AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

Using Nanopore Sequencing to Obtain Complete Bacterial Genomes from Saliva Samples

Jonathon L. Baker[a,b]

[a]Genomic Medicine Group, J. Craig Venter Institute, La Jolla, California, USA
[b]Department of Pediatrics, UC San Diego School of Medicine, La Jolla, California, USA

ABSTRACT Obtaining complete, high-quality reference genomes is essential to the study of any organism. Recent advances in nanopore sequencing, as well as genome assembly and analysis methods, have made it possible to obtain complete bacterial genomes from metagenomic (i.e., multispecies) samples, including those from the human microbiome. In this study, methods are presented to obtain complete bacterial genomes from human saliva using complementary Oxford Nanopore (ONT) and Illumina sequencing. Applied to 3 human saliva samples, these methods resulted in 11 complete bacterial genomes: 3 *Saccharibacteria* clade G6 (also known as *Ca.* Nanogingivalaceae HMT-870), 1 *Saccharibacteria* clade G1 HMT-348, 2 *Rothia mucilaginosa*, 2 *Actinomyces graevenitzii*, 1 *Mogibacterium diversum*, 1 *Lachnospiraceae* HMT-096, and 1 *Lancefieldella parvula*; and one circular chromosome of Ruminococcaceae HMT-075 (which likely has at least 2 chromosomes). The 4 *Saccharibacteria* genomes, as well as the *Actinomyces graeventizii* genomes, represented the first complete genomes from their respective bacterial taxa. Aside from the complete genomes, the assemblies contained 147 contigs of over 500,000 bp each and thousands of smaller contigs, together representing a myriad of additional draft genomes including many which are likely nearly complete. The complete genomes enabled highly accurate pangenome analysis, which identified unique and missing features of each genome compared to its closest relatives with complete genomes available in public repositories. These features provide clues as to the lifestyle and ecological role of these bacteria within the human oral microbiota, which will be particularly useful in designing future studies of the taxa that have never been isolated or cultivated.

IMPORTANCE Obtaining complete and accurate genomes is crucial to the study of any organism. Previously, obtaining complete genomes of bacteria, including those of the human microbiome, frequently required isolation of the organism, as well as low-throughput, manual sequencing methods to resolve repeat regions. Advancements in long-read sequencing technologies, including Oxford Nanopore (ONT), have made it possible to obtain complete, closed bacterial genomes from metagenomic samples. This study reports methods to obtain complete genomes from the human oral microbiome using complementary ONT and Illumina sequencing of saliva samples. Eleven complete genomes were obtained from 3 human saliva samples, with genomes of *Saccharibacteria* HMT-870, *Saccharibacteria* HMT-348, and *Actinomyces graeventzii* being the first complete genomes from their respective taxa. Obtaining complete bacterial genomes in a high-throughput manner will help illuminate the metabolic and ecological roles of important members of the human microbiota, particularly those that have remained recalcitrant to isolation and cultivation.

KEYWORDS oral microbiome, genomics, metagenomics, pangenomics

Sequencing the genomes of the taxa comprising the human microbiome is fundamental to our understanding of how these species live and affect our health (1, 2). For a given taxon, obtaining a high-quality reference genome is crucial for several reasons. First, a high-quality genome allows researchers to quantify the abundance of a particular taxon, or its mRNA transcripts, in microbiome samples accurately through

Editor Marnix Medema, Wageningen University

Address correspondence to jobaker@jcvi.org.

The authors declare no conflict of interest.

Received 25 May 2022
Accepted 29 July 2022

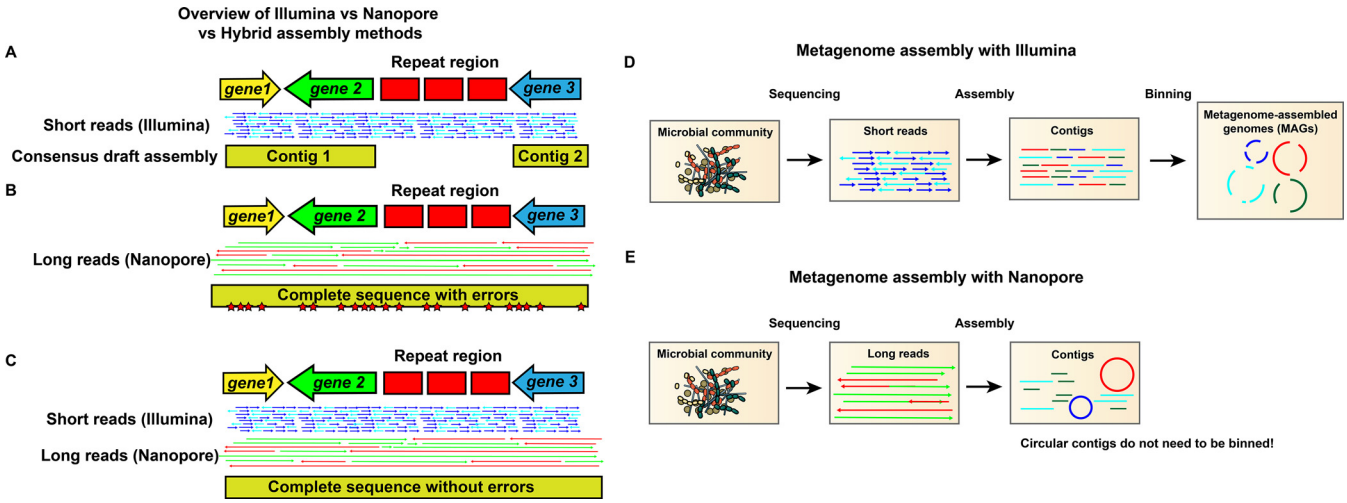

FIG 1 Getting complete and accurate genomes from metagenomes using Illumina and Nanopore sequencing. Illustrations showing assembly of genomes using (A) Illumina only, (B) Nanopore only, or (C) both Illumina and Nanopore. (A) Assembly with short reads yields high accuracy, but short reads map nonspecifically in repeat regions, which therefore cannot be resolved. The result is an assembly fragmented into contigs. (B) Assembly with nanopore reads resolves repeat regions, because individual reads map all the way through, resulting in a complete, contiguous assembly. Errors will be present due to the higher error rate of ONT sequencing, most frequently in homopolymeric tracts and short repeats. (C) Assembly with complementary Illumina and nanopore sequencing allows for nanopore errors to be corrected using the more accurate short reads, yielding an assembly that is complete, contiguous, and highly accurate. (D) Metagenome assembly with Illumina sequencing requires binning that frequently produces contamination in the resulting fragmented draft metagenome-assembled genome. (E) Metagenome assembly with ONT produces circular contigs that represent complete chromosomes and therefore do not need to be binned.

read-mapping (3). Second, with a complete genome in hand, researchers can predict the metabolic and ecological role of the taxon, which can be especially important for species that have not yet been isolated or cultivated in the lab (which still represents the majority of bacteria) (4). Finally, a high-quality genome is critical to guide wet lab experimentation, such as genome editing and mutagenesis (4). Despite the vast and continuously growing number of microbial genomes in databases such as NCBI RefSeq and the Human Oral Microbiome Database (HOMD), many bacteria, including those of the human oral microbiome, lack reference genomes. Complete genomes, in the case of bacteria, are most frequently circular chromosomes, although some taxa have linear chromosomes and some taxa, such as *Prevotella* spp., have more than one chromosome (5).

For the past ~15 years, Illumina shotgun sequencing has been the industry workhorse, revolutionizing the life sciences by lowering the cost and increasing the throughput of sequencing by several orders of magnitude, while providing very high accuracy (6, 7). The main drawback to this technology is the short length of the reads, which are generally 150 or 300 bp. To obtain genomes or metagenomes (i.e., genomic sequencing of samples containing the genomes of multiple taxa or isolates, such as a microbiome sample), these short reads must be assembled. However, most genomes have repetitive regions such as rRNA regions or invertases, that can be well over 10,000 bp in length (and may be contiguous or in distant loci; both are problematic). The short-reads from within these regions all map nonspecifically; therefore, one cannot tell how many repeats exist within the repeat structure, or what part of the genome connects on the other side of the repeat (Fig. 1). Consequently, these regions cannot be elucidated with confidence during assembly, causing the production of separate fragments, known as contigs, rather than a complete chromosome in the resulting draft genome (8). When dealing with metagenomes, rather than isolate genomes, several additional problems are introduced (Fig. 1). First, as there are a finite number of reads generated in a sequencing run, more genomes in the sample will mean fewer reads produced per genome (3). Because accurate genome assembly is, in part, dependent in on coverage depth, fewer reads supporting a given genome will lead to a more fragmented, incomplete, and error-prone assembly. Second, it can be difficult to know which particular contigs are from the same genome. "Binning" is the process to sort

the metagenomic contigs into discrete, fragmented draft genomes, and many computational tools have been developed to do this, typically using k-mer frequency, coverage depth, GC content, and/or alignment to references (9–11). However, even the most exhaustive automated and manual binning strategies can "misplace" contigs into the wrong genome bin, leading to "contamination" (12). It can be especially problematic when these contaminated draft genomes make it as far as public repositories, as downstream researchers will usually accept these assemblies as ground truth and use the erroneous data to design further experiments (13). Complete genomes (i.e., contiguous chromosomes) are, by definition, free of these types of errors, and are therefore the most useful tools for scientists, as high confidence can be placed in their accuracy. It is imperative that researchers examine whether the reference genomes in use are "complete" or "draft," and recognize the limitations of draft genomes (12, 13).

Until recently, due to the constraints inherent in short-read sequencing just described above, obtaining complete genomes usually required (i) manual steps beyond computational assembly, such as PCR and further sequencing, due to the repeat regions mentioned above, and (ii) pure cultures to eliminate the problems inherent with sequencing metagenomes. These requirements significantly limited the output of complete genomes, particularly since the majority of bacterial species have yet to be successfully isolated and cultured in the lab. However, recent advancements in long-read sequencing technology, particularly that of Oxford Nanopore Technologies (ONT), have been revolutionary (14). Unlike the sequencing-by-synthesis techniques employed by most Illumina and PacBio technologies, ONT sequences native DNA or RNA molecules, and read length is only limited by the size of input DNA, allowing for read lengths of over 1 Mbp under the right conditions (14). The long reads produced by ONT sequencing easily map all the way through problematic repeat regions, significantly improving assembly contiguity and frequently resulting in complete, circular genomes/chromosomes, even from metagenomic (i.e., nonisolate) samples, eliminating the need for binning (Fig. 1) (15–17). Recently, ONT sequencing was instrumental in obtaining the first telomere-to-telomere complete human genome (18).

In ONT sequencing, native nucleic acids are ratcheted through a nanopore embedded in a synthetic membrane, and the changes in electrical current are monitored (14). Sequences of bases or windows of bases cause specific and predictable perturbations in the electrical potential; therefore, the sequence of bases passing through the nanopore can be reconstructed by analyzing the current (8, 19). Although nanopore sequencing technology is not new, its applicability was limited for many years by a high error rate. These error rates have dropped substantially in recent years, from 35–40% in 2015 to $<$1% with current ONT instrumentation and software (20). These improvements have been due to major advancements both in nanopore chemistry and in machine learning technology, which underpins the basecalling algorithms. The errors introduced by nanopore sequencing are not random, but typically occur during homopolymeric tracts or short repeats, where the basecalling software has difficulty identifying how many consecutive iterations of a given base or bases have passed through the nanopore (19). This leads to insertions or deletions (i.e., indels) in these homopolymeric tracts, which can cause apparent frameshifts and therefore can have a significant impact on downstream gene calling and gene annotation (19). Increasing depth of sequencing coverage mitigates, but does not eliminate, this issue. Consequently, at this time ONT sequencing can fully stand on its own for many applications, such as RNA-seq and assembly of draft genomes with a high degree of completeness (8). ONT sequencing is also useful to perform 16S amplicon sequencing to profile microbial communities, as the longer ONT reads can span the entire 16s rRNA gene, therefore providing coverage across all the variable regions, which allows for increased taxonomic specificity and less taxonomic bias compared to methods targeting only one variable region (21). However, for producing error-free, publication-quality complete genomes, complementary Illumina sequencing is still helpful, as the substantially lower error rate in Illumina reads can be used to correct errors in the ONT-based assembly (Fig. 1) (8).

