Published in partnership with Nanyang Technological University

https://doi.org/10.1038/s41522-024-00634-1

Saliva microbiome profiling by full-gene 16S rRNA Oxford Nanopore Technology versus Illumina MiSeq sequencing

Molecular characterization of the oral microbiome is a crucial first step in experiments which aim to understand the complex dynamics of the oral microbiome or the interplay with host health and disease. Third-generation Oxford Nanopore Technology (ONT) offers advanced long-read sequencing capabilities, which hold promise for improved molecular characterization by distinguishing closely related microbial species within oral ecosystems in health and disease states. However, the performance of ONT sequencing of oral samples requires validation, and the consistency of this approach across different analytical and sampling conditions is not well understood. This study evaluates various factors that may influence the ONT sequencing outputs of saliva microbiota and compares results with those from Illumina MiSeq's v3v4 amplicon sequencing. Our analysis includes assessments of various stages in the workflow, including different collection and extraction methods, such as robot-extracted saliva DNA used in population-based biobanks, the effects of limited DNA quantities, different bioinformatics pipelines, and different 16S rRNA gene databases. The results demonstrate that ONT provides superior resolution in identifying oral species and subspecies compared to Illumina MiSeq, though the choice of bioinformatics strategy significantly affects the outcomes. Additionally, we confirm the suitability of biobank saliva DNA for large-scale cohort studies, which facilitates the mapping of oral bacterial phylotypes associated with disease states, including less prevalent conditions. Overall, our findings confirm a markedly improved resolution of oral microbiomes by ONT and offer an evidence base to guide the conduct of experiments using this method.

The human oral cavity hosts a variety of ecological niches, each colonized by taxonomically and functionally distinct bacterial species. This forms complex and diverse microbial communities, historically studied primarily within the contexts of oral health and disease[1]. Recent research has expanded our understanding, linking the oral microbiota to a range of systemic diseases, including metabolic and inflammatory disorders[1–3], and reported compositional correlations[4,5]. Given these associations, the oral microbiota is increasingly considered both a biomarker for systemic diseases and a potential mediator linking poor oral health to broader systemic outcomes[6].

Numerous methodologies exist for sampling oral microbiota, ranging from scraping of supra-gingival or sub-gingival biofilms to swabbing of the tongue or mucosal surfaces. Among these, saliva sampling offers a straightforward and non-invasive technique. Although saliva is sterile at the point of secretion, it rapidly accumulates bacterial species shed from mucosal and tooth surfaces as it enters the oral cavity, thus mirroring the microbial diversity of key oral niches[7]. Saliva microbiota profiles vary between individuals[8] but show considerable temporal stability[8,9]. Moreover, these profiles are reported resilient to methodological variations in sampling and DNA extraction techniques[10], making saliva an optimal medium for general oral microbiota characterization, except when research questions necessitate site-specific sampling. Salivary samples predominantly contain species from the *Prevotella, Neisseria, Haemophilus, and Streptococcus* genera[11], indicating a rich presence of both gram-positive and gram-negative bacteria.

Traditionally, research on saliva microbiota has involved small to medium-sized cohorts, focusing predominantly on its local role within the oral cavity and its link to common dental conditions like caries and periodontal disease. Recently, however, the scope of the investigation has expanded to explore the saliva microbiome's potential connections to less common non-oral diseases[12,13]. This shift may be supported by the development of large-scale biobanks, which collect comprehensive medical, lifestyle, and genetic data, and biosamples/biofluids, to enable both common and rare disease biomarker studies. For human genome-wide association studies (GWAS), extracted DNA from blood or saliva is commonly used. In Sweden, significant examples include the Swedish Infrastructure for Medical Population-based Life-course and Environmental Research (SIMPLER, https://www.simpler4health.se/about-us/) and the Swedish Twin Register[14] (STR), both of which utilize commercially available saliva collection kits and DNA extraction protocols optimized for lysing human cells. Details are described in the methods section.

Improved methods for the characterization of microbiomes were crucial to unlocking microbiome research in larger studies. In this context, second-generation sequencing platforms, notably the Illumina MiSeq and HiSeq platforms (San Diego, CA, USA), have been widely employed to profile the 16S rRNA genes of microorganisms within human microbiomes. These platforms boast a high-throughput capability and produce around 250 base pair (bp) reads, which are assembled into complete or partial sequences. Most studies focusing on taxonomic profiling of oral microbiota have opted for partial sequencing of the 16S rRNA gene's variable regions, mainly due to cost considerations. Typically, forward and reverse sequence fragments are overlapped and matched against 16S rRNA gene sequences in public databases. However, these shorter fragments with a maximum amplicon size of ~500 bp lack the resolution needed for accurate species-level identification in phylogenetically close bacterial groups[2]. Despite this limitation, Illumina amplicon sequencing has significantly enhanced our understanding of the human microbiome, and its widespread application has led to its recognition as a reference in the field.

Nanopore Sequencing, developed by Oxford Nanopore Technologies (ONT), represents a newer approach that relies on electric signals generated as nucleotide strands pass through membrane pores[15,16]. This method contrasts with the light induction techniques used in second-generation sequencing platforms. ONT sequencing can process short to very long DNA sequences, and its capital cost is lower, making it more accessible for smaller labs. While the ONT offers several benefits, it has been labeled as having a higher sequencing error rate compared to some conventional sequencing methods. Over time, the false-positive and false-negative rates have been reduced by improvements in both chemistry used (the R10.4 system)[17,18] and in bioinformatic pipelines explicitly allocated to handle high error rate sequencing methods[19,20]. Still, there is an ongoing need to verify the refined data analysis algorithms to ensure the accuracy and reliability of the results obtained through this method.

The primary aim of this study was to assess the performance and consistency of saliva microbiomes using the ONT sequencing platform under various conditions. These included (i) evaluate various sequence denoising pipelines and existing or novel combinations for taxonomic naming; (ii) compare taxonomic characterizations from full-length 16S rRNA gene sequencing by ONT to Illumina MiSeq sequencing of the v3v4 variable regions using bacterial isolates, mock communities, and DNA extracted from 407 human whole saliva samples; (iii) assess the effect of different DNA extraction protocols on microbiome profiles; and (iv) assess effect of diminishing quantities of saliva DNA. While a similar comparison between ONT and MiSeq sequencing has previously been conducted for the gut microbiome[21], this study is the first to contrast these sequencing methods for the saliva-microbiome and to evaluate the usability of biobank robot-extracted DNA. The findings may pave the way for improved resolution of oral bacterial phylotypes in studies targeting oral microbiomes.

Results

Bioinformatic pipeline optimization and database selection. Various bioinformatics pipelines are available for processing ONT sequences, including Kraken2[22], Minimap2[23,24], and Emu[19]. To evaluate the performance of these together with assigned taxa databases, we sequenced the ZymoBIOMICS Microbial Community (D6305) comprising seven bacterial species (Supplementary Table 1; and two fungal species not discussed here) and compared the outcomes from Kraken2, Minimap2, and Emu pipelines. When linked to the default NCBI database, Kraken2, and Minimap2 matched approximately 50% of the obtained reads accurately. In contrast, the Emu pipeline linked to the default RDP v11.5 database matched 97.0% of the reads correctly (Fig. 1a). Using the extended Human Oral Microbiome Database (eHOMD, version 3.1) database targeting species in the oral cavity and upper airways[25] matching accuracy for Kraken2 and Minimap2 increased to 91.1% and 97.1%, respectively, and Emu to 97.2% (Fig. 1a). Repeated sequencings of the ZymoBIOMICS Microbial Community with Emu-eHOMD path showed excellent stability with interclass and intraclass correlation coefficient >0.969.

As a more accurate reflection of the profiles of the oral microbiome, analyses were repeated for a mixture of 33 oral species (Supplementary Table 1). Here, Kraken2 with NCBI identified 32 species and achieved a read accuracy of 87.3%. Minimap2 with NCBI identified all 33 species with a read accuracy of 87.1%, whereas the Emu pipeline with RDP v11.5 successfully identified all 33 species with a read accuracy of 95.5% (Fig. 1b). Employing the eHOMD database enhanced the accuracy for all three pipelines, with Kraken2 identifying all 33 strains with a read accuracy of

