Illuminating the oral microbiome and its host interactions: recent advancements in omics and bioinformatics technologies in the context of oral microbiome research

Jonathon L. Baker[1,2]

[1] Genomic Medicine Group
J. Craig Venter Institute
4120 Capricorn Lane
La Jolla, CA 92037

[2] Department of Pediatrics
UC San Diego School of Medicine
La Jolla, CA

*Corresponding Author: JLB: jobaker@jcvi.org

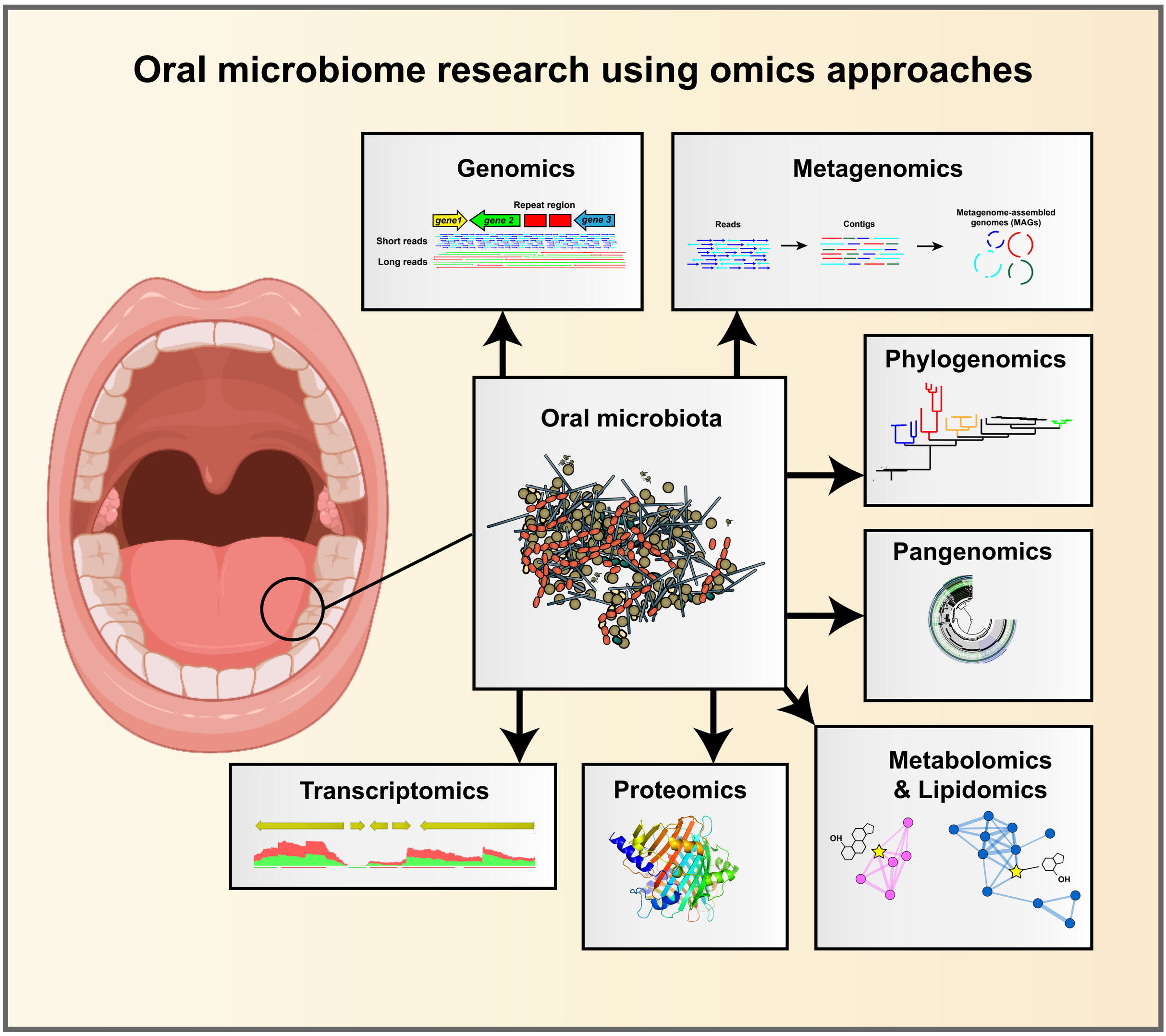
ORCID: JLB: 0000-0001-5378-322X

Keywords: oral microbiome, genomics, metagenomics, pangenomics, transcriptomics, proteomics, metabolomics, lipidomics

Abstract

The oral microbiota has an enormous impact on human health, with oral dysbiosis now linked to many oral and systemic diseases. Recent advancements in sequencing, mass spectrometry, bioinformatics, computational biology, and machine learning are revolutionizing the oral microbiome research, enabling analysis at an unprecedented scale and level of resolution using omics approaches. This review contains a comprehensive perspective of the current state-of-the-art tools available to perform genomics, metagenomics, phylogenomics, pangenomics, transcriptomics, proteomics, metabolomics, lipidomics, and multi-omics analysis on (all) microbiomes, and then provides examples of how the techniques have been applied to research of the oral microbiome, specifically. Key findings of these studies and remaining challenges for the field are highlighted. Although the methods discussed here are placed in the context of their contributions to oral microbiome research specifically, they are pertinent to the study of any microbiome, and the intended audience of this includes researchers would simply like to get an introduction to microbial omics and/or an update on the latest omics methods. Continued research of the oral microbiota using omics approaches is crucial and will lead to dramatic improvements in human health, longevity, and quality of life.

Graphical Abstract



Main Text

INTRODUCTION

The oral microbiota is a unique and diverse community of bacteria, viruses, fungi, and archaea that plays a major role in human health (Baker *et al.*, 2017). Distinct microenvironments within the oral cavity, such as the hard surface of the tooth, keratinized hard palate, or soft surface of the tongue result in the establishment of unique and highly structured communities at each site (Human Microbiome Project, 2012, Lamont *et al.*, 2018). The health-associated oral microbiota exhibits colonization resistance and plays an active role in preventing dysbiosis and associated disease (He *et al.*, 2014, Radaic & Kapila, 2021). Meanwhile, dysbiosis of the oral microbiome, even on a highly localized scale, is responsible for dental caries and periodontal disease, both extremely prevalent and costly (Bowen *et al.*, 2018). Furthermore, the majority of oral cancers are driven by oral infection with viruses such as human papilloma virus (HPV) and Epstein-Barr virus (EBV, formerly known as human gammaherpesvirus 4/HHV-4) (Tsao *et al.*, 2017, Economopoulou *et al.*, 2020). In addition to oral diseases, there are increasing lines of evidence linking the oral microbiota to a myriad of extra-oral and systemic diseases, such obesity, diabetes, cardiovascular disease, inflammatory bowel disease, nonalcoholic fatty liver disease, rheumatoid arthritis, colorectal cancers, and Alzheimer's Disease (Hajishengallis & Chavakis, 2021). The oral microbiome has also served as an important model system for researching microbiomes broadly, as diverse taxa across all kingdoms of life co-exist and interact at a site which is easily accessible to observe the processes of biofilm and community assembly and succession (Baker *et al.*, 2017). Despite significant progress in our understanding of the human oral microbiota, continued research is essential, and will lead to improvements in human health and overall quality of life.

Prior to the development of culture-independent analysis methods, such as untargeted (i.e., "shotgun") sequencing and mass spectrometry, study of the oral microbiome, and its role in human health, was limited to taxa that could be isolated and cultivated in the laboratory. Using these classic microbiological techniques, key members of the community including both

pathogens (e.g., *Streptococcus mutans* and *Porphyromonas gingivalis*) and commensals (e.g., *Streptococcus gordonii* and *Streptococcus sanguinis*) were discovered, became well-studied, and mechanisms of caries and periodontal disease pathogenesis were elucidated. However, the overall picture of the oral microbiota (and indeed all microbiomes), and its role in human health, was still relatively incomplete and had a very narrow focus.

Over the past 20 years, culture-independent analysis methods have enabled the formation and subsequent explosive growth of microbiome research, including that of the human oral microbiome. The development of these methods was due to major advancements in sequencing technology, mass spectrometry, bioinformatics, computational biology, and computer science/machine learning. In concert with the development of microbiome research has been the development of the omics fields of study. Especially pertinent to microbiome research are genomics, metagenomics, phylogenomics, pangenomics, and transcriptomics, which are based on nucleic acid sequencing, as well as metabolomics, proteomics, and lipidomics, which are based on mass spectrometry. Traditional omics analyzes populations of cells within samples in aggregate, getting an average for the population, which may not reflect the true profiles of a given analyte across individual cells in the population. Single cell analysis techniques are rapidly addressing this issue, already becoming a mainstay in eukaryotic transcriptomics. Single-cell analysis is much more challenging in bacteria, as cells, and therefore the amount of input material, are orders of magnitude smaller. However, the first single-cell analyses of bacteria have been described in the past several years. Meanwhile, multi-omics research, examining datasets from two or more omics fields, presents great potential for new discovery, but also additional challenges. The continued evolution of these fields of research has enabled study of the oral microbiome at an unprecedented scale and level of resolution. This review will provide an overview of these omics disciplines and explain some of the most used and state-of-the-art technologies and techniques. The review will then discuss how these approaches have been applied to the study of the oral microbiome, highlighting some of the major recent discoveries that

have been facilitated. Since several recent reviews have excellently summarized the results of the use of omics techniques in both dental caries (Moussa *et al.*, 2022) and periodontal disease (Nguyen *et al.*, 2020) research, this review will have focus more on the omics techniques and tools themselves, including historical context and current state of the technology. While it is not possible to include all of the technologies, tools, and research worthy of inclusion, this review provides the reader with reference to further comprehensive reviews on more specific topics where possible.

SEQUENCING-BASED OMICS

Historical background: Next Generation Sequencing (NGS) revolutionizes the life sciences and enables early microbiome research in the 2000s and early 2010s

"Next Generation Sequencing" (NGS) methods, including sequencing-by-synthesis (Illumina), pyrosequencing (454 Life Sciences), and sequencing by oligonucleotide ligation and detection (SOLiD; Applied Biosystems), revolutionized the life sciences in the 2000s and early 2010s by enabling accurate, high throughput, untargeted sequencing (Bennett, 2004, Margulies *et al.*, 2005, Bentley *et al.*, 2008, McKernan *et al.*, 2009). For the first time, microbiological samples could be analyzed for all microbial DNA or RNA content, regardless of the cultivability of the taxa present (Venter *et al.*, 2004, Ley *et al.*, 2005, Gill *et al.*, 2006). This led to the establishment of microbiome research as a scientific field, and the subsequent explosion of microbiome studies, including large, concerted efforts such as the Human Microbiome Project (Human Microbiome Project, 2012). The vast majority of this early microbiome research was conducted using amplicon sequencing-based analysis methods, largely of the 16S rRNA gene (termed "16S sequencing" or "16S analysis"). This was because 16S analysis allows many more samples to be analyzed with a sufficient depth to acquire microbiome data on a sequencing run, compared to metagenomics sequencing. As a result, 16S sequencing is higher throughput and significantly cheaper on a per-sample basis. It is important to note that advancements during this period were

not limited to sequencing instrumentation, and that there were also major developments in mass spectrometry, computer science, and computational biology that were foundational to many of the modern technologies discussed in this review. Among these were the algorithms and suites of analysis tools which were the first versions and/or predecessors of some of the tools still most widely used in microbiome studies today, including the precursors to the DADA2 (Callahan *et al.*, 2016), QIIME2 (Bolyen *et al.*, 2019), Kraken (Lu *et al.*, 2022), bioBakery (Beghini *et al.*, 2021), SEQUEST (Brodbelt & Russell, 2015), and SPAdes (Prjibelski *et al.*, 2020) suites of software.

