FEMS Microbiology Reviews, 2023, **47**, 1–15
DOI: 10.1093/femsre/fuad051
Advance access publication date: 4 September 2023
Review Article

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 [1,2,3,*]

[1]Department of Oral Rehabilitation & Biosciences, School of Dentistry, Oregon Health & Science University, 3181 Sam Jackson Park Road, Portland, OR 97202, United States
[2]Genomic Medicine Group, J. Craig Venter Institute, La Jolla, CA 92037, United States
[3]Department of Pediatrics, UC San Diego School of Medicine, La Jolla, CA 92093, United States
*Corresponding author. Department of Oral Rehabilitation & Biosciences, School of Dentistry, Oregon Health & Science University, 3181 Sam Jackson Park Road, Portland, OR 97202, United States. E-mail: bakerjo@ohsu.edu
Editor: [Dennis Nielsen]

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 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.

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

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 and 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 as obesity, diabetes, cardiovascular disease, inflammatory bowel disease, non-alcoholic fatty liver disease, rheumatoid arthritis, colorectal cancers, and Alzheimer's disease (Hajishengallis and 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 that 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 (MS), the 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. *S. gordonii* and *S. 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.

Received 31 January 2023; **revised** 2 August 2023; **accepted** 1 September 2023

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, MS, 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 MS. 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, multiomics 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 the 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 use of omics techniques in both dental caries (Bostanci et al. 2021, Moussa et al. 2022) and periodontal disease (Nguyen et al. 2020, Bostanci et al. 2021, Kumar et al. 2021) research, this review will focus more on the omics techniques and tools themselves, including historical context and the 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. This review will also go into more depth on sequencing-based omics rather than MS-based omics, mainly because the former approaches have been more extensively employed by the field of microbiome research.

Sequencing-based omics

Historical background: next-generation sequencing (NGS) revolutionizes the life sciences and enables early microbiome research in the 2000s and early 2010s

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 MS, 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 that 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 and Russell 2015), and SPAdes (Prjibelski et al. 2020) algorithms and 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 2023, 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

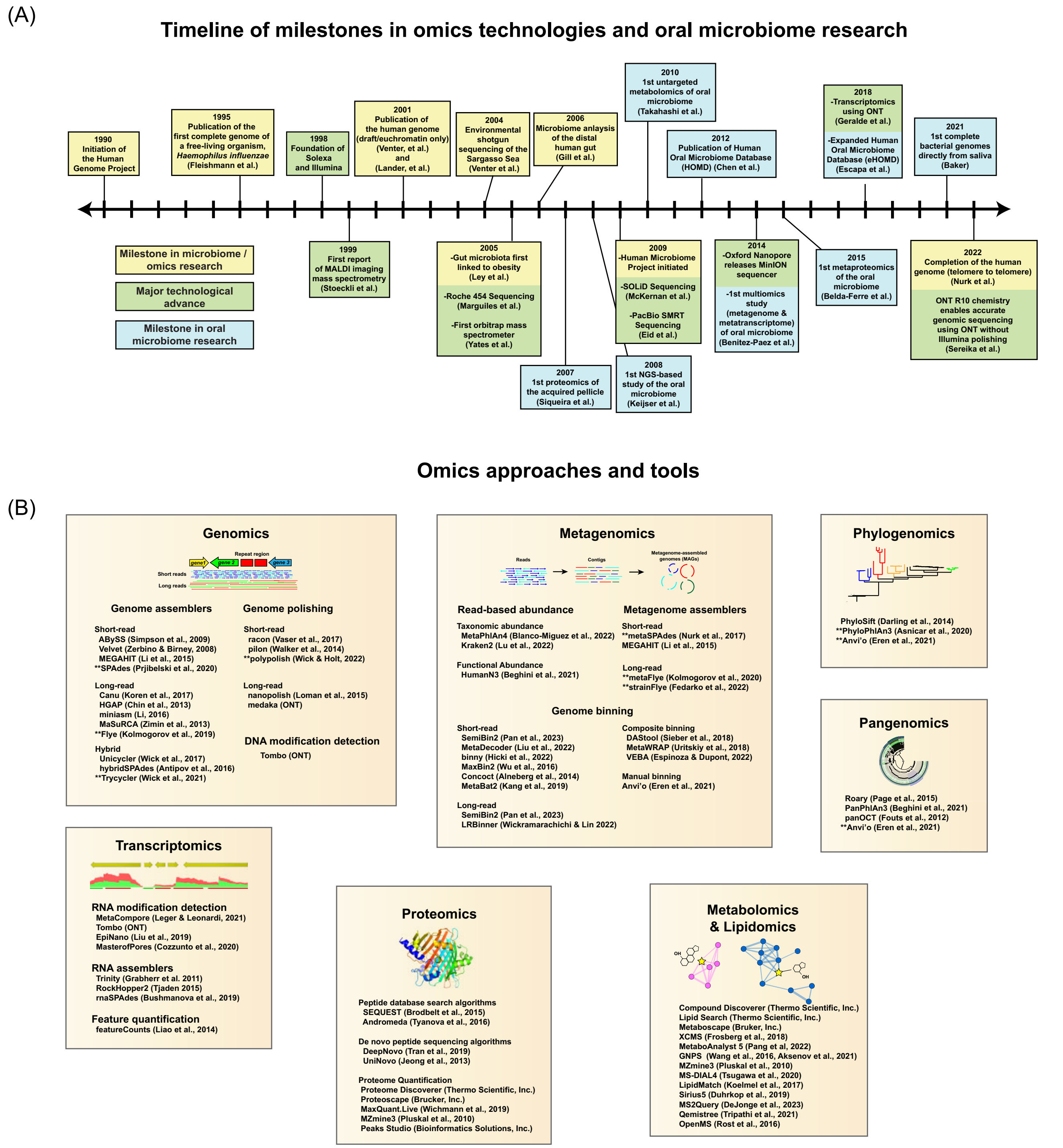
(A)
Timeline of milestones in omics technologies and oral microbiome research
1990
Initiation of the Human Genome Project
1995
Publication of the first complete genome of a free-living organism, *Haemophilus influenzae* (Fleishmann et al.)
1998
Foundation of Solexa and Illumina
1999
First report of MALDI imaging mass spectrometry (Stoeckli et al.)
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.)
2005
-Gut microbiota first linked to obesity (Ley et al.)
-Roche 454 Sequencing (Marguiles et al.)
-First orbitrap mass spectrometer (Yates et al.)
2006
Microbiome anlaysis of the distal human gut (Gill 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.)
2010
1st untargeted metabolomics of oral microbiome (Takahashi et al.)
2012
Publication of Human Oral Microbiome Database (HOMD) (Chen 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.)
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)
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.)
Milestone in microbiome / omics research
Major technological advance
Milestone in oral microbiome research
Omics approaches and tools
(B)
Genomics
Repeat region
gene1
gene 2
gene 3
Short reads
Long reads
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
Reads
Contigs
Metagenome-assembled genomes (MAGs)
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
SemiBin2 (Pan et al., 2023)
MetaDecoder (Liu et al., 2022)
binny (Hicki et al., 2022)
MaxBin2 (Wu et al., 2016)
Concoct (Alneberg et al., 2014)
MetaBat2 (Kang et al., 2019)
Long-read
SemiBin2 (Pan et al., 2023)
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 database search algorithms
SEQUEST (Brodbelt et al., 2015)
Andromeda (Tyanova et al., 2016)
De novo peptide sequencing algorithms
DeepNovo (Tran et al., 2019)
UniNovo (Jeong et al., 2013)
Proteome Quantification
Proteome Discoverer (Thermo Scientific, Inc.)
Proteoscape (Brucker, Inc.)
MaxQuant.Live (Wichmann et al., 2019)
MZmine3 (Pluskal et al., 2010)
Peaks Studio (Bioinformatics Solutions, Inc.)
Metabolomics & Lipidomics
OH
OH
Compound Discoverer (Thermo Scientific, Inc.)
Lipid Search (Thermo Scientific, Inc.)
Metaboscape (Bruker, Inc.)
XCMS (Frosberg et al., 2018)
MetaboAnalyst 5 (Pang et al, 2022)
GNPS (Wang et al., 2016, Aksenov et al., 2021)
MZmine3 (Pluskal et al., 2010)
MS-DIAL4 (Tsugawa et al., 2020)
LipidMatch (Koelmel et al., 2017)
Sirius5 (Duhrkop et al., 2019)
MS2Query (DeJonge et al., 2023)
Qemistree (Tripathi et al., 2021)
OpenMS (Rost et al., 2016)

