Novel Microbial Diversity and Functional Potential in the Marine Mammal Oral Microbiome

Natasha K. Dudek,[1] Christine L. Sun,[2,3] David Burstein,[4] Rose S. Kantor,[5] Daniela S. Aliaga Goltsman,[2,3] Elisabeth M. Bik,[2,8] Brian C. Thomas,[4] Jillian F. Banfield,[4,6] and David A. Relman[2,3,7,9,] *

[1]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA [2]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA [3]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA [4]Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA 94720, USA [5]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA 94720, USA [6]Earth and Environmental Science, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA [7]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA 94304, USA [8]Present address: uBiome, San Francisco, CA 94105, USA [9]Lead Contact *Correspondence: relman@stanford.edu

Summary

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.

Keywords: oral microbiota, microbial ecology, candidate phyla, metagenomics, Tursiops truncates, dolphin, marine mammal, CRISPR, Cas9, bacteriophage

Introduction

The vast majority of bacterial diversity is found within phylum-level lineages that lack isolated representatives [1], 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 [1, 2, 3]; 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 [4]. Genome-resolved metagenomic studies offer unique and unprecedented insights into the biology of these uncultured, poorly understood lineages and their biochemical diversity [3, 4, 5, 6, 7, 8]. 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 [6, 9]. Candidate phyla genomes may also reveal novel functional diversity, as phylogenetic diversity is correlated with novel proteomic diversity and biological properties [10, 11].

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 [12]. 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 [12]. 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 [12].

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

Results

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. [12], which indicated that these two samples (DolJOral78 and DolZOral124) contained representatives from nine candidate phyla at relative abundances of ≥0.05% (Table S1). Paired-end Illumina HiSeq reads were generated, filtered, assembled, and used to recover microbial genomes, as described in STAR 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 1; Data S1; see STAR 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 16S rRNA gene had been recovered [2]. Additionally, low-coverage (≤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 [13, 14, 15]), such as coral heads [16] and human skin [17], but were not originally recognized as affiliated with the Woesearcheota phylum or placed within a comprehensive phylogeny.

