THE UNIVERSITY OF CHICAGO

HIGH-RESOLUTION MICROBIAL 'OMICS AT SCALE TO ILLUMINATE ENIGMATIC AND RATHER PICKY MICROBIAL RESIDENTS OF THE HUMAN ORAL CAVITY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

AND

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY

ALON SHAIBER

CHICAGO, ILLINOIS

MARCH 2020

Dedicated to Rebecca.

Rebecca, I am eternally grateful for your love.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Microbes are the most common form of life on Earth and play a crucial role in biogeochemical processes that sustain all forms of life. Similar to every other habitat on Earth, microbes occupy almost every part of the human body and play an important role in health and disease. Our understanding of the ecology and evolution of microbes has been significantly changed due to the recent revolution in DNA sequencing technology and the rise of ‘omics data, which has transformed microbiology to a data-rich science. But new challenges are arising as computational tools and training that enable effective utilization of ‘omics data are lacking. Here I present my efforts to solve bottlenecks in the analysis of microbial ‘omics, and to empower microbiologists engaged in ‘omics data science.

My work in developing computational tools has been driven by specific questions in microbial ecology. By utilizing high resolution ‘omics analysis approaches, I illuminated the evolutionary journey of cryptic microbial residents of the human oral cavity, with a focus on members of the candidate division TM7. My analysis revealed that TM7s split into groups of tongue specialists and dental plaque specialists, indicating that oral TM7s are “picky” regarding their desired habitat within the mouth. While plaque specialists associated with TM7 from environmental samples from an evolutionary and functional perspectives, tongue specialists associated with TM7 from animal gut. These findings indicate an ecological resemblance between the plaque environment and non-host environments such as soil and sediment from a microbial point of view, suggesting that the plaque environment may have served as a stepping stone for environmental microbes to adapt to host environments for some clades of human associated microbes. Additionally, I revealed that prophages are widespread amongst oral-associated TM7, while absent from environmental TM7, suggesting that prophages may have played a role in adaptation of TM7 to the host environment, perhaps by facilitating horizontal gene transfer. An in-depth description of my findings from the oral cavity is followed by a discussion of novel tools along with examples of their applications, and a discussion of good practices for scalable, high resolution exploration of 'omics data.

CHAPTER 1 INTRODUCTION

1.1 Diversity, abundance, and importance of microbial life

They are hard to notice and easy to ignore as we go about our daily lives, and yet microbes are everywhere, and are not only the most common form of life on Earth (Whitman, Coleman, and Wiebe 1998), but also perform biogeochemical processes essential in recycling molecules and making them available to sustain all forms of life on Earth (Falkowski, Fenchel, and Delong 2008; Planavsky et al. 2014). Microbes are profoundly abundant and occupy every niche on Earth, from soil (Torsvik, Øvreås, and Thingstad 2002; Delmont et al. 2015) to oceans (Béjà et al. 2002; Delmont et al. 2018), and as far as we can tell, also within and on top of every plant (Hardoim et al. 2015; Vorholt 2012; Reinhold-Hurek et al. 2015) and animal (Amato et al. 2019; Reveillaud et al. 2019; Dudek et al. 2017; Dewhirst et al. 2012; Bahrndorff et al. 2016).

1.2 The human oral microbiome

Similar to every habitat on Earth, we are also colonized by microbes, that form the human microbiome, and that are abundantly found across our body (Turnbaugh et al. 2007); and each person is estimated to contain as many microbial cells as human cells (Sender, Fuchs, and Milo 2016). Microbial community structure and its variations have been associated with health and disease (Martinez-Guryn, Leone, and Chang 2019), hence our understanding of the composition and distribution of microbes across body sites is highly important from a medical perspective. The oral cavity is amongst the richest reservoirs of microbes in the human body, and is approximated to harbor more than 600 microbial species (Dewhirst et al. 2010) that are found in high densities (Sender, Fuchs, and Milo 2016) across anatomically diverse sites within the mouth (Welch, Dewhirst, and Borisy 2019), and play an important role both in oral and non-oral diseases (Wade 2013).

1.3 Opportunities and challenges in sequencing-enabled study of microbial life

The recent revolution in the field of microbiome has been largely driven by the emergence of new DNA sequencing technologies that allow access to large-scale genomic information. Studies utilizing the accessibility of sequencing data are producing deep insights into naturally occurring microbial populations, and are changing our understanding of the Tree of Life (Brown et al. 2015; Spang et al. 2015), transform our view of microbes performing key biogeochemical processes (Koch, van Kessel, and Lücker 2019; Delmont et al. 2018), leading to discovery of novel biosynthetic pathways (Libis et al. 2019) and novel antibiotics (Hover et al. 2018), and much more (Quince et al. 2017).

As microbiology is transforming into a data-rich science, microbiologists are faced with new challenges (Kyrpides, Eloe-Fadrosh, and Ivanova 2016). The complexity of the data requires novel algorithmic solutions, and a myriad of computational tools developed by the scientific community strive to address this need (List of Bioinformatics Software - omicX), but there are no established standards to guide researchers toward the appropriate tools for their specific needs (Quince et al. 2017). On the other hand, heavy reliance on standard workflows with rigid analysis steps, limits the creative exploration of researchers and prevents the utilization of the full potential of data. Moreover, as the field evolves, the requirement to integrate multiple sources of information, such as genomics, transcriptomics, proteomics and other 'omics data in a multi-'omics approach increases, but tools that allow such integration are lacking (Kyrpides, Eloe-Fadrosh, and Ivanova 2016). Proper training that would enable microbiologists to take advantage of the surge in 'omics data, and the infrastructure to support efficient use of data are lacking as well (Kyrpides, Eloe-Fadrosh, and Ivanova 2016). Efforts are being made by the scientific community to put forth standards of analysis (Bowers et al. 2017), but awareness of these guidelines amongst researchers and reviewers of studies that heavily rely on 'omics approaches is still lacking (Shaiber and Eren 2019). Improper analysis due to the complexity of the data could yield false conclusions (Koutsovoulos et al. 2016), and errors propagate as erroneous data are deposited to public databases (Shaiber and Eren 2019; Chen et al. 2019c). Sequencing technologies continue to evolve at a rapid pace with long read (van der Helm et al. 2017; Bertrand et al. 2019), droplet microfluidics (Zilionis et al. 2017), and Hi-C technologies (Belton et al. 2012), to name a few, suggesting that these challenges are likely to persist.

1.4 Anvi'o - an integrated analysis and visualization platform for 'omics data

Throughout my graduate studies I addressed these challenges primarily by taking a leading role in the development of anvi'o. Anvi'o is open source software with more than 65,000 lines of code for the analysis and visualization of 'omics data (Eren et al. 2015). The latest version of anvi'o (v6 ''esther") includes 125 programs that each perform a unique task, and an object-oriented design, allows for these programs to be extended, as well as combined together. To execute this modularity, anvi'o relies on a collection of databases that are created, modified, merged, split, and queried through the various atomic programs (Figure 1). The anvi'o databases allow researchers to combine information from various 'omics data-types, including genomes, metagenomes, and meta-transcriptomes, and apply a variety of 'omics analysis approaches, including various metagenomic, pagenomic, meta-pan-genomic, and phylogenomic approaches (Yeoman et al. 2019; Eren et al. 2015; Delmont et al. 2019, 2018; Reveillaud et al. 2019; Delmont and Eren 2018). Along with the flexibility in the design of each analysis, offered by this design, the anvi'o databases generated per project can be shared as stand-alone files in addition to standard summary tables and plots. Sharing an anvi'o database allows other scientists to easily reproduce results, and moreover, to explore novel questions by utilizing the anvi'o interactive interface.

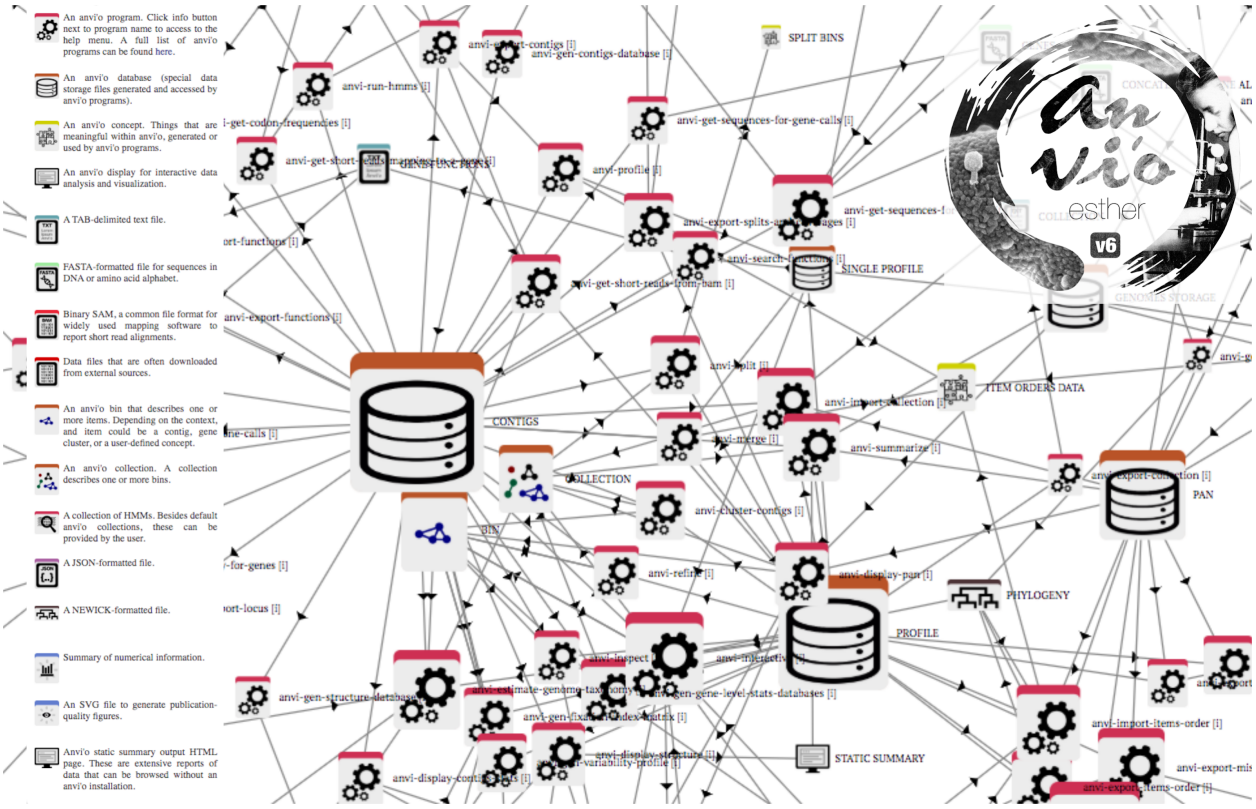

Figure 1: The anvi'o programs, databases and concepts form an interconnected network. This screenshot taken from http://merenlab.org/software/anvio/network/ presents how atomic anvi'o programs interact with anvi'o databases and relate to concepts in microbial 'omics. On the top right is the anvi'o symbol for anvi'o v6 'esther' (https://github.com/merenlab/anvio/releases/tag/v6).

The interactive interface is perhaps what distinguishes anvi'o more than anything when comparing to other 'omics analysis tools. The complexity of 'omics data often means that relying on summary statistics or on a single type of visualization is not sufficient. But most workflows available for the analysis of 'omics data produce static figures and summary tables, and each researcher is required to "dig" into the data within these tables. Due to the magnitude and complexity of these datasets, independent exploration requires high proficiency in computational approaches of data science, which is not necessarily an expertise held by every microbiologist. Anvi'o circumvents this predicament by allowing users to manually explore their data using an interactive interface that allows switching between a variety of visualization strategies seamlessly.

Although, the flexibility and breadth of the analyses offered by anvi'o provide a steep learning curve for a novice user. To help microbiologists take advantage of the variety of offered functionalities, anvi'o tutorials

include more than 115,000 words in total, of which I personally contributed more than 10,000 words spread across four tutorials (http://merenlab.org/2018/07/09/anvio-snakemake-workflows/; http://merenlab.org/2016/11/08/pangenomics-v2/; http://merenlab.org/2019/03/14/ncbi-genome-download-magic/; http://merenlab.org/2019/10/17/export-locus/). In addition, I have composed and taught a workshop to graduate students interested in learning approaches to the analysis of microbial 'omics data using anvi'o (for which material is provided at http://merenlab.org/2018/09/09/microbial-omics-workshop/) and I am actively engaged with the community of anvi'o users through github (https://github.com/merenlab/anvio), Slack (https://anvio.slack.com/), and Google Group (anvio@googlegroups.com).

In summary, anvi'o offers flexible and interactive analysis of 'omics data that empowers microbiologists to take an active role in data analysis and utilize the depth of knowledge offered by complex 'omics data. By contributing to the development of anvi'o and providing training to members of the scientific community I strived to empower scientists engaged in data-rich microbiology.

1.5 The anvi'o workflows - increasing the accessibility of large-scale and reproducible analyses using anvi'o

The flexibility offered by the atomic programs included in anvi'o comes with a price. Typical analysis steps become very numerous and grow in proportion to the number of samples/genomes that are being analyzed. Identifying this bottleneck, I implemented the anvi'o workflows, a collection of commonly-used analysis strategies for microbial 'omics. The anvi'o workflows rely on the Snakemake workflow management system (Köster and Rahmann 2012), which offers easy deployment to any computing system, automatic parallelization of independent analysis steps, and the ability to seamlessly resume interrupted workflows without repeating steps that were previously completed. Extensive documentation, helpful error messages, draft configuration files that can be edited by users to suit their analysis needs, and the reliance on Snakemake allow users with minimal knowledge of command line tools to perform analyses at scale. The anvi'o workflows are similar to other existing tools in many ways (Dean et al. 2018; Clarke et al. 2019; Uritskiy, DiRuggiero, and Taylor 2018; Kieser et al. 2019), but instead of offering static figures and tables, anvi'o workflows produce the aforementioned anvi'o databases and hence allow scientists to reach the initial steps of interactive exploration of 'omics data in a streamlined manner.

1.6 High resolution microbial ‘omics at scale to study questions in microbial ecology

My efforts in developing computational tools were strongly driven by my focus on specific questions in microbial ecology. The following chapters expand on applications of these tools to study specific ecosystems, as well as include an in-depth description of the anvi’o workflows. Chapter 2 describes the application of high resolution microbial ‘omics to investigate the ecology, evolution, and mobilome of poorly understood, yet prevalent members of the oral microbiome. In particular in this study we reveal dental-plaque specialists and tongue specialists amongst oral-associated TM7, and show that while plaque specialists are functionally and phylogenetically associated with environmental TM7, tongue-specialists are associated with other host-associated TM7 from animal gut, suggesting that at least for TM7, plaque resembles non-host environments. Chapter 3 expands on the functionality and design of anvi’o workflows. Chapter 4 includes descriptions of applications of anvi’o workflows in a variety of contexts, including the reanalysis of previously published data to highlight limitations and offer solutions for metagenomics analyses, as well as the analysis of newly generated data that led to the identification of a *Wolbachia* plasmid that could provide exciting possibilities for genomic engineering with potential application in the population control of mosquitoes that carry and transmit dengue, West Nile, and Zika viruses.

Overall, this work provides insights into the genomes, ecology, evolution and mobilome of cryptic microbes in the context of multiple ecosystems, including the human oral cavity, human blood samples, and insect ovaries, and includes a discussion of novel tools and good practices for high resolution exploration of large scale 'omics data.

CHAPTER 2 Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome

Alon Shaiber[1,2], Amy D. Willis[3], Tom O. Delmont[4], Simon Roux[5], Abigail Schmid[6], Mahmoud Yousef[7], Andrea Watson[1,8], Özcan C. Esen[1], Sonny T. M. Lee[9], Hilary Morrison[10], Floyd E Dewhirst[11,12], Jessica Mark Welch[10,*], A. Murat Eren[1,2,8,10,*]

[1] Department of Medicine, University of Chicago, Chicago, IL 60637, USA
[2] Graduate Program in Biophysical Sciences, University of Chicago, Chicago, IL 60637, USA
[3] Department of Biostatistics, University of Washington, Seattle WA 98195, USA
[4] Genoscope, Center of Atomic Energy, Évry 91000, France
[5] Department of Energy Joint Genome Institute, Walnut Creek CA 94598, USA
[6] Undergraduate Student, Computational and Applied Mathematics, University of Chicago, Chicago, IL 60637, USA
[7] Undergraduate Student, Computer Science, University of Chicago, Chicago, IL 60637, USA
[8] Committee on Microbiology, University of Chicago, Chicago, IL 60637, USA
[9] Division of Biology, Kansas State University, Manhattan, KS 66506, USA
[10] Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA
[11] Department of Microbiology, The Forsyth Institute, Cambridge, MA 02142, USA
[12] Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Boston, MA 02115, USA

* Correspondence: meren@uchicago.edu and jmarkwelch@mbl.edu

Author contribution

A Shaiber conceived the study and performed the primary data analysis, prepared figures and tables, and wrote the manuscript; ADW provided data analysis tools; TOD, SR, and A Schmid performed data analysis; MY provided data analysis tools; AW performed data analysis; OCE provided data analysis tools; STML performed data analysis; HGM processed, and analyzed the sequencing data; FED performed data analysis; JMW designed the study, collected and processed samples, and wrote the manuscript; AME designed the study, supervised the research, provided data analysis tools, performed data analysis, prepared figures and tables, and wrote the manuscript.

2.1 Abstract

Microbial residents of the human oral cavity have long been a major focus of microbiology due to their influence on host health and their intriguing patterns of site specificity amidst the lack of dispersal limitation. Yet, the determinants of niche partitioning in this habitat are yet to be fully understood, especially among the taxa that belong to recently discovered branches of microbial life. Here we used daily tongue and dental plaque metagenomes from multiple individuals and reconstructed 790 non-redundant genomes, 43 of which resolved to TM7 that formed six monophyletic clades distinctly associated either with plaque or with tongue. Both pangenomic and phylogenomic analyses grouped tongue-specific TM7 clades with other host-associated TM7 genomes. In contrast, plaque-specific TM7 grouped together with environmental TM7 genomes. Besides offering deeper insights into the ecology, evolution, and the mobilome of cryptic members of the oral microbiome, our study reveals an intriguing resemblance between dental plaque and non-host environments indicated by the TM7 evolution, suggesting that plaque may have served as a stepping stone for environmental microbes to adapt to host environments for some clades of human associated microbes. Additionally, we identify that prophages are widespread amongst oral-associated TM7, while absent from environmental TM7, suggesting that prophages may play a role in adaptation of TM7 to the host environment.

2.2 Introduction

Since the inception of microbiology as a new discipline following Antoni van Leeuwenhoek's historical observation of the animalcules (Lane, 2015), the human mouth has remained a major focus among microbiologists. The oral cavity is a rich environment with multiple distinct niches in a relatively small space partially due to (1) its diverse anatomy with hard and soft tissue structures (German and Palmer, 2006), (2) the differential influence of the host immunity throughout the oral tissue types (Moutsopoulos and Konkel, 2018), (3) its constant exposure to exogenous factors. Microbial residents of the oral cavity complement their environment with their own sophisticated lifestyles. Oral microbes form complex communities that show remarkable patterns of horizontal and vertical transmission across humans and animals (Ferretti et al., 2018; Song et al., 2013), temporal dynamism (Caporaso et al., 2011; Hall et al., 2017; Mark Welch et

al., 2014), spatial organization (Mark Welch et al., 2016), and site-specificity (Dewhirst et al., 2010; Eren et al., 2014; Mark Welch et al., 2019), where they influence the host health (Lamont et al., 2018) and the ecology of the gastrointestinal tract (Schmidt et al., 2019). Altogether, the oral cavity offers a powerful environment to study ecology and evolution of microbial systems.

One of the fundamental pursuits of microbiology is to understand the determinants of microbial colonization and niche partitioning that govern the distribution of microbes in their natural habitats. Despite the low dispersal limitation in the human oral cavity that ensures everything to be everywhere, extensive site-specificity among oral microbes has been observed since the earliest studies that used microscopy and cultivation (Socransky and Manganiello, 1971), DNA-DNA hybridization (Mager et al., 2003) and cloning (Aas et al., 2005) strategies. Factors influencing microbial site-specificity include (1) the nature of the underlying substrate (permanent teeth vs. mucosal surfaces), (2) keratinization and other features of the surface topography, (3) proximity to sources of saliva, gingival crevicular fluid, and oxygen, (4) and ability of microbes to adhere both to the substrate and to one another (Gibbons and Houte, 1975; Simón-Soro et al., 2013; Socransky and Manganiello, 1971), overall creating a fascinating ecological environment to study microbial colonization.

Our understanding of the ecology of oral microbes leapfrogged thanks to the Human Microbiome Project (HMP) (Human Microbiome Project Consortium, 2012), which generated extensive sequencing data from more than 200 healthy individuals and 9 oral sites. Studies focused on the HMP data confirmed major taxonomic differences between microbial communities associated with dental plaque and mucosal sites in the mouth (Lloyd-Price et al., 2017; Segata et al., 2012). Recruiting metagenomic short reads using single-copy core genes, Donati et al. demonstrated that while some members of the genus Neisseria were predominantly found in tongue dorsum samples, others were predominant in plaque samples (Donati et al., 2016), and Eren et al. revealed that even populations of the same species that differed by as little as one nucleotide in 16S rRNA gene amplicons could show extensive site specificity (Eren et al., 2014). Strong associations between oral sites and their microbial residents even at the finest levels of resolution raise questions regarding the drivers of such exclusiveness (Mark Welch et al., 2019). However, identifying genetic or functional determinants of site-specificity require insights into microbial pangenomes.

The human oral cavity is one of the most well characterized microbial habitats of the human body. The Human Oral Microbiome Database (HOMD) (Chen et al., 2010) describes more than 750 oral phylotypes based on full-length 16S rRNA gene sequences, 70% of which have cultured representatives, enabling genome-resolved analyses that cover a considerable fraction of oral metagenomes (Nayfach et al., 2016). Yet, one-third of the known oral taxa are missing or poorly represented in culture collections and genomic databases, and include some that are common in the oral cavity (Vartoukian et al., 2016), including members of the Candidate Phyla Radiation (CPR) (Brown et al., 2015), such as Saccharibacteria (TM7), Absconditabacteria (SR1), and Gracilibacteria (GN02). CPR bacteria form distinct branches in the Tree of Life both based on their phylogenetic origins (Hug et al., 2016) and functional makeup (Méheust et al., 2019); they lack many biological pathways that are considered essential (Brown et al., 2015) and have been shown to rely on epibiotic lifestyles (Bor et al., 2019), with a complex and poorly understood relationship with a microbial host (Bor et al., 2018). Their unique lifestyle (He et al., 2015), diversity and prevalence in the oral cavity (Camanocha and Dewhirst, 2014), association with distinct oral sites (Bor et al., 2019), and potential role in disease (Abusleme et al., 2013; Brinig et al., 2003) make them important clades to characterize for a fuller understanding of the ecology of the oral cavity.

Successful efforts targeting these enigmatic members of the oral microbiome produced the first genomic evidence to better understand their functional potential and ecology. The first genomes for oral TM7 emerged from single-amplified genomics studies (Marcy et al., 2007) and were followed by He et al.'s pioneering work that brought the first TM7 population into culture (He et al., 2015), establishing a deeper understanding of its relationship with an Actinomyces host. Additional recent cultivation efforts are proving successful in providing access to a wider variety of oral TM7 (Collins et al., 2019; Cross et al., 2019; Murugkar et al., 2019). Recent genome-resolved and single-amplified genomics studies have also produced genomes for oral GN02 and SR1 (Campbell et al., 2013; Espinoza et al., 2018), and recently the first targeted isolation of oral SR1 strains has been reported, but genomes were not produced (Cross et al., 2019). Despite the promise of these studies, our understanding of the ecology and evolution of these fastidious oral clades is incomplete.

Here we investigated phylogenetic and functional markers of niche partitioning of enigmatic members of the oral cavity, with a focus on members of the candidate phylum TM7. We used a metagenomic assembly and binning approach to recover metagenome-assembled genomes (MAGs) from the supragingival plaque and tongue dorsum of healthy individuals. Our genomes represented prevalent and abundant lineages that lack genomic representation in the HOMD and National Center for Biotechnology Information (NCBI) genomic databases, including members of the CPR. Using a multi-omics approach we show that oral TM7 species are split into plaque and tongue specialists, and that plaque TM7 phylogenetically and functionally associate with environmental TM7, while tongue TM7 associate with TM7 from animal guts. To assess the generality of our results we carried out read recruitment from approximately 200 tongue and 200 plaque Human Microbiome Project (HMP) samples; results confirm that the genomes we identified are prevalent, abundant, and site-specific. In order to associate MAGs with 16S rRNA sequences and hence associate MAGs with phylotypes previously identified based on 16S rRNA, we used long-read sequencing (nanopore sequencing). Our findings suggest that at least for TM7, dental plaque resembles non-host habitats, while tongue- and gut-associated TM7s are more strongly shaped by the host. In addition, our results shed light on other understudied members of the oral cavity, and allow for better genomic insight into prevalent, yet poorly understood members of the oral microbiome.

2.3 Results and Discussion

To create a genomic collection of oral microbes, we sampled supragingival plaque and tongue dorsum of seven individuals on four to six consecutive days. Shotgun metagenomic sequencing of the resulting 71 samples yielded 1.7 billion high-quality short-reads (Supplementary table 1a at doi:10.6084/m9.figshare.11634321). We independently co-assembled plaque and tongue samples from each individual to improve our ability to detect rare organisms and to minimize errors associated with single-assemblies (Chen et al. 2019c). The resulting 14 co-assemblies (7 people x 2 sites) contained 267,456 contigs longer than 2,500 nts that described approximately 1,163 million nucleotides and 1,554,807 genes (Supplementary table 1b at doi:10.6084/m9.figshare.11634321). To reconstruct genomes from these metagenomes we used a combination of automatic and manual binning strategies that resulted in 2,463 genome bins. Independent assembly and binning of metagenomes from similar habitats can result in the

recovery of multiple near-identical genomes (Raveh-Sadka et al. 2015; Delmont et al. 2018). To increase the accuracy of downstream analyses we employed only the 857 of 2,463 bins that were 0.5 Mbp or larger (Supplementary table 2g at doi:10.6084/m9.figshare.11634321), then removed redundancy by selecting a single representative for each set of genomes that shared an ANI > 99.8% (see Methods). This resulted in a final collection of 790 non-redundant genomes (Supplementary tables 2a-b, 3a-e at doi:10.6084/m9.figshare.11634321).

Automatic binning approaches can yield composite genomes that suffer from contamination, influencing downstream ecological and evolutionary insights (Shaiber and Eren 2019), even when single-copy core genes suggest the absence of an apparent contamination (Chen et al. 2019c). To minimize potential errors, we used anvi'o to manually inspect, and when necessary, further refine key genomes in our study by (1) visualizing the change in GC-content and gene taxonomy of each contig, (2) performing ad hoc searches of sequences in public databases, and (3) ensuring the agreement across all contigs with respect to sequence composition signal and differential coverage, the coverage of contigs by reads recruited from our metagenomes as well as metagenomes from other studies. In order to improve accuracy of genome assembly via analysis of differential coverage (Quince et al. 2017), we sampled each subject on at least 4 separate days. Our reproducible workflow includes each genome bin for interactive inspection (see Methods).

After removal of human host DNA-contamination, which accounted for 5%-45% of the reads per sample, competitive read recruitment revealed that the final list of genomes recruited 47% of the reads from our metagenomes, with a range of 10%-74% per sample. Confidently assessing the origins of the remaining short reads is difficult as many factors can explain unaccounted short reads including but not limited to the missing genomic context due to (1) host eukaryotic contamination, (2) poor assembly of strain mixtures, (3) incomplete metagenome-assembled genomes, and (4) mobile genetic elements such as viruses and plasmids that are often difficult to bin. A major driver of the variability we observed in the percentage of reads recruited by our MAGs across samples was the assembly quality, as we found a significant correlation (R^2: 0.67, p-value: $2e^{-18}$) between the percent of reads recruited by the assembled contigs and MAGs for each metagenome (Supplementary table 1a at doi:10.6084/m9.figshare.11634321, Figure 8). The

collection of 790 genomes recruited a significantly larger fraction of the reads in plaque metagenomes (51.6%) than in tongue metagenomes (38.3%) (z-score: 3.73, p-value: 0.0002), which may be partially due to the fact that a larger number of our genomes were derived from plaque samples (463 vs 327) (Supplementary table 2b at doi:10.6084/m9.figshare.11634321). Overall, despite variation between samples, our analysis shows that MAGs encompassed most of the microbial genomic content estimated to be included in each assembly, and represent a large (near 50%) portion of the reads after removal of human DNA.

2.3.1 Metagenome-assembled genomes reveal new lineages including members of the Candidate Phyla Radiation

In order to assess how taxons represented by our MAGs are distributed relative to known oral taxons, we performed a phylogenomic analysis using our genomes as well as the 1,332 genomes from the HOMD (accessed on August 1st 2018) (Supplementary table 6b at doi:10.6084/m9.figshare.11634321). Our strict criteria of inclusion of genomes with at least 18 of the 37 ribosomal proteins that we used for phylogenomics removed 539 genomes from the analysis, including 492 low completion (<70%) MAGs, 23 high completion (>=70%) MAGs, and 24 genomes from the HOMD. The 275 MAGs that passed this quality-control threshold covered much of the diversity at the abundant genera of the samples we collected, as evident by a comparison to taxonomic composition estimates using 16S rRNA amplicon data and metagenomic short-reads (Supplementary tables 2e-f, 4a-h, and 5a-j at doi:10.6084/m9.figshare.11634321, Supplementary Information file).

Some lineages contained members exclusively from our collection and not in the HOMD (Figure 2), including 51 genomes that we identified as members of the CPR, which formed a distinct branch, as expected (Figure 2). Our MAGs also included novel genomes from non-CPR lineages not represented in the HOMD (Figure 2). While some of these deeply branching MAGs clearly represent novel genomes, it is conceivable that others could be due to MAG contamination in which ribosomal proteins from distant populations were mixed together. To guard against this possibility we carried out three rounds of manual

refinement that benchmarked our genomes against multiple genomic and metagenomic resources (see Methods).

A large fraction of the CPR genomes in our collection belonged to the phylum Ca. Saccharibacteria (TM7; 43). The rest were affiliated with the phyla Ca. Absconditabacteria (SR1; 5) and Ca. Gracilibacteria (GN02; 3).

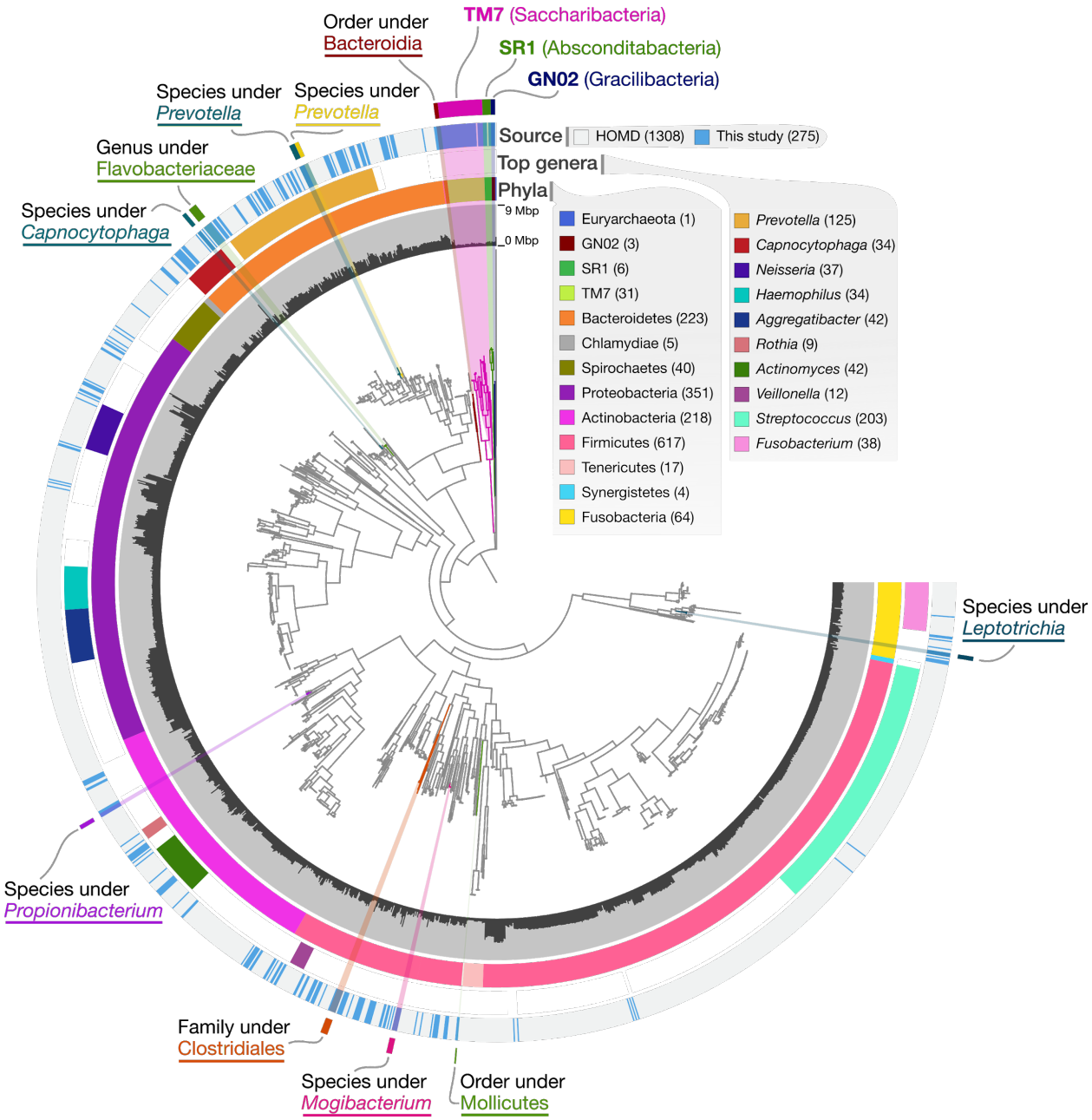

Figure 2: MAGs cover most of the abundant genera of the oral microbiome as well as represent lineages absent in public genomic databases. The dendrogram in the middle of the figure organizes 227 MAGs, 1582 genomes from the HOMD, and a single archeon, which was used to root the tree, according to their phylogenomic organization based on our collection of ribosomal proteins. The bars in the innermost circular layer represent the length of each genome. The second layer shows the phylum affiliation of each genome. The third layer shows the 10 most abundant genera in our samples as estimated by KrakenUniq. The fourth layer shows the affiliation of genomes as either MAGs from our study (blue) or genomes from HOMD (grey). The outermost layer marks novel genomes of lineages that lack representation in HOMD and NCBI. The lowest taxonomic level that could be assigned using CheckM and sequence search (see Methods) is listed for each novel lineage.