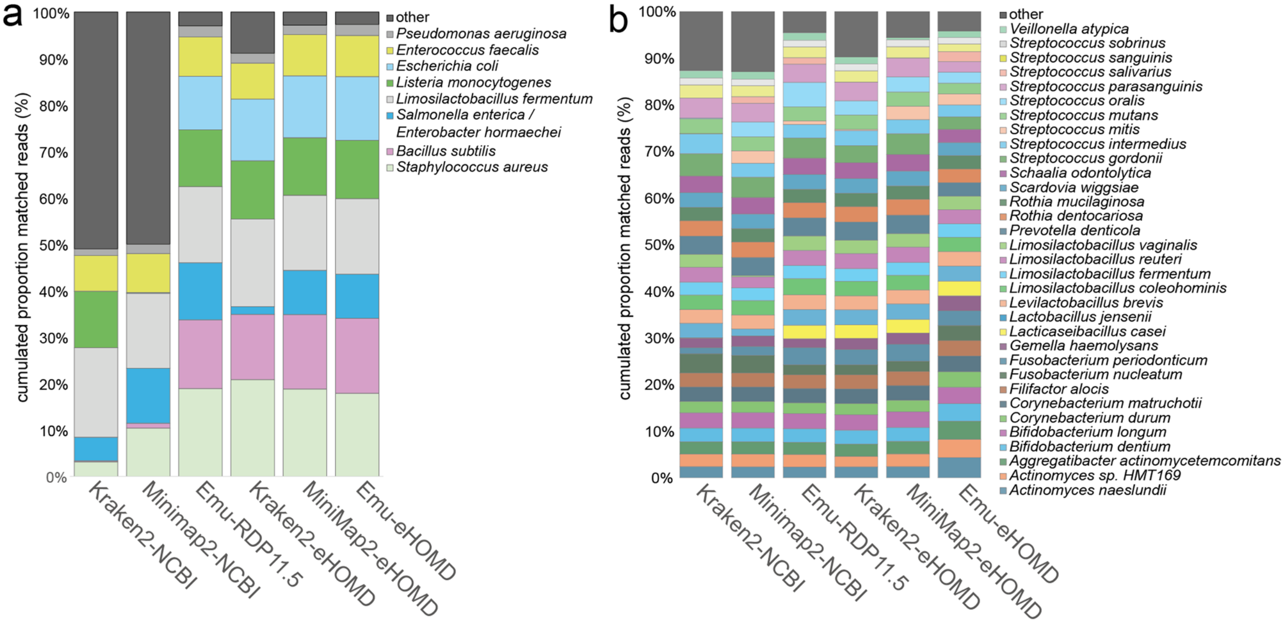

Fig. 1 | Evaluation of three bioinformatic pipelines and 16S rRNA gene databases for two mock communities. a The ZymoBIOMICS Microbial Community and **b** an oral-specific 33 mock community evaluated by the Kraken2, Minimap2, and Emu pipelines with their respective default 16S rRNA gene databases (NCBI and RDP v11.5) and the extended Human Oral Microbiome Database (eHOMD) database. The proportion of correctly matched reads for indicated pipeline, database, and species are shown in stacked bar plots. “Other” indicates the proportion of mismatched reads.

90.3%, Minimap2 improved to 94.5%, and Emu to 95.8% (Fig. 1b). Based on these findings the Emu pipeline was used for denoising and binning, and eHOMD was used for taxonomy classification for all further analyses.

Superior species-level resolution of ONT full-length 16S rRNA gene sequencing. Recent experiments characterizing the oral microbiomes in clinical settings have predominantly employed Illumina partial 16S rRNA gene amplicon sequencing. Though commonly recognized as a reference standard for alternative microbiota sequencing methodologies[21], this technique is limited by confinement to genus-level identification and limited ability to resolve phylogenetically close species. The potential of full-length 16S rRNA gene ONT sequences to separate 40 bacterial strains from the oral cavity was evaluated. Using the Emu-eHOMD setting, the ONT sequences were characterized, and the results contrasted with Illumina v3v4 sequences denoised and classified using DADA2 in QIIME2 and eHOMD. One strain was mismatched across both methodologies, potentially due to mislabeling or a novel taxonomic assignment not reflected in the existing database version, and was excluded from further comparisons, leaving 39 strains in the evaluation (Supplementary Table 1). Both sequencing platforms assigned the 39 strains accurately to their genus, but discrepancies emerged at the species level. While the ONT-Emu-eHOMD process achieved 92% accuracy, the Illumina sequences reached 74% (Supplementary Table 2). Notably, seven (out of ten) strains misclassified by Illumina involved closely related taxa in the *Streptococcus, Schaalia*, and *Veillonella* genera, i.e., three *Streptococcus oralis* and one *Streptococcus mitis* were classified as *Streptococcus infantis* clade 638, *Schaalia odontolytica* as *Actinomyces sp.* HMT180, and *Veillonella atypica* as *Veillonella parvula*. The capability of ONT to distinguish between closely related species and potential further resolution into subspecies within the *Streptococcus* and *Limosilactobacillus* genera is illustrated in Supplementary Fig. 1.

Comparison of ONT and Illumina sequencing of 407 saliva samples. The series of experiments on mocks or single species supported improved species identification and resolution for ONT with the Emu-eHOMD path compared to Illumina v3v4 amplicon sequencing. These also revealed a potential problem in recognizing taxa in the *Porphyromonas* genus. To explore this further and evaluate the profile and accuracy in a more complex context, we evaluated and compared microbiota profiling of ONT full-length 16S rRNA gene versus Illumina v3v4 sequencing in 407 saliva samples from healthy adults who had not used antibiotics for at least three months before sample collection. Population characteristics, including the 16–23 years olds with caries information, are presented in Supplementary Table 3). The mean read length from the ONT platform was 1540 bp after filtering out sequences shorter than 1000 bp compared to a read length of 388 bp (v3v4) from the MiSeq platform. After quality filtering, ONT generated between 17,526 and 220,488 (median 79,255) reads across the 407 samples, and Illumina v3v4 between 19,699 and 100,164 (median 48,897) reads.

ONT sequencing reveals higher saliva microbiome richness. Rarefaction analysis demonstrated that the species richness captured by ONT was greater than by Illumina v3v4 (Fig. 2a). This was supported by the Chao1 index (Fig. 2b), as well as indices that incorporate species evenness, such as the Shannon and Pielou indices (Fig. 2c, d). Composition profiles revealed that while most genera showed similar relative abundances between the two sequencing methods, there were notably higher proportions of *Streptococcus* and lower of *Haemophilus* detected with ONT (Fig. 2e). PCoA plotting of species Bray-Curtis distance matrix, where both presence and abundance are considered, distinctly separated the two sequence swarms (Fig. 2f).

Enhanced taxonomic resolution with ONT sequencing. Sequences from ONT encompassed 15 phyla and 131 genera, compared to the Illumina v3v4 sequences, which covered 80% (12 phyla) and 89% (116 genera) of these, respectively (Fig. 3a). Furthermore, ONT sequencing identified a total of 499 species/phylotypes, with 82% (409 species/phylotypes) also detected by the Illumina platform (Fig. 3a). ONT identified 30 species and Illumina v3v4 3 species in all 407 samples, i.e., the core microbiome in 100% participants. The mean abundances of phyla, genera, and species across the two sequencing methods were strongly correlated, with Spearman correlation coefficients ranging from 0.88 to 0.97, with notable exceptions for Streptococcus oralis and Haemophilus parainfluenzae, which showed lower correlations (Fig. 3b). Mean relative abundances for individual phyla, genera, and species assessed by the two platforms correlated strongly (≥0.6) for 58%, 71%, and 56%, respectively (Fig. 3c).

Further insights into differentially distributed features are illustrated for species in the core microbiome (Fig. 4a), those commonly reported as associated with caries (Fig. 4b) or periodontitis (Fig. 4c). Details on detection prevalence, relative abundance, and correlation metrics are available in Supplementary Tables 4–6.

Full-gene 16S rRNA sequencing via ONT revealed the predominance of five phyla, i.e., Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria, and Fusobacteria in all samples, with relative abundances ranging from 53% to 2%. Notably, Saccharibacteria (TM7), while present in 99.5% of samples, exhibited a low abundance of 0.6% (Supplementary Table 4). Comparable results were obtained with Illumina v3v4 sequencing. Correlation analysis showed strong Spearman correlation coefficients (≥0.6) for half of the 12 phyla detected by both methods, with an additional 25% demonstrating moderate coefficients (≥0.4).

At the genus level, ONT identified several taxa with 100% prevalence, including *Streptococcus, Prevotella, Veillonella, Rothia, Neisseria, Schaalia, Haemophilus, Granulicatella, Actinomyces*, and *Oribacterium*, with relative abundances ranging from 32.3% to 0.9% (Supplementary Table 5). Illumina v3v4 sequencing corroborated these results, ranking eight of these genera as predominant. Furthermore, 19 genera were identified in ≥99.0% of the samples by ONT, typically with relative abundances below 1%. Correlation coefficients for taxa detected by both methods indicated strong correlations (≥0.6) for 61.6% and moderate correlations (≥0.4) for 18.8% of taxa (Supplementary Table 5).

Species-level analysis showed that among the 65 species identified in at least 95% of the 407 samples by ONT, 23 belonged to *Streptococcus*, and multiple species were attributed to genera such as *Prevotella, Veillonella*, and *Schaalia*. In contrast, Illumina sequencing detected 24 species in 95% of samples, aligning with the species distribution observed via ONT. Spearman correlation coefficients between the relative abundances derived from the two methods demonstrated strong correlations (≥0.6) for 43.3% of the 409 species detected by both platforms and moderate correlations (≥0.4) for another 27.4% (Supplementary Table 6). The phylogenetic tree for ONT 16S rRNA full gene sequences versus those obtained from Illumina MiSeq sequencing of the v3v4 segment of the same gene is illustrated in Fig. 5.

