A high-quality genomic catalog of the human oral microbiome broadens its phylogeny and clinical insights

Jun Hyung Cha[1], Nayeon Kim[1], Junyeong Ma[1], Sungho Lee[1], Geon Koh[1], Sunmo Yang[1], Samuel Beck[2], Iksu Byeon[3], Byunguk Lee[3], and Insuk Lee[1,4,5]*

[1]Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul 03722, Republic of Korea

[2]Department of Dermatology, Center for Aging Research, Chobanian & Avedisian School of Medicine, Boston University, Boston, MA 02118, USA

[3]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 34141, Republic of Korea

[4]POSTECH Biotech Center, Pohang University of Science and Technology (POSTECH), Pohang 37673, Republic of Korea.

[5]DECODEBIOME Co., Ltd. Incheon 21983, Republic of Korea

*Corresponding authors:

Insuk Lee

Tel: +82-10-4186-8706, E-mail: insuklee@yonsei.ac.kr

Abstract

Understanding the taxonomic and functional diversity of the human oral microbiome requires comprehensive genomic catalogs. We present the human reference oral microbiome (HROM), with 72,641 high-quality genomes from 3,426 species, including 2,019 novel species, significantly improving metagenomic sequence read classification over existing catalogs. Notably, HROM unveils 1,137 novel candidate phyla radiation (CPR) species, establishing Patescibacteria as the most prevalent phylum in the oral microbiota, with oral Patescibacteria forming a distinct clade from environmental Patescibacteria. We also identify an oral CPR subclade associated with periodontitis, which complement Porphyromonas gingivalis in predicting the disease. Finally, comparing HROM with reference genomes of the human gut microbiome reveals significant taxonomic and functional divergence between the oral and gut microbiomes. We identify oral-gut transmitter species and demonstrate that their relative abundance in gut microbiota is predictive of intestinal, cardiovascular, and liver diseases, highlighting the clinical importance of the oral microbiota in systemic disorders.

Introduction

The human oral cavity serves as a gateway for microbial colonization. The microbial community within this environment is both structured and dynamic, with its collective functions having a significant impact on overall human health[1]. Dysbiosis of the oral microbiome has been linked to oral diseases such as dental caries and periodontitis[2]. Moreover, recent studies suggest a connection between oral microbiota and various systemic diseases, including gastrointestinal, cardiovascular, neurological, autoimmune diseases, and cancers[3]. As a result, understanding the diversity and functions of human oral microbiota is becoming increasingly important in healthcare research.

The initial survey of the human microbiome revealed that the gut and oral cavity are two primary body sites with highly diverse microbial communities[4]. Recent advances in genome-resolved metagenomics have uncovered numerous novel species through metagenome-assembled genomes (MAGs), a culture-independent method that reveals the true diversity of the biome and identifies bacterial genomes previously undetectable using culture-based approaches[5]. In the case of the human gut microbiome, several studies have demonstrated that the MAG strategy, involving sequential steps such as assembly and binning, has successfully identified under-represented clades and expanded our ecological and functional understanding of the previously unknown repertoire of the gut microbiome[6-8]. However, efforts to catalog the oral microbiome using MAGs have been limited in comprehensiveness and quality[9]. The most widely used genome database for the human oral microbiome, eHOMD[10], primarily provides genomes for oral prokaryotic species, which are predominantly based on isolate genomes.

We have assembled MAGs from publicly available whole metagenomic sequencing (WMS) data from 7,956 diverse oral samples, alongside isolate genomes from the human oral cavity, resulting in the Human Reference Oral Microbiome (HROM). This genomic catalog provides 145,149 prokaryotic genomes of medium quality (MQ, completeness ≥ 50%, contamination < 5%), with 72,641 near-complete genomes (NC, completeness ≥ 90%, contamination < 5%) comprising 3,426 species. The HROM expands the phylogeny of the oral microbiome, including numerous novel species within the phylum Patescibacteria, a candidate phyla radiation (CPR) group[11] known for its enigmatic lifestyle. Our analysis revealed that oral Patescibacteria species form a distinct clade compared to those found in the environment, with a subclade associated with the oral disease periodontitis. Additionally, the HROM facilitated a comparison between oral and gut microbiota,

uncovering significant differences in their taxonomic and functional landscapes. Notably, we identified microbial species that occasionally transmit from the oral to the gut environment, offering novel clinical insights into systemic diseases. With its extensive and high-quality genomic information, the HROM opens new avenues for innovative oral microbiome research, poised to reveal new dimensions of oral microbial ecology and its impact on human health.

Results

Cataloging oral microbial genomes using clade-adjusted quality evaluation methods

For cataloging microbial genomes in the human oral environment, we collected WMS data from 7,956 oral samples from publicly available resources such as ENA[12], SRA[13], and CNGBdb[14] (downloaded in May 2022, **Supplementary Table 1**). These data represent samples from 21 countries, with the majority from the United States and China **(Fig. 1a)**. The samples cover various biogeographic sources within the oral cavity, with saliva and dental samples being the most common **(Fig. 1b)**. Most of WMS data from oral samples show high contamination rate of human DNA sequence reads (**Extended Data Fig. 1a**). Using a bioinformatic pipeline previously employed for cataloging gut microbial genomes[7], we constructed MAGs **(Fig. 1c, Extended Data Fig. 1b)**. Additionally, we manually searched for prokaryotic genomes from the human oral cavity in public resources such as eHOMD (v.3)[10], GenBank[15], RefSeq[16], PATRIC[17], and a pre-existing catalog[18] **(Supplementary Table 2)**.

Genome quality is evaluated based on completeness and contamination using CheckM[19], which relies on lineage-specific markers. We found that the CheckM lineage module often underestimates completeness for relatively novel clades, as their lineage-specific marker genes are not well established. To address this, we identified clades with underestimated completeness by comparing them with universal bacterial marker genes[19] and re-estimated their genome quality (**Supplementary Table 3**). For the phylum Patescibacteria, which lacks many universal bacterial marker genes, CPR marker genes[20] were used for re-estimation. This quality re-estimation process significantly improved completeness of 12,606 genomes **(Fig. 1c, Extended Data Fig. 1c)**. The re-estimated completeness was further validated by showing high correlation with both the percentage of hits from GTDB 120 single markers, an independent marker gene set, and the

completeness estimated by CheckM2[21], an updated method that resolves the limitation of lineage-specific markers using a machine learning approach **(Extended Data Fig. 1d)**.

Additionally, genomes were subjected to quality control based on genome chimerism using the Genome UNClutterer (GUNC)[22], excluding genomes with clade separation scores (CSS) exceeding 0.45. This resulted in a total of 145,149 non-redundant genomes meeting the MQ genome criteria, represented by 5,113 species clusters **(Supplementary Table 4-5)**. Of these genomes, 72,641 NC genomes represent 3,426 species, with 2,019 (59%) unclassified at the species level by GTDB-tk[23], indicating their novelty. Species clusters consisting exclusively of MQ genomes typically contain only a single low-quality genome, making it unlikely that they correspond to real organisms. For example, clustering with MQ genomes yields 1,687 additional species; however, 1,374 of them (81.4%) are singletons. Furthermore, MQ genomes, which lack substantial portions of genomic content, can mislead functional analyses of species. Therefore, we focused on 3,426 species composed of NC genomes for HROM-based taxonomic profiling of microbiomes (**Fig. 1d**). Notably, only 486 of the 3,426 (~14.2%) species have isolate genomes, indicating that the majority of oral microbes remain uncultivated.

HROM outperforms pre-existing genomic catalogs in oral WMS read classification

We compared HROM to pre-existing genomic catalogs for the human oral microbiome. Notably, the expanded human oral microbiome database (eHOMD)[10], one of the most widely used human oral microbiome databases, predominantly consists of isolated/culturable genomes. Although eHOMD is frequently used as a reference genome database for oral microbiome analysis, it includes genomes not isolated from the human oral cavity. For instance, genome SEQF2350 (GenBank accession = GCA_000259485.1, biosample accession = SAMN00761858) is recorded as being human urogenital tract isolate. To address this issue, we manually inspected the genomes in eHOMD and selected only 752 confirmed oral microbial genomes from the 2,213 total genomes for inclusion in HROM, ensuring it comprises only authentic human oral-origin genomes.

HROM consists of 72,641 NC genomes, approximately 32.8 times the total number of genomes in eHOMD (**Fig. 2a**). When applying the same species-level clustering scheme, HROM included 7.37 times more species based on NC genomes (NC species) compared to eHOMD (**Fig. 2b**). We further compared HROM with other genome catalogs, including COGR[24], human-oral-v1-0

125 microbiome catalog in MGnify repository[25], and the largest existing MAG-based oral genome
126 catalog of Zhu et al[9], although the latter provides only representative genomes to the public. When
127 applying the same genome quality criteria as HROM, only 1,048 NC species were identified in the
128 Zhu et al. database. Overall, our analysis demonstrated that HROM contains substantially more
129 NC genomes and corresponding species compared to other genome catalogs included in this
130 comparison.

131 The improved comprehensiveness of genome catalogs may also enhance the taxonomic
132 classification of WMS reads. We aligned sequence reads from six independent public human oral
133 WMS datasets, comprising 1,129 samples not used in the construction of HROM **(Supplementary**
134 **Table 6)** to various genome catalogs using Kraken2[26]. HROM significantly outperformed other
135 genome catalogs in read classification rate **(Fig. 2c)**. Furthermore, using concatenated conspecific
136 non-redundant genomes for each HROM species further improved performance compared to using
137 representative genomes alone. Sequence reads unaligned by eHOMD were primarily classified
138 into the phyla Bacteroidota, Proteobacteria, Firmicutes, and Patescibacteria across most oral cavity
139 isolation sources **(Fig. 2d, Extended Data Fig. 2a)**. These findings suggest that the enhanced oral
140 WMS read classification capabilities of HROM will enable more accurate taxonomic profiling of
141 human oral microbiomes.

142

143 **DNA-based species profiling with HROM outperforms marker-based profiling of the human**
144 **oral microbiota**

145 One of the major applications of genomic catalogs of microbiomes is assembly-free taxonomic
146 profiling of WMS samples. Although a marker-based method, MetaPhlAn4[27], is widely used, we
147 expect that the improved read classification rate provided by HROM will enable more accurate
148 species-level taxonomic profiles through DNA-based methods, such as alignment using Kraken2[26]
149 followed by reassignment using Bracken[28]. To compare these two approaches of species profiling
150 of oral microbiome, we generated custom databases for both MetaPhlAn4 and Kraken2 based on
151 HROM. For Kraken2, we created two versions of custom databases: one based on species
152 representative genomes and the other on concatenated conspecific genomes for each species.
153 While building Kraken2 custom databases from a given genome catalog is relatively
154 straightforward, creating a custom database of species-specific markers for MetaPhlan4 requires

155 significant efforts. This may explain why most pre-existing genomic catalogs provide custom
156 databases for Kraken2 but not for MetaPhlan4. Based on the methodological principles of the
157 MetaPhlAn4 protocols, we constructed a species-specific marker database for HROM, but marker
158 genes could only be identified for 1,603 species (46.8% of HROM species; **Extended Data Fig.**
159 **2b**). We observed that species-specific marker genes could not be identified for species undergoing
160 rapid speciation, characterized by widely dispersed branches in the phylogenetic tree, such as those
161 in the phyla Patescibacteria and Actinobacteriota. These rapidly evolving species exhibit a high
162 density of single nucleotide variations (SNVs)[7] but lack gene-level diversification. The absence of
163 marker genes for these clades may reduce the accuracy of HROM-based species profiling of the
164 oral microbiome when using MetaPhlAn4.

165 The limited marker genes for fast-evolving clades may reduce the performance of species profiling
166 using the MetaPhlAn4 compared to Kraken2. To evaluate the accuracy of these methods under
167 varying species complexity, we generated simulated WMS samples using HROM genomes for
168 communities containing 40, 150, 600, 1,000, and 1,500 species, with either single or multiple
169 strains. Using WMS samples from healthy individuals, we surveyed the number of HROM species
170 across oral cavity biogeographic locations. Except for the throat, most environments harbored
171 between 700 and 1,100 species per individual (**Extended Data Fig. 2c**). To improve accuracy, we
172 applied a Kraken confidence threshold of 0.2 to reduce false positives and performed genome-size
173 normalization for better abundance measurement[29]. Both MetaPhlAn4 and Kraken2/Bracken with
174 concatenated genomes achieved near-perfect precision in species profiling. However, MetaPhlAn4
175 showed much lower recall than Kraken2/Bracken. Consequently, the F1 score, which balances
176 precision and recall, indicated that Kraken2/Bracken with the HROM custom database based on
177 concatenated genomes performed best (**Fig. 2e**). Furthermore, Kraken2/Bracken outperformed
178 MetaPhlAn4 in Bray-Curtis similarity-based evaluations of species abundance profiles. Finally,
179 we confirmed that the application of the Kraken confidence threshold and genome-size
180 normalization significantly improved species profiling accuracy (**Extended Data Fig. 2d**). These
181 results demonstrate that the Kraken2/Bracken pipeline using concatenated species genomes,
182 genome-size normalization, and a confidence threshold provides the most accurate species profiles
183 for the human oral microbiota.