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 seven 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.

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 the 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 the 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 and 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 reaching an inflection point where accurate genomics and metagenomics can be performed using ONT sequencing alone (Faulk 2022, Liu et al. 2022b, 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) and can detect methylation, enabling genome-wide epigenetic studies (Beaulaurier et al. 2018). In addition to Oxford Nanopore and PacBio, newcomers in the sequencing space, such as Element Biosciences (developing both innovations to short-read sequencing and long read LoopSeq) and Stratos Genomics (now owned by Roche, developing sequencing-by-expansion technologies), may indeed further disrupt the industry. In addition, synthetic long-read and linked-read approaches, such as TELL-seq, use labeling of short-reads adjacent on the genome to obtain contiguity of short-read-based assemblies similar to those obtained through long-read-based approaches (Wang et al. 2019). However, limitations, including incompatibility with metagenomic assemblies, continue to limit widespread use of these approaches (Wang et al. 2019).

Genomics

Figure 1B provides a list of bioinformatics tools, and their references, that are discussed in the following sections. Obtaining genomes that are both complete (i.e. contiguous chromosomes and plasmids) and accurate is of prime importance to microbiology research. High-quality, complete genomes (assuming they are publicly available to researchers in a database) enable: (1) accurate detection and quantification of a particular taxon, or its RNA transcripts, in an isolate or microbiome sample (Venter et al. 2004), (2) prediction of the metabolic pathways and therefore possible ecological and pathogenic roles of the taxon—particularly important for taxa that have not yet been isolated or cultivated (Naito et al. 2016), and (3) guiding wet-lab research, such as mutagenesis. It is important to recognize that many genomes in public repositories were assembled using short-read sequencing only, meaning that they are probably at an incomplete or draft stage and fragmented into contigs of various numbers and sizes. These genomes are likely to be missing sequences and may contain contaminant contigs. Therefore, it is crucial that researchers are cognizant of the limitations inherent with these assemblies if they are used as a reference.

To obtain a genome, sequencing reads that have passed quality control must be assembled. Note that the 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 and 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 necessitate 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 and 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 to get diversity metrics and abundance information on the taxa present. This can be done on unassembled reads using tools like MetaPhlAn4 (https://huttenhower.sph.harvard.edu/metaphlan/), [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 analyses 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 the 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 mainstream binning programs include MaxBin2 (Wu et al. 2016), Concoct (Alneberg et al. 2014), and MetaBat2 (Kang et al. 2019), along with high-performance newcomers such as SemiBin2 (Pan et al. 2023), Binny (Hickl et al. 2022), and MetaDecoder (Liu et al. 2022a). 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 and Dupont 2022). Manual bin inspection and refinement should be performed, where possible, and have been made much easier by the Anvi'o suite of microbiome analysis programs (Chen et al. 2020a, Eren et al. 2021). There are far fewer tools to perform binning on long-read datasets, with SemiBin2 (has algorithms for both short- and long-read binning) (Pan et al. 2023) and LRBinner being the most comprehensive and recently developed (Wickramarachchi and 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 (Chen et al. 2020b, 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 the evolution of gene regulation or when reconstructing evolutionary relationships over shorter time scales. However, the 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 the 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 of "species" in bacteria is not one with universally accepted traits. For the sake of ease when dealing examining massive numbers of MAGs, 95% average nucleotide identity (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 the 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 the phylogenetic data obtained while minimizing the 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 and 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, the 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 and Hesselberth 2022).

In addition to the many free and open-source sequencing-based bioinformatics tools mentioned above, it is worth mentioning that there are also several comprehensive software suites available from vendors, such as Geneious Prime (Dotmatics, Inc.) and CLC Genomics Workbench (Qiagen, Inc.), that can do many types of the above sequencing analyses in a user-friendly graphical user interface (GUI) format (in many cases using the aforementioned individual bioinformatics tools "under the hood"), which may benefit end users with limited experience with Linux/command line-based computing skills.

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 toward 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% and 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 that 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 that 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 radiation (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 *S. 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 the 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 *S. mutans* (Baker et al. 2022). Additional contemporary comparative genomics of *S. mutans* and *S. sobriunus* indicated a lack of phylogeographic differentiation for *S. mutans* but some for *S. sobrinus* (Achtman and Zhou 2020). Another recent study used an *S. mutans* pangenome to examine CRISPR spacers (Walker and 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 intraspecies 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 *S. mitis, S. oralis,* and *S. 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 *S. 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 the repertoire of immunoglobulin genes are likely to contribute to an 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.

