mammal microbiotas is the rich diversity of bacteria present in these communities, including representatives from numerous candidate phyla and other poorly characterized lineages (Nelson *et al.*, 2013; Bik *et al.*, 2016; Dudek *et al.*, 2017). This presents an exciting opportunity to study the ecology and evolution of such lineages. For example, genomes recovered from members of candidate phyla, such as those recovered from the mouths of dolphins, as discussed in Chapter three of this thesis, open the door to future comparative genomic studies of host-associated vs non-host-associated members of these phyla. This collection of genomes could further be supplemented with genomes from other marine mammal species such as sea otters. At present, such comparative studies are challenging due to an insufficient number of phylogenetically balanced, host-associated genomes from a given candidate phylum, especially when attempting to include representatives of candidate phyla from across multiple mammalian species (the vast majority of publicly available genomes are from non-host-associated environments such as aquifers - see Brown *et al.*, 2015; Anantharam *et al.*, 2017, and others). Another interesting area for future research would be to conduct a survey of the environmental distribution of bacteria within given candidate phyla, including how frequently they have been detected in mammalian-associated communities, since the global distribution of of such phyla is generally unknown. Extensions to such a study could include an investigation of what features are shared by environments in which members of a given candidate phylum are found and the extent to which representatives from these phyla are associated with animals with different lifestyles.

With regards to the two novel phyla proposed in Chapter three, next steps to-

wards characterizing the biology therein could include the following approaches: a) surveys of their global distribution (e.g., via 16S rRNA gene databases), b) increasing genomic coverage of these candidate phyla via sequencing of communities in which representatives have previously been detected through 16S rRNA gene surveys or through data mining of already published shotgun sequencing datasets in which these lineages are present, c) microscopy-based studies to characterize morphology. Ideally, though, cultured representatives would be obtained for in vitro experimentation. One hypothesis as to why many candidate phyla bacteria do not grow using traditional culturing techniques is that they are metabolically inter-dependent upon other members of their microbial communities, and that these metabolic interdependencies are difficult to mimic and/or maintain in laboratory settings (Kantor *et al.*, 2013; He *et al.*, 2015). Therefore, identifying potential symbionts in natural communities could potentiate research into the development of such bacterial cultures. For example, one could design fluorescence in situ hybridization (FISH) probes for members of a specific candidate phylum, and then apply those as well as other broad range probes (e.g., for other entire phyla) to samples of natural communities to look for patterns of spatial co-occurrence which could be indicative of cell:cell interactions.

Given that novel phylogenetic diversity is correlated with novel functional diversity (Wu *et al.*, 2009), a recurring theme in this dissertation was the search for novel genetic or morphological features associated with dolphin oral communities. In Chapter three, two classes of genetic systems were identified which we predicted to have novel functional properties: a biosynthetic gene cluster and CRISPR-Cas9 systems. For example, one of the Cas9 protein we recovered had a 304 amino acid insertion in the Ruv-III domain, while two CRISPR-Cas9 systems from Saccharibacteria genomes appeared to be missing the Cas4 protein typically found

in CRISPR-Cas9 systems (whether the systems were still functional is unknown). These findings underscore the potential for the discovery of novel CRISPR-Cas systems in nature, some of which may have biotechnological applications and/or may change our understanding of the ways in which bacteria can interact with phage and plasmids in their environment. Interesting follow-up experiments to the research presented in Chapter three would be to genetically engineer culturable bacteria to express these novel CRISPR-Cas9 systems *in vitro* and observe whether they have novel functional capacity.

The use of metagenomics in describing the functional potential of novel organisms is dependent upon the transfer of annotations from previously studied, well characterized bacteria. Therefore to study properties that have evolved in and are unique to 'microbial dark matter', additional approaches are typically required. In Chapter three, I used microscopy to identify and characterize the morphology of unusual, rectangular cell-like units found in the dolphin mouth. One of their most intriguing features was pili-like appendages that protruded from the segments, which are most likely the units of individual cells. The pili-like appendages consisted of thick stalks or bunches of hair-like structures that splay out at the tips. Future studies determining the genetic basis of these appendages may provide insight into whether they constitute a new type of pili. Given that form often follows function, follow-up questions include what is their function and why do they have a different morphology/architecture from known pili? Many questions regarding the lifestyle of the rectangular cell-like units will likely require cultured isolates to answer. In the meantime, once their identity is conclusively determined, metabolic reconstructions based on their genome will provide insight into the ecology and evolution of this taxon.

5.3 Coda

Overall, this thesis provided insights into the diversity, form, and function of microbes associated with marine mammals. Marine mammals represent a textbook case of convergent evolution of mammals to a new environment (reviewed in Berta *et al.*, 2005), and therefore present an interesting opportunity to study the co-evolution of mammals and their microbiota. Furthermore, these species represent a relatively unexplored environment in terms of their microbiota, meaning that there is great potential to add foliage to the tree of life and explore the diversity of genetic systems and functions that have evolved in previously uncharacterized microbial lineages (Bik *et al.*, 2016; Dudek *et al.*, 2017). Unfortunately, 25-37% of marine mammal species are in danger of extinction (Davidson *et al.*, 2012), as is their microbiota. Practically speaking, research into the interaction between mammalian hosts and their microbiota may assist in future management of marine mammal populations and of sick and/or vulnerable animals. Such research is also an important step towards towards creating a comprehensive picture of the evolutionary history of life on Earth.

Appendix A

Additional material for 'Characterization of the gingival and gut microbiota of the Southern sea otter *(Enhydra lutris nereis)*'

A.1 Additional figures and tables

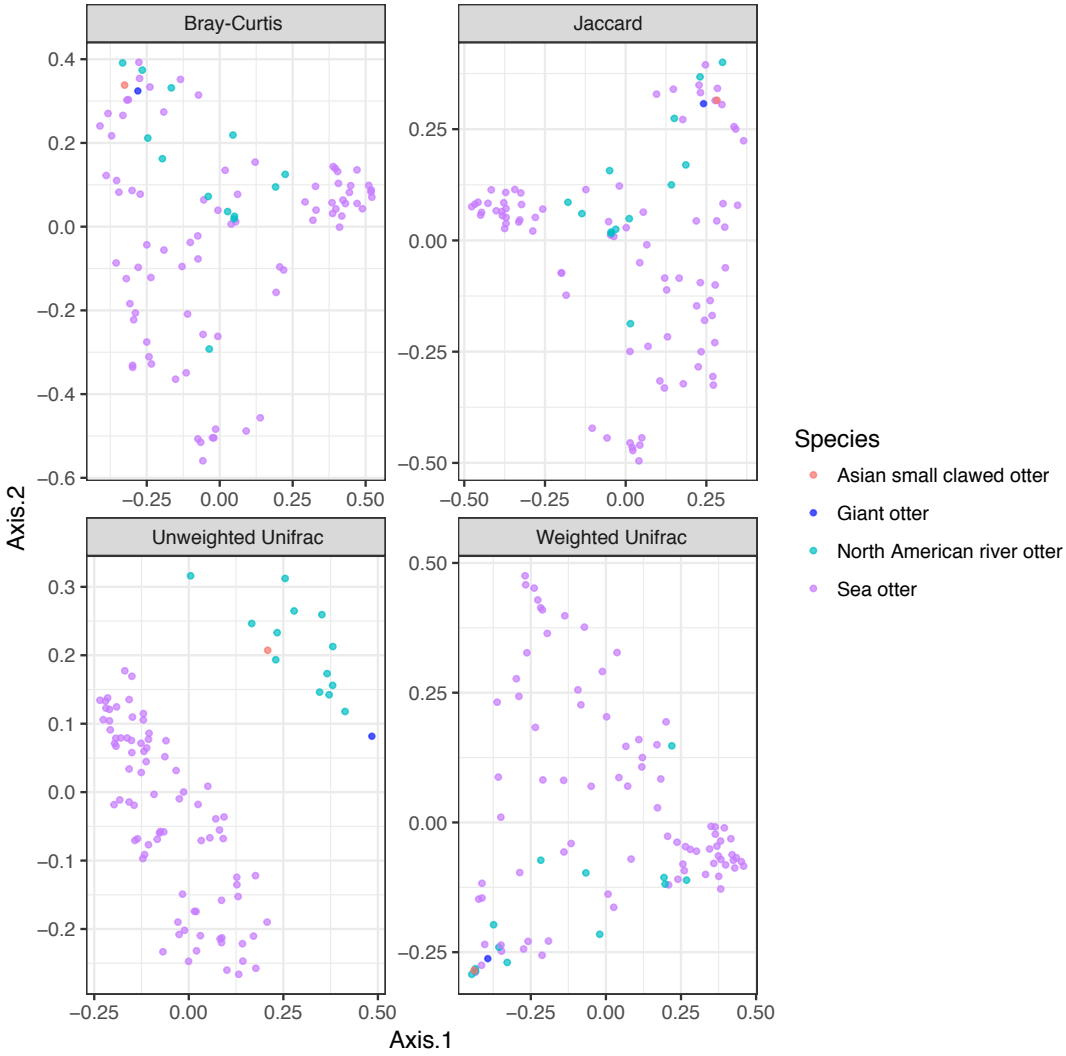


Figure A.1: Comparison of distal gut communities obtained from different otter species using multiple distance metrics. Principal coordinates analysis (PCoA) ordination of sea otter (n = 82), North American river otter (n = 13), Asian small-clawed otter (n = 1), and giant otter (n = 1) rectal microbiota composition based on 16S rRNA gene amplicon sequences. Taxa were collapsed at the genus level prior to comparison. Distance metrics used were Bray-Curtis, Jaccard, unweighted Unifrac, and weighted Unifrac.

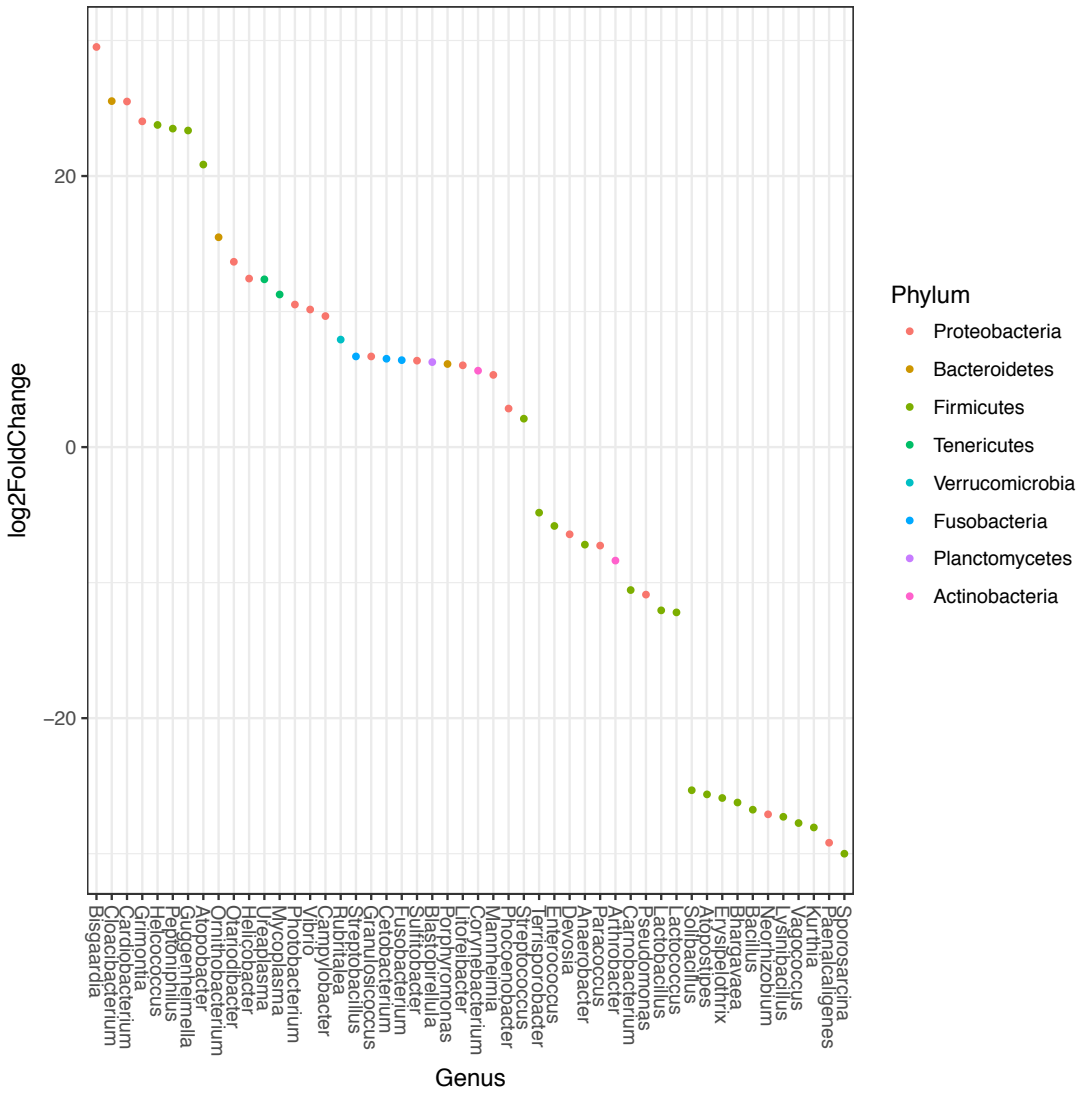


Figure A.2: Comparison of genera present in sea otter and North American river otter distal gut communities. Differential abundance testing revealed 50 genera that were differentially abundant in sea otter vs North American river otter distal gut communities. Genera with differentially abundant ASVs are shown on the x-axis. The extent to which each ASV was significantly different between host species is represented by its log2fold change on the y-axis. The more positive or more negative the log2fold change, the more differentially abundant a given genus between the two host species.

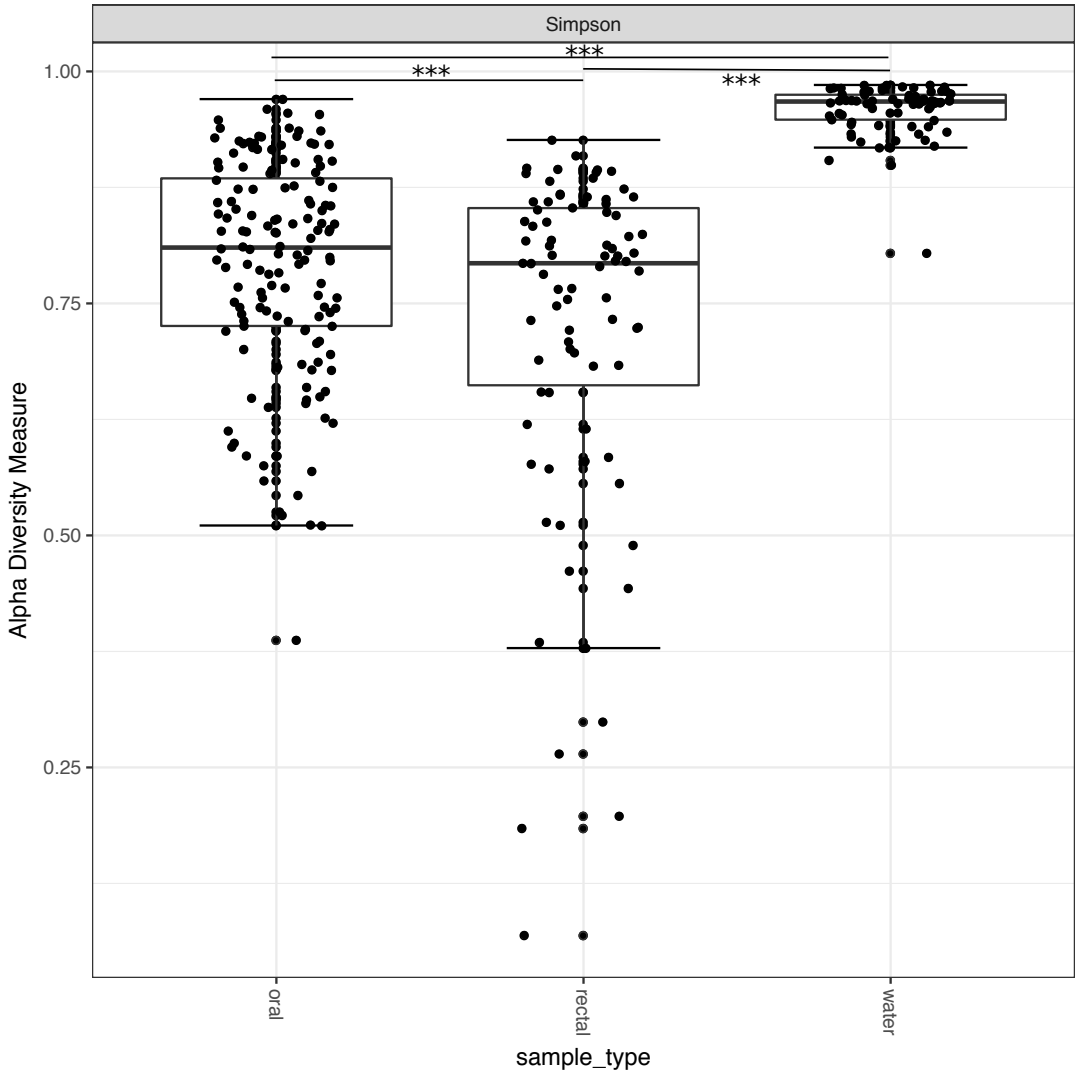


Figure A.3: Alpha diversity of gingival and rectal bacterial communities from wild sea otters, and adjacent seawater Simpson's diversity for gingival, rectal, and seawater samples. Bars denote significance. Three stars represent a p-value $<$ 0.001.