184

185 **Fast-evolving microbes dominate the oral microbiome**

186 One intriguing observation from the human oral microbiome is the high proportion of rapidly
187 evolving species, which hinders the identification of species-specific marker genes for marker-
188 based profiling. We were able to identify species-specific markers for fewer than half of the oral
189 microbial species. Similarly, in a previous study of the gut microbiome, we identified 619 species
190 with high divergence within the *Collinsella* genus, accounting for ~11% of gut microbial species[7].

191 To systematically identify rapidly evolving species in the oral microbiota, we analyzed the
192 distribution of average phylogenetic distances to the 10 nearest neighbors for each species. Rapidly
193 evolving species are expected to exhibit shorter phylogenetic distances to their close relatives
194 compared to those with typical evolutionary rates. Consistently, we observed that species lacking
195 species-specific marker genes had significantly lower distance scores, whereas species with such
196 markers fell within a higher distance range (**Extended Data Fig. 3a**). This positive association
197 between phylogenetic distance and the availability of marker genes highlights the limitations of
198 marker-based profiling for rapidly evolving species. Using a phylogenetic distance threshold
199 derived from the mean of the distance scores, which effectively separates the two modes, we
200 identified 2,637 fast-evolving species, accounting for 77.0% of the 3,426 oral microbial species.
201 This large proportion of fast-evolving species, which lack species-specific markers, underscores
202 the unsuitability of marker-based profiling for the oral microbiome.

203 From the HROM phylogenetic tree, we identified six clades enriched in fast-evolving species: the
204 phylum Patescibacteria, the family Campylobacteraceae, and four genera (*Rothia*, *Lancefieldelta*,
205 *Oribacterium*, and *Streptococcus*) (**Extended Data Fig. 3b**). We hypothesized that an accelerated
206 evolutionary rate is associated with increased genetic variation, anticipating a negative correlation
207 between phylogenetic distance to close species and single nucleotide variant (SNV) density,
208 consistent with findings from our previous study[7]. Among 514 species with more than 10
209 conspecific genomes, we observed this inverse correlation, indicating higher genetic variability in
210 fast-evolving species (**Extended Data Fig. 3c**). Bacterial genome evolution is driven not only by
211 genetic variability but also by adaptability to new environments. Accessory genes, which often
212 provide selective advantages, may play a significant role in adaptive evolution[30]. Consistent with
213 this, we observed a correlation between phylogenetic distance to close species and the proportion
214 of core genes in the pangenome for the 514 species, suggesting larger accessory genomes in fast-

evolving species (**Extended Data Fig. 3d**). Together, these findings confirm that species with shorter phylogenetic distances to close relatives exhibit faster evolutionary rates compared to others.

HROM reveals a novel functional repertoire of the human oral microbiome

To explore the functional landscape of the human oral microbiome, we identified protein-coding genes from 145,149 MQ genomes using Prokka[31]. The predicted protein sequences were clustered at 100%, 95%, 90%, 70%, and 50% sequence identity using MMseqs2[32], yielding approximately 76.2, 15.2, 8.5, 2.7, and 2.0 million protein clusters, respectively. We identified clusters of species within the same phylum based on Jaccard distance calculated from the presence/absence profiles of protein families at 50% sequence identity. This suggests that the phylogenetic relationships among HROM species align with their functional similarities (**Fig. 3a**). Notably, HROM cataloged many novel protein families that were not previously identified in bacterial protein databases. When comparing HROM protein families at 50% sequence identity with those in proGenomes3[33], the most comprehensive genome catalog containing nearly a million publicly available bacterial genomes, we found a large number of unique protein families in HROM that did not cluster with any known bacterial proteins (**Fig. 3b**). This highlights the substantial contribution of HROM to expanding the repertoire of bacterial protein families, particularly for the phyla Spirochaetota and Patescibacteria.

We performed functional annotation for these protein families using the homology-based method eggNOG-mapper[34], which assigns various functional terms including Cluster of Ortholog Groups (COG)[35], KEGG Orthology (KO)[36], KEGG Pathway[36], Gene Ontology (GO)[35]. We observed that most protein families lacked functional annotations across these terms (**Fig. 3c**). Notably, GO annotations covered only 6% and 3% of protein families at 90% and 50% sequence identity threshold, respectively. To address this limitation, we applied DeepGOPlus[37], a deep learning-based GO term prediction tool, which substantially improved the annotation rate, particularly for cellular component terms (**Fig. 3d**).

The landscape of KEGG pathways revealed that glycolysis and gluconeogenesis were highly prevalent across all HROM phyla, with other metabolic pathways also common in most phyla, while phylum Patescibacteria exhibited a marked depletion of metabolic modules (**Fig. 3e**,

245 **Supplementary Table 7**). In addition, amino acid and nucleotide metabolism were notably absent
246 in phyla Myxococcota and Verrucomicrobiota, and energy metabolism was lacking in phyla
247 Methanobacteriota and Thermoplasmatota. The landscape of antibiotic resistance genes showed
248 that glycopeptide resistance gene targeting vancomycin (vanT, vanY and vanW) were prevalent in
249 oral microbiome (**Fig. 3f**). Diverse defense mechanisms against phage infection were found across
250 phyla, with restriction modification systems were most prevalent **(Extended Data Fig. 4a)**. The
251 landscape of biosynthetic gene clusters (BGC) showed that categories of secondary metabolites
252 RiPP-like, arylpolyene, T3PKS, NRPS were relatively more prevalent in HROM **(Extended Data**
253 **Fig. 4b)**.

254 Many small peptides encoded by bacterial small open reading frames (sORFs) also exhibit broad-
255 spectrum antimicrobial activates and play key roles in microbial interactions[38]. Using Macrel[39],
256 we predicted anti-microbial peptides (AMPs) from HROM, identifying 15,368 AMPs, of which
257 only 2,744 (~17.9%) clustered at family level with 26,108 peptides in the DRAMP[18], a public
258 repository of AMPs **(Fig. 3g, Supplementary Table 8)**. This highlights that HROM includes a
259 substantial number of previously unreported AMPs. The density of AMP per genome was
260 relatively low overall, with phyla Patescibacteria and Bacteroidota showing comparatively higher
261 mean densities.

262 In conclusion, HROM unveils an extensive repertoire of functions in the human oral microbiome
263 by identifying novel protein families and antimicrobial peptides. With improved functional
264 annotation and comprehensive profiling of the antibiotic resistome, phage defensome, and
265 secondary metabolism gene clusters, HROM serves as a valuable resource for facilitating
266 functional studies of the oral microbiome.

267

268 **Distinct genomic and functional features of oral Patescibacteria unveiled by HROM**

269 Intriguingly, we collected 7,321 NC genomes and 1,402 species from the phylum Patescibacteria,
270 a member of the Candidate Phyla Radiation (CPR) group within the human oral microbiome. Of
271 these, 7,107 were assembled using our HROM pipeline, and 2,903 genomes correspond to 1,137
272 (81%) novel species unclassified at the species level by GTDB taxonomy. Previous studies using
273 concatenated protein markers have identified the CPR as a monophyletic group encompassing
274 numerous phyla[40,41]. However, the GTDB taxonomy, a standardized bacterial classification system

275 based on genome phylogeny and rank normalization, has reclassified CPR as a single phylum,
276 Patescibacteria[42]. Patescibacteria are characterized by ultrasmall cell sizes, reduced genomic
277 capabilities, small genome sizes, limited metabolic capacity[20], and an epi-symbiotic lifestyle[43] with
278 host bacteria. Their obligate symbiosis is thought to drive rapid evolutionary processes,
279 contributing to their high genetic divergence[44] (**Extended Data Fig. 3b**).

280 Notably, oral Patescibacteria species exhibit smaller genome sizes compared to other oral bacterial
281 species (**Extended Data Fig. 5a**). Representative genomes of Patescibacteria lack 21 of the 120
282 universal single-copy genes used as GTDB markers, with these 21 markers showing a maximum
283 prevalence of only 1.5% (**Extended Data Fig. 5b, Supplementary Table 9**). In contrast, the mean
284 prevalence of all GTDB markers across other oral bacterial species is 93.19%, with a minimum of
285 65.17%, highlighting the significantly reduced genomic functionality of oral Patescibacteria. The
286 division and cell wall (*dcw*) cluster, comprising 17 genes involved in peptidoglycan precursor
287 synthesis, is reported to be highly conserved across the bacterial domain, including the CPR
288 superphylum[45]. Interestingly, oral CPR bacteria have retained most of the genes in the *dcw* cluster
289 (**Extended Data Fig. 5c**), suggesting that the reduced genome size reflects an inherent property of
290 oral Patescibacteria rather than an artifact of incomplete genome assembly.

291 Previously, this group lacked isolated representative genomes until recent studies[46]. The protein
292 content of CPR bacterial genomes is distinct from that of other bacterial genomes[47], highlighting
293 their functional uniqueness. The functional distinctiveness of oral Patescibacteria was
294 demonstrated by their separate clustering from other oral bacterial species based on protein family
295 profile distances (see **Fig. 3a**). Consistent with their reduced genome sizes, Patescibacteria are
296 comparably depleted in metabolic capabilities relative to other oral bacterial species (**Fig. 4a**).
297 However, oral Patescibacteria genomes exhibit a higher proportion of genes encoding
298 transmembrane proteins compared to other species (**Fig. 4b, Extended Data Fig. 5d**). This
299 suggests that oral Patescibacteria genomes may prioritize genes related to transport, consistent
300 with their parasitic lifestyle. Using KO and COG terms, we identified several genes in oral
301 Patescibacteria species that are enriched for biofilm formation functions compared to other species:
302 VpsQ (K20950) is related to the production of exopolysaccharide, which are crucial for mature
303 biofilm formation[48]; resuscitation-promoting factor (K21688) is crucial for biofilm formation in
304 some species[49] or containing catalytic domains that mediates biofilm formation[50]; type IV pilus-
305 related genes such as pilM (K02662, COG4972) and VirB6 (COG3704) indirectly contribute to

306 biofilm formation[51-53] (**Extended Data Fig. 5e-f, Supplementary Table 10**). These findings
307 suggest that oral Patescibacteria species play a significant role in biofilm formation within the oral
308 microbiota, emphasizing their importance in community structure and function.

309 Given our extensive collection of CPR group bacterial genomes from the human oral microbiome,
310 we investigated the phylogenetic and functional divergence of mammalian host-associated (MHA)
311 CPR bacteria, including oral Patescibacteria, compared to those inhabiting environmental niches
312 such as soil and water. To achieve this, we gathered CPR bacterial genomes from other isolation
313 sources to analyze the functional and phylogenetic traits of oral Patescibacteria in the context of
314 other known Patescibacteria (**Supplementary Table 11**). Genomes belonging to the phylum
315 Patescibacteria were retrieved from previous studies[54] and public repositories and were filtered
316 using criteria identical to those applied to the HROM. This effort resulted in the collection of 3,721
317 genomes, which were subsequently de-replicated into 2,963 species clusters (**Fig. 4c**). Notably,
318 84.6% (2,507 species) were singletons, likely reflecting the rapid speciation characteristic of this
319 group. Phylogenetic analysis revealed that Patescibacteria species from the human oral cavity are
320 phylogenetically closer to genomes isolated from the human gut, skin, and other mammalian hosts
321 compared to those isolated from environmental sources such as soil and water. Interestingly, over
322 98% of the 1,553 MHA Patescibacteria species were classified within the class Saccharimonadia,
323 highlighting their taxonomic distinctness and adaptation to mammalian host environments. Using
324 a dataset of 22,977 protein families from the 2,963 representative genomes of Patescibacteria[47],
325 we performed clustering based on the Jaccard distance between protein family profiles. This
326 analysis showed clear, non-overlapping clusters among species associated with different
327 environments (**Fig. 4d**), further supporting the functional adaptation of Patescibacteria genomes
328 to their respective environments.

329 We examined functions that were more prevalent in MHA or environmental Patescibacteria
330 (**Supplementary Table 12, Extended Data Fig. 5g**). Among the identified genes, the
331 ribonucleoside-triphosphate reductase gene (*nrdD*), which has previously shown high
332 phylogenetic diversity in MHA Patescibacteria[55], was found to be more prevalent in MHA
333 genomes. This gene enables survival in anaerobic conditions, such as those present in the oral
334 cavity or gut. Similarly, L-lactate dehydrogenase (*ldh*), another gene previously reported to be
335 unique to the MHA group[55], was more prevalent in MHA Patescibacteria, particularly in oral
336 species. The *ldh* gene is a signature of oral bacteria, facilitating the production of lactate, a highly

337 abundant metabolite in dental plaque. Additionally, ABC-type transport systems were enriched in
338 MHA Patescibacteria, indicating their enhanced ability to import and export essential resources
339 necessary for survival in host environments (**Extended Data Fig. 5h**).