Identifying caries-associated bacteria using ONT and Illumina sequencing. Dental caries remains a critical global health challenge,

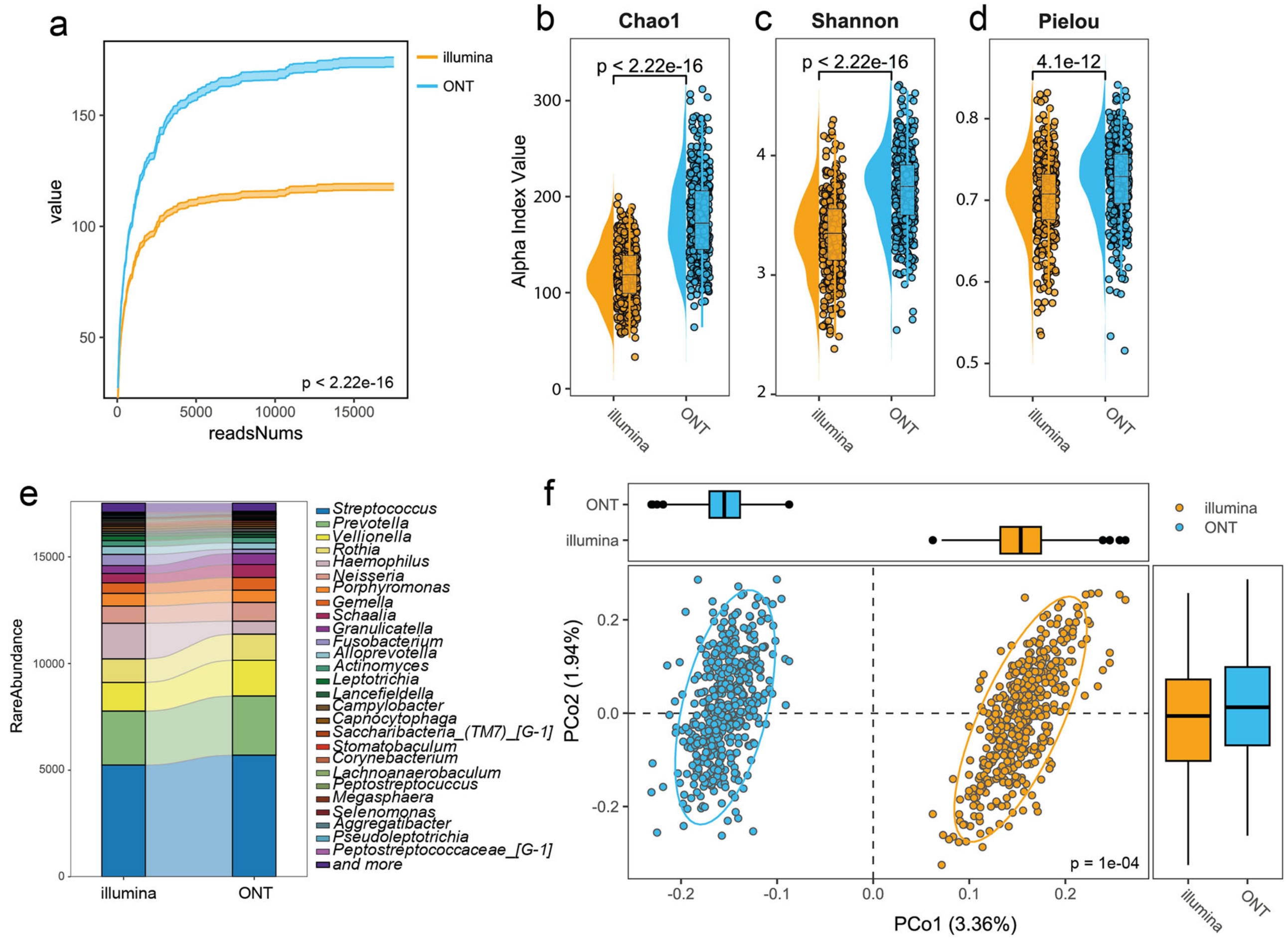

Fig. 2 | Diversity features from ONT and Illumina sequencing of 407 saliva samples. **a** Rarefactions curves (95% CI) of the number of observed species. Box and histogram plots for **b** Chao1, **c** Shannon index, and **d** Pielou indexes; **e** Rarified abundance of top 20 genera; **f** PCoA plot and explained variance for components 1 and 2 based on the Bray Curtis index. Results are based on rarefied sequences at a sequence depth of 17,520 reads.

characterized by the bacterial fermentation of carbohydrates leading to tooth tissue loss. While previous findings using Illumina MiSeq amplicon sequencing have been inconsistent, we searched associations with caries severity assessed as the number of caries-affected surfaces (DeFS, n = 209) and the microbial profiles of saliva using both ONT full 16S rRNA gene sequencing and Illumina MiSeq v3v4 sequencing. Our initial approach involved supervised, data-driven partial least squares (PLS) modeling, which did not yield a statistically significant model when analyzing species abundance data from either sequencing method. We shifted to linear regression analyses, adjusting for sex and age, which displayed nine ONT-identified species significantly associated with high DeFS after adjusting for multiple testing by the false discovery rate (FDR < 0.05) from the Benjamini-Hochberg method. These were *Streptococcus mutans* (HMT686), *Actinomyces gerencseriae* (HMT618), *Selenomonas sp.* HMT133, *Streptococcus vestibularis* (HMT021), *Streptococcus lactarius* (HMT948), *Streptococcus peroris* (HMT728), and *Leptotrichia sp.* HMT879, whereas *Stomatobaculum sp.* HMT097, and *Veillonella rogosae* (HMT158) were negatively associated with DeFS. Among the Illumina sequences only *S. mutans*, *A. gerencseriae*, and *Stomatobaculum* sp. HMT097 remained significantly associated with DeFS after FDR correction. Additionally, incorporating body mass index as a covariate did not alter these findings.

DNA extraction method comparison: consistent core microbiome despite methodological differences. Following the differences in age distributions, sample collection, sample storage time, and other uncontrolled cohort differences, we restricted the evaluation of the DNA extraction method to measures taking both taxa richness (number of species) and evenness (relative abundance) into account and rarefication at a sequences depth of 11,745 reads. Here, alpha-diversity from the Shannon indices tended to be lower among the Pure-Chem extracted samples from 13 to 15-year-old twins compared with Sigma-extracted samples from reference 16 or 17-year-olds (Fig. 6a), whereas the Qiagen-extracted DNA displayed higher alpha-diversity than Sigma extracted references in adults (Fig. 6a). Further, Sigma DNA extraction yielded higher proportions in the Firmicutes-phyla and *Streptococcus*-genus (gram positive) with an inverse shift in the Bacteriodales phylum and *Prevotella* genus (gram negative) (Fig. 6b, c). Despite the differences highlighted by the diversity analyses, a large core microbiome was shared across all samples. In total, 134 genera were detected among all samples, and of these, 84.3% were in common. Further, all top 20 genera and 98% of the top 50 genera, ranked similarly across the

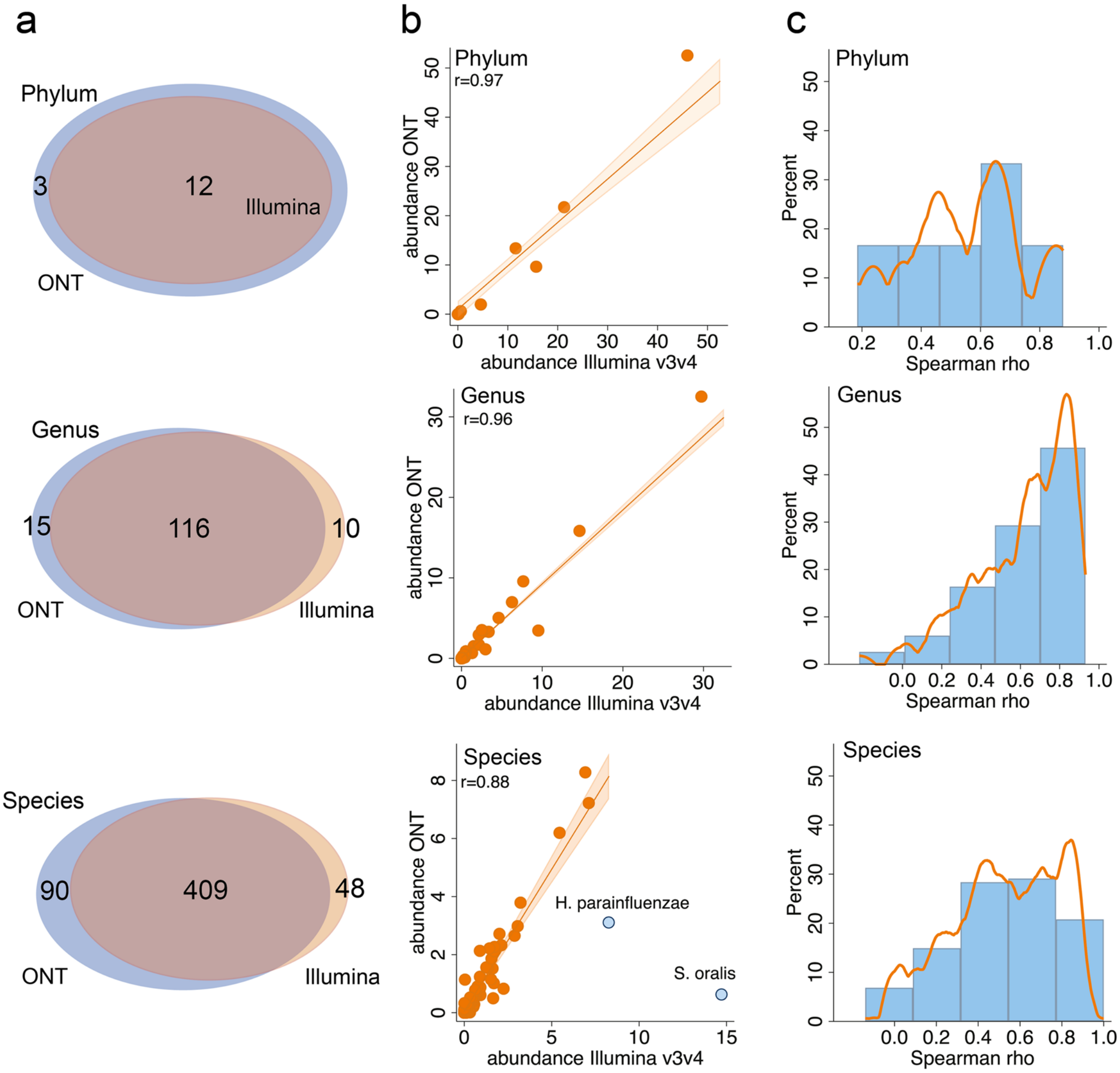

Fig. 3 | Overview of ONT versus Illumina sequencing results for 407 saliva-extracted DNA samples at the phylum, genus, and species levels. a Venn diagrams illustrating numbers of shared taxa; **b** Scatter plots with regression line and 95% confidence interval of mean abundances; **c** Histograms with Kernel density lines for Spearman correlation coefficients of individual taxa.

four groups based on their relative abundance. For the 21 genera that differed between the DNA extraction groups (LOD10 > 3, FDR adjusted $p < 0.05$, Fig. 6d), 14 were higher in the Qiagen group, 4 higher in the Sigma groups, and 3 in the Pure-Chem group. A large overlap could also be observed on the species level; all top 100 species were in common for the DNA extraction groups based on relative abundance, 98.5% of the top 200 species and 89.3% of the top 300 species. Even though the PCoA pot of the Bray-Curtis dissimilarity illustrates significant shifts in the composition between the DNA extraction groups ($p < 0.001$), it also indicates large overlaps between the extraction groups (Fig. 6e).