2.3.2 TM7 phylogenomic clades correspond to site of recovery

Our collection included 43 non-redundant TM7 MAGs (Supplementary table 2b at doi:10.6084/m9.figshare.11634321), presenting an opportunity to investigate associations between their lifestyles (i.e., cosmopolitan or site-specific) and their ancestral relationships. For this, we first examined the biogeography of TM7 populations by estimating their relative abundance in each of the 71 metagenomes through metagenomic read recruitment (Figure 3a, Supplementary tables 7a-c at doi:10.6084/m9.figshare.11634321). We defined a given TM7 population as detected in one of the 71 samples if at least 50% of the nucleotides of the genome were covered by at least one short read. We detected 42 of the 43 TM7 populations either only in plaque or only in tongue samples, but never in both (Figure 3a, Figure 9, Figure 10). The exception was T_C_M_Bin_00022, which we detected in 4/6 tongue samples and 6/6 plaque samples from participant C_M, but not in any other participant (Figure 2a). Interestingly, patterns of single nucleotide variations (SNVs) in samples of individual C_M suggest the existence of mixed sub-populations represented by T_C_M_Bin_00022 in tongue, while in plaque samples it appears monoclonal. To compare the variability of T_C_M_Bin_00022 we considered the 22,507 (of total 476,713) nucleotide positions at which both plaque and tongue samples had coverage of at least 20x, and found no variability in plaque samples, while there were 449 nucleotide positions (2%) in tongue samples that included variability, and where the ratio between the two competing nucleotides was at least 0.1 (median ratio 0.38), demonstrating intra-population diversity in tongue samples (Supplementary table 7t at doi:10.6084/m9.figshare.11634321). Other than this seemingly “cosmopolitan” population that was present in both tongue and plaque metagenomes, all TM7 genomes in our collection appeared to be specialists for plaque or tongue habitats.

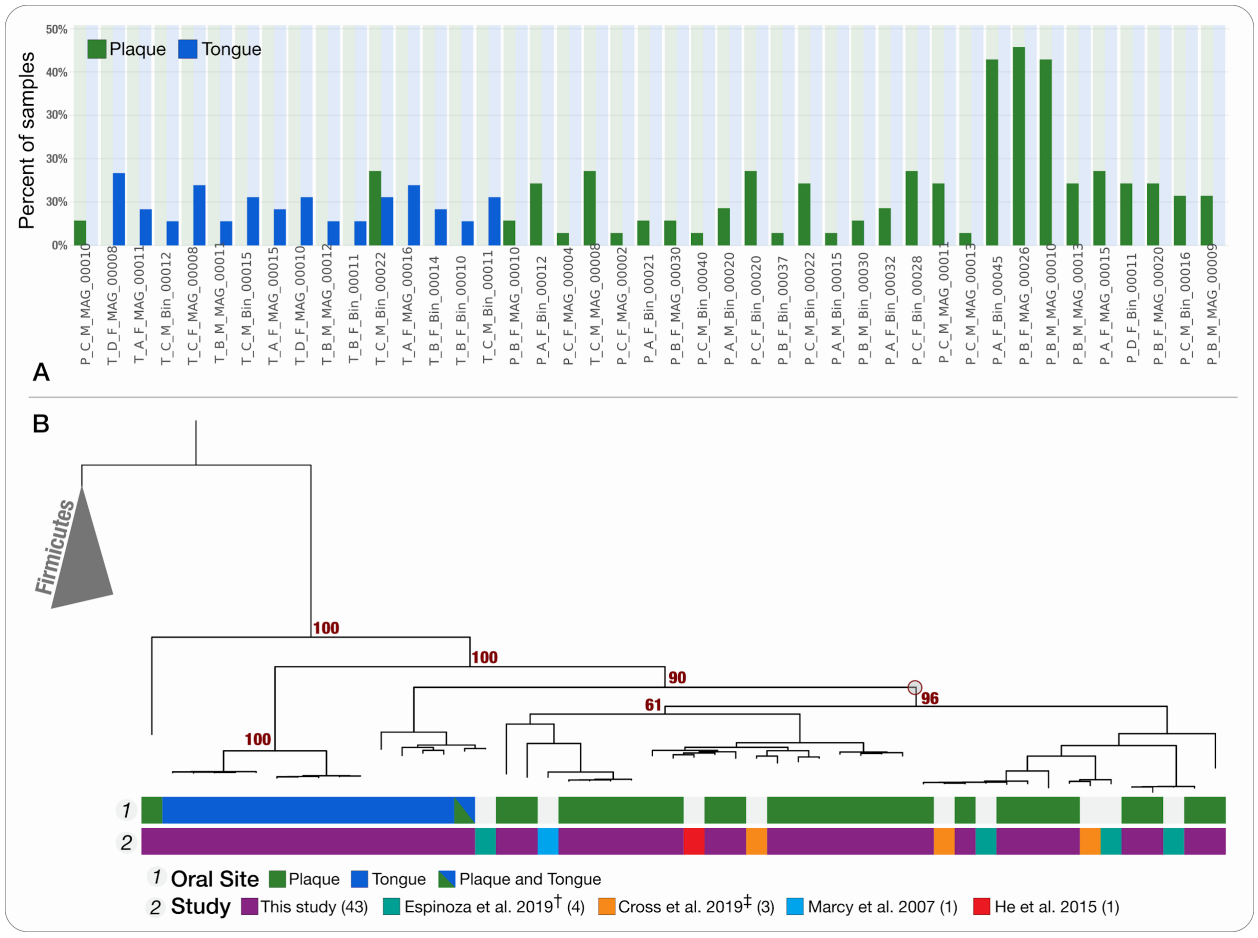

Figure 3: Detection of TM7 genomes across oral metagenomes and their phylogeny.(A) Most TM7 populations are exclusively detected in either tongue or plaque samples in our dataset. For each of the 43 MAGs (on the x-axis) the green and blue bars represent the portion of plaque and tongue samples, respectively, in which it is detected (detection > 0.5). (B) Phylogenetic organization of TM7 genomes reveals niche-associated oral clades. The phylogenetic tree at the top of the panel includes the 52 oral TM7 as well as 5 genomes of Firmicutes that root the tree. The layers below the tree describe (top to bottom): “Oral site” - the oral site to which each of our MAGs corresponded, where blue marks tongue dorsum, green marks supragingival plaque and a green-blue combination marks the “cosmopolitan” TM7; “Study” - the study associated with each genome: our MAGs (purple), Espinoza et al. 2019 (teal), Marcy et al. 2007 (blue), He et al. 2015 (red), and Cross et al. 2019 (orange). A red circle appears on the dendrogram and indicates the junction that separates the majority of plaque specialists from tongue specialists, and bootstrap values appear above branches that separate major clusters. † Refined versions of genomes, which we previously published (Shaiber and Eren 2019). ‡ Genomes from IMG that we refined in this study, but for which accession numbers for refined versions are available in Cross et al. 2019.

We then sought to investigate whether the ancestral relationships among TM7 genomes could explain their intriguing site-specificity. For this, we combined our 43 MAGs with 9 human oral TM7 genomes from the literature. In addition to 3 single amplified genomes that we downloaded from the Integrated Microbial Genomes and Microbiomes database (IMG/M) (Chen et al. 2019a) and refined (see Methods) and a MAG from Marcy et al. (Marcy et al. 2007), we included 4 MAGs from Espinoza et al. (Espinoza et al. 2018) after

manually refining composite TM7 genomes (Shaiber and Eren 2019), and the first cultivated strain of TM7, TM7x (He et al. 2015) (Supplementary table 7d at doi:10.6084/m9.figshare.11634321). The phylogenomic analysis of these 52 genomes separated tongue and plaque-associated genomes into distinct branches, where we could identify a single node on the tree that separated 41 of the 42 plaque associated genomes, suggesting that the site-specificity of TM7 is an ancestral trait. Another observation emerging from this analysis was that TM7x, which was cultivated from a saliva sample, clustered together with plaque-associated genomes, suggesting that its niche is most likely dental plaque rather than tongue (Figure 10).

2.3.3 TM7s found in plaque and tongue share exclusive ancestry with environment- and host-associated TM7s

Previous studies have shown that the human associated members of TM7 are polyphyletic, and cluster together with TM7 genomes of environmental origin (Camanocha and Dewhirst 2014; McLean et al. 2018). Taking advantage of the large number of genomes we have reconstructed, we revisited this observation by performing a phylogenomic analysis using all publicly available TM7 genomes in the NCBI's GenBank database as of 1/16/2019 (Figure 4). We identified six monophyletic human oral clades that were associated either with tongue (T1, T2) or plaque (P1, P2, P3, P4) (Figure 4). Using a pair-wise comparison of the average nucleotide identity (ANI) of oral TM7 genomes, we further identified sub-clades corresponding to genus and species level groups within the six monophyletic clades, including 12 species of TM7 represented each by at least 2 genomes in our collection (Figure 4, Supplementary tables 7f-h at doi:10.6084/m9.figshare.11634321, Supplementary Information). We then used a combination of long-read sequencing along with the phylogenetic analysis to compare our clades to the 6 previously described TM7 oral groups (G1-G6) based on 16S rRNA gene amplicons (Camanocha and Dewhirst 2014). We determined that our monophyletic clades T1, T2, and P4 correspond to G3, G6, and G5, respectively (Supplementary table 7e,i at doi:10.6084/m9.figshare.11634321). In contrast, clades P1, P2, and P3 all correspond to group G1, showing that G1 is likely composed of at least 3 distinct monophyletic oral clades. We have not recovered any MAGs for TM7 groups G2 and G4, which have been previously shown to have low prevalence as compared to other TM7 groups (B. Bor et al. 2019).

While tongue clades T1 and T2 clustered with genomes recovered from animal gut and together formed a deep monophyletic branch of an exclusively host-associated superclade shaded blue in Figure 4, plaque clades were interspersed with genomes from environmental sources (Figure 4). The exceptions to this clear distinction between plaque and tongue clades were T_C_M_Bin_00022, a cosmopolitan oral population that clustered within the clade T2, and the plaque-associated P_C_M_MAG_00010 (the only member of the clade P4) which was placed as a far outlier to all other oral TM7 and clustered together with genomes from animal gut (baboon feces). Animal-gut-associated genomes that grouped within the host-associated superclade were recovered predominantly from sheep and cow rumen samples, but also included genomes from termite gut, mouse colon, and elephant feces, suggesting an ancient association for members of the host-associated superclade and their host habitats (Figure 4, Supplementary table 7e at doi:10.6084/m9.figshare.11634321). Similarly, the inclusion of genomes recovered from dolphin dental plaque together with human-plaque-associated TM7 suggests an ancient association for plaque-specialists with the dental plaque environment. The phylogenetic clustering of tongue-associated TM7 genomes with TM7 genomes from animal gut, to the exclusion of environmental TM7, suggests that tongue and gut share a higher degree of ancestral relationship compared to those that are associated with plaque and with environments outside of a host. We know from previous studies that even though microbial community structures and membership in the human oral cavity and gut microbiomes are different, the ‘community types’ observed at these habitats are predictive of each other (Ding and Schloss, 2014), suggesting a level of continuity for host influences at these distinct sites that shape microbial community succession. Ancestral similarity between tongue- and gut-associated TM7s compared to those associated with non-host environments suggests that the host factors that influence microbial community succession may also have played a key role in the differentiation of host-associated and non-host-associated branches of TM7. We also know from previous studies that overall microbial community profiles in dental plaque dramatically differs from mucosal sites in the mouth with little overlap in membership (Eren et al., 2014; Segata et al., 2012). The strong ancestral associations between TM7 clades of plaque and non-host environments, as well as the depletion of plaque specialists from the host-associated superclade, suggest that from a microbial point of view, at least in the context of TM7, dental plaque resembles a non-host environment.

What led to the divergence of TM7 populations? Since TM7 have highly reduced genomes and have been found to be epibionts of other bacteria, primarily Actinobacteria (Bor et al., 2019; Kantor et al., 2013), one reasonable hypothesis is that the bacterial hosts of each TM7 clade are the drivers of the link between TM7 ecology and evolution. Such an hypothesis would imply that the similarity between tongue TM7 and gut TM7 is driven by the colonization of the gut and tongue environments by closely related bacterial hosts that provide a niche for TM7. Furthermore, it would imply the exclusion of such suitable hosts from the plaque environment, and vice versa, it would imply that plaque-specialist TM7 are dependent on bacterial hosts that are absent from the tongue and gut environments. In this context, it is notable that human oral Actinomyces species show strong site-specificity and little overlap in membership of dental plaque vs. tongue dorsum inhabitants (Mark Welch et al. 2019) and that Actinobacteria are rare in the human gut (Segata et al., 2012). An alternative hypothesis is that the mechanisms by which TM7 adapt to distinct habitats and distinct bacterial hosts are shaped by independent evolutionary events. While the existence of suitable bacterial hosts is likely an important factor, under this hypothesis, TM7 may acquire “local” bacterial hosts as they adapt to new environments. Our data are not suitable to evaluate either of these hypotheses. Yet given the ancestral similarity between dental plaque TM7 and TM7 from soils and sediments, it is conceivable to hypothesize that the dental plaque environment was able to support environmental TM7, while tongue and gut environments forced a distinct evolutionary path as suggested by the nested monophyletic superclade that is exclusively associated with host habitats. This depiction of TM7 evolution raises another question about the nature of dental plaque as a host habitat: why is dental plaque not as different from soil and sediment as tongue or gut? It is possible that fixed hard substrate of dental plaque renders it more similar to soils and sediments than to the constantly shedding epithelial surfaces of tongue and gut habitats from a microbial point of view. Whether dental plaque may have served as a stepping stone for environmental microbes by offering them a relatively safe harbor on the human body for host adaptation for some clades of human associated microbes is an intriguing question that warrants further study.

In summary, our data reveal the existence of at least 6 monophyletic oral TM7 clades with clear biogeography within the oral cavity, and a strong divide between the evolutionary history of host-associated

and non-host-associated TM7 genomes. Additionally, our analysis reveals 12 species of TM7 that are represented by multiple genomes in our collection and lays the groundwork for definition of taxonomic groups within this candidate phylum. The phylogenomic organization of genomes corresponds to their niche (tongue/plaque) in our dataset, suggesting a link between environmental distribution of these genomes and their evolutionary history in the context of ribosomal proteins.

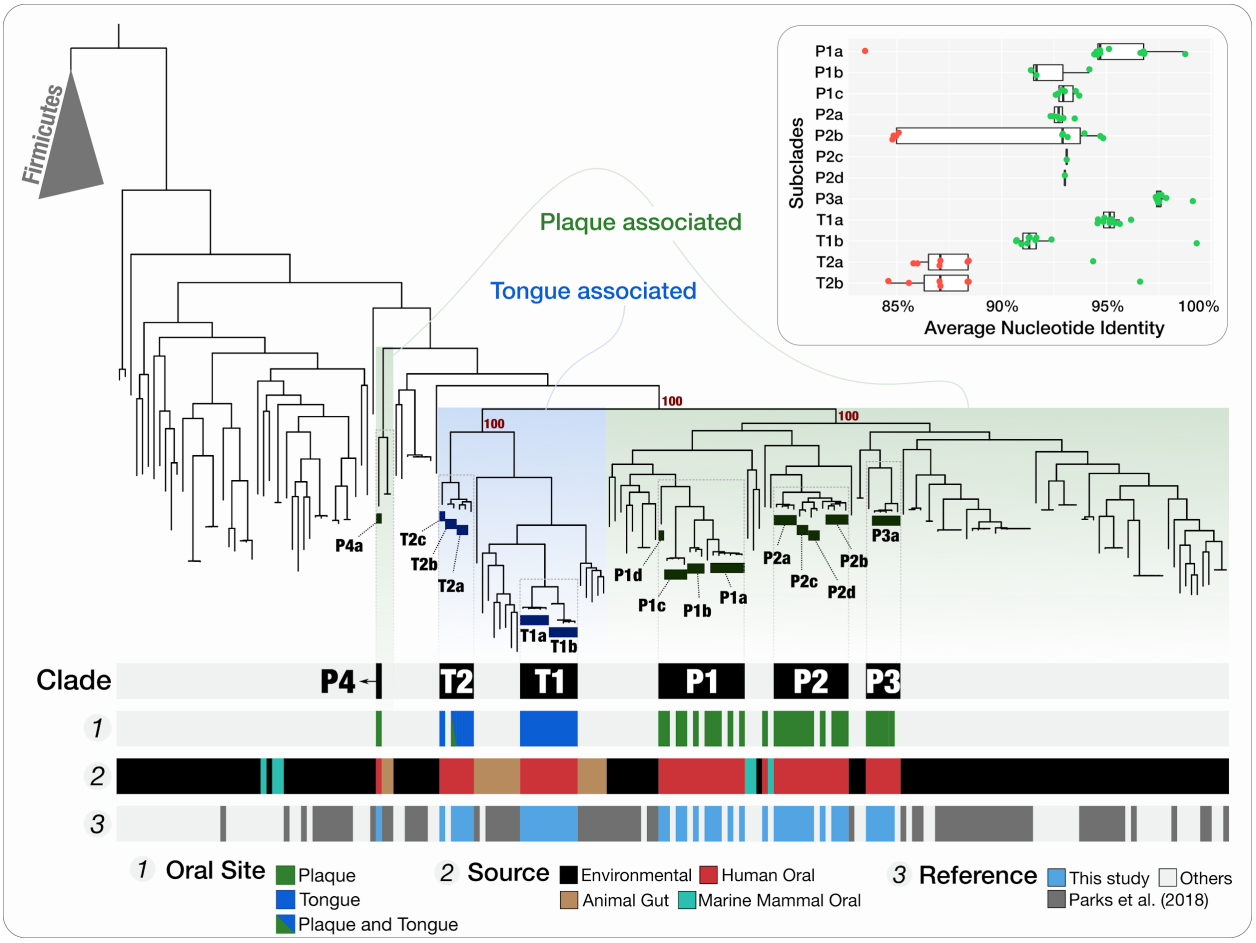

Figure 4: Phylogenetic analysis of human oral TM7 with all TM7 genomes on the NCBI's GenBank shows association of plaque TM7 with environmental genomes, and tongue TM7 with TM7 from animal stool. The phylogenetic tree at the top of the figure was computed using ribosomal proteins and includes 5 Firmicutes as an outgroup. Regions of the tree that are associated with either plaque or tongue clades from Figure 3 are marked with green or blue shaded backgrounds respectively. Bootstrap support values are shown next to branches separating major clusters of oral clades. Subclades are marked with rectangles below the branches they represent. The layers below the tree provide additional information for each genome. From top to bottom: Clade: the clade association is shown for each cluster of oral genomes. Oral Site: the oral site with which the genome is associated is shown for our MAGs in accordance with Figure 3. Source: the source of the genome, where red: human oral, brown: animal gut, cyan: dolphin oral, black: environmental samples. Reference: the genomes from this study in blue, and genomes from Parks et al. in grey (Parks et al. 2017). The majority of the rest of the genomes originate from various publications from the Banfield Lab at UC Berkeley. The insert at the top right of the figure shows boxplots for ANI results for genomes in each

Figure 4 (continued): subclades against all other genomes. Data points represent the ANI score for comparisons in which the alignment coverage was at least 25%. Within-subclade comparisons appear in green and between-subclades comparisons appear in red.

In summary, our analysis reveals 12 species of Saccharibacteria that are represented by multiple genomes in our collection and lays the groundwork for definition of taxonomic groups within this candidate phylum. The phylogenomic organization of genomes corresponds to their niche (tongue/plaque) in our dataset, suggesting a link between environmental distribution of these genomes and their evolutionary history in the context of ribosomal proteins. But the samples we used to generate our 43 Saccharibacteria MAGs represent only 7 individuals. We next sought to identify whether these patterns were representative of the distribution of TM7 among a wider cohort of healthy individuals.

2.3.4 Prevalence of TM7 across individuals is associated with TM7 clades, linking TM7 ecology and evolution

To assess the occurrence of these oral TM7 populations in a larger cohort of healthy individuals, we used a metagenomic short-read recruitment strategy to characterize the distribution of 52 oral TM7 genomes within 413 HMP oral metagenomes (with 30,005,746,488 pairs of reads) that included 196 samples from supragingival plaque and 217 tongue dorsum samples and were sampled from 131 individuals (Supplementary tables 7j-k at doi:10.6084/m9.figshare.11634321). We conservatively defined a genome to be present in a metagenome only if at least 50% of it was covered by at least one short read (see Methods). In addition to oral genomes, we also included three circular TM7 MAGs that were reconstructed from environmental samples and manually curated to circularity (Albertsen et al. 2013; Kantor et al. 2013; Brown et al. 2015). As expected, these 3 environmentally derived genomes (RAAC3, GWC2, and S_aal) were not detected in any oral metagenome (Figure 5, Supplementary tables 7l-n at doi:10.6084/m9.figshare.11634321). The occurrence pattern of TM7 genomes across the HMP individuals matched their occurrence in our seven participants, where all populations except the two genomes of sub-clade T2_b (T_C_M_Bin_00022, and TM7_MAG_III_B_1) were strongly associated with either tongue or plaque (Figure 5). Members of sub-clade T2_b indeed appeared to be cosmopolitan and were detected in both plaque and tongue samples (Figure 5). The most prevalent tongue-associated genome and plaque-associated genome were detected in samples from 45% and 50% of the HMP individuals, respectively

(Figure 5). In contrast, TM7x, the first cultured strain of TM7, was detected in only 5% of the HMP individuals. While the majority of the samples in the HMP dataset were taken from the tongue dorsum and supragingival plaque, there are additional oral sample types. Our analysis of these additional sample types suggested that certain TM7 populations have a preferential association with oral sites other than the tongue and supragingival plaque (Supplementary table 7o at doi:10.6084/m9.figshare.11634321, Supplementary Information file). Of particular notice, the single MAG of clade P4 (group G5), which was previously suggested to associate with periodontitis (Abusleme et al. 2013) appeared to associate with subgingival plaque, but occurred similarly in subgingival plaque metagenomes of patients with periodontitis and healthy individuals (Supplementary table 7p-s at doi:10.6084/m9.figshare.11634321). These results confirm that the exclusive association of most TM7 oral populations with either plaque or tongue is a general feature and not restricted to the participants of our study and reveal prevalent and abundant tongue and plaque specialists.

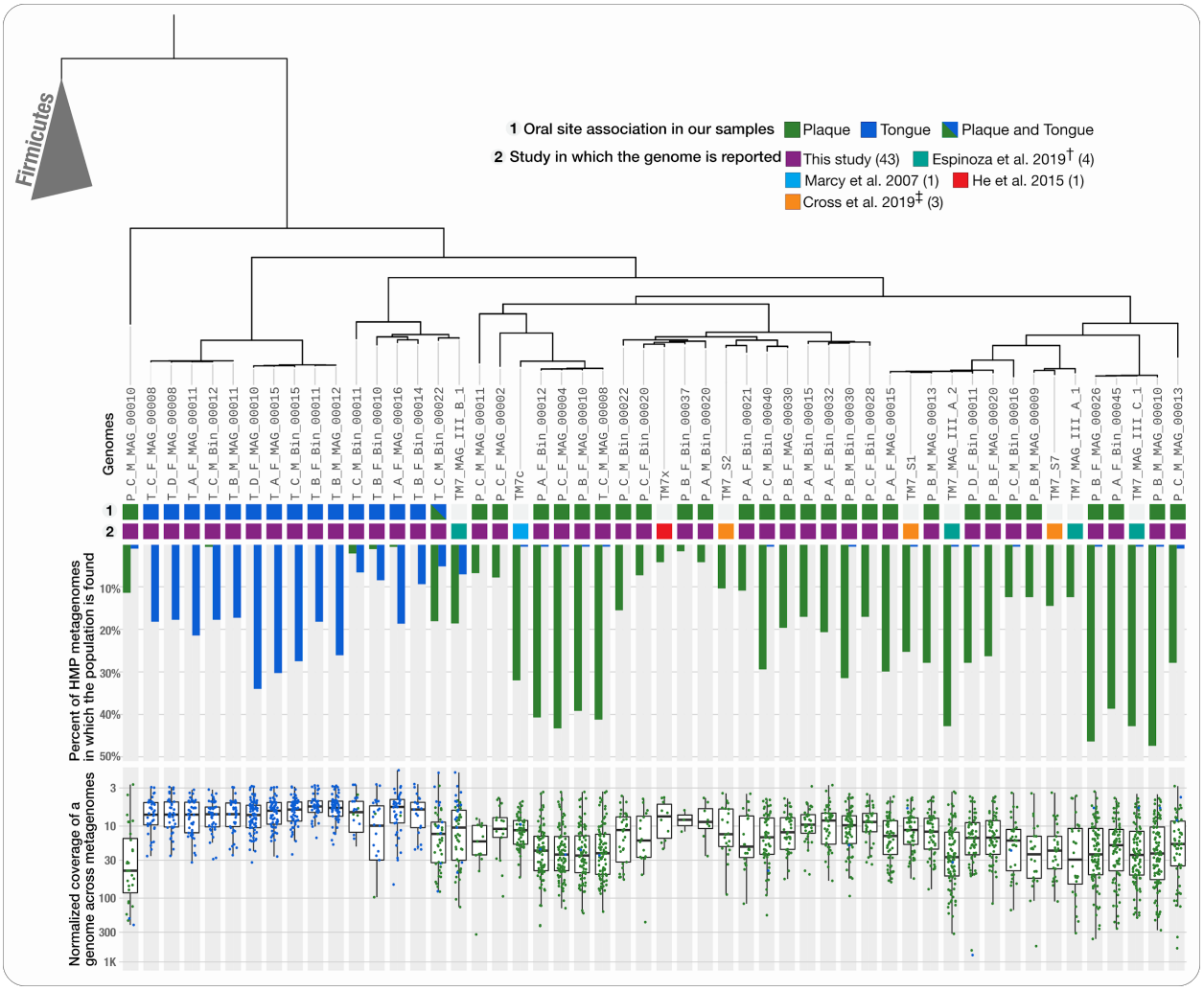

Figure 5: Detection and coverage of TM7 populations in the HMP plaque and tongue samples reveals abundant populations and niche specificity. The tree at the top of the figure and the two layers of information below it are identical to the one in Figure 3. Barplots below the tree show the portion of plaque (green) and tongue (blue) HMP samples in which each TM7 was detected, using a detection threshold of 0.5. Boxplots at the bottom of the figure show the normalized coverages of each TM7 in plaque (green) and tongue (blue) HMP samples in which it was detected.

2.3.5 TM7 pangenome reveals functional markers of niche specificity

We next sought to identify functional markers for the niche association of the plaque and tongue specialists. We utilized a pangenomic approach to identify functional determinants of niche specificity and investigate the functional differences between members of the various TM7 clades and subclades. Our analysis organized the total 40,832 genes across 55 genomes into 9,117 gene-clusters (GCs), 4,045 of which were non-singletons (i.e., occurred in at least 2 genomes) and included up to 162 homologous genes from the collection of 55 TM7 genomes described above (Figure 6, Supplementary tables 8a-b at

doi:10.6084/m9.figshare.11634321). The gene-clusters can themselves be clustered into groups that show similar distribution across genomes. By computing the hierarchical clustering of GCs based on their presence or absence in genomes we identified a collection of 205 core GCs that are found in nearly all genomes, as well as clusters of accessory GCs, many of which were exclusively associated with oral habitats or phylogenomic clades (Figure 6), confirming that the agreement between phylogenomics and ecology of these genomes was also represented by differentially occurring GCs. The proportion of genes with functional hits varied dramatically between the core and accessory TM7 genes. While more than 90% of core gene-clusters had functional annotations, COGs only annotated 29% of singletons, and 22% to 88% of non-singleton accessory gene-clusters (Supplementary table 8c at doi:10.6084/m9.figshare.11634321), revealing a vast number of unknown genes.

Whereas phylogenomics infers associations among genomes based on ancestral relationships, pangenomics reveals associations based on gene content (Dutilh et al. 2004), which can emphasize ecological similarities between genomes (Delmont and Eren 2018), primarily due to the fact that non-singleton accessory genes are the only drivers of hierarchical clustering based on gene content. The hierarchical clustering of TM7 genomes based on GCs predominantly matched their phylogenomic organization (Figure 12); however, it recapitulated their niche-association better than phylogenomics (Figure 12). Specifically, the plaque-associated genome P_C_M_MAG_00010 of the clade P4 (group ‘G5’), which is a distant outlier to all other oral TM7 according to phylogeny (Figure 3b), was placed together with all other plaque-associated TM7 (Figure 12). The data underlying this placement can be seen in the enrichment of P_C_M_MAG_00010 with GCs that belong to the 'Extended Core 2' cluster, generally characteristic of plaque TM7 and absent from tongue-associated TM7 such as clades T1 and T2 (Figure 6, Supplementary table 8c at doi:10.6084/m9.figshare.11634321). This enrichment appears to be responsible for the placement of P_C_M_MAG_00010 together with the other plaque-associated genomes and environmental genomes in Figure 12. In summary, these results show that the occurrence pattern of gene-clusters groups together phylogenetically-distinct clades of plaque-associated TM7s.

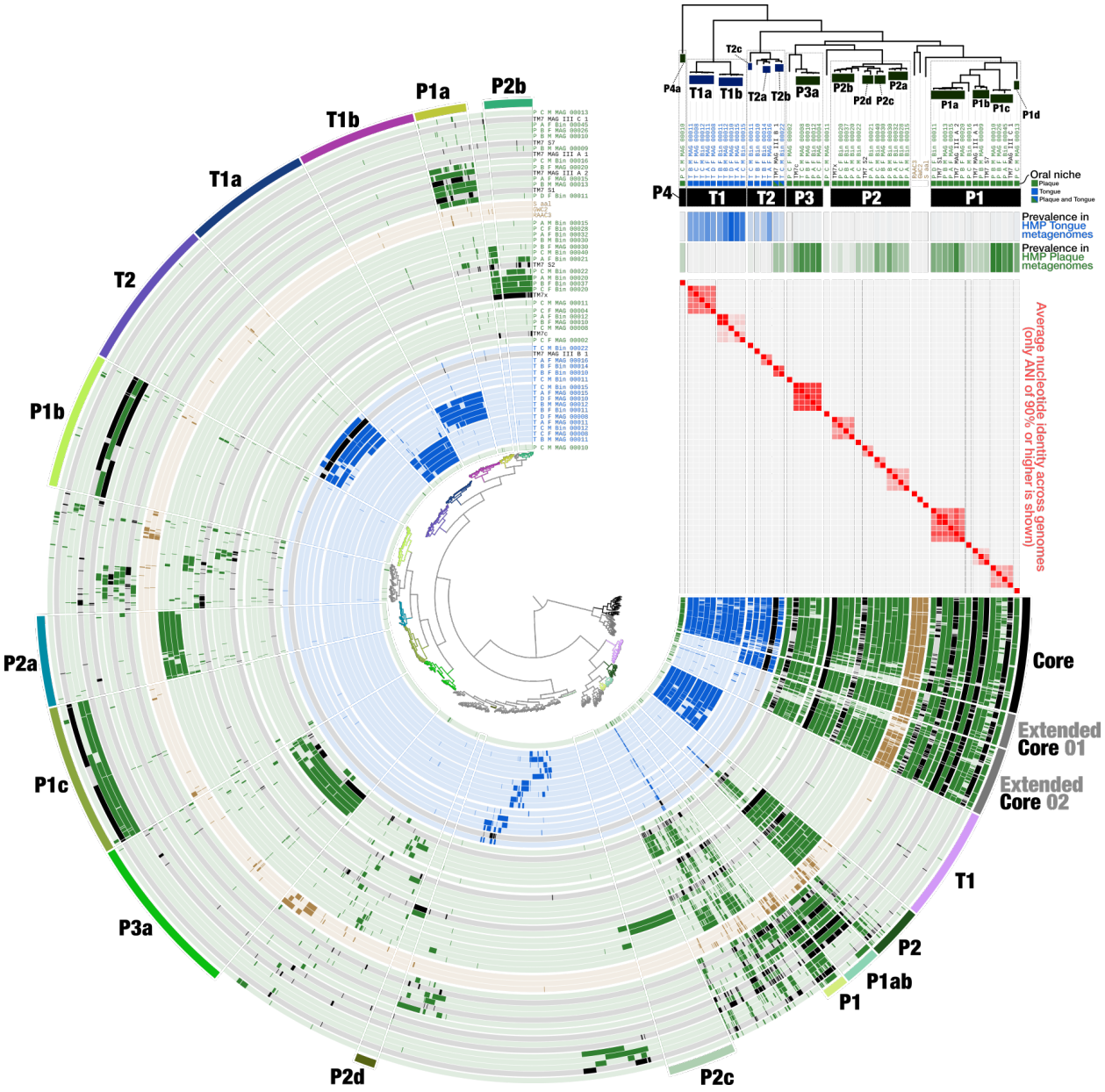

Figure 6: Pangenome of TM7 - Accessory gene-clusters include clade-specific and niche-specific markers. The dendrogram in the center of the figure organizes the 4,045 gene-clusters that occurred in more than one genome according to their frequency of occurrence in the 55 TM7 genomes. The 55 inner layers correspond to the 55 genomes, where our MAGs that associated with tongue and plaque are blue, and green, respectively; and previously published oral and environmental genomes are in black and brown, respectively. The colored regions in these 55 layers correspond to the presence of a gene-cluster in the corresponding genome. The circular layers of genomes are ordered according to their phylogenetic organization. The outermost circular layer highlights clusters of GCs that correspond to the core or to group-specific GCs. On the top right, the phylogenetic tree is shown and below it, the three top horizontal layers represent sub-clade, clade, and oral-site associations of genomes. The next three layers include statistics of coverage for each genome in the HMP oral metagenomes and show (from top to bottom) 1) the maximum interquartile mean coverage 2) occurrence in tongue samples 3) occurrence in supragingival plaque samples. The last two horizontal layers show the number of singleton GCs and the length for each genome.

The large number of TM7 genomes we recovered affords the opportunity to investigate key functional properties shared by all TM7s by examining the functions encoded by core GCs. As expected, the TM7 core GCs included many genes involved in translation, replication, and housekeeping (Supplementary table 8d at doi:10.6084/m9.figshare.11634321). The core GCs also included genes involved in amino-acid transport. Since TM7 lack the genes to produce their own amino acids (ref), these genes likely play an important role in scavenging amino acids from the environment or from the bacterial host. The core GCs also included several genes with potential roles in binding to the host, including components of a type IV pilus system that was identified in all genomes. Oral-associated TM7 have been shown to have a parasitic lifestyle in which they attach to the surface of their bacterial host (He et al. 2015; Cross et al. 2019), but the mechanism utilized for this attachment is unknown. Type IV pilus systems have been found to be enriched in CPR genomes as compared to other bacteria (Méheust et al. 2019) and were also specifically noted in TM7 genomes (Marcy et al. 2007). Type IV pilus systems are involved in many functions, including adherence (Craig, Forest, and Maier 2019), and could potentially be utilized by TM7 to attach to the bacterial host. Most of the components of the type IV pilus system we detected in the TM7 genomes occurred in a single operon with conserved gene synteny (Figure 7a). Additional copies of some of the type IV pilus proteins appear in various loci of the genome (Supplementary table 8a at doi:10.6084/m9.figshare.11634321). We found that while the cytosolic components of the type IV pilus system (PilT, PilB, PilC, PilM) were highly conserved across all genomes, components involved in the alignment of the system in the peptidoglycan (PilN) and the major and minor pilin proteins (PilE, and PilV) appeared in clade or sub-clade -specific gene-clusters and were completely absent from all genomes of clade T1 and from the single genome of clade P4 (Figure 7a, Supplementary table 8d at doi:10.6084/m9.figshare.11634321). Variability in PilV has been shown in the past to confer binding specificity (Ishiwa and Komano 2003) and in the case of TM7, the clade-specific nature of PilV and PilE sequences could be driven by host-specificity. While T1 genomes were lacking the components of the pilus system with known adhesive roles, they were highly enriched in proteins with a Leucine-rich repeat (LRR) (COG4886), which are often found in membrane bound proteins that are involved in adherence (Bella et al. 2008). 104 of the 207 proteins that were annotated with an LRR belonged to a single gene-cluster (GC_00000003) which was exclusively associated with T1 genomes, and each T1 genome had a total of

12-24 LRR proteins (COG4886) (Supplementary table 8a at doi:10.6084/m9.figshare.11634321). In summary, our analyses suggest that the diversity of pilin proteins could be driven by the host-specificity of TM7 species, and that TM7 species that lack pilin proteins could rely on alternative mechanisms such as LRR proteins for adherence.