In this study, a protocol for generating complete genomes of oral bacteria from saliva using complementary ONT and Illumina sequencing is reported. Sequencing of 3 saliva

samples resulted in 12 complete, circular chromosomes. Among these were 3 genomes representing two distinct species of clade G6 *Saccharibacteria* (HMT-870) (22, 23), 1 genome of clade G1 *Saccharibacteria* HMT-348 (24), 2 genomes of *Actinomyces graevenitzii*, 2 genomes of *Rothia mucilaginosa*, 1 genome of *Lachnospiraceae* HMT-096, 1 genome of *Mogibacterium diversum*, 1 genome of *Lancefieldella parvula*, and also a complete circular chromosome from *Ruminococcaceae* HMT-075 (this taxon likely has multiple chromosomes). The properties of these genomes are then highlighted.

RESULTS AND DISCUSSSION

Extracting high molecular weight genomic DNA (HMW gDNA). During ONT sequencing, the native DNA or RNA molecules are sequenced, and the read length of the sequences produced is limited mainly by the length of the input material. Therefore, to obtain the most complete genomes possible from genomic or metagenomic sequencing, it is important to obtain HMW gDNA with as little shearing of the molecules as possible. For the Ligation Sequencing Kit used to prepare the DNA library for ONT sequencing, 1 μg of HMW gDNA is required. Several gDNA preparation methods were tested here, described in more detail in Materials and Methods. The protocol that gave the best results, and was used subsequently here, was a phenol:chloroform-based protocol originally published by Chen and Burne (25), which was recently optimized specifically for ONT sequencing of *Streptococcus mutans* B04Sm5 (26). It should be emphasized that during the course of this study, several new HMW isolation kits/protocols have been released that may produce even longer reads (and therefore additional complete genomes) but were not tested here. These include the Ultra-Long DNA Sequencing Kit from Oxford Nanopore Technologies, Inc. and the Nanobind UHMW DNA Extraction protocol from Circulomics, Inc.

The 3 saliva samples sequenced with ONT in this study were SC23, SC24, and SC33, which were all collected from children with healthy dentition in Los Angeles, CA as previously described (NCBI accession number PRJNA624185) (22, 23). These samples were previously sequenced using Illumina technology and subjected to metagenomic analysis (27). The corresponding Illumina sequencing libraries from these samples are available in the Sequence Read Archive (SRA) with accession numbers SRX4318838 (SC23), SRX4318837 (SC24), and SRX4318835 (SC33). Here, HMW gDNA was extracted from 1-mL aliquots of the same saliva samples used to produce the original Illumina short-read libraries using the modified Chen and Burne protocol (25) described above and in further detail in Materials and Methods. The samples were sequenced on a GridION instrument with each sample using a full R9.4.1 flow cell. ONT sequencing of SC23 yielded 3.2 million reads with an N50 of 13,719 bp, while nanopore sequencing of SC24 yielded 3.9 million reads with an N50 of 14,073 bp, and nanopore sequencing of SC33 yielded 8.9 million reads with an N50 of 6,527 bp. Because the saliva contains human DNA, which will only complicate the assembly process for the microbial genomes, human reads were removed from all 3 long-read libraries by mapping the read libraries to the human genome using minimap2 (28) and removing the reads that mapped. Following removal of the human reads, SC23 contained 686K reads with an N50 of 19,172 bp, SC24 contained 1.07M reads with an N50 of 19,058 bp, and SC33 contained 1.2M reads with an N50 of 13,393 bp. The longest single reads in SC23, SC24, and SC33 were 88,225 bp, 109,498 bp, and 124,167 bp, respectively.

Metagenomic assembly. Each of the 3 long-read libraries was assembled using metaFlye (29). Tables S1–S3 contain a summary of the assembly info from Flye. In terms of putative complete genomes, considered here to be a circular contig >500,000 bp, SC23 had 2, SC24 had 5, and SC33 had 4. These assemblies are described in Table 1. The assemblies also contained many very large linear contigs (>500,000 bp); SC23 had 38, SC24 had 69, and SC33 had 40, with the largest linear contigs in all 3 assemblies being well over 2 Mbp, likely representing nearly complete genomes (or possibly complete genomes from organisms that have linear, rather than circular, genomes). As the circular contigs are much more likely to be complete genomes (although not error-free at this stage), only large circular contigs are the focus of this study.

Assembly of JB001, JB002, and JB003 (*Saccharibacteria* clade G6). *Candidatus* Nanogingivalaceae (a Clade G6 *Saccharibacteria*/HMT-870) strains JB001, JB002, and

TABLE 1 Genomes of interest produced by this study

Assembly name	Length	No. of genes/CDS	Nanopore coverage	Illumina coverage	Saliva sample	Flye contig name	NCBI accession	Notes
Rothia mucilaginosa strain JCVI-JB-Rm27	2,273,720	1,848	256	15	SC_23	contig_3145	CP097094	
Actinomyces graevenitzii strain JCVI-JB-Ag32	2,130,963	1,822	32	69	SC_33	contig_1	CP097095	
Mogibacterium diversum strain JCVI-JB-Md32	1,770,007	1,654	11	837	SC_33	contig_944	CP097093	
Lancefieldella parvula strain JCVI-JB-Lp32	1,624,536	1,535	20	1665	SC_33	contig_3802	CP097092	
HMT-348_TM7c-JB (Sacchraribacteria G1 HMT-348)	841,302	852	13	1103	SC_33	contig_847	CP090820	
Lachnospiraceae HMT-096 strain JCVI-JB-L28	2,401,457	2,599	31	0.3	SC_24	contig_1395		low Illumina coverage
Actinomyces graevenitzii strain JCVI-JB-Ag28	2,152,369	2,184	23	0.2	SC_24	contig_2578		low Illumina coverage
Rothia mucilaginosa strain JCVI-JB-Rm28	2,266,186	1,992	326	5	SC_24	contig_4121		low Illumina coverage
Ruminococcaceae HMT-075-chr1-JCVI-JB-R28	1,207,213	1,879	11	22	SC_24	contig_4465		1 of multiple chromosomes
JB001 (Saccharibacteria G6 HMT-870)	663,355	689	46	57	SC_23	contig_69	CP072208	Described in Baker et al. 2021
JB002 (Saccharibacteria G6 HMT-870)	637,739	651	NA	NA	SC_33	NA	CP076101	Described in Baker et al. 2021
JB003 (Saccharibacteria G6 HMT-870)	691,584	730	45	47	SC_24	contig_1024	CP076102	Described in Baker et al. 2021

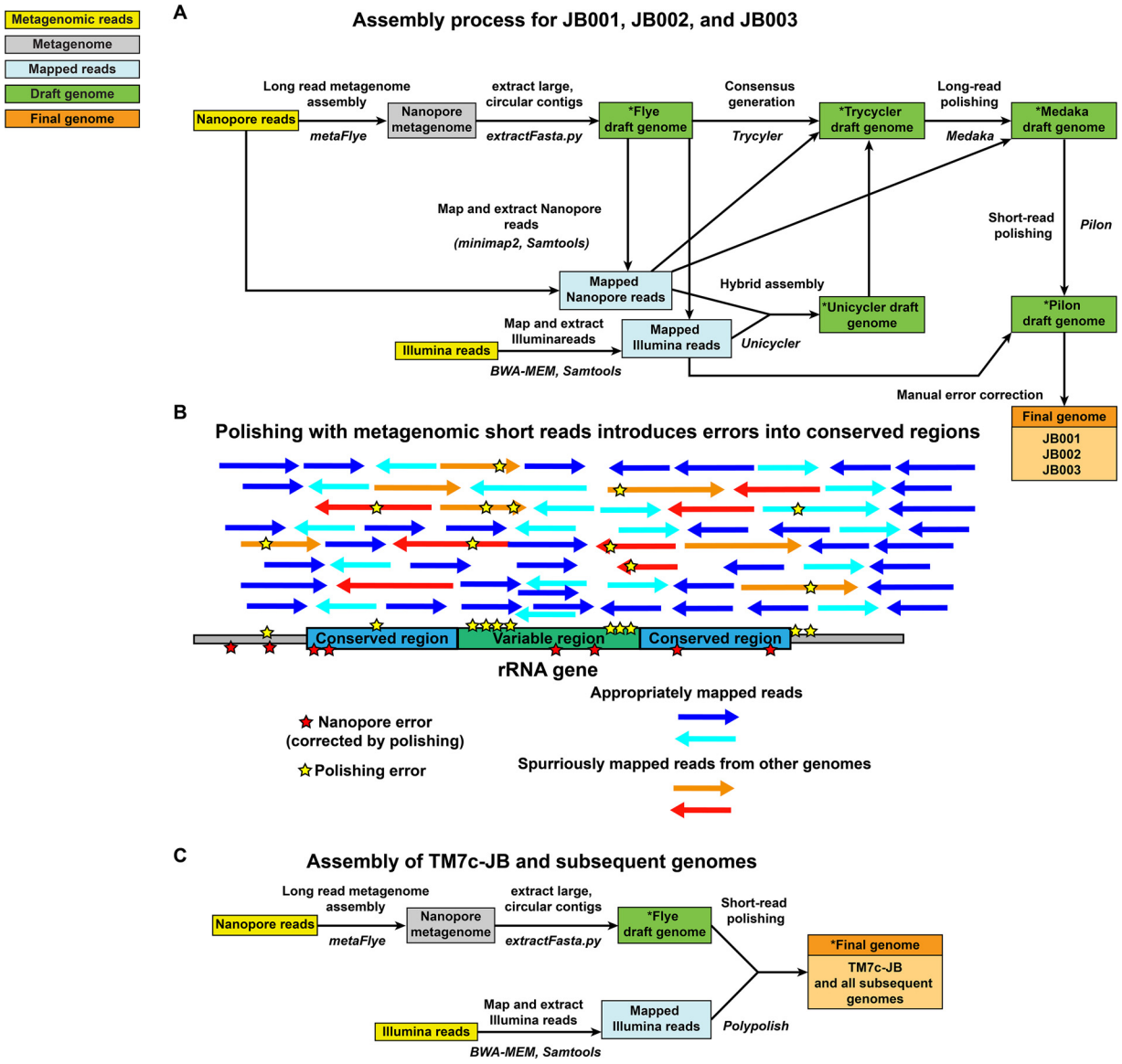

FIG 2 Genome assembly and polishing pipelines used in this study. (A) Original pipeline of assembling and correcting errors in the metagenome-assembled genomes JB001, JB002, and JB003. (B) Polishing with metagenomic short reads introduces errors into conserved regions. This issue was identified with the pipeline presented in panel A above. Metagenomic short reads from other species/genomes spuriously map to the draft genome because of conserved regions, such as those found in rRNA. These spuriously mapped reads introduce errors into the polished genome. The polishing is still useful for correcting ONT errors elsewhere in the genome, however. (C) Pipeline used on HMT-348-TM7c-JB and all other genomes in this study. Polypolish alleviates the issue described in panel B almost entirely. Draft genomes in Panels A and C labeled with asterisks are available on GitHub.

JB003 were derived from this data set and have already been described previously (22, 23); however, it is useful to summarize the assembly methods used on those genomes to explain how the assembly pipeline of the genomes described in this study matured and improved over the course of the study (Fig. 2A). Assembly methods using both long and short reads either (i) assemble the long reads into a draft and then remove errors in the long reads using the short reads or (ii) assemble the short reads and then use long reads to bridge together the disjointed contigs. Assembly with Flye, followed by short-read polishing uses the former strategy, while assembling both long and short reads mapping to the draft assembly using Unicycler (30), employs the latter. Both of these methods were attempted, and then Trycycler (31) was used to determine the best final consensus assembly using the 2 draft assemblies. The Trycycler consensus assembly was then polished with the long reads using Medaka (https://github.com/nanoporetech/medaka) and with the short reads using Pilon (32) (Fig. 2A).