MS-based omics

In addition to all the advances described above, which are dependent on nucleic acid sequencing, there have also been major improvements to MS-based omics analyses over the last decade. Recent innovations have made MS analyses significantly more sensitive, accurate, high-throughput, and able to detect a wider range of molecules. These have occurred via advancements at every stage of the analysis pipeline: sample preparation, ionization, separation, mass detection, and data analysis (Shuken 2023). Readers are pointed to an in-depth recent review for biological MS, broadly (Pade et al. 2023). MS-based analyses are typically either untargeted, measuring all possible analytes detectable with the given workflow, or targeted, where analysis is tailored to molecules with specific characteristics such as molecular weight and charge. Simultaneous quantitation and discovery (SQUAD) analysis, recently developed at Thermo Fisher Scientific, combines both targeted and untargeted workflows into a single-injection protocol, combining the strengths of each approach (Amer et al. 2023).

Imaging mass spectrometry (IMS) includes techniques such as matrix assisted laser desorption/ionization (MALDI), time-of-flight secondary ion mass spectrometry (TOF-SIMS), and electrospray-based desorption (DESI), which are utilized to visualize the spatial distribution and biogeography of analytes, and is at this point a mature field worthy of its own review (Chen et al. 2020a). MALDI imaging, in particular, has been widely applied to rapid clinical and diagnostic microbiology (Croxatto et al. 2012, Jang and Kim 2018). Tools such as PySM (Palmer et al. 2017), MSiReader (Nurk et al. 2017), and Ili (Protsyuk et al. 2018) have been developed to analyze and visualize IMS data.

Like sequencing-based omics, where both individual open-source tools and comprehensive vendor software suites are available, MS-based omics has individual tools as well as comprehensive software suites from vendors are options, with prime examples being Proteome Discoverer, Compound Discoverer, and Lipid Search from Thermo Fisher Scientific, and Proteoscape and Metaboscape from Brucker. Although proteomics, metabolomics, and lipidomics represent a more complete and “current” state of a given sample (i.e. rather than what is encoded for by DNA or soon-to-be translated RNA), there are unique challenges facing these omics approaches.

Proteomics

Proteins are typically higher molecular weight and more complex than metabolomic or lipidomic analytes; however, the relative wealth of proteome database data via translated RNA sequences, combined with the fact that proteins themselves are “sequences” from a finite pool of amino acids, makes proteomics datasets somewhat easier to annotate than untargeted metabolomics and lipidomics datasets. The current state of the proteomics field, including current approaches and challenges was excellently summarized recently (Shuken 2023), and bioinformatics tools for proteomics, recently comprehensively reviewed (Chen et al. 2020a). Briefly, proteomics is typically conducted with either a "bottom up" approach, which breaks proteins down into peptides prior to MS analysis, or a "top down" approach, which analyzes whole, native proteins to detect discrete proteoforms and chemical modifications (Donnelly et al. 2019). Peptide sequences are either queried in a database using an algorithm such as the workhorses SEQUEST (Eng et al. 1994) or Andromeda (Tyanova et al. 2016) or sequenced de novo using tools such as UniNovo (Jeong et al. 2013) or DeepNovo (Tran et al. 2017). Proteomics approaches can also be divided into data dependent analysis (DDA) and data-independent analysis (DIA) methods. DDA selects the most abundant peptides in each peak of the MS1 scan for the MS2 scan. Meanwhile, in DIA, ions are continuously collected and fragmented by collecting MS2 scans in overlapping m/z windows, thereby producing a complete record of all peptides in a sample (Xin et al. 2022). PEAKS Studio is a software that leverages all three techniques (peptide search, spectral library search, and de novo sequencing) (Xin et al. 2022). Proteins can be quantified using the area under the MS1 chromatogram (i.e. label-free quantification [LFQ]), which is somewhat problematic due to the compositional nature of sample-to-sample MS data, as discussed below. Alternatively, various labeling techniques such as stable isolate labeling by amino acids in cell culture (SILAC) or tandem mass tags (TMT) improve quantitative sample-to-sample reproducibility and increase throughput by allowing for multiplexing (Shuken 2023). In addition to the vendor suites of software, the MaxQuant (Wichmann et al. 2019) family of tools and MZmine 3 (Schmid et al. 2023) are free and open-source software able to perform various proteomics quantification workflows (Wichmann et al. 2019). Going forward, the application of machine learning and will further improve the sensitivity and dynamic range of proteomics via the implementation of deep learning-based spectral prediction and spectrum-centric DIA analysis (Zeng et al. 2022, Cox 2023, Neely et al. 2023).

Metabolomics and lipidomics

Several recent reviews have provided metabolomics best practices guidelines (Alseekh et al. 2021) and summarized lipidomics informatics (Ni et al. 2022), metabolomics/lipidomics separation methods (Harrieder et al. 2022), metabolite discovery (Giera et al. 2022), and the specific application of metabolomics in microbiome data (Bauermeister et al. 2022) in more depth than is provided here. In addition to vendor-specific tools such as Compound Discoverer, Lipid Search, and Metaboscape, a wealth of alternative tools (many of which are free and/or open source) are available for metabolomics analysis. These include platforms such as MetaboAnalyst 5.0 (Pang et al. 2022) and XCMS Online (Forsberg et al. 2018), which are web-based GUIs, as well as MZmine3 (Schmid et al. 2023), an open source tool to examine raw spectral files and perform custom downstream analysis, MS-DIAL (Tsugawa et al. 2020), and OpenMS (Rost et al. 2016). Unlike proteomics, metabolomics, and lipidomics data do not generate "sequences," with the molecules being detected occupying a comparatively unlimited chemical space. Furthermore, many of the databases used for dereplication (i.e. identification of known compounds) are not freely available. As a result, a much higher percentage of the features detected in metabolomics and lipidomics datasets are unknown, with annotation rates <10% routine (de Jonge et al. 2022). *In silico* analyses such as molecular networking and machine learning-based annotation have been instrumental in beginning to address this challenge (de Jonge et al. 2022). Molecular networking is a visualization of spectral alignment and correlation, which enables the prediction of the chemical structure of unknown features. One such landmark tool is the Global Natural Products Social Molecular Networking (GNPS), first published in 2016, created an open-access knowledge base for organizations and enabled the sharing of MS data, which is reanalyzed as the database grows, leveraging molecular networking to help identify novel spectra (Wang et al. 2016). The GNPS led to the development of a host of integrated analysis tools to further improve annotation and analysis. The first iteration of the GNPS utilized MS-MS data exclusively, while a subsequent improvement deploys "feature-based molecular networking," an approach that combines quantitative chromatographic peak data with qualitative MS/MS data (Nothias et al. 2020). Originally developed for liquid chromatography MS (LC-MS), the GNPS was also recently updated to enable the analysis of gas chromatography MS (GC-MS), which expands its utility to many GC-MS-based lipidomics and metabolomics analyses (Aksenov et al. 2021). Other recent innovations to MS-based omics include the use of metadata to enhance annotation of metabolomics (Gauglitz et al. 2022), native spray metal metabolomics to identify novel 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). MS2Query is another recently developed tool that utilizes machine learning to identify potential analogs to unknown spectra (de Jonge et al. 2023). SIRIUS5 (Duhrkop et al. 2019) predicts the chemical formula and molecular structure of query compounds, while Qemistree is a data exploration strategy using hierarchical organization of molecular fingerprints to visualize molecular relationships as a tree, enabling the use of many further analysis tools originally designed to analyze and visualize the relatedness of DNA, such as QIIME2 (Tripathi et al. 2021). Efforts to standardize MS data and databases, such as PeakForest (Paulhe et al. 2022) and ChemFONT (Wishart et al. 2023), seek to address that major issues of data reporting and reproducibility facing the MS field (Alseekh et al. 2021). Finally, MASST is a search tool that enables uses to query spectra against all small molecule tandem-MS data in public repositories, similar to how users can query NCBI-BLAST for the source of DNA sequences (Wang et al. 2020). Going forward, these advancements in MS analysis methods are poised to increase the scale and pace of discovery in the oral microbiota and lead to novel approaches to benefit human oral health.