Additional proteins that we identified to have a potential role in host attachment included proteins with a LysM repeat, which is a motif found in a wide range of proteins that are involved in binding to peptidoglycans (Buist et al. 2008). So far, the identified hosts of TM7 are all Gram-positive bacteria, and hence peptidoglycan binding could be a mechanism in which TM7 attach to their hosts. We found 33 GCs associated each with one of four COG functions that included LysM repeats and comprised a total of 169 genes (91 with COG0739, 6 with COG0741, 71 with COG1388, 1 with COG1652). We identified a Murein DD-endopeptidase MepM with a LysM domain (COG0739) in most genomes within a conserved operon, which included components of a Type IV Secretion system including VirB4 and VirB6 (Supplementary table 8a at doi:10.6084/m9.figshare.11634321). Similarly to what we observed for the type IV pilus system, the cytosolic component, Virb4, was highly conserved across all genomes, while the membrane bound Virb6 varied and appeared to be clade (and even sub-clade) -specific. This secretion system is also associated with motility in gram-positive bacteria (Marcy et al. 2007), and could potentially be used by TM7 for motion, and/or translocation from one host to another. We detected an additional protein with a LysM repeat (COG1388) in nearly all genomes. While in most genomes this proteins was flanked by genes involved in cell division, in the genomes of Clade T1_b, this locus included an insertion of 1-3 copies of a Leucine-rich repeat (LRR) protein, which as we mentioned above, also has a potential role in adherence. Overall, proteins with a LysM domain are common amongst oral TM7 and could provide another mechanism for attachment to the host surface.

The occurrences of functions across phylogenetic clades could reveal lifestyle differences that are not necessarily highlighted by the occurrences of gene-clusters. Since gene-clusters in a pangenome describe genes that are highly conserved in sequence space, identical functions can occur in distinct gene-clusters, rendering it difficult to describe core and accessory functions in a pangenome based on core and accessory gene-clusters. Here we developed a statistical approach that allows the identification of core and accessory

functions, and reveals enriched functions in any given subset of genomes in a pangenome (i.e., a phylogenomic clade). In this approach a logistic regression (binomial GLM) is fit to the occurrence of each COG function, using the clade affiliation as the explanatory variable. As this test is performed independently for each function, we computed q-values from p-values to account for multiple tests. We considered a function to be enriched if the q-value was below 0.05, hence setting the expected proportions of false discovery at 0.05. More information regarding this approach is available at http://merenlab.org/2016/11/08/pangenomics-v2/#making-sense-of-functions-in-your-pangenome.

Of the 972 unique functions, we identified 320 (34%) as the functional core, which included genes predominantly identified in all genomes, and 300 that were significantly enriched in specific clades (Figure 13 here and supplementary table 8q-r at doi:10.6084/m9.figshare.11634321). While there was a wide overlap between core functions and core GCs, 131 core functions occurred in clade-specific GCs, of these, 21 were exclusively associated with one GC from the 'Extended Core 1' cluster and one GC from the 'T1' cluster, further showing the uniqueness of clade T1 amongst the oral TM7 genomes. (Figure 13, supplementary table 8a at doi:10.6084/m9.figshare.11634321). Other cases also revealed functions that may have undergone selective pressure in a clade-specific manner. For example, a single copy of an RTX toxin-related Ca2+-binding protein, was highly conserved in nearly all genomes (gene-cluster GC_00000221), but appeared to have a slightly different variant in genomes P1_c (GC_00001826), and T2 (GC_00001332). Our examination of the top 100 most enriched functions revealed many membrane associated genes, including, but not limited to functions that were highlighted above by our examination of GCs (Supplementary table 8f at doi:10.6084/m9.figshare.11634321). For example, tongue and plaque clades appeared to be differentially enriched for transporters of ions and metals. Genes involved in respiration as well as genes involved in translation and stress-response were also differentially enriched for tongue and plaque clades. Overall, our analysis of the functional composition of oral TM7 shows that along with differences in accessory functions, sequence divergence of particular core genes distinguishes various clades, and in particular highlights members of clade T1 as outliers amongst the TM7 oral clades, matching their deep phylogenetic position. In addition, we identified functions that characterize tongue and plaque

clades and could provide targets for future endeavors to understand the unique biological features of members of each clade.

Overall, our data show that both accessory functions and core functions distinguish plaque and tongue specialists. While the core genome includes many functions common to all bacteria, it also includes many functions that are known to be enriched in CPR genomes. In particular, our data reveals proteins with potential roles in adherence, and suggests that while cytosolic components are highly conserved, extracellular proteins appear to be clade-specific, suggesting that interaction with the host and with the environment are important drivers in differentiating between TM7 oral clades. In addition, plaque-specialists that are phylogenetically distinct are functionally related and group together with environmental genomes based on GCs, while tongue-specialists group together with TM7 from animal gut. While members of clade T1 appear as outliers that differ both in functional composition and in the sequence divergence of many core functions as compared to other oral TM7, the functional composition of members of clade T2, which includes the cosmopolitan T2_b genomes, appears to represent an intermediate between the strictly host-associated group and the plaque/environmental group.

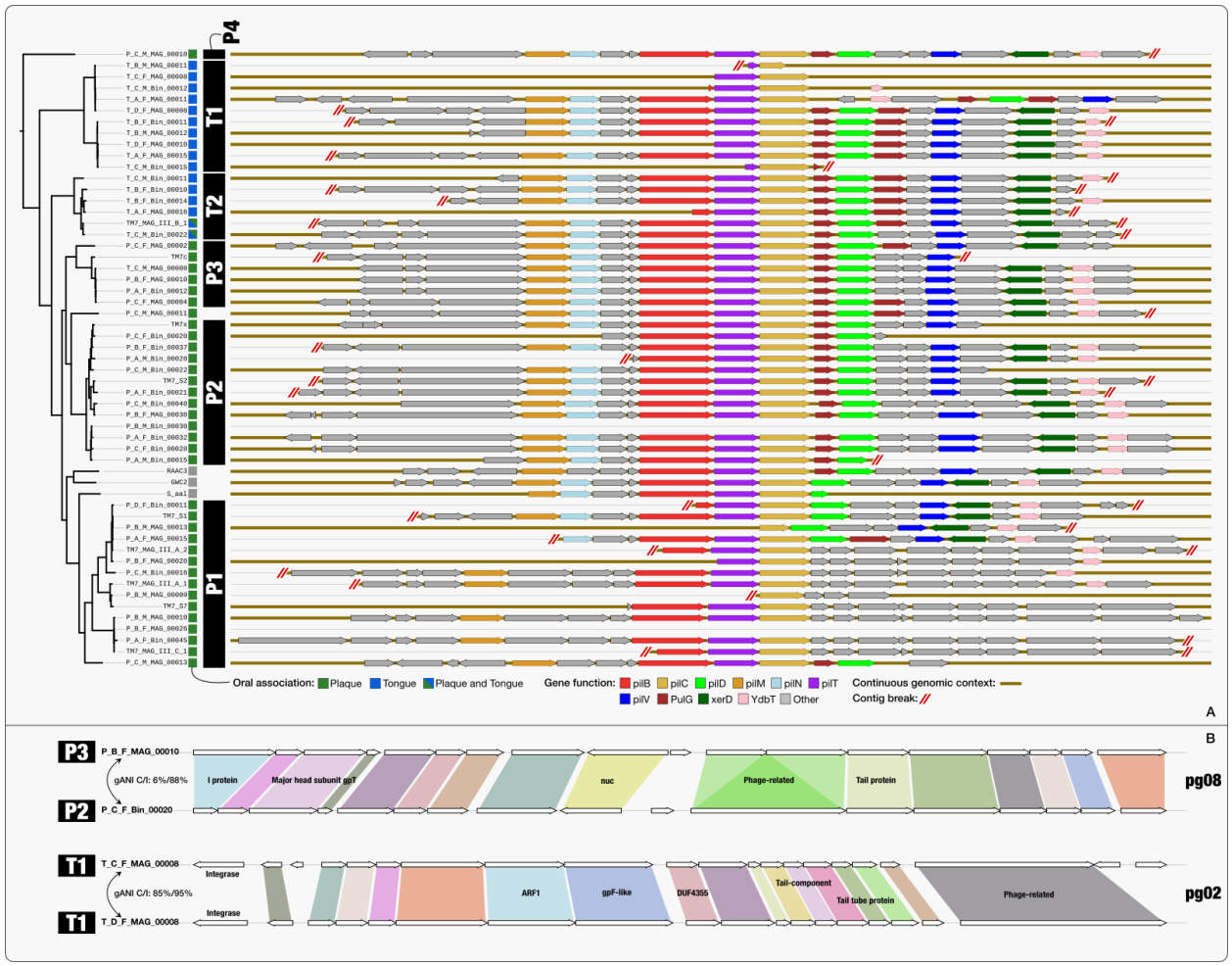

Figure 7: TM7 type IV pilus operon and TM7 prophages. A) Type IV pilus operon is highly conserved in TM7 genomes, but missing many components in genomes of the tongue-associated clade T1. Type IV pili operons from 52 of the 55 TM7 that included pilC are aligned according to pilC (yellow). Genomes are organized according to their phylogenetic organization shown in Figure 6. The top 10 functions identified in these operons appear with color filling, while the rest of the functions appear in grey. B) Some phage groups span phylogenetic clades, while other phage groups associate with specific clades. At the top of the panel the two prophages of phage group pg08 are compared and on the bottom of the panel the two prophages of the phage group pg02 are compared. White arrows signify genes as identified by Prodigal. Homologous genes, identified as belonging to the same gene-cluster, are connected by colored areas. A function name assigned by KEGG, COG or Pfam functional annotation source appears for genes for which it was available. On the left the phylogenetic clade of the TM7 host of each prophage is listed next to the host genome name. The genome-wide average nucleotide identity (gANI) appears for each pair of the host genomes, where C/I stands for alignment coverage / alignment identity.

2.3.6 Mobile elements and prophages in TM7 genomes

Little evidence for phage association with members of the CPR has been found so far (Chen, et al. 2019b). Dudek et al. recovered a phage associated with a TM7 genome from a dolphin plaque metagenome (Dudek et al. 2017) and Paez-Espino et al. identified phages with a predicted SR1 host (Paez-Espino et al. 2016)

in human oral metagenomes. A smaller genome size has been shown to correlate with the lack of lysogenic phages (Touchon, Bernheim, and Rocha 2016), and a lack of prophages in CPR genomes would fit this trend. To evaluate whether oral TM7 were indeed devoid of integrated prophages, we used an automatic approach based on VirSorter (Roux et al. 2015) and the recently published "inovirus detector" (Roux et al. 2019), along with a manual approach (see Supplementary Information), to identify 9 "phage groups" each composed of closely related prophages that were recovered from multiple TM7 genomes spanning all oral clades (Supplementary table 8g at doi:10.6084/m9.figshare.11634321). We did not identify any prophages in the three environmental genomes. Phage groups generally associated with closely related hosts but were not restricted to hosts of the same TM7 species, or even the same oral clades (Figure 7b, Supplementary table 8g at doi:10.6084/m9.figshare.11634321). A blast search of prophage nucleotide sequences against the NCBI's nr nucleotide collection returned no significant hits, confirming the novelty of these phage sequences. Using CRISPRCasFinder (Couvin et al. 2018) we identified CRISPR spacers targeting prophages of two "phage groups" in closely related hosts, validating the association of these prophages with their corresponding hosts. We identified CRISPR spacers and CRISPR related proteins in genomes representing clades P1, P2, P3, P4, and T2, but not in T1 nor in the three environmental genomes. The lack of CRISPR systems in the environmental TM7, despite their close affiliation with plaque TM7, raises the question whether these systems were recently acquired by oral clades. To investigate this hypothesis, we blasted cas9 proteins from 6 genomes representing all 5 CRISPR-containing clades, and found that these best matched cas9 protein from a variety of oral TM7 and a variety of Firmicutes, but no environmental TM7 nor any other CPR (Supplementary table 8p at doi:10.6084/m9.figshare.11634321). These results suggest that cas9 proteins might have been acquired by oral TM7 from Firmicutes. While some TM7 clades appear to lack CRISPR systems, we identified restriction modification (RM) systems in genomes representing all oral clades, including clade T1, as well as in the environmental genomes GWC2 and RAAC3 (Supplementary table 8a at doi:10.6084/m9.figshare.11634321). These RM systems could serve as alternative measures against foreign DNA for TM7 that lack CRISPR systems. Overall our data show that prophages are common amongst oral TM7, and appear to be a unique feature of oral TM7, while absent from environmental TM7. In addition, CRISPR systems appear to be common amongst specific clades of oral TM7, but not a common feature of all TM7. While additional analyses that include a larger collection of

environmental genomes will be required to verify this observation, a specific association of prophages with host-associated TM7 suggests that prophages may have played an important role in the adaptation of TM7 to the host environment, perhaps by facilitating horizontal gene transfer.

In search of other mobile genetic elements, we identified transposases in 18 TM7 genomes representing all oral clades and environmental genomes (Supplementary table 8n at doi:10.6084/m9.figshare.11634321). The varying location of the highly conserved transposases we identified in genomes of sub-clade T1_a suggests recent mobility, and that at least some of these elements are indeed active transposons (Supplementary tables 8a,o at doi:10.6084/m9.figshare.11634321). Blast search of genes annotated as transposases revealed that while the majority appear to be strongly associated with members of the CPR, two transposases had more close hits from non-CPR bacteria.

2.3.7 Additional members of the CPR are prevalent in the oral cavity, including a tongue-associated SR1

In addition to TM7, other members of the CPR have been commonly found in the human oral cavity, specifically members of the candidate phyla SR1 and GN02 (Camanocha and Dewhirst 2014). Using full length 16S rRNA, Camanocha and Dewhirst identified three clones corresponding to SR1 (HOT-345, HOT-874, and HOT-875) and three that corresponded to GN02 (HOT-871, HOT-872, and HOT-873) in the human oral cavity, of which, genomes have been previously published for all of these except SR1 HOT-875 (Camanocha and Dewhirst 2014; Campbell et al. 2013). While none of the GN02 and SR1 MAGs in our collection included 16S rRNA, which would allow a direct match to the Human Oral Taxon (HOT) designation, using a pangenomic analysis along with ANI statistics we were able to match MAGs to genomes representing HOT-871, HOT-873, HOT-345, and HOT-874 (Figure 14, Figure 17, Supplementary tables 9a-h at doi:10.6084/m9.figshare.11634321). Only a single tongue-associated SR1 (T_B_F_MAG_00004) did not match any previously published genome, and could potentially represent HOT-875, since it is the only known oral SR1 that currently lacks genomic representation. A recent study presented the successful isolation of an SR1 HOT-875, but a genome has not been sequenced (Cross et al. 2019).

In order to investigate the niche association of these CPR genomes, we mapped the short reads from the HMP metagenomes. While SR1 HOT-874 and HOT-345 were enriched in plaque samples, T_B_F_MAG_00004 was highly enriched in tongue samples, as it was detected in 37% of tongue samples (9% of plaque samples), and was highly abundant in some samples, recruiting 0.09% on average and up to 2.09% of the reads in tongue samples (Figure 15, Figure 16, Supplementary tables 9l-n at doi:10.6084/m9.figshare.11634321). Oral GN02 were all associated with plaque, and nearly absent from tongue samples (Figure 17, Figure 18, Supplementary tables 9i-l at doi:10.6084/m9.figshare.11634321). Our ANI analysis suggests that HOT-871 and HOT-872 represent the same genus as genomes from both of these lineages match with ANI>85% (alignment coverage>30%), while HOT-873 represents a separate genus and likely a separate family or order, as suggested by Camanocha & Dewhirst (Camanocha and Dewhirst 2014) (Supplementary tables 9e-f at doi:10.6084/m9.figshare.11634321). Overall our GN02 and SR1 MAGs extend the collection of genomes available for these under-studied members of the oral microbiome, and our analysis demonstrates their niche partitioning and reveals the prevalence of a tongue-associated SR1.

2.3.8 Novel non-CPR lineages represent prevalent members of the oral microbiome

Our collection included 34 MAGs that based on phylogenomics and blast sequence search represent 11 lineages with no representation on NCBI (from here on referred to as “novel MAGs”), and appear to include two unnamed species of the genus Prevotella, single unnamed species of the genera Mogibacterium, Propionibacterium, Leptotrichia, and Capnocytophaga each, as well as an unnamed genus in the family Flavobacteriaceae, an unnamed family within the class Clostridia, and unnamed families (and potentially unnamed orders) within the classes Bacteroidia and Mollicutes (Figure 2, Supplementary table 10a-d at doi:10.6084/m9.figshare.11634321, Supplementary Information file). Populations represented by these novel MAGs were absent from skin and gut samples, and in fact of our 790 MAGs, we found only two MAGs that were consistently detected in gut samples. Both of these MAGs belong to the species Dialister invisus, which were previously found to be the only abundant gut-associated microbes that were detected with considerable abundance in the oral cavity (Franzosa et al. 2014, Eren et al. 2014).

The oral microbiome is highly represented in genomic databases (Vartoukian et al. 2016; Nayfach et al. 2016), hence we next sought to check if the lack of genomic representation for these novel MAGs is due to low prevalence. We mapped short reads from the HMP metagenomes to these MAGs to estimate their prevalence and abundance across oral sites. Overall, these novel genomes presented strong tropism for either tongue or plaque, with the exception of three populations that appear to consistently recruit reads from both plaque and tongue samples, represented by the Flavobacteriaceae MAGs, T_A_M_MAG_00009 (Clostridiales), and three Capnocytophaga MAGs (Figure 20). While we found some populations to be rare, which could explain their lack of genomic representation in databases, other populations were extremely prevalent (Figure 21, Figure 22, Figure 23 Supplementary table 10e-h at doi:10.6084/m9.figshare.11634321). In addition to their high prevalence, some of these novel MAGs were highly abundant. P_B_M_MAG_00008 (Capnocytophaga) recruited on average 1% of the reads of plaque samples and two of the Propionibacterium MAGs recruited up to 18% of the reads of a single plaque metagenome, and on average 0.7% for plaque metagenomes (Supplementary table 10h at doi:10.6084/m9.figshare.11634321).

The most prevalent novel MAGs were five closely related MAGs of the family Flavobacteriaceae, which we detected in approximately 98.5% and 80% of HMP plaque and tongue samples, respectively, and reached high relative abundance, recruiting up to 2.98% of the reads of a single metagenome, and on average 0.19%, 0.62% of tongue, and plaque samples respectively (Supplementary tables 10e,g at doi:10.6084/m9.figshare.11634321). ANI comparison of these MAGs to each other and to representatives of all Flavobacteriaceae species on RefSeq suggested they represent a single new species in an unnamed genus, as within group ANI was >93.8% (with >80% alignment coverage), while they had no significant alignment with any other Flavobacteriaceae genome (Supplementary table 10i-j at doi:10.6084/m9.figshare.11634321). A phylogenomic analysis placed these MAGs in a subgroup of Flavobacteriaceae together with Cloacibacterium, Chryseobacterium, Bergeyella, Riemerella, Cruoricaptor, Elithabetkingia, and Soonwooa (Figure 23). While all five Flavobacteriaceae MAGs had high sequence similarity, both ANI results and the phylogenetic analysis clustered these genomes according to the site of recovery, suggesting the existence of a plaque and tongue-specific sub-population. Three of our

Flavobacteriaceae genomes were highly complete according to estimation by SCGs and were of length 1.7-1.8Mbp, considerably shorter than other Flavobacteriaceae genomes, as well as other commonly found oral microbes. The short length of these genomes as compared to other Flavobacteriaceae suggests a recent genomic reduction and possibly strong host-association. A strong host-association could lead to many auxotrophies and could explain why this species has never been isolated despite being an abundant and ubiquitous member of the oral microbiome. The recovery of novel genomes for these prevalent members of the oral microbiome could help shed light on their role and could assist future cultivation efforts.

2.4 Conclusions

Using genome resolved metagenomics, we have recovered much of the known diversity of the human oral cavity using samples from only 7 individuals, providing genomes for prevalent, yet uncultivated members of the microbiome, and highlighting phylogenetic and functional markers of niche partitioning of the cryptic candidate phylum TM7. Our findings group TM7 from the supragingival plaque with environmental TM7, both functionally and phylogenetically, while tongue-associated TM7 group together with lineages associated with animal gut, suggesting that at least for TM7, the supragingival plaque resembles non-host environments, while the tongue and gut TM7s are more strongly shaped by the host. Drivers of differentiation between the various microbial niches within the oral cavity are largely unknown, and could be revealed by applying similar approaches to study additional members of the oral microbiome.

2.5 Material and methods

Metagenomic assembly

Short reads from 71 metagenomes were quality filtered using the illumina-utils library (Eren et al. 2013) with the ‘iu-filter-quality-minoche’ program using default parameters, which removes noisy reads using the method described in (Minoche, Dohm, and Himmelbauer 2011). We then used MEGAHIT (D. Li et al. 2015) v1.0.6 to co-assemble the set of all quality filtered metagenomes originating from one oral site (either plaque

or tongue) of one donor, for a total of 14 co-assemblies. We used anvi-display-contigs-stats to get a summary of contigs statistics for each co-assembly.

To process assembly FASTA files we used the anvi'o contigs workflow which includes the following steps: we simplified the names of contigs in each one of the 14 assembly products using anvi'o (Eren et al. 2015) v5.5, and then used 'anvi-gen-contigs-database' to generate a contigs database in order to annotate the contigs. Briefly, anvi'o used Prodigal (Hyatt et al. 2010) v2.60 to find open reading frames. Centrifuge (Kim et al. 2016) was used to annotate genes with taxonomy. 'anvi-run-ncbi-cogs' was used to annotate genes with COG functions (Tatusov et al. 2000). 'anvi-run-hmms' was used to identify single copy core genes (SCGs) using a collection of built-in HMM profiles.

Metagenomic read recruitment, and initial automatic binning

In our metagenomic workflow we used Bowtie2 v2.3.4.3 (Langmead and Salzberg 2012) to recruit short reads from the set of metagenomes used for co-assembly to the assembly product; samtools (H. Li et al. 2009) was used to sort the output sam files into bam files; anvi'o was used to profile the bam files and compute coverage and detection statistics, and merge the profiles of each metagenomic sets. We then used CONCOCT (Alneberg et al. 2013) to create a preliminary collection of genomic bins. In short, CONCOCT uses coverage and composition to bin contigs together. We then used the anvi'o interactive interface to manually refine, using the method described below, the bins created by CONCOCT. Finally, we retained all MAGs of length greater than 0.5Mbp, and redundancy in SCGs below 10% for the rest of the analysis.

Sequence search

We used the NCBI nucleotide collection to search for nucleotide sequences, and the NCBI non-redundant protein sequences database to search for protein sequences. For 16S rRNA sequences, we used the online blast tool on the HOMD website (http://www.homd.org/?name=RNAblast&link=upload), where we used the 16S rRNA RefSeq Version 15.2 (starts at position 28) with default settings.

Manual bin refinement

We used the anvi'o interactive interface to refine our MAGs, as well as TM7 we downloaded from the IMG, which as previously reported (J. S. McLean et al. 2018), include contamination. Our refinement approach utilized the different clustering organizations available on the anvi'o interactive interface, which rely on sequence composition and differential coverage across multiple metagenomes. Our refinement was also assisted by the taxonomic assignments of contigs assigned based on Centrifuge annotation of genes. In cases in which we could not confidently distinguish contamination based on the clustering organizations, we used blast of specific sequences to assist us in making refinement decisions.

Refinement of our MAGs included between two to three rounds of refinement per MAG: 1) Refinement using the coverage information in the 4-6 samples used to assemble each MAG 2) Refinement of 63 MAGs which we identified as contaminated based on their coverage across our full collection of 71 metagenomes, and then used this coverage profile for refinement 3) Refinement of CPR MAGs and novel MAGs based on their coverage patterns in the HMP samples.

Refinement of TM7 genomes downloaded from IMG was done using coverage of their contigs across the HMP samples.

Naming scheme of MAGs

Names of the final MAGs included the prefix "ORAL", followed by a single letter to specify the type of samples used for the assembly of the MAG ("P" or "T" for plaque or tongue), followed by the ID of the individual (for example "C_M", which stands for "couple 'C', male"), followed by either "Bin" or "MAG" if the MAG had completion below or above 70% as estimated using the Campbell et al. collection of single copy core genes (SCGs) (Campbell et al. 2013), and followed by a number, where for each co-assembly the MAGs had a series of numbers from "00001" to the maximum number of MAGs that were retained from that co-assembly.

Removing redundancy and analysis of the non-redundant collection of MAGs

In order to identify near-identical MAGs, NUCmer (Delcher et al. 2002) was used to calculate the average nucleotide identity (ANI) between each pair of MAGs that were estimated by CheckM to belong to the same phylum. MAGs that had no phylum designation from CheckM were assigned phylum affiliation using phylogenomics (see below) and blast of protein sequences against the NCBI's non-redundant database. We determined that a pair of MAGs are redundant if their ANI was 99.8% with the alignment length covering at least 50% of the shorter of the two genomes. For each group of redundant genomes, the genome with the highest 'completion minus redundancy' was chosen as the representative of the group, where completion and redundancy were calculated by anvi'o based on single-copy core genes. If multiple redundant genomes had the same 'completion minus redundancy' then the longest genome was chosen.

We merged the sequences of the collection of non-redundant bins into one FASTA file, and processed this FASTA file using the anvi'o contigs workflow as mentioned above. We then also used this FASTA file to recruit reads from all 71 metagenomes, and used the anvi'o metagenomics workflow as mentioned above to generate a merged profile database. We used anvi-split to generate a profile database and contigs database for each MAG, followed by 'anvi-interactive' and inkscape in order to generate PNG images for all MAGs with contigs organized using a combined metric of differential coverage and sequence composition, and data points showing interquartile values of the mean coverage of contigs. We used these images to identify MAGs that required additional refinement.

Read recruitment from public metagenomes

We used 'anvi-run-workflow' with the 'metagenomics' workflow to recruit reads from oral samples of the Human Microbiome Project (HMP) (Human Microbiome Project Consortium 2012). The metagenomics workflow of 'anvi-run-workflow' uses Snakemake (Köster and Rahmann 2012) to execute the steps described above for our metagenomic read recruitment analysis. We used the same approach to also recruit reads from previously published metagenomes from periodontitis patients (Califf et al. 2017) to the

TM7 pangenome. The raw metagenomes of Califf et al. were obtained directly from the authors since the FASTQ files published by Califf et al. included only a single read for each pair of raw reads.

Quantifying human contamination in metagenomes

We ran the aforementioned metagenomics workflow using anvi-run-workflow and used the human genome build 38 (GRCh38) from NCBI to quantify the number of reads matching the human genome in each sample. We estimated the number of reads that originate from microbes (or “non-human” reads) in each sample as the total number of reads minus the number of reads that mapped to the human genome.

Relative abundance estimations of MAGs

For each MAG we used the number of reads that mapped to it, divided by the total number of non-human reads as the unnormalized abundance. All unmapped reads were counted as an UNKNOWN bin. In order to account for different genome lengths, which is expected to impact the number of reads expected from each population at a given true abundance, we divided each normalized abundance by the genome length. Since the genome length is unknown for the UNKNOWN bin, as it represents an agglomeration of whole genomes and portion of genomes that we did not recover, we used an arbitrary choice of 2Mbp as the normalization factor. The choice of this arbitrary factor changes the overall estimation of the portion of unknown reads, but not the observed trends.

Taxonomic profiles of metagenomes based on short reads

We used KrakenUniq (Florian P. Breitwieser and Salzberg 2018) to generate taxonomic profiles for all metagenomes. Briefly, KrakenUniq uses counts of unique k-mers to estimate the relative abundance of taxa in a sample, based on short-reads.

Phylogenomic analyses

For phylogenomic analysis we used our collection of 37 ribosomal proteins, which are in the overlap of the bacterial and archaeal single copy core gene collections created by Campbell et al. (Campbell et al. 2013) and Rinke et al. (Rinke et al. 2013): Ribosom_S12_S23, Ribosomal_L1, Ribosomal_L10, Ribosomal_L11, Ribosomal_L11_N, Ribosomal_L13, Ribosomal_L14, Ribosomal_L16, Ribosomal_L18e, Ribosomal_L18p, Ribosomal_L19, Ribosomal_L2, Ribosomal_L21p, Ribosomal_L22, Ribosomal_L23, Ribosomal_L29, Ribosomal_L2_C, Ribosomal_L3, Ribosomal_L32p, Ribosomal_L4, Ribosomal_L5, Ribosomal_L5_C, Ribosomal_L6, Ribosomal_S11, Ribosomal_S13, Ribosomal_S15, Ribosomal_S17, Ribosomal_S19, Ribosomal_S2, Ribosomal_S3_C, Ribosomal_S4, Ribosomal_S5, Ribosomal_S5_C, Ribosomal_S6, Ribosomal_S7, Ribosomal_S8, Ribosomal_S9. To compute phylogenetic trees based on these ribosomal proteins, we used 'anvi-run-workflow' with the 'phylogenomics' workflow. The phylogenomics workflow included running 'anvi-get-sequences-for-hmm-hits' to export a FASTA file with the concatenated and aligned ribosomal proteins with the following parameters: '--align-with famsa' to perform alignment of protein sequences using FAMSA (Deorowicz, Debudaj-Grabysz, and Gudyś 2016); '--concatenate-genes' to concatenate the sequences of the various ribosomal proteins; '--return-best-hit' to instruct the program to return only the best hit in case that a single HMM profile had two hits in one genome; `--get-aa-sequences' to output amino-acid sequences; '--hmm-sources Campbell_et_al' to use the Campbell_et_al HMM source (Campbell et al. 2013) to search for genes. For Figure 2 we also included the parameter '--max-num-genes-missing-from-bin 19' to only include genomes that contain at least 18 of the 37 ribosomal proteins. For the rest of the phylogenomics analyses we used '--min-num-bins-gene-occurs' to ensure that only ribosomal proteins that occur in at least 50% of the genomes are used for the analysis. The resulting alignments were trimmed using trimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) with the setting '-gt 0.5' to remove all positions that were gaps in more than 50% of sequences, and a maximum likelihood phylogenetic tree was computed using IQ-TREE (Nguyen et al. 2015) with the 'WAG' general matrix model (Whelan and Goldman 2001). Phylogeny of CPR genomes was computed with only 36 of the 37, excluding Ribosomal_L32p since it was absent from all TM7 genomes. In order to root phylogenetic trees we used an outlier genome in each analysis: for Figure 2 we used a genome of the archeal Methanobrevibacter oralis, and for all other phylogenomic analyses we used a collection of five members of the Firmicutes: Acidaminococcus intestini, Eubacterium rectale, Staphylococcus aureus, Streptococcus pneumoniae,

Veillonella parvula. To remove the Firmicutes from the trees in Figure 6, Figure 14, Figure 17we used the python package ete3 version 3.1.1 (Huerta-Cepas, Serra, and Bork 2016).

Processing publicly available genomes

To process FASTA files, we used 'anvi-run-workflow' with the 'contigs' workflow, which includes the steps of the anvi'o contigs workflow as described above. In order to generate the data in supplementary table 8a (at doi:10.6084/m9.figshare.11634321), our workflow also included running 'anvi-run-pfams' to annotate functions with Pfams (El-Gebali et al. 2019), and we used 'anvi-get-sequences-for-gene-calls' to get all protein sequences and used GhostKoala (https://www.kegg.jp/ghostkoala/) to annotate genes with KEGG functions (Kanehisa, Sato, and Morishima 2016).

Assessing the occurrence of populations in metagenomes

We used anvi-mcg-classifier with the settings '--get-samples-stats-only', '--alpha 0.1', which determines a threshold of 0.6 detection value for to determine occurrence, '--zeros-are-outliers', which considers positions with zero coverage as outlier coverage values when computing the non-outlier mean coverage. We used the anvi-mcg-classifier output to determine the occurrence of TM7 populations in our collection of 71 metagenomes. In order to account for the different number of reads per sample when comparing non-outlier mean coverage values, we normalized these values. To compute the normalization factor, we first divided the number of reads in each sample by the maximum number of reads in the biggest sample (so that the normalization factor would be ≤ 1 for all samples). We then divided the non-outlier mean coverage values in each sample by the normalization factor.

Pangenomic analyses

We used ‘anvi-run-workflow’ with the ‘pangenomics’ workflow to compute the pangenome. In this workflow, we used ‘anvi-gen-genome-storage’ to generate a genomes storage database. ‘Anvi-pan-genome’ accepts the genomes storage as input and uses BLAST (Altschul et al. 1990) to get similarity scores for all protein sequences of each pair of genes. Similarity scores are then used to form clusters of genes using the Markov Cluster algorithm (MCL) (Enright, Van Dongen, and Ouzounis 2002) using the default parameters of anvi-pan-genome (minbit of 0.5, and MCL inflation of 2). We used ‘anvi-script-add-default-collection’ to add a collection that includes all GCs, and then used ‘anvi-summarize’ to create a summary table. For the TM7 pangenome in Figure 6, when running ‘anvi-summarize’, we used the collection of GCs that we created by manual selections in the interactive interface. For visualization of pangenomes, we created a second pangenomic database using ‘--min-occurrence 2’ to exclude singleton GCs (GCs that occur only in a single genome), and used ‘anvi-display-pan’ to run the anvi’o interactive interface.

Average nucleotide identity (ANI)

We used anvi-compute-ani with the settings ‘--method ANIm’, in order to perform alignment using MUMmer (NUCmer) (cite), and ‘--min-alignment-fraction 0.25’ to only keep scores if the alignment fraction covers at least 25% percent of both genomes. For the ANI data presented in Figure 6, we first computed ANI without the flag ‘--min-alignment-fraction’ to get all alignment statistics, and then we imported ANI values only for pairs of genomes with alignment coverage of at least 25%.

Extraction of 16S rRNA sequences

To export all 16S rRNA sequences from contigs databases we used ‘anvi-get-sequences-for-hmm-hits’ with parameters ‘--hmm-sources Ribosomal_RNAs’ and `--no-wrap’.

Analysis of nanopore sequences

In order to filter human contamination, we mapped long read sequences to the human genome using minimap2 (H. Li 2018). The remaining contigs were used to generate anvi’o contigs databases as described

above. Sequences of 16S rRNA were extracted and blasted against HOMD, and the results were used to assign group affiliation to TM7 genomes as described below.

Group affiliation of TM7 based on 16S rRNA

We exported ribosomal RNA sequences from all TM7 genomes, including ones downloaded from NCBI. We then blasted 16S rRNA sequences against the eHOMD as explained above. For each genome, we identified the group affiliation (G-1, G-2, etc.) of the closest hit on HOMD. In addition, we blasted nanopore reads that matched to TM7 against the collection of oral TM7 genomes. We used blast hits to associate TM7 MAGs with a 16S rRNA group affiliation. The 16S rRNA group affiliations are summarized in Supplementary table 7i for oral genomes, and in Supplementary table 7e at doi:10.6084/m9.figshare.11634321 for the all TM7 downloaded from NCBI.