Robustness of ONT sequencing at low DNA concentrations. Samples from the oral cavity can be collected via swabbing or scraping mucosal surfaces or supra- or subgingival tooth surfaces, as well as through the collection of unstimulated or stimulated saliva. These methods yield a composite of host and microbial DNA, complicating the control of targeted starting material quantities. This is particularly challenging in microbiome studies where bacterial DNA constitutes only a minor fraction of the total DNA extracted from the samples. To investigate the effects of varying DNA input quantities on microbiota composition, we evaluated the yield from ONT sequencing of three samples of stimulated saliva systematically diluted to represent a gradient of total DNA amounts (host/microbial mixture) from 800 ng to 6.5 ng in the PCR amplification step. The analysis revealed no significant differences across the range of DNA inputs for the number of species identified (Fig. 7a), Shannon diversity index (Fig. 7b), or Bray-Curtis dissimilarity within samples (p = 0.999) (Fig. 7c). Additionally, the intra-sample species correlation was high, ranging from 0.961 to 0.942 ($p < 0.001$). These findings highlight the efficacy of ONT sequencing coupled with Emu-eHOMD denoising and classification in accurately profiling the saliva microbiota from samples with

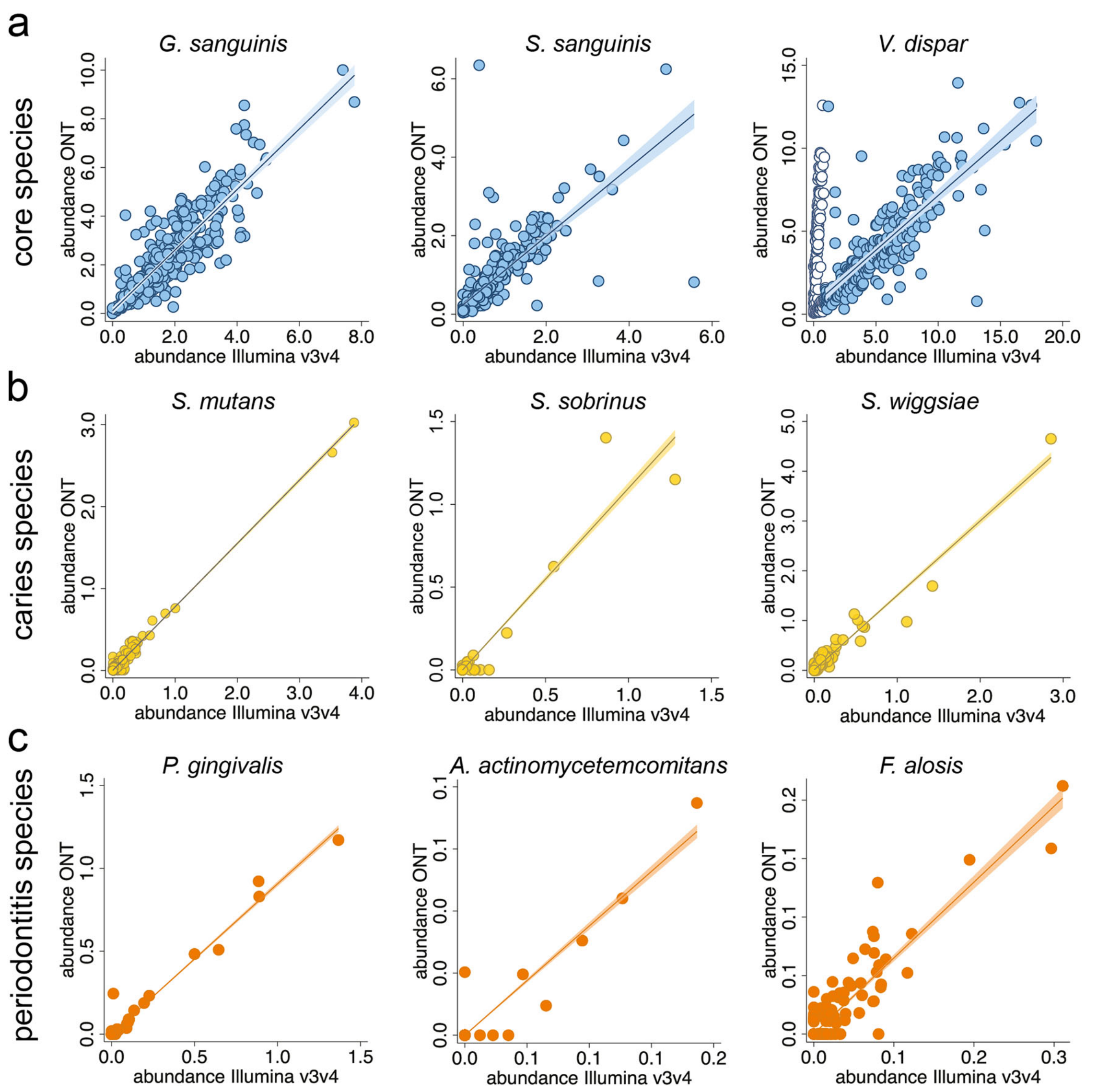

Fig. 4 | Panel of scatterplots for relative abundances determined from ONT and Illumina sequencing for species associated with the oral core microbiome, caries, and periodontitis risk. a Core microbiome; **b** Caries-associated species; **c** Periodontitis-associated species. A trendline with 95%CI is shown for each species.

low DNA concentrations down to 0.5 ng/µL (using 12 µL in the PCR reaction) without compromising sequencing integrity.

Discussion

Molecular characterization of the oral microbiota is a key step in understanding how oral microbiota relates to host health and disease. To date, most studies have used partial 16S rRNA gene amplification for oral microbiota characterization, which can be carried out relatively cheaply but offers limited taxonomic resolution, a potential problem in the oral microbiome where there are numerous phylogenetically highly similar taxa. Recognizing these limitations, this study utilized ONT to sequence the entire 16S rRNA gene, a method previously validated in e.g., gut microbiome studies but not extensively applied to the oral microbiome. We demonstrate that ONT sequencing of saliva-extracted DNA provides batch stability and precise taxonomic resolution, i.e., accurate classification of 36 of 39 test strains. Additionally, we analyzed over 400 clinical samples, achieving detailed delineation to the species and subspecies levels. This method notably enhanced the detection of clinically relevant species implicated in dental caries when compared to the commonly used Illumina MiSeq 16S rRNA v3v4 amplicon sequencing. We show that saliva-extracted DNA stored in biobanks from large, population-based cohorts in Sweden is suitable for extensive studies that require at least species-level detection. This capability is pivotal for advancing large-scale epidemiological research, potentially uncovering novel microbiological insights into oral and systemic health.

Three bioinformatics pipelines, Kraken2, Minimap2, and Emu, have been recommended for the processing of ONT sequences using either the 16S rRNA databases from the National Center for Biotechnology Information (NCBI) or the Ribosomal Database Project (RDP) for taxonomic classification. These databases include sequences sourced broadly, unlike the specialized eHOMD database, which focuses exclusively on