340 We further examined pathways relevant to CPR bacteria. The arginine deiminase (ADI) system,
341 known to protect both CPR bacteria and their host bacteria from acid stress[56] and essential for the
342 growth of CPR bacteria[57], was found to have core genes (*ArcA, ArcB, ArcC*) and the associated
343 metabolic gene cluster (MGC) highly prevalent in MHA Patescibacteria (**Fig. 4e**). In contrast, the
344 Type IVa pilus system (T4aP), which is a known trigger for epi-symbiotic relationships with host
345 bacteria[58], was prevalent across all Patescibacteria. Similarly, competence genes such as *comEC*
346 and *comFC*, required for the processing of extracellular DNA and facilitating potential horizontal
347 gene transfer (HGT)[59], were found across all Patescibacteria. Lastly, MHA Patescibacteria
348 demonstrated a comparatively more diverse array of viral defense systems but exhibited a
349 depletion of antibiotic resistance genes relative to their environmental counterparts. This suggests
350 that MHA Patescibacteria may rely on host bacteria for protection against antibiotics while
351 maintaining a robust viral defense system to counter phage threats in the oral cavity, which is more
352 exposed to phages. Notably, oral CPR bacteria lacking the ADI system exhibited a higher
353 abundance of phage defense systems, suggesting distinct subgroups of oral CPR bacteria that
354 enhance host bacteria ability to either resist acid stress or defend against phage infection.

355

356 **Unveiling an oral CPR subclade associated with periodontitis**

357 CPR bacteria has been implicated for periodontitis, a serious gum disease characterized by
358 inflammation and eventual destruction of the bone supporting the teeth[44]. However, no specific
359 CPR subclade or functional elements directly associated with periodontitis have been identified.
360 Leveraging our extensive collection of oral CPR species genomes, we aimed to identify CPR
361 bacterial species linked to periodontitis.

362 Given the high similarity in gene contents among CPR species due to rapid speciation, univariate
363 analysis of individual species may be suboptimal. To address this, we clustered 18 groups of oral
364 CPR bacterial species based on Jaccard distances of protein family presence profiles (**Extended**
365 **Data Fig. 6a**) and tested their association with periodontitis using publicly available WMS data.
366 This analysis identified cluster 16 CPR bacteria (**Fig. 4f, Supplementary Table 13**) as the most

367 enriched group in saliva-sampled periodontitis cohorts (**Fig. 4g**). The relative abundance of cluster
368 16 CPR bacteria was significantly higher in two of four periodontitis cohorts and showed a similar
369 but non-significant trend in another cohort (**Fig. 4h**). Despite the limited number of datasets,
370 disease states explained a higher proportion of variance than dataset-specific factors (**Fig. 4i**).
371 These findings suggest an association between the abundance of oral CPR bacteria in cluster 16
372 and periodontitis. Further, phylogenetic analysis revealed that cluster 16 CPR bacteria comprise a
373 subclade of the genus *Nanosynbacter* (**Extended Data Fig. 6b-c**).

374 A well-known oral bacterial species implicated in periodontitis is *Porphyromonas gingivalis*. We
375 observed that the relative abundance of *P. gingivalis* within the oral microbiome predicts the
376 occurrence of periodontitis with an area under the receiver operating characteristic curve (AUROC)
377 of 0.74 (**Fig. 4j**). Notably, the relative abundance of cluster 16 CPR bacteria provided slightly
378 better predictive performance (AUROC = 0.76) than *P. gingivalis*. Moreover, the combined
379 relative abundance of *P. gingivalis* and cluster 16 CPR bacteria further improved periodontitis
380 prediction, achieving an AUROC of 0.82. This finding suggests that *P. gingivalis* and cluster 16
381 CPR bacteria are independently associated with periodontitis. Supporting this, we observed a
382 relatively low correlation between their relative abundances in periodontitis patients (**Fig. 4k**).
383 These oral microbiomes of patients were generally characterized by either abundant *P. gingivalis*
384 or abundant cluster 16 CPR bacteria.

385 We found fam08340 as the most prevalent protein family among cluster 16 species **(Fig. 4l,**
386 **Extended Data Fig. 6d-e).** This protein family lacks functional annotation, and no homologous
387 proteins with known functions were identified. Analysis of the ~150 amino acid N-terminal region
388 of fam08340 revealed a signal peptide linked to transmembrane region. To explore potential
389 structure-based functions, we predicted the 3-dimensional structure of the outer-membrane region
390 of the protein using ESMFold[60] and conducted a structural homology search against AlphaFold
391 database(UniProt50)[61] with Foldseek[62] (**Extended Data Fig. 6f**). Among the top hits, we identified
392 polycystic kidney disease (PKD) domain-containing proteins (**Extended Data Fig. 6g**). The PKD
393 domain, an 80-90 amino acid module characterized by a β-sandwich fold, facilitates ligand-binding
394 interactions and is critical in the pathogenesis of autosomal dominant polycystic kidney disease
395 (ADPKD)[63]. Notably, a carboxypeptidase from *P. gingivalis* contains three tandem PKD domains,
396 which are implicated in its role in disease pathogenesis[64]. This suggests that fam08340 of cluster

397 16 CPR bacteria may similarly contribute to periodontitis through mechanisms analogous to the
398 carboxypeptidase of *P. gingivalis*.

399 In summary, leveraging a substantially expanded catalog of CPR species genomes and their
400 proteins, we identified a subclade of oral CPR bacteria implicated in periodontitis and uncovered
401 proteins potentially involved in its pathogenesis.

402

403 **Functional divergence and taxonomic overlap between oral and gut microbiomes**

404 The human oral cavity and gut are physiologically distinct yet directly interconnected body sites.
405 This interconnection may lead to the divergence of microbiomes between two sites while allowing
406 microbial exchange and cross-site influence[65]. For example, a significant number of oral bacteria
407 are transferred to the gut daily, primarily through the ingestion of saliva[66-68]. To explore the
408 taxonomic and functional relationship between the microbiomes of the oral cavity and
409 gastrointestinal tract, we compared HROM with HRGM[7], a human reference gut microbial
410 genome catalog. To improve reliability of this study, we utilized HRGM2, an improved version of
411 HRGM exclusively composed of high-quality genomes from 4,824 species (manuscript submitted).

412 We first observed a substantially smaller genome size in the oral microbiota compared to the gut
413 microbiota (**Fig. 5a**). Since both HROM and HRGM2 contain only high-quality genomes with
414 completeness greater than 90%, this difference cannot be attributed to the quality of the genome
415 catalogs. This skewed distribution toward smaller genome sizes appears to be largely attributable
416 to oral CPR bacteria.

417 Next, we compared protein families at 90% and 50% sequence identity and found that the majority
418 (~88% and ~65%, respectively) were unique to the oral microbiome (**Fig. 5b**), indicating that the
419 oral microbiome is functionally distinct from the gut microbiome. The overlap of protein families
420 between the two microbiomes varied among phyla; notably, Spirochaetota and Patescibacteria
421 exhibited particularly low overlap, suggesting that these two phyla represent the most functionally
422 divergent taxonomic groups in the oral microbiome (**Fig. 5c**). These findings highlight that the
423 symbiotic microbes of the oral cavity and gut have adapted to their distinct environments through
424 functional divergence.

425 The anatomical connection between the oral cavity and gastrointestinal tract suggests the potential
426 for shared microbial species. Based on a criterion of 95% ANI (average nucleotide identity) of
427 genomes, we identified 330 species present in both the oral and gut microbiomes (**Supplementary**
428 **Table 14**). Most of these oral-gut shared species were classified within the phylum Firmicutes,
429 followed by Actinobacteriota and Proteobacteria (**Fig. 5d**).

430 We hypothesized that these shared species might possess functional characteristics that enable their
431 ubiquitous presence. Specifically, we propose that microbes capable of inhabiting multiple
432 environments are more likely to be metabolically independent. To test this, we used a metabolic
433 independence score, calculated as the summation of the completeness of 33 KEGG metabolic
434 modules known to be enriched in bacteria identified as good colonizers during fecal microbiota
435 transplantation (FMT)[69]. Comparing the metabolic independence scores of the oral-gut shared
436 species with those of non-shared species revealed that shared species exhibited significantly higher
437 metabolic independence (**Fig. 5e**).

438 Additionally, shared species are likely to survive in the harsh gastric environment during
439 transmission from the oral cavity to the gut. Since endospore formation enhances bacterial survival
440 in adverse conditions, we evaluated the sporulation capacity of these species using a previously
441 established sporulation score, which is based on the presence of 66 sporulation-predictive genes
442 in bacterial genomes[70]. Shared species showed significantly higher sporulation scores than non-
443 shared species (**Fig. 5f**), supporting their enhanced probability of successful transmission from the
444 oral cavity to the gut.

445 Furthermore, we found that oral-gut shared species exhibit a higher tendency for an open
446 pangenome compared to non-shared species. This was evidenced by a higher pangenome
447 regression coefficient (**Fig. 5g**), a lower proportion of core genome (**Fig. 5h**), and a higher
448 proportion of accessory genome (**Fig. 5i**). These findings align with previous observations that
449 bacteria with open pangenomes are more adaptable to diverse environments.

450

451 **Oral-gut transmitter species abundance in gut microbiota linked to systemic diseases**

452 Recent evidence suggests that oral microbes influence not only oral diseases but also a range of
453 systemic disorders, including colorectal cancer (CRC), inflammatory bowel disease (IBD), and

454 cardiovascular diseases[3]. Although the roles of oral microbes in systemic diseases are not yet fully
455 understood, growing evidence indicates that oral bacteria can migrate into the gastrointestinal tract
456 and contribute to pathological processes there[65]. Based on this, we hypothesized that some of the
457 330 oral-gut shared species are transmitters rather than dual-site residents. To distinguish oral-gut
458 transmitter species from oral-gut shared species, we focused on those prevalent in healthy oral
459 microbiomes but not in healthy gut microbiomes. Specifically, we selected shared species present
460 in more than 50% of healthy oral microbiome samples but less than 25% of healthy gut microbiome
461 samples on average across cohorts (**Fig. 6a**, **Supplementary Tables 15–16**). This approach
462 identified 42 oral-gut transmitter species (**Supplementary Table 17**).

463 To validate these oral-gut transmitter species, we focused on the gut microbiome of CRC, an
464 intestinal disease known to be influenced by oral bacteria such as *Fusobacterium nucleatum*[71]. We
465 analyzed 12 publicly available CRC gut microbiome datasets (comprising 910 healthy individuals
466 and 877 patients) taxonomically profiled with HRGM2 (**Supplementary Table 18-19**). Using
467 microbial species abundance profiles from nine datasets for training and three datasets for
468 validation, we constructed disease classifier with a random forest algorithm. Species were then
469 prioritized for feature importance based on their random forest feature weights. Remarkably, 55%
470 of the top 20 important species associated with CRC were identified as oral-gut transmitter species
471 (**Fig. 6b**). Furthermore, CRC prediction using the sum of the relative abundances of 17 oral species
472 (oral-gut shared species and transmitter species) among the top 20 important species achieved a
473 significant performance, with a mean AUROC value of 0.834 (**Fig. 6c**). Additionally, we observed
474 that the co-abundance networks of these 17 oral species were enhanced in the gut microbiomes of
475 CRC patients compared to healthy controls (**Fig. 6d**, **Extended Data Fig. 7**), suggesting that the
476 co-translocation of these oral microbial species into the gut microbiome is associated with CRC.

477 Next, we tested whether the relative abundance of these 42 oral-gut transmitter species increased
478 in the gut microbiome of systemic diseases, including CRC. To do this, we measured the fold
479 change in the sum of relative abundance of the oral-gut transmitters in patient fecal microbiomes
480 compared to healthy controls across multiple disease-associated cohorts, encompassing 24 disease
481 types (**Supplementary Table 18-19**). We observed a substantial and consistent increase in the
482 relative abundance of oral-gut transmitters species in gut microbiome samples from patients with
483 intestinal diseases (e.g., CRC, Crohn's disease, ulcerative colitis, irritable bowel syndrome),
484 cardiovascular diseases (e.g., atrial fibrillation, subclinical carotid atherosclerosis, atherosclerotic

485 cardiovascular disease), and liver cirrhosis (**Fig. 6e**). Furthermore, the sum of the relative
486 abundance of the 42 transmitter species in the gut microbiome showed high predictive power for
487 these diseases (**Fig. 6f**). For other diseases such as COVID-19, Parkinson's disease, and type 2
488 diabetes, we also observed an increased relative abundance of oral-gut transmitters in patient gut
489 microbiomes compared to healthy controls. However, the effect sizes were smaller or inconsistent
490 across cohorts, resulting in only moderate disease prediction accuracy.

491 In summary, we identified 42 oral-gut transmitter species by comparing oral and gut bacterial
492 reference genomes and analyzing their differential prevalence between healthy oral and gut
493 microbiome samples. We uncovered their significant associations with various systemic diseases,
494 particularly intestinal diseases, cardiovascular diseases, and liver diseases. These findings
495 highlight the value of the genomic catalog of the oral microbiome in advancing our understanding
496 of how oral microbes influence systemic health.