Functional enrichment analysis

We used 'anvi-get-enriched-functions-per-pan-group' to find enriched functions per TM7 clade. This program fits a logistic regression (binomial GLM) to the occurrence of each COG function across genomes, using clade affiliation as the explanatory variable. It tests for equality of proportions across clade affiliation using a Rao score test, which gives a test statistic ("enrichment score") and p-value. q-values are estimated from p-values using the R package "qvalue" (Storey, Taylor, and Siegmund 2004). We considered a function to be enriched if the q-value was below 0.05; this controls the expected proportion of false positives at 0.05. More details on how to use this method are provided here: http://merenlab.org/2016/11/08/pangenomics-v2/#making-sense-of-functions-in-your-pangenome.

Identifying prophages in TM7 genomes

We used Virsorter (Roux et al. 2015) and the "Inovirus detector" (Roux et al. 2019) to identify contigs that include phage sequences. Contigs predicted as viral were manually inspected, and all contigs which gene

content was also consistent with a plasmid or another mobile genetic element, i.e. did not include either a viral hallmark gene or capsid-related gene(s) were excluded.

We further examined all remaining contigs to verify their placement in the prospective genomes, using the data in Supplementary table 8a at doi:10.6084/m9.figshare.11634321, as well as blast searches of protein sequences (see the notes in Supplementary table 8g at doi:10.6084/m9.figshare.11634321 for more details). We used functional annotations to identify additional contigs containing phage-related functions that were not identified by VirSorter/Inovirus detector. In addition, we identified additional phages by searching for contigs with many homologs (according to GC occurrence) to the identified phages. We repeated this process recursively and identified 11 more contigs that contain partial or complete prophages.

To identify start and end positions of prophages, we relied on identifying genes that appear to be TM7 genes as per their association with GCs. When possible, we used closely related TM7 genomes that lacked the prophage genes, to identify the position of the genes flanking the prophage, and hence confirming the insertion position of the prophage.

Identifying CRISPRs

We used the web service CRISPRCasFinder at https://crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index (Couvin et al. 2018) to search for CRISPR spacers in the 55 TM7 genomes. Along with a summary of the results, the web application allows the direct download of a FASTA file of all high confidence spacers (evidence level 3 or 4 as defined by Couvin et al). We used the FASTA file of high confidence spacers to blast spacer sequences against the TM7 genomes.

Statistics and visualization

We used ggplot2 version 3.2.1 to generate boxplots and barplots of abundances, as well as barplots of occurrences across metagenomes. To compare the number of reads recruited by our MAGs from our

plaque and tongue metagenomes, we ran a two-sided Z-test, using the Python package statsmodels (Seabold and Perktold 2010).

Access to previously published sequences

We downloaded all oral genomes from the HOMD FTP site (ftp://ftp.homd.org/HOMD_annotated_genomes/, and ftp://ftp.homd.org/NCBI_annotated_genomes/). Notice that while the TM7 genomes we downloaded from IMG had no accessions associated with them at the time we accessed them on the IMG, there have since then been refined versions of these genomes published and accession numbers for these refined genomes are available in Cross et al. 2019.

We used ncbi-genome-download (https://github.com/kblin/ncbi-genome-download) to download genomes from GenBank. We used anvi-script-process-genbank-metadata to process the metadata produced by ncbi-genome-download, and generate input files that we then used to run the contigs workflow of anvi-run-workflow. TM7 genomes from GenBank were downloaded on 1/16/2019; GN02 and SR1 on 12/17/2018; Flavobacteriaceae on 9/20/2019; Clostridiales on 9/25/2019;

2.6 Supplementary Material

2.6.1 Supplementary Figures

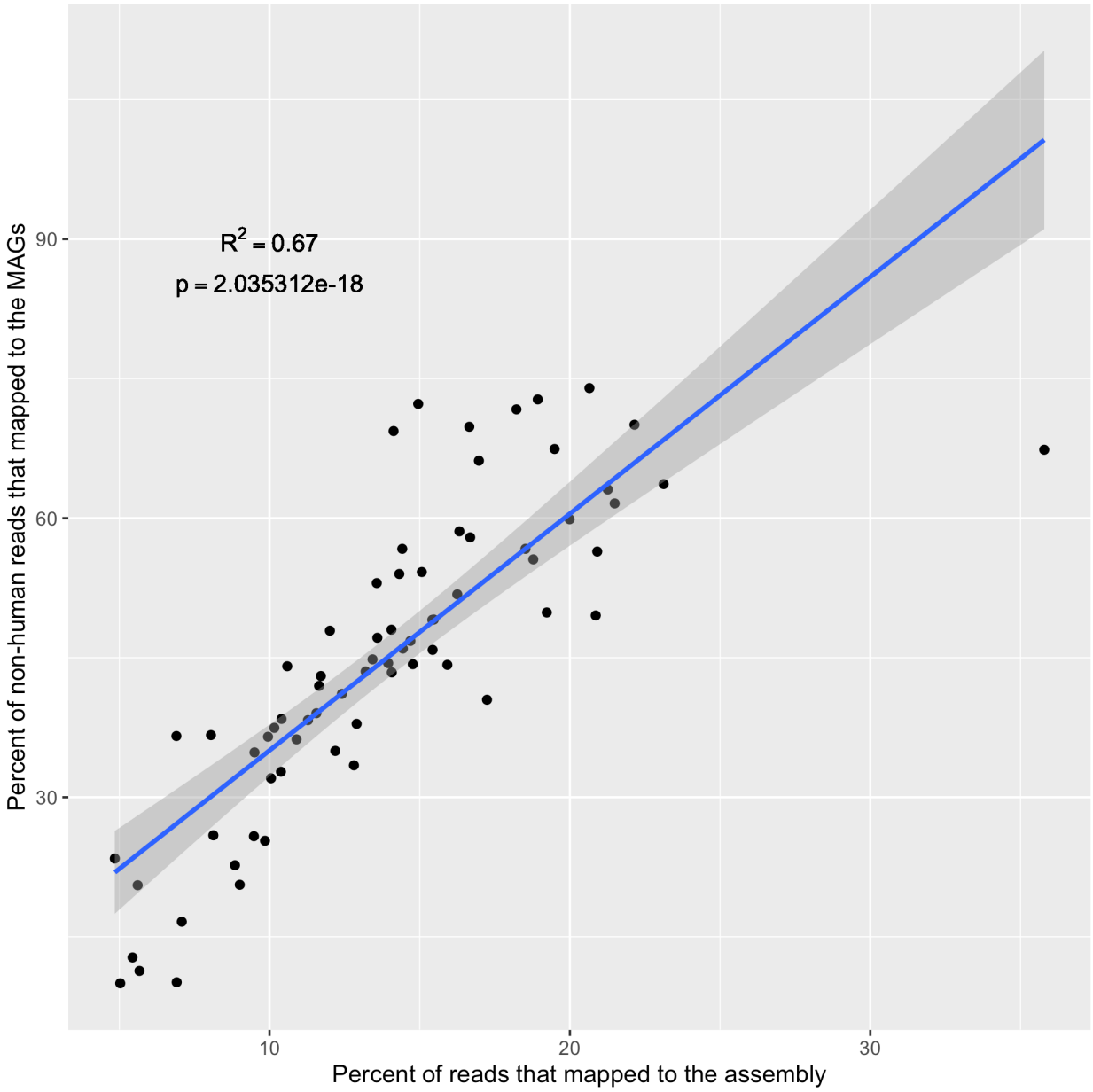

Figure 8: The percent of reads that map to MAGs is correlated with the quality of the assembly. The percent of reads that mapped to the non-redundant collection of MAGs out of the total number of reads, excluding reads that mapped to the human genome is presented for each of the 71 metagenomes as a function of the percent of reads that mapped to all contigs in the assembly. Blue curve represents a linear regression model with the grey shaded area marking the 95% confidence intervals. R-squared value and p-value for the linear regression appear above the curve.

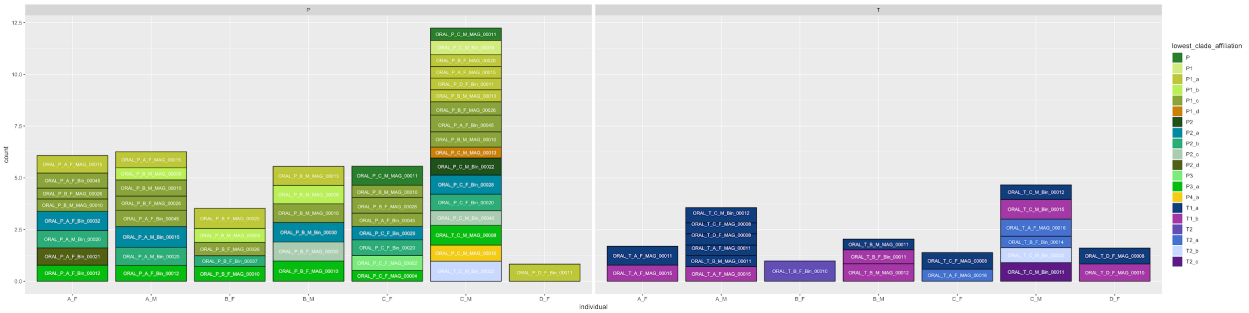

Figure 9: Normalized relative abundances of TM7 population per individual for the participants of our study. For those cases in which multiple closely related populations were recovered from multiple participants, each population is detected only in the participant from which it was recovered. The exceptions are when a closely related population exists, but assembly or binning failed to recover this population. In those cases of assembly/binning failure, each of the closely related population is recovered with similar abundance

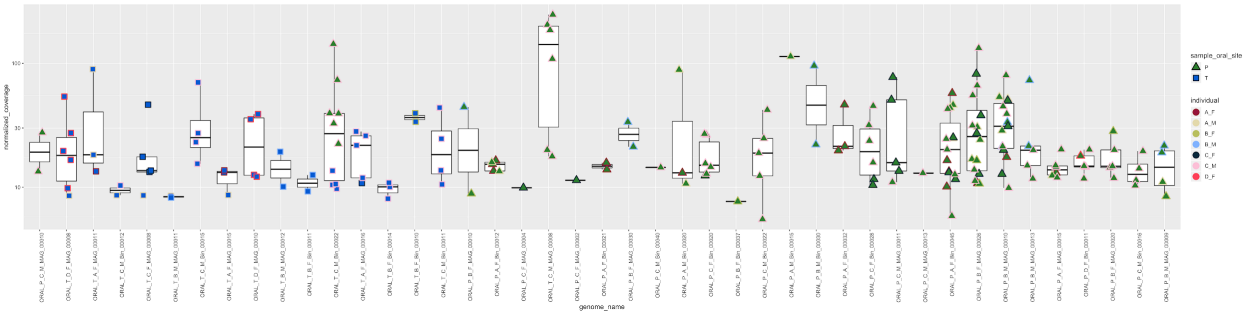

Figure 10: Normalized relative abundances of each of our 43 TM7 MAGs in the 71 metagenomes. The shape and fill color of each dot is according to the sample type (tongue/plaque), while the stroke color is according to the participant ID from which the sample was taken.

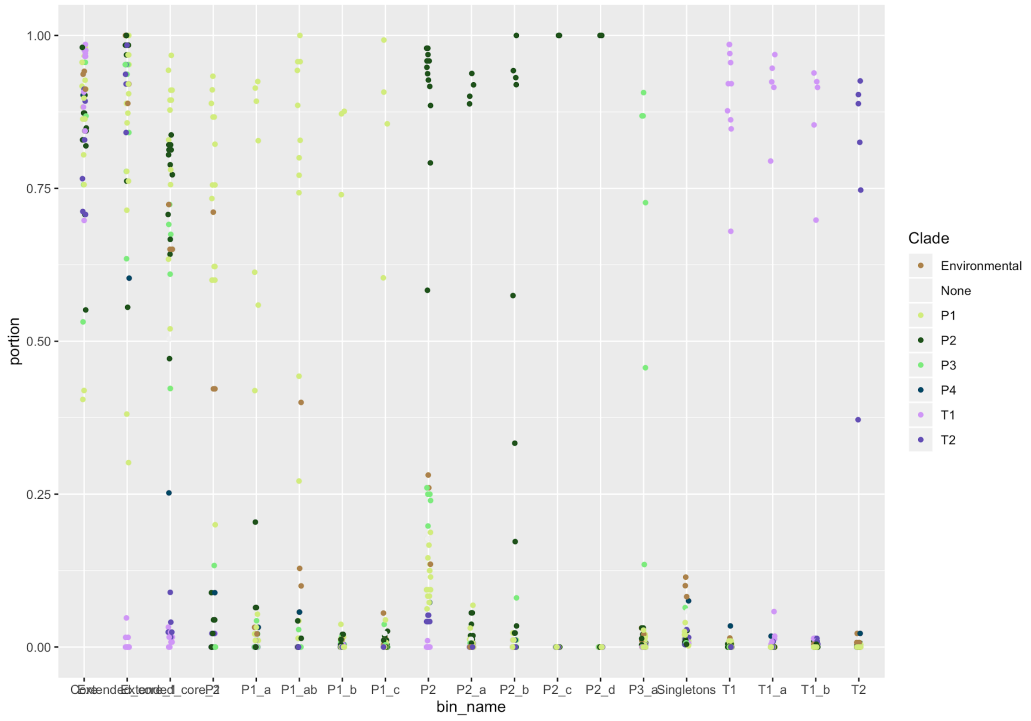

Figure 11: GC clusters represent clade-specific GCs. Data points represent the portion of the GCs of a GC bin that occur in each genome and colored according to clade designations of genomes.

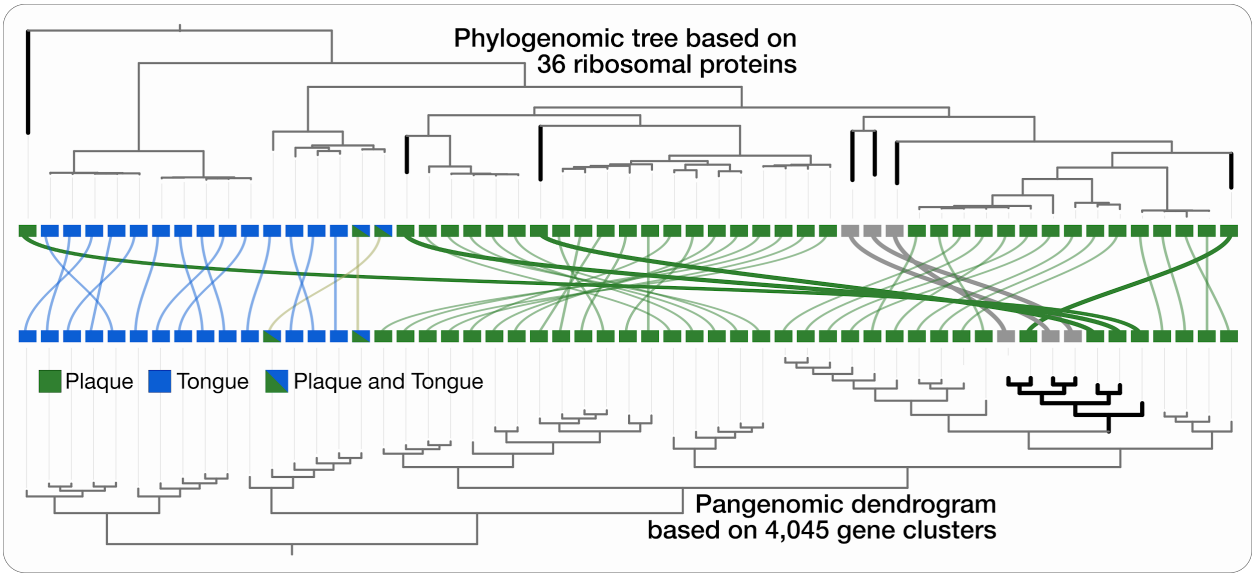

Figure 12: Organization of TM7 genomes according to the occurrence of gene-clusters clusters oral genomes according to oral site affiliation. The dendrogram at the top represents the phylogenetic organization based on ribosomal proteins, while the dendrogram on the bottom represents the hierarchical organization of genomes based on the gene-cluster frequency of occurrence across genomes using euclidean distance and ward ordination. The information at the center of the figure shows the site affiliation of each oral TM7 in accordance with Figure 5. Branches that appear in bold black color represent environmental and plaque-associated genomes that are phylogenetically-distinct, but that are grouped together based on their gene content, and nested together with plaque-associated genomes.

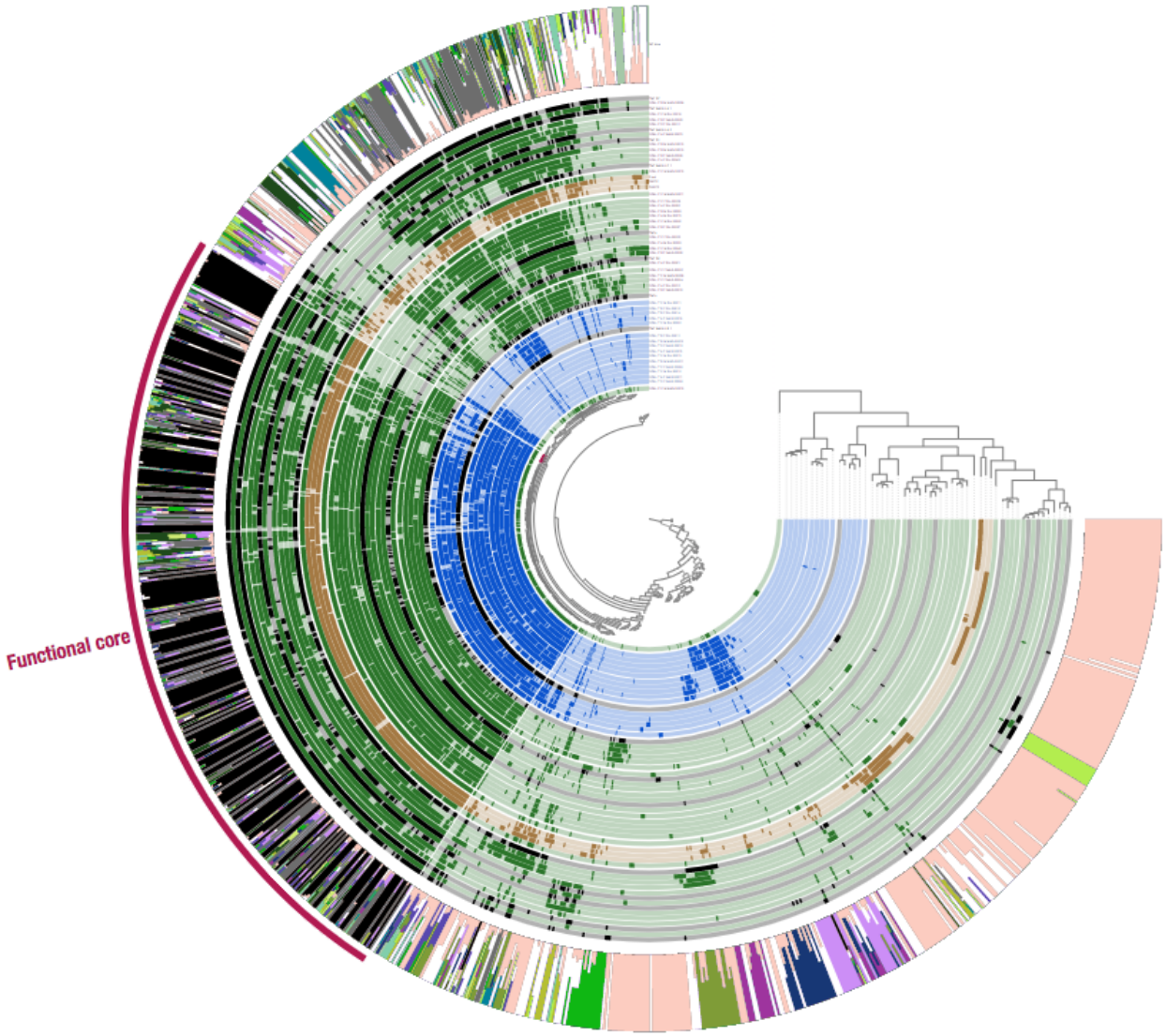

Figure 13: Functional core includes mostly core GCs, but also many clade specific GCs. Each of the 970 functions are organized in the tree in the center of the figure according to their occurrence in the 55 genomes (using Euclidean distance and Ward's method) . The first 55 layers correspond to the TM7 genomes, where layers corresponding to tongue MAGs are blue, plaque MAGs are green, and previously published genomes are black. Bars in these 55 layers represent the presence of a function in the genome. The layers are ordered using the phylogenetic tree from Figure 3b. The next layer includes a stacked bar representing the portion of GC bin affiliation of each gene associated with a function. The red arc in the outermost layer marks the functions that were defined as part of the core for this TM7 pangenome. Notice that while the majority of the core functions are associated with core GCs, there are many that are associated with clade-specific GCs.

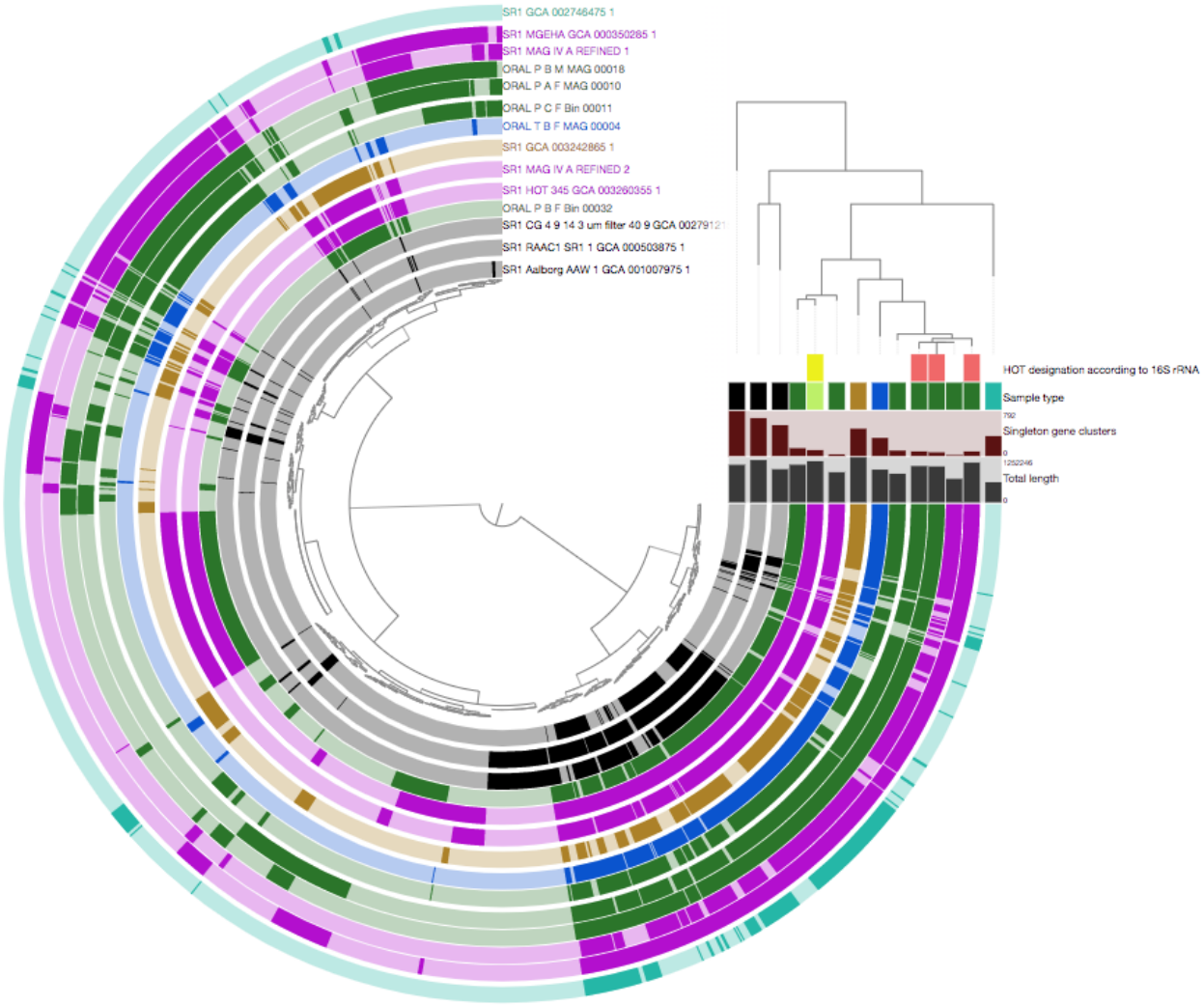

Figure 14: pangenomic analysis of SR1 genomes. The dendrogram at the center of the figure organizes gene-clusters according to their occurrence across the 14 SR1 genomes. The circular layers correspond to the 14 SR1 genomes and are ordered according to their phylogenetic organization. In these circular layers, colored sections mark the presence of gene-clusters in the corresponding genome. On the top right, the phylogenetic tree is shown and below it, the four horizontal layers correspond to (top to bottom) 1) Human Oral Taxon designation according to 16S rRNA sequences 2) Sample type (environmental: black, plaque: dark green, saliva: light green, canine supragingival plaque: brown, tongue: blue, dolphin gingival sulcus: cyan) 3) Number of singleton gene-clusters 4) Total length of the genome.

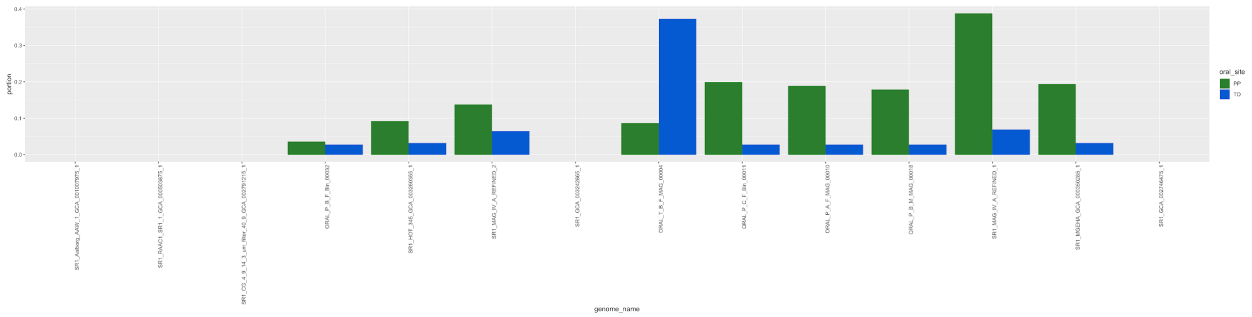

Figure 15: Detection of SR1 populations in the HMP plaque and tongue samples reveals prevalent populations and niche specificity. Barplots showing the portion of plaque (green) and tongue (blue) HMP samples in which each SR1 was detected, using a detection threshold of 0.5.

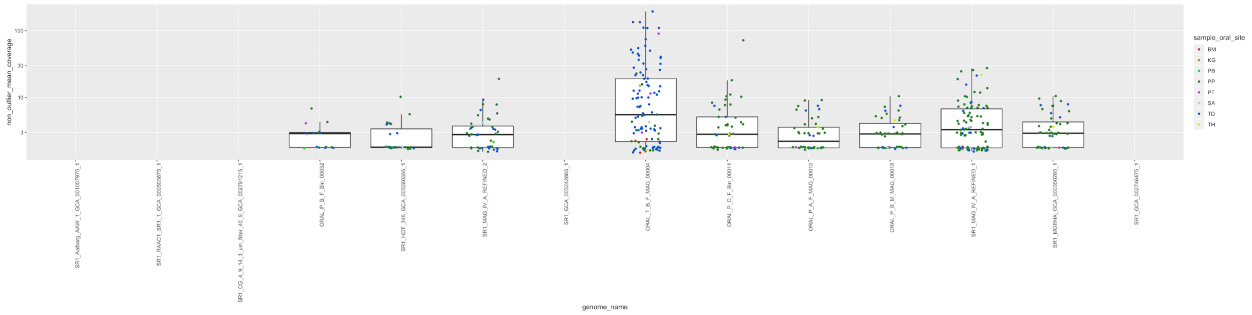

Figure 16: Normalized coverage of SR1 populations in HMP oral samples according to sample type. Boxplots showing the normalized coverages of each SR1 in plaque (green) and tongue (blue) HMP. For each genome, data is only shown for samples in which it was detected, according to the same criteria of detection used in Figure 15.

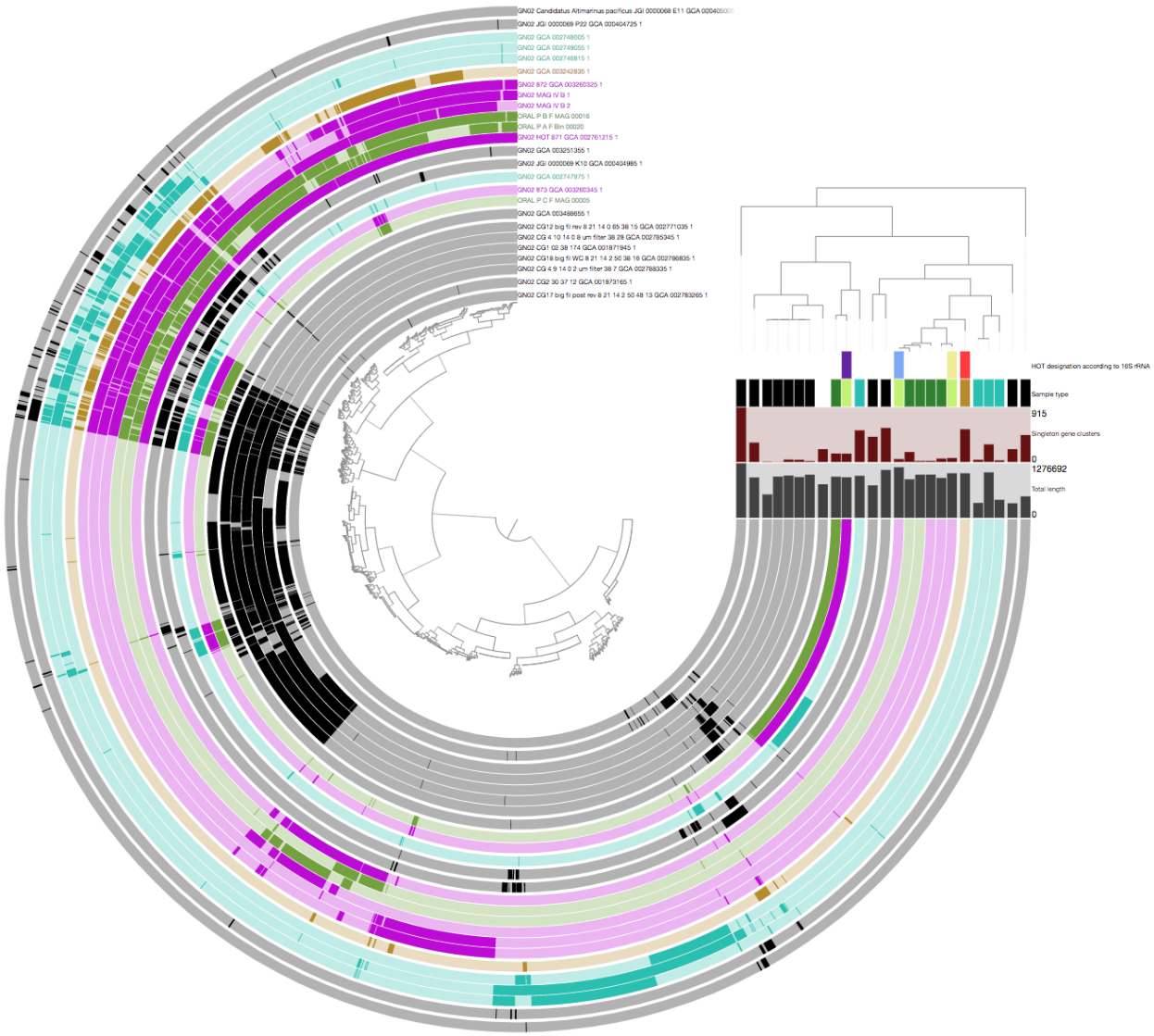

Figure 17: pangenomic analysis of GN02 genomes. The dendrogram at the center of the figure organizes gene-clusters according to their occurrence across the 25 SR1 genomes. The circular layers correspond to the 25 SR1 genomes and are ordered according to their phylogenetic organization. In these circular layers, colored sections mark the presence of gene-clusters in the corresponding genome. On the top right, the phylogenetic tree is shown and below it, the four horizontal layers correspond to (top to bottom) 1) Human Oral Taxon designation according to 16S rRNA sequences 2) Sample type (environmental: black, plaque: dark green, saliva: light green, canine supragingival plaque: brown, tongue: blue, dolphin gingival sulcus: cyan) 3) Number of singleton gene-clusters 4) Total length of the genome.

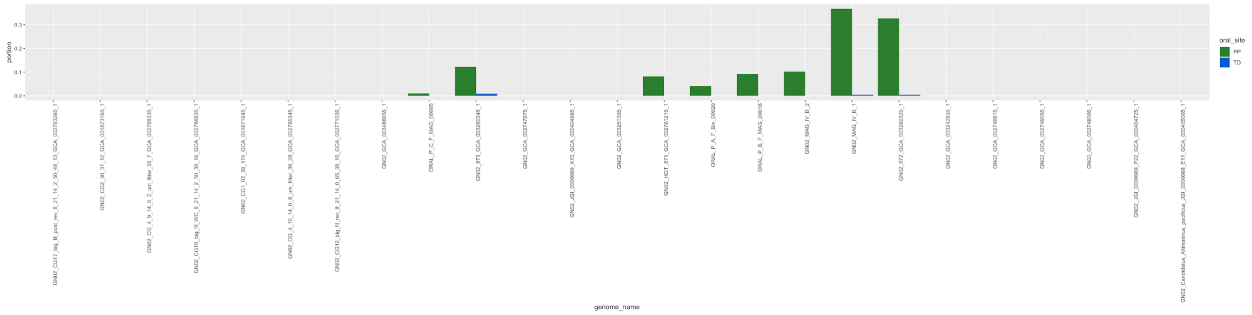

Figure 18: Detection of GN02 populations in the HMP plaque and tongue samples reveals the plaque specificity of oral members of this candidate phylum. Barplots showing the portion of plaque (green) and tongue (blue) HMP samples in which each GN02 was detected, using a detection threshold of 0.5.