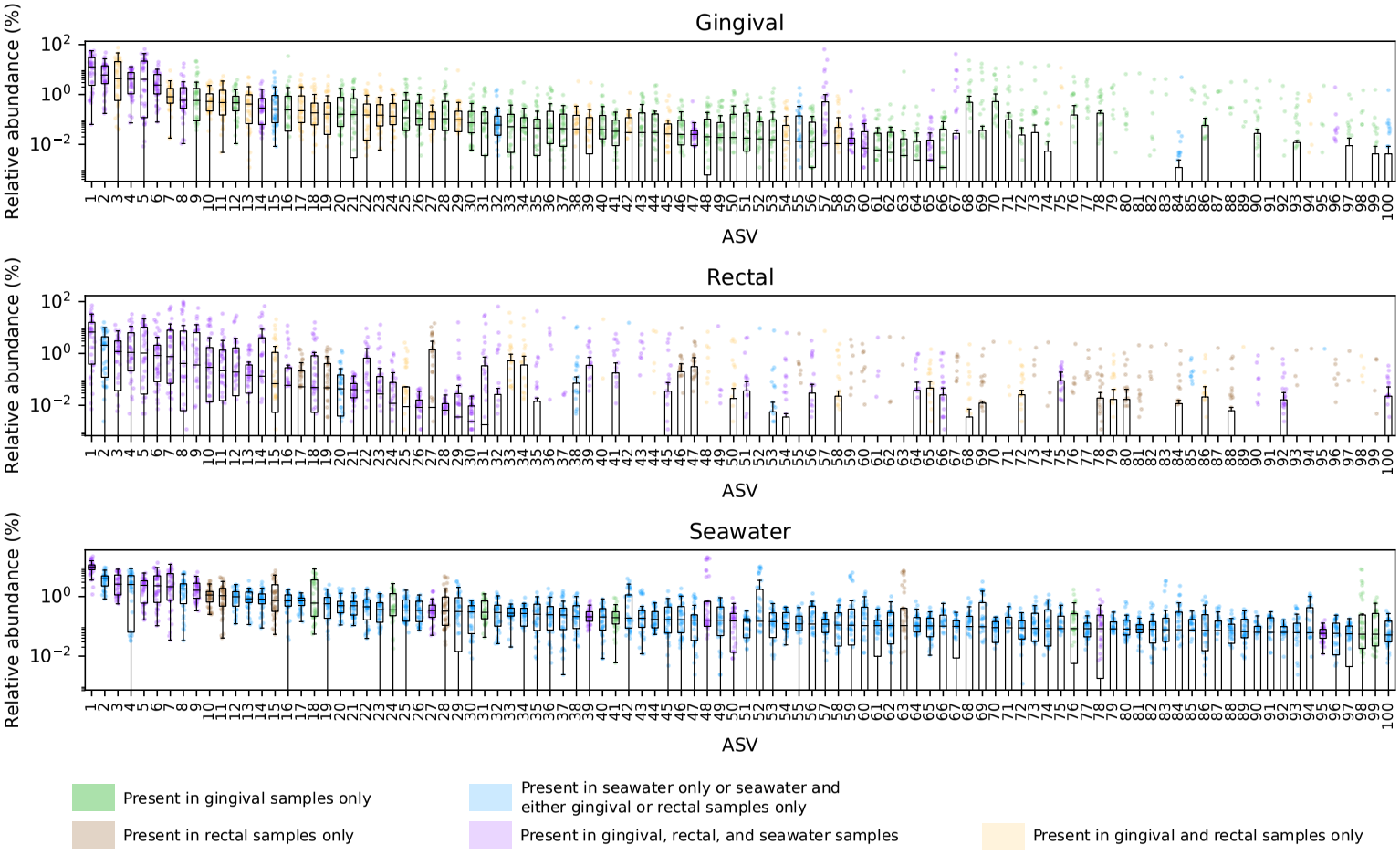

Gingival
Rectal
Seawater
Relative abundance (%)
ASV
Present in gingival samples only
Present in seawater only or seawater and either gingival or rectal samples only
Present in rectal samples only
Present in gingival, rectal, and seawater samples
Present in gingival and rectal samples only

Figure A.4: Relative abundance and overlap of ASVs between gingival, rectal, and seawater communities. ASVs from gingival, rectal, and seawater are plotted in order of decreasing median relative abundance in each environment. Gingival ASVs are colour coded based on whether they overlapped with seawater ASVs (note: overlap between gingival and rectal ASVs is not shown, based on the assumption that gingival ASVs are not seeded by rectal communities). Rectal and seawater ASVs are colour coded based on whether they are shared by any other environment.

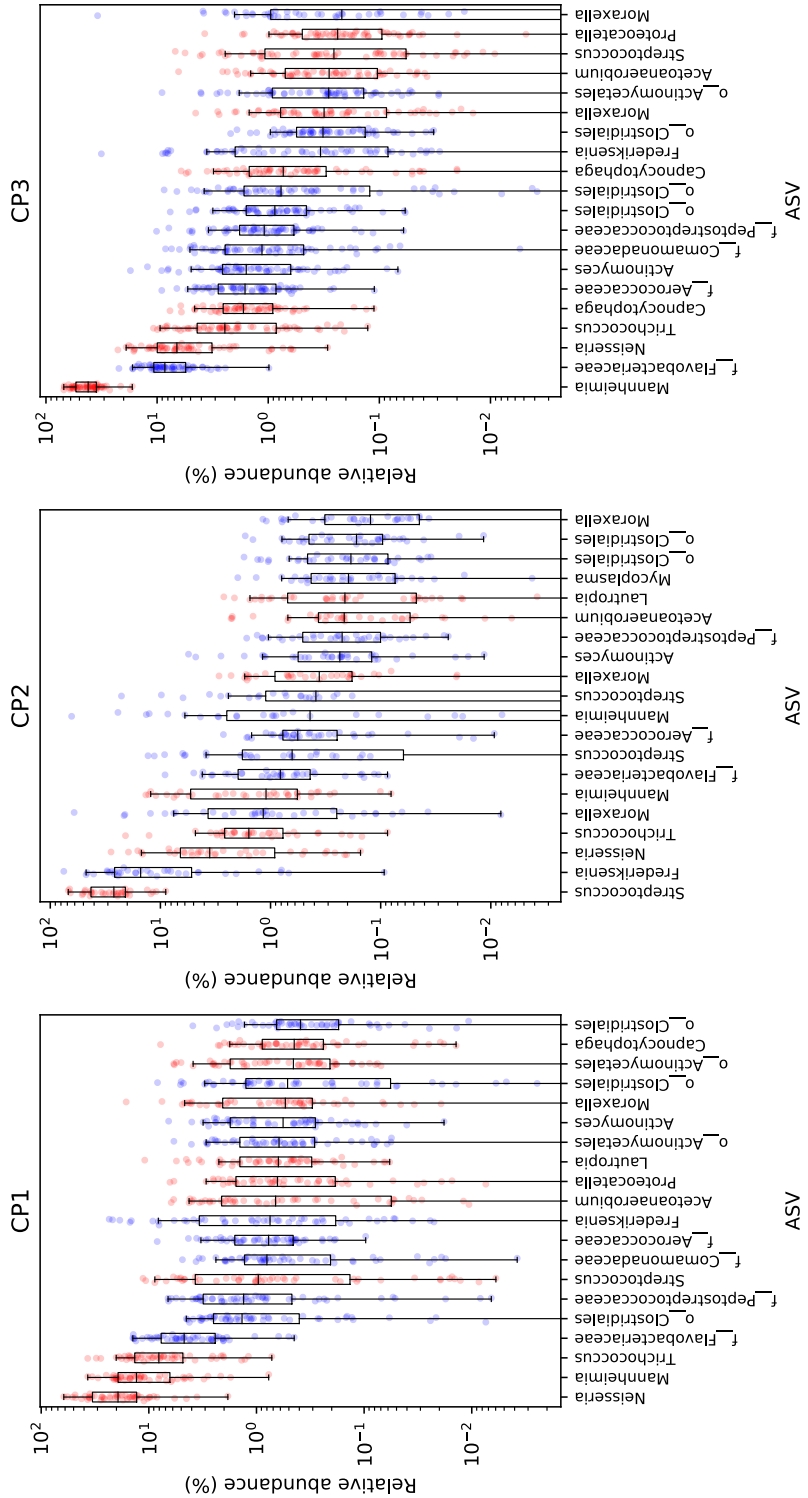

CP1
CP2
CP3
Relative abundance (%)
ASV
Neisseria
Mannheimia
Trichococcus
f__Flavobacteriaceae
o__Clostridiales
f__Peptostreptococcaceae
Streptococcus
f__Comamonadaceae
f__Aerococcaceae
Frederiksenia
Acetoanaerobium
Proteocatella
Lautropia
o__Actinomycetales
Actinomyces
Moraxella
o__Clostridiales
o__Actinomycetales
Capnocytophaga
o__Clostridiales
Myroplasma
Mannheimia
f__Flavobacteriaceae
Neisseria
Trichococcus
Capnocytophaga
f__Aerococcaceae
Actinomyces
f__Comamonadaceae
f__Peptostreptococcaceae
o__Clostridiales
o__Clostridiales
Capnocytophaga
Frederiksenia
o__Clostridiales
Moraxella
o__Actinomycetales
Acetoanaerobium
Streptococcus
Proteocatella
Moraxella

Figure A.5: Composition of gingival community profiles. The top 25 most prevalent ASVs in each CP are plotted along the x-axis, labeled at the genus level or lowest taxonomic level possible. ASVs are ordered by decreasing median abundance across all samples in a given CP. For each ASV, the relative abundance of that ASV in each sample is plotted on the y-axis. ASVs highlighted in red are those whose presence or absence was significantly different between CPs.

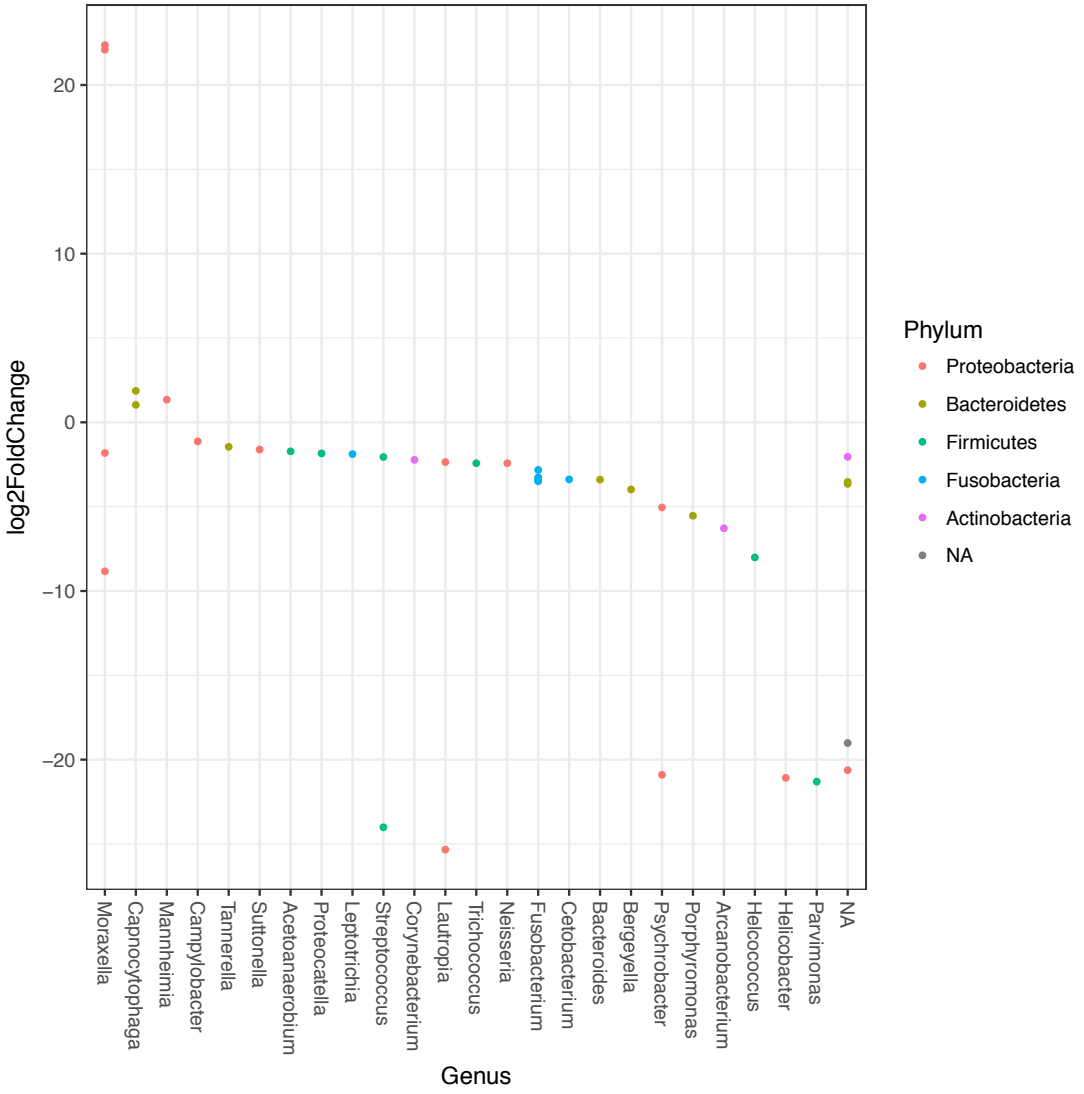


Figure A.6: ASVs are differentially abundant between sea otter gingival CPs. 40 ASVs were identified as differentially abundant between gingival CPs. Genera are represented on the x-axis, with dots representing each ASV within a given Genus that differed between CPs. The extent to which each ASV was significantly different between CPs is represented by its log2fold difference on the y-axis. The more positive or more negative the log2fold difference, the more differentially abundant the ASV between CPs.

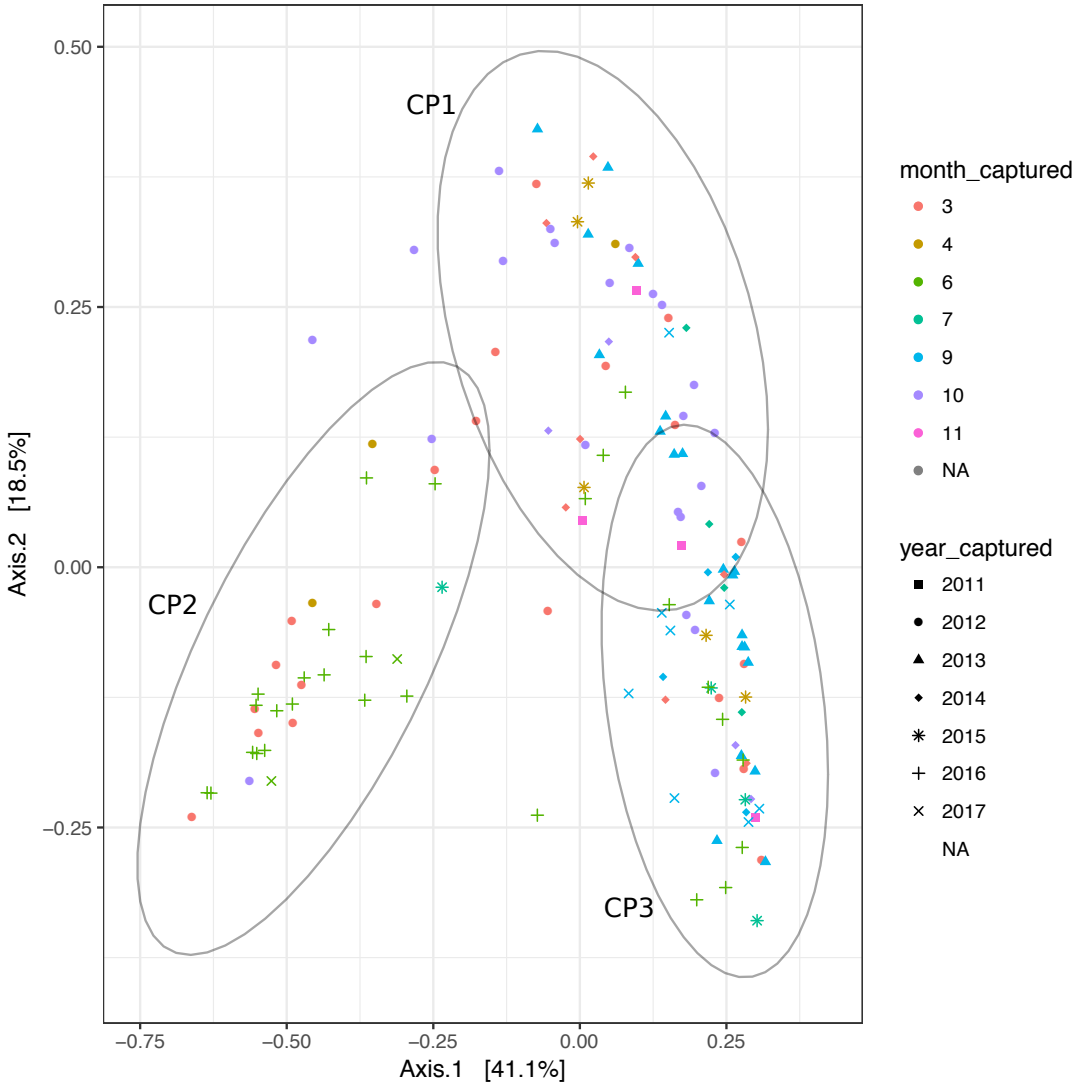


Figure A.7: Principal coordinates analysis (PCoA) of sea otter gingival samples, calculated with the Bray-Curtis distance metric. Colours indicate the month in which samples were obtained while shapes indicate the year in which samples were obtained. CPs groupings are indicated with 95% confidence ellipse.