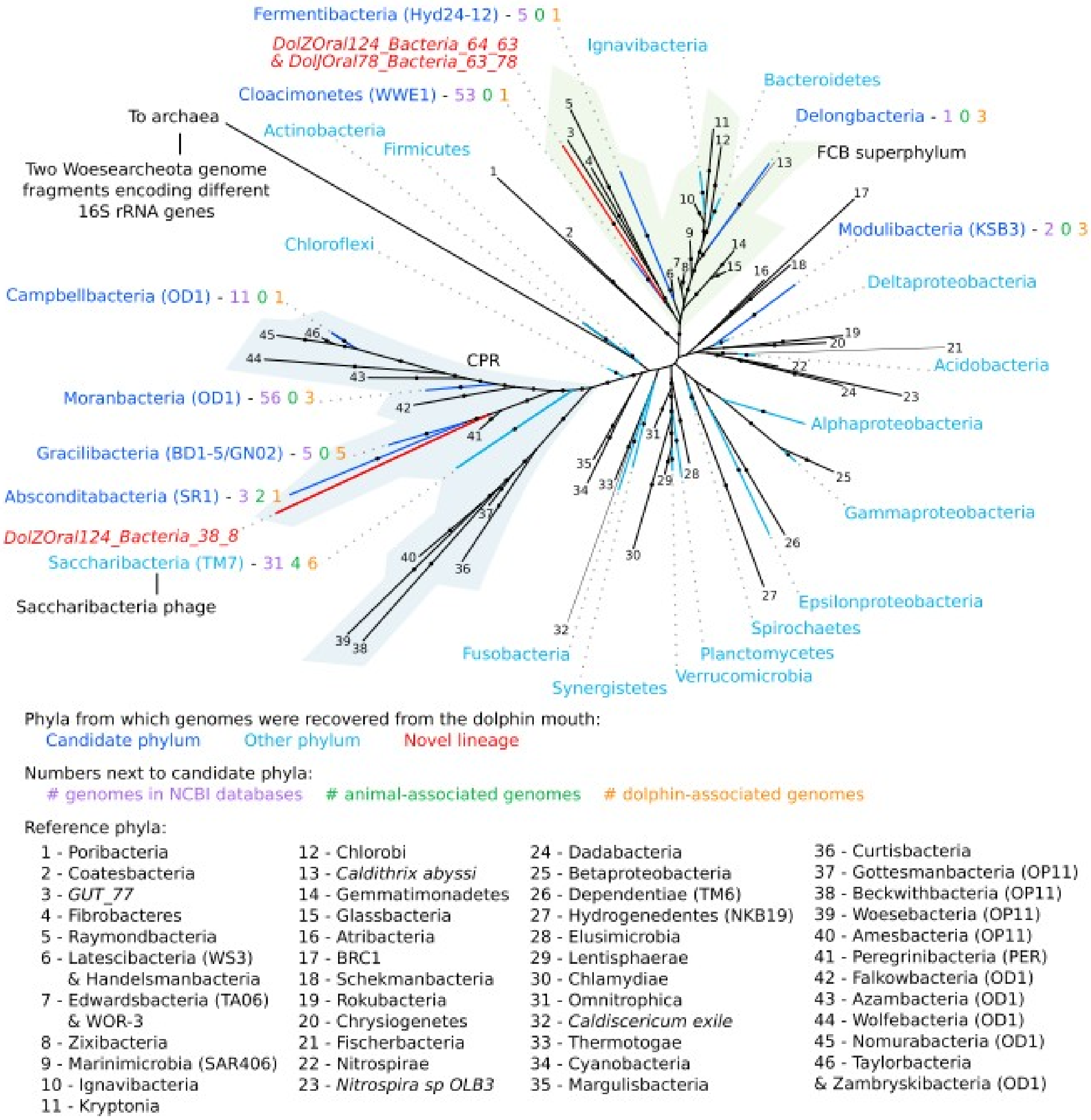


Figure 1: Phylogenetic Relationships among Genomes Recovered from the Dolphin Mouth. The maximum-likelihood tree includes representation from all genomes that contained ≥8 of 15 ribosomal proteins used to infer the phylogeny (with the exception of one Delongbacteria genome with 7 ribosomal proteins) 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. [1], presumably due to lower sampling depth reducing the ability to resolve the branching order of the deepest lineages. See also Figures S1–S3, Table S1, and Data S1.

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 [12] (Figure 2; Figure S1; Table S1). 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 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 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 [12], 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 [4]. 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. [12]) 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.

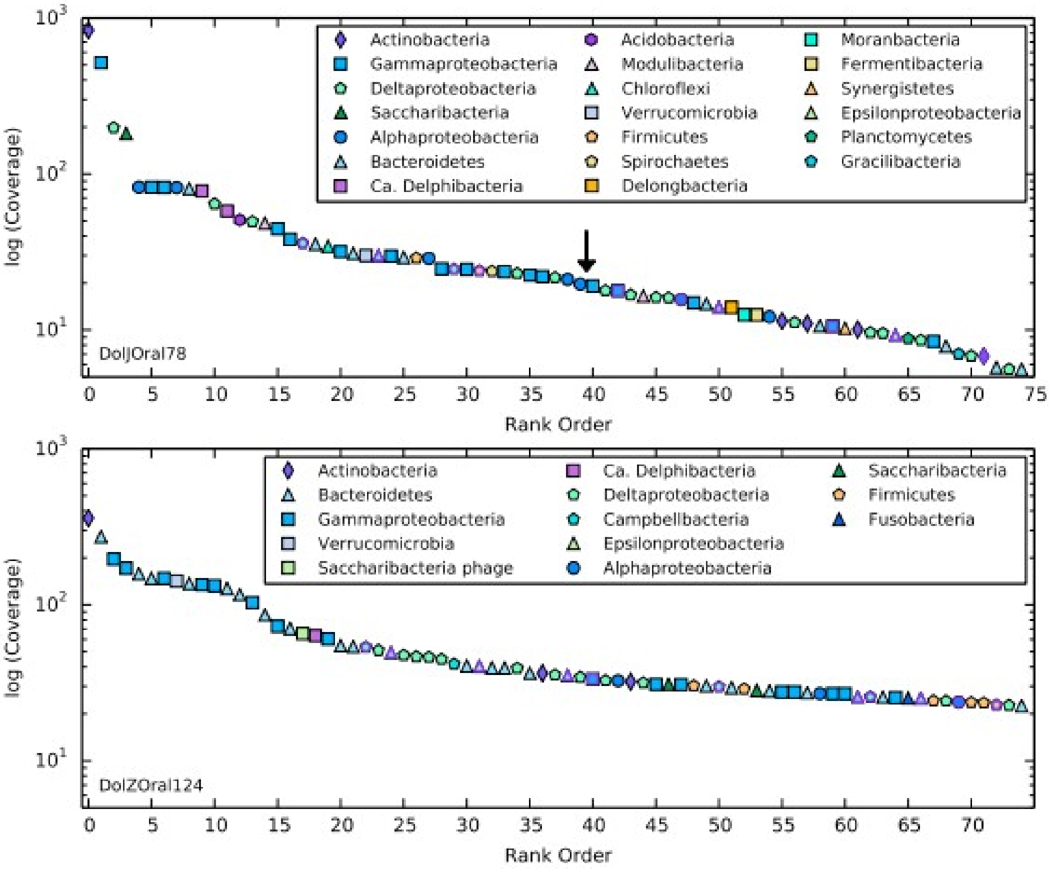


Figure 2: Community Structure of the Dolphin Oral Microbiota. The top panel presents the community structure of the DolJOral78 sample, and the bottom panel presents that of the DolZOral124 sample. Each symbol represents a bin, which is a set of scaffolds that share similar genomic signatures. In most 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 Figure S1 and Table S1.

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 STAR 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 [18, 19] 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 [18, 19] 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.

Novel, Deeply Divergent Phylum-Level Lineages

The concatenated ribosomal protein tree enabled determination of the phylum-level identity of recovered genomes (Figure 1). Within this tree, three genomes 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. [20]).