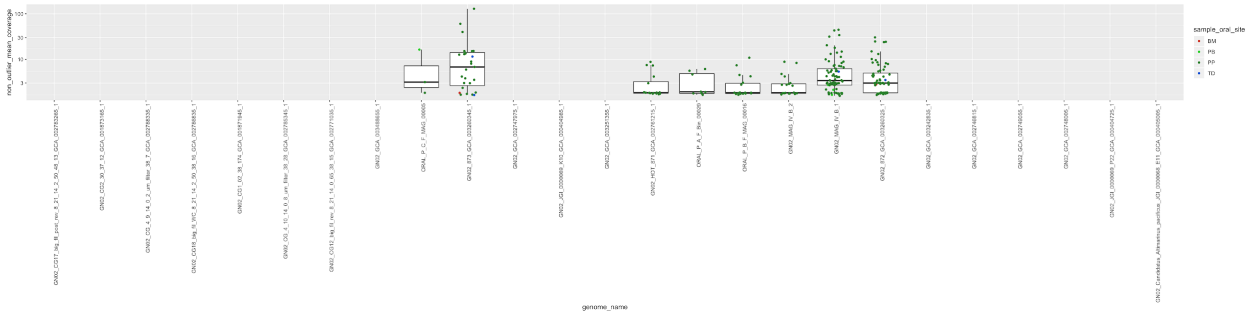

Figure 19: Normalized coverage of GN02 populations in HMP oral samples according to sample type. Boxplots showing the normalized coverages of each GN02 in plaque (green) and tongue (blue) HMP. For each genome, data is only shown for samples in which it was detected, according to the same criteria of detection used in Figure 18.

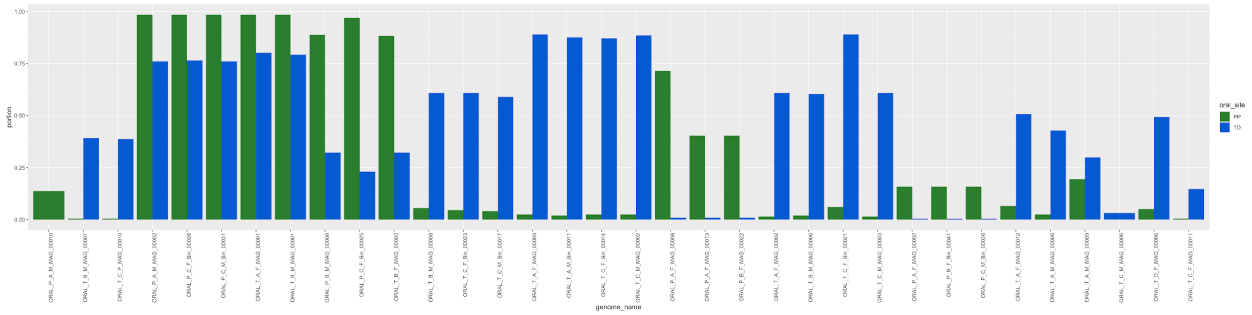

Figure 20: Presence of the novel populations in HMP tongue and plauqe samples. Barplots of the portion of plaque (green) and tongue (blue) samples in which each of the novel genomes occur. The presence of a population in a sample was determined according to a threshold of 0.5 detection value.

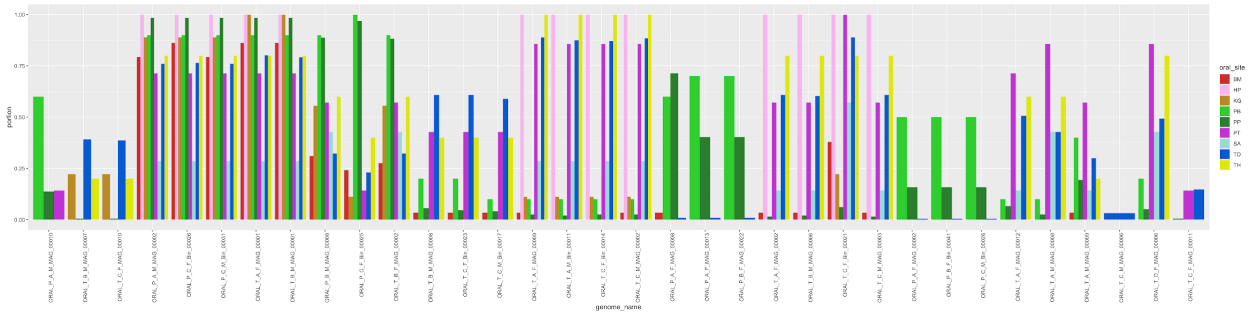

Figure 21: Presence of the novel populations in HMP oral samples by sample type. Barplots of the portion of samples in which each of the novel genomes occur, plotted by sample type for all 9 HMP sample types in which at least one novel population was detected. The presence of a population in a sample was determined according to a threshold of 0.5 detection value.

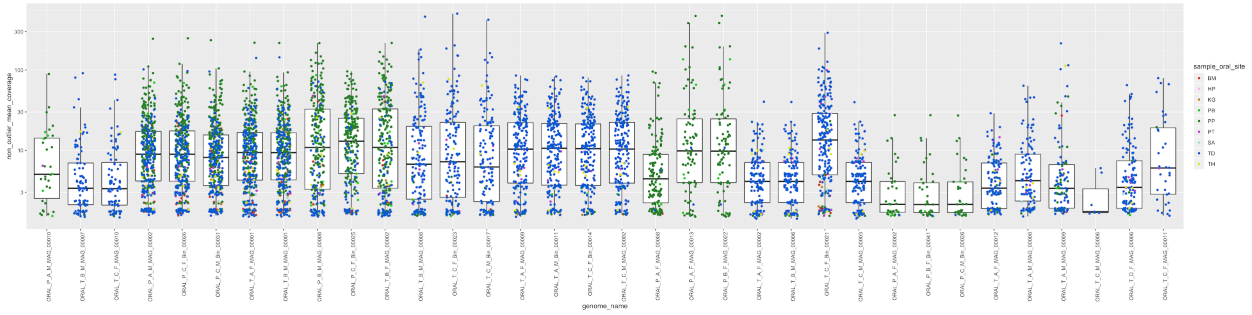

Figure 22: Normalized coverage of the novel populations in HMP oral samples according to sample type. Boxplots of the normalized coverage of the novel population. Color of data-points are according to the sample type. For each genome, data points are only shown for samples in which the genome was detected, according to the same detection threshold used in Figure 21.

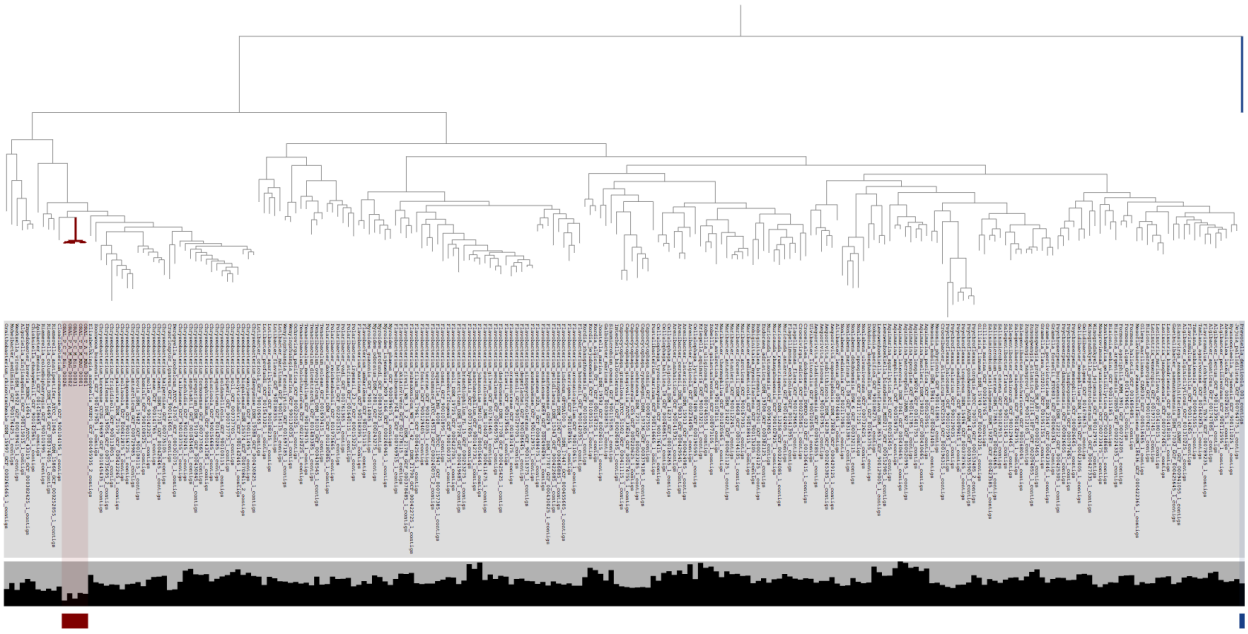

Figure 23: Phylogenomic analysis of Flavobacteriaceae genomes indicates oral MAGs represent an unnamed species in an unnamed genus within Flavobacteriaceae. Below the dendrogram, layers include the name and length of each genome. The 5 novel Flavobacteriaceae MAGs are indicated with red color and the Prevotella genome that was used to root the tree is indicated with blue color.

2.7 Supplementary information

2.7.1 Comparison of taxonomic composition using three methods

In order to investigate how the recovery of MAGs spans taxonomic units, we compared the estimation of taxonomic composition (at the genus level) of samples based on our MAGs with two other methods, KrakenUniq (F. P. Breitwieser, Baker, and Salzberg 2018), which utilizes short-reads, and hence circumvents potential challenges due to assembly and binning, and Minimum Entropy Decomposition (Eren, Morrison, et al. 2015) combined with GAST-based (Huse et al. 2008) taxonomic assignment of 16S rRNA amplicon sequence variants. While KrakenUniq lists 441 genera with above zero abundance in at least one sample (Supplementary table 4f at doi:10.6084/m9.figshare.11634321), GAST identified 40 (Supplementary table 5e at doi:10.6084/m9.figshare.11634321) and our genomes represented 37 distinct genera (Supplementary table 2f at doi:10.6084/m9.figshare.11634321). We included the 15 most abundant genera according to each method, which amounted to a list of 19 genera, and to which we added TM7, in a comparison of relative abundance estimations by the three methods. Overall, the three methods presented similar trends for most of these 20 taxa, but also revealed further discrepancies (Figure 24,

Figure 25, Figure 26). While 16S rRNA amplicons allow the taxonomic assignment of each sequenced amplicon (to various levels of resolution), it suffers from primer biases for specific taxa (Eloe-Fadrosh et al. 2016). While the study of metagenomes does not suffer from these primer biases, the ability to assign taxonomy to every sequenced read is limited by the reference database, leaving many reads either unidentified, or worse, wrongly classified (Escobar-Zepeda et al. 2018). While MAGs allow a confident taxonomic assignment (to known taxa), normalizing coverages to estimate relative abundance is challenging, especially when it is required to account for many unassigned reads. In addition, the occurrence of populations that undergo genomic reorganizations, and the occurrence of populations with large within-population variability, limits the ability to assemble short reads into large contigs and hence our ability to generate high quality MAGs. In conclusion, we could examine trends of particular taxons as these are revealed by a particular method, but none of these methods is likely to inform us of actual relative abundances. With these limitations in mind, our data shows that while the abundance profiles at the genus level are similar for the majority of the abundant genera, there are specific taxa for which there are major differences, such as Actinomyces, Rothia, and Fusobacterium (Figure 24, Figure 25, Figure 26).

To process the amplicon sequencing data mentioned above, we used the Oligotyping (Eren, Murat Eren, et al. 2013) command o-pad-with-gaps to pad sequences with gaps and eliminate length variation. We used Minimum Entropy Decomposition (MED) (Eren, Morrison, et al. 2015) to identify amplicon sequence variants (ASVs) across samples and determine microbial community structure, and we used Global Alignment for Sequence Taxonomy (GAST) (Huse et al. 2008) to assign taxonomic affiliation to each ASV.

We selected the genera used for the comparison of the relative abundance estimation between the three methods (MAGs, KrakenUniq, and 16SrRNA) by identifying the 15 most abundant genera according to each method and then merging these to a list of a total of 20 genera: Actinomyces, Aggregatibacter, Campylobacter, Capnocytophaga, Corynebacterium, Derxia, Fusobacterium, Gemella, Genus, Granulicatella, Haemophilus, Leptotrichia, Neisseria, Porphyromonas, Prevotella, Pseudomonas, Rothia, Streptococcus, Streptomyces, TM7, Veillonella. We considered TM7 as a “genus” for the sake of this analysis, despite the fact that it includes multiple genera. Of these “top genera”, Derxia was completely absent from both KrakenUniq and MAGs, and Gemella and Granulicatella were completely absent from

KrakenUniq. On the other hand, Pseudomonas, and Streptomyces appear in the top 15 abundant genera of the KrakenUniq results but were completely absent from the MAGs and 16S rRNA ASVs. Lastly, TM7 was completely absent from the 16S rRNA ASVs, despite being amongst the top abundant genera according to MAGs. We used ggplot2 (Wickham 2016) to generate relative abundance plots per sample per method. The tables used to generate relative abundance plots based on MAGs, KrakenUniq and 16S rRNA are available in Supplementary tables 2f, 4f, and 5e at doi:10.6084/m9.figshare.11634321, respectively. Tables with relative abundance for various taxonomic levels for MAGs, KrakenUniq and 16S rRNA are available in Supplementary tables 2e, 4a-e, and 5a-d at doi:10.6084/m9.figshare.11634321, respectively.

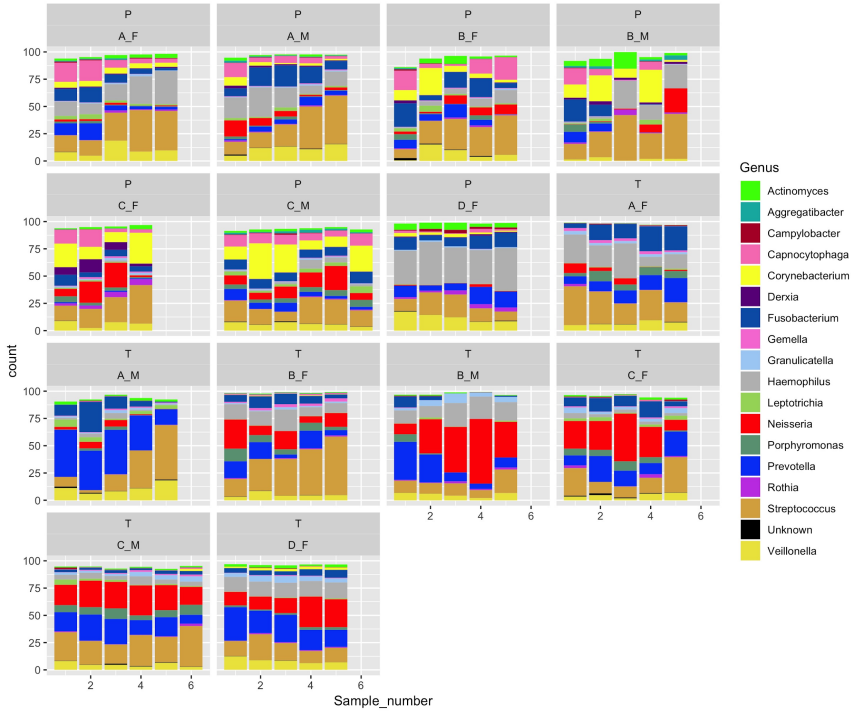

Figure 24: Taxonomic profiles using 16S rRNA gene amplicon sequence variants (ASVs) produced by MED with taxonomic assignment from GAST.

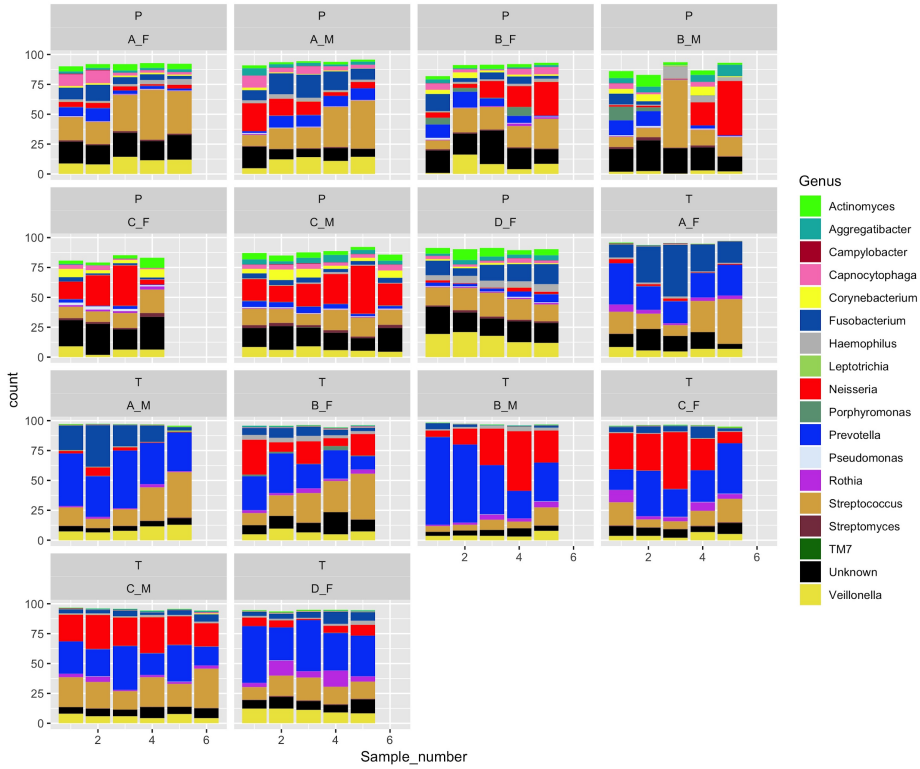

Figure 25: Taxonomic profiles based on metagenomic short reads using KrakenUniq.

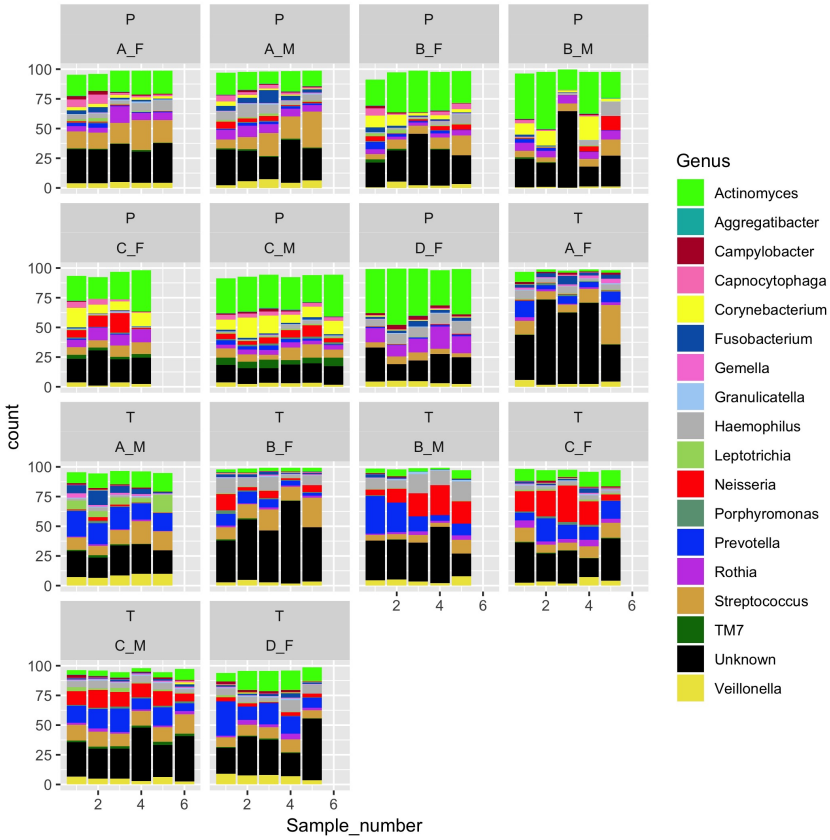

Figure 26: Taxonomic profiles based on coverages of MAGs.

2.7.2 Phylogenomic analysis of MAGs and HOMD genomes

P_C_M_Bin_00033 presents such an example of a deeply branching genome. In fact this genome is placed in phylogeny as a deep branch within Tannerella (of phylum bacteroidetes), but CheckM assigned this genome to the genus Granulicatella of phylum Firmicutes. This is likely due to a composition of at least two genomes that contribute SCGs to this genome. We also identified such issues with a certain genome from HOMD, “Capnocytophaga_sp__003”, which has an atypical genome length greater than 6Mbp, and indeed seems composite as it forms an unusually deep branch within Capnocytophaga, and in fact CheckM failed to assign any phylum affiliation to this genome.

2.7.3 Aveage Nucleotide Identity (ANI) of oral TM7

Each of the monophyletic clades that we identified include diverse sub-clades as evident by multiple sub clusters within each clade (Figure 4), hence we sought to search for genomic identity boundaries that could

allow the definition of distinct species within these clades. To examine whether phylogenetic clusters within the clades we identified correspond to species of TM7, we computed the average nucleotide identity (ANI) between each pair of genomes. Multiple studies have suggested a 95% cutoff using ANI to determine bacterial species (Jain et al. 2018; Konstantinidis and Tiedje 2005). Our analysis revealed 12 sub-clades that included at least 2 genomes each and separated according to a within-group alignment coverage of >25% and identity >90% (Figure 4, Supplementary tables 7f, 7g, 7h, and 7i at doi:10.6084/m9.figshare.11634321). We hypothesize that each of these represent a separate species, despite the slightly lower than the aforementioned 95% identity cutoff. Genomes of sub-clades T2_a and T2_b aligned between each other with alignment coverage of 50%-70% and identity of 85%-88%, suggesting that these two represent two species of the same genus (Figure 4, Supplementary table 7h at doi:10.6084/m9.figshare.11634321). There were only two other cases in which outgroup members had alignment coverage above 25%. P_C_M_Bin_00016 had 30% alignment coverage and 83% identity to P_B_M_MAG_00013 (P1_a), suggesting that it could belong to the same genus as the genomes of sub-clade P1_a. Similarly, P_C_M_Bin_00022 appears to be a single representative amongst our genomes of a species that belongs to the same genus as P2_b, as it aligned with ~50% coverage and ~85% identity with all four members of P2_b (including TM7x). Since we found no other significant alignment between members of distinct sub-clades, these TM7 genomes potentially represent at least 11 distinct genera.

2.7.4 Occurrence of TM7 across additional oral sample types, other than supragingival plaque and tongue dorsum, and including samples from patients with periodontitis

In order to examine the occurrence of the TM7 populations across the oral cavity, we used 68 HMP samples with a total of 7 additional sample types (Supplementary table 7j at doi:10.6084/m9.figshare.11634321), as well as 24 subgingival samples from 9 patients with periodontitis. The number of reads per sample was comparable across sample types with the exception of saliva samples, which had a lower number of reads per sample by an order of magnitude as compared to other sample types (Figure 27). TM7 populations were detected in all sample types except for the single hard palate sample (Figure 28, Supplementary table 4o at doi:10.6084/m9.figshare.11634321). While presence of populations in the subgingival plaque mostly matched with their presence in supragingival plaque, some populations were found in a larger portion of

the 10 subgingival plaque samples as compared to supragingival plaque (Figure 28). Moreover, we found that occurrence in subgingival plaque did not imply occurrence in supragingival plaque. For example, from the 5 individuals for which P_C_M_Bin_00016 (clade P1) was detected in the subgingival plaque, we only detected this population in the supragingival plaque of one individual. P_C_M_MAG_00010 (sub-clade P4_a) also appeared to be enriched in subgingival plaque vs. supragingival plaque. This genome belongs to group 'G5', which has been previously suggested to be enriched in patients with periodontitis based on studies of 16S rRNA amplicons (Abusleme et al. 2013). Our analysis of subgingival samples from patients with periodontitis revealed a similar occurrence as compared to the 10 subgingival plaque samples of the 8 healthy HMP individuals (Figure 30, Figure 31, Supplementary table 7p-s at doi:10.6084/m9.figshare.11634321). In Palatine tonsils and throat samples we detected only tongue-associated TM7, while in Keratinized gingiva samples only members of clade T2, and sub-clade P1_c were detected. T_C_M_Bin_00011 (sub-clade T2_c) appeared more prevalent and abundant in keratinized gingiva samples than in tongue samples, and T_B_F_Bin_00010 (clade T2) was more abundant in buccal mucosa samples than in tongue samples (Figure 29, Supplementary table 4o at doi:10.6084/m9.figshare.11634321). Due to the low number of HMP samples per sample type (other than tongue dorsum and supragingival plaque) further investigation would be required in order to confidently determine whether such associations exist.

The pair-end reads of the 24 subgingival plaque samples from patients with periodontitis from the study by Cliff et al. (Califf et al. 2017) were received directly from the authors, since the samples that were deposited on MG-RAST with the original Califf et al. publication included only one of the pairs of reads. Raw sequences were analyzed and the occurrence of TM7 MAGs in these samples were assessed as described in the Methods section.

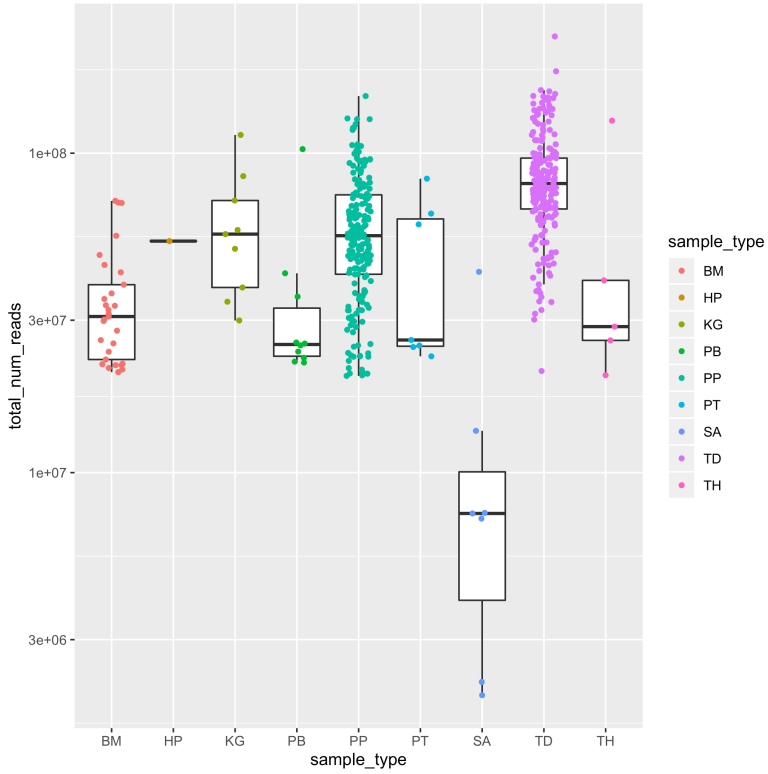

Figure 27: Number of reads per metagenome. Each data point represents the number of reads in a single sample for the 9 sample types.

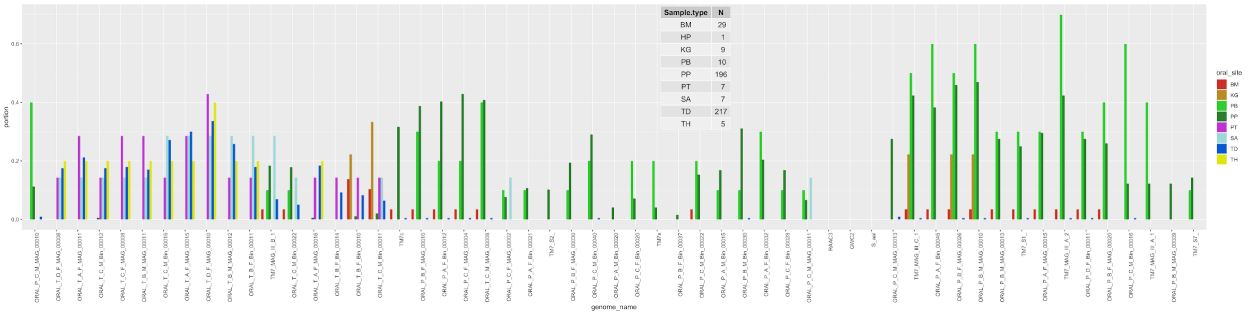

Figure 28: Occurrence of TM7 across oral sample types. For each of the 55 genomes (on the x-axis) the colored bars represent the portion of samples per sample type, in which it is detected (detection > 0.5).

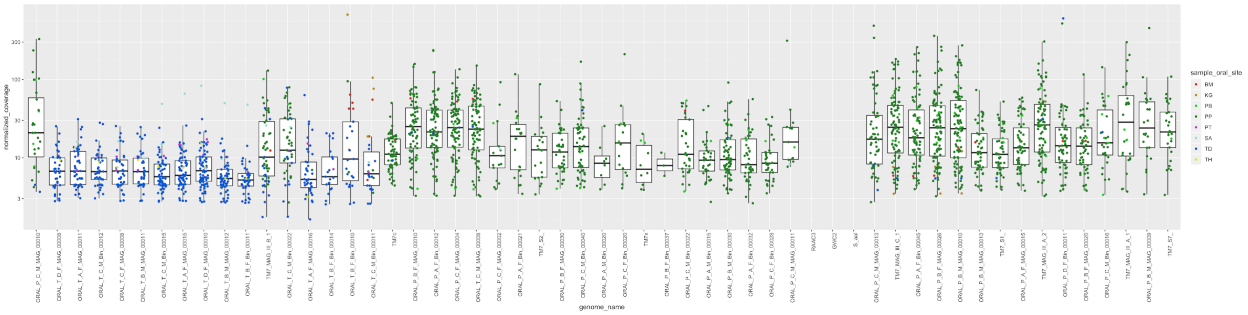

Figure 29: Coverage of TM7 across oral sample types. Boxplots of the normalized coverages of each TM7 across samples. Data points are colored according to sample type.

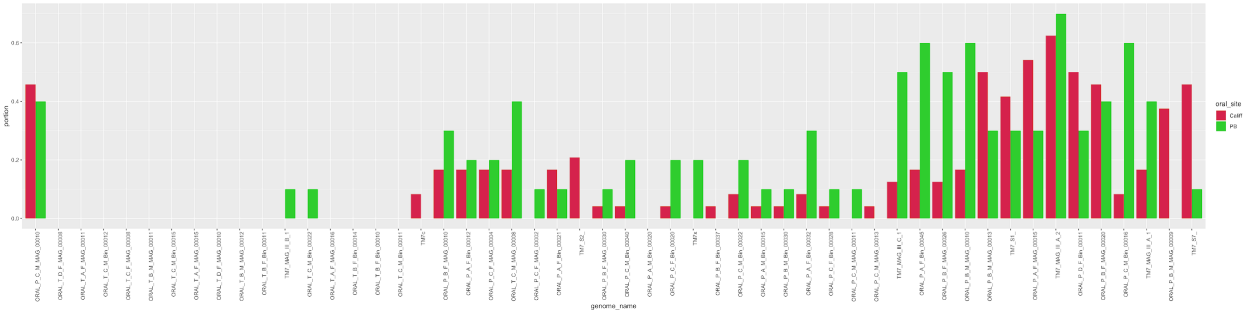

Figure 30: Occurrence of TM7 in subgingival plaque samples of healthy individuals and individuals with periodontitis is mostly matching. Bars indicate the portions of subgingival plaque samples from healthy individuals (green) and individuals with periodontitis in which each of the 55 TM7 are detected.

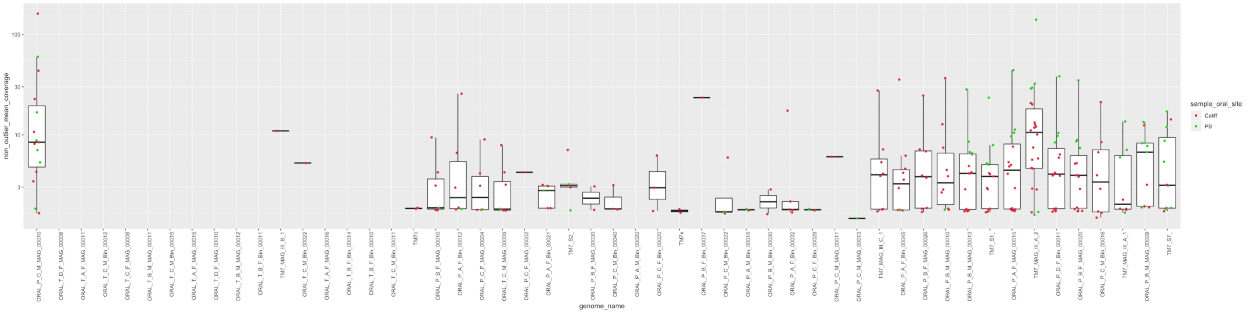

Figure 31: Coverage of TM7s in subgingival plaque. Boxplots of the normalized mean coverage of TM7 in samples of healthy individuals (green) and individuals with periodontitis (red).

2.7.5 Mobile elements and prophages in TM7 genomes

In order to systematically search TM7 genomes for evidence of prophages we used VirSorter (Roux et al. 2015) and the “inovirus detector” (Roux et al. 2019) to automatically detect contigs that potentially include prophages in the TM7 genomes and detected 47 contigs with potential prophages (Supplementary table 8g at doi:10.6084/m9.figshare.11634321). We extended this list to a total of 58 contigs by manually

identifying additional contigs using functional annotations as markers for phages, and by searching for contigs with GCs that associate with the contigs detected by VirSorter/"inovirus detector" (Supplementary table 8g at doi:10.6084/m9.figshare.11634321). We manually examined these contigs, and identified 36 contigs that include partial or complete prophages, which we manually curated to determine the likely start and end nucleotide positions of the prophages (Supplementary table 8g at doi:10.6084/m9.figshare.11634321). In order to search for conserved sequences amongst these phages, we employed a pangenomic approach. Our pangenomic analysis revealed contigs that likely represent different fragments of the same prophage (Figure 33), we merged these contigs, and removed 9 contigs that were mostly composed of singleton gene-clusters to generate a second pangenomic analysis with a refined collection of 25 prophages (Figure 32). Clustering this refined collection of prophages according to the occurrence of gene-clusters revealed 9 "phage groups" of closely related prophages present in two or more TM7 genomes (Figure 32).

Functional annotation is lacking for most virus genes, and the sequence diversity amongst the viral proteins is high, as is demonstrated in the lack of shared GCs across phages in Figure 32. Hence, it is challenging to find suitable targets for phylogenetic analysis of phages. In an effort to study the phylogenetic relationships of the phages we used two hallmark genes of (pro)phages: (1) integrase, and (2) terminase to compute phylogenies. We performed a phylogenetic analysis using the 13 integrases we identified in our collection of prophages (Figure 33). Our results reveal cases in which phages that associate with highly divergent hosts rely on similar integrases, while phages that otherwise appear to be closely related (i.e. belong to the same "phage group") often rely on divergent integrases (Figure 33). The phylogenetic tree we computed using the 10 tail terminase large subunit identified in the prophages showed a better overall concordance with the organization according to GCs (Figure 32, Figure 35). Genomes of phage groups "pg02", "pg07", and "pg08" had high within-group identity of the terminase large subunit, but "pg01", which also shows large variability in the pagenomic analysis (Figure 32) included prophages with divergent terminase large subunit, despite the fact that their hosts belonged to the same species (P1_a). While it appears that distantly related phages, infecting distantly related hosts, can use very similar integrases (Figure 35), our data does not include an case in which distantly related phages harbor similar terminases

(Figure 36). To examine the novelty of these prophages we searched for similar nucleotide sequences using Blast against the NCBI's nr nucleotide collection, but this search had no results, emphasizing the novelty of these sequences.