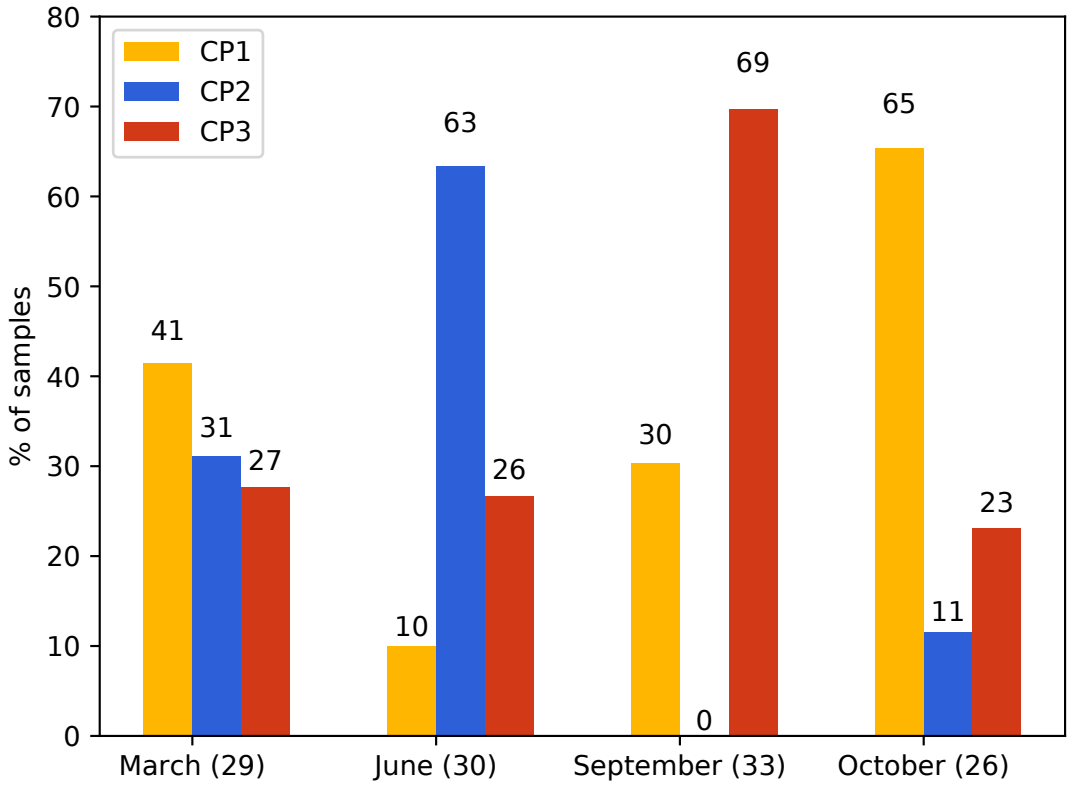


Figure A.8: Barplots indicate the percentage of samples from each month/set of years that were assigned to each CP. Only months in which ≥ 10 samples were collected are shown. The number of samples collected per month is denoted in brackets next to the name of the month.

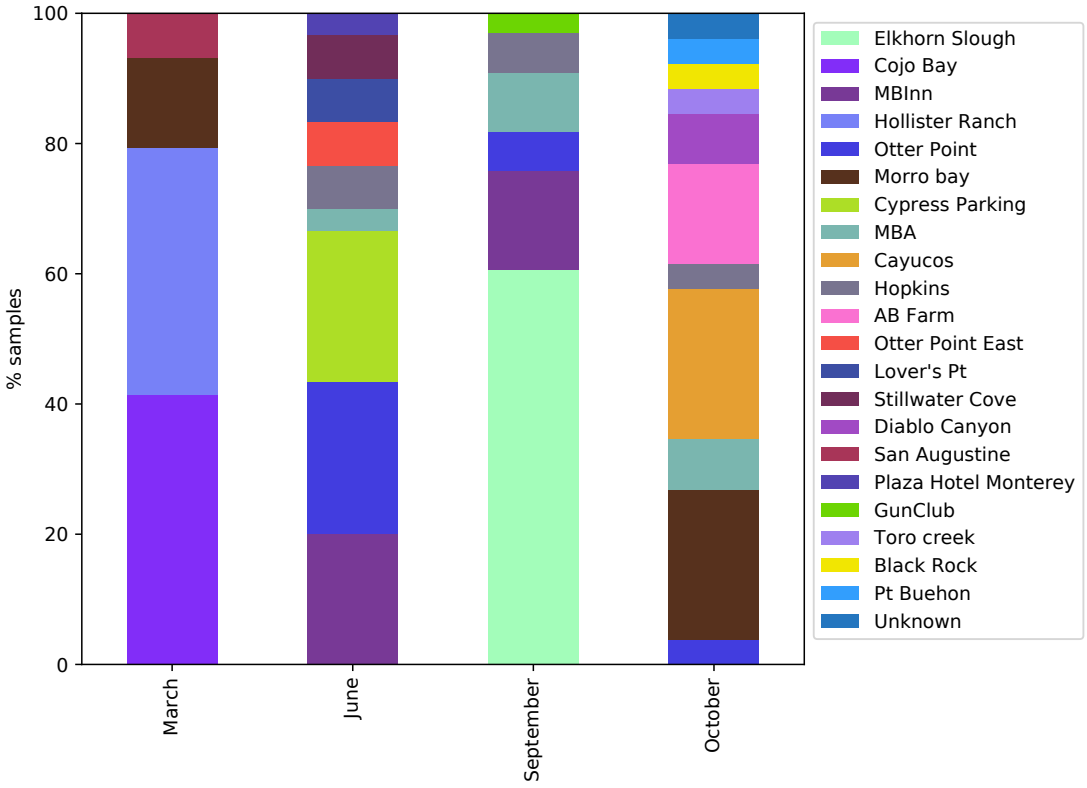


Figure A.9: For each month in which ≥ 10 samples were collected, the percentage of samples collected at each sampling location is shown.

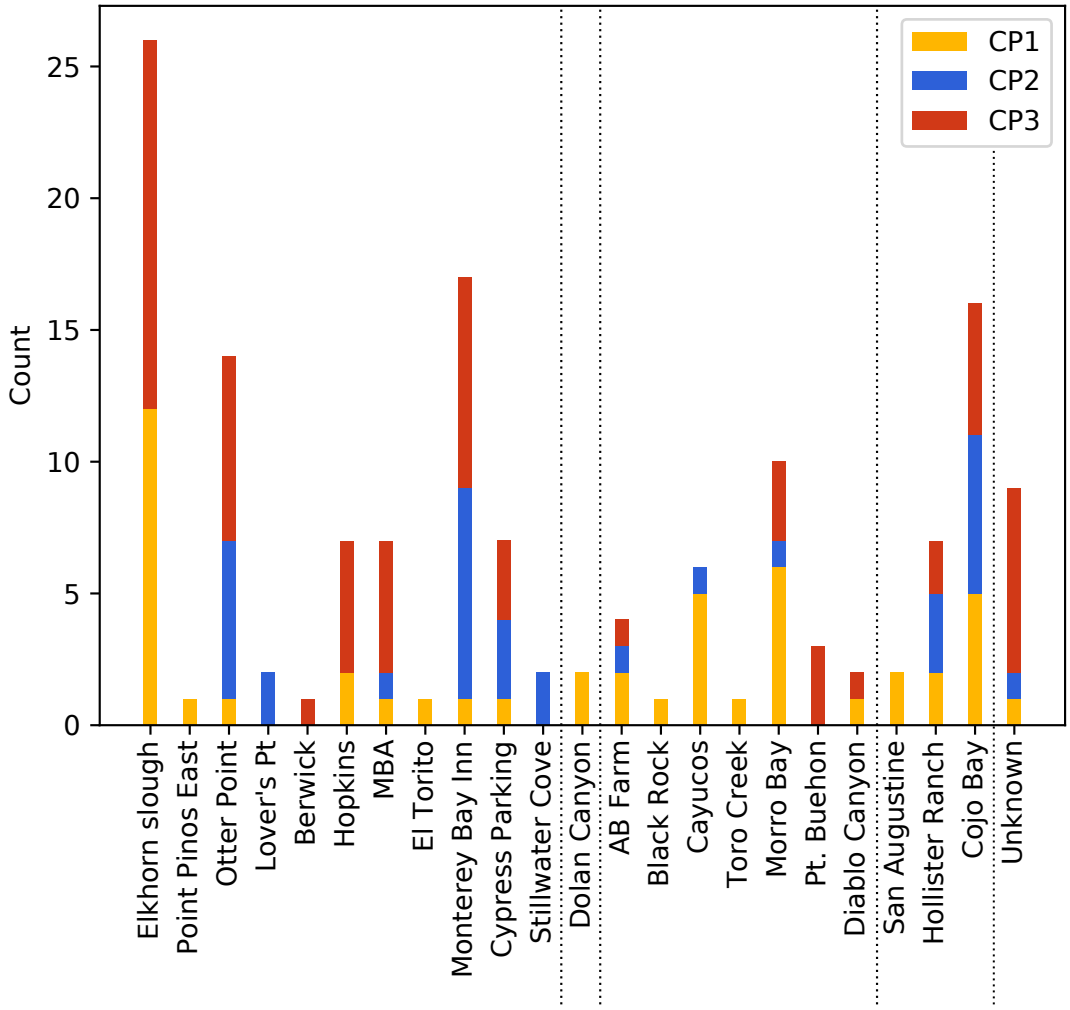


Figure A.10: For each sampling location, the number of samples assigned to CP1, CP2, or CP3 is shown. Locations are ordered along the x-axis by decreasing latitude (i.e. from North to South). Dotted lines denote groups of samples that are geographically close to one another as seen in Figure 1.

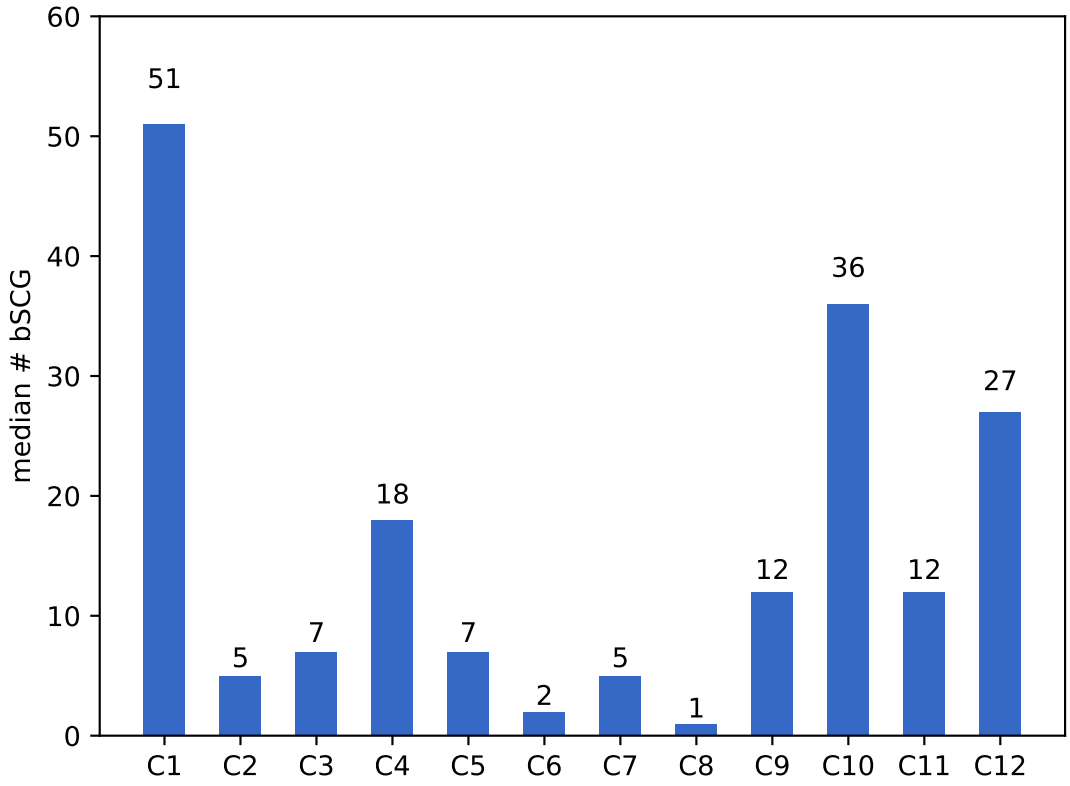


Figure A.11: Approximate number of bacterial genomes assembled per sea otter fecal metagenome. The median number of each bSCG assembled per sample was used as a proxy for the number of bacterial genomes assembled per sample. For example, if 5 bacterial genomes were sequenced, there should be 5 of each bSCG present and the median number of each bSCG present should be 5.

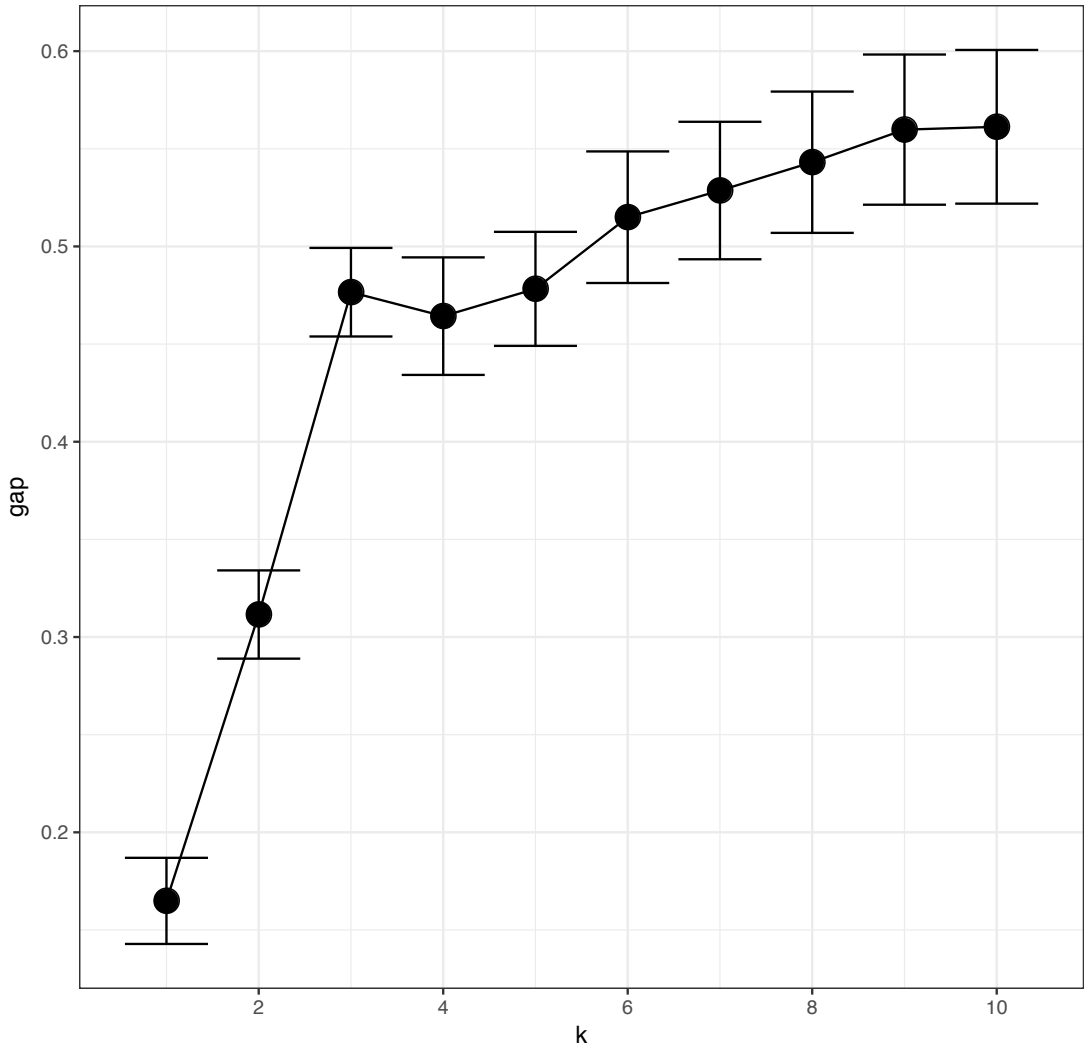


Figure A.12: Gap statistic for the sea otter gingival dataset. The gap statistic was used to estimate the number of clusters in the sea otter gingival dataset. The number of clusters is represented by "k".

Sample	Assembly estimate	Centrifuge estimate
C1	10203720	3802978
C2	183480	218013
C3	142560	44416
C4	4761300	1518824
C5	152400	22360
C6	29280	28458
C7	100080	150959
C8	0	20050
C9	975600	249172
C10	2921400	1055453
C11	855000	1337795
C12	6301500	1749776

Table A.1: Estimate of the number of bacterial reads using assembly-driven metagenomics vs read-based metagenomics. The number of bacterial reads per sample was estimated using two independent approaches. The first was based on assembly-driven metagenomics (see Methods) while the second consisted of using Centrifuge (Kim *et al.*, 2016), a read-based taxonomic classification algorithm, to identify bacterial reads.

gh19	Gut	Marine	gh20	Gut	Marine
fisher (1)	10	2	coyote (1)	30	3
hippo (1)	14	1	fisher (1)	10	6
mouse (1)	13	2	hippo (1)	56	2
rabbit (1)	12	1	rabbit (1)	17	2
humpback (1)	10	4	humpback (1)	58	6
sea otter (12)	95	43	sea otter (12)	239	52
right whale (6)	907	499	right whale (6)	2776	500
seiwhale (1)	78	23	seiwhale (1)	743	138
seawater (5)	1184	1646	seawater (5)	3519	1587

Table A.2: Number of marine-like vs gut-like glycoside hydrolase per mammalian species or seawater. Species with fewer than 10 chitin-degrading genes per glycoside hydrolase family are not shown. The number beside species/environment name is the number of individuals per species that were used in the metagenomic comparison.