One lineage, for which we propose the name “Delphibacteria” (rationale in Supplemental 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 63×). 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 [13, 14, 15] (see Supplemental Discussion, Figure S2, and Data S1). 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 [2]. Nearly all members of the Delphibacteria lineage share <75% sequence identity across the 16S rRNA gene with members of the Eisenbacteria phylum (Figure S2A) and <78.5% sequence identity with members of the Latescibacteria phylum (Figure S2B). Predicted proteins in the near-complete genome from this lineage were most similar to those from the Deltaproteobacteria phylum (Figure S3A). 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. [12], 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 Supplemental 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 Supplemental Discussion and Data S1). It is part of a well-supported clade with <75% sequence identity to the rest of the PER phylum, including PER-ii (Figure S2C). Predicted proteins from this lineage are most similar to those from the Peregrinibacteria phylum (Figure S3B), 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. [12], 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. [12] 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 (8×) in this study, with a relative abundance of 0.09% in the DolZOral124 sample.

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 dataset contained 49 of 51 universal bacterial single-copy genes used to assess completeness [21]. It comprised 3,362,850 bp and is predicted to contain 3,011 protein-coding genes. It 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; Data S2). 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.

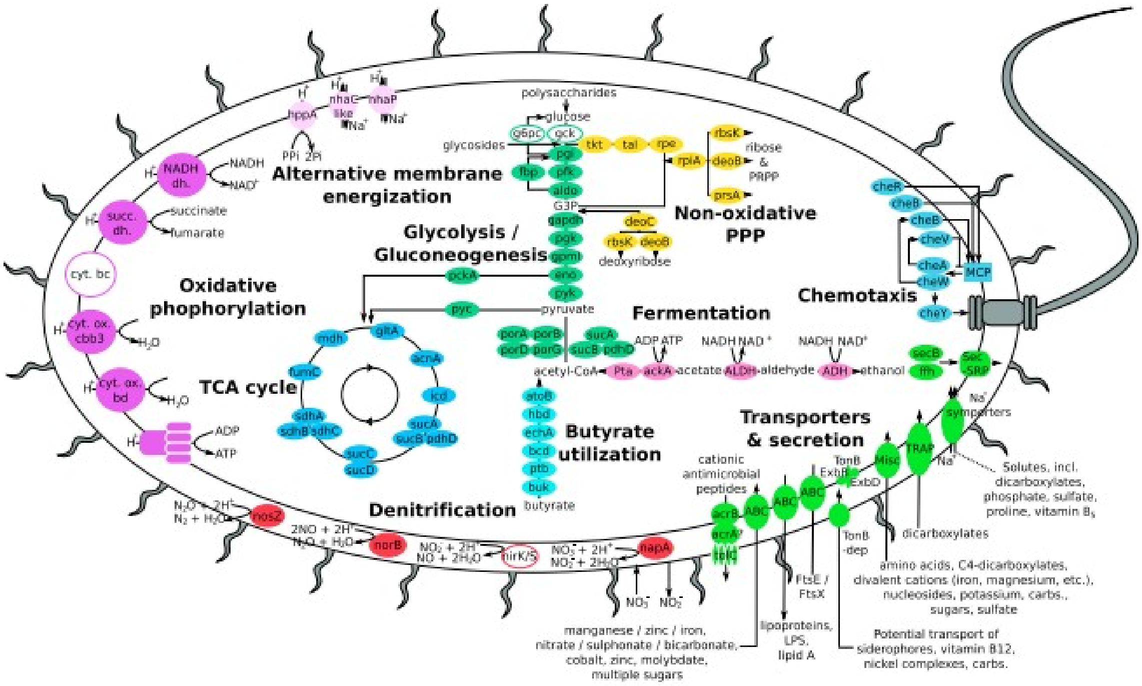


Figure 3: Functional Profile of Delphibacteria. Key predicted metabolic and functional features are depicted. Genes of interest are denoted by abbreviations in the colored shapes. Filled shapes represent genes predicted to be present or likely to be present, whereas unfilled shapes represent genes that were not identified. See also Figure S3 and Data S2.

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 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.

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 [22, 23]. 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) (Figure S4A) and a distinct tetranucleotide composition (Figure S4B). Its read coverage is consistent with the rest of the genome (Figure S4C). These findings suggest that the BGC was acquired through a relatively recent horizontal gene transfer event. Notably, the BGC is predicted to produce a relatively long non-ribosomal peptide (NRP) of 17 amino acids (Figure 4). NRPs are synthesized by NRP synthetase enzyme complexes, independent of the ribosome. In the MIBiG database [27], the average size of NRPs synthesized by BGCs is only 6 amino acids long (SD ±4.5) (Figure S4D). 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.

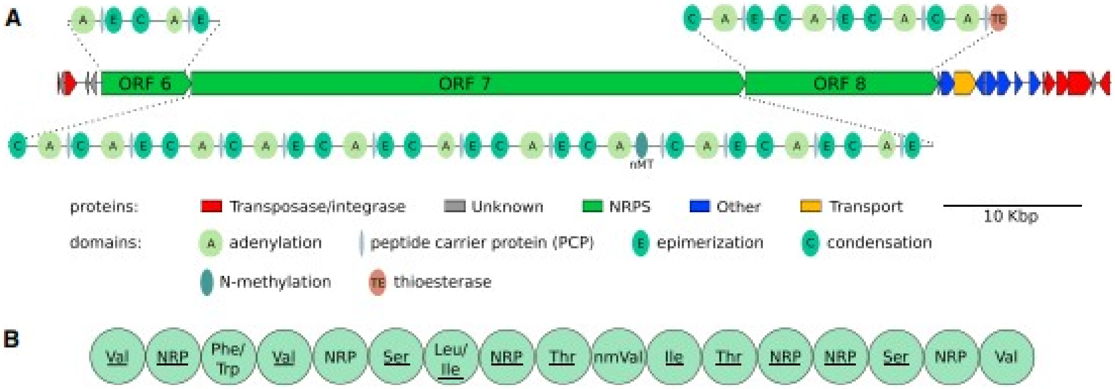


Figure 4: Novel Non-ribosomal Peptide Synthesis BGC Encoded by the Dominant Actinobacteria Genome. (A) Predicted protein and biosynthetic domain structure in the ~80.5-kbp genomic region comprising the BGC. Open reading frames along the 80.5-kbp genomic region are color coded by function: red, transposase 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 [24, 25, 26]. 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.