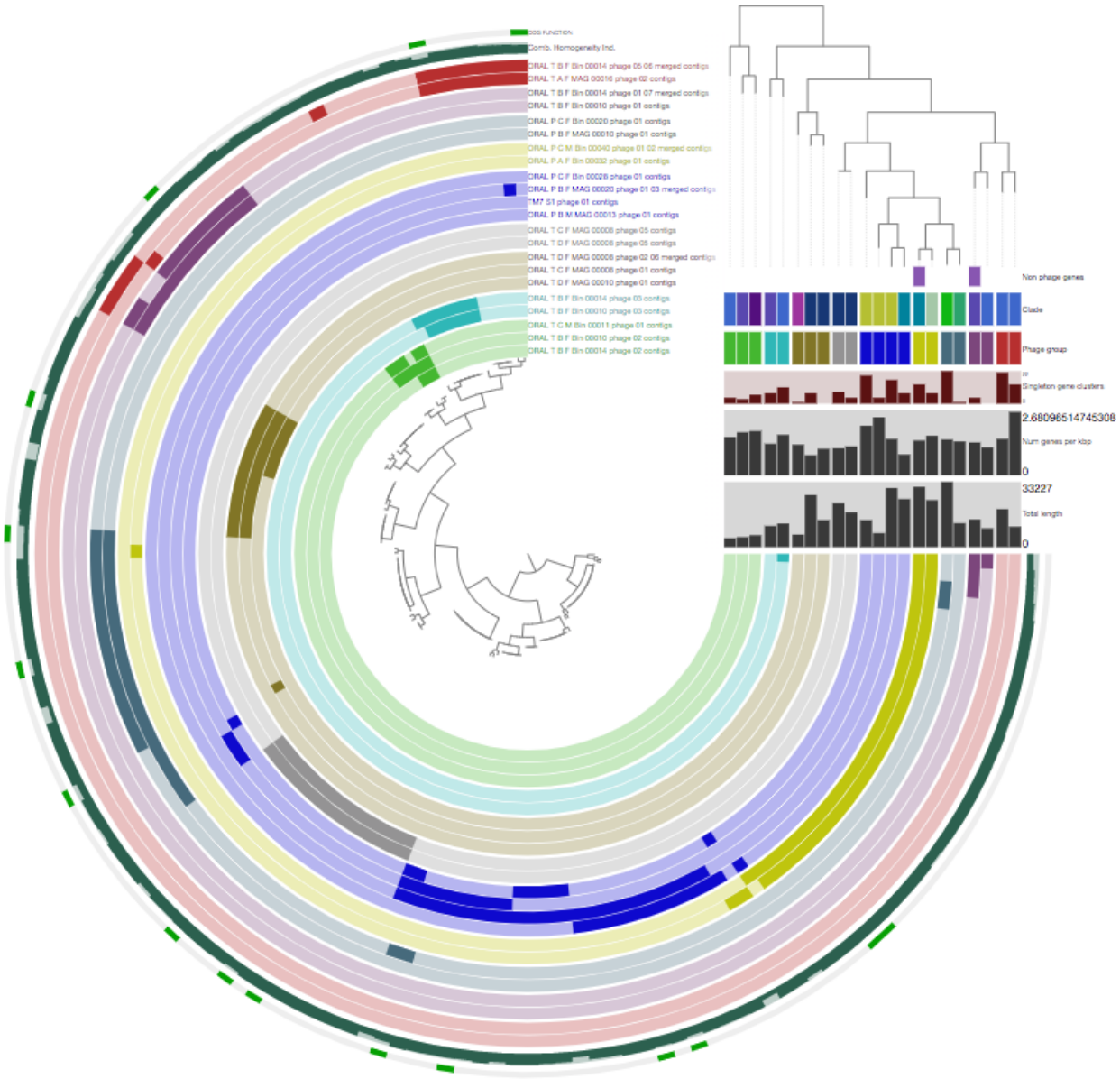

Figure 32: Pangenomic analysis of TM7 prophages reveals 9 "phage groups" of closely related phages. The dendrogram at the center of the figure represents the hierarchical clustering, using euclidean distance and Ward's method, based on the frequency of occurrence of 143 GCs, each containing at least two homologous genes from at least two prophage sequences. The 22 inner circular layers represent prophage sequences, where each data point marks the presence or absence of a protein that belongs to the corresponding GC. Colors of these 22 layers are according to their "phage group" affiliation. The two outermost circular layers represent the combined homogeneity index for each GC, and the GCs that were

Figure 32 (continued): annotated with a COG function (green). A low homogeneity index signifies higher sequence diversity amongst the proteins that comprise a GC. The dendrogram at the top right represents the hierarchical clustering of the prophage sequences according to the GC frequency of occurrence using Euclidean distance and Ward's method. The first horizontal layer below the dendrogram marks the two prophages that include a TM7 core protein. The next two layers show the clade affiliation of the TM7 genomes, and the "phage group" affiliation. The lowest three horizontal layers show the number of singletons, number of genes per kbp, and the total length for each prophage sequence.

The recovery of multiple closely related phages from TM7 genomes, as well as the presence of host (TM7) genes on the same contigs that contain the phage genes provide strong evidence for the association of these phages with the TM7 genomes. To further enforce this association, we used CRISPRCasFinder (Couvin et al. 2018) to search the TM7 genomes for CRISPR spacers and survey existing spacers for ones that match our collection of prophages. CRISPRCasFinder identified 66 CRISPR arrays, of which 14 had evidence level 3 or 4 as defined by Couvin et al. (Couvin et al. 2018) (Supplementary table 8l at doi:10.6084/m9.figshare.11634321), and originated from 12 genomes spanning clades P1, P2, P3, P4, and T2, but not T1 nor any of the environmental genomes. We blasted the set of 14 CRISPR arrays against the TM7 genomes and found a total of 9 spacers with blast hits that were not self-hits (i.e. not a blast match of the spacer to itself), which included 7 spacers with a single external match (i.e. a match outside of the genome where the spacer was found), 1 spacer with two external matches, and 1 spacer with 2 external matches and one internal match, showing that this spacer was self targeting (Supplementary table 8m at doi:10.6084/m9.figshare.11634321). 5 of these 9 spacers had hits to pg01 prophages, and revealed that this family of prophages targets a wide variety of TM7 species within the 'G1' oral caldes P1, P2, and P3 (Supplementary table 8m at doi:10.6084/m9.figshare.11634321). Another spacer matched a pg06 prophage. While we found pg06 prophages in genomes of sub-clades P2_a and P2_c, this spacer was found in a P3_a genome. An additional spacer from a P3_a genome matched a prophage from a P1_a genome suggesting the existence of multiple phage groups that target a variety of 'G1' oral genomes. Two additional spacers had hits across G1 genomes, but these matched sequences that we did not identify as prophages and were composed of singleton GCs with no functional annotation, deeming it hard to determine whether these are prophages or other mobile genetic elements. As mentioned above, we found a spacer from P_A_F_Bin_00032 to be self-targeting. Despite being potentially detrimental and confer autoimmunity, self-targeting spacers are fairly common (Stern et al. 2010). In this case, the spacer matched 3 of the 4 genes in our dataset that comprise GC_00002421 in P2_a genomes. This GC had no COG

function, but was recognized to have a 'PEGA domain' by Pfam, which is found in surface layer proteins. While this GC was unique to members of P2_a, it seems that this protein is conserved and represents a core function in the TM7 pangenome, since a protein with this annotation was found in nearly all genomes, and almost always flanked by a "Sortase (surface protein transpeptidase)". The apparent viability of the P_A_F_Bin_00032 population as evident by the recovery of the genome, despite the CRISPR self-targeting of a core function might suggest that this core function is not strictly required for the survival of TM7 in the oral cavity.

In contrast to the oral clades P1, P2, P3, P4, and T2, we found no evidence for CRISPR-cas systems in T1 genomes nor in the three environmental genomes. The CRISPRCasFinder output included contigs from T1 genomes, but these only had evidence level 1 or 2, suggesting that they could be spurious identifications (Supplementary table 8l at doi:10.6084/m9.figshare.11634321). Indeed, many of these appeared to fall within genes that belong to a single GC, suggesting that something about the sequence of these specific genes confuses the CRISPRCasFinder algorithm. There was only one contig from one of the three environmental genomes (GWC2) that was included in the output of CRISPRCasFinder, but it had evidence level 1, and the identification fell within a TM7 core protein, and hence is likely an erroneous identification. In accordance with the lack of CRISPR arrays, we did not find any of the CRISPR associated proteins in the environmental genomes nor in genomes of clade T1, but we did find these proteins in genomes of the oral clades P1, P2, P3, P4, and T2. We find the lack of prophages and the lack of CRISPRs in environmental genomes to be highly interesting, since these fall within the G1 group to which the P1, P2, and P3 clades belong, which could imply that these CRISPR-cas systems are unique to oral-associated (or more generally to animal-associated) TM7, but an analysis of a wider variety of environmental TM7 would be required to test this hypothesis. To search for the potential source for CRISPR proteins in oral TM7, we blasted cas9 proteins from 6 genomes representing all 5 CRISPR-containing clades, and representing the three GCs annotated as cas9 proteins, against the NCBI's nr protein sequences. All 6 cas9 proteins were matching the same collection of proteins from oral TM7, but no environmental TM7. The top non- TM7 matches were of Firmicutes (Bacilli and Clostridia), suggesting that these proteins were once horizontally transfered from Firmicutes to oral-associated TM7. Future investigations could include a phylogenetic analysis of CRISPR

associated proteins of TM7 along with ones from other CPR and non-CPR (including human-associated) genomes to further shed light on the source of CRISPR systems in TM7 genomes, and whether these are unique to mammalian-associated TM7.

While T1 and environmental genomes lacked CRISPR-cas systems, they could alternatively rely on restriction modification systems to defend against phages. Based on COG annotations, we identified Type I and/or Type II restriction-modification systems in 34 TM7 genomes spanning all identified oral clades and two of the three environmental genomes, GWC2 and RAAC3. In addition to lacking CRISPR-cas systems, ,members of clade T1 were also lacking a protein annotated with the COG function “Phage shock protein PspC (stress-responsive transcriptional regulator)”, which was found in nearly all genomes from all other oral clades and in two of the three environmental genomes.

In addition to prophages, we identified other mobile genetic elements in many TM7 genomes. 33 genes coding for various transposases were detected in 18 genomes, covering all oral clades and the three environmental TM7. These genes comprised a total of 22 GCs, and up to four transposases per genome (Supplementary table 8n at doi:10.6084/m9.figshare.11634321). The transposases were predominantly associated with GCs unique to specific lineages. 19 of the 22 GCs were singletons (i.e. identified in a single genome), the three other GCs, GC_00003909, GC_00002371, and GC_00001084 were identified in two, three and seven genomes, respectively. GC_00001084 was annotated as an “ISXO2-like transposase domain” by Pfam and was identified in most P3_a and three P1_b genomes. GC_00002371 was identified in 3 (out of 5) T1_a genomes and was annotated with the COG function “Transposase InsO and inactivated derivatives”. While the transposases in T1_a genomes were highly conserved in protein sequences, they occurred in differing positions within the genomes (Supplementary table 8a at doi:10.6084/m9.figshare.11634321), suggesting recent mobility of these elements. GC_00003909 was detected in the two P1_c genomes with the COG function “Transposase and inactivated derivatives, IS30 family”. In both P1_c genomes, this transposase occurred in the same exact position within the genome, suggesting that this might represent an inactive transposon.

In order to examine the potential origin of the TM7 transposases, we searched for similar sequences in NCBI's non-redundant protein sequence database (Supplementary table 8o at doi:10.6084/m9.figshare.11634321). The vast majority matched best to transposases from other TM7 genomes or other CPR genomes, including many genomes recovered from environmental samples. For example, the single transposase from T_C_M_MAG_00008 had best matches to other oral TM7, but also matched many other CPR, including CPR MAGs recovered by Probst et al. from an aquifer (Probst et al. 2018). In contrast, T_C_M_Bin_00011 included what appears to be only the N-terminal region of an IS30-family transposase which matched best to transposases from a Streptococcus agalactiae genome (89% coverage and 52% identity in protein sequence). Examination of the contig on which this transposase was detected showed that it is not likely to be explained by a binnig error, as this transposase was flanked by many core proteins of TM7 on one side, but on the other side, it was flanked by three short proteins that belonged to singleton GCs (i.e. with no homologs in the TM7 pangenome) and no functional annotation (gene ids 21837-21839 in Supplementary table 8a at doi:10.6084/m9.figshare.11634321). A blast search of protein sequences matched these three proteins with a surprisingly high identity (94%-100%) to genes from other oral bacteria representing various phyla, including Firmicutes, Fusobacteria, and Proteobacteria. The presence of a partial transposase next to genetic elements that appear to be widely shared between oral microbes could reflect a mechanism for horizontal gene transfer between TM7 and non-CPR oral microbes, but requires further validation. In summary, these results suggest that the transposases carried by oral TM7 genomes are predominantly anciently associated with CPR genomes, but also include transposases that were likely transferred to oral TM7 from other mammalian-associated bacteria more recently, and could potentially be used to incorporate proteins that are widely shared by oral bacteria.

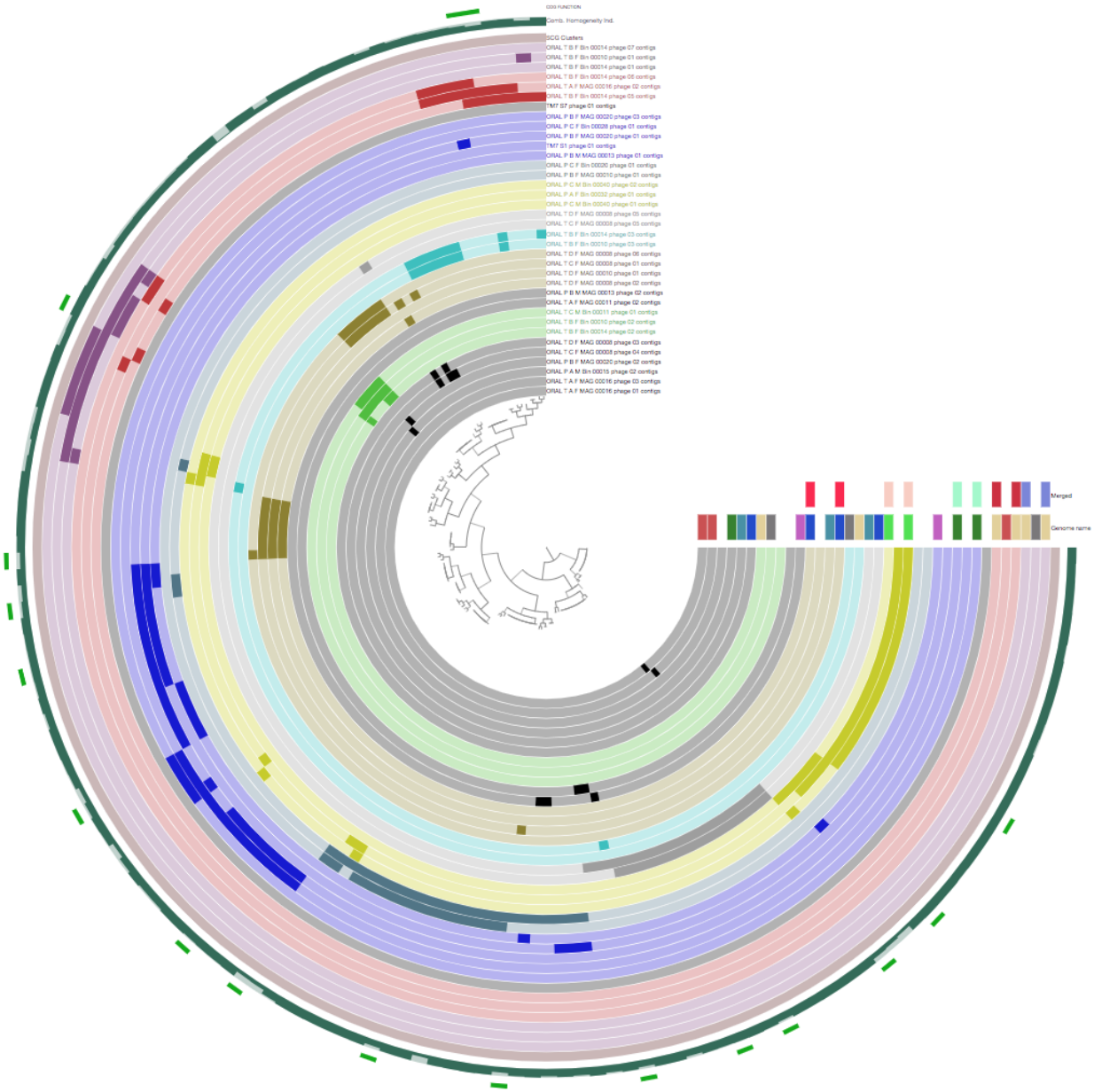

Figure 33: Pangenomic analysis of a potential prophages includes multiple contigs that likely represent fragments of the same prophage. The gene content of each prophage is represented by an individual layer, and the 9 main groups of TM7-associated prophages are highlighted in different colors across layers. Layers that are in black color are ones that consisted mostly of singletons and were hence excluded from subsequent analysis. On the top right of the figure, the color bars in the top horizontal layer highlight pairs of contigs that belong to the same genome and that we identified as fragments of the same prophage and merged for the subsequent pangenomic analysis (Figure 32). In next horizontal layer, each genome for which we identified multiple prophage contigs is associated with a unique color, so that contigs that are in the same genome can be identified.

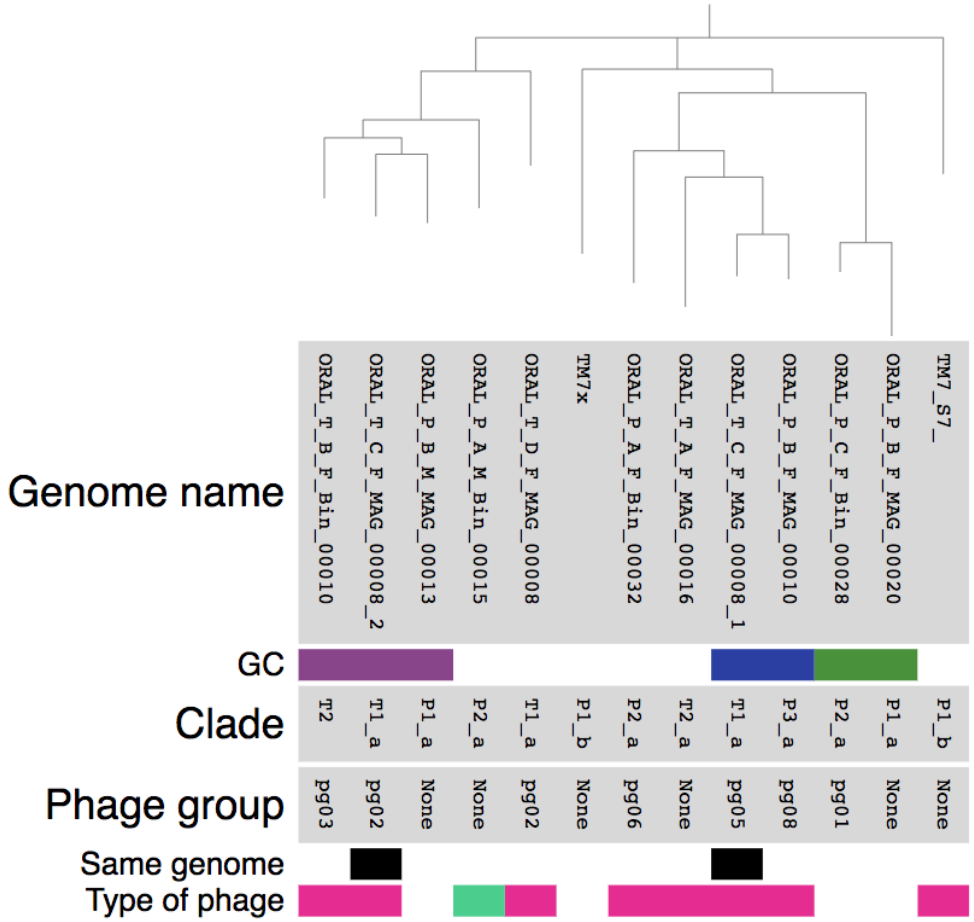

Figure 34: phylogeny of phages based on integrases. The dendrogram at the top of the figure represents the maximum likelihood phylogenetic tree of the prophages based on protein sequences of integrases. The names of genomes in which the phage was identified appear below the dendrogram, and a suffix of "_1" and "_2" marks the two prophages that were identified in T_C_F_MAG_00008. "GC": marks the integrases that were in non-singleton GCs. "Clade": the clade or subclade (if one exists) association of the host of each prophage. "Phage group": phage group designation. "Same genome": highlights two prophages from T_C_F_MAG_00008. "Type of phage": either inovirus (green) or caudovirales (pink).

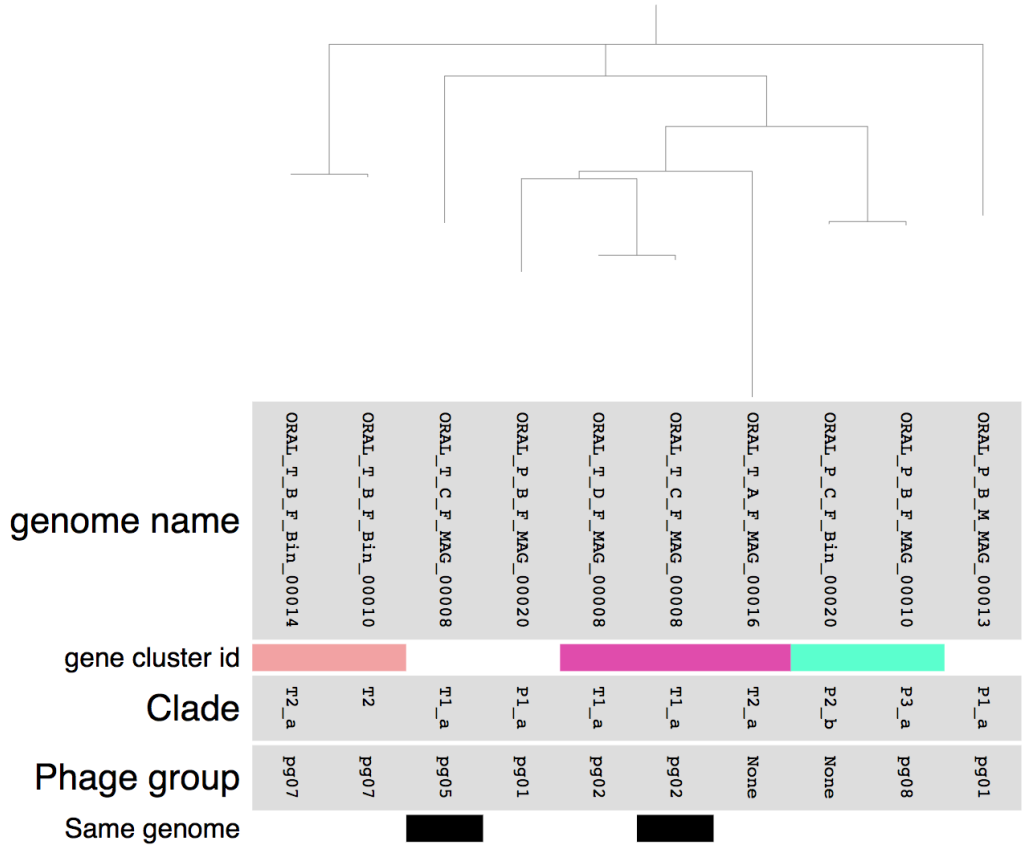

Figure 35: Phylogeny of phages based on terminases. The dendrogram at the top of the figure represents the maximum likelihood phylogenetic tree of the prophages based on protein sequences of terminase large subunit. The names of genomes in which the phage was identified appear below the dendrogram. “Gene cluster id”: marks the integrases that were in non-singleton GCs. “Clade”: the clade or subclade (if one exists) association of the host of each prophage. “Phage group”: phage group designation. “Same genome”: highlights two prophages from T_C_F_MAG_00008.

2.7.6 Novel non-CPR MAGs

Our collection of MAGs included 43 genomes with no closely related genome in HOMD (Figure 2, Supplementary table 10a at doi:10.6084/m9.figshare.11634321). In order to test the novelty of these genomes, we blasted the protein sequences of the ribosomal proteins of these populations against the NCBI non redundant protein sequences database. In conjunction with the phylogenetic analysis (Figure 2), blast results confirmed that 34 of these genomes represent 11 lineages with no representation on NCBI (from here on referred to as “novel MAGs”), while the additional 9 genomes belong to two lineages from the family Eubacteriaceae and matched genomes of Stomatobaculum longum and Lachnospiraceae bacterium oral taxon 096 on the NCBI, which were absent from the HOMD at the time that we downloaded

the HOMD genomes, but have since then been added (Supplementary tables 10b, 10c at doi:10.6084/m9.figshare.11634321).

2.7.7 A novel MAG for a member of the Mollicutes

Members of the Mollicutes, a class of bacteria that lack cell wall (Davis et al. 2013) are known to be commonly found in the human oral cavity. In particular, Mycoplasma are highly ubiquitous members of the oral microbiome (Dewhirst et al. 2010) and include some pathogens. Studies based on 16S rRNA amplicons identified two taxons, HMT-504 and HMT-906, as potential members of the Mollicutes on a deep phylogenetic branch between other known Mollicutes and members of the class Erysipelotrichia (Dewhirst et al. 2010). T_C_F_MAG_00011 has no closely related genome on GenBank (Supplementary table 10c at doi:10.6084/m9.figshare.11634321) and our phylogenomic analysis with representatives of all taxa under the classes Mollictutes and Erysipelotrichia as available on GenBank on 12/24/2018. (Figure 36) placing it deeply branching between these two classes, suggesting it could represent either HMT-504 or HMT-906. Notice that we excluded two GenBank genomes annotated as Erysipelotrichia (GCF.900120365.1, GCF.000178255.1) from our analysis, since our preliminary phylogenetic analysis showed these are likely not members of Erysipelotrichia. The closest genomes to T_C_F_MAG_00011 on were members of the genus acholeplasma, including many plant pathogens, but also including a horse oral pathogen (Atobe, Watabe, and Ogata 1983). Our analysis using the HMP metagenomes showed that T_C_F_MAG_00011 is associated with the tongue and occurs in 20% of HMP individuals for which tongue samples are available (Figure 20, Supplementary table 10c at doi:10.6084/m9.figshare.11634321).

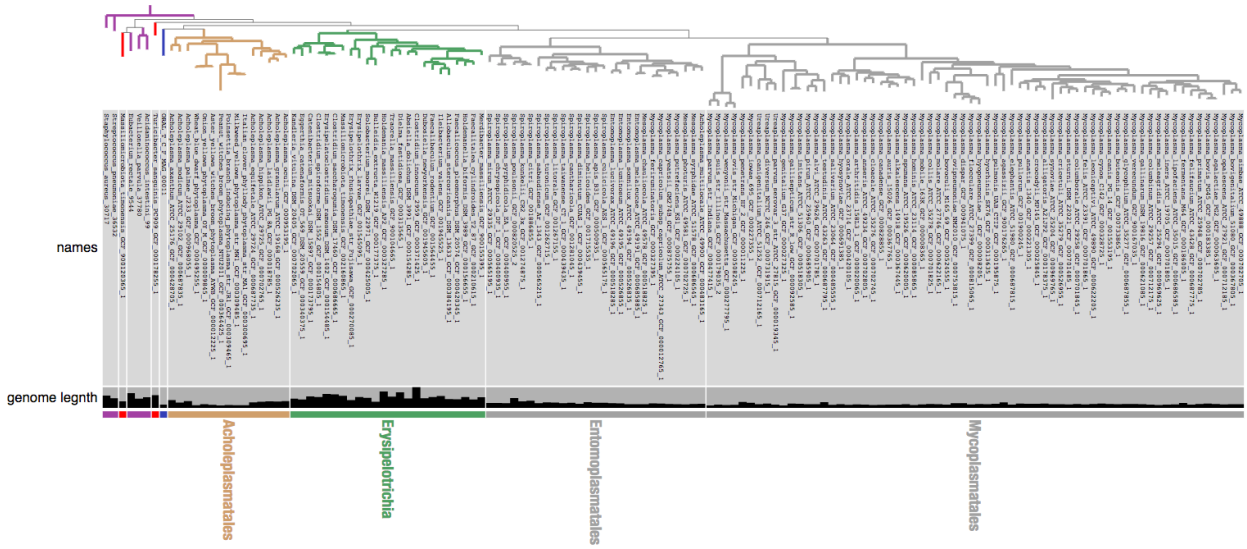

Figure 36: Phylogeny based on ribosomal proteins places T_C_F_MAG_00011 closest to genomes of Acholeplasmatales. Phylogenetic tree of T_C_F_MAG_00011 (blue) together with RefSeq genomes of class Erysipelotrichia (green), phylum Tenericutes, including class Mollicutes, and within it orders Entomoplasmatales and Mycoplasmatales (grey), and Acholeplasmatales (brown), along with five other Firmicutes, representing classes Bacilli, Clostridiales, and Negativicutes as outliers to root the phylogeny (purple). Two genomes wrongly annotated as Erysipelotrichia appear in red color.

2.7.8 Novel Clostridiales MAGs represent prevalent tongue-associated populations

We recovered 5 Clostridiales MAGs for which we could not assign a family designation (Figure 37). 3 MAGs were closely related and seem to represent a prevalent tongue-associated species, and were detected in >50% of HMP tongue metagenomes (Figure 20). We detected an additional population (T_A_M_MAG_00009) in 30% of tongue samples and 20% of plaque samples, while T_C_M_MAG_00006 was detected only in seven HMP tongue samples (3%), and were each distant phylogenetically from any other genome on our phylogenomic analysis using all Clostridiales genomes from (Supplementary tables 10e-h at doi:10.6084/m9.figshare.11634321).

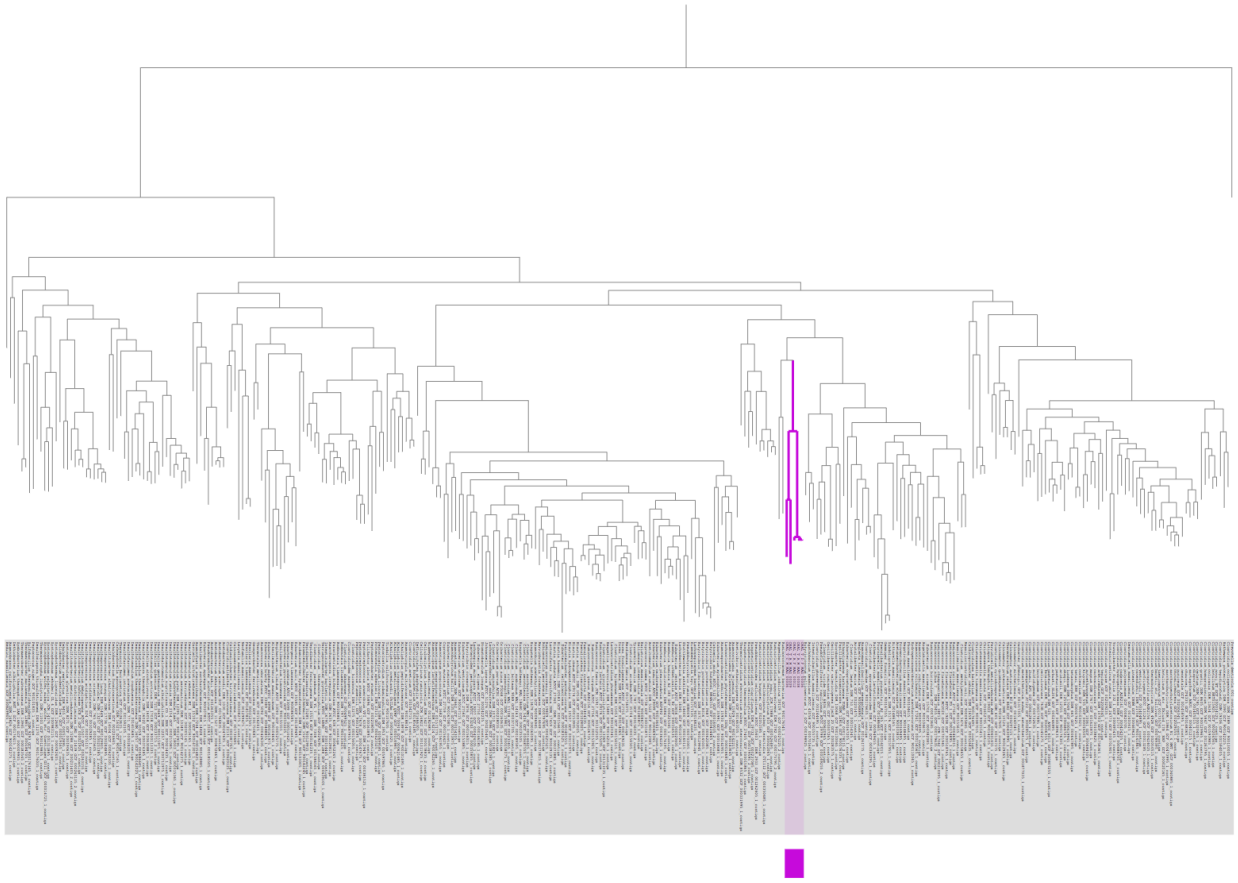

Figure 37: Phylogenomic analysis of Clostridiales genomes from NCBI with our Clostridiales MAGs. A maximum likelihood phylogenetic tree was computed based on our collection of ribosomal proteins using representative genomes for all taxa of order Clostridiales in RefSeq. Our MAGs are highlighted with purple color. The tree was rooted using a Prevotella genome.

2.7.9 Novel Bacteroidia MAGs include a tongue-specialist and a subgingival plaque specialist

One of our Bacteroidia MAGs (P-A-M_MAG_00010) matched a genome recently recovered from a metagenomic sample of periodontal pockets of a patient with periodontitis (McLean et al. 2015)) and seems to represent the same species. Mclean et al. named this population Candidatus Bacteroides periocalifornicus (CBP), an odd choice given the fact that phylogenomic analyses show that it is not a member of the genus bacteroides (McLean et al. 2015). Torres et al. (Torres et al. 2019) showed that this CBP is enriched in subgingival plaque samples as compared to supragingival plaque samples, which our analysis also confirms (Figure 21, Figure 22), an expected result as both analyses relied on the same HMP samples. Two closely related Bacteroidia (T_B_M_MAG_00007, T_C_F_MAG_00010) were prevalent in tongue samples, and detected in 40% of HMP tongue samples (Figure 20, Supplementary table 10f at

doi:10.6084/m9.figshare.11634321). CBP was the closest relative to these MAGs, but with an average of 76% identity of the amino-acid sequences of ribosomal proteins, suggesting that these two lineages are distant and potentially represent distinct genera or families within Bacteroidia.

CHAPTER 3 The anvi'o workflows: extensible, scalable, integrated microbial 'omics

Alon Shaiber[1,2], A. Murat Eren[1,2,3,4,*]

[1] Department of Medicine, University of Chicago, Chicago, IL, USA
[2] Graduate Program in Biophysical Sciences, University of Chicago, Chicago, IL 60637, USA
[3] Committee on Microbiology, University of Chicago, Chicago, IL 60637, USA
[4] Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA

* Correspondence: meren@uchicago.edu

Author contribution

AS conceived the study, designed and implemented the software, and wrote the manuscript. AME supervised the work and wrote the manuscript.