	Lane 1		Lane 2		Lane 3	
Sample ID	Average (bp)	Range (bp)	Average (bp)	Range (bp)	Average (bp)	Range (bp)
C1	410	150-1121	-	-	-	-
C2	332	147-1000	-	-	-	-
C3	539	200-2739	-	-	-	-
C4	623	185-3841	236	165-429	-	-
C5	644	200-3871	291	176-549	-	-
C6	447	200-2797	-	-	-	-
C7	300	142-1000	-	-	-	-
C8	283	148-1000	-	-	-	-
C9	453	200-1820	-	-	-	-
C10	492	200-1681	-	-	-	-
C11	560	200-3301	-	-	-	-
C12	513	200-3271	-	-	459	169-3235

Table A.3: gDNA sizes for final libraries. Library statistics for each sample are shown. In some cases, different libraries were used for a given sample for different sequencing lanes. If new libraries were prepared, they were prepared for the given lane and used for all subsequent lanes.

Sample ID	Lane 1 (2 x 250 bp)	Lane 2 (2 x 250 bp)	Lane 3 (2 x 100bp)	Total # pairs	Total Gbp
C1	21,594,507	9,955,394	-	31,549,901	63.10
C2	9,343,780	10,863,960	-	20,207,740	40.42
C3	7,798,939	10,169,556	-	17,968,495	35.94
C4	9,942,344	12,541,891	29,703,245	52,187,480	104.37
C5	8,565,032	10,372,952	-	18,937,984	37.88
C6	11,950,764	10,653,681	-	22,604,445	45.21
C7	14,095,757	10,857,893	-	24,953,650	49.91
C8	9,253,267	12,325,571	-	21,578,838	43.16
C9	9,476,682	9,703,485	36,386,319	55,566,486	111.13
C10	19,828,106	10,736,445	73,266,417	103,830,968	207.66
C11	12,082,679	11,457,754	-	23,540,433	47.08
C12	17,668,918	13,976,502	14,976,981	46,622,401	93.24
Total	151,600,775	133,615,084	154,332,962	439,548,821	879.10

Table A.4: Amount of sequencing data generated per sample. The number of pairs of reads generated per sample is shown (i.e. the total number of reads per sample is 2X the number shown). The total number of bp sequenced is also shown in Gbp.

Appendix B

Additional material for 'Novel microbial diversity and functional potential in the marine mammal oral microbiome'

B.1 Additional figures and table

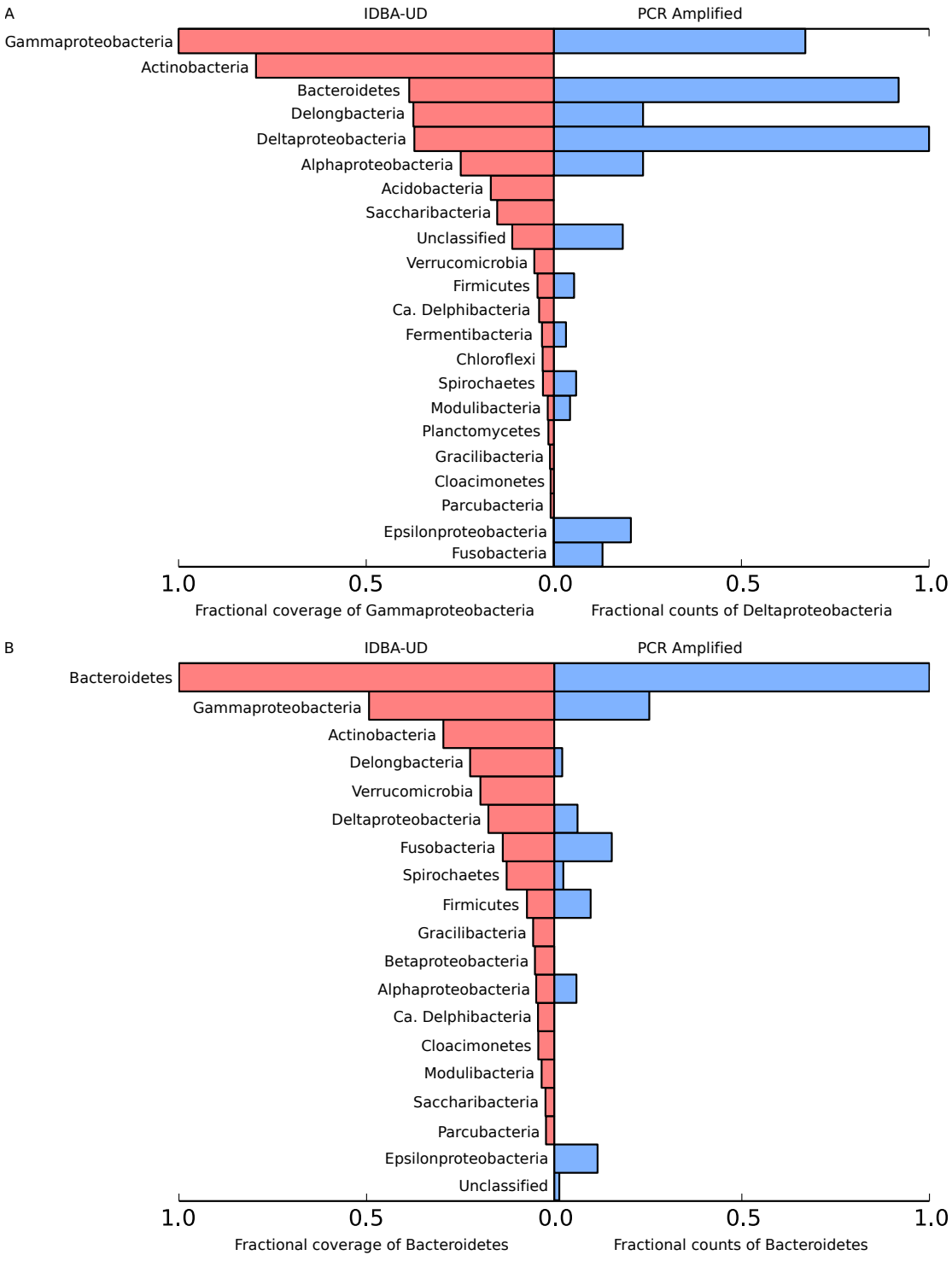

A
IDBA-UD
PCR Amplified
Gammaproteobacteria
Actinobacteria
Bacteroidetes
Delongbacteria
Deltaproteobacteria
Alphaproteobacteria
Acidobacteria
Saccharibacteria
Unclassified
Verrucomicrobia
Firmicutes
Ca. Delphibacteria
Fermentibacteria
Chloroflexi
Spirochaetes
Modulibacteria
Planctomycetes
Gracilibacteria
Cloacimonetes
Parcubacteria
Epsilonproteobacteria
Fusobacteria
1.0
0.5
0.0
0.5
1.0
Fractional coverage of Gammaproteobacteria
Fractional counts of Deltaproteobacteria
B
IDBA-UD
PCR Amplified
Bacteroidetes
Gammaproteobacteria
Actinobacteria
Delongbacteria
Verrucomicrobia
Deltaproteobacteria
Fusobacteria
Spirochaetes
Firmicutes
Gracilibacteria
Betaproteobacteria
Alphaproteobacteria
Ca. Delphibacteria
Cloacimonetes
Modulibacteria
Saccharibacteria
Parcubacteria
Epsilonproteobacteria
Unclassified
1.0
0.5
0.0
0.5
1.0
Fractional coverage of Bacteroidetes
Fractional counts of Bacteroidetes

Figure B.1: Comparison of the relative abundances of the most common 16S rRNA gene sequences identified with PCR and pyrosequencing (Bik *et al.*, 2016) with those assembled from Illumina paired-end reads using IDBA-UD (this work), for each of two dolphin oral metagenomes. Related to Figures 3.1 and 3.2. The rank abundance of the most abundant bacterial phyla is shown in descending order. Abundances of phyla were calculated using the top 50 most abundant 16S rRNA gene sequences from each study. Sequences assembled using the IDBA-UD assembly algorithm are shown in red on the left and sequences amplified in the PCR survey are shown in blue on the right. The x-axis represents relative abundance as a percentage of the most abundant phylum present; in other words, the relative abundance of each phylum was divided by the relative abundance of the most abundant phylum. The 'unclassified' group consists of 16S rRNA gene sequences without phylogenetic assignment. (A) Data from DolJOral78 metagenome. (B) Data from DolZOral124 metagenome.

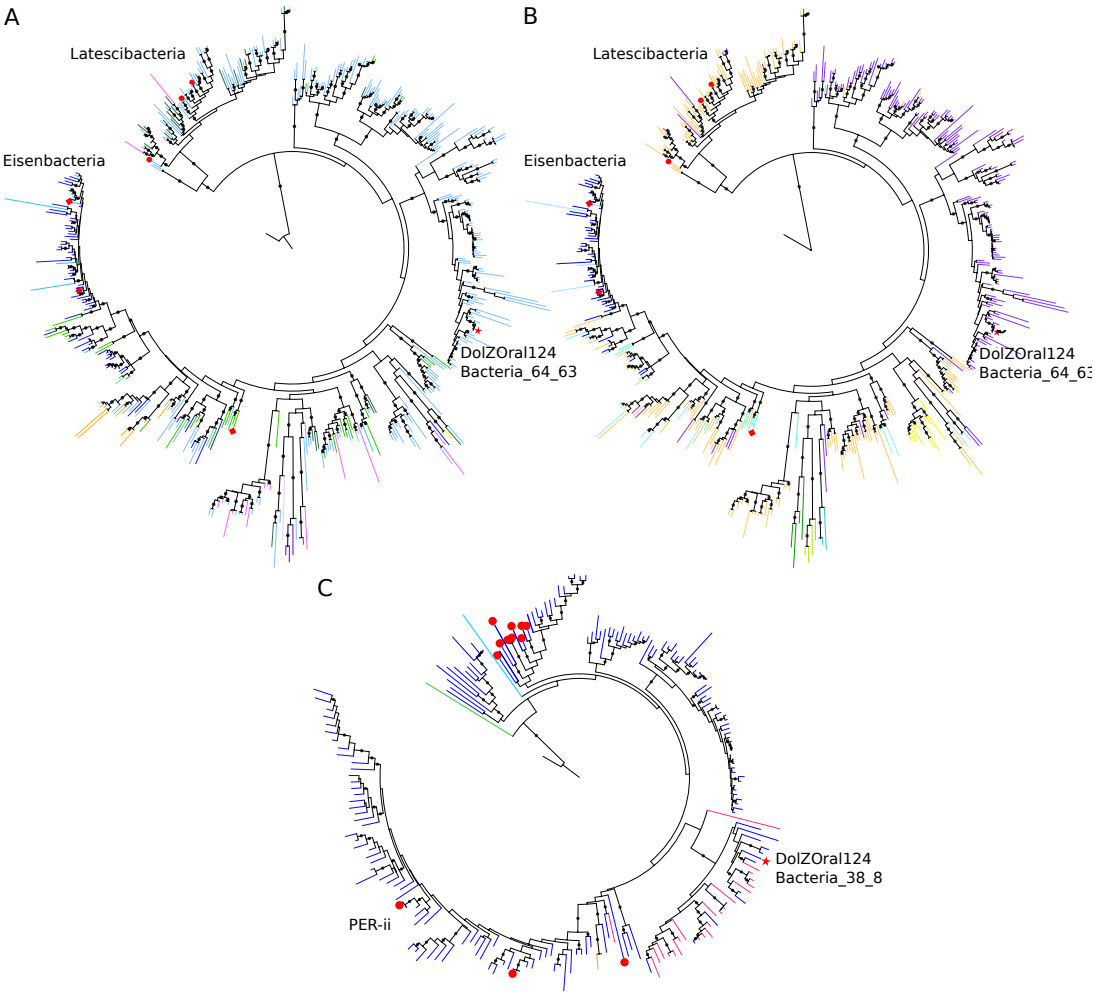

A
Latescibacteria
Eisenbacteria
DolZOral124
Bacteria_64_63
B
Latescibacteria
Eisenbacteria
DolZOral124
Bacteria_64_63
C
DolZOral124
Bacteria_38_8
PER-ii

Figure B.2: Maximum likelihood 16S rRNA gene phylogenies. Related to Figure 3.1 and Appendix B Data B.1. Branches are colour-coded according to clustering of 16S rRNA gene sequences based on percent identity using USEARCH (Edgar, 2010). Bootstrap support values $\geq 70\%$ are represented by closed black circles. (A) Latescibacteria group. Sequences were clustered into groups sharing $\geq 75\%$ sequence identity. These sequences consist of all those currently grouped in the Latescibacteria phylum in the SILVA NR Ref 99 database (Pruesse *et al.*, 2007; Quast *et al.*, 2013; Yilmaz *et al.*, 2014)), 16S rRNA gene sequences from Latescibacteria genome assemblies (red circles), Eisenbacteria genome assemblies (red diamonds), and the 16S rRNA gene sequence for the dolphin mouth lineage (red star). Sequences from the Fermentibacteria phylum were used as an outgroup (not shown). (B) Alternative clustering threshold for Latescibacteria group. Sequences were clustered into groups sharing $\geq 78.5\%$ sequence identity. The same set of sequences was used as in Appendix B Figure B.2A. (C) Peregrinibacteria group. Sequences were clustered into groups sharing $\geq 75\%$ sequence identity. These sequences consist of all those currently affiliated with the Peregrinibacteria phylum in the SILVA NR Ref 99 database (Pruesse *et al.*, 2007; Quast *et al.*, 2013; Yilmaz *et al.*, 2014), PER 16S rRNA sequences used in Hug *et al.* (2016) (red circles), and the 16S rRNA gene sequence for the dolphin mouth lineage (red star). Sequences from the Saccharibacteria phylum were used as an outgroup (not shown).

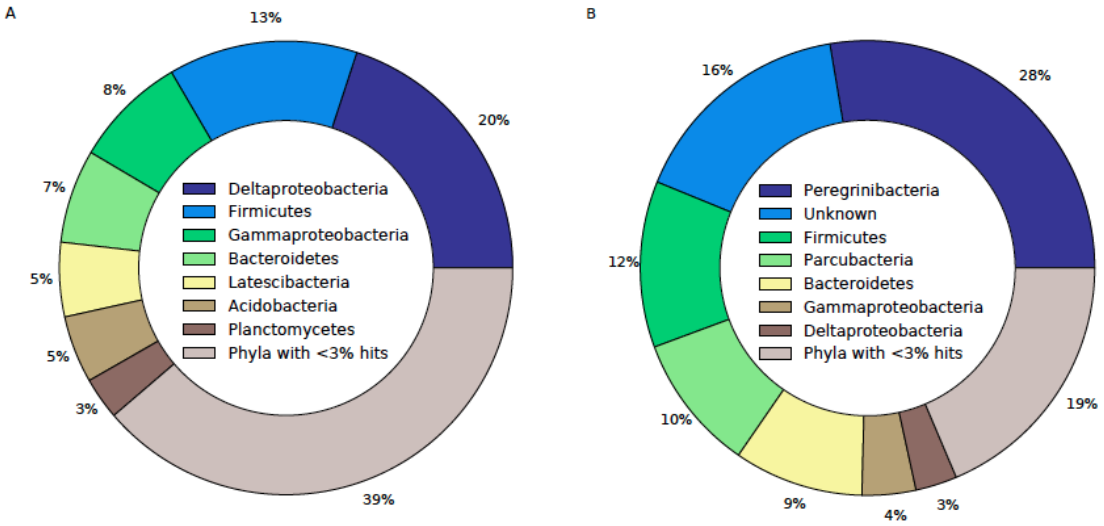


Figure B.3: Distribution of the taxonomic identity of top protein matches to the novel Fibrobacteres-Chlorobi-Bacteroidetes superphylum lineage and the CPR lineages. Related to Figures 3.1 and 3.3. Predicted ORFs from the DolZOral124_Bacteria_64_63 and DolZOral124_Bacteria_38_8 genomes were searched against the NCBI nr database (e-value $\leq$1e-10) using BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2008). The taxonomic affiliation of the top hit for each ORF was recorded. (A) DolZOral124_Bacteria_64_63 genome. Fractional representation for a total of 2,231 predicted proteins with significant hits. (B) DolZOral124_Bacteria_38_8 genome. Fractional representation for a total of 575 predicted proteins with significant hits.