A notable advance in the study of the oral microbiome, specifically, during this period was the development of the Human Oral Microbiome Database (HOMD), first published in 2010 (Chen *et al.*, 2010). This not only provided a free, public, large-scale database of 16S rRNA sequences specific to microbes from the human oral cavity, but also began to illustrate how limited previous understanding of the oral microbiota had been, highlighting that 53% of the 619 species-level taxa identified in the project had not been properly named, and 35% had never been isolated or cultivated (Chen *et al.*, 2010). The Human Microbiome Project also significantly advanced our understanding of the inhabitants of the oral microbiome at specific niches (Human Microbiome Project, 2012). Figure 1A is a timeline illustrating many of the major milestones in omics, microbiome, and oral microbiome research over the last several decades.

Current developments sequencing technologies

In the present day, new advancements in sequencing technology are in the process of revolutionizing microbiome research once more. Throughout the 2010s, Illumina emerged as the dominant player in sequencing, holding about 80% of the market share as of 2020, with improvements to their sequencing-by-synthesis technology increasing throughput dramatically while greatly reducing the cost of sequencing. This decrease in sequencing cost has even eclipsed Moore's Law (which posited that the number of transistors on an integrated circuit doubles about every two years, therefore dropping the cost of computer power to the consumer

in a log-linear manner), with the cost of sequencing one million base pairs falling from $10 million in 2001 to $0.10 by 2016 (Wetterstrand, Muir *et al.*, 2016). Interestingly, this phenomenon has led some scientists to hypothesize that computing power and storage will ultimately become the limiting cost factors in sequencing-based research, rather than the sequencing itself (Muir *et al.*, 2016). This dramatic reduction in sequencing cost has enabled many more oral health and microbiome researchers to perform larger-scale 16S sequencing projects, metagenomics, whole genome sequencing, and RNA-seq.

At the same time, emerging third generation sequencing technologies, especially long-read technologies such as nanopore sequencing (Oxford Nanopore [ONT]) (Jain *et al.*, 2015), single molecule real time sequencing (SMRT; Pacific Biosciences [PacBio]) (Roberts *et al.*, 2013), and LoopSeq (Element Biosciences) (Callahan *et al.*, 2021) are in the process of transforming the landscape of sequencing yet again and are challenging Illumina's preeminence. Although Illumina sequencing is highly accurate, the reads produced typically only 150 or 300 bp in length. With read lengths this short, repeat and nonspecific regions significantly hamper efforts to assemble complete genomes, with Illumina-based genome (or metagenome) assemblies typically being split into fragments, which are called contigs (Athanasopoulou *et al.*, 2021). Using ONT sequencing, the length of reads produced is theoretically only limited by the length of the input material, and single reads of over 1 mbp are now routinely reported (Jain *et al.*, 2015). These long reads span the entirety of repeat regions, enabling assembly of circular chromosomes and complete genomes with much greater ease. RNA can also be sequenced using ONT, where sequencing of the full-length transcripts easily provides transcriptome-wide information on co-transcribed genes and identification of novel RNA isoforms (Garalde *et al.*, 2018). In addition ONT sequencing can sequence native molecules, reducing bias by sidestepping the PCR and/or cDNA synthesis steps that are required in many sequencing library preparation protocols (Garalde *et al.*, 2018). Crucially, sequencing native molecules also enables the detection of base modifications and noncanonical bases (e.g., methylated bases, inosine, pseudouridine, etc.),

allowing these phenomena to be studied on a genome, metagenome, transciptome, or metatranscriptome scale for the first time (Garalde *et al.*, 2018). These epigenetic modifications have been particularly understudied in the context of microbiology. The most substantial drawback to ONT sequencing is a relatively low accuracy. Errors in ONT sequencing are not random, but usually occur during homopolymeric tracts, where the basecalling software has difficulty identifying how many consecutive iterations of a given base or bases have passed through the nanopore, as the rate of processivity through the channel is saltatory, not constant (Amarasinghe *et al.*, 2020). This leads to insertions or deletions, which are nontrivial as they are likely to cause apparent frameshifts and therefore impact downstream gene calling and annotation (Watson & Warr, 2019). As a result, ONT sequencing data is frequently combined with Illumina sequencing data of the same sample, where the long reads enable accurate large-scale assembly of contigs and scaffolds, and the short reads are used to polish out the errors inherent to the ONT reads (Koren *et al.*, 2012). Crucially, the accuracy of ONT sequencing has rapidly improved in recent years, falling from 30-40% in 2015 to <0.1% in raw-reads (or <0.001% in a consensus assembly with >=20X coverage) using current instrumentation and software (Sereika *et al.*, 2022). As a result of these recent, massive improvements in accuracy, several recent studies have shown that the field is close an inflection point where accurate genomics and metagenomics can be performed using ONT sequencing alone (Faulk, 2022, Liu *et al.*, 2022, Sereika *et al.*, 2022).

SMRT sequencing technology from PacBio represents a “middle ground” between Illumina and ONT sequencing technologies, combining relatively long reads, averaging 10-25 kb, and an error rate of <0.1% for raw reads and <0.003% for 25-30X consensus assemblies (Wenger *et al.*, 2019). While SMRT sequencing was able to produce accurate genomes and metagenomes independently of short-read polishing much earlier than nanopore sequencing, the significantly higher cost per base of PacBio sequencing, and the much higher cost of the PacBio sequencing machines themselves, have remained a barrier for many researchers (Sereika *et al.*, 2022). Like ONT sequencing, PacBio sequencing can also sequence full-length RNAs (Leung *et al.*, 2021)

198 and can detect methylation, enabling genome-wide epigenetic studies (Beaulaurier *et al.*, 2018).
199 In addition to Oxford Nanopore and PacBio, newcomers in the sequencing space, such as
200 Element Biosciences (developing both innovations to short read sequencing and long read
201 LoopSeq) and Stratos Genomics (now owned by Roche, developing sequencing-by-expansion
202 technologies) may indeed further disrupt the industry. In addition, synthetic long read and linked-
203 read approaches, such as TELL-seq, use labeling of short reads adjacent on the genome to obtain
204 contiguity of short-read based assemblies similar to those obtained through long read-based
205 approaches (Wang *et al.*, 2019). However, limitations, including incompatibility with metagenomic
206 assemblies, continue to limit widespread use of these approaches (Wang *et al.*, 2019).
207
208 **Genomics**
209 Figure 1B provides a list of bioinformatics tools, and their references, that are discussed in the
210 following sections. Obtaining genomes that are both complete (i.e., contiguous chromosomes
211 and plasmids) and accurate is of prime importance to microbiology research. High-quality,
212 complete genomes (assuming they are publicly available to researchers in a database) enable:
213 1) accurate detection and quantification of a particular taxon, or its RNA transcripts, in an isolate
214 or microbiome sample (Venter *et al.*, 2004), 2) prediction of the metabolic pathways, and therefore
215 possible ecological and pathogenic roles of the taxon—particularly important for taxa that have
216 not yet been isolated or cultivated (Naito *et al.*, 2016), and 3) guiding wet-lab research, such as
217 mutagenesis. It is important to recognize that many genomes in public repositories were
218 assembled using short read sequencing only, meaning that they are probably at an incomplete or
219 draft stage, and fragmented into contigs of various numbers and sizes. These genomes are likely
220 to be missing sequences, and may contain contaminant contigs. Therefore, it is crucial that
221 researchers are cognizant of the limitations inherent with these assemblies if they are used as a
222 reference.

To obtain a genome, sequencing reads that have passed quality control must be assembled. Note that assembly of multispecies (i.e., microbiome) samples is discussed in the following section on Metagenomics. A range of assembly tools and algorithms are available to assemble microbial genomes. For Illumina short reads, these include ABySS (Simpson *et al.*, 2009), Velvet (Zerbino & Birney, 2008), MEGAHIT (Li *et al.*, 2015), and SPAdes (Prjibelski *et al.*, 2020). SPAdes tends to give the highest quality assemblies, but is more computationally expensive and time-consuming than its competitors (van der Walt *et al.*, 2017). The significantly longer read length and higher error rate of ONT and PacBio sequencing datasets necessitates different assembly algorithms. Long read assemblers include Canu (Koren *et al.*, 2017), HGAP (Chin *et al.*, 2013), miniasm (Li, 2016), MaSuRCA (Zimin *et al.*, 2013), and Flye (Kolmogorov *et al.*, 2019). The innovative, repeat graph approach employed by Flye performs well relative to its competitors and is rapidly becoming a tool of choice for the field (Kolmogorov *et al.*, 2019). As mentioned above, long-read-only assemblies (particularly from ONT) have traditionally had higher error rates, and benefit from a complementary Illumina dataset (although the latest ONT technology can produce accurate assemblies of microbial taxa on its own, as mentioned above). For datasets where both long read and short read sequencing data is available, Unicycler (Wick *et al.*, 2017), Trycycler (Wick *et al.*, 2021), and hybridSPAdes (Antipov *et al.*, 2016) are available hybrid assembly tools, however these were all developed for isolate (i.e. not metagenomic) sequencing. Draft assemblies can also be polished to further remove errors using long reads via tools including nanopolish (Loman *et al.*, 2015) and medaka (https://github.com/nanoporetech/medaka), and/or with short reads via tools including racon (Vaser *et al.*, 2017), pilon (Walker *et al.*, 2014), and polypolish (Wick & Holt, 2022). Polypolish was a particularly helpful advance, greatly improving short read-based polishing in repeat and highly conserved regions, such as the rRNA genes (Baker, 2022). The combination of these bioinformatics tools with third generation long read sequencing technologies has made it relatively easy and inexpensive to obtain accurate and complete genomes, enabling researchers to monitor

reference strains for mutations, and study genome-wide evolution, physiology, and pathogenesis in novel clinical and environmental isolates.