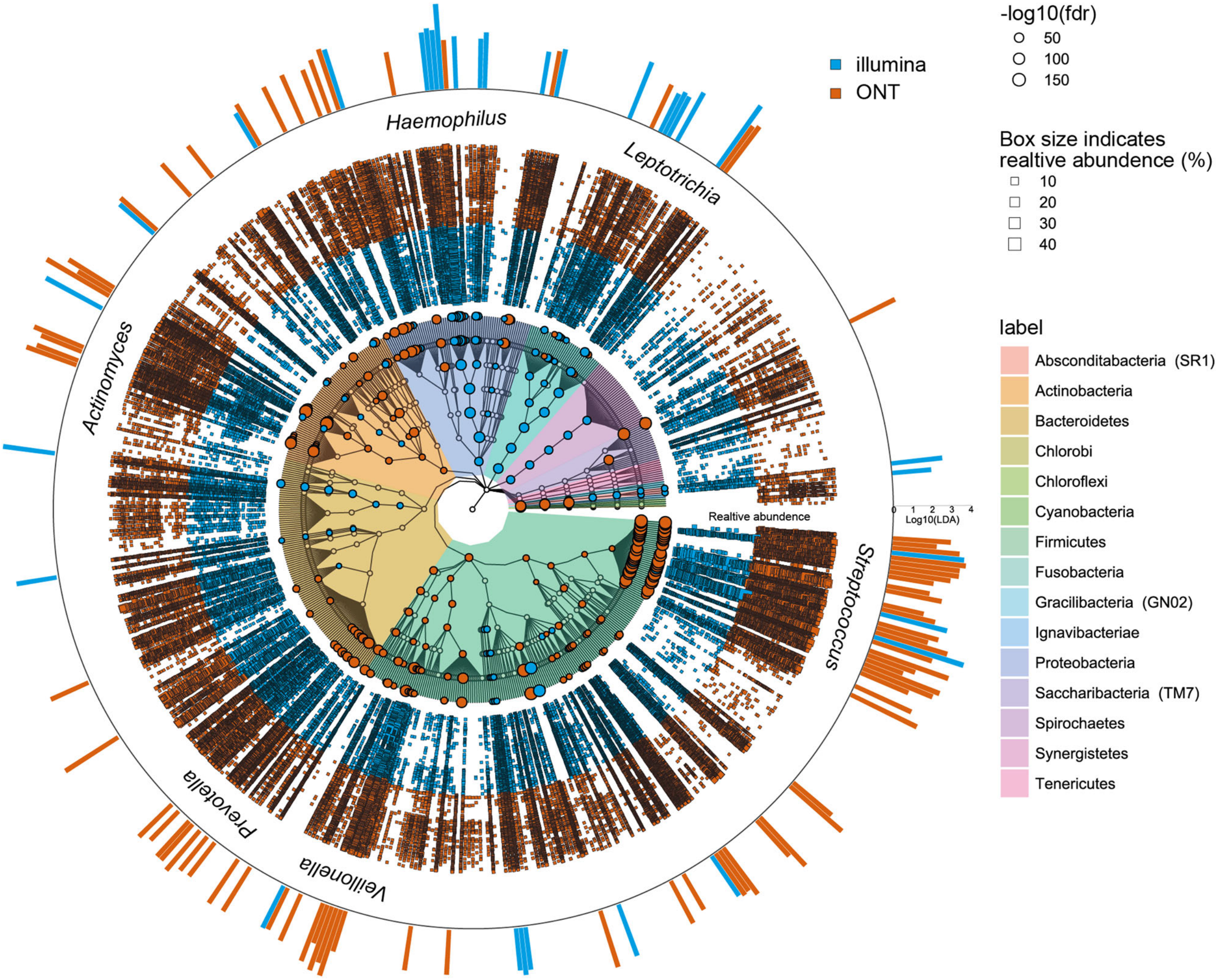

Fig. 5 | Phylogenetic tree comparing detection prevalences by ONT 16S rRNA sequencing versus Illumina MiSeq 16S rRNA v3v4 sequencing of 407 clinical saliva samples. The inner circle shows a phylogenetic tree with color-marked phyla, and filled circles indicating significant phyla. The middle circle with boxes indicates relative abundance for each sample and subgroup. The bar graph on the outermost circle represents LDA coefficients exceeding 2.0 and differing significantly between sequence platforms (Wilcox test). All tests are adjusted for multiple comparisons (FDR).

bacteria from oral and upper gastrointestinal/airway environments. Our experiments involved several setups of bacteria, i.e., a commercial mock bacterial community, a custom mix of oral bacteria, and individual taxa representing genetically similar oral phylotypes, to cover general and specific aspects of oral and other bacterial communities. Through these studies, we evaluated various combinations for denoising, binning, and taxonomic classification of saliva-derived ONA sequences. Our results demonstrated that the Emu pipeline, when paired with default 16S gene databases, significantly outperforms Kraken2 and Minimap2 in characterizing oral microbiota using full-length 16S rDNA sequences. Integration of the eHOMD database substantially enhanced taxonomic resolution relative to NCBI and RDP v11.5, achieving species-level identification and, in some cases, classification down to subspecies (HMT) level. Consequently, for oral microbiome samples, the combination of the Emu pipeline and the eHOMD database emerged as most effective and therefore is our recommendation for oral samples. Nevertheless, it is imperative to acknowledge the limitations in the diversity of species variants within the eHOMD database, which may impact the breadth of microbial characterization. A broader discussion regarding the selection of appropriate 16S rRNA databases for different microbiome studies is available in the recent literature[26,27]. Additionally, our findings were consistent even when the amount of extracted DNA was reduced to as little as 0.5 ng/μL. While DNA yields from whole saliva and pooled dental samples are typically sufficient, sampling challenges persist, particularly in contexts like infants or individual tooth surfaces, which potentially could affect the comprehensiveness of microbiome sequence coverage.

Recent advancements in third-generation multiplex DNA sequencing methods, such as ONT, have mitigated historical challenges related to high error rates and inadequate specificity through improvements in sequencing chemistry and sequence error filtering. These improvements are crucial in studies of complex characterizations of microbiomes like the oral microbiome, which is densely populated by genetically similar species across several genera, including *Actinomyces, Lacticaseibacillus*

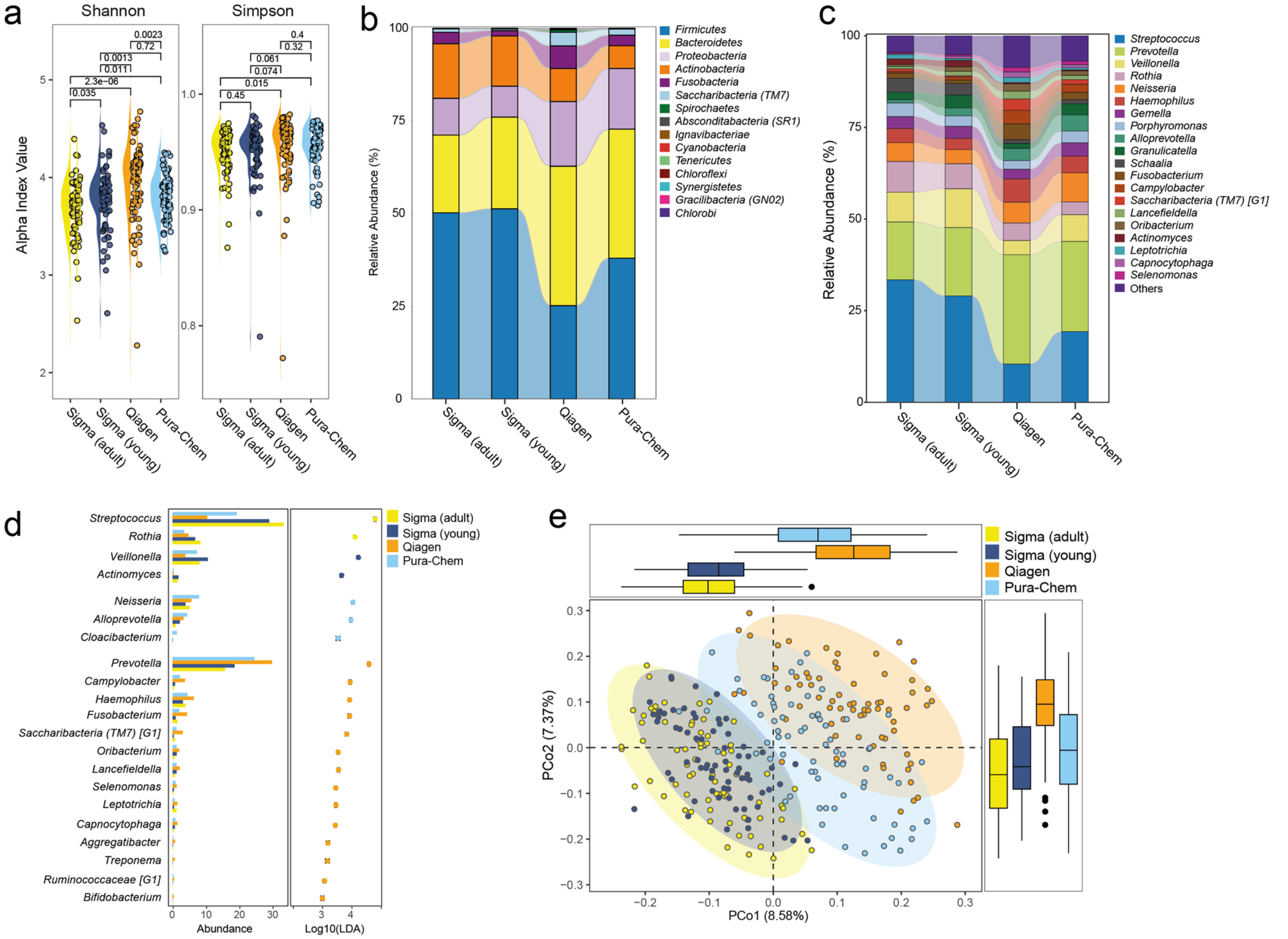

Fig. 6 | DNA extraction kit and microbiota composition. a Shannon and Simpson diversity indices, **b** Phyla and **c** Genus level comparison of Sigma (adult), Sigma (young), Qiagen, and Puregene/Chemagen (Pure-Gene) DNA extraction kits. **d** Genera differentiating between extraction kits significant after FDR adjusted p-values and LOD10 > 3. **e** Bray Curtis distance PCoA plot (explained variance for components 1 and 2). Results are based on rarefied sequences at a depth of 11 745 reads.

(formerly *Lactobacillus*), *Streptococcus, Veillonella, Prevotella*, and *Fusobacterium*. These genera play pivotal roles in various oral and systemic diseases, and accurate and detailed mapping is crucial. Although full genome or shotgun sequencing offers superior resolution for analyzing bacterial communities, their high costs often limit their application in extensive studies. Conversely, if full-length 16S rRNA gene sequencing can accurately delineate phylogenetic relationships at both species and subspecies levels, it could substantially enhance our understanding of oral bacteria's contributions to health and disease. For example, distinguishing between the *Aggregatibacter actinomycetemcomitans* JP2 genotype, and other genotypes is critical for understanding the pathogenesis of periodontal diseases[28]. Additionally to the improvements by the company, new algorithms for sequence denoising in the Emu pipeline have further decreased error rates[17,18]. Hence, the Emu pipeline, tailored for error-prone sequences, is a dual-stage algorithm that initially aligns reads accurately to a reference database, then applies an EM-based error correction method that iteratively refines species-level relative abundances using total read-mapping counts[19,20]. Additional notations worth reflecting on is that we used the ZymoBIOMICS Microbial Community Standard, where one theoretically would anticipate equal representation among the eight bacterial species. Although variations in cell size can affect volumetric counts at standardized optical densities, an equitable distribution of 10–15% per species should be feasible to expect. Notably, only the combination of the Emu pipeline with RDP v11.5 or eHOMD, and Minimap2 with eHOMD, achieved this balance.