However, comparison of annotated versions of the draft genomes both before and after polishing steps, along with reference rRNA genes from *Candidatus* Nanogingivalaceae, revealed a problem (Fig. 2B) (23). While genomic rRNA sequences from the original Flye assembly were nearly identical to the HOMD reference rRNA sequences, the genomic rRNA sequences produced following polishing steps contained many more mismatches (23). The likely explanation is that because regions of the rRNA are highly conserved (which is why 16S amplicon sequencing is a widely used and cost-effective means of microbiome analysis), and the read libraries used for Medaka and Pilon polishing came from a metagenomic sample, reads from other species/genomes align to the conserved rRNA regions and get spuriously mapped to the draft assembly (23). The portions of the reads that do not perfectly align the assembly then cause errors to be introduced during the polishing (Fig. 2B). Outside of the widely conserved rRNA operons, however, the Medaka and Pilon polishing steps were useful in correctly removing errors, especially the frameshift-causing indels within homopolymeric tracts, described above as a drawback to ONT sequencing (23). To solve this issue, the sequence of the rRNA regions within the consensus genome was manually changed back to that of the original, correct Flye assembly, while the remainder of the genome kept the changes introduced by polishing using Pilon (23). This protocol was used in the production of JB001, JB002, and JB003 (all *Candidatus* Nanogingivalacae HMT-870) (22, 23).

Assembly of HMT-348-TM7c-JB. The next genome examined in this data set was *Candidatus* Nanosynbacter HMT-348 strain HMT-348-TM7c-JB (contig_847 in the SC33 metaFlye assembly [Table S3]), which was circular and 841,704 bp in length (24). This Flye draft genome appeared to be the cognate long-read draft genome to the original short-read based draft genome, which was Candidatus_Nanosynbacter_sp._isolate_JCVI_32_bin.19, a 793,808 bp assembly fragmented into 7 contigs (27). Following the publication of JB001, JB002, and JB003, the Polypolish short-read polishing tool was published (33). Polypolish used a novel approach to circumvent the issues in the rRNA and other repeat/conserved regions reported in the paragraph above (33). Unlike its predecessors, such as Pilon, which map each short read to the best matching location in the draft assembly (randomly assigning a single mapping if the read maps equally well to multiple places in the assembly), Polypolish maps each short read to all matching places in the draft genome (33). This greatly improves error correction in repeat and conserved regions (Fig. 2C) (33).

To compare the number of errors present in the methods used in the manual pipeline used for JB001, JB002, and JB003 versus Polypolish (i.e., compare the results of the pipelines in Fig. 2A and C), and determine the optimal pipeline moving forward, each stage of the HMT-348-TM7c-JB assembly was examined for nucleotide identity to the HOMD reference 16S rRNA sequence for HMT-348 (to detect spurious assembly and/or polishing), and was examined for missing or truncated open reading frames (ORFs) (mainly due to ONT basecalling errors, which can be corrected by polishing). The Medaka, Flye, and Trycycler assemblies only had one mismatch with the HOMD reference 16S region, while Polypolish had 3, Pilon had 23, Unicycler had 24, and the original Illumina-based SPAdes assembly had 41 (Table 2). This was logical, as the metaSPAdes assembly, using short reads only, would be expected to have difficulty assembling the 16S region from a metagenomic pool of short reads, and the Pilon and Unicycler (Unicycler itself also uses SPAdes and Pilon as part of its pipeline) polish errors into the rRNA regions due to the conserved regions causing spurious mapping of reads, disrupting variable regions (Fig. 2B). It should be noted that the HOMD reference 16S rRNA sequence for HMT-348 is not itself linked to a good quality genome, and there is little knowledge about species and strain diversity within HMT-348; therefore, one cannot be sure that the 16S rRNA sequence of HMT-348-TM7c-JB would be expected to match the HOMD reference exactly (i.e., the Polypolish, Medaka, Flye, or Trycycler sequences may, in fact, be correct). However, obviously metaSPAdes, Pilon, and Unicycler perform poorly in assembling a correct 16S rRNA gene from a metagenomic read library.

Missing and disrupted open reading frames (ORFs) were counted by visualizing the annotated alignment of all 7 assemblies and identifying premature stop codons, split genes, and missing genes. Polypolish performed the best, with only 3 truncated/missing ORFs,

TABLE 2 Identifying and comparing errors in HMT-348-TM7c-JB assemblies[a]

Assembly	% identity	No. of mismatches
16S rRNA sequence % identity to HMT-348 16S rRNA on HOMD		
Polypolish	99.77%	3
Illumina-spades	96.84%	41
Medaka	99.92%	1
Pilon	98.23%	23
Unicycler	98.15%	24
Flye	99.92%	1
Trycycler	99.92%	1
Assembly	No. missing or truncated ORFs	Notes
Broken ORFs		
Polypolish	3	only missing small hypothetical proteins
Illumina-spades	53	1 large region that was missing accounted for ~35
Medaka	137	
Pilon	5	only 2 are the same
Unicycler	5	
Flye	189	
Trycycler	124	
Assembly	CDSs	Length
No. of CDSs		
Polypolish	804	841,260
Illumina-spades	774	788,149
Medaka	1,008	841,361
Pilon	807	841,266
Unicycler	807	841,180
Flye	1,102	841,704
Trycycler	1,021	841,085

[a]HOMD, Human Oral Microbiome Database; ORFs, open reading frames; CDSs, coding DNA sequences.

which were all encoding very small missing hypothetical proteins (which therefore may not even be bona fide genes and errors) (Table 2). The Pilon and Unicycler assemblies each had 5 disrupted ORFs (Table 2). The Illumina/metaSPAdes assembly had 53 truncated/missing ORFs, 35 of which were missing due to the assembly lacking that entire region (Table 2). As to be expected due to a lack of short-read polishing, the assemblies leaning the most heavily on ONT assembly and polishing did poorly in this regard, with the Medaka, Flye, and Pilon assemblies having 137, 189, and 124 disrupted/missing ORFs, respectively (Table 2). The number of disrupted ORFs is also apparent when examining the total number of predicted coding DNA sequences (CDSs) in each assembly following annotation with Prokka, with the Flye assembly having 298 more ORFs than the Polypolish assembly (Table 2). To further validate the metaFlye followed by Polypolish pipeline, JB001 and JB003 were reassembled using metaFlye and Polypolish and compared to the JB001 and JB003 genomes from GenBank that had used the manual Pilon-based approach shown in Fig. 2A. JB001 and JB003 quite likely represent different isolates of the same species, and only differed in a few single nucleotide polymorphisms (SNPs) and two regions that appear to be prophages or other types of mobile elements, which may legitimately be present in only a subset of the metagenomic populations sampled (23). When compared to the manually polished JB001 and JB003, the Polypolish-polished JB001 and JB003 had >99.99% average nucleotide identity (ANI) of aligned fractions, which had at worst 96.96% alignment percent (AP) of the genome. Polypolish produced identical sequences to the manually polished genomes in one of the 16S rRNA genes and had only 1 mismatch in JB001 and 3 in JB003 in the second 16S rRNA gene, indicating once again that it largely alleviates the issues observed with the older Pilon polishing tool, which had dozens of mismatches to the HOMD reference. Outside of the 16S mismatches, there were 13 other differences in the assemblies; two were the large putative mobile elements, which were noted previously (23), and the other 11 were limited to homopolymeric tracts and short repeats, as might be expected from ONT assembly. Overall, assembly with metaFlye, followed by short-read polishing using Polypolish, combined the "best of both worlds" in terms of accuracy, and was used to polish the remaining

circular draft genomes described below. The manual polishing approach used on JB001, JB002, and JB003 was also not considered as a viable option moving forward because it is not amenable to high throughput and is more subject to human error. Since HMT-348-TM7c-JB and all other genomes reported here were assembled using the same methods, the following sections will discuss details of the novel genomes themselves.

***Saccharibacteria* G1-HMT-348 strain HMT-348-TM7c-JB.** As recently described (24), HMT-348-TM7c-JB was 841,302 bp. This genome represents the first complete genome from *Saccharibacteria* clade G1 HMT-348, which is one of the most common members of *Saccharibacteria* present in supra- and subgingival plaque, and on the buccal mucosa (34, 35). The first draft genome from this clade, TM7c, was published in 2007 (36); however, that assembly was later found to contain a significant amount of contamination. More recent studies have binned draft genomes of HMT-348 out of oral metagenomes or single-cell amplified genomes (SAGs); however, they were still fragmented into >10 contigs. The exception was the original Illumina-only assembly of HMT-348-TM7c-JB, which was fragmented into 7 contigs (27). All of these draft genomes were still significantly incomplete or contaminated (27, 35, 37). Note that the species-level genome bin (SGB) of HMT-348-TM7c-JB in the original metagenomics study also contained two other genomes, likely representing the same species with an ANI of >95% compared to HMT-348-TM7c-JB (27). To better examine phylogeny of HMT-348-TM7c-JB and HMT-348, a pangenome including HMT-348-TM7c-JB and 39 other complete *Saccharibacteria* genomes (all complete nonduplicate *Saccharibacteria* genomes on NCBI as of April 2022) was created using anvi'o (38). Only 12 single-copy core genes were common to all 40 genomes. To minimize the effect of gaps on phylogeny, the minimum geometric homogeneity index was set to 0.95, and a maximum functional homogeneity index was set to 0.85 to ensure identical or nearly identical protein sequences were not used (as described in the pangenomics tutorial at anvio.org). This left 4 genes, the ribosomal protein subunits L6 and L27, SecG, and a peptide deformylase, with which to perform a bespoke phylogenetic comparison. Concatenated protein sequences of these 4 genes were used to construct a phylogenetic tree of all 40 complete Saccharibacteria genomes, which illustrated that HMT-348-TM7c-JB has its own fairly distinct branch (Fig. 3A). To examine unique features of HMT-348-TM7c-JB compared to other clades of oral Saccharibacteria, a more focused pangenome was constructed comparing HMT-348-TM7c-JB with 7 other complete genome representatives from HMT352, HMT-952, HMT-957, HMT-488, HMT-955, and HMT-349 (Fig. 3B). The analysis identified 218 pan-G1 core genes, 327 genes that were unique to HMT-348-TM7c-JB, and 21 genes that were missing in HMT-348-TM7c-JB but present in all other clades (Fig. 3B; pangenome summary table available at: https://github.com/jonbakerlab/nanopore-oral-genomes). Note that the pangenome data tables generated by this study are too large (>50MB, in some cases) to be included as supplemental material, and therefore have been made accessible on GitHub (https://github.com/jonbakerlab/nanopore-oral-genomes). Interestingly, many of the features uniquely missing in HMT-348-TM7c-JB were either chaperones or involved in DNA repair, including *xseA*, *uvrD*, and *dnaK*. In addition to the well-conserved Type IV and Type II secretion systems encoded by all the G1 Saccharibacteria genomes, HMT-348-TM7c-JB appeared to have its own distinct Type VI and II secretion systems. Further examination of the HMT-348 genome and comparison to other Saccharibacteria will provide insight into the host–epibiont dynamics and ecological role of this abundant yet enigmatic and uncultivated bacterial taxa.