Metagenomics

Metagenomics is the study of DNA recovered directly from environmental or clinical samples, thereby containing multiple taxa (i.e. multiple genomes), which of course includes microbiome analysis. In-depth recommendations for the design and execution of a microbiome study have been expertly provided (Knight *et al.*, 2018). Metagenomics data can be analyzed do get diversity metrics and abundance information on the taxa present. This can be done on unassembled reads using tools like MetaPhlAn4 (Blanco-Miguez, 2022, biorxiv) (based on marker genes and part of the bioBakery suite of tools (Beghini *et al.*, 2021)) and the Kraken family of tools (based on k-mers (Lu *et al.*, 2022)). In addition to taxonomic abundance information, tools such as HumanN3 (also a BioBakery tool) (Beghini *et al.*, 2021) can obtain information regarding the metabolic pathways present in a microbiome sample, enabling analysis such as contributional diversity. This provides a significant advantage over 16S sequencing, where the functional metagenomics are not directly examined and may only be inferred linking a 16S sequence to a reference genome in a database (using a tool such as PICRUSt2 (Douglas *et al.*, 2020)). A species may have one 16S rRNA sequence, but a significant amount of strain-to-strain intraspecies functional diversity, which will be missed in any 16S sequencing analysis. A disadvantage to most methods analyzing unassembled metagenomic reads is dependency on databases, where novel taxa or functions are likely to end up in an “unknown” bucket which is routinely discarded by investigators (although this issue continues to decrease substantially with each subsequent version of the tools and databases).

Beyond data generated by the unassembled reads, metagenomic datasets can be assembled to produce metagenome-assembled genomes (MAGs). A recent review covers these principles and methods in greater depth (Goussarov *et al.*, 2022). Most of the aforementioned

assembly algorithms now have versions specifically designed to handle metagenomic read sets, with metaSPAdes (Nurk *et al.*, 2017) and MEGAHIT (Li *et al.*, 2015) being the most commonly employed for short reads and metaFlye (Kolmogorov *et al.*, 2020) and strainFlye (Fedarko *et al.*, 2022) becoming the standard for long reads. Following assembly, a problem inherent with metagenomic datasets is not knowing which assembled contigs go together to form a given genome. Binning is the process of solving this problem, placing metagenomic contigs into discrete draft genomes or “bins”, and binning typically utilizes data like k-mer frequency, GC content, and coverage and/or alignment to references to do so. Many tools are available to perform binning or short read-based assemblies, and several of the most mainstream include MaxBin2 (Wu *et al.*, 2016), Concoct (Alneberg *et al.*, 2014), and MetaBat2 (Kang *et al.*, 2019). Recently, strategies for binning that leverage the methylation data provided by third generation sequencing methods have been reported (Wilbanks *et al.*, 2022). Different binning algorithms appear to produce better bins in different datasets, and indeed tools combining composite and/or iterative binning strategies are available including DAStool (Sieber *et al.*, 2018), MetaWRAP (Uritskiy *et al.*, 2018), and VEBA (Espinoza & Dupont, 2022). Manual bin inspection and refinement should be performed, where possible, and has been made much easier by the Anvi’o suite of microbiome analysis programs (Chen *et al.*, 2020, Eren *et al.*, 2021). There are far fewer tools to perform binning on long read datasets, with LRBinner being the most comprehensive and recently developed (Wickramarachchi & Lin, 2022). However, contigs in long read assemblies are so much longer, and draft genomes so much more contiguous, that manual binning is much more feasible. In fact, circular (and therefore complete) chromosomes are routinely obtained using long read metagenomic sequencing, and of course these do not need to be binned. The ability to obtain complete and accurate genomes from metagenomic samples represents a major advance, and has only become possible in a high throughput fashion following the development of long read sequencing (Moss *et al.*, 2020, Cusco *et al.*, 2021, Sereika *et al.*, 2022).

Phylogenomics

Phylogenomics is the practice of inferring evolutionary history and relatedness between different taxa, and can be done using a number of different strategies. DNA sequences, including whole genome alignment, can be used, and may be useful when studying evolution of gene regulation, or when reconstructing evolutionary relationships over shorter time scales. However, use of amino acid sequences is more widely used, as they are more directly affected by natural selection, less influenced by processes such as gene duplication and horizontal gene transfer, and evolve more slowly, making it easier to reconstruct evolutionary relationships over longer time scales. PhyloSift (Darling *et al.*, 2014), PhyloPhlAn3 (Asnicar *et al.*, 2020), and Anvi'o (Eren *et al.*, 2021) are widely-used pipelines for performing microbial phylogenomics. These pipelines are underpinned by sequence alignment tools, such as muscle (Edgar, 2004), mafft (Nakamura *et al.*, 2018), and famsa (Deorowicz *et al.*, 2016), as well as phylogenetic inference software, such as RAxML (Stamatakis, 2014), FastTree (Price *et al.*, 2009), and IQ-Tree (Nguyen *et al.*, 2015). PhyloPhlAn3 (part of BioBakery3 (Beghini *et al.*, 2021)) can easily provide taxonomic assignment to newly assembled MAGs and can perform phylogenomic analysis scalable from strain level analysis using clade-specific markers, to widely disparate clades such as whole gut microbiome phylogenomic analysis. Because phylogenomics depends, in many cases, on alignment of widely conserved homologous core genes, it inevitably intersects with pangenomics, which is needed to identify these genes. Ideally, the lowest number of genes that still allows accurate differentiation between each taxon in the analysis should be used to reduce the computational expense of the phylogenetic inference software. The most frequent use of phylogenomics in oral microbiome research is determining the species-level taxa of a newly assembled genome or MAG. It is important to note that the concept a "species" in bacteria is not one with universally accepted traits. For the sake of ease when dealing examining massive numbers of MAGs, 95% ANI is the cutoff used to estimate the species-level which has been adopted by the field, however this cutoff is not absolute, and remains controversial (Jain *et al.*, 2018, Murray *et al.*, 2021).

Pangenomics

Pangenomics is analysis of pangenomes, which are the collections of genes across multiple genomes. Pangenomics analysis typically identifies orthologous genes across a set of genomes and provides a list of core genes (genes present in every genome or 90-100% of the genomes in the analysis), cloud genes (found in only a minority of genomes in the analysis), and shell genes (found in many, but not all of the genomes, e.g. less than core genes, but more than cloud genes)., however there are no universally accepted thresholds to determine these groups. Pangenomics is especially useful for tracing horizontal gene transfer and the evolution of specific gene clusters, including pathogenicity islands and antimicrobial resistance genes. Tools, such as Roary (Page *et al.*, 2015), PanPhlAn3 (Beghini *et al.*, 2021), panOCT (Fouts *et al.*, 2012) and Anvi'o (Eren *et al.*, 2021), have allowed pangenomics analysis at an exceptional scale and resolution. A pangenome can be parsed to identify optimal genes for phylogenetic analysis of a given dataset. These would typically be single-copy core genes that also that have maximum sequence differences across orthologs in the pangenome (so that as few genomes are identical or have a flat line in the resulting tree), but also have minimal gaps in the alignment (because phylogenetic analysis tools struggle with where to place gaps in the alignment) (described in detail at anvio.org). This type of approach will yield a bespoke phylogenetic analysis that will maximize phylogenetic data obtained while minimizing computational resources used and time required to perform the analysis.

Transcriptomics

Transcriptomics is the study of gene expression via sequencing of RNA, and may be performed on isolates of a given taxon, or multispecies samples (i.e. a metatranscriptome). For short read-based RNAseq, gene quantification can be performed by either mapping reads to an annotated reference genome (or genomes, in the case of a metatranscriptome), or mapping reads to an

annotated de novo assembly of the transcriptome (useful when reference genomes are lacking). Commonly used mapping tools for short reads include BWA-MEM (Li, 2014), Bowtie2 (Langmead & Salzberg, 2012), and minimap2 (Li, 2018), while minimap2 can also map long reads. Common transcriptome assemblers include Trinity (Grabherr *et al.*, 2011), RockHopper2 (Tjaden, 2015), and rnaSPAdes (Bushmanova *et al.*, 2019). Once mapped, the number of reads mapping to genes and other features can be analyzed using featureCounts (Liao *et al.*, 2014) or a similar tool. As described in the section on Sequencing Technologies, major recent advancements to transcriptomics have come in the form of long-read RNA sequencing, the ability to detect RNA modifications and noncanonical bases, and single cell RNAseq (scRNAseq). At this time, application of these technologies to bacteria remains an area of active development. Current out-of-the-box RNA library preparation protocols for ONT require polyA-tailed RNA as input (eukaryotic mRNA has polyA tails but prokaryotic mRNA does not), therefore polyA tails must be added in addition to the recommended depletion of rRNAs. Several research groups have pioneered using ONT technology for bacterial RNA-seq and their publications provide protocols on how to do so (Pitt *et al.*, 2020, Baker *et al.*, 2022, Grunberger *et al.*, 2022). Tools used to detect DNA and RNA modifications and noncanonical bases in ONT-based transcriptomics include Tombo (Oxford Nanopore Technologies, Inc.), MetaCompore, EpiNano, and MasterofPores, which have been recently benchmarked and reviewed (Wang *et al.*, 2021, White & Hesselberth, 2022).

The impact of sequencing-based omics on oral microbiome research

The sequencing-based omics approaches detailed above have had an extraordinary impact on our understanding of the oral microbiome. Complete genomes of oral taxa are being published at an ever-accelerating rate, making databases such as NCBI and HOMD even more useful to researchers, and allowing for in-depth and accurate downstream phylogenomics and pangenomics. A number of studies have now described the oral microbiome in the context of

dental caries and/or periodontal disease using shotgun metagenomics (Belda-Ferre *et al.*, 2012, Shi *et al.*, 2015, Yost *et al.*, 2015, Belstrom *et al.*, 2017, Al-Hebshi *et al.*, 2019, Baker *et al.*, 2021). Furthermore, several recent studies have released large numbers of oral MAGs into the public domain (Escapa *et al.*, 2018, Pasolli *et al.*, 2019, Baker *et al.*, 2021, Zhu *et al.*, 2022). While the MAGs in these large-scale, short read-based studies are draft genomes, they represent significant progress towards identifying all of the taxa within the microbiome, as the largest study allowed mapping of ~95% of all oral microbiome reads to the draft genomes, with only <5% of the reads being unmapped and coming from an unknown bacterial genome (Zhu *et al.*, 2022). Crucially, between 30-77% of the species identified in these studies had no genomes in public repositories, illustrating that our understanding of the oral microbiota is still limited, and thousands of novel taxa are still awaiting study and naming (Pasolli *et al.*, 2019, Baker *et al.*, 2021, Zhu *et al.*, 2022). It is likely that many of these unknown taxa have been observed, and perhaps even given a designation, at the 16S level. Unfortunately, the 16S rRNA gene, due to the highly conserved elements, is only very rarely recovered in MAGs derived using short read sequencing. Long read metagenomic sequencing will be useful to link MAGs of novel species with their respective 16S sequences, allowing for previous 16S-based data to be leveraged for additional functional and taxonomic insight, with fewer data ending up in the “unknown taxa” bucket. Long read-based metagenomics of the oral microbiome has been limited, but the studies which have used it were highly successful in identifying novel oral phages and examining phage pangenomics (Yahara *et al.*, 2021), as well as obtaining complete genomes straight from saliva (Baker, 2021, Baker, 2022).