Encouraged by the taxonomic accuracy of Emu-eHOMD processed 16S rRNA gene sequences by ONT we applied the method to chewing-stimulated saliva collected in a clinical setting and also compared the profiling with corresponding Illumina MiSeq v3v4 sequences. In our experimental assessments of mock communities and the 407 saliva samples using ONT full-length 16S rRNA gene sequencing, we noted a potential limitation in that ONT failed to detect any species within the *Pseudomonas* genus. This observation could potentially be explained by primer mismatch with the 16S rRNA gene used for ONT. In contrast, Illumina's v3v4 sequencing successfully identified *Pseudomonas fluorescens* in the same set of 407 saliva samples, further supporting the possibility of primer inefficacy. The design of primers for amplifying the v1-v9 regions of the 16S rRNA gene aims to achieve the best possible match across many taxa in the non-variable regions flanking the targeted gene segment. Nevertheless, these regions are not entirely

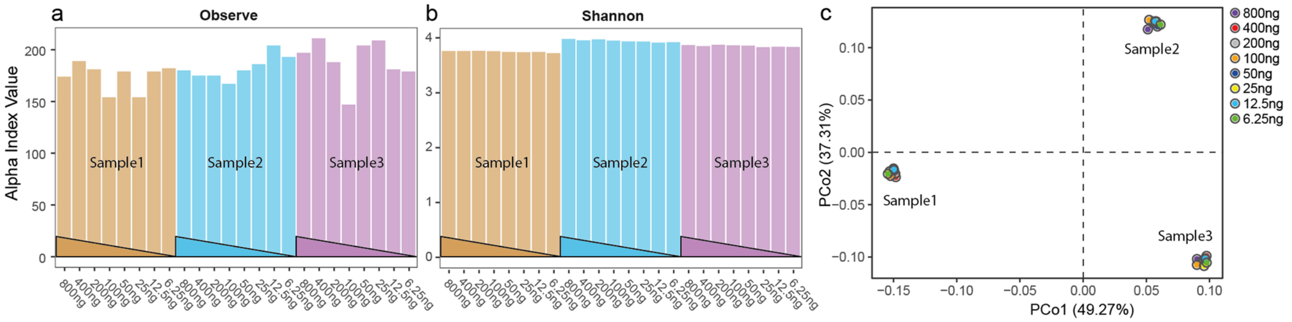

Fig. 7 | Effect of total saliva DNA used for the PCR amplification step prior to library preparation and ONT sequencing. a Barplots of the observed number of species; **b** Shannon index; and **c** Bray Curtis distance PCoA plot (explained variance for components 1 and 2) based on serially diluted DNA samples from three clinical samples ranging from 800 ng to 6.5 ng. Results are generated at a sequencing depth of 49,365 reads.

conserved, occasionally leading to mismatches and reduced efficiency. It is important to note that other researchers have also reported detection failures of *Pseudomonas* using ONT's recommended primers for v1-v9 amplification, indicating a broader issue that may affect reproducibility and accuracy[29]. Moreover, despite this specific detection limitation, the results from ONT's full-gene sequencing revealed a broader diversity and enhanced taxonomic resolution in the saliva microbiota compared to Illumina v3v4 reads. This finding aligns with previous studies on gut microbiota, illustrating the potential of ONT sequencing for detailed microbiome profiling[21]. The overall microbial profiles obtained by ONT, while more comprehensive, generally corresponded with those observed in studies using Illumina v3v4 amplicon sequencing[21].

Systematic reviews underscore the recurrent identification of certain cariogenic bacteria across studies, yet also highlight the significant variability in these associations due to differences in study populations, sample processing, and analytical methods. For example, one review noted that *S. mutans*, *Streptococcus sobrinus*, *Scardovia wiggsiae*, and *Prevotella denticola* were found to be more prevalent in adolescents with caries in at least two out of 20 studies[30]. Another review reported similar findings for *S. mutans*, and additionally identified *Veillonella dispar* as enriched in caries-affected adolescents in more than one study[31], however, both reviews highlight the lack of consensus in the available literature. The detailed resolution and facilitated deeper insights into the role of oral bacteria in host well-being inspired us to evaluate the associations between ONT-derived taxa and dental caries in the subgroup where such information was available. Though the specific findings should be seen as pilot findings to be evaluated in independent settings they suggest that advancement in taxa detection and resolution capabilities will drive the present understanding of microbial caries determinants, and the ecological balance at a refined taxonomic level. Notably, identification of species that potentially prevent tooth demineralization remains unclear, and future investigations may focus on clarifying the roles of species and subspecies that metabolize lactic acid and those with arginolytic capabilities, such as *Streptococcus oralis subspecies dentisani* (HMT058)[32] which may contribute to tooth demineralization prevention.

This study aimed to assess the quality of saliva DNA multiplex sequences obtained using Oxford Nanopore Technologies and to evaluate the potential of saliva DNA preserved in major Swedish biobanks for expansive studies on the mouth-gut interactome and the role of oral bacteria in disease risk. The comparison was made possible by the sample availability across different DNA extraction methods (extracted from drooling saliva) and the use of chewing-stimulated saliva as reference samples. Despite source differences and revealing a shift in bacterial population dynamics between gram-negative and gram-positive species, our results were promising. We successfully detected the majority of bacterial species across all DNA extraction methods. These findings suggest that both STR and SIMPLER extraction methods, which do not use specific cell wall lysing enzymes, are effective for characterizing oral microbiota within each biobank. This demonstrates their potential for future large-scale studies on the oral microbiome; however, significant variability associated with extraction methods necessitates careful experimental design and data analysis.

In conclusion, the present study, which builds upon previous research focused on the gut microbiome, advances our understanding of the oral microbiome using the ONT platform. It plays a pivotal, and pioneer, role in demonstrating enhanced resolution of oral bacterial phylotypes by ONT, and superior accuracy using the Emu pipeline for denoising and binning, coupled with the eHOMD database for taxonomic annotation. This was manifested as the successful differentiation of the very closely related *Streptococcus* species in the mitis group, such as *Streptococcus oralis*, and *Streptococcus mitis*, from *Streptococcus infantis* clade 638. This refinement in classification by ONT minimizes traditional misclassifications and offers new perspectives on the roles of specific oral bacterial species and subtypes in both dental and systemic health. A further significant achievement of the study is the demonstration of the usefulness of biobank saliva-extracted DNA. In concert, these advancements hold promise for improving diagnostic precision and enhancing our understanding of the microbiological and genetic underpinnings of health and disease.

Methods

Bacteria samples and DNA extraction. The bacterial DNA used in the present project was either from a commercial standard kit, saliva DNA stored in biobanks, or from in-house extraction from 33 oral species in a custom-made mock community, 39 single oral bacterial strains, 407 fresh chewing stimulated saliva samples, and ultra-pure water (negative control). In-house extracted DNA used the GenElute™ Bacterial Genomic DNA kit (Sigma-Aldrich, St. Louis, MO, USA) with lysozyme, mutanolysin, and Proteinase K and was treated with RNase and purified as described previously[33]. For the commercial mock community (ZymoBIOMICS Microbial Community DNA Standard, D6305, NordicBiosite, Stockholm, Sweden) DNA was provided pre-purified. The biobank DNA was from the SIMPLER cohort where DNA was extracted from saliva at Eurofins Genomics (Ebersberg, Germany) using a Qiagen spin column kit and 96 well setting with proteinase K (no mechanical disruption) with initial steps done manually. Saliva DNA in the STR biobank was extracted by the Puregene for blood kit (QIAgen, Hilden, Germany) or Chemagen (Revvity chemagen Technology, Baesweiler, Germany). STR used robot extraction and according to the manufacturer, these kits remove contaminants and enzyme inhibitors[34]. The quality of the extracted DNA was estimated using a NanoDrop 1000 Spectrophotometer and the quantity by the Qubit 4 Fluorometer (Thermo Fisher Scientific, Uppsala, Sweden) before use.

ONT sequencing and processing. For the Nanopore ONT sequencing, the v1 through v9 variable regions of the 16S rRNA gene were amplified using 50 ng DNA, KAPA 2x HiFi ready mix KAPA HiFi HotStart ReadyMix (2X) (New England Biolabs, Ipswich, MA, USA), and the primers 27 F 5′-AGAGTTTGATCMTGGCTCAG-3′ and 1,492 R 5′-CGGTTACCTTGTTACGACTT-3′. Amplification was performed on a MiniAmp™ Thermal Cycler (Thermo Fisher Scientific, Uppsala, Sweden) using the program: 1 min denaturation at 98 °C, 35 cycles of 95°C 20 s, 55 °C 15 s, 72 °C 1.5 min, and a final extension step of 1.5 min at 72 °C in 25 μL reactions. Creation of a single fragment of the expected 1465 bp was confirmed by separation on a 0.8% agarose gel in 0.5x Tris/Borate/EDTA buffer with SYBR™ Green (Fisher scientific, Göteborg, Sweden). Sample amplicons were purified using the 0.8% AMPure XP Beads (Beckman Coulter, Brea, CA), washed twice in 80% ethanol and eluted in EB buffer (Fisher scientific, Göteborg, Sweden), and quantified using the Qubit dsDNA HS Assay Kit and Qubit 4.0 Fluorometer (Thermo Fisher Scientific, Oregon, USA).