***Actinomyces graevenitzii*.** There were two genomes of *Actinomyces graevenitzii* assembled by this study, JCVI-JB-Ag32 from SC33 at 2,088,213 bp and JCVI-JB-Ag28 from SC24 at 2,122,055 bp. These two genomes had an ANI of 97.03% across an AP of 93.97%. JCVI-JB-Ag28 had very poor coverage of Illumina reads (0.2×), while JCVI-JB-Ag32 had good Illumina coverage, 68.5×. Therefore, JCVI-JB-Ag28 genome likely contained many ONT errors, undoubtedly contributing to the ANI difference between the genomes. JCVI-JB-Ag32 had a 16S rRNA gene with 99.963% identity (4 mismatches) to *Actinomyces graevenitzii* HMT-866 strain F0530 (from HOMD), while JCVI-JB-Ag28 had 99.344% identity (5 mismatches) to the same reference 16S rRNA. There are currently no complete genomes of *Actinomyces graevenitzii* available on HOMD or NCBI, with the two draft genomes on HOMD being fragmented into 10 and 29 contigs with a 2.09–

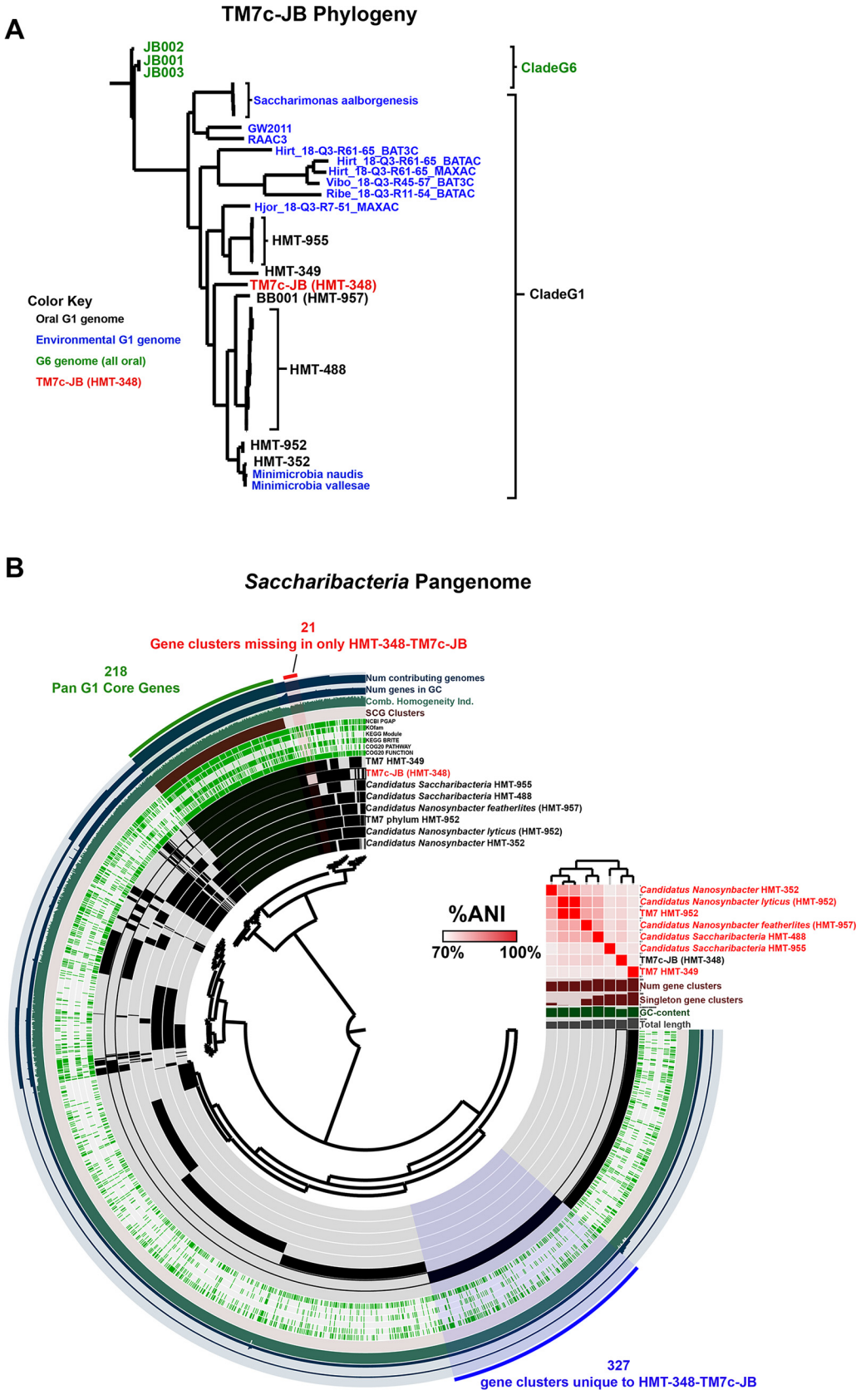

FIG 3 Phylogeny and pangenome of *Saccharibacteria* and HMT-348-TM7c-JB. (A) Updated phylogeny of complete *Saccharibacteria* genomes. Phylogenetic tree based upon concatenated protein sequences of 4 single-copy core genes

(Continued on next page)

2.21 Mbp size. The two genomes from this study had an ANI of 95–96% against the HOMD genomes with an AP of 91–92%. In the previous metagenomics study, there were 2 other *A. graevenitzii* in the SGB ANI >95% SGB with the draft genomes of JCVI-JB-Ag28 and JCVI-JB-Ag32 (27). Pangenome analysis of JCVI-JB-Ag32 and 26 *Actinomyces* complete genomes from GenBank identified 655 gene clusters unique to *A. graevenitzii* and 8 gene clusters missing in only *A. graevenitzii* (Fig. 4; pangenome summary table available at https://github.com/jonbakerlab/nanopore-oral-genomes). JCVI-JB-Ag28 was not included in the pangenome analysis due to the likely high rate of ONT errors, which would affect gene calling. The pangenome analysis indicated that *A. naeslundii* GCA002355915 is likely misidentified, as it was much more closely related to *A. oris*, rather than other genomes of *A. naeslundii*, in terms of both %ANI and gene cluster presence/absence (Fig. 4). The pangenome analysis also indicated that *A. graevenitzii* was comparatively more divergent from other *Actinomyces*, with %ANI < 76% compared to all other genomes examined. *A. graevenitzii* is frequently isolated from the oral cavity (39), occasionally from the gut (40), and also causes infections of the lung (41). Recently, *A. graevenitzii* was shown to inhibit the growth of *Streptococcus* and *Staphylococcus* while also cooperating with *Staphylococcus aureus* to evade neutrophil attack (42). Deeper examination of the genes present in *A. graevenitzii* will give insight into its metabolic capabilities and ecological role.

***Rothia mucilaginosa*.** There were two complete genomes of *Rothia mucilaginosa* obtained in this study, JCVI-JB-Rm27 from SC27 at 2,258,635 bp and 16.6× Illumina coverage, and JCVI-JB-Rm28 from SC28 at 2,259,930 bp with 5.3× Illumina coverage. Compared to each other, the two genomes obtained in this study had an ANI of 95.22% and an AP of 95.58%. Both strains have a 16S rRNA gene with 99.775% identity (3 mismatches) to the *Rothia mucilaginosa* HMT-681 reference strain, DY-18. There are currently 4 complete genomes of *Rothia mucilaginosa* on NCBI. Compared to these other *R. mucilaginosa* genomes, 27 and 28 have ANIs of ~93% and 90–92% AP. This is lower than might be expected, and may indicate the presence of multiple subspecies or genospecies within *R. mucilaginosa*. In the previous metagenomic study, there were 13 other *Rothia* bins in the SGB with the JCVI-JB-Rm27 and JCVI-JB-Rm28 draft genomes, with 11 of these having an ANI of >95% to JCVI-JB-Rm27 and JCVI-JB-Rm28 (27). Pangenome analysis was performed using JCVI-JB-Rm27 and 17 other complete *Rothia* genomes from GenBank (Fig. 5; pangenome summary table available at https://github.com/jonbakerlab/nanopore-oral-genomes). This analysis identified 1,291 gene clusters unique to *R. mucilaginosa* and 27 gene clusters missing only in *R. mucilaginosa* (Fig. 5). Compared to the other complete *R. mucilaginosa* genomes, JCVI-JB-Rm27 had 51 unique gene clusters, and there were 27 gene clusters missing in only JCVI-JB-Rm27 (Fig. 5). Although *Rothia* are known to cause various types of opportunistic infections (43), they have also received attention recently as a potential probiotic in the context of dental caries (44), as they and other nitrate-reducing oral bacteria have been associated with good dental health (27, 45).

FIG 3 Legend (Continued)
with optimal geometric and functional homogeneity indices. All complete *Saccharibacteria* genomes on GenBank (April 2022) were included, and the clade G6 genomes were used to root the tree (all remaining genomes were clade G1). To improve readability, several species-level clades were collapsed to one label. Clade G6 genomes, which are all from human oral sources, are labeled in green, human oral G1 genomes are labeled in black (except HMT-348-TM7c-JB), environmental G1 genomes are labeled in blue, and HMT-348-TM7c-JB is labeled in red. (B) Clade G1 *Saccharibacteria* pangenome. The dendrogram in the center organizes the 2,545 gene clusters identified across the genomes represented by the innermost 8 layers: TM7 HMT-349, HMT-348-TM7c-JB (HMT-348), *Candidatus* Saccharibacteria HMT-955, *Candidatus* Saccharibacteria HMT-488, *Candidatus* Nanosynbacter featherlites (HMT957), TM7 HMT-952, *Candidatus* Nanosynbacter lyticus (HMT-952), and *Candidatus* Nanosynbacter HMT-352. The data points within these 8 layers indicate the presence of a gene cluster in a given genome. From inside to outside, the next 6 layers indicate known versus unknown COG function, COG pathway, KEGG Brite, KEGG module, KOfam, and NCBI PGAP annotation. The next 4 layers indicate single copy core gene (SCG) clusters, the combined homogeneity index, the number of genes in the gene cluster, and the number of contributing genomes. The outermost layer indicates the gene clusters present in the following groups: genes missing in only HMT-348, Pan-G1 core genes, and genes unique to HMT-348. The 8 genome layers are ordered based on the tree of the %ANI comparison, which is displayed with the red and white heat map. The layers underneath the %ANI heat map, from top to bottom, indicate number of gene clusters, number of singleton gene clusters, GC content, and total length of each genome.