Impact of MS-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 and 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 and 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 MS-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 is 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. 2019b), 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 ALDEx2 (Fernandes et al. 2014), ANCOM (Mandal et al. 2015), and Songbird (Morton et al. 2019b); 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 a transformational technology, allowing for the omics analysis of individual cells and the identification of discrete biological dynamics that are obscured by the averages obtained by traditional bulk analysis. Single-cell proteomics, lipidomics, and metabolomics, based on MS, 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 and 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 also scRNA-seq to examine the 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. 2019a). mmvec, a recently developed approach for analyzing multi-omics data, uses co-occurrence probabilities rather than correlations (Morton et al. 2019a). When applied to metagenome and metabolome data, it allowed researchers to identify the most likely microbe-metabolite interactions (Morton et al. 2019a). Another tool, iNetModels2, was recently developed for interactively visualizing multi-omics data (Arif et al. 2021). A recent review also comprehensively discussed tools for proteomics-centric multi-omics analyses (Rajczewski et al. 2022).

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 the pH of the culture and subsequent microecology (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 reanalysis 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 the 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 MS 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 the 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 the 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 MS 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 the 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.

Acknowledgments

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

Conflict of interest: The author declares no conflicts of interest.

References

Achtman M, Zhou Z. 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* 2020;**375**:20190573.

Aksenov AA, Laponogov I, Zhang Z *et al.* Auto-deconvolution and molecular networking of gas chromatography-mass spectrometry data. *Nat Biotechnol* 2021;**39**:169–73.

Al-Hebshi NN, Baraniya D, Chen T *et al.* Metagenome sequencing-based strain-level and functional characterization of supragingival microbiome associated with dental caries in children. *J Oral Microbiol* 2019;**11**:1557986.

Alneberg J, Bjarnason BS, de Bruijn I *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014;**11**:1144–6.

Alseekh S, Aharoni A, Brotman Y *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat Methods* 2021;**18**:747–56.

Amarasinghe SL, Su S, Dong X *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020;**21**:30.

Amer B, Deshpande RR, Bird SS. Simultaneous quantitation and discovery (SQUAD) analysis: combining the best of targeted and untargeted mass spectrometry-based metabolomics. *Metabolites* 2023;**13**:648.

Ansbro K, Wade WG, Stafford GP. *Tannerella serpentiformis* sp. nov., isolated from the human mouth. *Int J Syst Evol Microbiol* 2020;**70**:3749–54.

Antipov D, Korobeynikov A, McLean JS *et al.* hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 2016;**32**:1009–15.

Arif M, Zhang C, Li X *et al.* iNetModels 2.0: an Interactive visualization and database of multi-omics data. *Nucleic Acids Res* 2021;**49**:W271–6.

Aron AT, Petras D, Schmid R *et al.* Native mass spectrometry-based metabolomics identifies metal-binding compounds. *Nat Chem* 2022;**14**:100–9.

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

Athanasopoulou K, Boti MA, Adamopoulos PG *et al.* Third-generation sequencing: the spearhead towards the radical transformation of modern genomics. *Life (Basel)* 2021;**12**:30.

Baker JL, Bor B, Agnello M *et al.* Ecology of the oral microbiome: beyond bacteria. *Trends Microbiol* 2017;**25**:362–74.

Baker JL, Morton JT, Dinis M *et al.* Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. *Genome Res* 2021;**31**:64–74.

Baker JL, Tang X, LaBonte S *et al. mucG*, *mucH*, and *mucI* modulate production of mutanocyclin and reutericyclins in *Streptococcus mutans* B04Sm5. *J Bacteriol* 2022;**204**:e0004222.

Baker JL. Complete genomes of clade G6 *Saccharibacteria* suggest a divergent ecological niche and lifestyle. *mSphere* 2021;**6**:e0053021.

Baker JL. Using nanopore sequencing to obtain complete bacterial genomes from saliva samples. *Msystems* 2022;**7**:e0049122.

Balachandran M, Cross KL, Podar M. Single-cell genomics and the oral microbiome. *J Dent Res* 2020;**99**:613–20.

Bao K, Li X, Poveda L *et al.* Proteome and microbiome mapping of Human gingival tissue in health and disease. *Front Cell Infect Microbiol* 2020;**10**:588155.

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

Bauermeister A, Mannochio-Russo H, Costa-Lotufo LV *et al.* Mass spectrometry-based metabolomics in microbiome investigations. *Nat Rev Micro* 2022;**20**:143–60.

Beall CJ, Campbell AG, Dayeh DM *et al.* Single cell genomics of uncultured, health-associated *Tannerella* BU063 (Oral Taxon 286) and comparison to the closely related pathogen *Tannerella forsythia*. *PLoS One* 2014;**9**:e89398.

Beaulaurier J, Zhu S, Deikus G *et al.* Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* 2018;**36**:61–9.

Beghini F, McIver LJ, Blanco-Miguez A *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* 2021;**10**:e65088.

Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R *et al.* The oral metagenome in health and disease. *ISME J* 2012;**6**:46–56.

Belda-Ferre P, Williamson J, Simon-Soro A *et al.* The human oral metaproteome reveals potential biomarkers for caries disease. *Proteomics* 2015;**15**:3497–507.