Library preparations were performed by barcoding amplicons using the Native Barcoding Kit 96 V14 (SQK-NBD114.96) kit (Oxford Nanopore Technologies Nanopore, Oxford, UK). For this, 200 fmol PCR products were end-repaired using NEB Next® Ultra™ II End Repair/dA-Tailing Module (NEB, E7546L) of which 1/20 was used for ligation to unique barcodes for each sample using NEB Blunt/TA Ligase Master Mix (NEB, M0367). Barcoded samples were pooled and cleaned using 0.4x AMPure XP Beads (Beckman Coulter, Brea, CA). Purified and pooled samples were finally fused to the Native Adapter using T4 DNA Ligase (NEB, E6056) and finally cleaned with 0.4x AMPure XP Beads (Beckman Coulter, Brea, CA). Libraries were quantified using a Qubit 4 fluorometer (Thermo Fisher Scientific, Oregon, USA).

Sequencing of Native Adapter barcoded pools was performed by loading 100 ng into a pre-primed R10.4.1 flow cell (Oxford Nanopore Technologies) and sequenced using a GridION nanopore sequencer (Oxford Nanopore Technologies) for 72 h. Base-calling of nanopore signals and demultiplexing was performed on the GridION using the MinKNOW (Nanopore, Oxford, UK), Dorado base callers Super accurate (SUP) model and Porechop (version 0.2.4, https://github.com/rrwick/porechop) generating demultiplexed FastQ files, with a quality score (QC) score ≥10 with a read length between 1350 and 1800 bp.

Illumina sequencing and sequence processing. For Illumina sequencing, the v3-v4 variable regions of the 16S rRNA gene were amplified and sequenced at the Miseq platform using the KAPA 2x HiFi ready mix KAPA HiFi HotStart ReadyMix (2X) (New England Biolabs, Ipswich, MA) with 341 F (CCTACGGGNGGCWGCAG) forward, and 806 R (GGACTACHVGGGTWTCTAAT) reverse primers[33]. Briefly, amplicons of the target region were generated by PCR using fusion primers with forward and reverse primers and sample barcodes as described by Caporaso[35]. Equimolar amplicon libraries were pooled and purified using AMPure XP beads (Beckman Coulter). Pools adjusted to 4 nM, with 5% PhiX (Illumina, Eindhoven, the Netherlands), were denatured and diluted according to Illumina instructions and were loaded and sequenced using MiSeq cartridges (Illumina, San Diego, CA). Obtained sequences were de-multiplexed, pair-end reads fused, primers, ambiguous, chimeric, and PhiX sequences removed, and amplicon sequence variants (ASVs) retrieved using the open-source software package DADA2[36] in the QIIME2 microbiome bioinformatics platform (https://qiime2.org accessed November 2020)[37]. ASVs were taxonomically classified against the expanded Human Oral Microbiome Database (version HOMD_16S_rRNA_RefSeq_V15.23) (http://www.homd.org accessed December 2023)[38,39]. ASVs with 98.5% identity with a named species or unnamed phylotype in HOMD and with at least two reads were retained, and those with the same Human Microbial Taxon (HMT) number were aggregated. An HMT present in at least two samples was retained for downstream analyses. Illumina sequencing was either done at the Swedish Defense Research Agency (Umeå, Sweden) or at Eurofins (Ebersberg, Germany).

Bioinformatic pipelines and 16S databases. Three publicly available pipelines for ONT sequence filtering and denoising were used. The Kraken2 and Minimap2 pipelines are two pipelines available within the ONT processing Epi2me Lab platform with the NextFlow workflow wf-16s (https://github.com/epi2me-labs/wf-16s). The Emu pipeline[19,20] was executed using the open-source software package (https://github.com/treangenlab/emu). Kraken2 (k-mer-based) and Minimap2 (alignment-based) pipelines were run using the setting read length 1400–1800, and Q-score >10; in addition, for Kraken2, the bracken_length was set to 1000, and for Minimap2 the [min_percent_identity] was set to 98.5% and [min_ref_coverage] to 90%. For the Emu pipeline, the relative abundance filter was set to >0.0001 [--min-abundance 0.0001].

Within the Epi2me-Kraken2 and -Minimap2 pipelines, the 16S-18S rRNA gene NCBI database is available, and for the Emu pipeline, the filtered and quality control reads RDP v11.5 database is available (https://github.com/treangenlab/emu?tab=readme-ov-file)[40]. In addition, the curated extended Human Oral Microbiome Database[38,39] was formatted for use within the Nextflow and Emu pipelines. The 16S rRNA gene NCBI database (16S rRNA gene RefSeq Nucleotide sequence records) matches against 27,155 RefSeq sequences. The RDP (Ribosomal Database Project)) provides quality-controlled, aligned, and annotated rRNA sequences and a suite of analysis tools containing 4779 RNA sequences. The Extended Human Oral Microbiome Database RefSeq (eHOMD 16S rRNA gene RefSeq Version 15.23) provides comprehensive curated information on bacteria in the human mouth and aerodigestive tract, including the pharynx, nasal passages, sinuses, and esophagus, and contains 1015 16S rRNA gene sequences covering 774 oral bacterial species (table2023-11-03_1698999909).

Saliva microbiota characterization. Whole saliva was collected from 427 16 to 79-year-old donors. The donors were healthy, had not taken antibiotics in the past two months, and were instructed not to brush or floss their teeth on the morning of saliva collection. They were also advised to abstain from eating or drinking anything except still water for at least 3 h before the sample was collected. All donors completed a questionnaire regarding their lifestyle habits, and for participants between 16 and 21 years of age (mean age 17.8 years, 95% CI 17.6–18.0) ($n = 209$), data on caries status were also available. Caries status was documented as previously described[41] and quantified by the number of decayed (in both dentin and enamel) and filled tooth surfaces (DeFS score). Missing teeth were due to orthodontic treatment or trauma and were not included in the caries score. Saliva flow was stimulated by chewing on a 1 g piece of paraffin and collected into ice-chilled sterile test tubes. The samples were kept on ice until they were transferred to a −80 °C freezer within six hours. Of the 427 donors, 15 samples were excluded due to insufficient saliva or DNA quantity. From the remaining samples, the ONT platform identified fewer than 10,000 reads in 4 (0.9%) samples, and the Illumina MiSeq platform (v3v4) sequencing identified fewer than 10,000 reads in 1 (0.2%) sample, leading to their exclusion. Consequently, 407 samples were retained for statistical analysis.

Additionally, 72 saliva-extracted DNA samples were retrieved from each of the SIMPLER (https://www.simpler4health.se/about-us/) and the Swedish Twin Cohort[14] biobanks. The 72 STR samples were equally distributed on their two extraction kits and were evaluated separately (STR-C or STR-P) and merged (STR).

Extraction of Biobank saliva DNA. The oral microbiome is a complex community, composed of gram-negative and gram-positive bacteria, with a significant representation of the latter[42]. The robust cell walls of gram-positive bacteria present considerable challenges in achieving optimal DNA extraction, impacting the profile of microbiological data obtained from, e.g., oral samples. In Sweden, several large cohorts with repeated samplings have collected saliva primarily for human DNA analyses, such as genome-wide screenings and twin zygosity determination. However, standard DNA extraction kits optimized for lysing human cells may not effectively lyse bacterial cells and limit comprehensive microbial characterizations[34]. To explore the potential of the Swedish cohort resources for large-scale and prospective studies on the associations between oral microbiota and medical conditions, we assessed the microbial profiles from saliva-extracted DNA by ONT in two distinct cohorts: SIMPLER and STR. The SIMPLER provided 72 samples of unstimulated saliva from individuals aged 56–79 years (mean 66.5), with DNA extractions performed at Eurofins Genomics (Ebersberg, Germany) using a Qiagen spin column kit supplemented with proteinase K and partial robotic assistance. The STR provided 72 samples of unstimulated saliva from 13 to 15-year-old twins (mean 14.1 years). DNA extraction for these samples was conducted with two different methodologies: 36 samples using the Puregene kit and another 36 using the Chemagen kit, both without enzyme addition and with full robotic assistance. As a reference, one set of 72 samples from 16 to 19-year-olds (mean 16.8 years) and one set of 72 samples from 43 to 79-year-olds (mean 58.1 years) were selected from the 407 samples of chewing-stimulated saliva, with DNA manually extracted with the Sigma Bacterial DNA Isolation kit.

Evaluating DNA quantities in sequencing. Serial dilutions of 800 ng to 6.5 ng DNA from three saliva samples were used for the full-length 16S rRNA gene PCR amplification step, followed by ONT sequencing and Emu-eHOMD bioinformatic processing as described above. All evaluation steps were done using the MicrobiotaProcess pipeline (https://github.com/YuLab-SMU/MicrobiotaProcess), and the commands used are given below. To adjust for variation in sequencing depth, all samples were rarefied to a depth of 49,365 reads using the [mp_rrarefy] and plotted using [mp_cal_rarecurve]. The number of HMTs and Shannon diversity index were retrieved using the [mp_cal_alpha] on RareAbundance and plotted using the [mp_plot_alpha] command, and the Bray-Curtis indexes were calculated using the [mp_cal_dist] with Hellinger standardization method. The [mp_adonis] command with 9999 permutations was used for evaluating Bray-Curtis distance between DNA amount subgroups. Alpha, a two-way mixed consistency model was used for the assessment within sample agreement over the range of DNA amount (IBM SPSS Statistics, v 29.0.1.0).