Summary. Affordable high-throughput sequencing strategies and rapidly emerging new 'omics approaches revolutionize microbiology and offer unprecedented access to the ecology and evolution of naturally occurring microbial life. However, accelerated progress in microbiology amidst this data revolution requires the empowerment of microbiologists with software tools that enable integrated analyses of complex 'omics data. Anvi'o is an open-source, community-driven analysis and visualization platform that empowers microbiologists to work with multiple 'omics strategies, perform exploratory data analyses, and visualize large datasets interactively. Yet, implementing an 'omics workflow in anvi'o starting from raw sequencing data requires the orchestration of a large number of atomic computational tasks, which can be discouraging. Here we implement an easy-to-use and extensible workflow management strategy for anvi'o to lower barriers for complex 'omics analyses. **Availability.** The URL http://github.com/merenlab/anvio serves the codebase for the anvi'o snakemake workflows and the URL http://merenlab.org/anvio-workflows serves a comprehensive user tutorial.

3.1 Introduction

Advances in molecular approaches and sequencing chemistry have turned every corner of biology into a 'data-enabled' discipline, including microbiology, the study of the most diverse and numerous forms of life (Whitman, Coleman, and Wiebe 1998) that makes our planet continue to tick (Falkowski, Fenchel, and Delong 2008). New data emerging form increasingly popular 'omics data generation approaches (i.e., metagenomics, metatranscriptomics, metaproteomics, etc) offer new insights into the ecology and evolution of microbes through new 'omics strategies (i.e., genome-resolved metagenomics, pangenomics, phylogenomics, etc).

We previously have introduced anvi'o (Eren et al. 2015), a comprehensive software platform that affords in-depth analyses of 'omics data (Delmont et al. 2018; Reveillaud et al. 2019; Yeoman et al. 2019) through interactive visualization strategies and extensive online tutorials. As of today, anvi'o comprises more than hundred programs, each of which performs individual tasks that can be flexibly combined to build complex analytical workflows (represented as a network at http://merenlab.org/nt). However, preparing raw sequencing data for exploratory analyses in anvi'o typically require many atomic steps of computation that dramatically increase with number of samples (i.e., quality filtering, assembly of short reads, read recruitment, etc). For instance, our recent genome-resolved metagenomics survey of 7 genomes in the context of 88 metagenomes resolved to more than 3,000 atomic steps of computation (Shaiber and Eren 2019), which demonstrates that even a relatively simple 'omics analysis can become intractable for those who do not have substantial training in bioinformatics.

3.2 The anvi'o workflows

Here we present the anvi'o workflows, a collection of commonly-used bioinformatics strategies for microbial 'omics. The anvi'o workflows rely on Snakemake (Köster and Rahmann 2012), which offers easy deployment to any computer system, automatic parallelization of independent analysis steps, and the ability to resume an interrupted workflow without repeating steps that were successfully executed. In many ways the anvi'o workflows are comparable to previous studies that offered means to streamline 'omics analyses

(Dean et al. 2018; Uritskiy, DiRuggiero, and Taylor 2018; Arkin et al. 2018; Clarke et al. 2019; Stewart et al. 2019; Murovec, Deutsch, and Stres 2019; P.-E. Li et al. 2017; Kieser et al. 2019; Naccache et al. 2014), but instead of static figures and tables, our workflows yield data products for the anvi'o ecosystem, enabling interactive exploration of the initial analyses.

The URL http://merenlab.org/anvio-workflows serves an online user tutorial.

3.3 General design

The main entrance point of the anvi'o workflows is the command line program anvi-run-workflow, which distributes within the codebase of anvi'o v5 or later and currently gives access to four workflows: contigs, metagenomics, pangenomics, and phylogenomics. These workflow management system is a collection of Python modules designed with object oriented principles in mind and use multiple inheritance models to extend any workflow with another, whether they are 'built-in' workflows described here, or 'external' workflows that can be implemented and specified by users. The anvi'o workflows dynamically generate template JSON configuration files with default options for users to edit, processes user-provided configuration files, sanity checks the input data, and imports Snakemake (Köster and Rahmann 2012) Python modules to resolve task dependencies and task scheduling within the boundaries of user-defined computational resources. A detailed description of each workflow is provided below.

3.4 Contigs workflow

The contigs workflow includes steps for annotating FASTA files using the anvi'o contigs database. The only mandatory step includes running anvi-run-contigs-database, which includes running prodigal (Hyatt et al. 2010) for gene calling amongst other steps described in Eren et al. (Eren et al. 2015). Optional steps include identifying single copy core genes (SCGs), functional annotation, taxonomy assignment and more (supplementary text 01). To enable handling FASTA files, the contigs workflow is inherited by all other built-in workflows.

3.5 Metagenomics workflow

Metagenomes are rich with information and highly complex, and as such their analysis could take many forms. Accordingly, the metagenomics workflow, includes two modes: 1. Assembly-based analysis (“default mode”) 2. Reference/s-based analysis (“references mode”). At the core of both modes is the generation and annotation of an anvi’o profile database that can be used to explore metagenomic data using the anvi’o interactive interface. The entry point to the default mode is a collection of FASTQ files of pair-end reads, and the output is an annotated merged profile database that is ready for manual binning and curation. This workflows includes all steps from quality filtering, assembly, automatic binning, mapping, taxonomic profiles, and more (supplementary text 02). Along with assembly and binning, metagenomes are often used to explore occurence of individual genes or whole genomes across metagenomes (Delmont and Eren 2017). The “references mode” is intended for this purpose, and takes a collection of FASTQ files and a collection of FASTA files as input. With the exception of the assembly steps, all other steps are performed as described for the “default mode”.

3.6 Phylogenomics workflow

Phylogenomics is a widely used approach to study the evolutionary relationships of organisms using genomic sequences. The contigs workflow is used to perform any required steps such as generating contigs databases and identifying SCGs, which are then exported as amino-acid sequences, concatenated, aligned, and trimmed prior to the calculation of a maximum likelihood phylogenetic tree (see supplementary text 03 for more details).

3.7 Pangenomics workflow

A pangenomic analysis includes the comparison of the set of genes encoded in a collection of genomes. Running a pangenomic analysis using anvi’o is simple and includes two steps, assuming contigs databases have been generated (see supplementary text 04). By inheriting the contigs and phylogenomics workflow, anvi-run-workflow can take a list of FASTA files as input, and generate a pangenomic database, ready for

visualization in the interactive interface, and would optionally include functional annotation, average nucleotide identity (ANI), and a phylogenetic tree.

3.8 Conclusion

The anvi'o workflows streamline the analysis of microbial 'omics data. The utilization of the Snakemake workflow management system along with an easy-to-use interface allows for scientists with minimal computational expertise to process large 'omics datasets, and thus enjoy the wide range of visualization and analysis approaches that anvi'o offers. More information, including examples for common use cases, and answers to frequently asked questions is available on the tutorial at: http://merenlab.org/2018/07/09/anvio-snakemake-workflows/.

Acknowledgements

We thank Jarrod J. Scott (ORCiD: 0000-0001-9863-1318) and Ryan Bartelme (ORCiD: 0000-0002-6178-2603) for testing and reporting their experience with our workflows on anvi'o Slack channel. We also thank the members of the Meren Lab for their input and patience with us. This work was supported in part by the NIH RC2DK122394 and the Mutchnik Family Fund.

3.9 Supplementary text 01 - contigs workflow

The contigs workflow includes the mandatory step of generating an anvi'o contigs database using anvi-gen-contigs-database, which computes and stores tetra-nucleotide frequencies and GC-content of contigs, and uses Prodigal (Hyatt et al. 2010) to identify and store information regarding open reading frames. Optional steps of the workflow include 'anvi-script-reformat-fasta', which is run prior to generating a contigs database, in order to reformat FASTA and simplify the names of contigs and/or remove short contigs; 'anvi-run-hmms', which by default runs built-in HMM profiles, for the identification of single-copy core genes (SCGs) and ribosomal RNAs, but also allows users to provide custom HMM profiles; Centrifuge (Kim et al. 2016) to annotate genes with taxonomy; functional annotation using one or more of the following: 'anvi-run-

ncbi-cogs', 'anvi-run-pfams', or eggNOG-mapper (Huerta-Cepas et al. 2017); 'anvi_run_scg_taxonomy' to annotate SCGs with taxonomy. Along with the config file, the contigs workflow requires a "fasta.txt" input, which is a TAB-delimited file to specify paths to the relevant input file. In addition to performing all the aforementioned steps, the contigs workflow could be easily utilized to work with genomes available on the NCBI's genomic databases in conjunction with 'ncbi-genome-download' (https://github.com/kblin/ncbi-genome-download) as is described here: http://merenlab.org/2019/03/14/ncbi-genome-download-magic/.

3.10 Supplementary text 02 - metagenomics workflow

Mandatory steps of the "default mode" of the metagenomics workflow include running assembly with MEGAHIT (D. Li et al. 2015), IDBA-UD (Peng et al. 2012), or metaSPAdes (Nurk et al. 2017); resulting FASTA files are processed using the aforementioned contigs workflow; short reads are mapped to the assembly using Bowtie2 (Langmead and Salzberg 2012); SAM files are converted to BAM files using SAMtools (H. Li et al. 2009); BAM files are sorted and indexed using 'anvi-init-bam', and together with the contigs databases are used to generate profile databases for each metagenome using 'anvi-profile' (Eren et al. 2015). Individual profile databases are merged using 'anvi-merge'. Optional steps include quality filtering using 'iu-filter-quality-minoche' and generation of a tabular summary of quality filtering results; the execution of one or more automatic binning algorithms using anvi-cluster-contigs, which currently clusters contigs using CONCOCT (Alneberg et al. 2013), METABAT2 (Kang et al. 2019), MAXBIN2 (Wu, Simmons, and Singer 2016), and/or BINSANITY (Graham, Heidelberg, and Tully 2017), and refines clustering results using DAS Tool (Sieber et al. 2018); taxonomic profiles of metagenomes created using KrakenUniq (Breitwieser, Baker, and Salzberg 2018) and imported into the profile databases; removal of short reads based on mapping using Bowtie2 to one or more reference FASTA files, which for example, could be used to remove human contamination from gut metagenomes by mapping to the human genome; summarizing profile databases using 'anvi-summarize'; and splitting self-contained profile and contigs databases using 'anvi-split'.

3.11 Supplementary text 03 - phylogenomics workflow

The phylogenomics workflow (which is extensively discussed here: http://merenlab.org/2018/07/09/anvio-snakemake-workflows/#phylogenomics-workflow) accepts three kinds of input:

1. An Internal genomes file
2. External genomes file
3. A “fasta.txt” file (same as for the contigs workflow)

The format of internal and external genomes files is described here: http://merenlab.org/2016/11/08/pangenomics-v2/#generating-an-anvio-genomes-storage. The contigs workflow is then used to perform any required steps, so that protein sequences of user-specified SCGs could be extracted from contigs databases using ‘anvi-get-sequnces-for-hmm-hits’, which aligns the protein sequences with either FAMSA (Deorowicz, Debudaj-Grabysz, and Gudyś 2016) or MUSCLE (Edgar 2004). Protein alignment is trimmed using trimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009), and a maximum likelihood phylogenetic tree is computed using IQ-TREE (Nguyen et al. 2015). When inherited by the pangenomics workflow (see below), the phylogeny could alternatively be computed using sequences exported using ‘anvi-get-sequences-for-gene-clusters’, which exports and aligned protein sequences using qualifying criteria that allow the identification of single copy core gene-clusters that are suitable for phylogenomics (See http://merenlab.org/2016/11/08/pangenomics-v2/#scrutinizing-phylogenomics).

3.12 Supplementary text 04 - pangenomics workflow

The anvi’o pangenomic workflow includes two steps: generating a genomes storage using ‘anvi-gen-genomes-storage’, and generating a pangenomic database using ‘anvi-pan-genome’, assuming the existence of ‘external genomes’ or ‘internal genomes’ as the entry point (http://merenlab.org/2016/11/08/pangenomics-v2/). The pangenomic workflow of ‘anvi-run-workflow’ allows the option of providing a collection of FASTA files as input (“fasta.txt”) in addition to internal and external genomes files. If a collection of FASTA files were provided, then the inherited contigs workflow is executed

with all the user-specified steps to generate annotated contigs databases, and an external genomes file will also be automatically produced. The phylogenomics workflow is inherited as well, and computed phylogenetic trees are automatically imported into the pangenomic database, and subsequently included in the interactive interface. Genome similarity is optionally computed using 'anvi-compute-genome-similarity', which currently includes sequence similarity calculations using PYANI (Pritchard et al. 2016), fastANI (Jain et al. 2018), or sourmash (Brown and Irber 2016). Genome similarity scores are then imported into the pangenomic database and presented in the interactive interface.

CHAPTER 4 EXAMPLES OF APPLICATIONS OF ANVI'O WORKFLOWS

Our motivation in developing the anvi'o workflows was originally driven by the need to perform metagenomic read recruitment studies, and the realization that performing such analyses at scale and with minimal effort, is an important need in the scientific community. We have expanded the workflow to cover additional common analysis types such as phylogenomics and pangenomics. By streamlining preprocessing steps, the anvi'o workflows allow researchers to easily utilize the anvi'o interactive interface for the exploratory investigation required to make sense of complex sequencing data. The following sections provide descriptions of applications of the anvi'o workflows to address various questions in microbial ecology, with a focus on genome resolved metagenomics, and thus demonstrate the utility of this tool to promote reproducibility and accessibility of microbial 'omics analysis at scale. Section 4.1 expands on the refinement of metagenome assembled genomes (MAGs) of cryptic members of the oral cavity, and demonstrates the importance of adhering to MAG quality guidelines set by the scientific community. Section 4.2 provides an example of the dangers in heavy reliance on MAG quality metrics with no manual exploration of 'omics data. While sections 4.1 and 4.2 serve as warnings against the misleading potential of poorly constructed MAGs, in section 4.3 we demonstrate the advantage of generating MAGs versus studying raw assemblies of metagenomes by expanding on the recovery of a Candidatus Parcubacteria genome from blood samples of pregnant women. Finally, section 4.4 includes an additional application of anvi'o workflows to study metagenomes of mosquito overies, and the discovery of a putative plasmid in the widespread arthropod parasite *wolbachia*.

4.1 Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories

Work published in mBio (Shaiber and Eren 2019)

In their recent study, Espinoza et al. employ genome-resolved metagenomics to study supragingival plaque metagenomes of 88 individuals (1). The 34 metagenome-assembled genomes (MAGs) that the authors report include those that resolve to clades that have largely evaded cultivation efforts, such as Gracilibacteria (formerly GN02) and Saccharibacteria (formerly TM7) of the recently described Candidate

Phyla Radiation (2). Generating new genomic insights into the understudied members of the human oral cavity is of critical importance for a comprehensive understanding of the microbial ecology and functioning of this biome, and we acknowledge the contribution of the authors on this front. However, the redundant occurrence of bacterial single-copy core genes suggest that more than half of the MAGs Espinoza et al. report are composite genomes that do not meet the recent quality guidelines suggested by the community (3). Composite genomes that aggregate sequences originating from multiple distinct populations can yield misleading insights when treated and reported as single genomes (4).

To briefly demonstrate their composite nature, we refined some of the key Espinoza et al. MAGs through a previously described approach (5) and the data the authors kindly provided (1). We found that MAG IV.A, MAG IV.B, and MAG III.A described multiple discrete populations with distinct distribution patterns across individuals (Figure 38). A phylogenomic analysis of refined MAG IV.A genomes resolved to the candidate phylum Absconditabacteria (formerly SR1), and not to Gracilibacteria as reported by Espinoza et al. (Figure 38D). A pangenomic analysis of the original and refined MAG III.A genomes with other publicly available Saccharibacteria genomes showed 7-fold increase in the number of single-copy core genes (Figure 38E). These findings demonstrate the potential implications of composite MAGs in comparative genomics studies where single-copy core genes are commonly used to infer diversity, phylogeny, and taxonomy (6). Composite MAGs can also lead to inaccurate ecological insights through inflated abundance and prevalence estimates. For instance, the original MAG III.A recruited a total of 1,849,593 reads from Espinoza et al. metagenomes, however, the most abundant refined III.A genome (MAG III.A.2, Figure 38C), recruited only 629,291 reads.

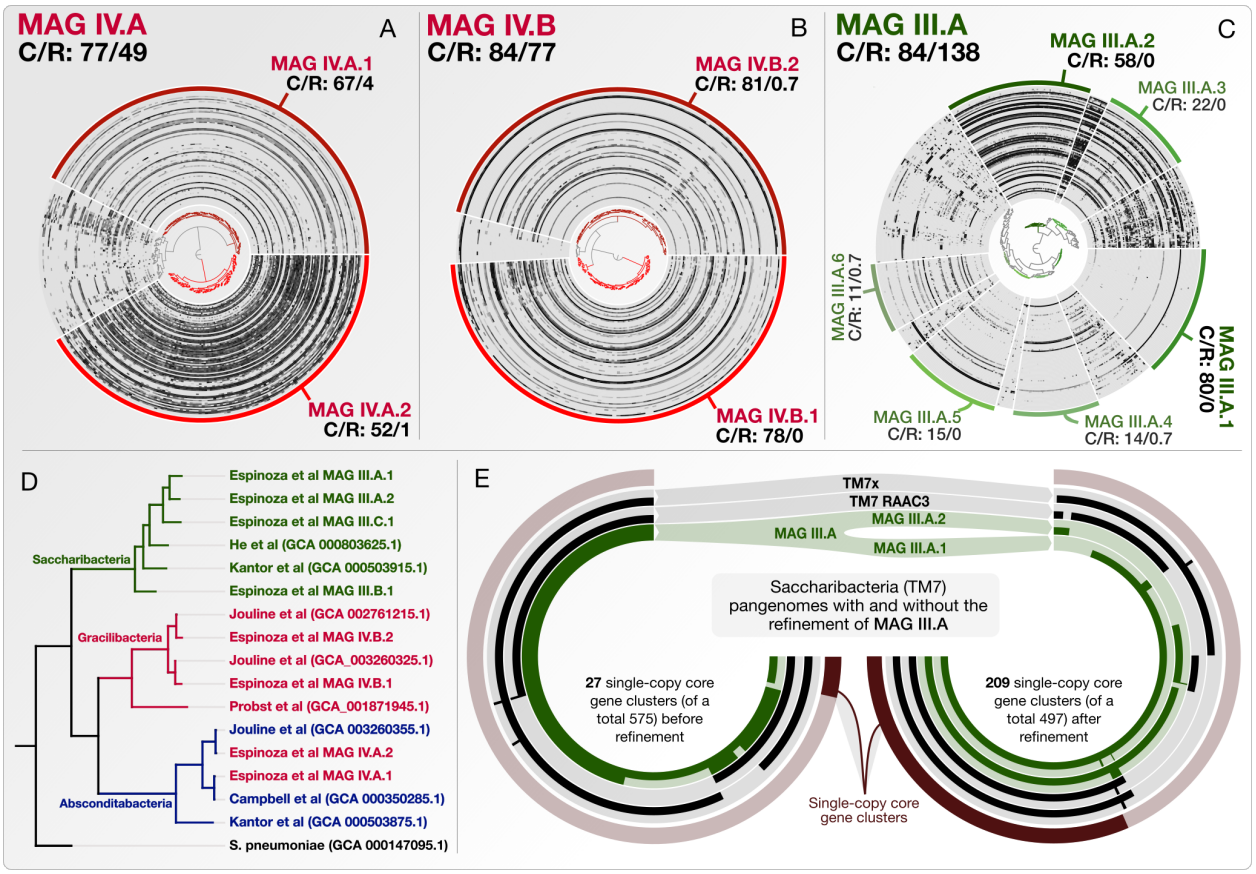

Figure 38: Refinement of three composite genome bins. (A-C) The top-left corners of these panels display the original name of a given Espinoza et al. MAG (see Table 1 in the original study) and its estimated completion and redundancy (C/R) based on a bacterial single-copy core gene collection (7). Each concentric circle represents one of the 88 metagenomes in the original study, dendrograms show hierarchical clustering of contigs based on sequence composition and differential mean coverage across metagenomes (using Euclidean distance and Ward's method), and each data point represents the read recruitment statistic of a given contig in a given metagenome. Arcs at the outmost layers mark contigs that belong to a refined bin along with their new completion and redundancy estimates (C/R). (D) The phylogenomic tree organizes genomes based on 37 concatenated ribosomal proteins. Coloring of genome names match their taxonomy on NCBI, and branch colors match the consensus taxonomy of genomes they represent. Espinoza et al. reported MAG IV.A as Gracilibacteria (hence the red color), however this phylogenomic analysis places refined MAGs under Absconditabacteria. (E) Pangenomic analysis of Espinoza et al. Saccharibacteria MAG III.A before (left) and after (right) refinement together with the Saccharibacteria genomes from panel D. Pangenomes describe 575 and 497 gene clusters, respectively, where each concentric circle represents a genome and bars correspond to the number of genes a given genome contributing to a given gene cluster (the maximum value is set to 2 for readability). Outermost layers mark single-copy core gene clusters to which every genome contributes precisely a single gene. We used Bowtie2 (8) to recruit reads from metagenomes, and anvi'o (9) to visualize and refine Espinoza et al. MAGs. FAMSA (10) aligned anvi'o-reported ribosomal protein amino acid sequences, trimAl (11) curated them, and IQ-TREE (12) computed the tree for the phylogenomic analysis. Anvi'o used DIAMOND (13) and MCL (14) algorithms to determine pangenomes. A reproducible bioinformatics workflow and FASTA files for refined MAGs are available at http://merenlab.org/data/refining-espinoza-mags.

Co-assembly of a large number of metagenomes that contain very closely related populations often hinders confident assignments of shared contigs into individual bins. Nevertheless, even when proper refinement is not possible, reporting composite MAGs as single genomes should be avoided. As of today, highly composite Espinoza et al. MAGs (Figure 38 in this letter and Table 1 in Espinoza et al.) are available as single genomes in public databases of the National Center for Biotechnology Information (NCBI).

The rapidly increasing number of MAGs in public databases already competes with the total number of microbial isolate genomes (3), and increasingly frequent studies that report large collections of MAGs offer a glimpse of the future (15–17). Despite their growing availability, metagenomes are inherently complex and demand researchers to orchestrate an intricate combination of rapidly evolving computational tools and approaches with many alternatives to reconstruct, characterize, and finalize MAGs. We must continue to champion studies such as the one by Espinoza et al. for their contribution to our collective effort to shed light on the darker branches of the ever-growing Tree of Life. At the same time, editors and reviewers of genome-resolved metagenomics studies should properly scrutinize the quality and accuracy of MAGs prior to their publication. A systematic failure at this will reduce the quality of public genome repositories while yielding adverse effects such as misleading insights into novel microbial groups and reduced trust among scientists in findings that emerge from genome-resolved metagenomics.

4.2 Standard Quality Measures For Metagenome Assembled Genomes Can Fail To Properly Predict the Quality of MAGs

Work described in preprint at bioRxiv (Chen et al. 2019c)

Recent studies employing single-assembly strategies and automatic binning are generating hundreds of thousands of metagenome-assembled genomes (MAGs), while heavily relying on metrics of MAG quality that are primarily based on occurrence of single copy core genes (SCGs), without the manual verification of MAG quality. (Almeida et al. 2019; Nayfach et al. 2019; Pasolli et al. 2019). While such studies expose previously unknown branches of the Tree of Life (Leviatan and Segal 2019), occurrence of SCGs might not be sufficient to support claims of MAG quality, and low quality MAGs could yield false conclusions (Shaiber

and Eren 2019). Pasolli et al. suggest that the MAGs they have reconstructed using this approach were comparable in quality to genomes of isolates or MAGs that are refined through manual processes (Pasolli et al. 2019). To highlight the potential pitfalls of automatic binning with no manual refinement, and in particular, the shortcomings of heavily relying on occurrence of SCGs to infer MAG quality, we examined one of the MAGs published by Pasolli et al. (Pasolli et al. 2019) (hereafter referred to as Pasolli MAG), that resolves to the candidate phylum TM7, a member of a poorly understood branch of the Tree of Life (Brown et al. 2015), that contains commonly found members of the oral microbiome (B. Bor et al. 2019).

Our recent publication with Chen and colleagues (Chen et al. 2019c) includes a description of the application of the anvi'o workflows to recruit reads from 481 Human Microbiome Project (HMP) oral samples, including the HMP sample that was originally used by Pasolli et al. to assemble and bin Pasolli MAG. Following read recruitment, we utilized the anvi'o interactive interface to identify contigs in Pasolli MAG that are contamination, originating from non-TM7 genomes, based on coverage patterns of contigs in Pasolli MAG. Using sequence search against the National Center for Biotechnology Information (NCBI) genomic databases, we further showed that the contaminating contigs primarily originated from *Veillonella*, and that these contaminating contigs were transparent to the quality measures applied by Pasolli et al. due to lack of SCGs. This work demonstrates that lack of SCGs does not imply lack of contamination in a MAG, and that heavily relying on SCGs to estimate MAG quality could lead to erroneous insight.

4.3 Binning Contigs Into Metagenome Assembled Genomes Can greatly improve data interpretation

Work described in preprint at bioRxiv (Chen et al. 2019c)

Analysis of shotgun metagenomes could take many forms, and common applications include assembly of short reads into contigs followed by either an analysis of these contigs as independent units, or binning of contigs into metagenome-assembled genomes (MAGs) (Quince et al. 2017). While analysis of contigs without binning could be appealing as an approach to circumvent challenges presented by the process of binning MAGs (Quince et al. 2017), claims made based on analysis of contigs that are not binned according to genomic affiliations may lead to erroneous conclusions.

In addition to the example mentioned in section 4.2, in Chen et al. 2019 (Chen et al. 2019c) we discuss a case study in which we demonstrate the contrast between a contigs-centric analysis (i.e. without binning) and genome-resolved analysis (i.e. with binning of contigs into MAGs) by reanalyzing the data of Kowarsky et al. (Kowarsky et al. 2017). In order to explore microbial diversity in blood samples, Kowarsky et al. (Kowarsky et al. 2017) performed shotgun sequencing of circulating cell-free DNA from more than 1000 samples. Kowarsky et al. identify a total of 3,761 novel contigs that do not match any known bacteria or virus in public databases with sequence homology, and by assigning taxonomy independently to each of these contigs, they conclude that these represent at least 1000 novel taxa of the human microbiome that represent both bacteria and viruses. Using a genome-resolved approach, we showed that a single Parcubateria genome is the only dominant bacterial source for novel contigs, contrasting with Kowarsky et al.'s finding (Chen et al. 2019c). In our re-analysis of the Kowarsky et al. samples we utilized the anvi'o workflows to streamline read recruitment of the Kowarsky et al. cell free DNA metagenomes to the novel contigs as well as non-novel contigs published in the original Kowarsky et al. study. The read recruitment analysis allowed us to utilize coverage patterns, along with sequence composition when clustering contigs in order to identify genomic bins confidently (Quince et al. 2017). Due to the low coverage of contigs in these metagenomes, we used differential detection, rather than the more common differential coverage of contigs in order to cluster contigs (see the reproducible workflow at http://merenlab.org/data/parcubacterium-in-hbcfdna/ for full details).

In summary, our reanalysis of the Kowarsky et al. samples, contrasts with their suggestion of more than 1,000 novel species found in blood samples, and instead suggests that a single Parcubacteria population is the only dominant source for novel bacterial contigs in these blood samples.

4.4 A genome resolved metagenomics strategy to explore the intra-species diversity and mobilome of *Wolbachia*

Work published in Nature Communications (Reveillaud et al. 2019)

Wolbachia are intracellular bacteria that are common parasites of nematodes and arthropods, including mosquitoes that are vectors that transmit diseases such as dengue, West Nile, and Zika viruses (Taylor, Bordenstein, and Slatko 2018; LePage and Bordenstein 2013; Stouthamer, Breeuwer, and Hurst 1999). New promising vector control strategies have been developed using *Wolbachia* due to their natural ability, through the temperate bacteriophage WO, to modify their mosquito host reproductive behavior (Bourtzis et al. 2014; Mains et al. 2016; O'Neill et al. 2018). But a lack of isolates provides challenges in studying the *Wolbachia* genome, and most prior metagenomic studies of *Wolbachia* relied on pooled samples of laboratory grown insects due to the low infection rate (Klasson et al. 2008; Iturbe-Ormaetxe et al. 2011). Studying pooled samples from multiple individuals can obscure variability across populations of *Wolbachia*. In Reveillaud et al. (Reveillaud et al. 2019) we used samples from ovaries of individual wild mosquitoes captured in France to overcome previous limitations, and along with discovering viral genes missing in previously published *Wolbachia* genomes, we identified a putative plasmid (pWCP). Al preprocessing steps required for the genomic binning of *Wolbachia* MAGs, including assembly and read recruitment were executed using the anvi'o workflows. We further utilized the anvi'o workflows to assess the occurrence of pWCP across published metagenomes, and showed that it was widespread and found in samples from Turkey, Algeria and Tunisia.

The discovery of a *Wolbachia* plasmid provides exciting avenues for future genome-editing strategies of *Wolbachia*, which has been recalcitrant to genetic modification to this date. Successful genomic manipulation of *Wolbachia* could enhance the ability to utilize *Wolbachia* for vector control.

4.5 Discussion

The anvi'o interactive interface allows the visualization of complex 'omics data needed for exploratory and effective data mining. By solving a major bottleneck in preprocessing steps required prior to visualization, the anvi'o workflows empower microbiologists by promoting microbial 'omics analyses at scale, and make it so that, pending on accessibility to appropriate computing infrastructure (Kyrpides, Eloe-Fadrosh, and Ivanova 2016), human involvement required for analyzing thousands of samples is as minimal as that required for analyzing a few samples. Utilization of the same data in multiple studies is crucial not only in

order to verify findings, but also since the complex nature of sequencing data guarantees that a single study will not be sufficient to extract the entire value out of the data (Kyrpides, Eloe-Fadrosh, and Ivanova 2016). Promoting reproducibility and reuse of data in microbiology will accelerate discovery even further in this fast evolving field. Indeed, our reanalysis of Espinoza et al., Pasolli et al., and Kowarsky et al. datasets was only possible due to the minimal effort required to process their data, along with the immediate insight provided by exploring these data in the anvi'o interactive interface.

But these advantages are not exclusive to reanalysis projects, by utilizing the anvi'o workflows in studying insect ovary metagenomes our time and effort remained invested in novel exploration of the newly generated data, rather than on the execution of the initial steps of analysis which are largely repetitive and standard. Automated processing of samples at scale also lowers the bar for additional exploratory work. For example, the assessment of the occurrence of TM7 populations in multiple datasets, including HMP oral samples, and samples from patients with periodontitis, which is discussed in Chapter 1, was made easy due to utilization of the anvi'o workflows.

Our work demonstrates that utilization of the anvi'o workflows streamlines the path from raw sequences to interactive visualization that allows high resolution exploratory investigations of 'omics data, and thus promoting reproducibility, and the democratization of data analysis in modern, data-rich microbiology.

CHAPTER 5 CONCLUSIONS

Microbes are abundant, and are abundantly involved in processes of ecological importance (Cavicchioli et al. 2019), and of medical and biotechnological importance (Quince et al. 2017). Advances in sequence technology have significantly enhanced our understanding of microbial ecology and evolution (Quince et al. 2017), and have transformed microbiology into a data-rich science (Kyrpides, Eloe-Fadrosh, and Ivanova 2016). But this transformation provides new challenges to microbiologists, and solutions that allow high resolution exploratory investigations of 'omics data at scale, along with computational training for microbiologists are lacking. The work presented here summarizes my efforts throughout my graduate studies to address these challenges. While focusing on specific questions in microbial ecology, I made efforts to streamline the analysis of microbial 'omics data and solve bottlenecks by striving to develop computational tools that are well-designed, and provide adequate documentation and tutorials to allow for 1) extensibility 2) the accessibility of these tools to people with minimal computational training.

Through my investigations of the oral microbiome, I have expanded our genomic insight into understudied members of the oral cavity. My work revealed plaque-associated TM7 to be much more similar to environmental TM7, rather than to tongue and gut-associated TM7. These findings suggest that at least for TM7 the plaque environment is similar to non-host environments. Applying the approaches presented here to study other taxa could reveal whether plaque resembles non-host environments for other members of the oral microbiome, and could shed light on the adaptation process of environmentally-derived microbes to the host environment.

CHAPTER 6 REFERENCES

Aas, Jørn A., Bruce J. Paster, Lauren N. Stokes, Ingar Olsen, and Floyd E. Dewhirst. 2005. “Defining the Normal Bacterial Flora of the Oral Cavity.” *Journal of Clinical Microbiology* 43 (11): 5721–32.

Abusleme, Loreto, Amanda K. Dupuy, Nicolás Dutzan, Nora Silva, Joseph A. Burleson, Linda D. Strausbaugh, Jorge Gamonal, and Patricia I. Diaz. 2013. “The Subgingival Microbiome in Health and Periodontitis and Its Relationship with Community Biomass and Inflammation.” *The ISME Journal* 7 (5): 1016–25.

Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L. Nielsen, Gene W. Tyson, and Per H. Nielsen. 2013. “Genome Sequences of Rare, Uncultured Bacteria Obtained by Differential Coverage Binning of Multiple Metagenomes.” Nature Biotechnology 31 (6): 533–38.

Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. 2019. “A New Genomic Blueprint of the Human Gut Microbiota.” Nature 568 (7753): 499–504.

Alneberg, Johannes, Brynjar Smari Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. 2013. “CONCOCT: Clustering cONtigs on COverage and ComposiTion.” arXiv [q-bio.GN]. arXiv. http://arxiv.org/abs/1312.4038.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” Journal of Molecular Biology 215 (3): 403–10.

Amato, Katherine R., Jon G Sanders, Se Jin Song, Michael Nute, Jessica L. Metcalf, Luke R. Thompson, James T. Morton, et al. 2019. “Evolutionary Trends in Host Physiology Outweigh Dietary Niche in Structuring Primate Gut Microbiomes.” The ISME Journal 13 (3): 576–87.

Atobe, Hisae, Junko Watabe, and Manabu Ogata. 1983. “Acholeplasma Parvum, a New Species from Horses.” International Journal of Systematic and Evolutionary Microbiology 33 (2): 344–49.

Bahrndorff, Simon, Tibebu Alemu, Temesgen Alemneh, and Jeppe Lund Nielsen. 2016. “The Microbiome of Animals: Implications for Conservation Biology.” International Journal of Genomics and Proteomics 2016 (April): 5304028.

Béjà, Oded, Marcelino T. Suzuki, John F. Heidelberg, William C. Nelson, Christina M. Preston, Tohru Hamada, Jonathan A. Eisen, Claire M. Fraser, and Edward F. DeLong. 2002. “Unsuspected Diversity among Marine Aerobic Anoxygenic Phototrophs.” Nature. https://doi.org/10.1038/415630a.

Bella, J., K. L. Hindle, P. A. McEwan, and S. C. Lovell. 2008. “The Leucine-Rich Repeat Structure.” Cellular and Molecular Life Sciences: CMLS 65 (15): 2307–33.

Belton, Jon-Matthew, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. 2012. “Hi-C: A Comprehensive Technique to Capture the Conformation of Genomes.” Methods 58 (3): 268–76.

Bertrand, Denis, Jim Shaw, Manesh Kalathiyappan, Amanda Hui Qi Ng, M. Senthil Kumar, Chenhao Li, Mirta Dvornicic, et al. 2019. “Hybrid Metagenomic Assembly Enables High-Resolution Analysis of Resistance Determinants and Mobile Elements in Human Microbiomes.” Nature Biotechnology 37 (8): 937–44.