Belstrom D, Constancias F, Liu Y *et al.* Metagenomic and metatranscriptomic analysis of saliva reveals disease-associated microbiota in patients with periodontitis and dental caries. *npj Biofilms Microbiomes* 2017;**3**:23.

Bennett S. Solexa Ltd. *Pharmacogenomics* 2004;**5**:433–8.

Bentley DR, Balasubramanian S, Swerdlow HP *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;**456**:53–9.

Blattman SB, Jiang W, Oikonomou P *et al.* Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. *Nat Microbiol* 2020;**5**:1192–201.

Bolyen E, Rideout JR, Dillon MR *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019;**37**:852–7.

Bostanci N, Grant M, Bao K *et al.* Metaproteome and metabolome of oral microbial communities. *Periodontol 2000* 2021;**85**:46–81.

Bowen WH, Burne RA, Wu H *et al.* Oral biofilms: pathogens, matrix, and polymicrobial interactions in microenvironments. *Trends Microbiol* 2018;**26**:229–42.

Brennan MA, Rosenthal AZ. Single-cell RNA sequencing elucidates the structure and organization of microbial communities. *Front Microbiol* 2021;**12**:713128.

Brodbelt JS, Russell DH. Focus on the 20-year anniversary of SEQUEST. *J Am Soc Mass Spectrom* 2015;**26**:1797–8.

Bushmanova E, Antipov D, Lapidus A *et al.* rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-seq data. *Gigascience* 2019;**8**.

Caetano AJ, D'Agostino EM, Sharpe P *et al.* Expression of periodontitis susceptibility genes in human gingiva using single-cell RNA sequencing. *J Periodontal Res* 2022;**57**:1210–8.

Callahan BJ, Grinevich D, Thakur S *et al.* Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome* 2021;**9**:130.

Callahan BJ, McMurdie PJ, Rosen MJ *et al.* DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;**13**:581–3.

Camelo-Castillo A, Benitez-Paez A, Belda-Ferre P *et al. Streptococcus dentisani* sp. nov., a novel member of the mitis group. *Int J Syst Evol Microbiol* 2014;**64**:60–5.

Campbell AG, Campbell JH, Schwientek P *et al.* Multiple single-cell genomes provide insight into functions of uncultured *Deltaproteobacteria* in the human oral cavity. *PLoS One* 2013;**8**:e59361.

Campbell AG, Schwientek P, Vishnivetskaya T *et al.* Diversity and genomic insights into the uncultured *Chloroflexi* from the human microbiota. *Environ Microbiol* 2014;**16**:2635–43.

Chen C, Hou J, Tanner JJ *et al.* Bioinformatics methods for mass spectrometry-based proteomics data analysis. *Int J Mol Sci* 2020a;**21**:2873.

Chen LX, Anantharaman K, Shaiber A *et al.* Accurate and complete genomes from metagenomes. *Genome Res* 2020b;**30**:315–33.

Chen P, Wu H, Yao H *et al.* Multi-omics analysis reveals the systematic relationship between oral homeostasis and chronic sleep deprivation in rats. *Front Immunol* 2022;**13**:847132.

Chen T, Yu WH, Izard J *et al.* The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)* 2010;**2010**:baq013.

Chin CS, Alexander DH, Marks P *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013;**10**:563–9.

Cornejo OE, Lefebure T, Bitar PD *et al.* Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. *Mol Biol Evol* 2013;**30**:881–93.

Couvillion SP, Zhu Y, Nagy G *et al.* New mass spectrometry technologies contributing towards comprehensive and high throughput omics analyses of single cells. *Analyst* 2019;**144**:794–807.

Cox J. Prediction of peptide mass spectral libraries with machine learning. *Nat Biotechnol* 2023;**41**:33–43.

Cross KL, Campbell JH, Balachandran M *et al.* Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat Biotechnol* 2019;**37**:1314–21.

Croxatto A, Prod'hom G, Greub G. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiol Rev* 2012;**36**:380–407.

Cusco A, Perez D, Vines J *et al.* Long-read metagenomics retrieves complete single-contig bacterial genomes from canine feces. *BMC Genomics* 2021;**22**:330.

Darling AE, Jospin G, Lowe E *et al.* PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2014;**2**:e243.

Dawes C, Wong DTW. Role of saliva and salivary diagnostics in the advancement of oral health. *J Dent Res* 2019;**98**:133–41.

de Jonge NF, Louwen JJR, Chekmeneva E *et al.* MS2Query: reliable and scalable MS(2) mass spectra-based analogue search. *Nat Commun* 2023;**14**:1752.

de Jonge NF, Mildau K, Meijer D *et al.* Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics* 2022;**18**:103.

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

Do T, Sheehy EC, Mulli T *et al.* Transcriptomic analysis of three *Veillonella* spp. Present In carious dentine and in the saliva of caries-free individuals. *Front Cell Infect Microbiol* 2015;**5**:25.

Donnelly DP, Rawlins CM, DeHart CJ *et al.* Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat Methods* 2019;**16**:587–94.

Douglas GM, Maffei VJ, Zaneveld JR *et al.* PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol* 2020;**38**:685–8.

Duhrkop K, Fleischauer M, Ludwig M *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* 2019;**16**:299–302.

Duran-Pinedo AE, Chen T, Teles R *et al.* Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. *ISME J* 2014;**8**:1659–72.

Duran-Pinedo AE, Solbiati J, Teles F *et al.* Subgingival host-microbiome metatranscriptomic changes following scaling and root planing in grade II/III periodontitis. *J Clin Periodontol* 2022.**50**:316–330.

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

Ebersole JL, Nagarajan R, Kirakodu S *et al.* Oral microbiome and gingival gene expression of inflammatory biomolecules with aging and periodontitis. *Front Oral Health* 2021;**2**:725115.

Economopoulou P, Kotsantis I, Psyrri A. Special issue about head and neck cancers: HPV positive cancers. *Int J Mol Sci* 2020;**21**:3388.

Edgar RC. MUSCLE: a Multiple sequence alignment method with reduced time and space complexity. *BMC Bioinf* 2004;**5**:113.

Edlund A, Yang Y, Yooseph S *et al.* Meta-omics uncover temporal regulation of pathways across oral microbiome genera during *in vitro* sugar metabolism. *ISME J* 2015;**9**:2605–19.

Edlund A, Yang Y, Yooseph S *et al.* Uncovering complex microbiome activities via metatranscriptomics during 24 hours of oral biofilm assembly and maturation. *Microbiome* 2018;**6**:217.

Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;**5**:976–89.

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

Escapa IF, Chen T, Huang Y *et al.* 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* 2018;**3**:e00187–18.