497

498 Discussion

499 In this study, we present the most comprehensive genomic catalog of the human oral microbiome
500 to date, HROM, including 72,641 NC genomes assembled from 3,426 species, of which 2,019
501 (59%) are novel. This represents a substantial expansion over preexisting catalogs, increasing the
502 classification rate of metagenomic sequence reads significantly. This improvement facilitates more
503 accurate taxonomic profiling of WMS data from oral microbiota samples. Notably, we identified
504 a high prevalence of fast-evolving species among oral microbiota, which limits the identification
505 of reliable species-specific markers. This challenge results in suboptimal accuracy for marker-
506 based profiling methods like MetaPhlAn4. Therefore, we recommend DNA-based methods, such
507 as Kraken-Bracken, for species profiling of oral microbiota. To ensure versatility, we have
508 provided custom genome databases for both DNA-based and marker-based profiling methods in
509 the accompanying companion webserver.

510 Unexpectedly, we identified 1,137 novel CPR bacterial species from oral microbiota through
511 metagenome assembly, establishing Patescibacteria as one of the most prevalent phyla in the
512 human oral microbiota. Functional and phylogenetic analyses revealed that oral Patescibacteria
513 form a clade distinct from environmental Patescibacteria. This suggests that CPR bacteria, initially
514 introduced from exposed environments like groundwater (a major habitat for CPR bacteria), have

515 adapted to the host environment through functional repurposing of their genomes. Functional
516 analysis of oral Patescibacteria genomes revealed key roles, including (i) biofilm formation, aiding
517 microbial communities in the oral cavity; (ii) protection against acid stress, supporting host
518 bacteria survival; (iii) nutrient supply for host bacteria through specialized transport systems; (iv)
519 defense against phage infections, through a diverse and abundant defensome. Given their epi-
520 symbiotic relationships with host bacteria, the diversification of oral CPR bacterial species may
521 be driven by the diversity of their host bacteria. Identifying CPR-host species pairs and
522 understanding their co-evolution represents an important avenue for future oral microbiota
523 research. Additionally, we identified a novel oral CPR subclade associated with periodontitis and
524 demonstrated that they complement *P. gingivalis* in predicting the disease. These findings expand
525 our genomic understanding of oral pathogens and suggest that the oral CPR subclade may facilitate
526 early diagnosis of periodontitis in clinical settings.

527 Although HROM database includes MQ genomes, we recommend using only NC genomes and
528 species containing at least one NC genome (referred to as "NC genome species"). MQ genomes
529 can lose up to 50% of their genomic content, which may result in inaccuracies in taxonomic and
530 functional analyses. For instance, including MQ genomes may identify 1,687 additional species;
531 however, 81.4% of these are singletons, likely representing technical artifacts rather than true
532 biological diversity. We believe the exclusive use of high-quality genomes enhances the credibility
533 of conclusions regarding functional divergence and taxonomic overlap between oral and gut
534 microbiomes.

535 The transmission of oral bacteria into the gastrointestinal tract and their influence on systemic
536 health has been suggested in previous studies. However, to our knowledge, no prior studies have
537 systematically investigated oral bacteria associated with systemic diseases. By comparing HROM
538 with HRGM, we first identified 330 bacterial species shared between the oral and gut environments.
539 Among these, some may represent different subspecies adapted to each body site, while others
540 likely consist of oral species that occasionally transmit to the gut. These oral-gut transmitters may
541 contribute to or exacerbate pathogenic conditions, such as inflammation. For the first time, we
542 identified 42 oral-gut transmitter species based on their prevalence in oral and gut samples from
543 healthy individuals. We further demonstrated that their relative abundance in gut microbiota is
544 predictive of intestinal diseases, cardiovascular diseases, and liver diseases. This pathogenic effect
545 is likely mediated by systemic inflammation. In addition, we observed weak or inconsistent

546 associations between the 42 oral-gut transmitter species and diseases like COVID-19, Parkinson's
547 disease, and type 2 diabetes, which are also influenced by systemic inflammation. Consistent with
548 this, a recent study of patient cohorts undergoing allogeneic hematopoietic cell transplantation or
549 suffering from IBD revealed that the relative abundance of oral bacterial species in the gut
550 increases as a marker of depleted gut bacterial diversity, while their absolute abundance does not
551 significantly change[72]. Future research should confirm whether the 42 oral-gut transmitter species
552 similarly increase their relative abundance without a corresponding increase in absolute abundance.
553 This distinction could provide deeper insights into their role in systemic health and disease.

554 This study has certain limitations. While we compiled the most comprehensive collection of WMS
555 data from public repositories, the majority of the data originated from the USA and China. This
556 geographic bias may impact the representativeness of the oral reference microbiome, particularly
557 concerning ethnicity, geography, and lifestyle. Future updates should prioritize sampling from
558 underrepresented regions to reduce these biases. Additionally, since most HROM genomes were
559 assembled from short-read sequencing data, approximately 37% of species (1,265 out of 3,426)
560 lack the 16S rRNA region, one of the most challenging regions to assemble. Future efforts
561 employing long-read sequencing technologies could help resolve this limitation, enabling more
562 complete genomic assemblies of the human oral microbiome.

563

564 # Methods

565 **Collection of whole metagenome sequencing (WMS) data and bacterial isolate genomes**
566 **from the human oral cavity**

567 The WMS data were downloaded from public repositories, including ENA[12], SRA[13], and
568 CNGBdb[14]. Each dataset underwent manual inspection to identify samples originating from the
569 human oral cavity. Metadata, such as NCBI SRA biosample annotations, bioproject metadata, and
570 its corresponding literature, were scrutinized to confirm sample origins. To collect pre-assembled
571 genomes, we surveyed eHOMD (v3)[10], GenBank[15], RefSeq[16], and PATRIC[17]. For eHOMD,
572 genomes annotated as originating from non-oral environments were excluded. Biosample data,
573 GCF/GCA accessions, and relevant literature were reviewed to verify that genomes were isolated
574 from the human oral cavity. A similar process was applied to genomes from GenBank, RefSeq,
575 and PATRIC, ensuring only genomes with verified isolate records in biosamples or literature were

576 included. Pre-assembled MAGs derived from short-read sequencing were excluded, as most
577 samples previously used for MAG assembly were reanalyzed in this study. Additionally,
578 representative genomes from the oral bacterial genome catalog by Zhu et al[9] were included, but
579 genomes not isolated from oral cavity were manually identified and removed from the dataset.

580

581 **Metagenome-assembled genome (MAG) assembly**

582 For the collected WMS data, adapter sequences were trimmed, and low-quality reads were
583 removed using Trimmomatic (v0.39)[73]. The filtered reads were aligned to human genome
584 (vCRCh38[74]) using Bowtie2 (v2.3.5.1)[75] with the "--very-sensitive" option to remove contaminant
585 human DNA sequence reads. Most processed samples were assembled into scaffolds using
586 metaSPAdes (v3.15.3)[76]. For single-end sequencing samples and samples where metaSPAdes
587 failed to complete assembly for unknown reasons, MEGAHIT (v1.2.9)[77] was used instead.

588 Assembled scaffolds were binned to genome bins using three binning tools, MetaBAT2 (v2.12.1)[78],
589 MaxBin2 (v2.2.7)[79], and CONCOCT (v1.1.0)[80]. To estimate read coverage and depth for scaffolds,
590 processed reads were aligned back to the assembled contigs using Bowtie2 with the "--very-
591 sensitive-local" option. Binning with MetaBAT2 was performed with a minimum contig size of
592 1,500, while MaxBin2 and CONCOCT were used with a minimum contig size of 1,100. Bin
593 refinement was carried out using the result of all three binning methods as input for the
594 bin_refinment module of MetaWRAP (v1.3.2)[81].

595

596 **16S rRNA sequence prediction from bacterial genomes**

597 To identify 16S rRNA sequences from bacterial genomes, we employed Barrnap (v0.9)[82] with an
598 e-value threshold of 1e-5, resulting in the identification of 16S sequences in 1,802 genomes.
599 Predicting 16S sequences is particularly challenging due to the difficulty of assembling 16S
600 regions from short-read sequence data. For species where 16S rRNA sequences could not be
601 predicted, we expanded our search to genome sequences from the same species in the GTDB r207[83]
602 database. This approach enabled the prediction of 16S rRNA sequences for an additional 670
603 species. As a result, we successfully identified 16S rRNA sequences for 2,161 out of 3,426 species
604 (63%) in the HROM catalog.

605

606 **Evaluating genome quality and filtering chimeric genome**

607 Genome completeness and contamination were initially assessed using the "lineage_wf" module
608 of CheckM (v1.1.9)[19]. Additionally, tRNA sequences were identified and annotated using
609 itRNAscan-SE 2 (v2.0.9)[84] with the "-B" option for bacterial genomes and the "-A" option for
610 archaeal genomes. For rRNA annotation, Barrnap (v0.9)[82] was employed to identify 5S, 16S, and
611 23S rRNA sequences, applying e-value threshold of 1e-5 for 16S and 23S rRNA and 1e-4 for 5S
612 rRNA. Genomes were categorized into three quality tiers based on specific criteria[85]. Medium-
613 quality (MQ) genomes were defined as those with completeness ≥ 50%, contamination < 5%, and
614 a genome quality score (Completeness – 5 × Contamination) ≥ 50. Near-complete (NC) genomes
615 had a completeness ≥ 90%, contamination < 5%, and a genome quality score ≥ 50. High-quality
616 (HQ) genomes met the same criteria as NC genomes but also contained 5S, 16S, and 23S rRNA
617 sequences along with more than 18 tRNAs. To identify and exclude chimeric genomes, all
618 genomes were analyzed with GUNC (v1.0.4)[22]. Chimeric genomes were defined as those with a
619 clade separation score (CSS) > 0.45 and were removed from further analysis. Taxonomic
620 classification of the remaining non-chimeric genomes was performed using the "gtdbtk
621 classify_wf" module of GTDB-Tk2 (v2.1.1)[23].

622

623 **Re-estimation of genome completeness**

624 To address potential underestimation of genome completeness by the original CheckM[19]
625 "lineage_wf" workflow, which uses lineage-specific marker genes, we compared completeness
626 values from "lineage_wf" with those from CheckM "taxonomy_wf" for the domain "Bacteria,"
627 which relies on universal bacterial marker genes. Starting at the phylum level, we identified
628 taxonomic clades with significantly lower completeness in the "lineage_wf" results using the
629 Wilcoxon paired rank-sum test. We then focused on clades where genome quality classifications
630 improved based on re-estimated completeness. For the underestimated clades, genome
631 completeness was re-estimated using universal bacterial marker gene sets. In the case of the
632 phylum Patescibacteria, which lacks many universal bacterial marker genes, we used CPR-specific
633 marker genes[20] for re-estimation. The newly estimated completeness values were used for the final
634 genome quality classification. To validate the re-estimated completeness, we applied CheckM2[21],

635 a recently updated method that incorporated machine learning to address limitations of earlier
636 workflows. We assessed the correlation between completeness values from CheckM "lineage_wf"
637 and CheckM2, as well as the proportion of hits from the 120 single-copy marker genes used in
638 GTDB-Tk, using Spearman correlation analysis.

639

640 **Species-level clustering and removal of redundant genomes**

641 For genomes classified within the same order, MinHash distances were calculated using MASH[86]
642 (MASH sketch = 10,000), implemented in the dRep (v3.3.0)[87] compare module with the options
643 "--MASH_sketch 10000 --P_ani 0.8 --SkipSecondary". Primary clusters were formed through
644 hierarchical clustering with a distance threshold of 0.2. Within each cluster, the average nucleotide
645 identity (ANI) was calculated for every genome pair using MUMmer4 (v4.0.0rc1)[88]. Genomes
646 were then grouped into species clusters using average-linkage hierarchical clustering, implemented
647 in R with the fastcluster pachage[89]. Clustering was based on ANI distance with a species-level
648 cutoff of 95% identity (ANI distance of 0.05) and an alignment coverage threshold of 30% (0.3).
649 For each species cluster, the representative genome was selected as the genome with the highest
650 intactness score (S), calculated as $S = Completeness - 5 \times Contamination + 0.5 \times log_{10}(N50)$.
651 This scoring method follows the approach previously used in cataloging human gut microbial
652 genomes[7]. To further refine the dataset, redundant genomes within species clusters were removed
653 using average-linkage hierarchical clustering, applying an ANI cutoff of 99.9% identity (ANI
654 distance of 0.001) and an alignment coverage cutoff of 81% (0.81). This process identified and
655 removed 9,738 redundant genomes. Finally, a phylogenetic tree of the representative genomes was
656 constructed using IQ-TREE2 (v2.1.3)[90], based on the multiple sequence alignment of the 120
657 concatenated marker genes provided by GTDB-Tk (v2).