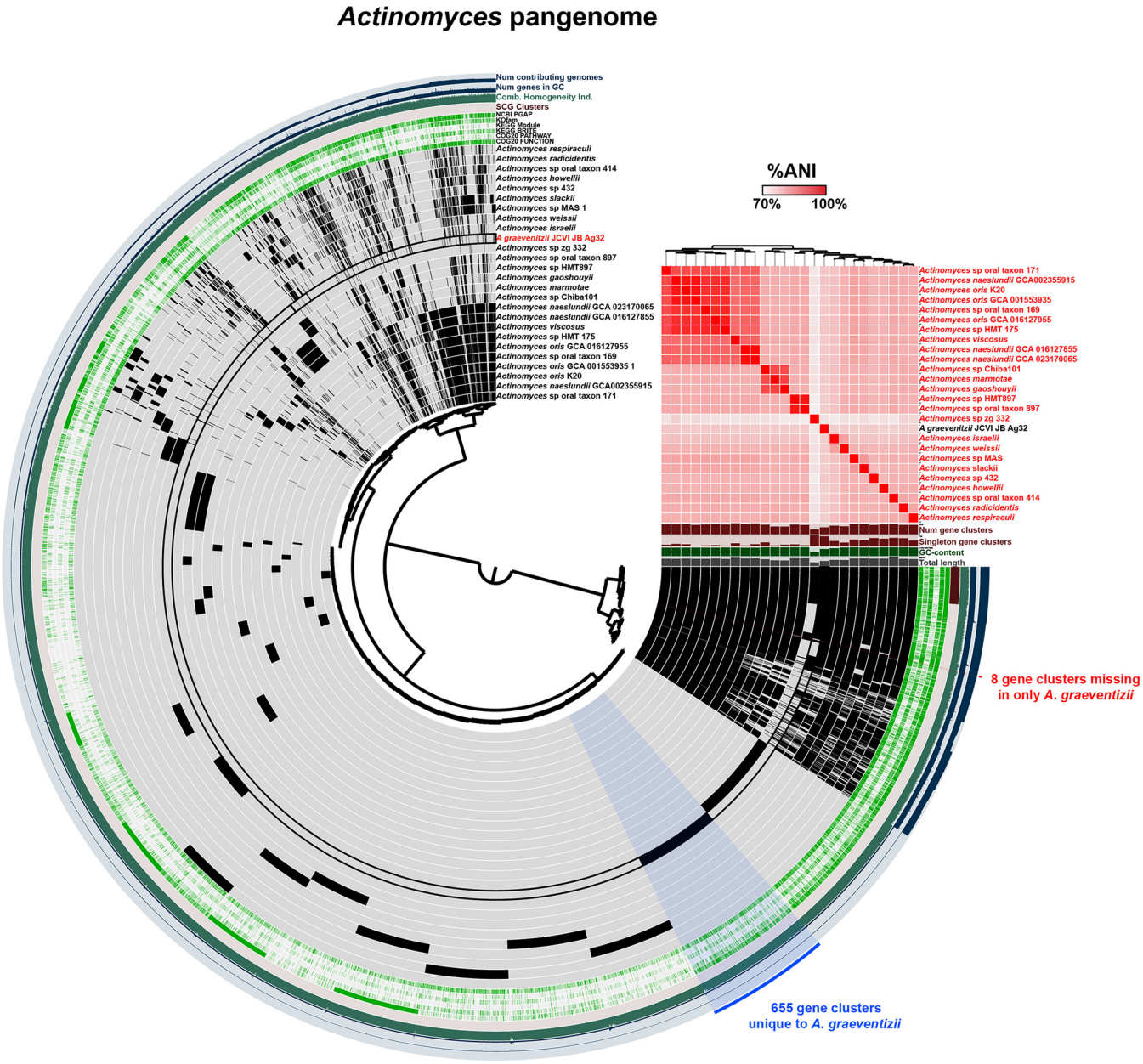

FIG 4 The *Actinomyces* pangenome. The dendrogram in the center organizes the 12,168 gene clusters identified across the indicated genomes represented by the innermost 26 layers. The data points within these 26 layers indicate the presence of a gene cluster in a given genome. From inside to outside, the next 6 layers indicate known versus unknown COG function, COG pathway, KEGG Brite, KEGG module, KOfam, and NCBI PGAP annotation. The next 4 layers indicate single copy core gene (SCG) clusters, the combined homogeneity index, the number of genes in the gene cluster, and the number of contributing genomes. The outermost layer indicates the gene clusters present in the following groups: gene clusters missing in only *A. graevenitzii*, and gene clusters unique to *A. graevenitzii*. The 26 genome layers are ordered based on the tree of the %ANI comparison, which is displayed with the red and white heat map. The layers underneath the %ANI heat map, from top to bottom, indicate number of gene clusters, number of singleton gene clusters, GC content, and total length of each genome.

Mogibacterium diversum. From SC33, a circular 1,770,007 bp genome of *Mogibacterium diversum*, JCVI-JB-Md32, was obtained. The 16S rRNA gene from JCVI-JB-Md32 had 100% identity to *Mogibacterium diversum* HMT-593 strain ATCC 700923. JCVI-JB-Md32 had an ANI of 96.08% and AP of 89.40% to *M. diversum* strain CCUG47132, the only other complete *M. diversum* genome on HOMD and NCBI. JCVI-JB-Md32 had excellent coverage with the Illumina reads at 837×. *M. diversum* belongs to the very poorly understood family, Eubacteriales Family XIII, *insertae sedis*, which also contains the prevalent oral resident, (Eubacterim) *sulci*. Interestingly, the JCVI-JB-Md32 draft genome had ANI >95% with 10 other genomes in the same SGB in the previous study (27). Having 10

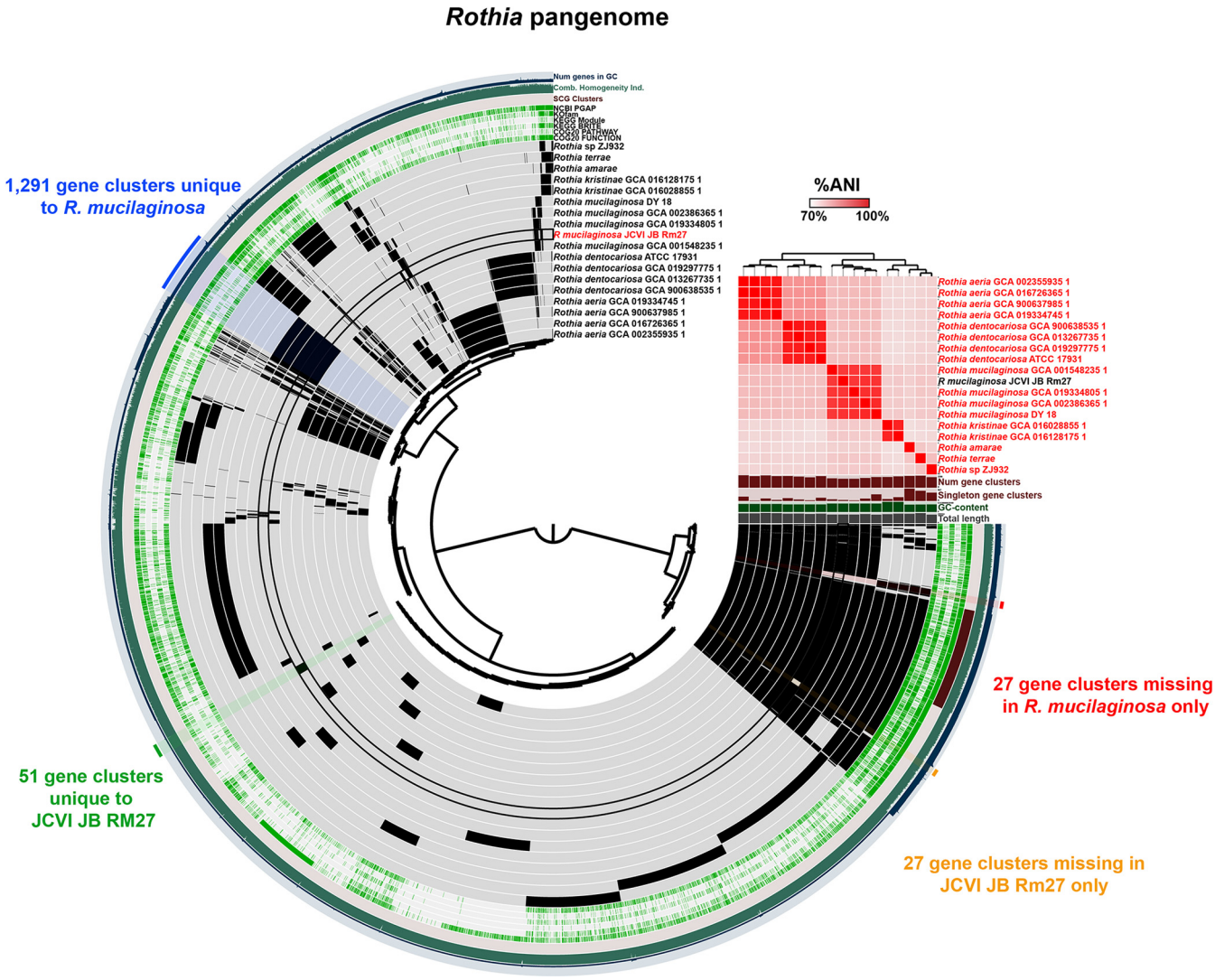

FIG 5 The *Rothia* pangenome. The dendrogram in the center organizes the 8,118 gene clusters identified across the indicated genomes represented by the innermost 18 layers. The data points within these 18 layers indicate the presence of a gene cluster in a given genome. From inside to outside, the next 6 layers indicate known versus unknown COG function, COG pathway, KEGG Brite, KEGG module, KOfam, and NCBI PGAP annotation. The next 4 layers indicate single copy core gene (SCG) clusters, the combined homogeneity index, the number of genes in the gene cluster, and the number of contributing genomes. The outermost layer indicates the gene clusters present in the following groups: gene clusters missing in only *R. mucilaginosa*, gene clusters unique to *R. mucilaginosa*, gene clusters unique to JCVI-JB-Rm27, and gene clusters missing in only JCVI-JB-Rm27. The 18 genome layers are ordered based on the tree of the %ANI comparison, which is displayed with the red and white heat map. The layers underneath the %ANI heat map, from top to bottom, indicate number of gene clusters, number of singleton gene clusters, GC content, and total length of each genome.

genomes of this species independently assembled and binned from a study of 47 human subjects indicates that this taxon is rather prevalent, despite being relatively unstudied. Pangenome analysis of JCVI-JB-Md32 and all 6 other complete genomes within this family was performed (Fig. 6; pangenome summary table available at https://github.com/jonbakerlab/nanopore-oral-genomes). This analysis identified 743 gene clusters unique to *M. diversum* and 162 gene clusters missing in only *M. diversum*. Compared to CCUG47132, JCVI-JB-Md32 had 163 unique gene clusters and only 3 gene clusters missing (Fig. 6). These complete genome and pangenome data will be a useful resource in the study of *M. diversum*, as only 1 study of this common member of the oral flora has been reported (46).

***Lancefieldella parvula*.** JCVI-JB-Lp32, a complete genome of *Lancefieldella parvula* (formerly known as *Atopobium parvulum*), was 1,624,536 bp. The JCVI-JB-Lp32 16S rRNA gene is 99.544% identical (6 mismatches) to *Lancefieldella parvula* HMT-723 strain ATCC22793. One other complete genome of *L. parvula*, IPP1246, is available and was