As these new oral genomes become available, phylogenomics analyses have identified many new species, and have led to the several major phylogenetic reorganizations of taxa in the oral microbiome. Most prominent was perhaps the 2020 reorganization of the family, *Lactobacillaceae* (Zheng *et al.*, 2020). This effort reclassified over 300 species in 7 genera and 2 families into one family, *Lactobacillaceae,* which contains 31 genera, including 23 new genera which were all formerly classified as the genus *Lactobacillus.* The reclassification was only

possible after high quality genome sequences became available for all the type strains, as the 16S sequences were inadequate to illustrate the real phylogenetic relationships (Zheng *et al.*, 2020). Similarly, the phylum *Actinobacteria* was re-classified in 2018 to include 2 orders, 10 families, and 17 genera, with over 100 species within the phylum being moved into a different genus (Nouioui *et al.*, 2018). Diverse phylogeny within *Saccharibacteria,* a candidate phylum within the Candidate Phyla Radation (CPR), continues to be resolved as new genomes become available to augment earlier 16S-based analysis (Cross *et al.*, 2019, McLean *et al.*, 2020, Shaiber *et al.*, 2020, Baker, 2021). On a smaller scale, phylogenomics has resolved the phylogeny of novel species within important oral taxa such as *Streptococcus dentisani* (Camelo-Castillo *et al.*, 2014), Candidatus *Bacteroides periocalifornicus* (Torres *et al.*, 2019), *Tannerella serpentiformis* (Ansbro *et al.*, 2020), and novel taxa within Actinobacteridae (Treerat *et al.*, 2022).

Linked closely with phylogenomics is pangenomics, and there has been no shortage of pangenome studies of oral taxa in recent years. A highlight of early pangenomics of oral bacteria was analysis of 57 *S. mutans* strains to gain insight on the links phylogeny and phenotypic/virulence traits (Cornejo *et al.*, 2013, Palmer *et al.*, 2013). More recent work reported a detailed, updated pangenome across 244 near complete genomes of *Streptococcus mutans* (Baker *et al.*, 2022). Additional contemporary comparative genomics of *S. mutans* and *S. sobriunus* indicated lack of phylogeographic differentiation for *S. mutans*, but some for *S. sobrinus* (Achtman & Zhou, 2020). Another recent study used an *S. mutans* pangenome to examine CRISPR spacers (Walker & Shields, 2022). Beyond *S. mutans,* several recent studies have analyzed other *Streptococcus* pangenomes. A pangenome of 113 genomes from 10 *Streptococcus* species was utilized to gain insight into ammonia production via the arginine deiminase system, and identified significant intra-species phenotypic heterogeneity (Velsko *et al.*, 2018). Site tropism of streptococci in the oral microbiome was examined using an approach that leveraged phylogenetic and pangenomic analysis, illustrating that even closely-related species such as *Streptococcus mitis, Streptococcus oralis,* and *Streptococcus infantis* specialized in

different sites within the oral cavity (McLean *et al.*, 2022). There was also substantial overlap in the core genomes of these 3 species, indicating that site-specialization is likely determined by subtle differences across the pangenome (McLean *et al.*, 2022). Other pangenome studies examined *Streptococcus intermedius* and its relationship to virulence at various body sites (Sinha *et al.*, 2021), identified homologs of adhesion and immune evasion across endocarditis and oral isolates of *S. sanguinis* and *S. gordonii,* (Iversen *et al.*, 2020), identified genomic factors influencing defense from phage and mobile genetic elements in *Dolosigranulum pigrum* (Flores Ramos *et al.*, 2021), and discovered that carbohydrate utilization pathways are well-conserved across *Veillonella* (Mashima *et al.*, 2021). Pangenome-based approaches also identified candidate genes involved in oral niche habitat adaptation for *Rothia mucilaginosa* and *Haemophilus parainfluenzae* (Utter *et al.*, 2020), and illustrated niche partitioning and vast differences in metabolic repertoires between clades of oral *Saccharibacteria* (Shaiber *et al.*, 2020, Baker, 2021, Baker *et al.*, 2021).

Dozens of studies have utilized transcriptomics (i.e., RNAseq) to study both individual oral bacteria under various conditions, as well as communities and the entire microbiome. Early analysis of the oral metatranscriptome was provided through several studies examining both caries (Peterson *et al.*, 2014, Do *et al.*, 2015) and periodontal disease (Duran-Pinedo *et al.*, 2014, Jorth *et al.*, 2014, Yost *et al.*, 2015, Belstrom *et al.*, 2017, Nowicki *et al.*, 2018), illustrating changes in both the taxonomy and functional expression in the microbiome in health versus disease. These findings were summarized in a recent review (Duran-Pinedo, 2021). Metatranscriptome changes following scaling and root planning as treatment for periodontal disease were examined, showing that there was a significant effect on progressing sites, but not so much in stable and fluctuating sites (Duran-Pinedo *et al.*, 2022). Transcriptomics was used to examine the relationship between the epibiont *Saccharibacteria, Nanosynbacter lyticus* and its host, *Schaalia odontolytica* (Hendrickson *et al.*, 2022). A transcriptomic time course of an in vitro dental plaque biofilm maturation provided insight of transcriptional inflection points in the community associated

with pH drops and blooms of acidophilic taxa such as *Limosilactobacillus fermentum* (Edlund *et al.*, 2018). Recent work has illustrated the transcriptome in periodontitis in a nonhuman primate model, which supported a significant role of the adaptive immune response in the kinetics of periodontal disease progression, and that aging effects on repertoire of immunoglobulin genes is likely to contribute to increased prevalence and severity of periodontal disease with age (Gonzalez *et al.*, 2022). Furthermore, that the same bacterial taxa interface with host immunology differently at a healthy site compared to a diseased site (Ebersole *et al.*, 2021). Other recent work explored the role of health-associated oral bacteria on the transcriptome of oral squamous cell carcinoma cell lines (Baraniya *et al.*, 2022). As the oral microbiology field begins to adopt third generation RNA sequencing, a wealth of data regarding transcriptional isoforms and RNA modification will soon become available. Several additional studies using sequencing-based omics as part of multi-omics are discussed in the Multi-omics section below.

MASS SPECTROMETRY-BASED OMICS

Proteomics, Metabolomics, and Lipidomics

In addition to all the advances described above, which are dependent on nucleic acid sequencing, there have been major recent advances to omics analyses that depend on mass spectrometry. Recent innovations in mass spectrometry-based research have come through advancements in mass spectrometry instrumentation, machine-learning based spectral analysis methods, molecular networking, and crowd-sourced spectral annotation. Matrix-assisted laser desorption/ionization (MALDI; including MALDI imaging), orbitrap, and native mass spectrometry technologies have made mass spectrometry analysis significantly more sensitive, accurate, and able to detect a wider range of molecules. Highlights of technological advances and available methods/tools, guidelines for best practices, and challenges facing the field have been provided by several excellent and recent reviews on proteomics (Chen *et al.*, 2020), metabolomics (Alseekh *et al.*, 2021), lipidomics (Ni *et al.*, 2022), and the specific application of metabolomics in

483 microbiome data (Bauermeister *et al.*, 2022). Briefly, although lipidomics, proteomics, and
484 metabolomics represent the most complete and “current” state of a given sample (i.e., rather than
485 what is encoded for by DNA or soon-to-be translated RNA), there are several unique challenges
486 facing mass spectrometry analysis, particularly metabolomics and lipidomics. Like nucleic acid
487 sequencing, proteomics ultimately generates data that has a more standardized “sequence” (i.e.,
488 the proteins are limited to an amino acid sequence) which can be compared against large,
489 relatively comprehensive databases (e.g., NCBI, UniProt, etc.). A large number of tools exist to
490 perform protein sequence identification against databases or de novo peptide sequencing, and
491 execute downstream analyses such as quantification. Several tools of note are listed in Figure
492 1B, and these and others have been reviewed in depth (Chen *et al.*, 2020). On the other hand,
493 metabolomics and lipidomics data do not generate sequences, the with the molecules being
494 detected occupying a comparatively unlimited chemical space. Furthermore, many of the
495 databases used for dereplication (i.e., identification of known compounds) are not freely available.
496 As a result, a much higher percentage of the features detected in metabolomics and lipidomics
497 datasets are unknown. The Global Natural Products Social Molecular Networking (GNPS) was
498 published in 2016 to help address some of these issues, creating an open-access knowledge
499 base for organization and sharing of mass spectrometry data, which is reanalyzed as the
500 database grows, leveraging molecular networking to help identify novel spectra (Wang *et al.*,
501 2016). Molecular networking is a visualization of spectral alignment and correlation, which
502 enables prediction of the chemical structure of unknown features. Originally developed for liquid
503 chromatography mass spectrometry (LC-MS), the GNPS was recently updated to enable analysis
504 of gas chromatography mass spectrometry (GC-MS), which epand its utility to many GC-MS-
505 based lipidomics and metabolomics analyses (Aksenov *et al.*, 2021). Other recent innovations to
506 mass spectrometry-based omics include the use of metadata to enhance annotation of
507 metabolomics (Gauglitz *et al.*, 2022), native spray metal metabolomics to identify novel
508 siderophores and other metal-binding compounds (Aron *et al.*, 2022), and ion identity molecular

networking (IIMN) to integrate chromatographic peak shape into molecular networking, enhancing annotation with molecular networks (Schmid *et al.*, 2021). Going forward, these advancements in mass spectrometry analysis methods are poised to increase the scale and pace of discovery in the oral microbiota.

Impact of mass spectrometry-based omics on oral microbiome research

Proteomics was utilized to study stress responses of the caries pathogen, *S. mutans* as early as 2004 (Len *et al.*, 2004), and many other studies have examined single oral taxa using proteomics and metabolomics. A recent study examined the *S. mutans* proteome during acid and oxidative stress, illustrating modules of co-expressed proteins under various stress conditions (Tinder *et al.*, 2022). A landmark metaproteomics study of the oral microbiome identified potential biomarkers for caries (Belda-Ferre *et al.*, 2015). Beyond the strictly microbial constituents of the oral microbiota, saliva has great diagnostic potential, due to its accessibility and the large number biomarkers that can be measured using proteomics and/or metabolomics (Dawes & Wong, 2019). Along those lines, the Human Salivary Proteome Wiki was recently developed, and serves as a public data platform for researching and retrieving custom-curated data knowledge of the salivary proteome (Lau *et al.*, 2021). Although lipidomics of single species, such as *S. mutans* (Fozo & Quivey, 2004), have been performed and used to study physiology, the lipidome of the oral microbiota as a community is in need of further study. Several studies that have used mass spectrometry-based omics in oral microbiome research are also mentioned in the multi-omics section below.