658

659 **Evaluation of sequence read classification performance**

660 To assess the read classification performance of genome catalogs, including HROM, we obtained
661 six additional WMS datasets from public repositories that were not used in the construction of
662 HROM. Each dataset contains at least 30 samples, originating from various oral biogeographic
663 sources. The datasets include buccal swab samples (PRJNA917836, n = 317), in-vitro oral biofilm

samples (PRJNA983519, n = 75), oral samples (PRJNA983219, n = 30), saliva samples (PRJNA917836, n = 27; PRJNA997379, n = 45), supragingival plaque samples (PRJEB54673, n = 554), and tongue samples (PRJNA879058, n = 81), resulting in a total of 1,129 samples. Using these validation data, we tested the performance of several human oral microbial genome catalogs: eHOMD v3[10], the Human Oral Genome Database by Zhu et al[9], the Cultivated Oral Bacteria Genome Reference (COGR)[24], and the human-oral-v.1.0.1 database from the MGnify repository[25]. For databases other than eHOMD, we applied the identical species clustering method and genome filtering scheme as used in HROM to build the near-complete species-level reference database built with representative genomes for Kraken2. All samples were aligned to each database using Kraken2[26] with a confidence score threshold of 0.2 to assess their read classification performance.

Constructing HROM custom database for marker-based species profiling

Marker-based species profiling was conducted using MetaPhlAn4[27] with a custom database derived from HROM genomes. To construct database of species-specific marker genes, we adapted the MetaPhlAn4 computational pipeline. Panaroo (v1.3.0)[91] was applied to 72,641 NC genomes and 3,246 species using the following options: "-c 0.90 --core_threshold 0.90 -f 0.5 --clean-mode strict --merge_paralogs", generating species-level pangenomes. For singleton species, genes/proteins from their representative species were used. Coding genes between 150–1,500 amino acids in length were selected from each species pangenomes and subsequently clustered using the linclust module of MMseqs2 (v2.1.9)[32] with a 90% identity. To identify clade-specific, non-spurious markers, initial candidates with a coreness (prevalence among conspecific genomes) greater than 50% were selected, with the threshold increased to 60% for species with more than 100 conspecific genomes. To avoid duplicate genes or paralogs, markers with a mean count exceeding 1.5 per genome within a species were excluded. Additionally, markers that clustered with genes from more than 10 other species were discarded. The remaining marker candidates were fragmented into 150-nt segments and aligned to other species and conspecific genomes using Bowtie2. Markers aligned to more than 50% of conspecific genomes but fewer than 1% of genomes from any other species were retained. Only perfectly unique markers (aligning to no other species) or quasi-markers (aligning to fewer than 1% of genomes from other species) were

693 included. After these filtering steps, up to 200 markers were selected per species to complete the
694 custom marker database.

695

696 **Benchmarking species profiling methods**

697 To evaluate taxonomic profiling methods using custom databases, we generated 5Gbp simulation
698 datasets with CAMISIM (v1.3)[92]. From the 72,641 NC species in HROM, we randomly selected
699 sets of 40, 150, 600, 1,000, and 1,500 species to simulate varying levels of species complexity
700 within microbial communities. Simulations were performed using the CAMISIM
701 "metagenomesimulation.py" module with the following configuration: "readsim = art_illumina,
702 profile = mbarc, size=5". For each genome set size, we generated 75 random samples under two
703 strain complexity scenarios: single strain (selecting one genome per species) and multiple strains
704 (selecting up to three genomes per species).

705 Taxonomic abundances were determined using both DNA-based and marker-based approaches.
706 For the DNA-based approach, we employed Kraken2 paired with Bracken[28] for species-level
707 reassignment. Custom Kraken2 databases were generated using either representative genomes or
708 concatenated conspecific genomes from HROM. Simulated reads were classified using Kraken2
709 with the options "--confidence = 0.2, --paired" for the respective HROM database. To minimize
710 potential biases in DNA-based taxonomic abundance profiling, abundances were normalized by
711 multiplying by the factor $10^6 / l_{species}$, where $l_{species}$ is the mean genome length of the species[9].
712 Relative abundances were calculated by dividing normalized read counts by their total sum.

713 For marker-based profiling, microbiome profiles of simulated reads were estimated using
714 MetaPhlAn4[27] with the HROM custom marker database.

715 Performance metrics, including precision, recall, and F1-score, were calculated using Scikit-
716 learn[93] (v0.24.2). Bray-Curtis distances between ground-truth profiles and estimated profiles were
717 computed using SciPy[94] (v1.5.3). Hits to conspecific genomes within each species were treated as
718 hits to the respective species.

719 Finally, to estimate the species count for each sample source from healthy individuals, we analyzed
720 WMS samples with a sequencing depth greater than 5 Gbp using a custom pipeline and the HROM

721 concatenated database. Species counts were derived from the species-level profiles generated for
722 each sample.

723

724 **Construction of human reference oral microbial protein catalog**

725 Protein coding sequence and its corresponding amino acid sequences were predicted for each
726 HROM genome using Prokka (v1.14.6)[31]. The predicted amino acid sequences were concatenated
727 and clustered at identity thresholds of 100%, 95%, 90%, 70%, and 50% using MMseqs2[32] with the
728 options "--kmer-per-seq 80 --cluster-mode 2". This process generated protein catalogs for each
729 identity threshold. Functional annotation of the protein catalogs was performed using eggNOG-
730 mapper v2 (v2.1.9)[34,95,96].

731

732 **Deep learning-based gene ontology annotation for microbial proteins**

733 DeepGOPlus[37](v1.0.1), a deep learning-based tool for gene ontology (GO) prediction, was used to
734 annotate representative proteins from the protein clusters in the HROM dataset. Only predictions
735 with a confidence score above 0.3 were retained. The GO term "cellular anatomical entity
736 (GO:0110165)", representing structural components within a cell, was excluded as it annotated
737 87.5% of the protein clusters, leading to an inflated prediction rate. Since DeepGOPlus was trained
738 on data that includes eukaryotic protein sequences, some predictions might reflect functions
739 specific to eukaryotes, potentially resulting in mispredictions for bacterial proteins. To address this,
740 only bacteria-associated GO terms were retained. Proteins derived from bacteria were identified
741 using the UniRef90 database (downloaded in September 2023), and their corresponding GO terms
742 were designated as bacteria associated.

743

744 **Protein functional landscape analysis**

745 For the protein cluster analysis, which included protein presence/absence and clustering of HROM
746 with other protein catalogs, protein sequences inferred from representative genomes were clustered
747 using MMseqs2 with the options "--min-seq-id 0.5 --kmer-per-seq 80 --cluster-mode 2". The
748 resulting presence/absence matrix, based on the clustering results, was imported as a table artifact
749 into QIIME2 (v2022.11)[97]. The "q2-umap"[98] module of QIIME2 was utilized to perform Uniform

750 Manifold Approximation and Projection (UMAP) analysis, using the Jaccard distance matrix
751 derived from protein content presence/absence data. Visualization of the UMAP results was
752 conducted with ggplot2[99].

753 To estimate the percentage of proteins clustered per species with the human gut microbiome,
754 protein sequences from HROM representative genomes were clustered with the protein catalog of
755 HRGM2 using MMseqs2 with the same option. For each HROM representative genome, the
756 number of proteins clustered with HRGM2 proteins was identified, and the ratio of shared proteins
757 was calculated by dividing the number of clustered proteins by the total protein count of the species.
758 For clustering percentage with proGenomes3[33], the representative protein sequence of
759 proGenomes3 were obtained, and the same process was applied to infer clustering results.

760

761 **Analysis of species-level pangenome and single nucleotide variants (SNVs)**

762 For species-level pangenome analysis, we conducted Panaroo (v1.3.0)[91] with the options of "-c
763 0.90 --core_threshold 0.90 -f 0.5 --clean-mode strict --merge_paralogs", applied to GFF files
764 produced by Prokka for NC conspecific genomes per species. From the pangenome gene
765 presence/absence matrix, we generated the rarefaction curves by randomly removing genomes.
766 The resulting curve was fitted to the equation $\boldsymbol{y \sim \log(x)}$, where x is the number of genome and y
767 is the size of the pangenome. For the analysis of single nucleotide variants (SNVs) in HROM, we
768 utilized scripts and pipeline provided in previous research[6] (available at
769 https://github.com/zjshi/snv_analysis_almeida2019). For species with more than five conspecific
770 genomes, each non-representative conspecific genome (g) was aligned against the representative
771 genome (r) of the species cluster using nucmer from MUMmer4[88]. Best bi-directional alignments
772 were identified using the delta-filter program with options "-q –r". SNVs were annotated with the
773 show-snp program of MUMmer4. For each species with a representative genome (r), the SNV
774 density was calculated by normalizing the total number of SNVs by the aligned genome length
775 and the number of conspecific genomes (G) as follows:

776
$$\boldsymbol{SNV\ per\ kb} = \frac{\sum_{g \in (G-\{r\})} \frac{\#SNV_{r,g}}{\frac{Aligned\ length_{r,g}}{1000}}}{n(G)-1}$$

777 For downstream analyses and SNV density calculations, only species with more than 10 NC
778 conspecific genomes were included.

779

780 **Identifying biosynthetic gene clusters (BGCs) and metabolic gene clusters (MGCs)**

781 Biosynthetic and metabolic gene clusters were identified for all NC genomes in HROM. MGCs
782 were detected using gutSMASH (v1.0.0)[100],while secondary metabolite BGCs were identified
783 using antiSMASH (v6.1.1)[101].

784

785 **Profiling division and cell wall (dcw) gene clusters**

786 To profile division and cell wall (dcw) gene clusters, we obtained 17 dcw gene cluster HMM
787 models from the original research dataset (available at https://doi.org/10.17632/4y5mzppzmb.1).
788 The HMM models were used with hmmsearch (e-value threshold = 1e-5) to identify hits for each
789 marker within the genomes.

790

791 **Profiling genomes for antibiotic resistance and viral defense systems**

792 Antibiotic resistance genes in the 72,641 NC genomes of HROM were identified using the "RGI
793 main" function (with the "strict" option) of RGI (v6.0.3)[102]. Viral defense systems were inferred
794 using DefenseFinder (v1.0.9)[103].

795

796 **Identifying anti-microbial peptides (AMPs)**

797 To identify antimicrobial peptides (AMPs), amino acid sequence of small open reading frame
798 (smORF) were inferred from all NC genomes in HROM using Prodigal (v2.3.0) with the option "-
799 -min-gene = 33"[104]. The smORFs were filtered to include sequences raging between 15 and 100
800 amino acids. To remove spurious smORFs, we employed the AntiFam[105] HMM model and
801 performed sequence filtering using hmmsearch. Redundant smORFs were then removed to
802 generate a non-redundant dataset. The non-redundant smORFs were subject to AMP prediction
803 using Macrel (v1.2.0)[39]. Singleton smORFs identified as AMP were excluded. To estimate the
804 proportion of novel AMPs among those derived from HROM genomes, we obtained reference

805 AMPs from the database of Antimicrobial Peptides (DRAMP v3.0)[18], ensuring that the sequences
806 were correctly labeled with amino acids. AMPs derived from HROM were clustered with AMPs
807 retrieved from DRAMP at the family-level using cd-hit of CD-HIT (v4.6.8)[106] with the options "-
808 c 0.75 -M 0 -d 0" after converting the amino acids sequences into a reduced alphabet of eight
809 letters[107]. The AMP density for each genome was defined as number of AMPs identified divided
810 by the total genome length.

811

812 **Patescibacteria genomes from different isolation sources**

813 We collected genomes annotated as Patescibacteria from GenBank. Additionally, we collected
814 those isolated from groundwater ecosystems[54]. These genomes were classified using GTDB-Tk[23]
815 and only those confirmed to belong to the phylum Patescibacteria were retained. The genomes
816 were subjected to the same quality filtering criteria used for HROM to ensure only non-chimeric,
817 NC genomes were included. Metadata were manually inspected to determine the isolation source
818 of each genome. Genomes were annotated as MHA (mammalian host-associated) if isolated from
819 animal or human associated environments; otherwise, they were categorized as "environment".
820 Using the HROM representative genomes classified as Patescibacteria, we applied the same
821 species-level de-replication pipeline as HROM to infer species-level genome bins within
822 Patescibacteria. For protein family analysis, protein sequences were predicted from the
823 representative genome of each species using Prokka[31]. These sequences were then clustered and
824 analyzed using the same methodology described earlier to infer the protein family of
825 Patescibacteria.

826

827 **Inference of protein family from Patescibacteria genomes**

828 Protein families were inferred using a pipeline and script from a previous study (available at
829 https://github.com/raphael-upmc/proteinClusteringPipeline)[47]. Briefly, an all-vs-all similarity
830 search was conducted using MMseqs2[32] with the following parameters: e-value = 0.001, sensitivity
831 = 7.5, and coverage = 0.5. The input consisted of protein sequences derived from the representative
832 genome of 2,963 Patescibacteria genomes. A similarity-based network was constructed and used
833 as input for clustering protein sequences into subfamilies using the greedy set cover algorithm of

834 MMseqs2. Subfamilies containing more than two protein members were selected for downstream
835 analysis. These selected subfamilies were aligned using the "mmseqs result2msa" module of
836 MMseqs2, and multiple sequence alignments were used to construct HMM profiles with the
837 HHpred suite[108,109]. Pairwise comparisons of the HMM profiles were performed using hhblits from
838 the HHpred suite with the following parameters: -v 0 -p 50 -z 4 -Z 32000 -B 0 -b 0. Subfamilies
839 with probability scores ≥95% and coverage ≥0.50 were integrated into the input network, with
840 similarity scores serving as edge weights. Protein families were defined by clustering the network
841 using the Markov Clustering (MCL) algorithm with an inflation parameter of 2.0.