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 Supplemental Discussion, Figure S5, and Data S3 and S4). We identified a total of 67 unique predicted Cas9 proteins (see STAR Methods). Interestingly, two are longer than all Cas9 protein sequences in the RefSeq database [28] (accessed December 2016) (Figure 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 5B). AnaCas9 was selected as a reference because it has a resolved crystal structure and it is a type II-C Cas9, as are the three novel predicted proteins in the present study (Figure S6; Data S1). 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 (≥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 Supplemental Discussion) [29, 30].

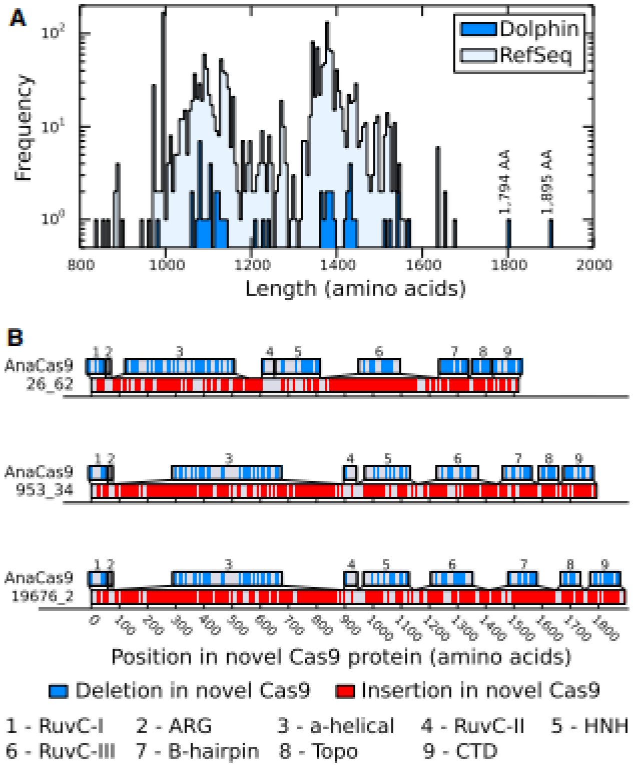


Figure 5: Unusual Predicted Cas9 Protein Sequences in the Dolphin Oral Samples. (A) Length distribution of 1,799 complete Cas9 proteins from the RefSeq database [28] (light blue) and 53 complete Cas9 proteins from the dolphin datasets (dark blue). The longest Cas9 protein in the RefSeq database [28] 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.

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. [31] 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 Supplemental 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 [32]. Although the two complete Saccharibacteria Cas9 proteins are affiliated with a single clade of type II-C Cas9 proteins (Figure S6), 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 [31, 33, 34]. Using CRISPRFinder [35] and Crass [36], we identified a total of 42 unique spacers from Saccharibacteria CRISPR arrays (see Supplemental Discussion and Data S4). 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 [37]. The 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 6; Data S5). Predicted functions of these ORFs include host cell lysis, phage packaging, and DNA recombination.