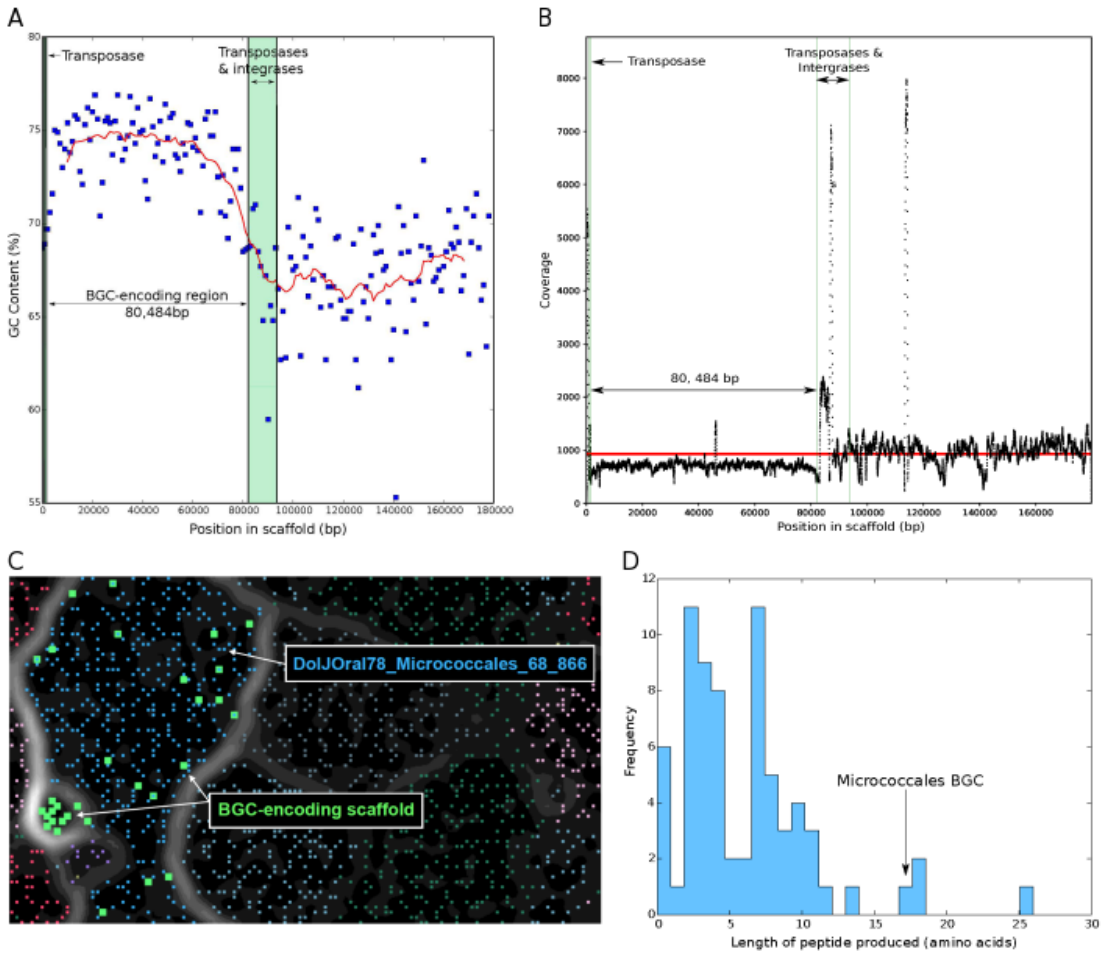
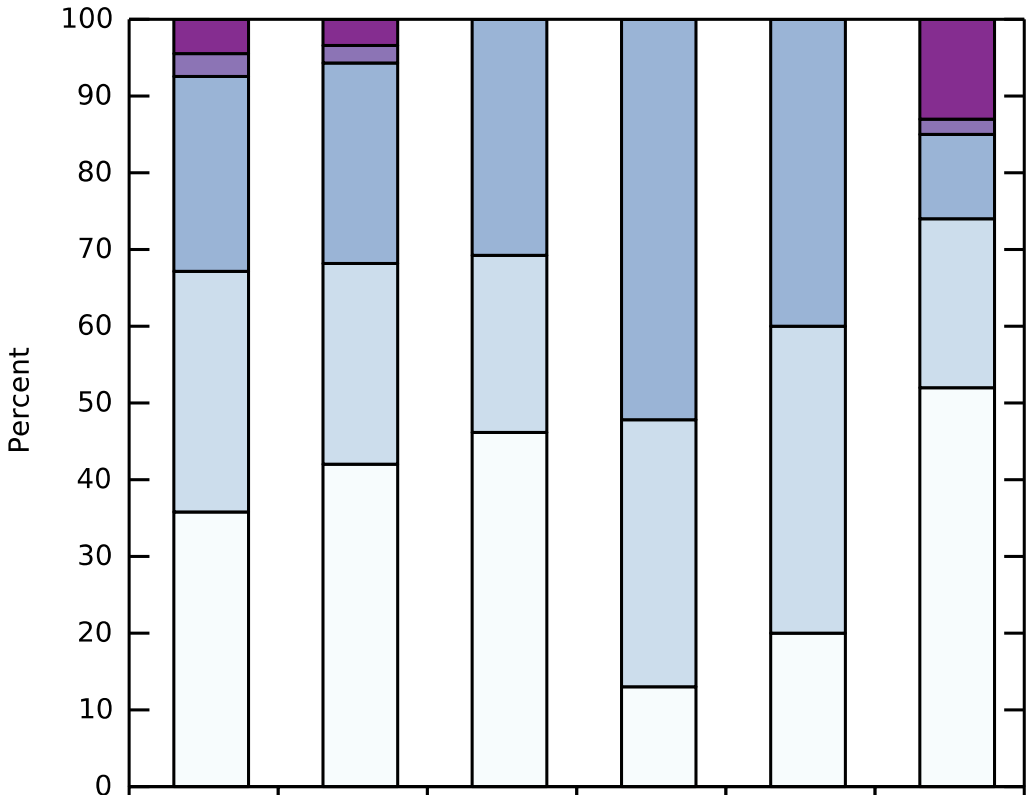


Figure B.4: Characterization of the BGC-encoding scaffold. Related to Figure 3.4. (A) GC content for the BGC-encoding region of the scaffold. GC content was calculated using a window size of 1000 bp. Regions encoding transposases and integrases are coloured green. (B) Tetranucleotide frequency of the BGC-encoding scaffold (green keys) and the Micrococcales genome (blue keys). (C) Coverage along the BGC-encoding region of the scaffold. Regions encoding transposases and integrases are bounded by green bars. (D) Distribution of non-ribosomally produced peptide product lengths from NRP synthesis clusters in the MIBiG database (Medema *et al.*, 2015).



	dolphin J	dolphin Z	seal	human A	human B	ref.
Type I	36% (24)	42% (37)	46% (6)	13% (3)	20% (1)	52%
Type III	31% (21)	26% (23)	23% (3)	35% (8)	40% (2)	22%
Type II	25% (17)	26% (23)	31% (4)	52% (12)	40% (2)	11%
Type IV & V	3% (2)	2% (2)	0% (0)	0% (0)	0% (0)	2%
Other	4% (3)	3% (3)	0% (0)	0% (0)	0% (0)	13%

Figure B.5: Distribution of CRISPR-Cas types across dolphin, harbour seal, and human oral environments. Related to Figure 3.5 and Appendix B Data B.3, B.4. The percentage of CRISPR-Cas types within each sample is shown. The 'reference' column is from Makarova *et al.* (2015) ,who surveyed all bacterial genomes available in NCBI databases as of February 2014. The table underneath the bar plot indicates the percentage of each CRISPR-Cas type within each sample. In brackets is the corresponding number of CRISPR-Cas operons of a given type that were assembled. For samples from this study, the operons in the 'other' group are those with both a *cas3* and a *cas10* gene that do not have the type I-D operon structure, which is the only defined subtype with both *cas3* and *cas10* components. For the Makarova *et al.* (2015) study, the "other" group consists of ambiguous or incomplete operons, 48% of which are incomplete type I operons and 25% of which are incomplete type III operons. See Appendix A Methods for information on harbour seal and human samples.

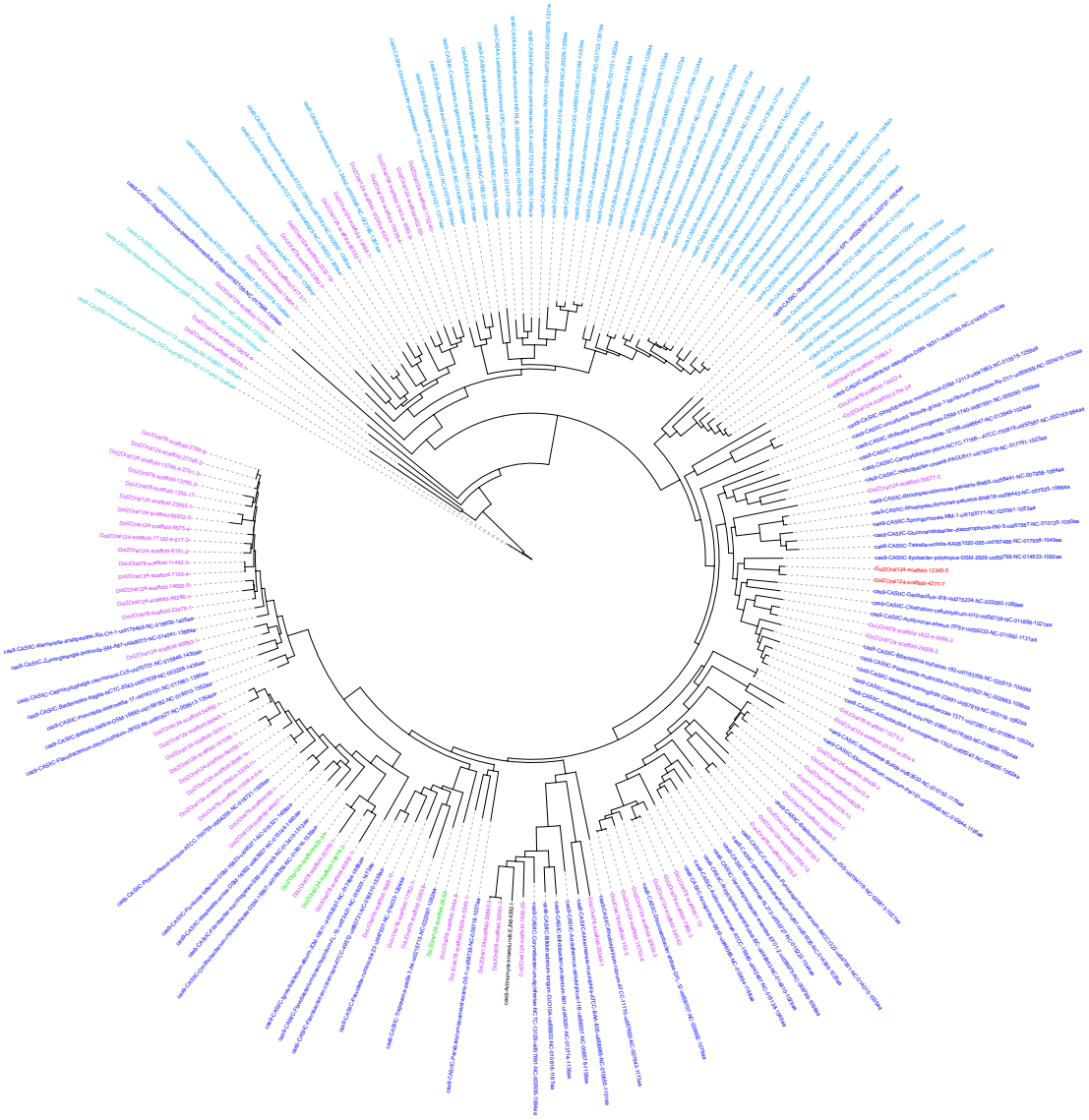


Figure B.6: Phylogeny of Cas9 proteins. Related to Figure 3.5 and Appendix B Data B.1, B.3. The tree includes Cas9 sequences from the dolphin oral microbiomes that are ≥ 800 amino acids, Cas9 proteins that have been classified as type II-A, II-B, or II-C (Makarova *et al.*, 2015), and AnaCas9 from *Actinomyces naeslundii*. Colour coding is as follows: orange-red, dolphin Saccharibacteria Cas9; green, one of the three 'unusual' dolphin Cas9 proteins; purple, all other dolphin Cas9 proteins; light blue, type II-A as defined in (Makarova *et al.*, 2015); blue-green, type II-B as defined in (Makarova *et al.*, 2015); dark blue, type II- C as defined in (Makarova *et al.*, 2015); black, AnaCas9. Bootstrap support values are shown.

	DolJOral78			DolZOral124				DolJOral78			DolZOral124		
Phylum	Amplicon	IDBA-UD	Genome	Amplicon	IDBA-UD	Genome	Phylum	Amplicon	IDBA-UD	Genome	Amplicon	IDBA-UD	Genome
Absconditabacteria	0	1	N	0	2	Y	Fusobacteria	5	0	N	14	1	Y
Acidobacteria	0	1	Y	0	1	Y	Gammaproteobacteria	10	7	Y	41	10	Y
Actinobacteria	0	5	Y	2	8	Y	**Gracilibacteria**	**1**	**1**	**Y**	**4**	**3**	**Y**
Alphaproteobacteria	7	3	Y	5	3	Y	Ignavibacteriae	-	-	-	0	1	Y
Aminicenantes	-	-	-	**1**	**1**	**N**	Microgenomates	-	-	-	0	1	N
Bacteroidetes	62	16	Y	90	23	Y	**Modulibacteria**	**2**	**1**	**Y**	**2**	**2**	**Y**
Betaproteobacteria	-	-	-	0	1	N	Parcubacteria	0	1	Y	2	5	Y
Chlorobi	1	0	N	3	2	N	Planctomycetes	1	1	Y	-	-	-
Chloroflexi	0	2	Y	0	1	Y	**Saccharibacteria**	0	1	Y	**1**	**5**	**Y**
Cloacimonetes	0	1	N	0	1	Y	Spirochaetes	10	4	Y	19	9	Y
Delongbacteria	**3**	**2**	**Y**	**4**	**1**	**Y**	Synergistetes	1	1	Y	3	3	N
Ca.Delphibacteria	**1**	**1**	**Y**	**1**	**1**	**Y**	Tenericutes	0	2	N	0	6	N
Deltaproteobacteria	27	11	Y	28	9	Y	Verrucomicrobia	0	1	Y	1	1	Y
Epsiloproteobacteria	8	2	Y	11	0	Y	Woesearchaeota	-	-	-	0	2	N
Fermentibacteria	**1**	**1**	**Y**	-	-	-	Unclassified Proteobacteria	3	1	-	6	4	-
Ca. Fertabacteria	-	-	-	0	1	Y	Unclassified Bacteria	21	5	-	49	1	-
Firmicutes	14	6	Y	27	16	Y	Total phyla	18	25	19	21	30	22

Table B.1: Comparison of phyla detected in a pyrosequencing amplicon survey (Bik *et al.*, 2016) versus those identified after assembly of Illumina paired-end reads with IDBA-UD (this work). Related to Figures 3.1 and 3.2. The number of unique amplicon OTUs/16S rRNA genes recovered from each phylum is shown, out of a total of 901 and 2,538 amplicons from DolJOral78 and DolZOral124, respectively. The final column indicates whether genomes of any level of completeness were recovered and identified from the corresponding phylum. Candidate phyla with relative abundances of $\geq 0.05\%$ in the amplicon dataset are shown in bold for each sample and grey shading highlights presence/absence discrepancies.

B.2 Additional Discussion

B.2.1 Naming novel lineages

For the novel FCB superphylum lineage, we propose the name *Candidatus* Delphibacteria in recognition of the first genomic representatives having been recovered from the dolphin (family *Delphinidae*) mouth and due to its ubiquity within dolphin mouths based on Bik *et al.* (2016). For the CPR lineage, we propose the name *Candidatus* Fertabacteria, where 'ferta' is the Latin word for 'tricky'. This name alludes to the multiple mismatches between the 16S rRNA gene sequence of the first genomic representative of this lineage and the commonly used PCR primers used for detection via 16S rRNA gene amplicon surveys.

B.2.2 Linking 16S rRNA genes to genomes from novel lineages

For the DolZOral124_Bacteria_64_63 genome (*Candidatus* Delphibacteria), our confidence in the 16S rRNA gene being correctly binned derives from two lines of evidence. First, the DolZOral124_Bacteria_64_63 16S rRNA gene is assembled on a 33.9 Kbp scaffold with 63% GC content, 62X coverage, and matching tetranucleotide frequency, providing strong support for inclusion within this bin. Second, there are only 5 other genomes in this sample with GC content ranging from 53% to 73% and coverage greater than 52X. These genomes fall within the Actinobacteria, Gammaproteobacteria, Verrucomicrobia, and Alphaproteobacteria phyla, all of which are well characterized and have highly divergent 16S rRNA gene sequences from the one in question. For the DolZOral124_Bacteria_38_8 genome (*Candidatus* Fertabacteria), our confidence in the 16S rRNA gene being correctly binned reflects its assembly on a scaffold that is 263 Kbp long and

whose inclusion within the bin is strongly supported based on 38% GC content, 9X coverage, and matching tetranucleotide frequency.

B.2.3 High proportion of type II CRISPR-Cas systems in the dolphin oral microbiome

Approximately 25% of all CRISPR-Cas systems in the dolphin oral microbiomes were type II systems, which are defined by the presence of a *cas9* gene. This is in contrast to a previous survey of all bacterial genomes available on NCBI as of February 2014, which found that only 13% of complete bacterial CRISPR-Cas systems are type II (Makarova *et al.*, 2015). To determine whether a high proportion of type II systems is a unique characteristic of the dolphin oral habitat or a more common feature of mammalian oral habitats, we compared the distribution of CRISPR-Cas types within samples from dolphin, harbour seal, and human mouths (Appendix B Figure B.5, Data B.3, B.4; also see Appendix B Methods). Across all three environments, type II systems were 25-52% of all CRISPR-Cas systems, suggesting that type II systems may be enriched in the mammalian oral microbiome when compared to bacterial genomes from diverse environments.

B.2.4 Additional information on the insertion in Cas9 from DolZOral124_scaffold_26_62

Five of the BLAST (Altschul *et al.*, 1990; Camacho *et al.*, 2008) matches to the 304 amino acid insertion in the Cas9 protein sequence from DolZOral124_scaffold_26_62 were from Bacteroidetes genomes and two were from unclassified scaffolds. Environments of origin range from aquifer groundwater to the intestinal tract of humans. We infer that the unbinned DolZOral124_scaffold_26 is from a Bac-

teroidetes genome, as 78% of predicted proteins on the 178,508 bp long scaffold had the highest homology to proteins found in other Bacteroidetes genomes.

In an attempt to infer the function of the DolZOral124_scaffold_26_62 RuvC-III insertion, we modeled its secondary structure using HHpred (Soding *et al.*, 2005) and Phyre2 (Kelley *et al.*, 2015). In both cases, the results consisted of very low confidence models (for HHpred, probability of 60%; for Phyre2, confidence of 33% over 9% coverage) matching the insertion sequence to the cytochrome C biogenesis protein CCME and the glycylpeptide n-tetradecanoyltransferase chain of n-myristoyltransferase, respectively. While these proteins are involved in different biologic processes, both act as transferases. None of the eight spacers in the DolZOral124_scaffold_26 target any scaffolds in the dolphin metagenomes.

B.2.5 Identity of Saccharibacteria genomes with type II CRISPR-Cas systems

A complete type II system was identified in each of the DolZOral124_Saccharibacteria_54_13_A and DolZOral124_Saccharibacteria_45_28 genomes. In addition, two partial *cas9* genes are encoded by the DolZOral124_Saccharibacteria_55_12_B genome, separated by a transposase. The two partial genes are complementary halves of a complete *cas9* gene. The *cas9* genes are adjacent to *cas1* and *cas2* genes and are on the same scaffold as a CRISPR array.