COMPOSITIONAL ANALYSIS, SINGLE-CELL OMICS, AND MULTI-OMICS

A note on compositional data and analysis tools

Nearly all omics data is compositional in nature, meaning that it a quantitative description of parts of some whole, therefore conveying relative information. The limitations of compositional data have been excellently reviewed (Gloor *et al.*, 2017, Morton *et al.*, 2017, Knight *et al.*, 2018, Morton *et al.*, 2019), and it is imperative that researchers are aware that omics data is compositional, perform analysis using tools designed to handle compositional data, and be cognizant of the limitations inherent to compositional data. Determining correlation is particularly intractable with compositional data, with conventional methods producing unacceptably high false discovery rates. Numerous approaches have been developed to address these problems, including LEfSe (Segata *et al.*, 2011), DESeq (Love *et al.*, 2014), ALDEx2 (Fernandes *et al.*, 2014), ANCOM (Mandal *et al.*, 2015), and Songbird (Morton *et al.*, 2019); however, none are "perfect". Ultimately, it is generally best to analyze compositional data using multiple approaches and take all results with a grain of salt when forming hypotheses.

Single-cell omics

Single-cell analysis is transformational technology, allowing for the omics analysis of individual cells and identification of discrete biological dynamics that are obscured by the averages obtained by traditional bulk analysis. Single-cell proteomics, lipidomics, and metabolomics, based on mass spectrometry, is an advancing field, however it is still at a nascent stage even for eukaryotes and therefore will not be discussed (Couvillion *et al.*, 2019, Perkel, 2021, Tajik *et al.*, 2022). Meanwhile, driven by advancements in microfluidics, sample handling, labeling, imaging, bioinformatics, computational biology, and machine learning, companies like 10X Genomics and Standard Biotools are making single-cell analysis of eukaryotic genomes and transcriptomes (scRNA-seq) commonplace. Challenges facing scRNA-seq in bacteria include low content of mRNA, lack of a polyA tail on mRNAs, diverse cell walls, and small size hindering microfluidic single-cell isolation (Kuchina *et al.*, 2021). Early attempts at bacterial scRNA-seq involved using fluorescence activated cell sorting (FACS) to distribute individual cells to wells in 96-well plates,

however this technique is low throughput, with a very high cost to examine only several hundred bacterial cells (Imdahl *et al.*, 2020). Two concurrently developed, yet technically similar, approaches to deal with these issues are MicroSPLiT (Kuchina *et al.*, 2021) and PETRI-seq (Blattman *et al.*, 2020) which do not depend on single-cell isolation. Cells are permeabilized and then labeled with several rounds of split-pool barcoding of cDNA to ensure that nearly every cell has a unique barcode prior to sequencing (Blattman *et al.*, 2020, Kuchina *et al.*, 2021). These approaches were able to differentiate multiple transcriptional states in *Bacillus subtilis* and *Escherichia coli,* respectively (Blattman *et al.*, 2020, Kuchina *et al.*, 2021). More recent approaches have modified other eukaryotic scRNA-seq protocols such as multiple annealing and dC-tailing-based quantitative single-cell RNA-seq (MATQ-seq)(Homberger *et al.*, 2023), and made use of the 10X Genomics Chromium microfluidic device (Brennan & Rosenthal, 2021) to perform bacterial scRNA-seq.

In oral microbiome research, single-cell techniques have been used to isolate cells and amplify DNA to generate single-cell amplified assembled genomes (SAGs) of *Saccharibacteria* (Cross *et al.*, 2019), Chloroflexi and Chlorobi (Campbell *et al.*, 2014), *Tannerella* (Beall *et al.*, 2014), *Porphyromonas* (McLean *et al.*, 2013), and *Desulfovibrio* and *Desulfobulbus* (Campbell *et al.*, 2013), and these techniques and findings were recently reviewed (Balachandran *et al.*, 2020). Most of these organisms were present in such low numbers in the original sample that getting a substantial portion of the respective genome sequence would have been impossible without the single-cell methods. Although at this time, no studies have leveraged single-cell technology to study oral bacteria at the transcriptional level, a recent landmark study generated an atlas of human oral mucosa cells using scRNA-seq, examining healthy individuals versus periodontitis, revealing exaggerated responsiveness of stromal cells and enhanced immune cell infiltration in periodontitis (Williams *et al.*, 2021). A recent study used scRNA-seq to examine expression of periodontitis susceptibility genes in human gingival cells (Caetano *et al.*, 2022).

Multi-omics

While integrating multiple types of omics analysis is critical for microbiome research, this type of analysis introduces several additional statistical challenges as now multiple datasets that are each compositional are now being compared. Crucially, many tools specifically developed for handling compositional data lose scale invariance when applied to multi-omics datasets (Morton *et al.*, 2019). mmvec, a recently developed approach for analyzing multi-omics data uses co-occurrence probabilities rather than correlations (Morton *et al.*, 2019). When applied to metagenome and metabolome data, it allowed researchers to identify the most likely microbe-metabolite interactions (Morton *et al.*, 2019). Another tool, iNetModels2, was recently developed for interactively visualizing multi-omics data (Arif *et al.*, 2021).

Several examples exist of published research used multi-omics data to examine the oral microbiota in various contexts. Multi-omics analysis of an in vitro oral biofilm community following a glucose pulse revealed temporal regulation of fermentation pathways affected pH of the culture and subsequent micro ecology (Edlund *et al.*, 2015). Multi-omics of dental plaque from patients with diabetes and periodontal disease identified both proteins and lipids that were associated with disease, and also showed that *Lautropia mirabilis* synthesizes monomethyl phosphatidylethanolamine, which is rarely produced by bacteria (Overmyer *et al.*, 2021). Multi-omics of germ-free and specific pathogen-free mice indicated that the oral microbiota influenced the permeability of the oral epithelial barrier, vis-à-vis keratinization and cell adhesion (Long *et al.*, 2022). The relationship between the oral microbiome and chronic sleep deprivation was examined in rats, observing both taxonomic changes in the microbiota, as well as modulation of host immunological molecules (Chen *et al.*, 2022). Finally, a recent study examined the proteome and microbiome of diseased gingival tissue (Bao *et al.*, 2020).

PERSPECTIVES

Omics approaches have transformed our understanding of the oral microbiome and its relationship to human health, allowing for studies with a scale and resolution unimaginable 20 years ago. The HOMD now contains genomes, in addition to 16S sequences, and now includes the taxa from the aerodigestive tract and the oral cavity (Escapa *et al.*, 2018). Some of the main challenges currently facing omics-based microbiome research are standardization and deposition of data in public repositories, as well as re-analysis of old data with updated reference databases. Although repositories such as the Sequence Read Archive (SRA), RefSeq, and GenBank are highly useful and do enforce some level of standardization, journals and reviewers do not always enforce deposition of published data into these databases. Furthermore, unified repositories and data file formats are significantly more limited (and many times are vendor-specific/proprietary) for mass spectrometry data. Efforts to make public databases into "living data" will also be highly useful. For example, as more and more accurate and complete genomes get deposited into the databases used to analyze taxonomy of sequencing reads, older raw microbiome datasets can be periodically re-analyzed, and reads representing newly identified taxa can be moved from the "unknown taxa" to the proper newly identified taxa (which may change interpretation of the results and/or identify new data trends). This is being implemented to some extent in the SRA, with entries now having a "Taxonomy Analysis" tab included in the Run Browser (Katz *et al.*, 2021). The same is true for mass spectrometry datasets, as new reference spectra get identified and added to public databases. The GNPS already has implemented "living data" using periodic reanalysis of metabolomics data stored in its repository (Wang *et al.*, 2016). Additionally, to help reduce some of issues in equity and reproducibility facing the field, enforcement of the publication of all analysis tools, settings, and code used in omics-based research on public repositories such as GitHub would be helpful. Continued research of the oral microbiome using omics-based approaches is needed, especially those sampling more diverse populations and performing longitudinal analysis. The discoveries enabled by this type of research will significantly improve human health.

637

Acknowledgements

This work was supported by NIH/NIDCR K99-DE029228 and the Research Council of Norway (Norges Forskningsråd) INTPART-32282.

Conflict of Interest Statement

The author declares no conflicts of interest.

Figure 1

A.

Timeline of milestones in omics technologies and oral microbiome research

1990
Initiation of the Human Genome Project

1995
Publication of the first complete genome of a free-living organism, *Haemophilus influenzae* (Fleishmann et al.)

1998
Foundation of Solexa and Illumina

2001
Publication of the human genome (draft/euchromatin only) (Venter, et al.) and (Lander, et al.)

2004
Environmental shotgun sequencing of the Sargasso Sea (Venter et al.)

2006
Microbiome anlaysis of the distal human gut (Gill et al.)

2010
1st untargeted metabolomics of oral microbiome (Takahashi et al.)

2012
Publication of Human Oral Microbiome Database (HOMD) (Chen et al.)

2018
-Transcriptomics using ONT (Geralde et al.)
-Expanded Human Oral Microbiome Database (eHOMD) (Escapa et al.)

2021
1st complete bacterial genomes directly from saliva (Baker)

Milestone in microbiome / omics research

Major technological advance

Milestone in oral microbiome research

1999
First report of MALDI imaging mass spectrometry (Stoeckli et al.)

2005
-Gut microbiota first linked to obesity (Let et al.)
-Roche 454 Sequencing (Marguiles et al.)
-First orbitrap mass spectrometer (Yates et al.)

2007
1st proteomics of the acquired pellicle (Siqueira et al.)

2008
1st NGS-based study of the oral microbiome (Keijser et al.)

2009
-Human Microbiome Project initiated
-SOLiD Sequencing (McKernan et al.)
-PacBio SMRT Sequencing (Eid et al.)

2014
-Oxford Nanopore releases MinION sequencer
-1st multiomics study (metagenome & metatranscriptome) of oral microbiome (Benitez-Paez et al.)

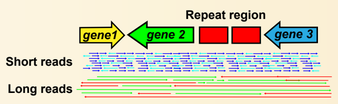
2015
1st metaproteomics of the oral microbiome (Belda-Ferre et al.)

2022
Completion of the human genome (telomere to telomere) (Nurk et al.)
ONT R10 chemistry enables accurate genomic sequencing using ONT without Illumina polishing (Sereika et al.)

B.