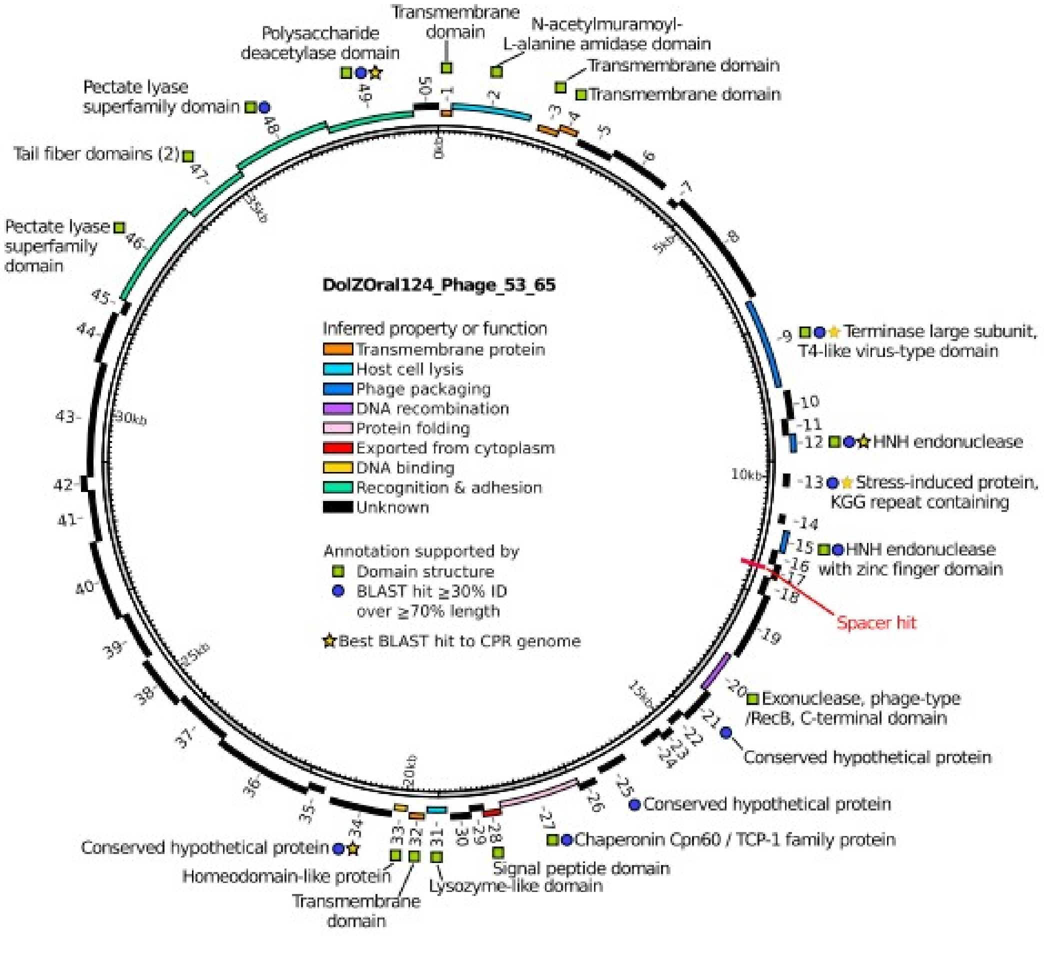

DolZOral124_Phage_53_65
Inferred property or function
Transmembrane protein
Host cell lysis
Phage packaging
DNA recombination
Protein folding
Exported from cytoplasm
DNA binding
Recognition & adhesion
Unknown
Annotation supported by
Domain structure
BLAST hit ≥30% ID over ≥70% length
Best BLAST hit to CPR genome
Transmembrane domain
N-acetylmuramoyl-L-alanine amidase domain
Transmembrane domain
Transmembrane domain
Polysaccharide deacetylase domain
Pectate lyase superfamily domain
Tail fiber domains (2)
Pectate lyase superfamily domain
Terminase large subunit, T4-like virus-type domain
HNH endonuclease
Stress-induced protein, KGG repeat containing
HNH endonuclease with zinc finger domain
Spacer hit
Exonuclease, phage-type /RecB, C-terminal domain
Conserved hypothetical protein
Conserved hypothetical protein
Chaperonin Cpn60 / TCP-1 family protein
Signal peptide domain
Lysozyme-like domain
Transmembrane domain
Homeodomain-like protein
Conserved hypothetical protein
0kb
5kb
10kb
15kb
20kb
25kb
30kb
35kb
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

Figure 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 [18, 19] hit of ≥30% identity over ≥70% length of the ORF with an e value ≤1e-05, and yellow stars denote annotations whose top BLAST [18, 19] 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 Data S3, S4, and S5.

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 [12]. 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 [38, 39]. 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 [40]. At least one Saccharibacteria strain, an obligate endobiont of an Actinobacterium, has the ability to modify human immune responses *in vitro* [9]. 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 [41]. 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 [10].

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.

Contact for Reagent and Resource Sharing

Further information and requests for reagents and resources should be directed to, and will be fulfilled by the Lead Contact, David A. Relman (relman@stanford.edu).

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. [12], 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 Supplemental Discussion and Figure S5), 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.

Method Details

DNA extraction, sequencing, and quality filtering

We used the same DNA preparations from MMP dolphin gingival sulcus samples as used by Bik et al. [12]. 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 × 250bp) 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 [66] 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) [43] were considered to be host contamination and were removed from the dataset using bowtie2 version 2.2.4 [50]. 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.

Metagenome assembly, annotation, and binning

Assembly of read-pairs from each sample was performed using IDBA-UD version 1.1.1 [56]. 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 [62]. USEARCH version 7.0.1 [69] was used to compare protein sequences from all predicted ORFs against the UniRef 90

[49] and KEGG [44, 45, 46] 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 [4] and tRNA genes were predicted using tRNAscan version 1.23 [68].

A bin is a set of scaffolds that share similar genomic features, and is typically 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. [70]), and ESOMs were computed and visualized with the Databionics ESOM Tools software [53].