B.2.6 Analysis of spacer sequences from Saccharibacteria CRISPR arrays

Nine spacers from Saccharibacteria CRISPR arrays have a match to a scaffold in either dataset, using a threshold of $\geq$ 95% identity over 100% sequence length

or 100% identity over $\geq$ 95% sequence length, though only three matches are to scaffolds that are long enough to identify after extension with PRICE (Ruby *et al.*, 2013). Aside from the spacer from DolZOral124_Saccharibacteria_55_12_B that matches the phage genome, there are two more Saccharibacteria spacers with matches to long scaffolds (>5kb) in the dolphin metagenomes. They were assembled using CRASS (Skennerton *et al.*, 2013) and come from an array with a direct repeat sequence identical to that of the DolZOral124_TM7_54_13_A genome. One of the spacers matches two scaffolds. The first is a scaffold (DolZOral124_scaffold_1162980) which binned with the DolZOral124_Saccharibacteria_55_12_B genome, and the second is the DolZOral124_Phage_53_65 genome. To gain insight into which DNA sequence is more likely to be the in vivo target of the spacer, we searched for a PAM sequence in the flanking DNA of the four spacer matches from the DolZOral124_TM7_54_13_A genome: two of the scaffolds have an 'ACA' sequence five base pairs away from the spacer match. While the sample size is too small to have a high degree of confidence in the possible PAM sequence, the 'ACA' 3-mer is present in the DolZOral124_scaffold_1162980 spacer match flanking sequence but not in the DolZOral124_Phage_53_65 spacer match flanking sequence. The DolZOral124_scaffold_1162980 spacer match is in a region of DNA that encodes integrases and a transposase, as well as ribosomal proteins further upstream, suggesting that it may be a mobile element inserted into the DolZOral124_Saccharibacteria_55_12_B genome. The second spacer with the DolZOral124_TM7_54_13_A direct repeat sequence matches a scaffold within the DolZOral124_TM7_54_13_A genome. This scaffold does not have the 'ACA' 3-mer and therefore it is unclear whether this match might indicate targeting of self by the CRISPR-Cas system.

Appendix C

Additional material for 'Previously uncharacterized rectangular microbial units in the marine mammal oral cavity'

C.1 Additional figures

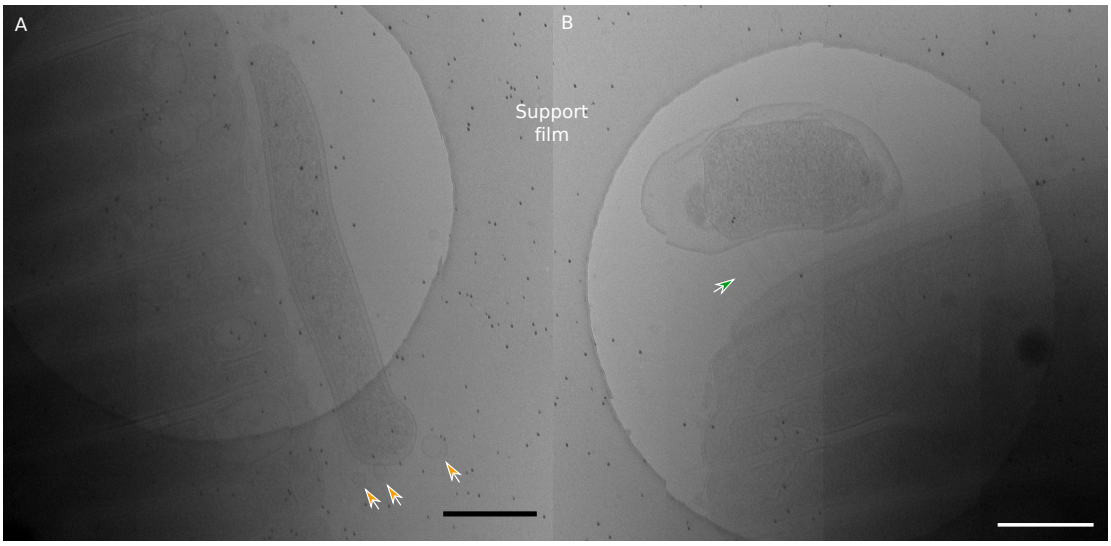


Figure C.1: Cryo-TEM image documenting rectangular cell-like unit in association with other cells. A rectangular cell-like unit (same one in both images) was in close proximity to other cells. (A) Small bubble-like structures were seen near the exterior of both cells; for example, see orange arrows. No pili-like appendages were seen. (B) A cell was potentially connected to the rectangular cell-like unit via pili-like appendages; see green arrow. The cell appeared to be shriveled compared to morphologically similar ones seen in Figure 4. Scale bars: 500 nm.

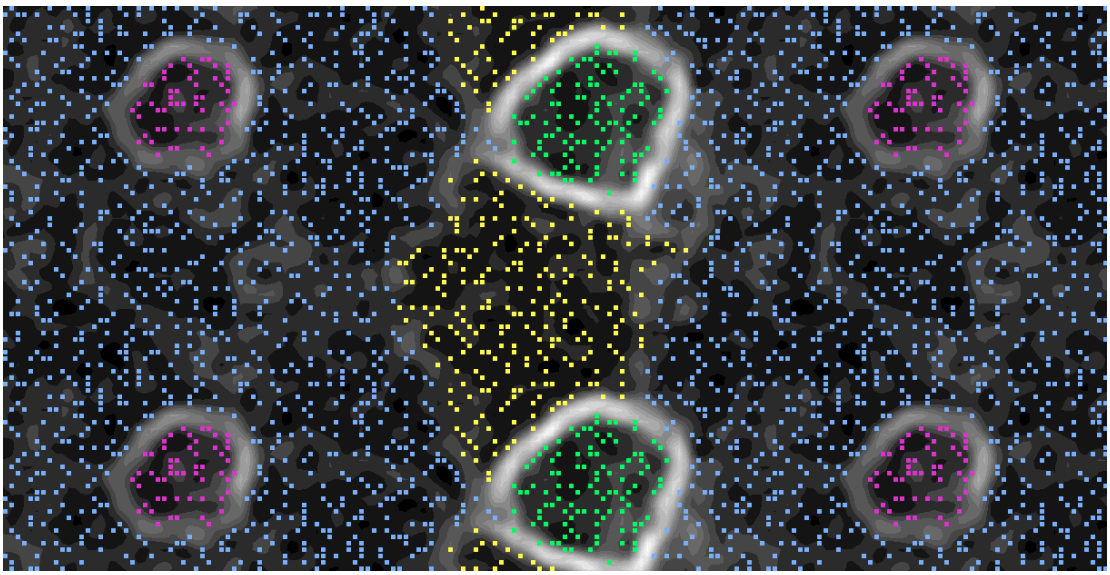


Figure C.2: Tetranucleotide ESOM of genomic material recovered the single-cell genomics experiment. Scaffolds $\geq$ 5kb were binned with a tetranucleotide ESOM over windows of 5 kb. The ESOM was created with tiling mode enabled (i.e. the same tile / image repeats). Colours of keys on the map indicate user-defined genome bins as follows: Bin 1 = blue, Bin 2 = yellow, Bin 3 = green, Bin 4 = pink.

List of supplemental data files

Supplemental Data File A.1: Sea otter metagenome statistics using different assembly parameters. Assembly statistics for sea otter metagenomes assembled using different Megahit (Li et al., 2015, Li et al., 2016) assembly parameters. Megahit parameters modified consisted of k-min, k-max, and k-step. Statistics used to assess assembly quality were the number of contigs generated, the number of these contigs that were ≥1kbp, and the number of three bacterial single copy genes that were assembled on contigs ≥1kbp (ribosomal proteins S3, S10, L24).

Supplemental Data File B.1: Phylogenetic trees in Newick format and the corresponding alignments.

Supplemental Data File B.2: Genome overview of DolZOral124_Bacteria_64_63 (Delphibacteria).

Supplemental Data File B.3: Cas operons recovered from the dolphin oral microbiome. Cas and accessory proteins (such as DinG) were detected using the HMMer suite version 3.1b2 (Eddy, 2011), e-value = 0.01. Tabs labeled "hits" contain ORFs identified as putative members of Cas loci, whereas tabs labeled

"loci" contain sets of these ORFs grouped into Cas loci.

Supplemental Data File B.4: CRISPR arrays detected by CRISPERFinder (Grissa *et al.*, 2007) and CRASS (Skennerton *et al.*, 2013).

Supplemental Data File B.5: Genome overview of Saccharibacteria-infecting phage, DolZOral124_Phage_53_65.

Bibliography

[1] Abdelrhman, K.F.A, *et al.* 'A first insight into the gut microbiota of the sea turtle Caretta caretta.' *Frontiers in microbiology* 7 (2016): 1060.

[2] Alam, M, *et al.* 'Flagella and motility behaviour of square bacteria.' *The EMBO journal* 3.12 (1984): 2899-2903.

[3] Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., Nielsen, P.H. 'Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes.' *Nature biotechnology* 31.6 (2013): 533-538.

[4] Altschul, Stephen F., *et al.* 'Basic local alignment search tool.' *Journal of molecular biology* 215.3 (1990): 403-410.

[5] Anantharaman, K., *et al.* 'Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system.' *Nature communications* 7 (2016): 13219.

[6] Andersson, G. E., *et al.* 'On the origin of mitochondria: a genomics perspective.' *Philosophical Transactions of the Royal Society B: Biological Sciences* 358.1429 (2003): 165-179.

[7] Apprill, A., *et al.* 'Humpback whales harbour a combination of specific and variable skin bacteria.' *Environmental microbiology reports* 3.2 (2011): 223-232.

[8] Baker, B.J., *et al.* 'Enigmatic, ultrasmall, uncultivated Archaea.' *Proceedings of the National Academy of Sciences* 107.19 (2010): 8806-8811.

[9] Bai, Y., *et al.* 'Genomic comparison of chitinolytic enzyme systems from terrestrial and aquatic bacteria.' *Environmental microbiology* 18.1 (2016): 38-49.

[10] Bankevich, A., *et al.* 'SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.' *Journal of computational biology* 19.5 (2012): 455-477.

[11] Berta, A., Sumich, J.L., Kovacs, K.M. Marine mammals: evolutionary biology. *Elsevier*, 2005.

[12] Blin, K., *et al.* 'antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers.' *Nucleic acids research* 41.W1 (2013): W204-W212.

[13] Bik, E.M., *et al.* 'Marine mammals harbor unique microbiotas shaped by and yet distinct from the sea.' *Nature communications* 7 (2016): 10516.

[14] Bowen, W. D. 'Role of marine mammals in aquatic ecosystems.' *Marine Ecology Progress Series* (1997): 267-274.

[15] Braak, T., Cajo, J.F. 'Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis.' *Ecology* 67.5 (1986): 1167-1179.

[16] Brinig, M.M., *et al.* 'Prevalence of bacteria of division TM7 in human subgingival plaque and their association with disease.' *Applied and environmental microbiology* 69.3 (2003): 1687-1694.

[17] Brister, J.R., *et al.* 'NCBI viral genomes resource.' *Nucleic acids research* 43.D1 (2014): D571-D577.

[18] Broberg, P. 'SAGx: Statistical Analysis of the GeneChip.' *R package version 1.0* (2009): 2010.

[19] Brown, C.T., *et al.* 'Unusual biology across a group comprising more than 15% of domain Bacteria.' *Nature* 523.7559 (2015): 208.

[20] Brown, L., *et al.* 'Through the wall: extracellular vesicles in Gram-positive bacteria, mycobacteria and fungi.' *Nature Reviews Microbiology* 13.10 (2015): 620.

[21] Bruger, E., Waters, C. 'Sharing the sandbox: evolutionary mechanisms that maintain bacterial cooperation.' *F1000Research* 4 (2015).

[22] Buffie, C.G., Pamer, E.G. 'Microbiota-mediated colonization resistance against intestinal pathogens.' *Nature Reviews Immunology* 13.11 (2013): 790.

[23] Burns, D.G., *et al.* '*Haloquadratum walsbyi* gen. nov., sp. nov., the square haloarchaeon of Walsby, isolated from saltern crystallizers in Australia and Spain.' *International Journal of Systematic and Evolutionary Microbiology* 57.2 (2007): 387-392.

[24] Burstein, D., *et al.* 'Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems.' *Nature communications* 7 (2016): ncomms10613.

[25] Burstein, D., *et al.* 'New CRISPR–Cas systems from uncultivated microbes.' *Nature* 542.7640 (2017): 237.

[26] Callahan, B.J., *et al.* 'DADA2: high-resolution sample inference from Illumina amplicon data.' *Nature methods* 13.7 (2016): 581.

[27] Camacho, C., *et al.* 'BLAST plus :architecture and applications.' *BMC bioinformatics* 10.1 (2009):421.

[28] Campbell, J.H., *et al.* 'UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota.' *Proceedings of the National Academy of Sciences* 110.14 (2013): 5540-5545.

[29] Caporaso, J.G., *et al.* 'QIIME allows analysis of high-throughput community sequencing data.' *Nature methods* 7.5 (2010): 335.

[30] Cardoso, A.M., *et al.* 'Metagenomic analysis of the microbiota from the crop of an invasive snail reveals a rich reservoir of novel genes.' *PLoS One* 7.11 (2012): e48505.

[31] Castelle, C.J., *et al.* 'Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling.' *Current biology* 25.6 (2015): 690-701.

[32] Castelle, C.J., Banfield, J.F. 'Major new microbial groups expand diversity and alter our understanding of the tree of life.' *Cell* 172.6 (2018): 1181-1197.

[33] Chinn, S.M., *et al.* 'The high cost of motherhood: end-lactation syndrome in southern sea otters (Enhydra lutris nereis) on the Central California Coast, USA.' *Journal of wildlife diseases* 52.2 (2016): 307-318.

[34] Cho, I., *et al.* 'Antibiotics in early life alter the murine colonic microbiome and adiposity.' *Nature* 488.7413 (2012): 621.

[35] Church, D.M., *et al.* 'Lineage-specific biology revealed by a finished genome assembly of the mouse.' *PLoS biology* 7.5 (2009): e1000112.

[36] Cole, J.R., *et al.* 'Ribosomal Database Project: data and tools for high throughput rRNA analysis.' *Nucleic acids research* 42.D1 (2013): D633-D642.

[37] Comolli, L.R., Banfield, J.F. 'Inter-species interconnections in acid mine drainage microbial communities.' *Frontiers in microbiology* 5 (2014): 367.

[38] Costa, D.P. 'Energy, nitrogen, and electrolyte flux and sea water drinking in the sea otter *Enhydra lutris*.' *Physiological Zoology* 55.1 (1982): 35-44.

[39] Costa, D. P. 'Marine mammal energetics.' *The biology of marine mammals*. Smithsonian Institution Press: Washington, DC (1999): 176-217.

[40] Costa, D.P., Kooyman, G.L. 'Oxygen consumption, thermoregulation, and the effect of fur oiling and washing on the sea otter, *Enhydra lutris*.' *Canadian Journal of Zoology* 60.11 (1982): 2761-2767.

[41] Costa, D.P., Kooyman, G.L. 'Contribution of specific dynamic action to heat balance and thermoregulation in the sea otter *Enhydra lutris*.' *Physiological Zoology* 57.2 (1984): 199-203.

[42] Darriba, D., *et al.* 'ProtTest 3: fast selection of best-fit models of protein evolution.' *Bioinformatics* 27.8 (2011): 1164-1165.

[43] Davidson, A.D., *et al.* 'Drivers and hotspots of extinction risk in marine mammals.' *Proceedings of the National Academy of Sciences* 109.9 (2012): 3395-3400.

[44] Davis, N.M., *et al.* 'Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data.' *bioRxiv* (2017): 221499.

[45] Davison, N.J., *et al.* '*Helicobacter cetorum* infection in striped dolphin (*Stenella coeruleoalba*), Atlantic white-sided dolphin (*Lagenorhynchus acutus*), and short-beaked common dolphin (*Delphinus delphus*) from the southwest coast of England.' *Journal of wildlife diseases* 50.3 (2014): 431-437.

[46] de Duve, C., Osborn, M.J. 'Size limits of very small microorganisms: Proceedings of a workshop'. *National Academies Press*, 1999.

[47] Delport, T.C., *et al.* 'Colony location and captivity influence the gut microbial community composition of the Australian sea lion (*Neophoca cinerea*).' *Applied and environmental microbiology* 82.12 (2016): 3440-3449.

[48] DeSantis, T.Z., *et al.* 'Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.' *Applied and environmental microbiology* 72.7 (2006): 5069-5072.

[49] Dethlefsen, L., McFall-Ngai, M., Relman, D.A. 'An ecological and evolutionary perspective on human–microbe mutualism and disease.' *Nature* 449.7164 (2007): 811.

[50] Dick, G.J., *et al.* 'Community-wide analysis of microbial genome sequence signatures.' *Genome biology* 10.8 (2009): R85.

[51] DiGiulio, D.B., *et al.* 'Temporal and spatial variation of the human microbiota during pregnancy.' *Proceedings of the National Academy of Sciences* 112.35 (2015): 11060-11065.

[52] Dobro, M.J., *et al.* 'Uncharacterized bacterial structures revealed by electron cryotomography.' *Journal of bacteriology* 199.17 (2017): e00100-17.

[53] Donia, M.S., *et al.* 'A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics.' *Cell* 158.6 (2014): 1402-1414.

[54] Doroff, A., Burdin, A. 2015. '*Enhydra lutris*.' *The IUCN Red List of Threatened Species 2015*: e.T7750A21939518.

[55] Dudek, N.K., *et al.* 'Novel microbial diversity and functional potential in the marine mammal oral microbiome.' *Current Biology* 27.24 (2017): 3752-3762.

[56] Dworkin, M. 'Introduction to the myxobacteria.' *Prokaryotic Development*. American Society of Microbiology, 2000. 221-242.