842

843 **Functional enrichment analysis and metabolic comparison of Patescibacteria genomes**

844 For functional term enrichment and metabolic comparison, we profiled each representative
845 genomes of 2,963 Patescibacteria species using the "anvi-run-ncbi-cogs" and "anvi-run-kegg-
846 kofams" modules of Anvi'o. Additional functional terms were identified using MacSyFinder
847 (v2)[110] and InterProScan (v5.60-92.0)[111]. The completeness of KEGG metabolism modules was
848 assessed using "anvi-estimate-metabolism" module, with modules considered present if their
849 completeness was equal to or greater than the default threshold of 0.75. Terms and proteins were
850 defined as prevalent if their prevalence among MHA genomes was ≥ 50% and their prevalence
851 among environmental genomes was ≤ 20%.

852

853 **Analyses of the periodontitis-associated Patescibacteria**

854 For species clustering based on protein profiles, we used the "makeSNNGraph" function from the
855 R package bluster (v1.9.1)[112] to construct a nearest-neighbor graph. Louvain clustering was
856 performed using the "cluster_louvain" function from the R package igraph (v1.3.1)[113]. Four public
857 datasets, comprising both healthy and periodontitis samples, were included in the analysis.
858 Differentially enriched Patescibacteria clusters were identified using MaAsLin2 (v1,8)[114], with the
859 batch variable treated as a random effect. Batch correction for the saliva periodontitis datasets was
860 performed using MMUPHin (v1.16)[115].

861 To identify protein families enriched in periodontitis-associated Patescibacteria cluster 16, we
862 selected families with a prevalence lower than 5% in other clusters but higher than 50% in cluster

863 16. WMS reads were realigned to the selected protein families using DIAMOND (v2.1.9)[96] with
864 threshold of 80% coverage and 80% identity. Any aligned reads to these proteins were considered
865 hit. The relative abundance of protein families was estimated by dividing the number of hits by the
866 total number of reads.

867 For protein structure prediction, we utilized ESMFold (v1.0.3)[60], followed by a protein structure
868 search using Foldseek (v9.427df8a)[62] against the Alphafold UniProt50[61] database, employing the
869 "easy-search" mode. Hits were retrieved if they met the following criteria: an e-value less than 10-
870 4, a TM score ≥ 0.3, and query coverage ≥ 0.5.

871

872 **Identification of oral-gut shared species**

873 Species-level de-replication was performed for representative genomes from HROM (oral) and
874 HRGM2 (gut) using the previously described pipeline. In brief, all representative genomes were
875 taxonomically classified with GTDB-Tk[23]. For genomes within the same order, primary clusters
876 were defined using the MASH tool from the dRep compare module, applying the same parameters
877 as those used in HROM species definition. Within each primary cluster, ANI was calculated, and
878 hierarchical clustering was conducted using the same parameter to achieve species-level de-
879 replication. Genome bins that included at least one genome from HRGM2 and one from HROM
880 were identified as oral-gut shared species.

881

882 **Estimation of sporulation capacity**

883 To quantify the sporulation score, we followed a scheme similar to that used in a previous study[70].
884 From the 66 known sporulation-predictive genes, we successfully retrieved protein sequences for
885 65 of them. The presence of 65 sporulation-predictive genes in each genome was measured using
886 tblastn from BLAST (v2.12.0+)[116] with the parameters e-value = 1e-5 and 30% identity. The
887 sporulation score was calculated as the number of sporulation-predictive genes detected in the
888 genome divided by 65. For each species, the highest sporulation score observed among its
889 conspecific genomes was used as the representative sporulation score.

890

891 **Estimation of metabolic independence**

892 To estimate metabolic independence, we followed a scheme similar to that described in a previous
893 study[69]. The metabolic independence score was calculated as the sum of the completeness scores
894 of 33 KEGG modules associated with metabolic independence. KEGG modules were identified
895 using anvi-run-kegg-kofams function from Anvi'o (v7.1)[117]. The completeness of each KEGG
896 metabolism module was then quantified using the anvi-estimate-metabolism module of Anvi'o,
897 with a default threshold of 0.75 applied to confirm module presence. The metabolic independence
898 score was defined as total completeness scores of the 33 KEGG modules divided by 33. For each
899 species, the representative metabolic independence score was calculated as the mean score of all
900 its conspecific genomes.

901

902 **Identification of oral-gut transmitter species**

903 To identify gut-oral transmitter species, which are presumed to be prevalent in the healthy human
904 oral cavity but not in the healthy gut microbiota, we analyzed 15 healthy oral microbiome datasets
905 (**Supplementary Table 15**) and 9 healthy gut microbiome datasets (**Supplementary Table 16**),
906 each containing at least 30 samples. For oral microbiome datasets, if a single dataset included
907 samples from more than two isolation sources, each subset was treated as a separate dataset. This
908 resulted in the collection of 3,377 healthy oral samples and 956 healthy gut samples for analysis.
909 For the healthy oral microbiome datasets, species-level classification of WMS reads was
910 conducted using the HROM custom database based on concatenated genomes. Similarly, for
911 healthy gut microbiome datasets, species-level classification was performed using the HRGM2
912 custom database. From these species-level profiles, we identified 330 oral-gut shared species and
913 measured their prevalence using a relative abundance threshold of 1e-6 to determine species
914 presence. For cases where a single HROM species was clustered with multiple HRGM genomes,
915 the highest prevalence value of the HRGM2 species was selected. Conversely, when multiple
916 HROM species were clustered with a single HRGM2 genome, the prevalence of the HRGM2
917 genome was assigned to all clustered species. The mean prevalence was calculated across all
918 datasets for each species within oral and gut samples. Thresholds of 50% prevalence in healthy
919 oral microbiomes and 25% prevalence in healthy gut microbiomes were applied to define the oral-
920 gut transmitter species.

921

922 **Identifying important bacterial species for colorectal cancer (CRC) classification**

923 We analyzed 12 cohorts of human gut microbiome samples from colorectal cancer (CRC) patients
924 and healthy controls (**Supplementary Tables 18–19**). These datasets were processed using a
925 workflow similar to the preprocessing of WMS data for HROM. Taxonomic classification was
926 performed with Kraken2 using the HRGM2 custom genome database. Species-level counts were
927 estimated using Bracken, and reads were normalized by multiplying by $10^6 / l_{species}$, where $l_{species}$
928 represents the mean genome length of the species. Relative abundance was calculated by dividing
929 the normalized abundance by their total sum. From the multi-cohort profile, we identified
930 differentially abundant features associated with CRC using the "check.associations" function
931 (option = "lm", formula = "feat ~ cohort + disease") of SIAMCAT (v2.6.0)[118], applying an FDR-
932 adjusted *p*-value ≤ 0.05. Additionally, we employed ANCOM-BC[119] in individual cohorts to
933 identify CRC-enriched taxa with the same FDR-adjusted *p*-value threshold. The variance
934 explained by cohort and disease labels was assessed using "check.confounders" function of
935 SIAMCAT. For CRC classification, we trained a random forest model on gut species profiles from
936 nine datasets using SIAMCAT functions "create.data.split" and "train.model" (option = "method
937 = randomForest"). CRC prediction was then conducted on the remaining three datasets not used
938 for training, using the SIAMCAT functions "make.predictions" and "evaluate.predictions". To
939 identify the key features of the model, we used the "feature_weights" function to extract the median
940 relative feature weights. The top 20 species with the highest median relative feature weights were
941 designated as the most important features. To construct microbial networks, we employed SPIEC-
942 EASI (v1.1.0)[120]. Specifically, we focused on 17 oral bacterial species among the top 20 important
943 species for CRC classification.

944

945 ## Data availability

946 By accessing the web server, www.decodebiome.org/HROM/ users can browse and download
947 sequences and annotations for oral microbial genomes, protein families, 16S regions, a custom
948 Kraken2 database, and a database of species-specific markers.

949

950 ## Acknowledgements

This work was supported by the National Research Foundation (NRF) funded by the Ministry of Science and ICT, Republic of Korea (2022M3A9F3016364, 2022R1A2C1092062), Ministry of Trade, Industry and Energy (200022947), Brain Korea 21 (BK21) FOUR program.

Author contributions

J.H.C. and I.L. conceived and designed the study. J.H.C. performed the construction of genome catalog and formulated the study hypothesis. N.K., J.M. co-developed bioinformatics pipelines. S.L., G.K. assisted data collection and compilation. Y.S. assisted building the data server. S.B., I.B., and B.L. assisted data analysis. I.L. supervised the project. J.H.C. and I.L. wrote and edited the manuscript.

Declaration of interest

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve language and readability, with caution. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Baker, J.L., Mark Welch, J.L., Kauffman, K.M., McLean, J.S., and He, X. (2024). The oral microbiome: diversity, biogeography and human health. Nat Rev Microbiol *22*, 89-104. 10.1038/s41579-023-00963-6.

2. Lamont, R.J., Koo, H., and Hajishengallis, G. (2018). The oral microbiota: dynamic communities and host interactions. Nat Rev Microbiol *16*, 745-759. 10.1038/s41579-018-0089-x.

977 3. Peng, X., Cheng, L., You, Y., Tang, C., Ren, B., Li, Y., Xu, X., and Zhou, X. (2022). Oral
978 microbiota in human systematic diseases. Int J Oral Sci *14*, 14. 10.1038/s41368-022-
979 00163-7.

980 4. Human Microbiome Project, C. (2012). Structure, function and diversity of the healthy
981 human microbiome. Nature *486*, 207-214. 10.1038/nature11234.

982 5. Kim, N., Ma, J., Kim, W., Kim, J., Belenky, P., and Lee, I. (2024). Genome-resolved
983 metagenomics: a game changer for microbiome medicine. Exp Mol Med *56*, 1501-1512.
984 10.1038/s12276-024-01262-7.

985 6. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard,
986 K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of
987 204,938 reference genomes from the human gut microbiome. Nat Biotechnol *39*, 105-
988 114. 10.1038/s41587-020-0603-3.

989 7. Kim, C.Y., Lee, M., Yang, S., Kim, K., Yong, D., Kim, H.R., and Lee, I. (2021). Human
990 reference gut microbiome catalog including newly assembled genomes from under-
991 represented Asian metagenomes. Genome Med *13*, 134. 10.1186/s13073-021-00950-7.

992 8. Hiseni, P., Rudi, K., Wilson, R.C., Hegge, F.T., and Snipen, L. (2021). HumGut: a
993 comprehensive human gut prokaryotic genomes collection filtered by metagenome data.
994 Microbiome *9*, 165. 10.1186/s40168-021-01114-w.

995 9. Zhu, J., Tian, L., Chen, P., Han, M., Song, L., Tong, X., Sun, X., Yang, F., Lin, Z., Liu,
996 X., et al. (2022). Over 50,000 Metagenomically Assembled Draft Genomes for the
997 Human Oral Microbiome Reveal New Taxa. Genomics Proteomics Bioinformatics *20*,
998 246-259. 10.1016/j.gpb.2021.05.001.

999 10. Escapa, I.F., Chen, T., Huang, Y., Gajare, P., Dewhirst, F.E., and Lemon, K.P. (2018).
1000 New Insights into Human Nostril Microbiome from the Expanded Human Oral
1001 Microbiome Database (eHOMD): a Resource for the Microbiome of the Human
1002 Aerodigestive Tract. mSystems *3*. 10.1128/mSystems.00187-18.

1003 11. Castelle, C.J., Brown, C.T., Anantharaman, K., Probst, A.J., Huang, R.H., and Banfield,
1004 J.F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and
1005 DPANN radiations. Nat Rev Microbiol *16*, 629-645. 10.1038/s41579-018-0076-2.

1006 12. Yuan, D., Ahamed, A., Burgin, J., Cummins, C., Devraj, R., Gueye, K., Gupta, D., Gupta,
1007 V., Haseeb, M., Ihsan, M., et al. (2024). The European Nucleotide Archive in 2023.
1008 Nucleic Acids Res *52*, D92-D97. 10.1093/nar/gkad1067.

1009 13. Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Brister, J.R., and O'Sullivan, C. (2022).
1010 The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res *50*,
1011 D387-D390. 10.1093/nar/gkab1053.

1012 14. Chen, F.Z., You, L.J., Yang, F., Wang, L.N., Guo, X.Q., Gao, F., Hua, C., Tan, C., Fang,
1013 L., Shan, R.Q., et al. (2020). CNGBdb: China National GeneBank DataBase. Yi Chuan
1014 *42*, 799-809. 10.16288/j.yczz.20-080.

1015 15. Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Schoch, C.L., Sherry, S.T., and
1016 Karsch-Mizrachi, I. (2021). GenBank. Nucleic Acids Res *49*, D92-D96.
1017 10.1093/nar/gkaa1023.

1018 16. Li, W., O'Neill, K.R., Haft, D.H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris,
1019 G., Chitsaz, F., Derbyshire, M.K., Durkin, A.S., et al. (2021). RefSeq: expanding the
1020 Prokaryotic Genome Annotation Pipeline reach with protein family model curation.
1021 Nucleic Acids Res *49*, D1020-D1028. 10.1093/nar/gkaa1105.