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

Faulk C. *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* 2022;**51**:17–28.

Fedarko MW, Kolmogorov M, Pevzner PA. Analyzing rare mutations in metagenomes assembled using long and accurate reads. *Genome Res* 2022;**32**:2119–33.

Fernandes AD, Reid JN, Macklaim JM *et al.* Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* 2014;**2**:15.

Flores Ramos S, Brugger SD, Escapa IF *et al.* Genomic stability and genetic defense systems in *Dolosigranulum pigrum*, a candidate beneficial bacterium from the Human microbiome. *mSystems* 2021;**6**:e0042521.

Forsberg EM, Huan T, Rinehart D *et al.* Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. *Nat Protoc* 2018;**13**:633–51.

Fouts DE, Brinkac L, Beck E *et al.* PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res* 2012;**40**:e172.

Fozo EM, Quivey RG. Shifts in the membrane fatty acid profile of *Streptococcus mutans* enhance survival in acidic environments. *Appl Environ Microb* 2004;**70**:929–36.

Garalde DR, Snell EA, Jachimowicz D *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* 2018;**15**:201–6.

Gauglitz JM, West KA, Bittremieux W *et al.* Enhancing untargeted metabolomics using metadata-based source annotation. *Nat Biotechnol* 2022;**40**:1774–9.

Giera M, Yanes O, Siuzdak G. Metabolite discovery: biochemistry's scientific driver. *Cell Metab* 2022;**34**:21–34.

Gill SR, Pop M, Deboy RT *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* 2006;**312**:1355–9.

Gloor GB, Macklaim JM, Pawlowsky-Glahn V *et al.* Microbiome datasets are compositional: and this is not optional. *Front Microbiol* 2017;**8**:2224.

Gonzalez OA, Kirakodu SS, Nguyen LM *et al.* Gingival transcriptomic patterns of macrophage polarization during initiation, progression, and resolution of periodontitis. *Clin Exp Immunol* 2022;**211**:248–68.

Goussarov G, Mysara M, Vandamme P *et al.* Introduction to the principles and methods underlying the recovery of metagenome-assembled genomes from metagenomic data. *Microbiologyopen* 2022;**11**:e1298.

Grabherr MG, Haas BJ, Yassour M *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol* 2011;**29**:644–52.

Grunberger F, Ferreira-Cerca S, Grohmann D. Nanopore sequencing of RNA and cDNA molecules in *Escherichia coli*. *RNA* 2022;**28**:400–17.

Hajishengallis G, Chavakis T. Local and systemic mechanisms linking periodontal disease and inflammatory comorbidities. *Nat Rev Immunol* 2021;**21**:426–40.

Harrieder EM, Kretschmer F, Bocker S *et al.* Current state-of-the-art of separation methods used in LC-MS based metabolomics and lipidomics. *J Chromatogr B Analyt Technol Biomed Life Sci* 2022;**1188**:123069.

He X, McLean JS, Guo L *et al.* The social structure of microbial community involved in colonization resistance. *ISME J* 2014;**8**:564–74.

Hendrickson EL, Bor B, Kerns KA *et al.* Transcriptome of epibiont *Saccharibacteria Nanosynbacter lyticus* strain TM7x during the establishment of symbiosis. *J Bacteriol* 2022;**204**:e0011222.

Hickl O, Queiros P, Wilmes P *et al.* binny: an automated binning algorithm to recover high-quality genomes from complex metagenomic datasets. *Brief Bioinform* 2022;**23**:bbac431.

Homberger C, Saliba AE, Vogel J. A MATQ-seq-based protocol for single-cell RNA-seq in bacteria. *Methods Mol Biol* 2023;**2584**:105–21.

Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;**486**:207–14.

Imdahl F, Vafadarnejad E, Homberger C *et al.* Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. *Nat Microbiol* 2020;**5**:1202–6.

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

Jain C, Rodriguez RL, Phillippy AM *et al.* High throughput ANI analysis of 90 K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;**9**:5114.

Jain M, Fiddes IT, Miga KH *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat Methods* 2015;**12**:351–6.

Jang KS, Kim YH. Rapid and robust MALDI-TOF MS techniques for microbial identification: a brief overview of their diverse applications. *J Microbiol* 2018;**56**:209–16.

Jeong K, Kim S, Pevzner PA. UniNovo: a universal tool for *de novo* peptide sequencing. *Bioinformatics* 2013;**29**:1953–62.

Jorth P, Turner KH, Gumus P *et al.* Metatranscriptomics of the human oral microbiome during health and disease. *mBio* 2014;**5**:e01012–01014.

Kang DD, Li F, Kirton E *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;**7**:e7359.

Katz KS, Shutov O, Lapoint R *et al.* STAT: a Fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biol* 2021;**22**:270.

Knight R, Vrbanac A, Taylor BC *et al.* Best practices for analysing microbiomes. *Nat Rev Micro* 2018;**16**:410–22.

Kolmogorov M, Bickhart DM, Behsaz B *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020;**17**:1103–10.

Kolmogorov M, Yuan J, Lin Y *et al.* Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;**37**:540–6.

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

Koren S, Walenz BP, Berlin K *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017;**27**:722–36.

Kuchina A, Brettner LM, Paleologu L *et al.* Microbial single-cell RNA sequencing by split-pool barcoding. *Science* 2021;**371**:eaba5257.

Kumar PS, Dabdoub SM, Ganesan SM. Probing periodontal microbial dark matter using metataxonomics and metagenomics. *Periodontol 2000* 2021;**85**:12–27.

Lamont RJ, Koo H, Hajishengallis G. The oral microbiota: dynamic communities and host interactions. *Nat Rev Micro* 2018;**16**:745–59.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.

Lau WW, Hardt M, Zhang YH *et al.* The Human Salivary Proteome Wiki: a community-driven research platform. *J Dent Res* 2021;**100**:1510–9.

Len ACL, Harty DWS, Jacques NA. Proteome analysis of *Streptococcus mutans* metabolic phenotype during acid tolerance. *Microbiology (Reading)* 2004;**150**:1353–66.

Leung SK, Jeffries AR, Castanho I *et al.* Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Rep* 2021;**37**:110022.

Ley RE, Backhed F, Turnbaugh P *et al.* Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA* 2005;**102**:11070–5.

Li D, Liu CM, Luo R *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**:1674–6.

Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* 2016;**32**:2103–10.

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100.

Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 2014;**30**:2843–51.

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30.

Liu CC, Dong SS, Chen JB *et al.* MetaDecoder: a novel method for clustering metagenomic contigs. *Microbiome* 2022a;**10**:46.