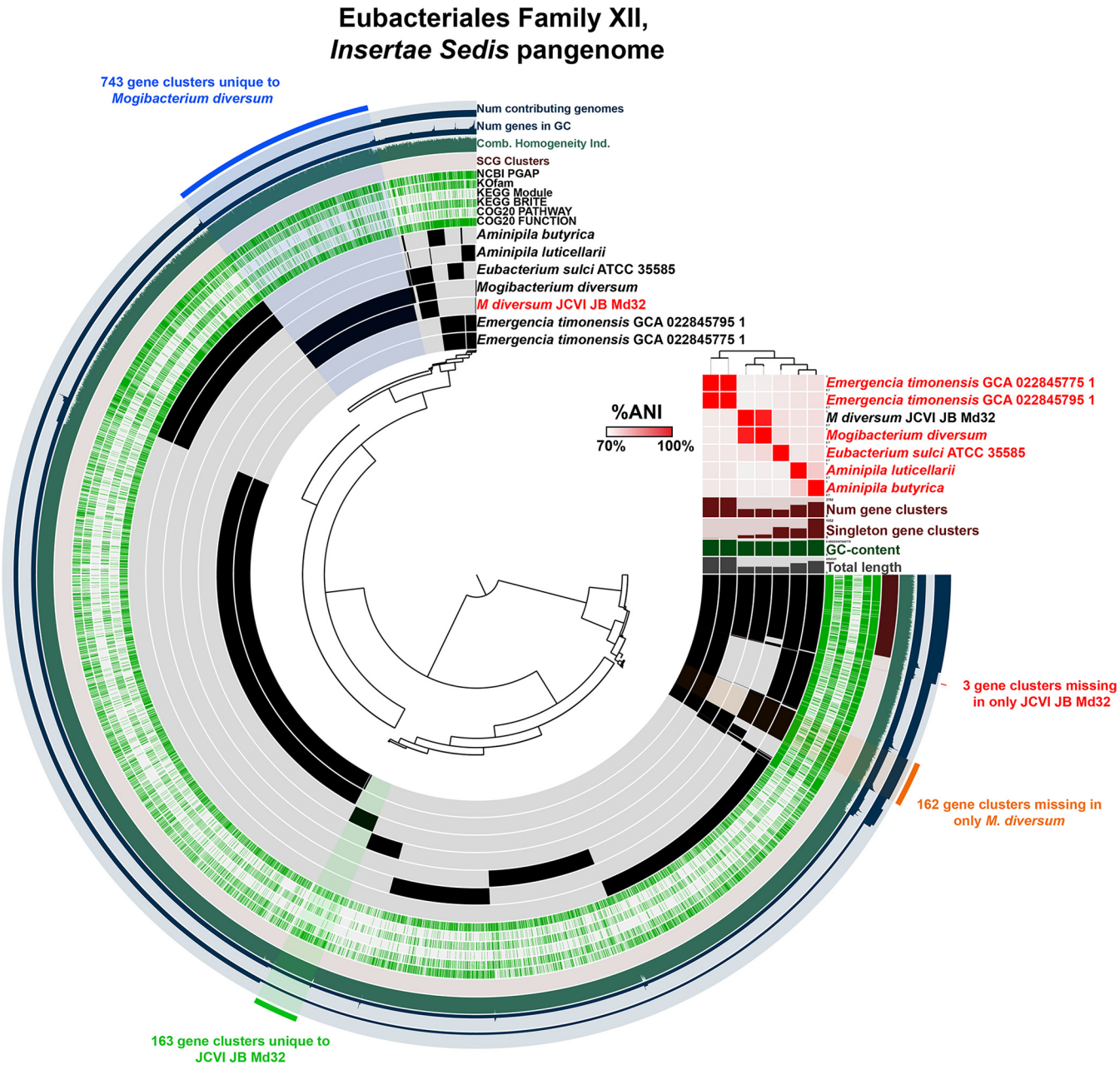

FIG 6 The Eubacteriales Family XIII, *insertae sedis* pangenome. The dendrogram in the center organizes the 8,084 gene clusters identified across the indicated genomes represented by the innermost 7 layers. The data points within these 7 layers indicate the presence of a gene cluster in a given genome. From inside to outside, the next 6 layers indicate known versus unknown COG function, COG pathway, KEGG Brite, KEGG module, KOfam, and NCBI PGAP annotation. The next 4 layers indicate single copy core gene (SCG) clusters, the combined homogeneity index, the number of genes in the gene cluster, and the number of contributing genomes. The outermost layer indicates the gene clusters present in the following groups: gene clusters missing in only *M. diversum*, gene clusters unique to *M. diversum*, gene clusters unique to JCVI-JB-Md32, and gene clusters missing in only JCVI-JB-Md32. The 7 genome layers are ordered based on the tree of the %ANI comparison, which is displayed with the red and white heat map. The layers underneath the %ANI heat map, from top to bottom, indicate number of gene clusters, number of singleton gene clusters, GC content, and total length of each genome.

published in 2009 (47). JCVI-JB-Lp32 and IPP1246 have an ANI of 88.7% over an AP of 84.13%. It is interesting that although the genome assembled here and the NCBI reference genome have an extremely high 16S rRNA identity, and certainly are the same species based upon 16S rRNA sequence, the overall genome ANI is significantly lower than 95%, which is atypical for genomes from the same species. Notably, although the ANI to IPP1246 genome was lower, in the original metagenomics study with the Illumina draft assembly of JCVI-JB-Lp32, there were 22 genomes in the same SGB with

ANI > 95% to JCVI-JB-Lp32 (27). This indicates that perhaps these 22 draft genomes and JCVI-JB-Lp32 are a distinct subspecies compared to IP1246, and that *L. parvula* has a relatively plastic genome, or that perhaps these genomes are a separate species entirely. Pangenome analysis of JCVI-JB-Lp32 and 6 other Atopobiaceae genomes identified 327 gene clusters unique to *L. parvula* and 62 gene clusters missing in only *L. parvula* (Fig. 7; pangenome summary table available at: https://github.com/jonbakerlab/nanopore-oral-genomes). Compared to IPP1246, JCVI-JB-Lp32 had 178 unique gene clusters and was missing 50 gene clusters uniquely compared to all other Atopobiaceae analyzed (Fig. 7). In the oral cavity, *L. parvulum* has been associated with occlusal lesions in caries (48), or with a healthy microbiome, in the context of periodontal disease (49). In the gut, *L. parvulum* has been associated with colorectal cancer (50) and the onset of Crohn's disease, due to its ability to produce H2S through the SufS cysteine desulfurase (51, 52).

Lachnospiraeceae HMT-096. JCVI-JB-L28 was assembled from SC28, was 2,401,457 bp, and had a 16S rRNA gene sequence 99.7% identical (3 mismatches) to Lachnospiraceae (G-2) HMT-096 from HOMD. There is one complete genome from Lachnospiraceae (G-2) HMT-096 on NCBI, CP073340, deposited in 2021 by The Forsyth Institute, which has an ANI of 98.585% to the JCVI-JB-L28 over an AP of 87.79%. As this genome did not have good coverage by the Illumina reads (0.3× coverage), this assembly currently represents a draft genome that was not further analyzed, and likely would have a higher ANI to the published genome with increased short-read coverage for polishing. Lachnospiraceae HMT-096 is very poorly understood, but its abundance in saliva was positively correlated with hemodialysis in patients with chronic and end-stage kidney disease (53).

Ruminococcaceae HMT-075 chr1. The 1,207,213 bp JCVI-JB-R28 Ruminococcaceae chromosome obtained had a 16S rRNA gene with 99.773% identity (3 mismatches) to Ruminococcaceae (G-1) bacterium HMT-075 clone F058. There are currently no genomes of this taxon on HOMD or NCBI. Due to the small size of this genome compared to other Ruminococcaceae, which typically have 2–4 Mbp genomes, and the fact that some other Ruminococcaceae genomes have two chromosomes (54), it is likely that this represents a single chromosome of a taxon with multiple chromosomes. Phylophlan3 (55) was used to examine all other large contigs (>700,000 bp, linear and circular) from SC28, to determine if they may represent the other chromosomes of this taxon; however, none were predicted to be from the same taxon. Although this is only an incomplete genome due to the fact that there is only one chromosome, it does represent the first genome sequence to be associated with the 16S sequence for Ruminococcaceae HMT-075, a taxon that almost nothing is known about other than its existence.

Conclusions. This study serves as proof of concept, and provides a protocol, for obtaining complete genomes from metagenomes derived from human saliva using complementary ONT and Illumina sequencing. This represents a major advance, as obtaining complete genomes previously required isolation of the microbe and/or manual PCR/sequencing steps due to repeat regions. In this manner, obtaining complete genomes, which include the 16S rRNA regions that are typically not accurately assembled and binned from Illumina reads alone, will finally provide genomes for the large number of taxa that are currently only identified by 16S rRNA sequences. The taxa for which complete genomes were obtained here were quite varied in their relative abundance in the saliva samples (anywhere from the 3rd most abundant to the 83rd most abundant taxa in the corresponding samples [27]), indicating that features other than abundance and coverage depth, including read length, read quality, the complexity of the genome in question, and the number of closely related species or strains present, are also crucial factors influencing whether a complete genome will be obtained for a given taxon in an ONT metagenomic assembly. Polypolish is highly effective at removing errors in ONT assemblies. Across 120 NCBI complete reference genomes, Polypolish reduced the average number of errors per genome from 3,266 in an unpolished long read assembly to 41, with an average of only 2 errors remaining outside of repeat regions (33). Its utility was further confirmed here, resulting in the fewest number of truncated or missing open reading frames of all pipelines tested on TM7c-JB while also assembling a 16S rRNA region that was 99.77% identical to the NCBI reference (which was not possible using Illumina sequencing alone). Deeper sequencing

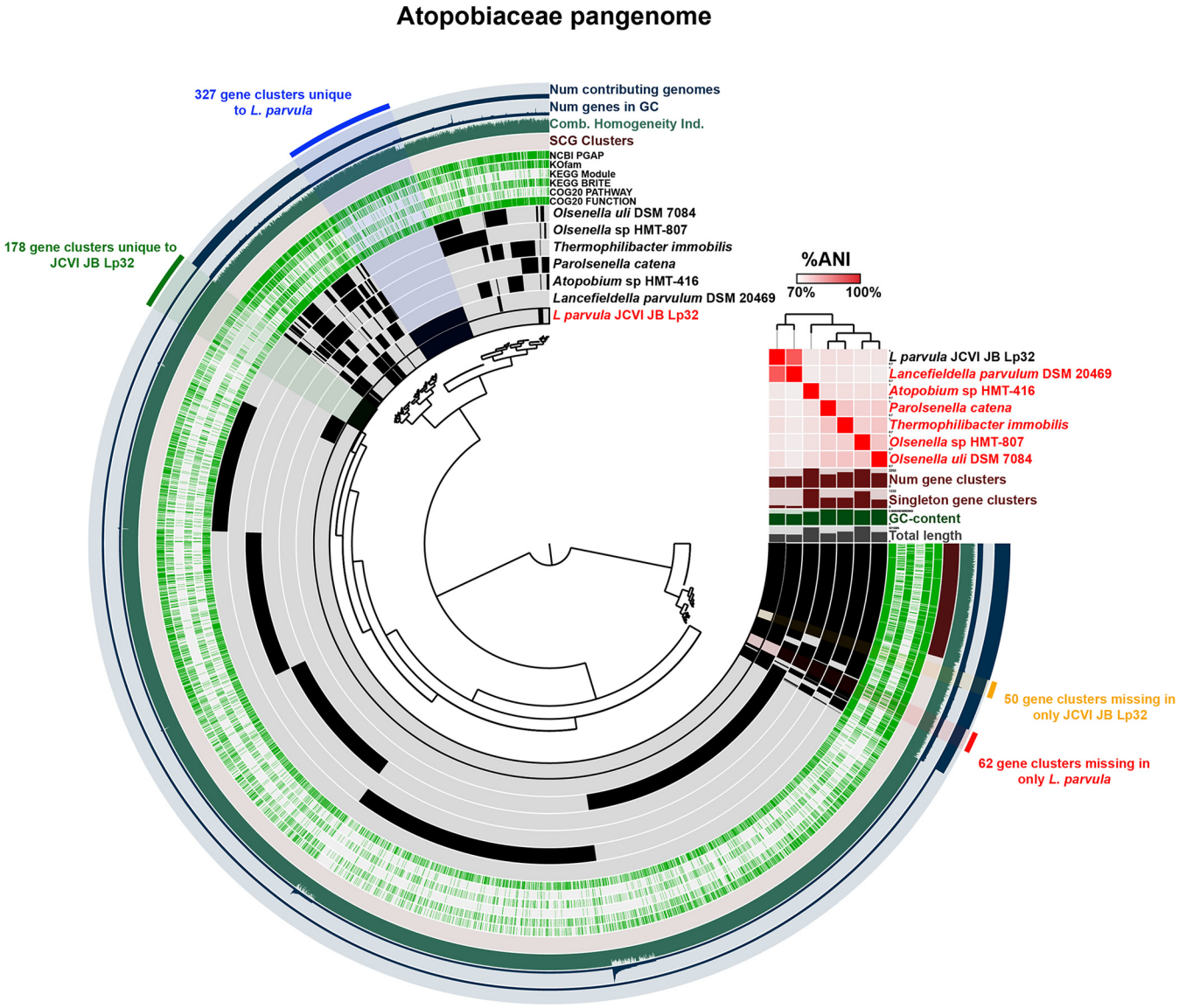

FIG 7 The Atopobiceae pangenome. The dendrogram in the center organizes the 6,488 gene clusters identified across the indicated genomes represented by the innermost 7 layers. The data points within these 7 layers indicate the presence of a gene cluster in a given genome. From inside to outside, the next 6 layers indicate known versus unknown COG function, COG pathway, KEGG Brite, KEGG module, KOfam, and NCBI PGAP annotation. The next 4 layers indicate single copy core gene (SCG) clusters, the combined homogeneity index, the number of genes in the gene cluster, and the number of contributing genomes. The outermost layer indicates the gene clusters present in the following groups: gene clusters missing in only *L. parvula*, gene clusters unique to *L. parvula*, gene clusters unique to JCVI-JB-Lp32, and gene clusters missing in only JCVI-JB-Lp32. The 7 genome layers are ordered based on the tree of the %ANI comparison, which is displayed with the red and white heat map. The layers underneath the %ANI heat map, from top to bottom, indicate: number of gene clusters, number of singleton gene clusters, GC content, and total length of each genome.

and improved methods for obtaining ultra-high molecular weight gDNA are likely to produce even more circular assemblies per sample. The information provided by the genomes obtained here will be useful to help determine the metabolic capabilities, ecological roles, and pathogenic potential of the cognate bacterial species, particularly those that have not been isolated or cultivated.