Identification of phage scaffolds

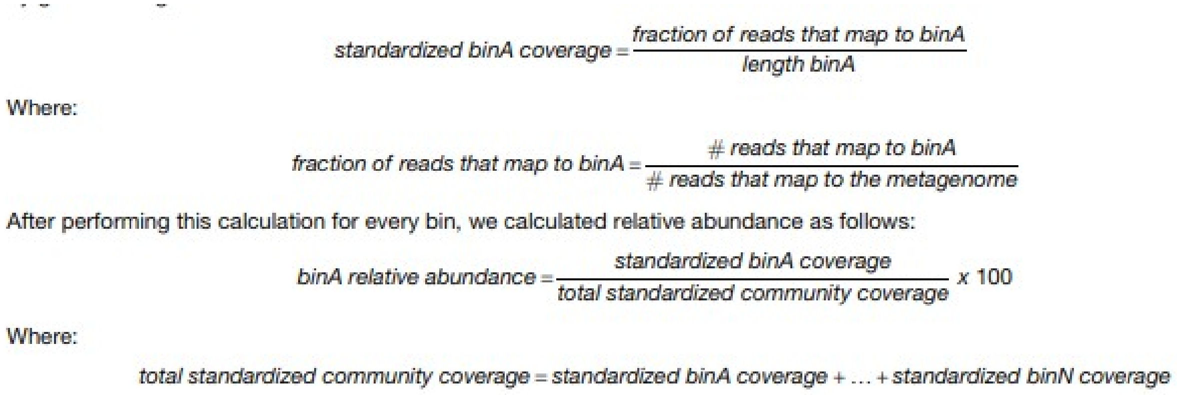
To identify candidate phage sequences, we required that scaffolds have two or more gene annotations containing virus-specific keywords from the list: “capsid, phage, terminase, base plate, baseplate, prohead, virion, virus, viral, tape measure, tapemeasure neck, tail, head, bacteriophage, prophage, portal, 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).

Refining selected scaffolds

The PRICE assembly algorithm [61] 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 [4]. 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) [4] 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 [4] 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 [56].

Bin completeness and characterization

From sample DolJOral78, we recovered 34 near complete bacterial genomes (≥80% complete), 16 draft-quality partial bacterial genomes (≥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 ≥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:



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

Where:

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

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

$$\textit{binA relative abundance} = \frac{\textit{standardized binA coverage}}{\textit{total standardized community coverage}}\ x\ 100$$

Where:

$$\textit{total standardized community coverage} = \textit{standardized binA coverage} + \ldots + \textit{standardized binN coverage}$$

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 [65]. 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 [18, 19] with OTUs from the previous 16S rRNA gene survey [12] and determining whether close relatives (≥95% identity) had been detected and identified.

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) [71]. 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 [48], 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 [60] 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 [54]. Evolutionary model selection for each of the ribosomal protein sets was performed using ProtTest3 [63, 72]. Protein sets were concatenated using Geneious version 7.1.9 [54]. A phylogenetic tree was created using RAxML [64] under the LG+G (PROTGAMMALG) evolutionary model with 100 bootstrap replicates. The tree was visualized using iTOL [58] 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 [13, 14, 15]. 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 [13, 14, 15], sequences from all genome assemblies from the Latescibacteria, Delphibacteria, and Eisenbacteria phyla with a 16S rRNA gene, and the top 20 BLAST [18, 19] 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 [13, 14, 15] 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 [67] with the SILVA SSU Ref NR 99 database release 128 [14, 15, 67] as a reference. Columns containing at least 3% gaps were removed using Geneious version 7.1.9 [54] and a phylogenetic tree was run under the GTR+G (PROTGAMMAGTR) evolutionary model in RAxML [64] 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. [20]. 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 [69] (-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 [58].

Metabolic reconstruction of DolZOral124_Bacteria_64_63 (*Candidatus* Delphibacteria)

Metabolic pathways were identified using KAAS [59]. 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 [18, 19] searches against the NCBI non-redundant protein database (https://www.ncbi.nlm.nih.gov/protein/), pHMMER [55], and/or InterProScan [57]. Searches for specific proteins of interest that were not identified by KAAS [59] 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 [73] and searching for it using the HMMER suite version 3.1b2 [

74], or by obtaining the corresponding protein sequence from the NCBI database and querying it against our genome with BLAST [18, 19], 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 [55] and by performing BLAST [18, 19] searches of candidates against the ABCdb CleanDB [42], 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/).

Biosynthetic gene cluster structural predictions

The structure of the dolphin Actinobacteria BGC was characterized using antiSMASH version 3.0 [24, 25, 26]. Figure 4 was based on output from antiSMASH, which was modified using Inkscape (https://inkscape.org/en/).

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 [74], using the Cas9 HMMs from Makarova et al. [75] 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 [51, 76] to cluster together similar protein sequences ≥ 800 amino acids, using cutoffs of ≥ 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 [77]. To compare the dolphin Cas9 protein sequences against previously sequenced Cas9 proteins, we downloaded all Cas9 proteins from the RefSeq database [28] 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. [75], and the AnaCas9 protein using MUSCLE [60]. 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 [63, 72] to determine the best fitting evolutionary model. A phylogenetic tree was constructed using RAxML [64], applying the VT + G + F (PROTGAMMAVTF) evolutionary model. The tree was visualized using iTOL [58]. 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 [60], 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 Supplemental Discussion and Figure S5), 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 [74] to search for Cas protein HMMs constructed based on alignments from [75]. 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 [35] and false positives were removed manually. These results were used to calculate the proportion of CRISPR-Cas types in mammalian oral microbiomes.

Identification and analysis of scaffolds targeted by CPR spacers

We identified spacers in assembled scaffolds using CRISPRFinder [35] and CRASS [36]. 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 [18, 19]) against both full metagenomic assemblies, the NCBI non-redundant nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/), and the NCBI virus database [47] to identify scaffolds targeted by spacers. Spacers were required to have a match of ≥ 95% sequence identity over 100% of the spacer or 100% identity over ≥ 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 [18, 19] searches against the NCBI non-redundant protein database (https://www.ncbi.nlm.nih.gov/protein/) and by submitting protein sequences to pHMMER [55] and InterProScan [57]. Figure 6 was created using Circos [52] and Inkscape (https://inkscape.org/en/).