[57] Eddy, S.R. 'Accelerated profile HMM searches.' *PLoS computational biology* 7.10 (2011): e1002195.

[58] Edgar, R.C. 'MUSCLE: multiple sequence alignment with high accuracy and high throughput.' *Nucleic acids research* 32.5 (2004): 1792-1797.

[59] Edgar, R.C. 'Search and clustering orders of magnitude faster than BLAST.' *Bioinformatics* 26.19 (2010): 2460-2461.

[60] Eigeland, K.A., *et al.* 'Bacterial community structure in the hindgut of wild and captive dugongs (*Dugong dugong*).' *Aquatic Mammals* 38.4 (2012): 402.

[61] Eloe-Fadrosh, E.A., *et al.* 'Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs.' *Nature communications* 7 (2016): 10476.

[62] Eren, A.M., *et al.* 'Anvi'o: an advanced analysis and visualization platform for 'omics data.' *PeerJ* 3 (2015): e1319.

[63] Erwin, P.M., *et al.* 'High diversity and unique composition of gut microbiomes in pygmy (*Kogia breviceps*) and dwarf (*K. sima*) sperm whales.' *Scientific reports* 7.1 (2017): 7205.

[64] Estes, J.A. 'Growth and equilibrium in sea otter populations.' *The Journal of Animal Ecology* (1990): 385-401.

[65] Estes, J. A., *et al.* 'Killer whale predation on sea otters linking oceanic and nearshore ecosystems.' *Science* 282.5388 (1998): 473-476.

[66] Estes, J.A., Palmisano, J.F. 'Sea otters: their role in structuring nearshore communities.' *Science* 185.4156 (1974): 1058-1060.

[67] Estes, J.A, *et al.* 'Causes of mortality in California sea otters during periods of population growth and decline.' *Marine Mammal Science* 19.1 (2003): 198-216.

[68] Ewels, P., *et al.* 'MultiQC: summarize analysis results for multiple tools and samples in a single report.' *Bioinformatics* 32.19 (2016): 3047-3048.

[69] Fadely, B. S., Zeligs, J.A, Costa, D.P. 'Assimilation efficiencies and maintenance requirements of California sea lions (*Zalophus californianus*) fed walleye pollock (*Theragra chalcogramma*) and herring (*Clupea harengus*).' Final Report to the National Marine Mammal Laboratory, Alaska Fisheries Science Center, National Marine Fisheries Service 7600.1994 (1994): 98115-0070.

[70] Fagan, R.P., Fairweather, N.F. 'Biogenesis and functions of bacterial S-layers.' *Nature Reviews Microbiology* 12.3 (2014): 211.

[71] Falasco, E., *et al.* 'Diatom teratological forms and environmental alterations: a review.' *Hydrobiologia* 623.1 (2009): 1-35.

[72] Fernandez, L.A., Berenguer, J. 'Secretion and assembly of regular surface structures in Gram-negative bacteria.' *FEMS microbiology reviews* 24.1 (2000): 21-44.

[73] Fichant, G., Basse, M-J., Quentin, Y. 'ABCdb: an online resource for ABC transporter repertories from sequenced archaeal and bacterial genomes.' *FEMS microbiology letters* 256.2 (2006): 333-339.

[74] Finn, R.D., Clements, J., Eddy, S.R. 'HMMER web server: interactive sequence similarity searching.' *Nucleic acids research* 39.2 (2011): W29-W37.

[75] Finn, R.D., *et al.* 'HMMER web server: 2015 update.' *Nucleic acids research* 43.W1 (2015): W30-W38.

[76] Finn, R.D., *et al.* 'The Pfam protein families database: towards a more sustainable future.' *Nucleic acids research* 44.D1 (2015): D279-D285.

[77] Foster, G., *et al.* 'Proposal of *Bisgaardia hudsonensis* gen. nov., sp. nov. and an additional genomospecies, isolated from seals, as new members of the family Pasteurellaceae.' *International journal of systematic and evolutionary microbiology* 61.12 (2011): 3016-3022.

[78] Fredricks, D.N., Fiedler, T.L., Marrazzo, J.M. 'Molecular identification of bacteria associated with bacterial vaginosis.' *New England Journal of Medicine* 353.18 (2005): 1899-1911.

[79] Fu, L., *et al.* 'CD-HIT: accelerated for clustering the next-generation sequencing data.' *Bioinformatics* 28.23 (2012): 3150-3152.

[80] Gaboriau-Routhiau, V., *et al.* 'The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses.' *Immunity* 31.4 (2009): 677-689.

[81] Georgiou, G., *et al.* 'Practical applications of engineering Gram-negative bacterial cell surfaces.' *Trends in biotechnology* 11.1 (1993): 6-10.

[82] Gerber, L.R., *et al.* 'Mortality sensitivity in life-stage simulation analysis: a case study of Southern sea otters' *Ecological Applications* 14.5 (2004): 1554-1565.

[83] Glad, T., *et al.* 'Bacterial diversity in faeces from polar bear (*Ursus maritimus*) in Arctic Svalbard.' *BMC microbiology* 10.1 (2010): 10.

[84] Godoy-Vitorino, F., *et al.* 'The microbiome of a striped dolphin (*Stenella coeruleoalba*) stranded in Portugal.' *Research in microbiology* 168.1 (2017): 85-93.

[85] Goldman, C.G., *et al.* '*Helicobacter* spp. from gastric biopsies of stranded South American fur seals (*Arctocephalus australis*).' *Research in veterinary science* 86.1 (2009): 18-21.

[86] Goldman, C. G., *et al.* 'Novel gastric *helicobacters* and oral campylobacters are present in captive and wild cetaceans.' *Veterinary microbiology* 152.1-2 (2011): 138-145.

[87] Griffen, A.L., *et al.* 'Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing.' *The ISME journal* 6.6 (2012): 1176.

[88] Grissa, I., Vergnaud, G., Pourcel, C. 'CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats.' *Nucleic acids research* 35.suppl2 (2007): W52-W57.

[89] Groenendijk, J., Duplaix, N., Marmontel, M., Van Damme, P., Schenck, C. 2015. '*Pteronura brasiliensis*.' *The IUCN Red List of Threatened Species* 2015: e.T18711A21938411.

[90] Guindon, S., Gascuel, O. 'A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.' *Systematic biology* 52.5 (2003): 696-704.

[91] Hammer, T.J., *et al.* 'Caterpillars lack a resident gut microbiome.' *Proceedings of the National Academy of Sciences* (2017): 201707186.

[92] Hansen, M.J., *et al.* '*Otariodibacter oris* gen. nov., sp. nov., a member of the family *Pasteurellaceae* isolated from the oral cavity of pinnipeds.' *International journal of systematic and evolutionary microbiology* 62.11 (2012): 2572-2578.

[93] Harper, C.G., *et al.* 'Isolation and characterization of novel *Helicobacter* spp. from the gastric mucosa of harp seals *Phoca groenlandica*.' *Diseases of aquatic organisms* 57.1-2 (2003): 1-9.

[94] Hasle, G.R., Fryxell, G.A. 'Diatoms: cleaning and mounting for light and electron microscopy.' *Transactions of the American Microscopical Society* (1970): 469-474.

[95] He, X., *et al.* 'Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle.' *Proceedings of the National Academy of Sciences* 112.1 (2015): 244-249.

[96] Hedlund, B.P., Kuhn, D.A. 'The genera Simonsiella and Alysiella.' *The Prokaryotes*. Springer New York, 2006. 828-839.

[97] Heithaus, M.R., *et al.* 'Predicting ecological consequences of marine top predator declines.' *Trends in ecology & evolution* 23.4 (2008): 202-210.

[98] Hooda, S., *et al.* 'Current state of knowledge: the canine gastrointestinal microbiome.' *Animal health research reviews* 13.1 (2012): 78-88.

[99] Hooper, L.V., Littman, D.R., Macpherson, A.J. 'Interactions between the microbiota and the immune system.' *Science* 336.6086 (2012): 1268-1273.

[100] Horner, R.A. 'A taxonomic guide to some common marine phytoplankton.' (2002).

[101] Hospenthal, M.K., Costa, T.R.D, Waksman, G. 'A comprehensive guide to pilus biogenesis in Gram-negative bacteria.' *Nature Reviews Microbiology* 15.6 (2017): 365.

[102] Hug, L.A., *et al.* 'A new view of the tree of life.' *Nature microbiology* 1 (2016): 16048.

[103] Hug, L.A., *et al.* 'Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling.' *Microbiome* 1.1 (2013): 22.

[104] Hug, L.A., *et al.* 'Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages.' *Environmental microbiology* 18.1 (2016): 159-173.

[105] Hyatt, D., *et al.* 'Prodigal: prokaryotic gene recognition and translation initiation site identification.' *BMC bioinformatics* 11.1 (2010): 119.

[106] Ichinohe, T., *et al.* 'Microbiota regulates immune defense against respiratory tract influenza A virus infection.' *Proceedings of the National Academy of Sciences* 108.13 (2011): 5354-5359.

[107] Janssen, S., *et al.* 'Phylogenetic Placement of Exact Amplicon Sequences Improves Associations with Clinical Information.' *mSystems* 3.3 (2018): e00021-18.

[108] Jinek, M., *et al.* 'A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity.' *Science* (2012): 1225829.

[109] Jones, P., *et al.* 'InterProScan 5: genome-scale protein function classification.' *Bioinformatics* 30.9 (2014): 1236-1240.

[110] Jones, S.J., *et al.* 'The Genome of the Northern Sea Otter (*Enhydra lutris kenyoni*).' *Genes* 8.12 (2017): 379.

[111] Joshi, N.A., and Fass, J.N. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files, version 1.33 (2011). https://github.com/najoshi/sickle.

[112] Kadioglu, A., *et al.* 'The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease.' *Nature Reviews Microbiology* 6.4 (2008): 288.

[113] Kantor, R.S. *et al.* 'Small genomes and sparse metabolism of sediment-associated bacteria from four candidate phyla.' *MBio* 4.5 (2013): e00708-13.

[114] Kaufman, L., Rousseeuw, P. 'Clustering by means of medoids'. North-Holland, 1987.

[115] Kelley, L.A., *et al.* 'The Phyre2 web portal for protein modeling, prediction and analysis.' *Nature protocols* 10.6 (2015): 845.

[116] Kenyon, K.W. 'The sea otter in the eastern Pacific Ocean.' *North American Fauna* (1969): 1-352.

[117] Kim, D., *et al.* 'Centrifuge: rapid and sensitive classification of metagenomic sequences.' *Genome research* 26.12 (2016): 1721-1729.

[118] Kanehisa, M., *et al.* 'KEGG as a reference resource for gene and protein annotation.' *Nucleic acids research* 44.D1 (2015): D457-D462.

[119] Kanehisa, M., *et al.* 'KEGG: new perspectives on genomes, pathways, diseases and drugs.' *Nucleic acids research* 45.D1 (2016): D353-D361.

[120] Kanehisa, M., Goto. S. 'KEGG: kyoto encyclopedia of genes and genomes.' *Nucleic acids research* 28.1 (2000): 27-30.

[121] Kearse, M., *et al.* 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.' *Bioinformatics* 28.12 (2012): 1647-1649.

[122] Kirkpatrick, C.M., Stullken, D.E., Jones, R.D. 'Notes on captive sea otters.' *Arctic* 8.1 (1955): 46-59.

[123] Kiszka, J.J., Heithaus, M.R., Wirsing, A.J. 'Behavioural drivers of the ecological roles and importance of marine mammals.' *Marine Ecology Progress Series 523* (2015): 267-281.

[124] Knoll A.H. 'Life on a Young Planet' (Princeton Univ Press, Princeton, NJ). (2003)

[125] Koepfli, K-P., *et al.* 'Multigene phylogeny of the Mustelidae: resolving relationships, tempo and biogeographic history of a mammalian adaptive radiation.' *BMC biology* 6.1 (2008): 10.

[126] Kolenbrander, P.E. 'Oral microbial communities: biofilms, interactions, and genetic systems.' *Annual Reviews in Microbiology* 54.1 (2000): 413-437.

[127] Koropatkin, N.M., Cameron, E.A., Martens, E.C. 'How glycan metabolism shapes the human gut microbiota.' *Nature Reviews Microbiology* 10.5 (2012): 323.

[128] Kremer, J.R., Mastronarde, D.N., McIntosh, J.R. 'Computer visualization of three-dimensional image data using IMOD.' *Journal of structural biology* 116.1 (1996): 71-76.

[129] Kreuder, C., *et al.* 'Patterns of mortality in southern sea otters (Enhydra lutris nereis) from 1998–2001.' *Journal of Wildlife Diseases* 39.3 (2003): 495-509.

[130] Krzywinski, M.I., *et al.* 'Circos: an information aesthetic for comparative genomics.' *Genome research* (2009).

[131] Kuehbacher, T., *et al.* 'Intestinal TM7 bacterial phylogenies in active inflammatory bowel disease.' *Journal of medical microbiology* 57.12 (2008): 1569-1576.

[132] Kuramitsu, H.K., *et al.* 'Interspecies interactions within oral microbial communities.' *Microbiology and molecular biology reviews* 71.4 (2007): 653-670.

[133] Land, M., *et al.* 'Insights from 20 years of bacterial genome sequencing.' *Functional & integrative genomics* 15.2 (2015): 141-161.

[134] Langmead, B., Salzberg, S.L. 'Fast gapped-read alignment with Bowtie 2.' *Nature methods* 9.4 (2012): 357.

[135] Lavery, T.J., *et al.* 'High nutrient transport and cycling potential revealed in the microbial metagenome of Australian sea lion (*Neophoca cinerea*) faeces.' *PLoS One* 7.5 (2012): e36478.

[136] Leewenhoeck A. Observation, communicated to the publisher by Mr. Antony van Leewenhoeck, in a Dutch letter of the 9 Octob. 1676 here English'd: concerning little animals by him observed in rain-well-sea and snow water; as also in water wherein pepper had lain infused. *Phil. Trans.* 12, 821–831. (doi:10.1098/rstl.1677.0003)

[137] Legendre, P., Legendre, L.F.J. 'Numerical ecology'. Vol. 24. *Elsevier*, 2012.

[138] Letunic, I., Bork, P. 'Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.' *Nucleic acids research* 39.suppl2 (2011): W475-W478.

[139] Ley, R.E., *et al.* 'Evolution of mammals and their gut microbes.' *Science* 320.5883 (2008): 1647-1651.

[140] Ley, R.E., *et al.* 'Worlds within worlds: evolution of the vertebrate gut microbiota.' *Nature Reviews Microbiology* 6.10 (2008): 776.

[141] Li, W., Godzik, A. 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.' *Bioinformatics* 22.13 (2006): 1658-1659.

[142] Li, H., *et al.* 'The sequence alignment/map format and SAMtools.' *Bioinformatics* 25.16 (2009): 2078-2079.

[143] Li, D., *et al.* 'MEGAHIT v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices.' *Methods* 102 (2016): 3-11.

[144] Li, D., *et al.* 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.' *Bioinformatics* 31.10 (2015): 1674-1676.

[145] Li, H., *et al.* 'The sequence alignment/map format and SAMtools.' *Bioinformatics* 25.16 (2009): 2078-2079.

[146] Lima, N., *et al.* 'Temporal stability and species specificity in bacteria associated with the bottlenose dolphins respiratory system.' *Environmental microbiology reports* 4.1 (2012): 89-96.

[147] Lindblad-Toh, K., *et al.* 'A high-resolution map of human evolutionary constraint using 29 mammals.' *Nature* 478.7370 (2011): 476.

[148] Lloyd-Price, J., *et al.* 'Strains, functions and dynamics in the expanded Human Microbiome Project.' *Nature* 550.7674 (2017): 61.

[149] Lonhienne, T., *et al.* 'Endocytosis-like protein uptake in the bacterium *Gemmata obscuriglobus*.' *Proceedings of the National Academy of Sciences* 107.29 (2010): 12883-12888.

[150] Love, M.I., Huber, W., Anders, S. 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.' *Genome biology* 15.12 (2014): 550.

[151] Lowe, T.M., Eddy, S.R. 'tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.' *Nucleic acids research* 25.5 (1997): 955.

[152] Luef, B., *et al.* 'Diverse uncultivated ultra-small bacterial cells in groundwater.' *Nature communications* 6 (2015): 6372.

[153] Lundberg, J.O., Weitzberg, E., Gladwin, M.T. 'The nitrate–nitrite–nitric oxide pathway in physiology and therapeutics.' *Nature reviews drug discovery* 7.2 (2008): 156.

[154] Makarova, K.S., *et al.* 'An updated evolutionary classification of CRISPR–Cas systems.' *Nature Reviews Microbiology* 13.11 (2015): 722.

[155] Markowitz, V.M., *et al.* 'IMG: the integrated microbial genomes database and comparative analysis system.' *Nucleic acids research* 40.D1 (2011): D115-D122.

[156] Martensson, P-E, Nordoy, E.S., Blix, A.S. 'Digestibility of krill (*Euphausia superba* and *Thysanoessa* sp.) in minke whales (*Balaenoptera acutorostrata*) and crabeater seals (*Lobodon carcinophagus*).' *British Journal of Nutrition* 72.05 (1994): 713-716.

[157] Mastronarde, D.N. 'SerialEM: a program for automated tilt series acquisition on Tecnai microscopes using prediction of specimen position.' *Microscopy and Microanalysis* 9.S02 (2003): 1182-1183.

[158] Mazmanian, S.K., *et al.* 'An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system.' *Cell* 122.1 (2005): 107-118.

[159] McCowan, R. P., Cheng, K. J., Costerton, J. W. 'Colonization of a portion of the bovine tongue by unusual filamentous bacteria.' *Applied and environmental microbiology* 37.6 (1979): 1224-1229.

[160] McFall-Ngai, M., *et al.* 'Animals in a bacterial world, a new imperative for the life sciences.' *Proceedings of the National Academy of Sciences* 110.9 (2013): 3229-3236.