Statistical analyses. Descriptive statistics, including mean, median, and 95% confidence interval (95%CI), intraclass correlations, and Spearman correlation were extracted using STATA (16.1). Epi2me Lab (labs.epi2me.io) was used together with NextFlow wf-16S workflow for the evaluation of Kraken2 and Minimap2 performance and accuracy to classify ONT reads. For assessment of sequencing depth, a- and b diversity, R and R-studio (23.12.1), with the package MicrobiotaProcess[43], was used with the included packages: forcats_1.0.0, ggstar_1.0.4, tidytree_0.4.6, treeio_1.26.0, ggtreeExtra_1.12.0, ggtree_3.10.1, shadowtext_0.1.3, phyloseq_1.46.0, ggplot2_3.5.0, knitr_1.45, and Patchwork_1.2.1) for processing the raw reads and evaluating similarities and difference between subgrous. More in-depth, a Metadata, ASV, and Taxonomy file were imported as MPSE objects into the MicrobiotaProcess pipeline. The [mp_rrarefy], [mp_cal_rarecurve], [mp_plot_rarecurve], [mp_cal_alpha], [mp_plot_alpha], [mp_cal_abundance], and [mp_plot_abundance] were used to rarefy all samples to the same sequencing depth and to calculate, evaluate, and plot taxonomy and a-diversity measurements including species richness, Shannon, Simpson, Chao1, and Pielou. Compositional comparison of sequencing platforms or DNA extraction kits were performed by the [mp_decostand], [mp_cal_dist], [mp_plot_dist], [mp_cal_pcoa], and [mp_adonis] using the Bray-Curtis distance matrix. The [mp_plot_ord] and [mp_diff_analysis] were used to extract feature differences between subgroups.

Data availability

The microbiota data generated during the current study are available in BioProject ID PRJNA1197754. Anonymized data related to the 407 saliva samples analyzed during the current study are not publicly available due to GDPR restrictions but are available from the corresponding author upon reasonable request and fulfillment of appropriate juridical requirements.

Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Anders Esberg[1] ✉, **Niklas Fries**[1], **Simon Haworth**[2] **& Ingegerd Johansson**[1]
[1]Department of Odontology, Umeå University, Umeå, Sweden. [2]Bristol Dental School, Bristol, UK. ✉e-mail: anders.esberg@umu.se

Received: 15 July 2024; Accepted: 6 December 2024
Published online: 18 December 2024

References

1. Tuganbaev, T., Yoshida, K. & Honda, K. The effects of oral microbiota on health. *Science* **3769**, 34–936 (2022).
2. van der Meulen, T. A. et al. Shared gut, but distinct oral microbiota composition in primary Sjögren's syndrome and systemic lupus erythematosus. *J. Autoimmun.* **97**, 77–87 (2019).
3. Santacroce, L. et al. Oral microbiota in human health and disease: A perspective. *Exp. Biol. Med.* **248**, 1288–1301 (2023).
4. Zhang, X. et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* **21**, 895–905 (2015).
5. Atarashi, K. et al. Ectopic colonization of oral bacteria in the intestine drives T_H1 cell induction and inflammation. *Science* **358**, 359–365 (2017).
6. Nagakubo, D. & Kaibori, Y. Oral Microbiota: The Influences and Interactions of Saliva, IgA, and Dietary Factors in Health and Disease. *Microorganisms* **11**, 2307 (2023).
7. Belstrøm, D. The salivary microbiota in health and disease. *J. Oral. Microbiol.* **12**, 1723975 (2020).
8. Leake, S. L., Pagni, M., Falquet, L., Taroni, F. & Greub, G. The salivary microbiome for differentiating individuals: proof of principle. *Microbes Infect.* **18**, 399–405 (2016).
9. Cameron, S. J., Huws, S. A., Hegarty, M. J., Smith, D. P. & Mur, L. A. The human salivary microbiome exhibits temporal stability in bacterial diversity. *FEMS Microbiol. Ecol.* **91**, fiv091 (2015).
10. Lim, Y., Totsika, M., Morrison, M. & Punyadeera, C. The saliva microbiome profiles are minimally affected by collection method or DNA extraction protocols. *Sci. Rep.* **7**, 8523 (2017).

11. Caselli, E. et al. Defining the oral microbiome by whole-genome sequencing and resistome analysis: the complexity of the healthy picture. *BMC Microbiol.* **20**, 120 (2020).
12. Baker, J. L., Mark Welch, J. L., Kauffman, K. M., McLean, J. S. & He, X. The oral microbiome: diversity, biogeography and human health. *Nat. Rev. Microbiol.* **22**, 89–104 (2024).
13. Hou, K. et al. Microbiota in health and diseases. *Signal Transduct. Target Ther.* **7**, 135 (2022).
14. Zagai, U., Lichtenstein, P., Pedersen, N. L. & Magnusson, P. K. E. The Swedish Twin Registry: Content and Management as a Research Infrastructure. *Twin Res. Hum. Genet* **22**, 672–680 (2019).
15. Branton, D. et al. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).
16. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).
17. Zhang, T. et al. The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. *Appl. Environ. Microbiol* **89**, e0060523 (2023).
18. Sereika, M. 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* **19**, 823–826 (2022).
19. Curry, K. D. et al. Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. *Nat. Methods* **19**, 845–853 (2022).
20. Curry, K. D. et al. Microbial Community Profiling Protocol with Full-length 16S rRNA Sequences and Emu. *Curr. Protoc.* **4**, e978 (2024).
21. Cha, T. et al. Gut microbiome profiling of neonates using Nanopore MinION and Illumina MiSeq sequencing. *Front. Microbiol.* **14**, 1148466 (2023).
22. Lu, J. et al. Metagenome analysis using the Kraken software suite. *Nat. Protoc.* **17**, 2815–2839 (2022).
23. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
24. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
25. Escapa, F. et al. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. *Microbiome* **8**, 65 (2020).
26. Chorlton, S. D. Ten common issues with reference sequence databases and how to mitigate them. *Front Bioinform* **4**, 1278228 (2024).
27. Hsieh, Y. P., Hung, Y. M., Tsai, M. H., Lai, L. C. & Chuang, E. Y. 16S-ITGDB: An Integrated Database for Improving Species Classification of Prokaryotic 16S Ribosomal RNA Sequences. *Front. Bioinform.* **2**, 905489 (2022).
28. Haubek, D., Poulsen, K., Westergaard, J., Dahlèn, G. & Kilian, M. Highly toxic clone of Actinobacillus actinomycetemcomitans in geographically widespread cases of juvenile periodontitis in adolescents of African origin. *J. Clin. Microbiol.* **34**, 1576–1578 (1996).
29. Bertolo, A., Valido, E. & Stoyanov, J. Optimized bacterial community characterization through full-length 16S rRNA gene sequencing utilizing MinION nanopore technology. *BMC Microbiol.* **24**, 58 (2024).
30. Veenman, F. et al. Oral microbiota of adolescents with dental caries: A systematic review. *Arch. Oral. Biol.* **161**, 105933 (2024).
31. Bhaumik, D., Manikandan, D. & Foxman, B. Cariogenic and oral health taxa in the oral cavity among children and adults: A scoping review. *Arch. Oral. Biol.* **129**, 105204 (2021).
32. Velsko, I. M., Chakraborty, B., Nascimento, M. M., Burne, R. A. & Richards, V. P. Species Designations Belie Phenotypic and Genotypic Heterogeneity in Oral Streptococci. *mSystems* **3**, e00158–18 (2018).
33. Eriksson, L., Lif Holgerson, P., Esberg, A. & Johansson, I. Microbial Complexes and Caries in 17-Year-Olds with and without *Streptococcus mutans*. *J. Dent. Res* **97**, 275–282 (2018).
34. Esberg, A., Haworth, S., Kuja-Halkola, R., Magnusson, P. K. E. & Johansson, I. Heritability of Oral Microbiota and Immune Responses to Oral Bacteria. *Microorganisms* **8**, 1126 (2020).
35. Caporaso, J. G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
36. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
37. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 [published correction appears in Nat Biotechnol. 2019 Sep;37(9):1091]. *Nat Biotechnol.* **37**, 852–857 (2019).
38. Chen, T. et al. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database*. **2010**, baq013 (2010).
39. Escapa, I. F. 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* **3**, e00187–18 (2018).
40. Cole, J. R. et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42**, D633–D642 (2014).
41. Fries, N. et al. A Polygenic Score Predicts Caries Experience in Elderly Swedish Adults. *J. Dent. Res.* **103**, 502–508 (2024).
42. Dewhirst, F. E. et al. The human oral microbiome. *J. Bacteriol.* **192**, 5002–5017 (2010).
43. Xu, S. et al. MicrobiotaProcess: A comprehensive R package for deep mining microbiome. *Innovation* **4**, 100388 (2023).

Acknowledgements

We acknowledge the Swedish Infrastructure for Medical Population-based Life-course and Environmental Research (SIMPLER) and The Swedish Twin Registry (STR) for access to data. The SIMPER infrastructure is managed by Uppsala University and receives funding through the Swedish Research Council under grant numbers 2017-00644 and 2021-00160. The Swedish Twin Registry is managed by Karolinska Institutet and receives funding through the Swedish Research Council under grant no 2017-00641. We also acknowledge Agnetha Rönnlund for assistance in sample collection and DNA extraction of the mock and clinical samples. This study was funded by TUA Research Funding (grant number RV-990187), County Council of Västerbotten, Sweden. SH receives support from Wellcome through grant 227534/Z/23/Z. The funder played no role in the study design, data collection, analysis, and interpretation of data or the writing of this manuscript.

Author contributions

A.E. and I.J. conceptualized the study. A.E. performed microbiota analyses and bioinformatics processing with support from N.F. A.E. and I.J. performed statistical analyses. A.E., I.J., and S.H. wrote the first draft. All authors reviewed and edited the manuscript, and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

The study was ethically approved by the Ethics Review Authority in Sweden (dnr 2018/335-31 and dnr 09-134 M). The parts involving saliva donors followed the Helsinki Declaration including that participation was voluntary and that all participants had given written informed consent to participate.