1022 17. Davis, J.J., Wattam, A.R., Aziz, R.K., Brettin, T., Butler, R., Butler, R.M., Chlenski, P.,
1023 Conrad, N., Dickerman, A., Dietrich, E.M., et al. (2020). The PATRIC Bioinformatics
1024 Resource Center: expanding data and analysis capabilities. Nucleic Acids Res *48*, D606-
1025 D612. 10.1093/nar/gkz943.

1026 18. Shi, G., Kang, X., Dong, F., Liu, Y., Zhu, N., Hu, Y., Xu, H., Lao, X., and Zheng, H.
1027 (2022). DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial
1028 peptides. Nucleic Acids Res *50*, D488-D496. 10.1093/nar/gkab651.

1029 19. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015).
1030 CheckM: assessing the quality of microbial genomes recovered from isolates, single
1031 cells, and metagenomes. Genome Res *25*, 1043-1055. 10.1101/gr.186072.114.

1032 20. Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins,
1033 M.J., Wrighton, K.C., Williams, K.H., and Banfield, J.F. (2015). Unusual biology across a
1034 group comprising more than 15% of domain Bacteria. Nature *523*, 208-211.
1035 10.1038/nature14486.

1036 21. Chklovski, A., Parks, D.H., Woodcroft, B.J., and Tyson, G.W. (2023). CheckM2: a rapid,
1037 scalable and accurate tool for assessing microbial genome quality using machine
1038 learning. Nat Methods *20*, 1203-1212. 10.1038/s41592-023-01940-w.

1039 22. Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt,
1040 T.S.B., and Bork, P. (2021). GUNC: detection of chimerism and contamination in
1041 prokaryotic genomes. Genome Biol *22*, 178. 10.1186/s13059-021-02393-0.

1042 23. Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2022). GTDB-Tk v2:
1043 memory friendly classification with the genome taxonomy database. Bioinformatics *38*,
1044 5315-5316. 10.1093/bioinformatics/btac672.

1045 24. Li, W., Liang, H., Lin, X., Hu, T., Wu, Z., He, W., Wang, M., Zhang, J., Jie, Z., Jin, X., et
1046 al. (2023). A catalog of bacterial reference genomes from cultivated human oral bacteria.
1047 NPJ Biofilms Microbiomes *9*, 45. 10.1038/s41522-023-00414-3.

1048 25. Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M.L., Burdett, T., Burgin,
1049 J., Caballero-Perez, J., Cochrane, G., Colwell, L.J., et al. (2023). MGnify: the
1050 microbiome sequence data analysis resource in 2023. Nucleic Acids Res *51*, D753-D759.
1051 10.1093/nar/gkac1080.

1052 26. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with
1053 Kraken 2. Genome Biol *20*, 257. 10.1186/s13059-019-1891-0.

1054 27. Blanco-Miguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M.,
1055 Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., et al. (2023). Extending and
1056 improving metagenomic taxonomic profiling with uncharacterized species using
1057 MetaPhlAn 4. Nat Biotechnol *41*, 1633-1644. 10.1038/s41587-023-01688-w.

1058 28. Lu, J., Breitwieser, F.P., Thielen, P., and Salzberg, S.L. (2017). Bracken: estimating
1059 species abundance in metagenomics data. PeerJ Computer Science *3*, e104.

1060 29. Sun, Z., Huang, S., Zhang, M., Zhu, Q., Haiminen, N., Carrieri, A.P., Vazquez-Baeza, Y.,
1061 Parida, L., Kim, H.C., Knight, R., and Liu, Y.Y. (2021). Challenges in benchmarking
1062 metagenomic profilers. Nat Methods *18*, 618-626. 10.1038/s41592-021-01141-3.

1063 30. Croll, D., and McDonald, B.A. (2012). The accessory genome as a cradle for adaptive
1064 evolution in pathogens. PLoS Pathog *8*, e1002608. 10.1371/journal.ppat.1002608.

1065 31. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*,
1066 2068-2069. 10.1093/bioinformatics/btu153.

1067 32. Steinegger, M., and Soding, J. (2017). MMseqs2 enables sensitive protein sequence
1068 searching for the analysis of massive data sets. Nat Biotechnol *35*, 1026-1028.
1069 10.1038/nbt.3988.

1070 33. Fullam, A., Letunic, I., Schmidt, T.S.B., Ducarmon, Q.R., Karcher, N., Khedkar, S.,
1071 Kuhn, M., Larralde, M., Maistrenko, O.M., Malfertheiner, L., et al. (2023).
1072 proGenomes3: approaching one million accurately and consistently annotated high-
1073 quality prokaryotic genomes. Nucleic Acids Res *51*, D760-D766. 10.1093/nar/gkac1078.

1074 34. Cantalapiedra, C.P., Hernandez-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J.
1075 (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and
1076 Domain Prediction at the Metagenomic Scale. Mol Biol Evol *38*, 5825-5829.
1077 10.1093/molbev/msab293.

1078 35. Galperin, M.Y., Wolf, Y.I., Makarova, K.S., Vera Alvarez, R., Landsman, D., and Koonin,
1079 E.V. (2021). COG database update: focus on microbial diversity, model organisms, and
1080 widespread pathogens. Nucleic Acids Res *49*, D274-D281. 10.1093/nar/gkaa1018.

1081 36. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M.
1082 (2023). KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids
1083 Res *51*, D587-D592. 10.1093/nar/gkac963.

1084 37. Kulmanov, M., and Hoehndorf, R. (2021). DeepGOPlus: improved protein function
1085 prediction from sequence. Bioinformatics *37*, 1187. 10.1093/bioinformatics/btaa763.

1086 38. Santos-Junior, C.D., Torres, M.D.T., Duan, Y., Rodriguez Del Rio, A., Schmidt, T.S.B.,
1087 Chong, H., Fullam, A., Kuhn, M., Zhu, C., Houseman, A., et al. (2024). Discovery of
1088 antimicrobial peptides in the global microbiome with machine learning. Cell *187*, 3761-
1089 3778 e3716. 10.1016/j.cell.2024.05.013.

1090 39. Santos-Junior, C.D., Pan, S., Zhao, X.M., and Coelho, L.P. (2020). Macrel: antimicrobial
1091 peptide screening in genomes and metagenomes. PeerJ *8*, e10555. 10.7717/peerj.10555.

1092 40. Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J.,
1093 Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016). A new view of the
1094 tree of life. Nat Microbiol *1*, 16048. 10.1038/nmicrobiol.2016.48.

1095 41. Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N.,
1096 Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-
1097 assembled genomes substantially expands the tree of life. Nat Microbiol *2*, 1533-1542.
1098 10.1038/s41564-017-0012-7.

1099 42. Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A.,
1100 and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome
1101 phylogeny substantially revises the tree of life. Nat Biotechnol *36*, 996-1004.
1102 10.1038/nbt.4229.

1103 43. Bor, B., Poweleit, N., Bois, J.S., Cen, L., Bedree, J.K., Zhou, Z.H., Gunsalus, R.P., Lux,
1104 R., McLean, J.S., He, X., and Shi, W. (2016). Phenotypic and Physiological
1105 Characterization of the Epibiotic Interaction Between TM7x and Its Basibiont
1106 Actinomyces. Microb Ecol *71*, 243-255. 10.1007/s00248-015-0711-7.

1107 44. Naud, S., Ibrahim, A., Valles, C., Maatouk, M., Bittar, F., Tidjani Alou, M., and Raoult,
1108 D. (2022). Candidate Phyla Radiation, an Underappreciated Division of the Human
1109 Microbiome, and Its Impact on Health and Disease. Clin Microbiol Rev *35*, e0014021.
1110 10.1128/cmr.00140-21.

1111 45. Megrian, D., Taib, N., Jaffe, A.L., Banfield, J.F., and Gribaldo, S. (2022). Ancient origin
1112 and constrained evolution of the division and cell wall gene cluster in Bacteria. Nat
1113 Microbiol *7*, 2114-2127. 10.1038/s41564-022-01257-y.

1114 46. Bor, B., Bedree, J.K., Shi, W., McLean, J.S., and He, X. (2019). Saccharibacteria (TM7)
1115 in the Human Oral Microbiome. J Dent Res *98*, 500-509. 10.1177/0022034519831671.

1116 47. Meheust, R., Burstein, D., Castelle, C.J., and Banfield, J.F. (2019). The distinction of
1117 CPR bacteria from other bacteria based on protein family content. Nat Commun *10*,
1118 4173. 10.1038/s41467-019-12171-z.

1119 48. Fong, J.C.N., Syed, K.A., Klose, K.E., and Yildiz, F.H. (2010). Role of Vibrio
1120 polysaccharide (vps) genes in VPS production, biofilm formation and Vibrio cholerae
1121 pathogenesis. Microbiology (Reading) *156*, 2757-2769. 10.1099/mic.0.040196-0.

1122 49. Ealand, C., Rimal, B., Chang, J., Mashigo, L., Chengalroyen, M., Mapela, L., Beukes, G.,
1123 Machowski, E., Kim, S.J., and Kana, B. (2018). Resuscitation-Promoting Factors Are
1124 Required for Mycobacterium smegmatis Biofilm Formation. Appl Environ Microbiol *84*.
1125 10.1128/AEM.00687-18.

1126 50. Ruggiero, A., Tizzano, B., Pedone, E., Pedone, C., Wilmanns, M., and Berisio, R. (2009).
1127 Crystal structure of the resuscitation-promoting factor (DeltaDUF)RpfB from M.
1128 tuberculosis. J Mol Biol *385*, 153-162. 10.1016/j.jmb.2008.10.042.

1129 51. Maldarelli, G.A., Piepenbrink, K.H., Scott, A.J., Freiberg, J.A., Song, Y., Achermann, Y.,
1130 Ernst, R.K., Shirtliff, M.E., Sundberg, E.J., Donnenberg, M.S., and von Rosenvinge, E.C.
1131 (2016). Type IV pili promote early biofilm formation by Clostridium difficile. Pathog Dis
1132 *74*. 10.1093/femspd/ftw061.

1133 52. Bahar, O., Goffer, T., and Burdman, S. (2009). Type IV Pili are required for virulence,
1134 twitching motility, and biofilm formation of acidovorax avenae subsp. Citrulli. Mol Plant
1135 Microbe Interact *22*, 909-920. 10.1094/MPMI-22-8-0909.

1136 53. Barken, K.B., Pamp, S.J., Yang, L., Gjermansen, M., Bertrand, J.J., Klausen, M.,
1137 Givskov, M., Whitchurch, C.B., Engel, J.N., and Tolker-Nielsen, T. (2008). Roles of type
1138 IV pili, flagellum-mediated motility and extracellular DNA in the formation of mature
1139 multicellular structures in Pseudomonas aeruginosa biofilms. Environ Microbiol *10*,
1140 2331-2343. 10.1111/j.1462-2920.2008.01658.x.

1141 54. He, C., Keren, R., Whittaker, M.L., Farag, I.F., Doudna, J.A., Cate, J.H.D., and Banfield,
1142 J.F. (2021). Genome-resolved metagenomics reveals site-specific diversity of
1143 episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. Nat
1144 Microbiol *6*, 354-365. 10.1038/s41564-020-00840-5.

1145 55. McLean, J.S., Bor, B., Kerns, K.A., Liu, Q., To, T.T., Solden, L., Hendrickson, E.L.,
1146 Wrighton, K., Shi, W., and He, X. (2020). Acquisition and Adaptation of Ultra-small
1147 Parasitic Reduced Genome Bacteria to Mammalian Hosts. Cell Rep *32*, 107939.
1148 10.1016/j.celrep.2020.107939.

1149 56. Noens, E.E., and Lolkema, J.S. (2017). Convergent evolution of the arginine deiminase
1150 pathway: the ArcD and ArcE arginine/ornithine exchangers. Microbiologyopen *6*.
1151 10.1002/mbo3.412.

1152 57. Wang, Y., Gallagher, L.A., Andrade, P.A., Liu, A., Humphreys, I.R., Turkarslan, S.,
1153 Cutler, K.J., Arrieta-Ortiz, M.L., Li, Y., Radey, M.C., et al. (2023). Genetic manipulation
1154 of Patescibacteria provides mechanistic insights into microbial dark matter and the
1155 epibiotic lifestyle. Cell *186*, 4803-4817 e4813. 10.1016/j.cell.2023.08.017.

1156 58. Xie, B., Wang, J., Nie, Y., Tian, J., Wang, Z., Chen, D., Hu, B., Wu, X.L., and Du, W.
1157 (2022). Type IV pili trigger episymbiotic association of Saccharibacteria with its bacterial
1158 host. Proc Natl Acad Sci U S A *119*, e2215990119. 10.1073/pnas.2215990119.

1159 59. Mell, J.C., and Redfield, R.J. (2014). Natural competence and the evolution of DNA
1160 uptake specificity. J Bacteriol *196*, 1471-1483. 10.1128/JB.01293-13.

1161 60. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O.,
1162 Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure
1163 with a language model. Science *379*, 1123-1130. 10.1126/science.ade2574.