Quantification and Statistical Analysis

Detailed descriptions of the quantitative and statistical methods used in this paper can be found in the Results and Method Details sections. Briefly, this includes the methods used for DNA extraction, sequencing, read quality filtering, metagenome assembly, annotation, genome binning and curation, assessment of bin completeness, phylogenetic analyses, functional analyses, and CRISPR-Cas-related analyses.

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.

Acknowledgments

We thank Mohamed Donia (Princeton University) for guidance in the analysis of the Actinobacteria biosynthetic gene cluster. We thank Stephanie Venn-Watson, Kevin Carlin, and Eric D. Jensen from the U.S. Navy Marine Mammal Program (San Diego, CA, USA) for providing the dolphin oral samples, Alix Switzer (Stanford University) for the contribution of DNA prepared from a harbor seal oral sample, and Frances Gulland and coworkers at The Marine Mammal Center (Sausalito, CA, USA) for collection of the harbor seal oral sample. We thank Beth Shapiro (University of California, Santa Cruz) for her support of this research, as well as members of the Relman laboratory for helpful comments and discussion, especially Elizabeth Costello. Funding was provided by a B1 fellowship from the Fonds de Recherche du Québec–Nature et Technologie (to N.K.D.), a long-term EMBO fellowship (to D.B.), a Frontiers in Science award from the Allen Institute (to J.F.B.), a cycle grant from the U.S. Department of Energy (to J.F.B.), Office of Naval Research grant 00014-07-1-0287 (to D.A.R.), and the Thomas C. and Joan M. Merigan Endowment at Stanford University (to D.A.R.). B.C.T. and J.F.B. are shareholders of Metagenomi.co. E.M.B. is an employee and shareholder of uBiome.

References

1. Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. Nat. Microbiol. 1, 16048. 2. Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J., Thomas, B.C., Singh, A., Wilkins, M.J., Karaoz, U., et al. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat. Commun. 7, 13219. 3. Eloe-Fadrosh, E.A., Paez-Espino, D., Jarett, J., Dunfield, P.F., Hedlund, B.P., Dekas, A.E., Grasby, S.E., Brady, A.L., Dong, H., Briggs, B.R., et al. (2016). Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. Nat. Commun. 7, 10476. 4. Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., and Banfield, J.F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523, 208–211. 5. Hug, L.A., Thomas, B.C., Sharon, I., Brown, C.T., Sharma, R., Hettich, R.L., Wilkins, M.J., Williams, K.H., Singh, A., and Banfield, J.F. (2016). Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. Environ. Microbiol. 18, 159–173. 6. Kantor, R.S., Wrighton, K.C., Handley, K.M., Sharon, I., Hug, L.A., Castelle, C.J., Thomas, B.C., and Banfield, J.F. (2013). Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. MBio 4, e00708–e00713. 7. Sekiguchi, Y., Ohashi, A., Parks, D.H., Yamauchi, T., Tyson, G.W., and Hugenholtz, P. (2015). First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking. PeerJ 3, e740. 8. Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J., VerBerkmoes, N.C., Wilkins, M.J., Hettich, R.L., Lipton, M.S., Williams, K.H., et al. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science 337, 1661–1665. 9. He, X., McLean, J.S., Edlund, A., Yooseph, S., Hall, A.P., Liu, S.Y., Dorrestein, P.C., Esquenazi, E., Hunter, R.C., Cheng, G., et al. (2015). Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc. Natl. Acad. Sci. USA 112, 244–249. 10. Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A., and Banfield, J.F. (2017). New CRISPR-Cas systems from uncultivated microbes. Nature 542, 237–241. 11. Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., Kunin, V.,