Bor, Batbileg, McLean, Kevin R. Foster, Lujia Cen, Thao T. To, Alejandro Serrato-Guillen, Floyd E. Dewhirst, Wenyuan Shi, and Xuesong He. 2018. “Rapid Evolution of Decreased Host Susceptibility

Drives a Stable Relationship between Ultrasmall Parasite TM7x and Its Bacterial Host." Proceedings of the National Academy of Sciences of the United States of America 115 (48): 12277–82.

Bor, B., J. K. Bedree, W. Shi, J. S. McLean, and X. He. 2019. "Saccharibacteria (TM7) in the Human Oral Microbiome." Journal of Dental Research 98 (5): 500–509.

Bourtzis, Kostas, Stephen L. Dobson, Zhiyong Xi, Jason L. Rasgon, Maurizio Calvitti, Luciano A. Moreira, Hervé C. Bossin, et al. 2014. "Harnessing mosquito–Wolbachia Symbiosis for Vector and Disease Control." Acta Tropica 132 (April): S150–63.

Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, et al. 2017. "Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." Nature Biotechnology 35 (8): 725–31.

Breitwieser, Florian P., and Steven L. Salzberg. 2018. "KrakenHLL: Confident and Fast Metagenomics Classification Using Unique K-Mer Counts." bioRxiv. https://doi.org/10.1101/262956.

Breitwieser, F. P., D. N. Baker, and S. L. Salzberg. 2018. "KrakenUniq: Confident and Fast Metagenomics Classification Using Unique K-Mer Counts." Genome Biology 19 (1): 198.

Brinig, Mary M., Paul W. Lepp, Cleber C. Ouverney, Gary C. Armitage, and David A. Relman. 2003. "Prevalence of Bacteria of Division TM7 in Human Subgingival Plaque and Their Association with Disease." Applied and Environmental Microbiology 69 (3): 1687–94.

Brown, Christopher T., Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H. Williams, and Jillian F. Banfield. 2015. "Unusual Biology across a Group Comprising More than 15% of Domain Bacteria." Nature 523 (7559): 208–11.

Buist, Girbe, Anton Steen, Jan Kok, and Oscar P. Kuipers. 2008. "LysM, a Widely Distributed Protein Motif for Binding to (peptido)glycans." Molecular Microbiology 68 (4): 838–47.

Califf, Katy J., Karen Schwarzberg-Lipson, Neha Garg, Sean M. Gibbons, J. Gregory Caporaso, Jørgen Slots, Chloe Cohen, Pieter C. Dorrestein, and Scott T. Kelley. 2017. "Multi-Omics Analysis of Periodontal Pocket Microbial Communities Pre- and Posttreatment." mSystems. https://doi.org/10.1128/msystems.00016-17.

Camanocha, A., and F. E. Dewhirst. n.d. "Host-Associated Bacterial Taxa from Chlorobi, Chloroflexi, GN02, Synergistetes, SR1, TM7, and WPS-2 Phyla/candidate Divisions. J Oral Microbiol. 2014; 6." Epub 2014/10/16. doi: 10.3402/jom. v6. 25468 PMID: 25317252.

Campbell, James H., Patrick O'Donoghue, Alisha G. Campbell, Patrick Schwientek, Alexander Sczyrba, Tanja Woyke, Dieter Söll, and Mircea Podar. 2013. "UGA Is an Additional Glycine Codon in Uncultured SR1 Bacteria from the Human Microbiota." Proceedings of the National Academy of Sciences of the United States of America 110 (14): 5540–45.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. "trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." Bioinformatics 25 (15): 1972–73.

Caporaso, J. Gregory, Christian L. Lauber, Elizabeth K. Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, et al. 2011. "Moving Pictures of the Human Microbiome." Genome Biology 12 (5): R50.

Cavicchioli, Ricardo, William J. Ripple, Kenneth N. Timmis, Farooq Azam, Lars R. Bakken, Matthew Baylis, Michael J. Behrenfeld, et al. 2019. “Scientists’ Warning to Humanity: Microorganisms and Climate Change.” Nature Reviews. Microbiology 17 (9): 569–86.

Chen, I-Min A., Ken Chu, Krishna Palaniappan, Manoj Pillay, Anna Ratner, Jinghua Huang, Marcel Huntemann, et al. 2019a. “IMG/M v.5.0: An Integrated Data Management and Comparative Analysis System for Microbial Genomes and Microbiomes.” Nucleic Acids Research 47 (D1): D666–77.

Chen, Lin-Xing, Basem Al-Shayeb, Raphaël Méheust, Wen-Jun Li, Jennifer A. Doudna, and Jillian F. Banfield. 2019b. “Candidate Phyla Radiation Roizmanbacteria From Hot Springs Have Novel and Unexpectedly Abundant CRISPR-Cas Systems.” Frontiers in Microbiology 10 (May): 928.

Chen, Lin-Xing, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. 2019c. “Accurate and Complete Genomes from Metagenomes.” bioRxiv. https://doi.org/10.1101/808410.

Chen, Tsute, Wen-Han Yu, Jacques Izard, Oxana V. Baranova, Abirami Lakshmanan, and Floyd E. Dewhirst. 2010. “The Human Oral Microbiome Database: A Web Accessible Resource for Investigating Oral Microbe Taxonomic and Genomic Information.” Database: The Journal of Biological Databases and Curation 2010 (July): baq013.

Clarke, Erik L., Louis J. Taylor, Chunyu Zhao, Andrew Connell, Jung-Jin Lee, Bryton Fett, Frederic D. Bushman, and Kyle Bittinger. 2019. “Sunbeam: An Extensible Pipeline for Analyzing Metagenomic Sequencing Experiments.” Microbiome 7 (1): 46.

Collins, Andrew J., Pallavi P. Murugkar, and Floyd E. Dewhirst. 2019. “Complete Genome Sequence of Strain AC001, a Novel Cultured Member of the Human Oral Microbiome from the Candidate Phylum Saccharibacteria (TM7).” Microbiology Resource Announcements. https://doi.org/10.1128/mra.01158-19.

Couvin, David, Aude Bernheim, Claire Toffano-Nioche, Marie Touchon, Juraj Michalik, Bertrand Néron, Eduardo P. C. Rocha, Gilles Vergnaud, Daniel Gautheret, and Christine Pourcel. 2018. “CRISPRCasFinder, an Update of CRISRFinder, Includes a Portable Version, Enhanced Performance and Integrates Search for Cas Proteins.” Nucleic Acids Research 46 (W1): W246–51.

Craig, Lisa, Katrina T. Forest, and Berenike Maier. 2019. “Type IV Pili: Dynamics, Biophysics and Functional Consequences.” Nature Reviews. Microbiology 17 (7): 429–40.

Cross, Karissa L., James H. Campbell, Manasi Balachandran, Alisha G. Campbell, Sarah J. Cooper, Ann Griffen, Matthew Heaton, et al. 2019. “Targeted Isolation and Cultivation of Uncultivated Bacteria by Reverse Genomics.” Nature Biotechnology 37 (11): 1314–21.

Davis, James J., Fangfang Xia, Ross A. Overbeek, and Gary J. Olsen. 2013. “Genomes of the Class Erysipelotrichia Clarify the Firmicute Origin of the Class Mollicutes.” International Journal of Systematic and Evolutionary Microbiology 63 (Pt 7): 2727–41.

Dean, Christopher, Noelle Noyes, Steven Lakin, Pablo Rovira-Sanz, Xiang Yang, Keith Belk, Paul S. Morley, Rick Meinersmann, and Zaid Abdo. 2018. “Tychus: A Whole Genome Sequencing Pipeline for Assembly, Annotation and Phylogenetics of Bacterial Genomes.” bioRxiv. https://doi.org/10.1101/283101.

Delcher, Arthur L., Adam Phillippy, Jane Carlton, and Steven L. Salzberg. 2002. “Fast Algorithms for Large-Scale Genome Alignment and Comparison.” Nucleic Acids Research 30 (11): 2478–83.

Delmont, Tom O., and A. Murat Eren. 2018. “Linking Pangenomes and Metagenomes: The Prochlorococcus Metapangenome.” PeerJ 6 (January): e4320.

Delmont, Tom O., A. Murat Eren, Lorrie Maccario, Emmanuel Prestat, Özcan C. Esen, Eric Pelletier, Denis Le Paslier, Pascal Simonet, and Timothy M. Vogel. 2015. “Reconstructing Rare Soil Microbial Genomes Using in Situ Enrichments and Metagenomics.” Frontiers in Microbiology 6 (April): 358.

Delmont, Tom O., Evan Kiefl, Ozsel Kilinc, Ozcan C. Esen, Ismail Uysal, Michael S. Rappé, Steven Giovannoni, and A. Murat Eren. 2019. “Single-Amino Acid Variants Reveal Evolutionary Processes That Shape the Biogeography of a Global SAR11 Subclade.” eLife 8 (September). https://doi.org/10.7554/eLife.46497.

Delmont, Tom O., Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny Tm Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker, and A. Murat Eren. 2018. “Nitrogen-Fixing Populations of Planctomycetes and Proteobacteria Are Abundant in Surface Ocean Metagenomes.” Nature Microbiology 3 (7): 804–13.

Deorowicz, Sebastian, Agnieszka Debudaj-Grabysz, and Adam Gudyś. 2016. “FAMSA: Fast and Accurate Multiple Sequence Alignment of Huge Protein Families.” Scientific Reports 6 (September): 33964.

Dewhirst, Floyd E., Tuste Chen, Jacques Izard, Bruce J. Paster, Anne C. R. Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G. Wade. 2010. “The Human Oral Microbiome.” Journal of Bacteriology 192 (19): 5002–17.

Dewhirst, Floyd E., Erin A. Klein, Emily C. Thompson, Jessica M. Blanton, Tsute Chen, Lisa Milella, Catherine M. F. Buckley, Ian J. Davis, Marie-Lousie Bennett, and Zoe V. Marshall-Jones. 2012. “The Canine Oral Microbiome.” PloS One 7 (4): e36067.

Ding, Tao, and Patrick D. Schloss. 2014. “Dynamics and Associations of Microbial Community Types across the Human Body.” Nature 509 (7500): 357–60.

Donati, Claudio, Moreno Zolfo, Davide Albanese, Duy Tin Truong, Francesco Asnicar, Valerio Iebba, Duccio Cavalieri, et al. 2016. “Uncovering Oral Neisseria Tropism and Persistence Using Metagenomic Sequencing.” Nature Microbiology 1 (7): 16070.

Dudek, Natasha K., Christine L. Sun, David Burstein, Rose S. Kantor, Daniela S. Aliaga Goltsman, Elisabeth M. Bik, Brian C. Thomas, Jillian F. Banfield, and David A. Relman. 2017. “Novel Microbial Diversity and Functional Potential in the Marine Mammal Oral Microbiome.” Current Biology: CB 27 (24): 3752–62.e6.

Dutilh, Bas E., Martijn A. Huynen, William J. Bruno, and Berend Snel. 2004. “The Consistent Phylogenetic Signal in Genome Trees Revealed by Reducing the Impact of Noise.” Journal of Molecular Evolution 58 (5): 527–39.

Eloe-Fadrosh, Emiley A., Natalia N. Ivanova, Tanja Woyke, and Nikos C. Kyrpides. 2016. “Metagenomics Uncovers Gaps in Amplicon-Based Detection of Microbial Diversity.” Nature Microbiology. https://doi.org/10.1038/nmicrobiol.2015.32.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. “An Efficient Algorithm for Large-Scale Detection of Protein Families.” Nucleic Acids Research 30 (7): 1575–84.

Eren, A. Murat, Gary G. Borisy, Susan M. Huse, and Jessica L. Mark Welch. 2014. “Oligotyping Analysis of the Human Oral Microbiome.” Proceedings of the National Academy of Sciences of the United States of America 111 (28): E2875–84.

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. “Anvi’o: An Advanced Analysis and Visualization Platform for

'Omics Data." PeerJ 3 (October): e1319.

Eren, A. Murat, Hilary G. Morrison, Pamela J. Lescault, Julie Reveillaud, Joseph H. Vineis, and Mitchell L. Sogin. 2015. "Minimum Entropy Decomposition: Unsupervised Oligotyping for Sensitive Partitioning of High-Throughput Marker Gene Sequences." The ISME Journal 9 (4): 968–79.

Eren, A. Murat, Joseph H. Vineis, Hilary G. Morrison, and Mitchell L. Sogin. 2013. "A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology." PloS One 8 (6): e66643.

Escobar-Zepeda, Alejandra, Elizabeth Ernestina Godoy-Lozano, Luciana Raggi, Lorenzo Segovia, Enrique Merino, Rosa María Gutiérrez-Rios, Katy Juarez, Alexei F. Licea-Navarro, Liliana Pardo-Lopez, and Alejandro Sanchez-Flores. 2018. "Analysis of Sequencing Strategies and Tools for Taxonomic Annotation: Defining Standards for Progressive Metagenomics." Scientific Reports 8 (1): 12034.

Espinoza, Josh L., Derek M. Harkins, Manolito Torralba, Andres Gomez, Sarah K. Highlander, Marcus B. Jones, Pamela Leong, et al. 2018. "Supragingival Plaque Microbiome Ecology and Functional Potential in the Context of Health and Disease." mBio 9 (6). https://doi.org/10.1128/mBio.01631-18.

Falkowski, Paul G., Tom Fenchel, and Edward F. Delong. 2008. "The Microbial Engines That Drive Earth's Biogeochemical Cycles." Science 320 (5879): 1034–39.

Ferretti, Pamela, Edoardo Pasolli, Adrian Tett, Francesco Asnicar, Valentina Gorfer, Sabina Fedi, Federica Armanini, et al. 2018. "Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome." Cell Host & Microbe 24 (1): 133–45.e5.

German, Rebecca Z., and Jeffrey B. Palmer. 2006. "Anatomy and Development of Oral Cavity and Pharynx." GI Motility Online, May. https://doi.org/10.1038/gimo5.

Hall, Michael W., Natasha Singh, Kester F. Ng, David K. Lam, Michael B. Goldberg, Howard C. Tenenbaum, Josh D. Neufeld, Robert G Beiko, and Dilani B. Senadheera. 2017. "Inter-Personal Diversity and Temporal Dynamics of Dental, Tongue, and Salivary Microbiota in the Healthy Oral Cavity." NPJ Biofilms and Microbiomes 3 (January): 2.

Hardoim, Pablo R., Leonard S. van Overbeek, Gabriele Berg, Anna Maria Pirttilä, Stéphane Compant, Andrea Campisano, Matthias Döring, and Angela Sessitsch. 2015. "The Hidden World within Plants: Ecological and Evolutionary Considerations for Defining Functioning of Microbial Endophytes." Microbiology and Molecular Biology Reviews: MMBR 79 (3): 293–320.

Helm, Eric van der, Lejla Imamovic, Mostafa M. Hashim Ellabaan, Willem van Schaik, Anna Koza, and Morten O. A. Sommer. 2017. "Rapid Resistome Mapping Using Nanopore Sequencing." Nucleic Acids Research 45 (8): e61.

He, Xuesong, McLean, Anna Edlund, Shibu Yooseph, Adam P. Hall, Su-Yang Liu, Pieter C. Dorrestein, et al. 2015. "Cultivation of a Human-Associated TM7 Phylotype Reveals a Reduced Genome and Epibiotic Parasitic Lifestyle." Proceedings of the National Academy of Sciences of the United States of America 112 (1): 244–49.

Hover, Bradley M., Seong-Hwan Kim, Micah Katz, Zachary Charlop-Powers, Jeremy G. Owen, Melinda A. Ternei, Jeffrey Maniko, et al. 2018. "Culture-Independent Discovery of the Malacidins as Calcium-Dependent Antibiotics with Activity against Multidrug-Resistant Gram-Positive Pathogens." Nature Microbiology 3 (4): 415–22.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. "ETE 3: Reconstruction, Analysis, and

Visualization of Phylogenomic Data." Molecular Biology and Evolution 33 (6): 1635–38.

Hug, Laura A., Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, et al. 2016. "A New View of the Tree of Life." Nature Microbiology 1 (April): 16048.

Human Microbiome Project Consortium. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." Nature 486 (7402): 207–14.

Huse, Susan M., Les Dethlefsen, Julie A. Huber, David Mark Welch, David A. Relman, and Mitchell L. Sogin. 2008. "Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing." PLoS Genetics 4 (11): e1000255.

Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." BMC Bioinformatics 11 (March): 119.

Ishiwa, Akiko, and Teruya Komano. 2003. "Thin Pilus PilV Adhesins of Plasmid R64 Recognize Specific Structures of the Lipopolysaccharide Molecules of Recipient Cells." Journal of Bacteriology 185 (17): 5192–99.

Iturbe-Ormaetxe, Iñaki, Megan Woolfit, Edwige Rancès, Anne Duplouy, and Scott L. O'Neill. 2011. "A Simple Protocol to Obtain Highly Pure Wolbachia Endosymbiont DNA for Genome Sequencing." Journal of Microbiological Methods 84 (1): 134–36.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. "High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries." Nature Communications. https://doi.org/10.1038/s41467-018-07641-9.

Kantor, Rose S., Kelly C. Wrighton, Kim M. Handley, Itai Sharon, Laura A. Hug, Cindy J. Castelle, Brian C. Thomas, and Jillian F. Banfield. 2013. "Small Genomes and Sparse Metabolisms of Sediment-Associated Bacteria from Four Candidate Phyla." mBio 4 (5): e00708–13.

Kieser, Silas, Joseph Brown, Evgeny M. Zdobnov, Mirko Trajkovski, and Lee Ann McCue. 2019. "ATLAS: A Snakemake Workflow for Assembly, Annotation, and Genomic Binning of Metagenome Sequence Data." bioRxiv. https://doi.org/10.1101/737528.

Kim, Daehwan, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. 2016. "Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences." Genome Research 26 (12): 1721–29.

Klasson, Lisa, Thomas Walker, Mohammed Sebaihia, Mandy J. Sanders, Michael A. Quail, Angela Lord, Susanne Sanders, et al. 2008. "Genome Evolution of Wolbachia Strain wPip from the Culex Pipiens Group." Molecular Biology and Evolution 25 (9): 1877–87.

Koch, Hanna, Maartje A. H. J. van Kessel, and Sebastian Lücker. 2019. "Complete Nitrification: Insights into the Ecophysiology of Comammox Nitrospira." Applied Microbiology and Biotechnology 103 (1): 177–89.

Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. "Genomic Insights That Advance the Species Definition for Prokaryotes." Proceedings of the National Academy of Sciences of the United States of America 102 (7): 2567–72.

Köster, Johannes, and Sven Rahmann. 2012. "Snakemake--a Scalable Bioinformatics Workflow Engine." Bioinformatics 28 (19): 2520–22.

Koutsovoulos, Georgios, Sujai Kumar, Dominik R. Laetsch, Lewis Stevens, Jennifer Daub, Claire Conlon, Habib Maroon, Fran Thomas, Aziz A. Aboobaker, and Mark Blaxter. 2016. “No Evidence for Extensive Horizontal Gene Transfer in the Genome of the Tardigrade Hypsibius Dujardini.” Proceedings of the National Academy of Sciences of the United States of America.

Kowarsky, Mark, Joan Camunas-Soler, Michael Kertesz, Iwijn De Vlaminck, Winston Koh, Wenying Pan, Lance Martin, et al. 2017. “Numerous Uncharacterized and Highly Divergent Microbes Which Colonize Humans Are Revealed by Circulating Cell-Free DNA.” Proceedings of the National Academy of Sciences of the United States of America 114 (36): 9623–28.

Kyrpides, Nikos C., Emiley A. Eloe-Fadrosh, and Natalia N. Ivanova. 2016. “Microbiome Data Science: Understanding Our Microbial Planet.” Trends in Microbiology 24 (6): 425–27.

Lamont, Richard J., Hyun Koo, and George Hajishengallis. 2018. “The Oral Microbiota: Dynamic Communities and Host Interactions.” Nature Reviews. Microbiology 16 (12): 745–59.

Lane, Nick. 2015. “The Unseen World: Reflections on Leeuwenhoek (1677) ‘Concerning Little Animals.’” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 370 (1666): 20140344.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” Nature Methods 9 (4): 357–59.

LePage, Daniel, and Seth R. Bordenstein. 2013. “Wolbachia: Can We Save Lives with a Great Pandemic?” Trends in Parasitology 29 (8): 385–93.

Leviatan, Sigal, and Eran Segal. 2019. “A Significant Expansion of Our Understanding of the Composition of the Human Microbiome.” mSystems 4 (1). https://doi.org/10.1128/mSystems.00010-19.

Libis, Vincent, Niv Antonovsky, Mengyin Zhang, Zhuo Shang, Daniel Montiel, Jeffrey Maniko, Melinda A. Ternei, et al. 2019. “Uncovering the Biosynthetic Potential of Rare Metagenomic DNA Using Co-Occurrence Network Analysis of Targeted Sequences.” Nature Communications 10 (1): 3848.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. “MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” Bioinformatics 31 (10): 1674–76.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” Bioinformatics 34 (18): 3094–3100.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” Bioinformatics 25 (16): 2078–79.

“List of Bioinformatics Software - omicX.” n.d. Accessed January 4, 2020. https://omictools.com/software.

Lloyd-Price, Jason, Anup Mahurkar, Gholamali Rahnavard, Jonathan Crabtree, Joshua Orvis, A. Brantley Hall, Arthur Brady, et al. 2017. “Strains, Functions and Dynamics in the Expanded Human Microbiome Project.” Nature 550 (7674): 61–66.

Mager, Donna L., Laurie Ann Ximenez-Fyvie, Anne D. Haffajee, and Sigmund S. Socransky. 2003. “Distribution of Selected Bacterial Species on Intraoral Surfaces.” Journal of Clinical Periodontology 30 (7): 644–54.

Mains, James W., Corey L. Brelsfoard, Robert I. Rose, and Stephen L. Dobson. 2016. “Female Adult

Aedes Albopictus Suppression by Wolbachia-Infected Male Mosquitoes." Scientific Reports. https://doi.org/10.1038/srep33846.

Marcy, Yann, Cleber Ouverney, Elisabeth M. Bik, Tina Lösekann, Natalia Ivanova, Hector Garcia Martin, Ernest Szeto, et al. 2007. "Dissecting Biological 'Dark Matter' with Single-Cell Genetic Analysis of Rare and Uncultivated TM7 Microbes from the Human Mouth." Proceedings of the National Academy of Sciences of the United States of America 104 (29): 11889–94.

Mark Welch, Jessica L., Floyd E. Dewhirst, and Gary G. Borisy. 2019. "Biogeography of the Oral Microbiome: The Site-Specialist Hypothesis." Annual Review of Microbiology 73 (September): 335–58.

Mark Welch, Jessica L., Blair J. Rossetti, Christopher W. Rieken, Floyd E. Dewhirst, and Gary G. Borisy. 2016. "Biogeography of a Human Oral Microbiome at the Micron Scale." Proceedings of the National Academy of Sciences of the United States of America 113 (6): E791–800.

Mark Welch, Jessica L., Daniel R. Utter, Blair J. Rossetti, David B. Mark Welch, A. Murat Eren, and Gary G. Borisy. 2014. "Dynamics of Tongue Microbial Communities with Single-Nucleotide Resolution Using Oligotyping." Frontiers in Microbiology 5 (November): 568.

Martinez-Guryn, Kristina, Vanessa Leone, and Eugene B. Chang. 2019. "Regional Diversity of the Gastrointestinal Microbiome." Cell Host & Microbe 26 (3): 314–24.

McLean, Jeffrey S., Batbileg Bor, Thao T. To, Quanhui Liu, Kristopher A. Kearns, Lindsey M. Solden, Kelly C. Wrighton, Xuesong He, and Wenyuan Shi. 2018. "Evidence of Independent Acquisition and Adaption of Ultra-Small Bacteria to Human Hosts across the Highly Diverse yet Reduced Genomes of the Phylum Saccharibacteria." https://doi.org/10.1101/258137.

McLean, Jeffrey S., Quanhui Liu, John Thompson, Anna Edlund, and Scott Kelley. 2015. "Draft Genome Sequence of ' Candidatus Bacteroides Periocalifornicus,' a New Member of the Bacteriodetes Phylum Found within the Oral Microbiome of Periodontitis Patients." Genome Announcements. https://doi.org/10.1128/genomea.01485-15.

McLean, J. S., B. Bor, T. T. To, Q. Liu, K. A. Kearns, and L. M. Solden. 2018. "Evidence of Independent Acquisition and Adaption of Ultra-Small Bacteria to Human Hosts across the Highly Diverse yet Reduced Genomes of the Phylum …." bioRxiv. https://www.biorxiv.org/content/early/2018/02/02/258137.abstract.

Méheust, Raphaël, David Burstein, Cindy J. Castelle, and Jillian F. Banfield. 2019. "The Distinction of CPR Bacteria from Other Bacteria Based on Protein Family Content." Nature Communications 10 (1): 4173.

Minoche, André E., Juliane C. Dohm, and Heinz Himmelbauer. 2011. "Evaluation of Genomic High-Throughput Sequencing Data Generated on Illumina HiSeq and Genome Analyzer Systems." Genome Biology 12 (11): R112.

Moutsopoulos, Niki M., and Joanne E. Konkel. 2018. "Tissue-Specific Immunity at the Oral Mucosal Barrier." Trends in Immunology 39 (4): 276–87.

Murugkar, Pallavi P., Andrew J. Collins, and Floyd E. Dewhirst. 2019. "Complete Genome Sequence of Strain PM004, a Novel Cultured Member of the Human Oral Microbiome from the Candidate Phylum Saccharibacteria (TM7)." Microbiology Resource Announcements 8 (42). https://doi.org/10.1128/MRA.01159-19.

Nayfach, Stephen, Beltran Rodriguez-Mueller, Nandita Garud, and Katherine S. Pollard. 2016. "An

Integrated Metagenomics Pipeline for Strain Profiling Reveals Novel Patterns of Bacterial Transmission and Biogeography." Genome Research 26 (11): 1612–25.

Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, and Nikos C. Kyrpides. 2019. "New Insights from Uncultivated Genomes of the Global Human Gut Microbiome." Nature 568 (7753): 505–10.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." Molecular Biology and Evolution 32 (1): 268–74.

O'Neill, Scott L., Peter A. Ryan, Andrew P. Turley, Geoff Wilson, Kate Retzki, Inaki Iturbe-Ormaetxe, Yi Dong, et al. 2018. "Scaled Deployment of Wolbachia to Protect the Community from Aedes Transmitted Arboviruses." Gates Open Research. https://doi.org/10.12688/gatesopenres.12844.1.

Paez-Espino, David, Emiley A. Eloe-Fadrosh, Georgios A. Pavlopoulos, Alex D. Thomas, Marcel Huntemann, Natalia Mikhailova, Edward Rubin, Natalia N. Ivanova, and Nikos C. Kyrpides. 2016. "Uncovering Earth's Virome." Nature 536 (7617): 425–30.

Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. "Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life." Nature Microbiology 2 (11): 1533–42.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. "Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle." Cell 176 (3): 649–62.e20.

Planavsky, Noah J., Christopher T. Reinhard, Xiangli Wang, Danielle Thomson, Peter McGoldrick, Robert H. Rainbird, Thomas Johnson, Woodward W. Fischer, and Timothy W. Lyons. 2014. "Earth History. Low Mid-Proterozoic Atmospheric Oxygen Levels and the Delayed Rise of Animals." Science 346 (6209): 635–38.

Probst, Alexander J., Bethany Ladd, Jessica K. Jarett, David E. Geller-McGrath, Christian M. K. Sieber, Joanne B. Emerson, Karthik Anantharaman, et al. 2018. "Differential Depth Distribution of Microbial Function and Putative Symbionts through Sediment-Hosted Aquifers in the Deep Terrestrial Subsurface." Nature Microbiology 3 (3): 328–36.

Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. "Shotgun Metagenomics, from Sampling to Analysis." Nature Biotechnology 35 (9): 833–44.

Raveh-Sadka, Tali, Brian C. Thomas, Andrea Singh, Brian Firek, Brandon Brooks, Cindy J. Castelle, Itai Sharon, et al. 2015. "Gut Bacteria Are Rarely Shared by Co-Hospitalized Premature Infants, regardless of Necrotizing Enterocolitis Development." eLife 4 (March). https://doi.org/10.7554/eLife.05477.

Reinhold-Hurek, Barbara, Wiebke Bünger, Claudia Sofía Burbano, Mugdha Sabale, and Thomas Hurek. 2015. "Roots Shaping Their Microbiome: Global Hotspots for Microbial Activity." Annual Review of Phytopathology 53: 403–24.

Reveillaud, Julie, Sarah R. Bordenstein, Corinne Cruaud, Alon Shaiber, Özcan C. Esen, Mylène Weill, Patrick Makoundou, et al. 2019. "The Wolbachia Mobilome in Culex Pipiens Includes a Putative Plasmid." Nature Communications 10 (1): 1051.

Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, Jan-Fang

Cheng, Aaron Darling, et al. 2013. “Insights into the Phylogeny and Coding Potential of Microbial Dark Matter.” Nature 499 (7459): 431–37.

Roux, Simon, Francois Enault, Bonnie L. Hurwitz, and Matthew B. Sullivan. 2015. “VirSorter: Mining Viral Signal from Microbial Genomic Data.” PeerJ 3 (May): e985.

Roux, Simon, Mart Krupovic, Rebecca A. Daly, Adair L. Borges, Stephen Nayfach, Frederik Schulz, Allison Sharrar, et al. 2019. “Cryptic Inoviruses Revealed as Pervasive in Bacteria and Archaea across Earth’s Biomes.” Nature Microbiology, July. https://doi.org/10.1038/s41564-019-0510-x.

Schmidt, Thomas Sb, Matthew R. Hayward, Luis P. Coelho, Simone S. Li, Paul I. Costea, Anita Y. Voigt, Jakob Wirbel, et al. 2019. “Extensive Transmission of Microbes along the Gastrointestinal Tract.” eLife 8 (February). https://doi.org/10.7554/eLife.42693.

Seabold, Skipper, and Josef Perktold. 2010. “Statsmodels: Econometric and Statistical Modeling with Python.” In Proceedings of the 9th Python in Science Conference, 57:61. Scipy.

Segata, Nicola, Susan Kinder Haake, Peter Mannon, Katherine P. Lemon, Levi Waldron, Dirk Gevers, Curtis Huttenhower, and Jacques Izard. 2012. “Composition of the Adult Digestive Tract Bacterial Microbiome Based on Seven Mouth Surfaces, Tonsils, Throat and Stool Samples.” Genome Biology 13 (6): R42.

Sender, Ron, Shai Fuchs, and Ron Milo. 2016. “Revised Estimates for the Number of Human and Bacteria Cells in the Body.” PLoS Biology 14 (8): e1002533.

Shaiber, Alon, and A. Murat Eren. 2019. “Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories.” mBio. https://doi.org/10.1128/mBio.00725-19.

Song, Se Jin, Christian Lauber, Elizabeth K. Costello, Catherine A. Lozupone, Gregory Humphrey, Donna Berg-Lyons, J. Gregory Caporaso, et al. 2013. “Cohabiting Family Members Share Microbiota with One Another and with Their Dogs.” eLife 2 (April): e00458.

Spang, Anja, Jimmy H. Saw, Steffen L. Jørgensen, Katarzyna Zaremba-Niedzwiedzka, Joran Martijn, Anders E. Lind, Roel van Eijk, Christa Schleper, Lionel Guy, and Thijs J. G. Ettema. 2015. “Complex Archaea That Bridge the Gap between Prokaryotes and Eukaryotes.” Nature 521 (7551): 173–79.

Stern, Adi, Leeat Keren, Omri Wurtzel, Gil Amitai, and Rotem Sorek. 2010. “Self-Targeting by CRISPR: Gene Regulation or Autoimmunity?” Trends in Genetics: TIG 26 (8): 335–40.

Storey, John D., Jonathan E. Taylor, and David Siegmund. 2004. “Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach.” Journal of the Royal Statistical Society: Series B (Statistical Methodology). https://doi.org/10.1111/j.1467-9868.2004.00439.x.

Stouthamer, R., J. A. Breeuwer, and G. D. Hurst. 1999. “Wolbachia Pipientis: Microbial Manipulator of Arthropod Reproduction.” Annual Review of Microbiology 53: 71–102.

Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin. 2000. “The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution.” Nucleic Acids Research 28 (1): 33–36.

Taylor, Mark J., Seth R. Bordenstein, and Barton Slatko. 2018. “Microbe Profile: Wolbachia: A Sex Selector, a Viral Protector and a Target to Treat Filarial Nematodes.” Microbiology 164 (11): 1345–47.

Torres, Pedro J., John Thompson, McLean, Scott T. Kelley, and Anna Edlund. 2019. “Discovery of a

Novel Periodontal Disease-Associated Bacterium." Microbial Ecology 77 (1): 267–76.

Torsvik, Vigdis, Lise Øvreås, and Tron Frede Thingstad. 2002. "Prokaryotic Diversity--Magnitude, Dynamics, and Controlling Factors." Science 296 (5570): 1064–66.

Touchon, Marie, Aude Bernheim, and Eduardo Pc Rocha. 2016. "Genetic and Life-History Traits Associated with the Distribution of Prophages in Bacteria." The ISME Journal 10 (11): 2744–54.

Uritskiy, Gherman V., Jocelyne DiRuggiero, and James Taylor. 2018. "MetaWRAP—a Flexible Pipeline for Genome-Resolved Metagenomic Data Analysis." Microbiome 6 (1): 158.

Vartoukian, Sonia R., Aleksandra Adamowska, Megan Lawlor, Rebecca Moazzez, Floyd E. Dewhirst, and William G. Wade. 2016. "In Vitro Cultivation of 'Unculturable' Oral Bacteria, Facilitated by Community Culture and Media Supplementation with Siderophores." PloS One 11 (1): e0146926.

Vorholt, Julia A. 2012. "Microbial Life in the Phyllosphere." Nature Reviews. Microbiology 10 (12): 828–40.

Wade, William G. 2013. "The Oral Microbiome in Health and Disease." Pharmacological Research: The Official Journal of the Italian Pharmacological Society 69 (1): 137–43.

Welch, Jessica L. Mark, Floyd E. Dewhirst, and Gary G. Borisy. 2019. "Biogeography of the Oral Microbiome: The Site-Specialist Hypothesis." Annual Review of Microbiology 73. https://www.annualreviews.org/doi/abs/10.1146/annurev-micro-090817-062503.

Whelan, S., and N. Goldman. 2001. "A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach." Molecular Biology and Evolution 18 (5): 691–99.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. "Prokaryotes: The Unseen Majority." Proceedings of the National Academy of Sciences of the United States of America 95 (12): 6578–83.

Wickham, Hadley. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer.

Yeoman, Carl J., Laura M. Brutscher, Özcan C. Esen, Furkan Ibaoglu, Curtis Fowler, A. Murat Eren, Kevin Wanner, and David K. Weaver. 2019. "Genome-Resolved Insights into a Novel Spiroplasma Symbiont of the Wheat Stem Sawfly (Cephus Cinctus)." PeerJ 7 (August): e7548.

Zilionis, Rapolas, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M. Klein, and Linas Mazutis. 2017. "Single-Cell Barcoding and Sequencing Using Droplet Microfluidics." Nature Protocols 12 (1): 44–73.