MATERIALS AND METHODS

DNA extraction. As described in Results and Discussion, several methods were attempted to optimize extraction of HMW gDNA: (i) the Chen and Burne phenol:chloroform method described in reference 25; (ii) the open-source Bio-On-Magnetic-Beads (BOMB) gDNA Extraction using guanidine isothiocyanate (GITC) lysis and purification using silane magnetic beads (detailed protocol available at bomb.bio); (iii) the lysis steps from the Chen and Burne protocol (25), followed by purification using silane beads (instead of phenol:

chloroform), and (iv) the Monarch Genomic DNA purification kit from New England Biolabs (using manufacturer's instructions). The first approach (i) gave the best combination of yield and size of HMW gDNA, and is therefore reported in full detail here (gDNA chromatograms and concentrations from all 4 methods are available on GitHub: https://github.com/jonbakerlab/nanopore-oral-genomes/blob/main/Supplemental-File-S1.pdf and https://github.com/jonbakerlab/nanopore-oral-genomes/blob/main/Supplemental-File-S2.pdf): Saliva (545 μL; from a frozen aliquot) was added to 545 μL Tris-EDTA (TE) buffer (100 mM, 20 mM) containing 10 mg/mL lysozyme and 300 U/mL mutanolysin, and was incubated for 30 min at 37°C. One hundred μL 20% SDS was added, and the sample was incubated at 65°C for 15 min. Three hundred μL TE (100 mM, 20 mM) was added, and the whole volume of lysate was transferred to a 2-mL screwcap tube containing 0.1-mm glass beads. The sample was homogenized for 30 s on a FastPrep-24 homogenizer (MP Biomedicals), and the lysate (none of the foam) was transferred to a new 1.5 microcentrifuge tube and cooled to 37°C. Two μL of proteinase K was added and the sample was incubated for 30 min at 37 C. One hundred μL of 5M NaCl was added, followed by 80 μL of 10 cetyltrimethylammonium bromide (CTAB) in 0.7M NaCl that had been warmed to 65°C. The sample was then incubated at 65°C for 20 min. Seven hundred fifty μL phenol:chloroform:isoamyl alcohol (25:24:1) was added, and the solution was mixed by inversion and then centrifuged at 12,000 $\times$ *g* for 1 min. The aqueous phase was extracted and extractions were repeated until the white interface between the aqueous and organic layers was gone (typically 2–3 exactions). The aqueous phase was then extracted once with 750 μL chloroform:isoamyl alcohol (24:1). Seven hundred fifty μL ice-cold 100% isopropanol was added and the sample was centrifuged at 12,000 $\times$ *g* for 30 min. The pellet was washed with 70% ethanol then rinsed with 100% ethanol. The pellet was then resuspended in 100 μL water. An additional final ethanol precipitation step was performed. Ten μL 3M sodium acetate pH 5.2 and 250 μL of ice-cold 100% ethanol were added to the sample, and the sample was centrifuged at 12,000 $\times$ *g* for 10 min at 4°C. The supernatant was decanted, then 250 μL of 70% was added, and the sample was centrifuged for 5 min at 12,000 $\times$ *g*. The supernatant was decanted and the pellet was dried in a speed-vac. The final pellet was resuspended in 50 μL of molecular-grade water. The quality of the DNA was checked using a TapeStation (Agilent Technologies, Santa Clara, CA, USA) and Qubit Fluorimeter (Thermo Fisher Scientific, Waltham, MA, USA).

Illumina sequencing. The short-read sequencing was performed as previously described (27). Briefly, libraries were generated using the Nextera XT DNA Library Preparation Kit (Illumina, Inc., San Diego, CA, USA) using the manufacturer's instructions, and the sequencing run was performed on a NextSeq500 (Illumina, Inc., San Diego, CA, USA).

Nanopore sequencing. ONT sequencing was performed as previously described (22, 23, 26). Briefly, the long-read library was prepared using a Ligation Sequencing Kit (Oxford Nanopore Technologies, Oxford, UK) and sequenced on a GridION using an R9.4.1 flow cell (Oxford Nanopore Technologies, Oxford, UK). Base calling, quality control, and adapter trimming were performed using Guppy v4.0.11/MinKNOW v20.06.9 (Oxford Nanopore Technologies, Oxford, UK).

Genome assembly. (i) JB001, JB002, JB003. Two independent methods generated improved draft assemblies (compared to the draft assemblies published in Baker et al., 2021 (27)). (i) Human reads were removed from the long-read assemblies using minimap2 v2.17-r941 (28), and the remaining long reads were assembled using metaFlye v2.8-b1674 (29). MegaBLAST v2.2.26 (56) was used to identify the circular contigs of interest within the metagenome assemblies. (ii) Long-reads mapping to the draft genomes of JB001, JB002, and JB003 were extracted using minimap2. These long reads, along with the short reads used to generate the original JB001, JB002, and JB003 draft assemblies, were used by Unicycler v0.4.8 (30) to obtain draft genomes. Short contigs in the Unicycler assemblies were removed based on disparate GC content, coverage, and BLAST hits to other organisms (anvi'o v7-dev) (38), leaving single circular contigs. Trycycler v0.3.0 (31) was used to develop consensus assemblies from the draft assemblies. The resulting assemblies were polished using Medaka v1.0.3 (https://github.com/nanoporetech/medaka), then Pilon v1.23 (32). Circlator v1.5.5 (57) was used to rotate the genome sequences such that the start sites were at the *dnaA* gene. Default parameters were used unless otherwise noted. JB001, JB002, and JB003 were annotated initially using Prokka (58), while the final genomes submitted to NCBI were annotated using the NCBI Prokaryotic Genome Annotation Pipeline v5.1. As noted above in Results and Discussion and in reference 23, further examination indicated the Medaka and Pilon polishing steps introduced errors into the rRNA regions, but properly removed errors from other locations in the genomes. Therefore, the rRNA versions were manually corrected to those obtained from the Flye assemblies.

(ii) HMT-348-TM7c-JB. The same approaches just detailed for JB001, JB002, and JB003 were also used for HMT-348-TM7c-JB. In addition, Polypolish (v0.4.3) (33), which was published during preparation of this study, was attempted in parallel on the HMT-348-TM7c-JB contig from the metaFlye assembly. The Polypolish tool utilized the short reads from the cognate short-read libraries that had mapped to the metaFlye draft contig. 16S rRNA sequences were compared using the HOMD 16S rRNA Sequence Identification tool (https://www.homd.org/refseq/refseq_blastn). Disrupted and missing ORFs were identified, and %ANI and %AP were calculated using CLC Genomics Workbench v21.0.3 (Qiagen, Inc., MD, USA). The results obtained from metaFlye followed by Polypolish closely resembled those from the final composite methods used in the JB001, JB002, and JB003 final assemblies (i.e., rRNA regions from metaFlye, remainder of the genome from Unicycler/Flye/Trycycler/Medaka/Pilon), but was much less computationally and temporally expensive (Table 2). Therefore, this pipeline, metaFlye, followed by Polypolish, was used for all other genomes in this study.

(iii) All other genomes. metaFlye was used to assemble the ONT metagenomic read libraries. Short reads were mapped to the circular contigs representing the draft genomes of interest using BWA-MEM (59), and Polypolish was used for error correction. The start sequences were rotated to *dnaA* using Circlator and annotated using NCBI PGAP.

Phylogenomics and pangenomics. The anvi'o (v-dev) pangenomics workflow (38, 60, 61) was implemented using Snakemake (62) and used to perform the pangenomics analysis. For the phylogenetic analysis of *Saccharibacteria*, HMT-348-TM7c-JB and HMT-348, a pangenome including HMT-348-TM7c-JB and 39 other complete *Saccharibacteria* genomes (all complete nonduplicate *Saccharibacteria* genomes on NCBI as of April 2022), were created. Only 12 single-copy core genes were common to all 40 genomes. To minimize the effect of gaps on phylogeny, the minimum geometric homogeneity index was set to 0.95, and a maximum functional homogeneity index was set to 0.85 to ensure nearly identical protein sequences were not used. This left 4 genes, the ribosomal protein subunits L6 and L27, SecG, and a peptide deformylase. Concatenated protein sequences of these 4 genes were used to construct a phylogenetic tree of all 40 complete *Saccharibacteria* genomes on NCBI using anvi'o.

Data availability. The short reads used to generate the assemblies were published previously (27) and are available in the SRA database with the accession numbers SRX4318838, SRX4318837, and SRX4318835. The long reads used to assemble the metagenomes are available in the SRA database with accession numbers SRX15103396, SRX15103397, and SRX15103398. The complete genomes of JCVI-JB-Rm27, JCVI-JB-Ag32, JCVI-JB-Md32, JCVI-JB-Lp32, and HMT-348-TM7c-JB are available on NCBI GenBank as accession numbers CP097094, CP097095, CP097093, CP097092, and CP090820, respectively. Since JCVI-JB-L28, JCVI-JB-Ag28, and JCVI-JB-Rm28 have low Illumina coverage and are likely to contain ONT errors, and JCVI-JB-R28-chr1 is likely not a complete genome, these sequences are available at https://github.com/jonbakerlab/nanopore-oral-genomes. The other draft assemblies of HMT-348-TM7c-JB (metaSPAdes, Unicycler, Trycycler, Medaka, Pilon, and metaFlye) and the Polypolish versions of JB001 and JB003 are also available at https://github.com/jonbakerlab/nanopore-oral-genomes. The pangenome data tables are too large (some >50MB) to be included as supplemental material, and therefore have been made available at https://github.com/jonbakerlab/nanopore-oral-genomes.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TABLE S1, XLSX file, 0.2 MB.

TABLE S2, XLSX file, 0.3 MB.

TABLE S3, XLSX file, 0.2 MB.

ACKNOWLEDGMENTS

I thank Karrie Goglin-Alemeida, Jelena Jablanovic, and Kara Riggsbee for performing the library preparation and sequencing. This research was supported by NIH/NIDCR K99-DE029228.

REFERENCES

1. Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. https://doi.org/10.1038/nature11234.
2. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. 2006. Metagenomic analysis of the human distal gut microbiome. Science 312:1355–1359. https://doi.org/10.1126/science.1124234.
3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74. https://doi.org/10.1126/science.1093857.
4. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512. https://doi.org/10.1126/science.7542800.
5. Naito M, Ogura Y, Itoh T, Shoji M, Okamoto M, Hayashi T, Nakayama K. 2016. 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 23:11–19.
6. Bennett S. 2004. Solexa Ltd. Pharmacogenomics 5:433–438. https://doi.org/10.1517/14622416.5.4.433.
7. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59. https://doi.org/10.1038/nature07517.
8. Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC, Scorilas A. 2021. Third-generation sequencing: the spearhead towards the radical transformation of modern genomics. Life (Basel) 12:30. https://doi.org/10.3390/life12010030.
9. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146. https://doi.org/10.1038/nmeth.3103.
10. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:e7359. https://doi.org/10.7717/peerj.7359.
11. Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607. https://doi.org/10.1093/bioinformatics/btv638.
12. Chen LX, Anantharaman K, Shaiber A, Eren AM, Banfield JF. 2020. Accurate and complete genomes from metagenomes. Genome Res 30:315–333. https://doi.org/10.1101/gr.258640.119.
13. Shaiber A, Eren AM. 2019. Composite metagenome-assembled genomes reduce the quality of public genome repositories. mBio 10:e00725-19. https://doi.org/10.1128/mBio.00725-19.
14. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. 2015. Improved data analysis for the MinION nanopore sequencer. Nat Methods 12:351–356. https://doi.org/10.1038/nmeth.3290.
15. Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. Nat Methods 12:733–735. https://doi.org/10.1038/nmeth.3444.
16. Cusco A, Perez D, Vines J, Fabregas N, Francino O. 2021. Long-read metagenomics retrieves complete single-contig bacterial genomes from canine feces. BMC Genomics 22:330. https://doi.org/10.1186/s12864-021-07607-0.