Liu L, Yang Y, Deng Y *et al.* Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes. *Microbiome* 2022b;**10**:209.

Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat Methods* 2015;**12**:733–5.

Long H, Yan L, Pu J *et al.* Multi-omics analysis reveals the effects of microbiota on oral homeostasis. *Front Immunol* 2022;**13**:1005992.

Lu J, Rincon N, Wood DE *et al.* Metagenome analysis using the Kraken software suite. *Nat Protoc* 2022;**17**:2815–39.

Mandal S, Van Treuren W, White RA *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* 2015;**26**:27663.

Margulies M, Egholm M, Altman WE *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005;**437**:376–80.

Mashima I, Liao YC, Lin CH *et al.* Comparative pan-genome analysis of oral *Veillonella* species. *Microorganisms* 2021;**9**:1775.

McKernan KJ, Peckham HE, Costa GL *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* 2009;**19**:1527–41.

McLean AR, Torres-Morales J, Dewhirst FE *et al.* Site-tropism of streptococci in the oral microbiome. *Mol Oral Microbiol* 2022;**37**:229–43.

McLean JS, Bor B, Kerns KA *et al.* Acquisition and adaptation of ultra-small parasitic reduced genome bacteria to mammalian hosts. *Cell Rep* 2020;**32**:107939.

McLean JS, Lombardo MJ, Ziegler MG *et al.* Genome of the pathogen *Porphyromonas gingivalis* recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform. *Genome Res* 2013;**23**:867–77.

Morton JT, Aksenov AA, Nothias LF *et al.* Learning representations of microbe-metabolite interactions. *Nat Methods* 2019a;**16**:1306–14.

Morton JT, Marotz C, Washburne A *et al.* Establishing microbial composition measurement standards with reference frames. *Nat Commun* 2019b;**10**:2719.

Morton JT, Sanders J, Quinn RA *et al.* Balance trees reveal microbial niche differentiation. *mSystems* 2017;**2**:e00162–16.

Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* 2020;**38**:701–7.

Moussa DG, Ahmad P, Mansour TA *et al.* Current state and challenges of the global outcomes of dental caries research in the meta-omics era. *Front Cell Infect Microbiol* 2022;**12**:887907.

Muir P, Li S, Lou S *et al.* The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol* 2016;**17**:53.

Murray CS, Gao Y, Wu M. Re-evaluating the evidence for a universal genetic boundary among microbial species. *Nat Commun* 2021;**12**:4059.

Naito M, Ogura Y, Itoh T *et al.* The complete genome sequencing of *Prevotella intermedia* strain OMA14 and a subsequent fine-scale, intra-species genomic comparison reveal an unusual amplification of conjugative and mobile transposons and identify a novel *Prevotella*-lineage-specific repeat. *DNA Res* 2016;**23**:11–9.

Nakamura T, Yamada KD, Tomii K *et al.* Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 2018;**34**:2490–2.

Neely BA, Dorfer V, Martens L *et al.* Toward an integrated machine learning model of a proteomics experiment. *J Proteome Res* 2023;**22**:681–96.

Nguyen LT, Schmidt HA, von Haeseler A *et al.* IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;**32**:268–74.

Nguyen T, Sedghi L, Ganther S *et al.* Host-microbe interactions: profiles in the transcriptome, the proteome, and the metabolome. *Periodontol 2000* 2020;**82**:115–28.

Ni Z, Wolk M, Jukes G *et al.* Guiding the choice of informatics software and tools for lipidomics research applications. *Nat Methods* 2022;**20**:193–204.

Nothias LF, Petras D, Schmid R *et al.* Feature-based molecular networking in the GNPS analysis environment. *Nat Methods* 2020;**17**:905–8.

Nouioui I, Carro L, Garcia-Lopez M *et al.* Genome-based taxonomic classification of the phylum *Actinobacteria*. *Front Microbiol* 2018;**9**:2007.

Nowicki EM, Shroff R, Singleton JA *et al.* Microbiota and metatranscriptome changes accompanying the onset of gingivitis. *mBio* 2018;**9**:e00575–18.

Nurk S, Meleshko D, Korobeynikov A *et al.* metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017;**27**:824–34.

Overmyer KA, Rhoads TW, Merrill AE *et al.* Proteomics, lipidomics, metabolomics, and 16S DNA sequencing of dental plaque from patients with diabetes and periodontal disease. *Mol Cell Proteomics* 2021;**20**:100126.

Pade LR, Stepler KE, Portero EP *et al.* Biological mass spectrometry enables spatiotemporal 'omics: from tissues to cells to organelles. *Mass Spectrom Rev* 2023;**16**:e21824.

Page AJ, Cummins CA, Hunt M *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;**31**:3691–3.

Palmer A, Phapale P, Chernyavsky I *et al.* FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat Methods* 2017;**14**:57–60.

Palmer SR, Miller JH, Abranches J *et al.* Phenotypic heterogeneity of genomically-diverse isolates of *Streptococcus mutans*. *PLoS One* 2013;**8**:e61358.

Pan S, Zhao XM, Coelho LP. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics* 2023;**39**:i21–9.

Pang Z, Zhou G, Ewald J *et al.* Using MetaboAnalyst 5.0 for LC-HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. *Nat Protoc* 2022;**17**: 1735–61.

Pasolli E, Asnicar F, Manara S *et al.* Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 2019;**176**:649–62.

Paulhe N, Canlet C, Damont A *et al.* PeakForest: a multi-platform digital infrastructure for interoperable metabolite spectral data and metadata management. *Metabolomics* 2022;**18**:40.

Perkel JM. Single-cell proteomics takes centre stage. *Nature* 2021;**597**:580–2.

Peterson SN, Meissner T, Su AI *et al.* Functional expression of dental plaque microbiota. *Front Cell Infect Microbiol* 2014;**4**:108.

Pitt ME, Nguyen SH, Duarte TPS *et al.* Evaluating the genome and resistome of extensively drug-resistant *Klebsiella pneumoniae* using native DNA and RNA nanopore sequencing. *Gigascience* 2020;**9**:giaa002.

Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 2009;**26**:1641–50.

Prjibelski A, Antipov D, Meleshko D *et al.* Using SPAdes *de novo* assembler. *Curr Protoc Bioinformatics* 2020;**70**:e102.

Protsyuk I, Melnik AV, Nothias LF *et al.* 3D molecular cartography using LC-MS facilitated by optimus and 'ili software. *Nat Protoc* 2018;**13**:134–54.

Radaic A, Kapila YL. The oralome and its dysbiosis: new insights into oral microbiome-host interactions. *Comput Struct Biotechnol J* 2021;**19**:1335–60.