Omics approaches and tools

Genomics



Genome assemblers

Short-read
ABySS (Simpson et al., 2009)
Velvet (Zerbino & Birney, 2008)
MEGAHIT (Li et al., 2015)
**SPAdes (Prjibelski et al., 2020)

Long-read
Canu (Koren et al., 2017)
HGAP (Chin et al., 2013)
miniasm (Li, 2016)
MaSuRCA (Zimin et al., 2013)
**Flye (Kolmogorov et al., 2019)

Hybrid
Unicycler (Wick et al., 2017)
hybridSPAdes (Antipov et al., 2016)
**Trycycler (Wick et al., 2021)

Genome polishing

Short-read
racon (Vaser et al., 2017)
pilon (Walker et al., 2014)
**polypolish (Wick & Holt, 2022)

Long-read
nanopolish (Loman et al., 2015)
medaka (ONT)

DNA modification detection

Tombo (ONT)

Metagenomics

Read-based abundance

Taxonomic abundance
MetaPhlAn4 (Blanco-Miguez et al., 2022)
Kraken2 (Lu et al., 2022)

Functional Abundance
HumanN3 (Beghini et al., 2021)

Metagenome assemblers

Short-read
**metaSPAdes (Nurk et al., 2017)
MEGAHIT (Li et al., 2015)

Long-read
**metaFlye (Kolmogorov et al., 2020)
**strainFlye (Fedarko et al., 2022)

Genome binning

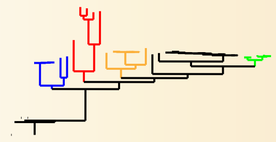
Short-read
MaxBin2 (Wu et al., 2016)
Concoct (Alneberg et al., 2014)
MetaBat2 (Kang et al., 2019)

Long-read
LRBinner (Wickramarachichi & Lin 2022)

Composite binning
DAStool (Sieber et al., 2018)
MetaWRAP (Uritskiy et al., 2018)
VEBA (Espinoza & Dupont, 2022)

Manual binning
Anvi'o (Eren et al., 2021)

Phylogenomics



PhyloSift (Darling et al., 2014)
**PhyloPhlAn3 (Asnicar et al., 2020)
**Anvi'o (Eren et al., 2021)

Pangenomics

Roary (Page et al., 2015)
PanPhlAn3 (Beghini et al., 2021)
panOCT (Fouts et al., 2012)
**Anvi'o (Eren et al., 2021)

Transcriptomics

RNA modification detection

MetaCompore (Leger & Leonardi, 2021)
Tombo (ONT)
EpiNano (Liu et al., 2019)
MasterofPores (Cozzunto et al., 2020)

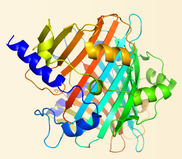
RNA assemblers

Trinity (Grabherr et al. 2011)
RockHopper2 (Tjaden 2015)
rnaSPAdes (Bushmanova et al., 2019)

Feature quantification

featureCounts (Liao et al., 2014)

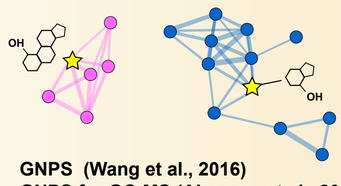
Proteomics



Peptide & protein identification
SEQUEST (Brodbelt et al., 2015)
MaxQuant (Tyanova et al., 2016)
DeepNovo (Tran et al., 2019)
UniNovo (Jeong et al., 203)

Protein quantification
Proteome Discoverer (Thermo Scientific)
MassIVE.quant (Choi et al., 2020)

Metabolomics & Lipidomics



GNPS (Wang et al., 2016)
GNPS for GC-MS (Aksenov et al., 2021)
MZmine3 (Pluskal et al., 2010)
MS-DIAL (Tsugawa et al., 2020)
LipidMatch (Koelmel et al., 2017)
Sirius5 (Duhrkop et al., 2019)

Figure Legend

Figure 1: (A) Timeline of milestones in omics technologies and oral microbiome research. This timeline highlights milestones in microbiome/omics research generally (yellow), major technological advances (green), and milestones in oral microbiome research, specifically (blue) over the past 33 years. **(B) Omics approaches and tools.** For each of the 7 omics approaches discussed here, a list of the most significant and/or commonly used bioinformatics tools is provided. Note that this list is not exhaustive and readers are referred in the main text to additional references on the specific software and benchmarking. **denotes a particularly useful or "gold standard" tool. ONT, Oxford Nanopore Technologies.

References

Achtman M & Zhou Z (2020) Metagenomics of the modern and historical human oral microbiome with phylogenetic studies on Streptococcus mutans and Streptococcus sobrinus. *Philos Trans R Soc Lond B Biol Sci* **375**: 20190573.

Aksenov AA, Laponogov I, Zhang Z*, et al.* (2021) Auto-deconvolution and molecular networking of gas chromatography-mass spectrometry data. *Nat Biotechnol* **39**: 169-173.

Al-Hebshi NN, Baraniya D, Chen T, Hill J, Puri S, Tellez M, Hasan NA, Colwell RR & Ismail A (2019) Metagenome sequencing-based strain-level and functional characterization of supragingival microbiome associated with dental caries in children. *J Oral Microbiol* **11**: 1557986.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF & Quince C (2014) Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**: 1144-1146.

Alseekh S, Aharoni A, Brotman Y*, et al.* (2021) Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat Methods* **18**: 747-756.

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME & Gouil Q (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* **21**: 30.

Ansbro K, Wade WG & Stafford GP (2020) Tannerella serpentiformis sp. nov., isolated from the human mouth. *Int J Syst Evol Microbiol* **70**: 3749-3754.

Antipov D, Korobeynikov A, McLean JS & Pevzner PA (2016) hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**: 1009-1015.

Arif M, Zhang C, Li X*, et al.* (2021) iNetModels 2.0: an interactive visualization and database of multi-omics data. *Nucleic Acids Res* **49**: W271-W276.

Aron AT, Petras D, Schmid R*, et al.* (2022) Native mass spectrometry-based metabolomics identifies metal-binding compounds. *Nat Chem* **14**: 100-109.

Asnicar F, Thomas AM, Beghini F*, et al.* (2020) Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun* **11**: 2500.

Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC & Scorilas A (2021) Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. *Life (Basel)* **12**.

Baker JL (2021) Complete Genomes of Clade G6 Saccharibacteria Suggest a Divergent Ecological Niche and Lifestyle. *mSphere* **6**: e0053021.

Baker JL (2022) Using Nanopore Sequencing to Obtain Complete Bacterial Genomes from Saliva Samples. *mSystems* **7**: e0049122.

Baker JL, Bor B, Agnello M, Shi W & He X (2017) Ecology of the Oral Microbiome: Beyond Bacteria. *Trends Microbiol* **25**: 362-374.

Baker JL, Tang X, LaBonte S, Uranga C & Edlund A (2022) mucG, mucH, and mucI Modulate Production of Mutanocyclin and Reutericyclins in Streptococcus mutans B04Sm5. *J Bacteriol* **204**: e0004222.

Baker JL, Morton JT, Dinis M, Alvarez R, Tran NC, Knight R & Edlund A (2021) Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *Genome Res* **31**: 64-74.

Balachandran M, Cross KL & Podar M (2020) Single-Cell Genomics and the Oral Microbiome. *J Dent Res* **99**: 613-620.

Bao K, Li X, Poveda L*, et al.* (2020) Proteome and Microbiome Mapping of Human Gingival Tissue in Health and Disease. *Front Cell Infect Microbiol* **10**: 588155.

Baraniya D, Chitrala KN & Al-Hebshi NN (2022) Global transcriptional response of oral squamous cell carcinoma cell lines to health-associated oral bacteria - an in vitro study. *J Oral Microbiol* **14**: 2073866.

Bauermeister A, Mannochio-Russo H, Costa-Lotufo LV, Jarmusch AK & Dorrestein PC (2022) Mass spectrometry-based metabolomics in microbiome investigations. *Nat Rev Microbiol* **20**: 143-160.

Beall CJ, Campbell AG, Dayeh DM, Griffen AL, Podar M & Leys EJ (2014) Single cell genomics of uncultured, health-associated Tannerella BU063 (Oral Taxon 286) and comparison to the closely related pathogen Tannerella forsythia. *PLoS One* **9**: e89398.

Beaulaurier J, Zhu S, Deikus G*, et al.* (2018) Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* **36**: 61-69.

Beghini F, McIver LJ, Blanco-Miguez A*, et al.* (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**.

Belda-Ferre P, Williamson J, Simon-Soro A, Artacho A, Jensen ON & Mira A (2015) The human oral metaproteome reveals potential biomarkers for caries disease. *Proteomics* **15**: 3497-3507.

Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simon-Soro A, Pignatelli M & Mira A (2012) The oral metagenome in health and disease. *ISME J* **6**: 46-56.

Belstrom D, Constancias F, Liu Y, Yang L, Drautz-Moses DI, Schuster SC, Kohli GS, Jakobsen TH, Holmstrup P & Givskov M (2017) Metagenomic and metatranscriptomic analysis of saliva reveals disease-associated microbiota in patients with periodontitis and dental caries. *NPJ Biofilms Microbiomes* **3**: 23.

Bennett S (2004) Solexa Ltd. *Pharmacogenomics* **5**: 433-438.

Bentley DR & Balasubramanian S & Swerdlow HP*, et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.

Blattman SB, Jiang W, Oikonomou P & Tavazoie S (2020) Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. *Nat Microbiol* **5**: 1192-1201.

Bolyen E & Rideout JR & Dillon MR*, et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**: 852-857.

Bowen WH, Burne RA, Wu H & Koo H (2018) Oral Biofilms: Pathogens, Matrix, and Polymicrobial Interactions in Microenvironments. *Trends Microbiol* **26**: 229-242.

Brennan MA & Rosenthal AZ (2021) Single-Cell RNA Sequencing Elucidates the Structure and Organization of Microbial Communities. *Front Microbiol* **12**: 713128.

Brodbelt JS & Russell DH (2015) Focus on the 20-year anniversary of SEQUEST. *J Am Soc Mass Spectrom* **26**: 1797-1798.

Bushmanova E, Antipov D, Lapidus A & Prjibelski AD (2019) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**.

Caetano AJ, D'Agostino EM, Sharpe P & Nibali L (2022) Expression of periodontitis susceptibility genes in human gingiva using single-cell RNA sequencing. *J Periodontal Res* **57**: 1210-1218.

Callahan BJ, Grinevich D, Thakur S, Balamotis MA & Yehezkel TB (2021) Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome* **9**: 130.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ & Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**: 581-583.

Camelo-Castillo A, Benitez-Paez A, Belda-Ferre P, Cabrera-Rubio R & Mira A (2014) Streptococcus dentisani sp. nov., a novel member of the mitis group. *Int J Syst Evol Microbiol* **64**: 60-65.

Campbell AG, Schwientek P, Vishnivetskaya T, Woyke T, Levy S, Beall CJ, Griffen A, Leys E & Podar M (2014) Diversity and genomic insights into the uncultured Chloroflexi from the human microbiota. *Environ Microbiol* **16**: 2635-2643.

Campbell AG, Campbell JH, Schwientek P, Woyke T, Sczyrba A, Allman S, Beall CJ, Griffen A, Leys E & Podar M (2013) Multiple single-cell genomes provide insight into functions of uncultured Deltaproteobacteria in the human oral cavity. *PLoS One* **8**: e59361.

Chen C, Hou J, Tanner JJ & Cheng J (2020) Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *Int J Mol Sci* **21**.

Chen LX, Anantharaman K, Shaiber A, Eren AM & Banfield JF (2020) Accurate and complete genomes from metagenomes. *Genome Res* **30**: 315-333.

Chen P, Wu H, Yao H, Zhang J, Fan W, Chen Z, Su W, Wang Y & Li P (2022) Multi-Omics Analysis Reveals the Systematic Relationship Between Oral Homeostasis and Chronic Sleep Deprivation in Rats. *Front Immunol* **13**: 847132.

Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A & Dewhirst FE (2010) The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)* **2010**: baq013.