17. Moss EL, Maghini DG, Bhatt AS. 2020. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol 38:701–707. https://doi.org/10.1038/s41587-020-0422-6.
18. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Chen NC, Cheng H, Chin CS, Chow W, de Lima LG, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, Fulton RS, Fungtammasan A, Garrison E, Grady PGS, Graves-Lindsay TA, Hall IM, Hansen NF, Hartley GA, Haukness M, Howe K, Hunkapiller MW, Jain C, Jain M, Jarvis ED, Kerpedjiev P, Kirsche M, Kolmogorov M, Korlach J, Kremitzki M, Li H, et al. 2022. The complete sequence of a human genome. Science 376:44–53. https://doi.org/10.1126/science.abj6987.
19. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21:30. https://doi.org/10.1186/s13059-020-1935-5.
20. Liu C, Yang X, Duffy BF, Hoisington-Lopez J, Crosby M, Porche-Sorbet R, Saito K, Berry R, Swamidass V, Mitra RD. 2021. High-resolution HLA typing by long reads from the R10.3 Oxford nanopore flow cells. Hum Immunol 82:288–295. https://doi.org/10.1016/j.humimm.2021.02.005.
21. Matsuo Y, Komiya S, Yasumizu Y, Yasuoka Y, Mizushima K, Takagi T, Kryukov K, Fukuda A, Morimoto Y, Naito Y, Okada H, Bono H, Nakagawa S, Hirota K. 2021. Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION nanopore sequencing confers species-level resolution. BMC Microbiol 21:35. https://doi.org/10.1186/s12866-021-02094-5.
22. Baker JL. 2021. Complete genome sequence of strain JB001, a member of *Saccharibacteria* Clade G6 ("*Candidatus* Nanogingivalaceae"). Microbiol Resour Announc 10:e0051721. https://doi.org/10.1128/MRA.00517-21.
23. Baker JL. 2021. Complete genomes of clade G6 *Saccharibacteria* suggest a divergent ecological niche and lifestyle. mSphere 6:e0053021. https://doi.org/10.1128/mSphere.00530-21.
24. Baker JL. 2022. Complete genome sequence of "*Candidatus* Nanosynbacter" strain HMT-348_TM7c-JB, a member of *Saccharibacteria* clade G1. Microbiol Resour Announc 11:e0002322. https://doi.org/10.1128/mra.00023-22.
25. Chen YY, Clancy KA, Burne RA. 1996. *Streptococcus salivarius* urease: genetic and biochemical characterization and expression in a dental plaque streptococcus. Infect Immun 64:585–592. https://doi.org/10.1128/iai.64.2.585-592.1996.
26. Baker JL, Edlund A. 2020. Composite long- and short-read sequencing delivers a complete genome sequence of B04Sm5, a reutericyclin- and mutanocyclin-producing strain of *Streptococcus mutans*. Microbiol Resour Announc 9:e01067-20. https://doi.org/10.1128/MRA.01067-20.
27. Baker JL, Morton JT, Dinis M, Alvarez R, Tran NC, Knight R, Edlund A. 2021. Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. Genome Res 31:64–74. https://doi.org/10.1101/gr.265645.120.
28. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
29. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, Pevzner PA. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods 17:1103–1110. https://doi.org/10.1038/s41592-020-00971-x.
30. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
31. Wick RR, Judd LM, Cerdeira LT, Hawkey J, Meric G, Vezina B, Wyres KL, Holt KE. 2021. Trycycler: consensus long-read assemblies for bacterial genomes. Genome Biol 22:266. https://doi.org/10.1186/s13059-021-02483-z.
32. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal.pone.0112963.
33. Wick RR, Holt KE. 2022. Polypolish: short-read polishing of long-read bacterial genome assemblies. PLoS Comput Biol 18:e1009802. https://doi.org/10.1371/journal.pcbi.1009802.
34. McLean JS, Bor B, Kerns KA, Liu Q, To TT, Solden L, Hendrickson EL, Wrighton K, Shi W, He X. 2020. Acquisition and adaptation of ultra-small parasitic reduced genome bacteria to mammalian hosts. Cell Rep 32:107939. https://doi.org/10.1016/j.celrep.2020.107939.
35. Shaiber A, Willis AD, Delmont TO, Roux S, Chen LX, Schmid AC, Yousef M, Watson AR, Lolans K, Esen OC, Lee STM, Downey N, Morrison HG, Dewhirst FE, Mark Welch JL, Eren AM. 2020. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biol 21:292. https://doi.org/10.1186/s13059-020-02195-w.
36. Marcy Y, Ouverney C, Bik EM, Losekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, Quake SR. 2007. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A 104:11889–11894. https://doi.org/10.1073/pnas.0704662104.
37. Cross KL, Campbell JH, Balachandran M, Campbell AG, Cooper SJ, Griffen A, Heaton M, Joshi S, Klingeman D, Leys E, Yang Z, Parks JM, Podar M. 2019. Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. Nat Biotechnol 37:1314–1321. https://doi.org/10.1038/s41587-019-0260-6.
38. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN, Yousef M, Fogarty EC, Trigodet F, Watson AR, Esen OC, Moore RM, Clayssen Q, Lee MD, Kivenson V, Graham ED, Merrill BD, Karkman A, Blankenberg D, Eppley JM, Sjodin A, Scott JJ, Vazquez-Campos X, McKay LJ, McDaniel EA, Stevens SLR, Anderson RE, Fuessel J, Fernandez-Guerra A, Maignien L, Delmont TO, Willis AD. 2021. Community-led, integrated, reproducible multi-omics with anvi'o. Nat Microbiol 6:3–6. https://doi.org/10.1038/s41564-020-00834-3.
39. Wang Y, Wang S, Wu C, Chen X, Duan Z, Xu Q, Jiang W, Xu L, Wang T, Su L, Wang Y, Chen Y, Zhang J, Huang Y, Tong S, Zhou C, Deng S, Qin N. 2019. Oral microbiome alterations associated with early childhood caries highlight the importance of carbohydrate metabolic activities. mSystems 4:e00450-19. https://doi.org/10.1128/mSystems.00450-19.
40. Ruigrok R, Collij V, Sureda P, Klaassen MAY, Bolte LA, Jansen BH, Voskuil MD, Fu J, Wijmenga C, Zhernakova A, Weersma RK, Vich Vila A. 2021. The composition and metabolic potential of the human small intestinal microbiota within the context of inflammatory bowel disease. J Crohns Colitis 15:1326–1338. https://doi.org/10.1093/ecco-jcc/jjab020.
41. Gliga S, Devaux M, Gosset Woimant M, Mompoint D, Perronne C, Davido B. 2014. *Actinomyces graevenitzii* pulmonary abscess mimicking tuberculosis in a healthy young man. Can Respir J 21:e75–e77. https://doi.org/10.1155/2014/841480.
42. Jalali F, Ellett F, Balani P, Duncan MJ, Dewhirst FE, Borisy GG, Irimia D. 2021. No man's land: species-specific formation of exclusion zones bordering *Actinomyces graevenitzii* microcolonies in nanoliter cultures. Microbiologyopen 10:e1137. https://doi.org/10.1002/mbo3.1137.
43. Fatahi-Bafghi M. 2021. Characterization of the *Rothia* spp. and their role in human clinical infections. Infect Genet Evol 93:104877. https://doi.org/10.1016/j.meegid.2021.104877.
44. Rosier BT, Moya-Gonzalvez EM, Corell-Escuin P, Mira A. 2020. Isolation and characterization of nitrate-reducing bacteria as potential probiotics for oral and systemic health. Front Microbiol 11:555465. https://doi.org/10.3389/fmicb.2020.555465.
45. Agnello M, Marques J, Cen L, Mittermuller B, Huang A, Chaichanasakul Tran N, Shi W, He X, Schroth RJ. 2017. Microbiome associated with severe caries in Canadian First Nations children. J Dent Res 96:1378–1385. https://doi.org/10.1177/0022034517718819.
46. Nakazawa F, Poco SE, Sato M, Ikeda T, Kalfas S, Sundqvist G, Hoshino E. 2002. Taxonomic characterization of *Mogibacterium diversum* sp. nov. and *Mogibacterium neglectum* sp. nov., isolated from human oral cavities. Int J Syst Evol Microbiol 52:115–122. https://doi.org/10.1099/00207713-52-1-115.
47. Copeland A, Sikorski J, Lapidus A, Nolan M, Del Rio TG, Lucas S, Chen F, Tice H, Pitluck S, Cheng JF, Pukall R, Chertkov O, Brettin T, Han C, Detter JC, Kuske C, Bruce D, Goodwin L, Ivanova N, Mavromatis K, Mikhailova N, Chen A, Palaniappan K, Chain P, Rohde M, Goker M, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk HP, Detter JC. 2009. Complete genome sequence of *Atopobium parvulum* type strain (IPP 1246). Stand Genomic Sci 1:166–173. https://doi.org/10.4056/sigs.29547.
48. Fakhruddin KS, Samaranayake LP, Hamoudi RA, Ngo HC, Egusa H. 2022. Diversity of site-specific microbes of occlusal and proximal lesions in severe- early childhood caries (S-ECC). J Oral Microbiol 14:2037832. https://doi.org/10.1080/20002297.2022.2037832.
49. Kumar PS, Griffen AL, Barton JA, Paster BJ, Moeschberger ML, Leys EJ. 2003. New bacterial species associated with chronic periodontitis. J Dent Res 82:338–344. https://doi.org/10.1177/154405910308200503.
50. Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, Watanabe H, Masuda K, Nishimoto Y, Kubo M, Hosoda F, Rokutan H, Matsumoto M, Takamaru H, Yamada M, Matsuda T, Iwasaki M, Yamaji T, Yachida T, Soga T, Kurokawa K, Toyoda A, Ogura Y, Hayashi T, Hatakeyama M, Nakagama H, Saito Y, Fukuda S, Shibata T, Yamada T. 2019. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 25:968–976. https://doi.org/10.1038/s41591-019-0458-7.

51. Karunakaran G, Yang Y, Tremblay V, Ning Z, Martin J, Belaouad A, Figeys D, Brunzelle JS, Giguere PM, Stintzi A, Couture JF. 2022. Structural analysis of *Atopobium parvulum* SufS cysteine desulfurase linked to Crohn's disease. FEBS Lett 596:898–909. https://doi.org/10.1002/1873-3468.14295.
52. Mottawea W, Chiang CK, Muhlbauer M, Starr AE, Butcher J, Abujamel T, Deeke SA, Brandel A, Zhou H, Shokralla S, Hajibabaei M, Singleton R, Benchimol EI, Jobin C, Mack DR, Figeys D, Stintzi A. 2016. Altered intestinal microbiota–host mitochondria crosstalk in new onset Crohn's disease. Nat Commun 7:13419. https://doi.org/10.1038/ncomms13419.
53. Duan X, Chen X, Gupta M, Seriwatanachai D, Xue H, Xiong Q, Xu T, Li D, Mo A, Tang X, Zhou X, Li Y, Yuan Q. 2020. Salivary microbiome in patients undergoing hemodialysis and its associations with the duration of the dialysis. BMC Nephrol 21:414. https://doi.org/10.1186/s12882-020-02009-y.
54. Wegmann U, Louis P, Goesmann A, Henrissat B, Duncan SH, Flint HJ. 2014. Complete genome of a new Firmicutes species belonging to the dominant human colonic microbiota (*"Ruminococcus bicirculans"*) reveals two chromosomes and a selective capacity to utilize plant glucans. Environ Microbiol 16:2879–2890. https://doi.org/10.1111/1462-2920.12217.
55. Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, Zhu Q, Bolzan M, Cumbo F, May U, Sanders JG, Zolfo M, Kopylova E, Pasolli E, Knight R, Mirarab S, Huttenhower C, Segata N. 2020. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat Commun 11:2500. https://doi.org/10.1038/s41467-020-16366-7.
56. Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214. https://doi.org/10.1089/10665270050081478.
57. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol 16:294. https://doi.org/10.1186/s13059-015-0849-0.
58. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
59. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
60. Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319. https://doi.org/10.7717/peerj.1319.
61. Delmont TO, Eren AM. 2018. Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. PeerJ 6:e4320. https://doi.org/10.7717/peerj.4320.
62. Koster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28:2520–2522. https://doi.org/10.1093/bioinformatics/bts480.