Rajczewski AT, Jagtap PD, Griffin TJ. An overview of technologies for MS-based proteomics-centric multi-omics. *Expert Rev Proteomics* 2022;**19**:165–81.

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

Rost HL, Sachsenberg T, Aiche S *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods* 2016;**13**:741–8.

Schmid R, Heuckeroth S, Korf A *et al.* Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat Biotechnol* 2023;**41**:447–9.

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

Sereika M, Kirkegaard RH, Karst SM *et al.* 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* 2022;**19**: 823–6.

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

Shi B, Chang M, Martin J *et al.* Dynamic changes in the subgingival microbiome and their potential for diagnosis and prognosis of periodontitis. *mBio* 2015;**6**:e01926–01914.

Shuken SR. An introduction to mass spectrometry-based proteomics. *J Proteome Res* 2023;**22**:2151–71.

Sieber CMK, Probst AJ, Sharrar A *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 2018;**3**:836–43.

Simpson JT, Wong K, Jackman SD *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009;**19**:1117–23.

Sinha D, Sun X, Khare M *et al.* Pangenome analysis and virulence profiling of *Streptococcus intermedius*. *BMC Genomics* 2021;**22**:522.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;**30**:1312–3.

Tajik M, Baharfar M, Donald WA. Single-cell mass spectrometry. *Trends Biotechnol* 2022;**40**:1374–92.

Tinder EL, Faustoferri RC, Buckley AA *et al.* 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* 2022;**7**:e0127221.

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

Torres PJ, Thompson J, McLean JS *et al.* Discovery of a novel periodontal disease-associated bacterium. *Microb Ecol* 2019;**77**:267–76.

Tran NH, Zhang X, Xin L *et al.* De novo peptide sequencing by deep learning. *Proc Natl Acad Sci USA* 2017;**114**:8247–52.

Treerat P, McGuire B, Palmer E *et al.* Oral microbiome diversity: the curious case of *Corynebacterium* sp. isolation. *Mol Oral Microbiol* 2022;**37**:167–79.

Tripathi A, Vazquez-Baeza Y, Gauglitz JM *et al.* Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. *Nat Chem Biol* 2021;**17**:146–51.

Tsao SW, Tsang CM, Lo KW. Epstein-Barr virus infection and nasopharyngeal carcinoma. *Philos Trans R Soc Lond B Biol Sci* 2017;**372**:20160270.

Tsugawa H, Ikeda K, Takahashi M *et al.* A lipidome atlas in MS-DIAL 4. *Nat Biotechnol* 2020;**38**:1159–63.

Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* 2016;**11**:2301–19.

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

Utter DR, Borisy GG, Eren AM *et al.* Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biol* 2020;**21**:293.

van der Walt AJ, van Goethem MW, Ramond JB *et al.* Assembling metagenomes, one community at a time. *BMC Genomics* 2017;**18**:521.

Vaser R, Sovic I, Nagarajan N *et al.* Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res* 2017;**27**:737–46.

Velsko IM, Chakraborty B, Nascimento MM *et al.* Species designations belie phenotypic and genotypic heterogeneity in oral streptococci. *mSystems* 2018;**3**:e00158–18.

Venter JC, Remington K, Heidelberg JF *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004;**304**:66–74.

Walker AR, Shields RC. Investigating CRISPR spacer targets and their impact on genomic diversification of *Streptococcus mutans*. *Front Genet* 2022;**13**:997341.

Walker BJ, Abeel T, Shea T *et al.* Pilon: an Integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;**9**:e112963.

Wang M, Carver JJ, Phelan VV *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* 2016;**34**:828–37.

Wang M, Jarmusch AK, Vargas F *et al.* Mass spectrometry searches using MASST. *Nat Biotechnol* 2020;**38**:23–6.

Wang O, Chin R, Cheng X *et al.* 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* 2019;**29**:798–808.

Wang Y, Zhao Y, Bollas A *et al.* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* 2021;**39**:1348–65.

Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol* 2019;**37**:124–6.

Wenger AM, Peluso P, Rowell WJ *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;**37**:1155–62.

Wetterstrand KA. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. 2023. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (6 September 2023, date last accessed).

White LK, Hesselberth JR. Modification mapping by nanopore sequencing. *Front Genet* 2022;**13**:1037134.

Wichmann C, Meier F, Virreira Winter S *et al.* MaxQuant. Live enables global targeting of more than 25,000 peptides. *Mol Cell Proteomics* 2019;**18**:982–94.

Wick RR, Holt KE. Polypolish: short-read polishing of long-read bacterial genome assemblies. *PLoS Comput Biol* 2022;**18**:e1009802.

Wick RR, Judd LM, Cerdeira LT *et al.* Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol* 2021;**22**:266.

Wick RR, Judd LM, Gorrie CL *et al.*: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;**13**:e1005595.

Wickramarachchi A, Lin Y. Binning long reads in metagenomics datasets using composition and coverage information. *Algorithms Mol Biol* 2022;**17**:14.

Wilbanks EG, Dore H, Ashby MH *et al.* Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity. *ISME J* 2022;**16**:1921–31.

Williams DW, Greenwell-Wild T, Brenchley L *et al.* Human oral mucosa cell atlas reveals a stromal-neutrophil axis regulating tissue immunity. *Cell* 2021;**184**:4090–104.

Wishart DS, Girod S, Peters H *et al.* ChemFOnt: the chemical functional ontology resource. *Nucleic Acids Res* 2023;**51**:D1220–9.

Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016;**32**:605–7.

Xin L, Qiao R, Chen X *et al.* A streamlined platform for analyzing tera-scale DDA and DIA mass spectrometry data enables highly sensitive immunopeptidomics. *Nat Commun* 2022;**13**:3108.

Yahara K, Suzuki M, Hirabayashi A *et al.* Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria. *Nat Commun* 2021;**12**:27.

Yost S, Duran-Pinedo AE, Teles R *et al.* Functional signatures of oral dysbiosis during periodontitis progression revealed by microbial metatranscriptome analysis. *Genome Med* 2015;**7**:27.

Zeng WF, Zhou XX, Willems S *et al.* AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nat Commun* 2022;**13**:7238.

Zerbino DR, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* 2008;**18**:821–9.

Zheng J, Wittouck S, Salvetti E *et al.* 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* 2020;**70**:2782–858.

Zhu J, Tian L, Chen P *et al.* Over 50,000 metagenomically assembled draft genomes for the Human oral microbiome reveal new taxa. *Genomics Proteomics Bioinformatics* 2022;**20**:246–59.

Received 31 January 2023; **revised** 2 August 2023; **accepted** 1 September 2023