Chin CS, Alexander DH, Marks P*, et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**: 563-569.

Cornejo OE, Lefebure T, Bitar PD*, et al.* (2013) Evolutionary and population genomics of the cavity causing bacteria Streptococcus mutans. *Mol Biol Evol* **30**: 881-893.

Couvillion SP, Zhu Y, Nagy G, Adkins JN, Ansong C, Renslow RS, Piehowski PD, Ibrahim YM, Kelly RT & Metz TO (2019) New mass spectrometry technologies contributing towards comprehensive and high throughput omics analyses of single cells. *Analyst* **144**: 794-807.

Cross KL, Campbell JH, Balachandran M*, et al.* (2019) Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat Biotechnol* **37**: 1314-1321.

Cusco A, Perez D, Vines J, Fabregas N & Francino O (2021) Long-read metagenomics retrieves complete single-contig bacterial genomes from canine feces. *BMC Genomics* **22**: 330.

Darling AE, Jospin G, Lowe E, Matsen FAt, Bik HM & Eisen JA (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**: e243.

Dawes C & Wong DTW (2019) Role of Saliva and Salivary Diagnostics in the Advancement of Oral Health. *J Dent Res* **98**: 133-141.

Deorowicz S, Debudaj-Grabysz A & Gudys A (2016) FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Sci Rep* **6**: 33964.

Do T, Sheehy EC, Mulli T, Hughes F & Beighton D (2015) Transcriptomic analysis of three Veillonella spp. present in carious dentine and in the saliva of caries-free individuals. *Front Cell Infect Microbiol* **5**: 25.

Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C & Langille MGI (2020) PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol* **38**: 685-688.

Duran-Pinedo AE (2021) Metatranscriptomic analyses of the oral microbiome. *Periodontol 2000* **85**: 28-45.

Duran-Pinedo AE, Solbiati J, Teles F & Frias-Lopez J (2022) Subgingival host-microbiome metatranscriptomic changes following scaling and root planing in grade II/III periodontitis. *J Clin Periodontol*.

Duran-Pinedo AE, Chen T, Teles R, Starr JR, Wang X, Krishnan K & Frias-Lopez J (2014) Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. *ISME J* **8**: 1659-1672.

Ebersole JL, Nagarajan R, Kirakodu S & Gonzalez OA (2021) Oral Microbiome and Gingival Gene Expression of Inflammatory Biomolecules With Aging and Periodontitis. *Front Oral Health* **2**: 725115.

Economopoulou P, Kotsantis I & Psyrri A (2020) Special Issue about Head and Neck Cancers: HPV Positive Cancers. *Int J Mol Sci* **21**.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Edlund A, Yang Y, Yooseph S, He X, Shi W & McLean JS (2018) Uncovering complex microbiome activities via metatranscriptomics during 24 hours of oral biofilm assembly and maturation. *Microbiome* **6**: 217.

Edlund A, Yang Y, Yooseph S*, et al.* (2015) Meta-omics uncover temporal regulation of pathways across oral microbiome genera during in vitro sugar metabolism. *ISME J* **9**: 2605-2619.

Eren AM, Kiefl E, Shaiber A*, et al.* (2021) Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* **6**: 3-6.

Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE & Lemon KP (2018) New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. *mSystems* **3**.

Espinoza JL & Dupont CL (2022) VEBA: a modular end-to-end suite for in silico recovery, clustering, and analysis of prokaryotic, microeukaryotic, and viral genomes from metagenomes. *BMC Bioinformatics* **23**: 419.

Faulk C (2022) De novo sequencing, diploid assembly, and annotation of the black carpenter ant, Camponotus pennsylvanicus, and its symbionts by one person for $1000, using nanopore sequencing. *Nucleic Acids Res*.

Fedarko MW, Kolmogorov M & Pevzner PA (2022) Analyzing rare mutations in metagenomes assembled using long and accurate reads. *Genome Res* **32**: 2119-2133.

Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR & Gloor GB (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**: 15.

Flores Ramos S, Brugger SD, Escapa IF*, et al.* (2021) Genomic Stability and Genetic Defense Systems in Dolosigranulum pigrum, a Candidate Beneficial Bacterium from the Human Microbiome. *mSystems* **6**: e0042521.

Fouts DE, Brinkac L, Beck E, Inman J & Sutton G (2012) PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res* **40**: e172.

Fozo EM & Quivey RG, Jr. (2004) Shifts in the membrane fatty acid profile of Streptococcus mutans enhance survival in acidic environments. *Appl Environ Microbiol* **70**: 929-936.

Garalde DR, Snell EA, Jachimowicz D*, et al.* (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201-206.

Gauglitz JM, West KA, Bittremieux W*, et al.* (2022) Enhancing untargeted metabolomics using metadata-based source annotation. *Nat Biotechnol* **40**: 1774-1779.

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM & Nelson KE (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312**: 1355-1359.

Gloor GB, Macklaim JM, Pawlowsky-Glahn V & Egozcue JJ (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* **8**: 2224.

Gonzalez OA, Kirakodu SS, Nguyen LM & Ebersole JL (2022) Gingival Transcriptomic Patterns of Macrophage Polarization during Initiation, Progression, and Resolution of Periodontitis. *Clin Exp Immunol*.

Goussarov G, Mysara M, Vandamme P & Van Houdt R (2022) Introduction to the principles and methods underlying the recovery of metagenome-assembled genomes from metagenomic data. *Microbiologyopen* **11**: e1298.

Grabherr MG, Haas BJ, Yassour M*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644-652.

Grunberger F, Ferreira-Cerca S & Grohmann D (2022) Nanopore sequencing of RNA and cDNA molecules in Escherichia coli. *RNA* **28**: 400-417.

Hajishengallis G & Chavakis T (2021) Local and systemic mechanisms linking periodontal disease and inflammatory comorbidities. *Nat Rev Immunol* **21**: 426-440.

He X, McLean JS, Guo L, Lux R & Shi W (2014) The social structure of microbial community involved in colonization resistance. *ISME J* **8**: 564-574.

Hendrickson EL, Bor B, Kerns KA*, et al.* (2022) Transcriptome of Epibiont Saccharibacteria Nanosynbacter lyticus Strain TM7x During the Establishment of Symbiosis. *J Bacteriol* **204**: e0011222.

Homberger C, Saliba AE & Vogel J (2023) A MATQ-seq-Based Protocol for Single-Cell RNA-seq in Bacteria. *Methods Mol Biol* **2584**: 105-121.

Human Microbiome Project C (2012) Structure, function and diversity of the healthy human microbiome. *Nature* **486**: 207-214.

Imdahl F, Vafadarnejad E, Homberger C, Saliba AE & Vogel J (2020) Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. *Nat Microbiol* **5**: 1202-1206.

Iversen KH, Rasmussen LH, Al-Nakeeb K*, et al.* (2020) Similar genomic patterns of clinical infective endocarditis and oral isolates of Streptococcus sanguinis and Streptococcus gordonii. *Sci Rep* **10**: 2728.

Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT & Aluru S (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**: 5114.

Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B & Akeson M (2015) Improved data analysis for the MinION nanopore sequencer. *Nat Methods* **12**: 351-356.

Jorth P, Turner KH, Gumus P, Nizam N, Buduneli N & Whiteley M (2014) Metatranscriptomics of the human oral microbiome during health and disease. *mBio* **5**: e01012-01014.

Kang DD, Li F, Kirton E, Thomas A, Egan R, An H & Wang Z (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**: e7359.

Katz KS, Shutov O, Lapoint R, Kimelman M, Brister JR & O'Sullivan C (2021) STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biol* **22**: 270.

Knight R, Vrbanac A, Taylor BC*, et al.* (2018) Best practices for analysing microbiomes. *Nat Rev Microbiol* **16**: 410-422.

Kolmogorov M, Yuan J, Lin Y & Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**: 540-546.

Kolmogorov M, Bickhart DM, Behsaz B*, et al.* (2020) metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* **17**: 1103-1110.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH & Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722-736.

Koren S, Schatz MC, Walenz BP*, et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* **30**: 693-700.

Kuchina A, Brettner LM, Paleologu L, Roco CM, Rosenberg AB, Carignano A, Kibler R, Hirano M, DePaolo RW & Seelig G (2021) Microbial single-cell RNA sequencing by split-pool barcoding. *Science* **371**.

Lamont RJ, Koo H & Hajishengallis G (2018) The oral microbiota: dynamic communities and host interactions. *Nat Rev Microbiol* **16**: 745-759.

Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Lau WW, Hardt M, Zhang YH, Freire M & Ruhl S (2021) The Human Salivary Proteome Wiki: A Community-Driven Research Platform. *J Dent Res* **100**: 1510-1519.

Len ACL, Harty DWS & Jacques NA (2004) Proteome analysis of Streptococcus mutans metabolic phenotype during acid tolerance. *Microbiology (Reading)* **150**: 1353-1366.
Leung SK, Jeffries AR, Castanho I*, et al.* (2021) Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Rep* **37**: 110022.
Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD & Gordon JI (2005) Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* **102**: 11070-11075.
Li D, Liu CM, Luo R, Sadakane K & Lam TW (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**: 1674-1676.
Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**: 2843-2851.
Li H (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103-2110.
Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.
Liao Y, Smyth GK & Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923-930.
Liu L, Yang Y, Deng Y & Zhang T (2022) Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes. *Microbiome* **10**: 209.
Loman NJ, Quick J & Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* **12**: 733-735.
Long H, Yan L, Pu J*, et al.* (2022) Multi-omics analysis reveals the effects of microbiota on oral homeostasis. *Front Immunol* **13**: 1005992.
Love MI, Huber W & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.
Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, Salzberg SL & Steinegger M (2022) Metagenome analysis using the Kraken software suite. *Nat Protoc* **17**: 2815-2839.
Mandal S, Van Treuren W, White RA, Eggesbo M, Knight R & Peddada SD (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* **26**: 27663.
Margulies M, Egholm M, Altman WE*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.
Mashima I, Liao YC, Lin CH, Nakazawa F, Haase EM, Kiyoura Y & Scannapieco FA (2021) Comparative Pan-Genome Analysis of Oral Veillonella Species. *Microorganisms* **9**.
McKernan KJ, Peckham HE, Costa GL*, et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**: 1527-1541.
McLean AR, Torres-Morales J, Dewhirst FE, Borisy GG & Mark Welch JL (2022) Site-tropism of streptococci in the oral microbiome. *Mol Oral Microbiol* **37**: 229-243.
McLean JS, Bor B, Kerns KA, Liu Q, To TT, Solden L, Hendrickson EL, Wrighton K, Shi W & He X (2020) Acquisition and Adaptation of Ultra-small Parasitic Reduced Genome Bacteria to Mammalian Hosts. *Cell Rep* **32**: 107939.
McLean JS, Lombardo MJ, Ziegler MG*, et al.* (2013) Genome of the pathogen Porphyromonas gingivalis recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform. *Genome Res* **23**: 867-877.
Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, Zengler K & Knight R (2019) Establishing microbial composition measurement standards with reference frames. *Nat Commun* **10**: 2719.