[161] McLaughlin, R. W., *et al.* 'Detection of *Helicobacter* in the fecal material of the endangered Yangtze finless porpoise *Neophocaena phocaenoides asiaeorientalis*.' *Diseases of aquatic organisms* 95.3 (2011): 241-245.

[162] McMurdie, P.J., Holmes, S. 'phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.' *PloS one* 8.4 (2013): e61217.

[163] McMurdie, P.J., Holmes, S. 'Waste not, want not: why rarefying microbiome data is inadmissible.' *PLoS computational biology* 10.4 (2014): e1003531.

[164] Medema, M.H., *et al.* 'antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences.' *Nucleic acids research* 39 (2011): W339-W346.

[165] Medema, M.H., *et al.* 'Minimum information about a biosynthetic gene cluster.' *Nature chemical biology* 11.9 (2015): 625.

[166] Merson, S.D., *et al.* 'Variation in the hindgut microbial communities of the Florida manatee, *Trichechus manatus latirostris* over winter in Crystal River, Florida.' *FEMS microbiology ecology* 87.3 (2014): 601-615.

[167] Moissl, C., *et al.* 'The unique structure of archaeal 'hami', highly complex cell appendages with nano-grappling hooks.' *Molecular microbiology* 56.2 (2005): 361-370.

[168] Monson, D.H., McCormick, C., Ballachey, B.E. 'Chemical anesthesia of

northern sea otters (*Enhydra lutris*): results of past field studies.' *Journal of Zoo and Wildlife Medicine* 32.2 (2001): 181-189.

[169] Moriya, Y., *et al.* 'KAAS: an automatic genome annotation and pathway reconstruction server.' *Nucleic acids research* 35.suppl2 (2007): W182-W185.

[170] Murgarella, M., *et al.* 'A first insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*.' *PLoS One* 11.3 (2016): e0151561.

[171] Nelson, T.M., *et al.* 'Diet and phylogeny shape the gut microbiota of Antarctic seals: a comparison of wild and captive animals.' *Environmental microbiology* 15.4 (2013): 1132-1145.

[172] Nelson, T.M., Rogers, T.L., Brown, M.V. 'The gut bacterial community of mammals from marine and terrestrial habitats.' *PLoS One* 8.12 (2013): e83655.

[173] Nikitin, D. I., Vasilyeva, L.V., Lokhmacheva, R.A. 'New and rare forms of soil microorganisms.' M.: Nauka (1966).

[174] Nowak, M.A. 'Five rules for the evolution of cooperation.' *Science* 314.5805 (2006): 1560-1563.

[175] O'Leary, N.A., *et al.* 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.' *Nucleic acids research* 44.D1 (2015): D733-D745.

[176] Olsen, M.A., *et al.* 'Chitinolytic bacteria in the minke whale forestomach.' *Candian journal of microbiology* 46.1 (1999): 95-94.

[177] Paez-Espino, D., *et al.* 'Uncovering Earth's virome.' *Nature* 536.7617 (2016): 425.

[178] Paez-Espino, D., *et al.* 'IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses.' *Nucleic acids research* (2016): gkw1030.

[179] Pangborn, J., Kuhn, D.A., Woods, J.R. 'Dorsal-ventral differentiation in Simonsiella and other aspects of its morphology and ultrastructure.' *Archives of microbiology* 113.3 (1977): 197-204.

[180] Pankhurst, C.L., Auger, D. W., Hardie, J.M. 'An ultrastructural study of adherence to buccal epithelial cells of *Simonsiella* sp.' *Letters in applied microbiology* 6.5 (1988): 125-128.

[181] Parks, D.H., *et al.* 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.' *Genome research* 25.7 (2015): 1043-1055.

[182] Peng, Y., *et al.* 'IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth.' *Bioinformatics* 28.11 (2012): 1420-1428.

[183] Perras, A. K., *et al.* 'S-layers at second glance? Altiarchaeal grappling hooks (hami) resemble archaeal S-layer proteins in structure and sequence.' *Frontiers in microbiology* 6 (2015): 543.

[184] Ponomarova, O., Patil, K.R. 'Metabolic interactions in microbial communities: untangling the Gordian knot.' *Current opinion in microbiology* 27 (2015): 37-44.

[185] Pride, D.T., *et al.* 'Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses.' *BMC genomics* 7.1 (2006): 8.

[186] Probst, A.J., Auerbach, A.K., Moissl-Eichinger, C. 'Archaea on human skin.' *PloS one* 8.6 (2013): e65388.

[187] Pruesse, Elmar, *et al.* 'SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.' *Nucleic acids research* 35.21 (2007): 7188-7196.

[188] Pruesse, E., Peplies, J., Glockner, F.O. 'SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes.' *Bioinformatics* 28.14 (2012): 1823-1829.

[189] Quast, C., *et al.* 'The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.' *Nucleic acids research* 41.D1 (2012): D590-D596.

[190] Raes, J., *et al.* 'Prediction of effective genome size in metagenomic samples.' *Genome biology* 8.1 (2007): R10.

[191] Ravel, J., *et al.* 'Vaginal microbiome of reproductive-age women.' *Proceedings of the National Academy of Sciences* 108.Supplement 1 (2011): 4680-4687.

[192] Relman, D.A., Falkow, S. 'The meaning and impact of the human genome sequence for microbiology.' *Trends in microbiology* 9.5 (2001): 206-208.

[193] Reynolds, A.P., *et al.* 'Clustering rules: a comparison of partitioning and hierarchical clustering algorithms.' *Journal of Mathematical Modelling and Algorithms* 5.4 (2006): 475-504.

[194] Rinke, C., *et al.* 'Insights into the phylogeny and coding potential of microbial dark matter.' *Nature* 499.7459 (2013): 431.

[195] Rodriguez, L.M., Konstantinidis, K.T. 'Bypassing cultivation to identify bacterial species.' *Microbe* 9.3 (2014): 111-8.

[196] Rognes, T., *et al.* 'VSEARCH: a versatile open source tool for metagenomics.' *PeerJ* 4 (2016): e2584.

[197] Roman, J., McCarthy, J.J. 'The whale pump: marine mammals enhance primary productivity in a coastal basin.' *PloS one* 5.10 (2010): e13255.

[198] Ross, R., Sims, P.A. 'The fine structure of the frustule in centric diatoms: a suggested terminology.' *British Phycological Journal* 7.2 (1972): 139-163.

[199] Ruby, G.J., Bellare, P., DeRisi, J.L. 'PRICE: software for the targeted assembly of components of (Meta) genomic sequence data.' *G3: Genes, Genomes, Genetics* (2013): g3-113.

[200] Russo, C.D., *et al.* 'Bacterial Species Identified on the Skin of Bottlenose Dolphins Off Southern California via Next Generation Sequencing Techniques.' *Microbial ecology* 75.2 (2018): 303-309.

[201] Russell, J.B., Rychlik, J.L. 'Factors that alter rumen microbial ecology.' *Science* 292.5519 (2001): 1119-1122.

[202] Sadd, B.M., *et al.* 'The genomes of two key bumblebee species with primitive eusocial organization.' *Genome biology* 16.1 (2015): 76.

[203] Sanders, J.G., *et al.* 'Baleen whales host a unique gut microbiome with similarities to both carnivores and herbivores.' *Nature communications* 6 (2015): 8285.

[204] Sato, Y., Willis, B.L., Bourne, D.G. 'Pyrosequencing-based profiling of archaeal and bacterial 16S r RNA genes identifies a novel archaeon associated

with black band disease in corals.' *Environmental microbiology* 15.11 (2013): 2994-3007.

[205] Schreiber, F., *et al.* 'Denitrification in human dental plaque.' *BMC biology* 8.1 (2010): 24.

[206] Schipper, J., *et al.* 'The status of the world's land and marine mammals: diversity, threat, and knowledge.' *Science* 322.5899 (2008): 225-230.

[207] Schulz, F., *et al.* 'Towards a balanced view of the bacterial tree of life.' *Microbiome* 5.1 (2017): 140.

[208] Sekiguchi, Y., *et al.* 'First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking.' *PeerJ* 3 (2015): e740.

[209] Serfass, T., Evans, S.S., Polechla, P. 2015. '*Lontra canadensis*.' *The IUCN Red List of Threatened Species 2015*: e.T12302A21936349.

[210] Sharon, I., *et al.* 'Accurate, multi-kb reads resolve complex populations and detect rare microorganisms.' *Genome research* (2015): gr-183012.

[211] Shen, Z., *et al.* 'Novel urease-negative *Helicobacter* sp.'*H. enhydrae* sp. nov.' isolated from inflamed gastric tissue of southern sea otters.' *Diseases of aquatic organisms* 123.1 (2017): 1-11.

[212] Shmakov, S., *et al.* 'Discovery and functional characterization of diverse class 2 CRISPR-Cas systems.' *Molecular cell* 60.3 (2015): 385-397.

[213] Simunek, J., Hodrova, B., Bartonova, H., Kopecny, J. 'Chitinolytic bacteria of the mammal digestive tract.' *Folia microbiologica* 46.1 (2001):76-78.

[214] Skennerton, C.T., Imelfort, M., Tyson, G.W. 'Crass: identification and reconstruction of CRISPR from unassembled metagenomic data.' *Nucleic acids research* 41.10 (2013): e105-e105.

[215] Sodergren, E., *et al.* 'The genome of the sea urchin *Strongylocentrotus purpuratus*.' *Science* 314.5801 (2006): 941-952.

[216] Soding, J., Biegert, A., Lupas, A.N. 'The HHpred interactive server for protein homology detection and structure prediction.' *Nucleic acids research* 33.suppl2 (2005): W244-W248.

[217] Soto, G. E., Hultgren, S.J. 'Bacterial adhesins: common themes and variations in architecture and assembly.' *Journal of bacteriology* 181.4 (1999): 1059-1071.

[218] Soverini, M., *et al.* 'The bottlenose dolphin (*Tursiops truncatus*) faecal microbiota.' *FEMS microbiology ecology* 92.4 (2016): fiw055.

[219] Stamatakis, A. 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.' *Bioinformatics* 30.9 (2014): 1312-1313.

[220] Steed, P.D.M. 'Simonsiellaceae fam. nov. with characterization of *Simonsiella crassa* and *Alysiella filiformis*.' *Microbiology* 29.4 (1962): 615-624.

[221] Stevens, E.C, Hume, I.D. 'Comparative physiology of the vertebrate digestive system'. Cambridge University Press, 2004.

[222] Stoermer, E. F., Pankratz, H.S., Bowen, C.C. 'Fine structure of the diatom *Amphipleura pellucida*. II. Cytoplasmic fine structure and frustule formation.' *American Journal of Botany* (1965): 1067-1078.

[223] Sunagawa, S., *et al.* 'Structure and function of the global ocean microbiome.' *Science* 348.6237 (2015): 1261359.

[224] Suzek, B.E., *et al.* 'UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches.' *Bioinformatics* 31.6 (2014): 926-932.

[225] Swanson, K.S., *et al.* 'Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice.' *The ISME journal* 5.4 (2011): 639.

[226] Tamaki, H., *et al.* '*Armatimonas rosea* gen. nov., sp. nov., of a novel bacterial phylum, Armatimonadetes phyl. nov., formally called the candidate phylum OP10.' *International journal of systematic and evolutionary microbiology* 61.6 (2011): 1442-1447.

[227] Tharanathan, R.N., Kittur, F.S. 'Chitin—the undisputed biomolecule of great potential.' *Critical reviews in food science and nutrition* (2003): 61-87.

[228] Thometz, N. M., *et al.* 'Energetic demands of immature sea otters from birth to weaning: implications for maternal costs, reproductive behavior and population-level trends.' *Journal of Experimental Biology* 217.12 (2014): 2053-2061.

[229] Tibshirani, R., Walther, G., Hastie, T. 'Estimating the number of clusters in a data set via the gap statistic.' *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001): 411-423.

[230] Tibshirani, R., Walther, G. and Hastie, T. (2000). 'Estimating the number of clusters in a dataset via the Gap statistic.' Technical Report. Stanford.

[231] Tinker, T.M., Hatfield, B. 'Annual California Sea Otter Census-2016 spring census summary.' *USGS annual report* (2016).

[232] Tinker, T.M., *et al.* 'Incorporating diverse data and realistic complexity into demographic estimation procedures for sea otters.' *Ecological Applications* 16.6 (2006): 2293-2312.

[233] Turnbaugh, P.J., *et al.* 'An obesity-associated gut microbiome with increased capacity for energy harvest.' *Nature* 444.7122 (2006): 1027.

[234] Turnbaugh, P.J., *et al.* 'The human microbiome project.' *Nature* 449.7164 (2007): 804.

[235] Ultsch, A., Morchen, F. 'ESOM-Maps: tools for clustering, visualization, and classification with emergent SOM'. Department of Mathematics and Computer Science, University of Marburg, Germany, Technical Report 46, 1–7 (2005).

[236] US Fish and Wildlife Service. 'Southern Sea Otter (*Enhydra lutris nereis*) 5-Year Review: Summary and Evaluation.' US Fish and Wildlife Service, Ventura Office, CA, USA (2015). https://www.fws.gov/ventura/endangered/species/info/sso.html

[237] Vasilyeva, L.V. '*Stella*, a New Genus of Soil Prosthecobacteria, with Proposals for *Stella humosa* sp. nov. and *Stella vacuolata* sp. nov.' *International Journal of Systematic and Evolutionary Microbiology* 35.4 (1985): 518-521.

[238] Voelz, H., Reichenbach, H. 'Fine structure of fruiting bodies of *Stigmatella aurantiaca* (*Myxobacterales*).' *Journal of bacteriology* 99.3 (1969): 856-866.

[239] Wang, Q., *et al.* 'Naive Bayesian classifier for rapid assignment of rRNA

sequences into the new bacterial taxonomy.' *Applied and environmental microbiology* 73.16 (2007): 5261-5267.

[240] Wanger, G., Onstott, T.C., Southam, G. 'Stars of the terrestrial deep subsurface: A novel 'star-shaped' bacterial morphotype from a South African platinum mine.' *Geobiology* 6.3 (2008): 325-330.

[241] Warnow, T. 'SATe-Enabled Phylogenetic Placement.' *Encyclopedia of Metagenomics*. Springer US, 2015. 619-621.

[242] Watt, J., Siniff, D.B., Estes, J.A. 'Inter-decadal patterns of population and dietary change in sea otters at Amchitka Island, Alaska.' *Oecologia* 124.2 (2000): 289-298.

[243] Wattam, A.R., *et al.* 'PATRIC, the bacterial bioinformatics database and analysis resource.' *Nucleic acids research* 42.D1 (2013): D581-D591.

[244] Weber, T., *et al.* 'antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters.' *Nucleic acids research* 43.W1 (2015): W237-W243.

[245] Welch, J.L.M., *et al.* 'Biogeography of a human oral microbiome at the micron scale.' *Proceedings of the National Academy of Sciences* 113.6 (2016): E791-E800.

[246] Wildhaber, I., Baumeister, W. 'The cell envelope of *Thermoproteus tenax*: three-dimensional structure of the surface layer and its role in shape maintenance.' *The EMBO journal* 6.5 (1987): 1475-1480.

[247] Woyke, T., *et al.* 'Decontamination of MDA reagents for single cell whole genome amplification.' *PloS one* 6.10 (2011): e26161.

[248] Wright, L., de Silva, P., Chan, B., Reza Lubis, I. 2015. '*Aonyx cinereus*.' *The IUCN Red List of Threatened Species 2015*: e.T44166A21939068.

[249] Wrighton, K.C., *et al.* 'Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla.' *Science* 337.6102 (2012): 1661-1665.

[250] Wu, D., *et al.* 'A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea.' *Nature* 462.7276 (2009): 1056.

[251] Xie, C-H., Yokota, A. 'Transfer of the misnamed [*Alysiella*] sp. IAM 14971 (ATCC 29468) to the genus *Moraxella* as *Moraxella oblonga* sp. nov.' *International journal of systematic and evolutionary microbiology* 55.1 (2005): 331-334.

[252] Xu, J., *et al.* 'Emerging trends for microbiome analysis: from single-cell functional imaging to microbiome big data.' *Engineering* 3.1 (2017): 66-70.

[253] Xu, Q., *et al.* 'A distinct type of pilus from the human microbiome.' *Cell* 165.3 (2016): 690-703.

[254] Xu, B., *et al.* 'Metagenomic analysis of the pygmy loris fecal microbiome reveals unique functional capacity related to metabolism of aromatic compounds.' *PLoS One 8.2* (2013): e56565.

[255] Xu, B., *et al.* 'Metagenomic analysis of the *Rhinopithecus bieti* fecal microbiome reveals a broad diversity of bacterial and glycoside hydrolase profiles related to lignocellulose degradation.' *BMC genomics* 16.1 (2015): 174.

[256] Yarza, P., *et al.* 'Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences.' *Nature Reviews Microbiology* 12.9 (2014): 635.

[257] Yilmaz, P., *et al.* 'The SILVA and 'all-species living tree project (LTP)' taxonomic frameworks.' *Nucleic acids research* 42.D1 (2013): D643-D648.

[258] Young, K.D. 'Bacterial morphology: why have different shapes?' *Current opinion in microbiology* 10.6 (2007): 596-600.

[259] Zilber-Rosenberg, I., Rosenberg, E. 'Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution.' *FEMS microbiology reviews* 32.5 (2008): 723-735.

[260] Zinder, S.H., Dworkin, M. 'Morphological and physiological diversity.' *The Prokaryotes*. Springer, New York, NY, 2006. 185-220.

[261] Zheng, Y., *et al.* 'Interleukin-22 mediates early host defense against attaching and effacing bacterial pathogens.' *Nature medicine* 14.3 (2008): 282.

[262] Zhu, B., Wang, X., Li, L. 'Human gut microbiome: the second genome of human body.' *Protein & cell* 1.8 (2010): 718-725.

[263] Zhu, L., *et al.* 'Evidence of cellulose metabolism by the giant panda gut microbiome.' *Proceedings of the National Academy of Sciences* 108.43 (2011): 17714-17719.