Goodwin, L., Wu, M., Tindall, B.J., et al. (2009). A phylogenydriven genomic encyclopaedia of Bacteria and Archaea. Nature 462, 1056–1060. 12. Bik, E.M., Costello, E.K., Switzer, A.D., Callahan, B.J., Holmes, S.P., Wells, R.S., Carlin, K.P., Jensen, E.D., Venn-Watson, S., and Relman, D.A. (2016). Marine mammals harbor unique microbiotas shaped by and yet distinct from the sea. Nat. Commun. 7, 10516. 13. Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glo¨ ckner, F.O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35, 7188–7196. 14. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glo¨ ckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596. 15. Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., and Glo¨ ckner, F.O. (2014). The SILVA and ‘‘All-species Living Tree Project (LTP)’’ taxonomic frameworks. Nucleic Acids Res. 42, D643–D648. 16. Sato, Y., Willis, B.L., and Bourne, D.G. (2013). Pyrosequencing-based profiling of archaeal and bacterial 16S rRNA genes identifies a novel archaeon associated with black band disease in corals. Environ. Microbiol. 15, 2994–3007. 17. Probst, A.J., Auerbach, A.K., and Moissl-Eichinger, C. (2013). Archaea on human skin. PLoS ONE 8, e65388. 18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410. 19. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics 10, 421. 20. Yarza, P., Yilmaz, P., Pruesse, E., Glo¨ ckner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzeby, J., Amann, R., and Rossello ´-Mo´ ra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. 12, 635–645. 21. Raes, J., Korbel, J.O., Lercher, M.J., von Mering, C., and Bork, P. (2007). Prediction of effective genome size in metagenomic samples. Genome Biol. 8, R10. 22. Donia, M.S., Cimermancic, P., Schulze, C.J., Wieland Brown, L.C., Martin, J., Mitreva, M., Clardy, J., Linington, R.G., and Fischbach, M.A. (2014). A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell 158, 1402–1414. 23. Kadioglu, A., Weiser, J.N., Paton, J.C., and Andrew, P.W. (2008). The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease. Nat. Rev. Microbiol. 6, 288–301. 24. Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., and Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 39, W339–W346. 25. Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E., and Weber, T. (2013). antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 41, W204–W212. 26. Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., Lee, S.Y., Fischbach, M.A., Mu¨ ller, R., Wohlleben, W., et al. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 43, W237–W243. 27. Medema, M.H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J.B., Blin, K., de Bruijn, I., Chooi, Y.H., Claesen, J., Coates, R.C., et al. (2015). Minimum information about a biosynthetic gene cluster. Nat. Chem. Biol. 11, 625–631. 28. O’Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745. 29. So¨ ding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33, W244–W248. 30. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. 10, 845–858. 31. Burstein, D., Sun, C.L., Brown, C.T., Sharon, I., Anantharaman, K., Probst, A.J., Thomas, B.C., and Banfield, J.F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat. Commun. 7, 10613. 32. Sharon, I., Kertesz, M., Hug, L.A., Pushkarev, D., Blauwkamp, T.A., Castelle, C.J., Amirebrahimi, M., Thomas, B.C., Burstein, D., Tringe, S.G., et al. (2015). Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res. 25, 534–543. 33. Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering Earth’s virome. Nature 536, 425–430. 34. Paez-Espino, D., Chen, I.A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V.M., Nielsen, T., et al. (2017). IMG/VR: a database of cultured and uncultured DNA viruses and retroviruses. Nucleic Acids Res. 45, D457–D465. 35. Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 35, W52–W57. 36. Skennerton, C.T., Imelfort, M., and Tyson, G.W. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res. 41, e105. 37. Pride, D.T., Wassenaar, T.M., Ghose, C., and Blaser, M.J. (2006). Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 7, 8. 38. Lundberg, J.O., Weitzberg, E., and Gladwin, M.T. (2008). The nitrate-nitrite-nitric oxide pathway in physiology and therapeutics. Nat. Rev. Drug Discov. 7, 156–167. 39. Schreiber, F., Stief, P., Gieseke, A., Heisterkamp, I.M., Verstraete, W., de Beer, D., and Stoodley, P. (2010). Denitrification in human dental plaque. BMC Biol. 8, 24. 40. Brinig, M.M., Lepp, P.W., Ouverney, C.C., Armitage, G.C., and Relman, D.A. (2003). Prevalence of bacteria of division TM7 in human subgingival plaque and their association with disease. Appl. Environ. Microbiol. 69, 1687–1694. 41. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816–821. 42. Fichant, G., Basse, M.J., and

Quentin, Y. (2006). ABCdb: an online resource for ABC transporter repertories from sequenced archaeal and bacterial genomes. FEMS Microbiol. Lett. 256, 333–339. 43. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al.; Broad Institute Sequencing Platform and Whole Genome Assembly Team; Baylor College of Medicine Human Genome Sequencing Center Sequencing Team; Genome Institute at Washington University (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478, 476–482. 44. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462. 45. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353–D361. 46. Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30. 47. Brister, J.R., Ako-Adjei, D., Bao, Y., and Blinkova, O. (2015). NCBI viral genomes resource. Nucleic Acids Res. 43, D571–D577. 48. Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 42, D581–D591. 49. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., and Wu, C.H.; UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926–932. 50. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. 51. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659. 52. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645. 53. Ultsch, A., and Mo¨ rchen, F. (2005). 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. 54. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. 55. Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A., and Eddy, S.R. (2015). HMMER web server: 2015 update. Nucleic Acids Res. 43, W30–W38. 56. Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428. 57. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240. 58. Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39, W475–W478. 59. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35, W182–W185. 60. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. 61. Ruby, J.G., Bellare, P., and Derisi, J.L. (2013). PRICE: software for the targeted assembly of components of (meta) genomic sequence data. G3 (Bethesda) 3, 865–880. 62. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. 63. Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27, 1164–1165. 64. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. 65. Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–5267. 66. Joshi, N.A., and Fass, J.N. (2011). Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files, version 1.33. https://github. com/najoshi/sickle. 67. Pruesse, E., Peplies, J., and Glo¨ ckner, F.O. (2012). SINA: accurate highthroughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28, 1823–1829. 68. Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964. 69. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461. 70. Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., and Banfield, J.F. (2009). Community-wide analysis of microbial genome sequence signatures. Genome Biol. 10, R85. 71. Hug, L.A., Castelle, C.J., Wrighton, K.C., Thomas, B.C., Sharon, I., Frischkorn, K.R., Williams, K.H., Tringe, S.G., and Banfield, J.F. (2013). Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome 1, 22. 72. Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704. 73. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279–D285. 74. Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195. 75. Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPRCas systems. Nat. Rev. Microbiol. 13, 722–736. 76. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. 77. Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., et al. (2015). Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. Mol. Cell 60, 385–397.