954 Morton JT, Sanders J, Quinn RA*, et al.* (2017) Balance Trees Reveal Microbial Niche
955 Differentiation. *mSystems* **2**.
956 Morton JT, Aksenov AA, Nothias LF*, et al.* (2019) Learning representations of microbe-metabolite
957 interactions. *Nat Methods* **16**: 1306-1314.
958 Moss EL, Maghini DG & Bhatt AS (2020) Complete, closed bacterial genomes from microbiomes
959 using nanopore sequencing. *Nat Biotechnol* **38**: 701-707.
960 Moussa DG, Ahmad P, Mansour TA & Siqueira WL (2022) Current State and Challenges of the
961 Global Outcomes of Dental Caries Research in the Meta-Omics Era. *Front Cell Infect*
962 *Microbiol* **12**: 887907.
963 Muir P, Li S, Lou S*, et al.* (2016) The real cost of sequencing: scaling computation to keep pace
964 with data generation. *Genome Biol* **17**: 53.
965 Murray CS, Gao Y & Wu M (2021) Re-evaluating the evidence for a universal genetic boundary
966 among microbial species. *Nat Commun* **12**: 4059.
967 Naito M, Ogura Y, Itoh T, Shoji M, Okamoto M, Hayashi T & Nakayama K (2016) The complete
968 genome sequencing of Prevotella intermedia strain OMA14 and a subsequent fine-scale,
969 intra-species genomic comparison reveal an unusual amplification of conjugative and
970 mobile transposons and identify a novel Prevotella-lineage-specific repeat. *DNA Res* **23**:
971 11-19.
972 Nakamura T, Yamada KD, Tomii K & Katoh K (2018) Parallelization of MAFFT for large-scale
973 multiple sequence alignments. *Bioinformatics* **34**: 2490-2492.
974 Nguyen LT, Schmidt HA, von Haeseler A & Minh BQ (2015) IQ-TREE: a fast and effective
975 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**:
976 268-274.
977 Nguyen T, Sedghi L, Ganther S, Malone E, Kamarajan P & Kapila YL (2020) Host-microbe
978 interactions: Profiles in the transcriptome, the proteome, and the metabolome. *Periodontol*
979 *2000* **82**: 115-128.
980 Ni Z, Wolk M, Jukes G*, et al.* (2022) Guiding the choice of informatics software and tools for
981 lipidomics research applications. *Nat Methods*.
982 Nouioui I, Carro L, Garcia-Lopez M, Meier-Kolthoff JP, Woyke T, Kyrpides NC, Pukall R, Klenk
983 HP, Goodfellow M & Goker M (2018) Genome-Based Taxonomic Classification of the
984 Phylum Actinobacteria. *Front Microbiol* **9**: 2007.
985 Nowicki EM, Shroff R, Singleton JA*, et al.* (2018) Microbiota and Metatranscriptome Changes
986 Accompanying the Onset of Gingivitis. *mBio* **9**.
987 Nurk S, Meleshko D, Korobeynikov A & Pevzner PA (2017) metaSPAdes: a new versatile
988 metagenomic assembler. *Genome Res* **27**: 824-834.
989 Overmyer KA, Rhoads TW, Merrill AE, Ye Z, Westphall MS, Acharya A, Shukla SK & Coon JJ
990 (2021) Proteomics, Lipidomics, Metabolomics, and 16S DNA Sequencing of Dental
991 Plaque From Patients With Diabetes and Periodontal Disease. *Mol Cell Proteomics* **20**:
992 100126.
993 Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane
994 JA & Parkhill J (2015) Roary: rapid large-scale prokaryote pan genome analysis.
995 *Bioinformatics* **31**: 3691-3693.
996 Palmer SR, Miller JH, Abranches J, Zeng L, Lefebure T, Richards VP, Lemos JA, Stanhope MJ
997 & Burne RA (2013) Phenotypic heterogeneity of genomically-diverse isolates of
998 Streptococcus mutans. *PLoS One* **8**: e61358.
999 Pasolli E, Asnicar F, Manara S*, et al.* (2019) Extensive Unexplored Human Microbiome Diversity
1000 Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and
1001 Lifestyle. *Cell* **176**: 649-662 e620.
1002 Perkel JM (2021) Single-cell proteomics takes centre stage. *Nature* **597**: 580-582.
1003 Peterson SN, Meissner T, Su AI, Snesrud E, Ong AC, Schork NJ & Bretz WA (2014) Functional
1004 expression of dental plaque microbiota. *Front Cell Infect Microbiol* **4**: 108.

Pitt ME, Nguyen SH, Duarte TPS, Teng H, Blaskovich MAT, Cooper MA & Coin LJM (2020) Evaluating the genome and resistome of extensively drug-resistant Klebsiella pneumoniae using native DNA and RNA Nanopore sequencing. *Gigascience* **9**.

Price MN, Dehal PS & Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641-1650.

Prjibelski A, Antipov D, Meleshko D, Lapidus A & Korobeynikov A (2020) Using SPAdes De Novo Assembler. *Curr Protoc Bioinformatics* **70**: e102.

Radaic A & Kapila YL (2021) The oralome and its dysbiosis: New insights into oral microbiome-host interactions. *Comput Struct Biotechnol J* **19**: 1335-1360.

Roberts RJ, Carneiro MO & Schatz MC (2013) The advantages of SMRT sequencing. *Genome Biol* **14**: 405.

Schmid R, Petras D, Nothias LF*, et al.* (2021) Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat Commun* **12**: 3832.

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS & Huttenhower C (2011) Metagenomic biomarker discovery and explanation. *Genome Biol* **12**: R60.

Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sorensen EA, Wollenberg RD & Albertsen M (2022) Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods* **19**: 823-826.

Shaiber A, Willis AD, Delmont TO*, et al.* (2020) Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol* **21**: 292.

Shi B, Chang M, Martin J, Mitreva M, Lux R, Klokkevold P, Sodergren E, Weinstock GM, Haake SK & Li H (2015) Dynamic changes in the subgingival microbiome and their potential for diagnosis and prognosis of periodontitis. *mBio* **6**: e01926-01914.

Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG & Banfield JF (2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* **3**: 836-843.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ & Birol I (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**: 1117-1123.

Sinha D, Sun X, Khare M, Drancourt M, Raoult D & Fournier PE (2021) Pangenome analysis and virulence profiling of Streptococcus intermedius. *BMC Genomics* **22**: 522.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312-1313.

Tajik M, Baharfar M & Donald WA (2022) Single-cell mass spectrometry. *Trends Biotechnol* **40**: 1374-1392.

Tinder EL, Faustoferri RC, Buckley AA, Quivey RG, Jr. & Baker JL (2022) Analysis of the Streptococcus mutans Proteome during Acid and Oxidative Stress Reveals Modules of Protein Coexpression and an Expanded Role for the TreR Transcriptional Regulator. *mSystems* **7**: e0127221.

Tjaden B (2015) De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biol* **16**: 1.

Torres PJ, Thompson J, McLean JS, Kelley ST & Edlund A (2019) Discovery of a Novel Periodontal Disease-Associated Bacterium. *Microb Ecol* **77**: 267-276.

Treerat P, McGuire B, Palmer E, Dahl EM, Karstens L, Merritt J & Kreth J (2022) Oral microbiome diversity: The curious case of Corynebacterium sp. isolation. *Mol Oral Microbiol* **37**: 167-179.

Tsao SW, Tsang CM & Lo KW (2017) Epstein-Barr virus infection and nasopharyngeal carcinoma. *Philos Trans R Soc Lond B Biol Sci* **372**.

Uritskiy GV, DiRuggiero J & Taylor J (2018) MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**: 158.

Utter DR, Borisy GG, Eren AM, Cavanaugh CM & Mark Welch JL (2020) Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biol* **21**: 293.
van der Walt AJ, van Goethem MW, Ramond JB, Makhalanyane TP, Reva O & Cowan DA (2017) Assembling metagenomes, one community at a time. *BMC Genomics* **18**: 521.
Vaser R, Sovic I, Nagarajan N & Sikic M (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737-746.
Velsko IM, Chakraborty B, Nascimento MM, Burne RA & Richards VP (2018) Species Designations Belie Phenotypic and Genotypic Heterogeneity in Oral Streptococci. *mSystems* **3**.
Venter JC, Remington K, Heidelberg JF*, et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.
Walker AR & Shields RC (2022) Investigating CRISPR spacer targets and their impact on genomic diversification of Streptococcus mutans. *Front Genet* **13**: 997341.
Walker BJ, Abeel T, Shea T*, et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.
Wang M & Carver JJ & Phelan VV*, et al.* (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**: 828-837.
Wang O, Chin R, Cheng X*, et al.* (2019) Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res* **29**: 798-808.
Wang Y, Zhao Y, Bollas A, Wang Y & Au KF (2021) Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**: 1348-1365.
Watson M & Warr A (2019) Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol* **37**: 124-126.
Wenger AM, Peluso P, Rowell WJ*, et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**: 1155-1162.
Wetterstrand KA DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).
White LK & Hesselberth JR (2022) Modification mapping by nanopore sequencing. *Front Genet* **13**: 1037134.
Wick RR & Holt KE (2022) Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS Comput Biol* **18**: e1009802.
Wick RR, Judd LM, Gorrie CL & Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**: e1005595.
Wick RR, Judd LM, Cerdeira LT, Hawkey J, Meric G, Vezina B, Wyres KL & Holt KE (2021) Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol* **22**: 266.
Wickramarachchi A & Lin Y (2022) Binning long reads in metagenomics datasets using composition and coverage information. *Algorithms Mol Biol* **17**: 14.
Wilbanks EG, Dore H, Ashby MH, Heiner C, Roberts RJ & Eisen JA (2022) Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity. *ISME J* **16**: 1921-1931.
Williams DW, Greenwell-Wild T, Brenchley L*, et al.* (2021) Human oral mucosa cell atlas reveals a stromal-neutrophil axis regulating tissue immunity. *Cell* **184**: 4090-4104 e4015.
Wu YW, Simmons BA & Singer SW (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**: 605-607.
Yahara K, Suzuki M, Hirabayashi A, Suda W, Hattori M, Suzuki Y & Okazaki Y (2021) Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria. *Nat Commun* **12**: 27.

Yost S, Duran-Pinedo AE, Teles R, Krishnan K & Frias-Lopez J (2015) Functional signatures of oral dysbiosis during periodontitis progression revealed by microbial metatranscriptome analysis. *Genome Med* **7**: 27.

Zerbino DR & Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821-829.

Zheng J, Wittouck S, Salvetti E*, et al.* (2020) A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. *Int J Syst Evol Microbiol* **70**: 2782-2858.

Zhu J, Tian L, Chen P*, et al.* (2022) Over 50,000 Metagenomically Assembled Draft Genomes for the Human Oral Microbiome Reveal New Taxa. *Genomics Proteomics Bioinformatics* **20**: 246-259.

Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL & Yorke JA (2013) The MaSuRCA genome assembler. *Bioinformatics* **29**: 2669-2677.