1164 61. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D.,
1165 Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database:
1166 massively expanding the structural coverage of protein-sequence space with high-
1167 accuracy models. Nucleic Acids Res *50*, D439-D444. 10.1093/nar/gkab1061.

1168 62. Barrio-Hernandez, I., Yeo, J., Janes, J., Mirdita, M., Gilchrist, C.L.M., Wein, T., Varadi,
1169 M., Velankar, S., Beltrao, P., and Steinegger, M. (2023). Clustering predicted structures at
1170 the scale of the known protein universe. Nature *622*, 637-645. 10.1038/s41586-023-
1171 06510-w.

1172 63. Bycroft, M., Bateman, A., Clarke, J., Hamill, S.J., Sandford, R., Thomas, R.L., and
1173 Chothia, C. (1999). The structure of a PKD domain from polycystin-1: implications for
1174 polycystic kidney disease. EMBO J *18*, 297-305. 10.1093/emboj/18.2.297.

1175 64. Chen, Y.Y., Cross, K.J., Paolini, R.A., Fielding, J.E., Slakeski, N., and Reynolds, E.C.
1176 (2002). CPG70 is a novel basic metallocarboxypeptidase with C-terminal polycystic
1177 kidney disease domains from Porphyromonas gingivalis. J Biol Chem *277*, 23433-23440.
1178 10.1074/jbc.M200811200.

1179 65. Kunath, B.J., De Rudder, C., Laczny, C.C., Letellier, E., and Wilmes, P. (2024). The oral-
1180 gut microbiome axis in health and disease. Nat Rev Microbiol *22*, 791-805.
1181 10.1038/s41579-024-01075-5.

1182 66. Schmidt, T.S., Hayward, M.R., Coelho, L.P., Li, S.S., Costea, P.I., Voigt, A.Y., Wirbel, J.,
1183 Maistrenko, O.M., Alves, R.J., Bergsten, E., et al. (2019). Extensive transmission of
1184 microbes along the gastrointestinal tract. Elife *8*. 10.7554/eLife.42693.

1185 67. Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human
1186 and Bacteria Cells in the Body. PLoS Biol *14*, e1002533. 10.1371/journal.pbio.1002533.

1187 68. Olsen, I., and Yamazaki, K. (2019). Can oral bacteria affect the microbiome of the gut? J
1188 Oral Microbiol *11*, 1586422. 10.1080/20002297.2019.1586422.

1189 69. Watson, A.R., Fussel, J., Veseli, I., DeLongchamp, J.Z., Silva, M., Trigodet, F., Lolans,
1190 K., Shaiber, A., Fogarty, E., Runde, J.M., et al. (2023). Metabolic independence drives
1191 gut microbial colonization and resilience in health and disease. Genome Biol *24*, 78.
1192 10.1186/s13059-023-02924-x.

1193 70. Browne, H.P., Almeida, A., Kumar, N., Vervier, K., Adoum, A.T., Viciani, E., Dawson,
1194 N.J.R., Forster, S.C., Cormie, C., Goulding, D., and Lawley, T.D. (2021). Host adaptation
1195 in gut Firmicutes is associated with sporulation loss and altered transmission cycle.
1196 Genome Biol *22*, 204. 10.1186/s13059-021-02428-6.

1197 71. Wang, N., and Fang, J.Y. (2023). Fusobacterium nucleatum, a key pathogenic factor and
1198 microbial biomarker for colorectal cancer. Trends Microbiol *31*, 159-172.
1199 10.1016/j.tim.2022.08.010.

1200 72. Liao, C., Rolling, T., Djukovic, A., Fei, T., Mishra, V., Liu, H., Lindberg, C., Dai, L.,
1201 Zhai, B., Peled, J.U., et al. (2024). Oral bacteria relative abundance in faeces increases
1202 due to gut microbiota depletion and is linked with patient outcomes. Nat Microbiol *9*,
1203 1555-1565. 10.1038/s41564-024-01680-3.

1204 73. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for
1205 Illumina sequence data. Bioinformatics *30*, 2114-2120. 10.1093/bioinformatics/btu170.

1206 74. Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.C., Kitts, P.A.,
1207 Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of
1208 GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of
1209 the reference assembly. Genome Res *27*, 849-864. 10.1101/gr.213611.116.

1210 75. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat
1211 Methods *9*, 357-359. 10.1038/nmeth.1923.

1212 76. Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new
1213 versatile metagenomic assembler. Genome Res *27*, 824-834. 10.1101/gr.213959.116.

1214 77. Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. (2015). MEGAHIT: an ultra-fast
1215 single-node solution for large and complex metagenomics assembly via succinct de
1216 Bruijn graph. Bioinformatics *31*, 1674-1676. 10.1093/bioinformatics/btv033.

1217 78. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019).
1218 MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
1219 reconstruction from metagenome assemblies. PeerJ *7*, e7359. 10.7717/peerj.7359.

1220 79. Wu, Y.W., Simmons, B.A., and Singer, S.W. (2016). MaxBin 2.0: an automated binning
1221 algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics *32*,
1222 605-607. 10.1093/bioinformatics/btv638.

1223 80. Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L.,
1224 Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by
1225 coverage and composition. Nat Methods *11*, 1144-1146. 10.1038/nmeth.3103.

1226 81. Uritskiy, G.V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP-a flexible pipeline for
1227 genome-resolved metagenomic data analysis. Microbiome *6*, 158. 10.1186/s40168-018-
1228 0541-1.

1229 82. Seemann, T. (2018). barrnap 0.9: rapid ribosomal RNA prediction.
1230 GitHub:https://github.com/tseemann/barrnap.

1231 83. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.A., and Hugenholtz,
1232 P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a
1233 phylogenetically consistent, rank normalized and complete genome-based taxonomy.
1234 Nucleic Acids Res *50*, D785-D794. 10.1093/nar/gkab776.

1235 84. Chan, P.P., Lin, B.Y., Mak, A.J., and Lowe, T.M. (2021). tRNAscan-SE 2.0: improved
1236 detection and functional classification of transfer RNA genes. Nucleic Acids Res *49*,
1237 9077-9096. 10.1093/nar/gkab688.

1238 85. Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy,
1239 T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al. (2017). Minimum
1240 information about a single amplified genome (MISAG) and a metagenome-assembled
1241 genome (MIMAG) of bacteria and archaea. Nat Biotechnol *35*, 725-731.
1242 10.1038/nbt.3893.

1243 86. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and
1244 Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using
1245 MinHash. Genome Biol *17*, 132. 10.1186/s13059-016-0997-x.

1246 87. Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and
1247 accurate genomic comparisons that enables improved genome recovery from
1248 metagenomes through de-replication. ISME J *11*, 2864-2868. 10.1038/ismej.2017.126.

1249 88. Marcais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A.
1250 (2018). MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol
1251 *14*, e1005944. 10.1371/journal.pcbi.1005944.

1252 89. Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R
1253 and Python. Journal of Statistical Software *53*, 1-18.

1254 90. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von
1255 Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for
1256 Phylogenetic Inference in the Genomic Era. Mol Biol Evol *37*, 1530-1534.
1257 10.1093/molbev/msaa015.

1258 91. Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J.A.,
1259 Gladstone, R.A., Lo, S., Beaudoin, C., Floto, R.A., et al. (2020). Producing polished
1260 prokaryotic pangenomes with the Panaroo pipeline. Genome Biol *21*, 180.
1261 10.1186/s13059-020-02090-4.

1262 92. Fritz, A., Hofmann, P., Majda, S., Dahms, E., Droge, J., Fiedler, J., Lesker, T.R.,
1263 Belmann, P., DeMaere, M.Z., Darling, A.E., et al. (2019). CAMISIM: simulating
1264 metagenomes and microbial communities. Microbiome *7*, 17. 10.1186/s40168-019-0633-
1265 6.

93. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research *12*, 2825-2830. 10.5555/1953048.2078195.

94. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods *17*, 261-272. 10.1038/s41592-019-0686-2.

95. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res *47*, D309-D314. 10.1093/nar/gky1085.

96. Buchfink, B., Reuter, K., and Drost, H.G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods *18*, 366-368. 10.1038/s41592-021-01101-x.

97. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol *37*, 852-857. 10.1038/s41587-019-0209-9.

98. Armstrong, G., Martino, C., Rahman, G., Gonzalez, A., Vazquez-Baeza, Y., Mishne, G., and Knight, R. (2021). Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data. mSystems *6*, e0069121. 10.1128/mSystems.00691-21.

99. Wilkinson, L. (2011). ggplot2: elegant graphics for data analysis by WICKHAM, H. Oxford University Press.

100. Pascal Andreu, V., Augustijn, H.E., Chen, L., Zhernakova, A., Fu, J., Fischbach, M.A., Dodd, D., and Medema, M.H. (2023). gutSMASH predicts specialized primary metabolic pathways from the human gut microbiota. Nat Biotechnol *41*, 1416-1423. 10.1038/s41587-023-01675-1.

101. Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., and Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res *49*, W29-W35. 10.1093/nar/gkab335.

102. Alcock, B.P., Huynh, W., Chalil, R., Smith, K.W., Raphenya, A.R., Wlodarski, M.A., Edalatmand, A., Petkau, A., Syed, S.A., Tsang, K.K., et al. (2023). CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res *51*, D690-D699. 10.1093/nar/gkac920.

1303 103. Tesson, F., Herve, A., Mordret, E., Touchon, M., d'Humieres, C., Cury, J., and Bernheim,
1304 A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat
1305 Commun *13*, 2561. 10.1038/s41467-022-30269-9.

1306 104. Larralde, M. (2022). Pyrodigal: Python bindings and interface to Prodigal, an efficient
1307 method for gene prediction in prokaryotes. Journal of Open Source Software *7*, 4296.
1308 10.21105/joss.04296.

1309 105. Eberhardt, R.Y., Haft, D.H., Punta, M., Martin, M., O'Donovan, C., and Bateman, A.
1310 (2012). AntiFam: a tool to help identify spurious ORFs in protein annotation. Database
1311 (Oxford) *2012*, bas003. 10.1093/database/bas003.

1312 106. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering
1313 the next-generation sequencing data. Bioinformatics *28*, 3150-3152.
1314 10.1093/bioinformatics/bts565.

1315 107. Murphy, L.R., Wallqvist, A., and Levy, R.M. (2000). Simplified amino acid alphabets for
1316 protein fold recognition and implications for folding. Protein Eng *13*, 149-152.
1317 10.1093/protein/13.3.149.

1318 108. Hildebrand, A., Remmert, M., Biegert, A., and Soding, J. (2009). Fast and accurate
1319 automatic structure prediction with HHpred. Proteins *77 Suppl 9*, 128-132.
1320 10.1002/prot.22499.

1321 109. Steinegger, M., Meier, M., Mirdita, M., Vohringer, H., Haunsberger, S.J., and Soding, J.
1322 (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC
1323 Bioinformatics *20*, 473. 10.1186/s12859-019-3019-7.

1324 110. Néron, B., Denise, R., Coluzzi, C., Touchon, M., Rocha, E.P., and Abby, S.S. (2023).
1325 MacSyFinder v2: Improved modelling and search engine to identify molecular systems in
1326 genomes. Peer Community Journal *3*. 10.24072/pcjournal.250.

1327 111. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H.,
1328 Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein
1329 function classification. Bioinformatics *30*, 1236-1240. 10.1093/bioinformatics/btu031.

1330 112. Lun, A., Hicks, S., Courbayre, B., Borman, T., and Lahti, L. (2024). bluster: Clustering
1331 Algorithms for Bioconductor. R package version 1.14.0.

1332 113. Csardi, G., and Nepusz, T. (2006). The igraph software. Complex syst *1695*, 1-9.

1333 114. Mallick, H., Rahnavard, A., McIver, L.J., Ma, S., Zhang, Y., Nguyen, L.H., Tickle, T.L.,
1334 Weingart, G., Ren, B., Schwager, E.H., et al. (2021). Multivariable association discovery
1335 in population-scale meta-omics studies. PLoS Comput Biol *17*, e1009442.
1336 10.1371/journal.pcbi.1009442.

1337 115. Ma, S., Shungin, D., Mallick, H., Schirmer, M., Nguyen, L.H., Kolde, R., Franzosa, E.,
1338 Vlamakis, H., Xavier, R., and Huttenhower, C. (2022). Population structure discovery in

meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin. Genome Biol *23*, 208. 10.1186/s13059-022-02753-4.

116. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421. 10.1186/1471-2105-10-421.

117. Eren, A.M., Esen, O.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ *3*, e1319. 10.7717/peerj.1319.

118. Wirbel, J., Zych, K., Essex, M., Karcher, N., Kartal, E., Salazar, G., Bork, P., Sunagawa, S., and Zeller, G. (2021). Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol *22*, 93. 10.1186/s13059-021-02306-1.

119. Lin, H., and Peddada, S.D. (2020). Analysis of compositions of microbiomes with bias correction. Nat Commun *11*, 3514. 10.1038/s41467-020-17041-7.

120. Kurtz, Z.D., Muller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., and Bonneau, R.A. (2015). Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol *11*, e1004226. 10.1371/journal.pcbi.1004226.

Figures

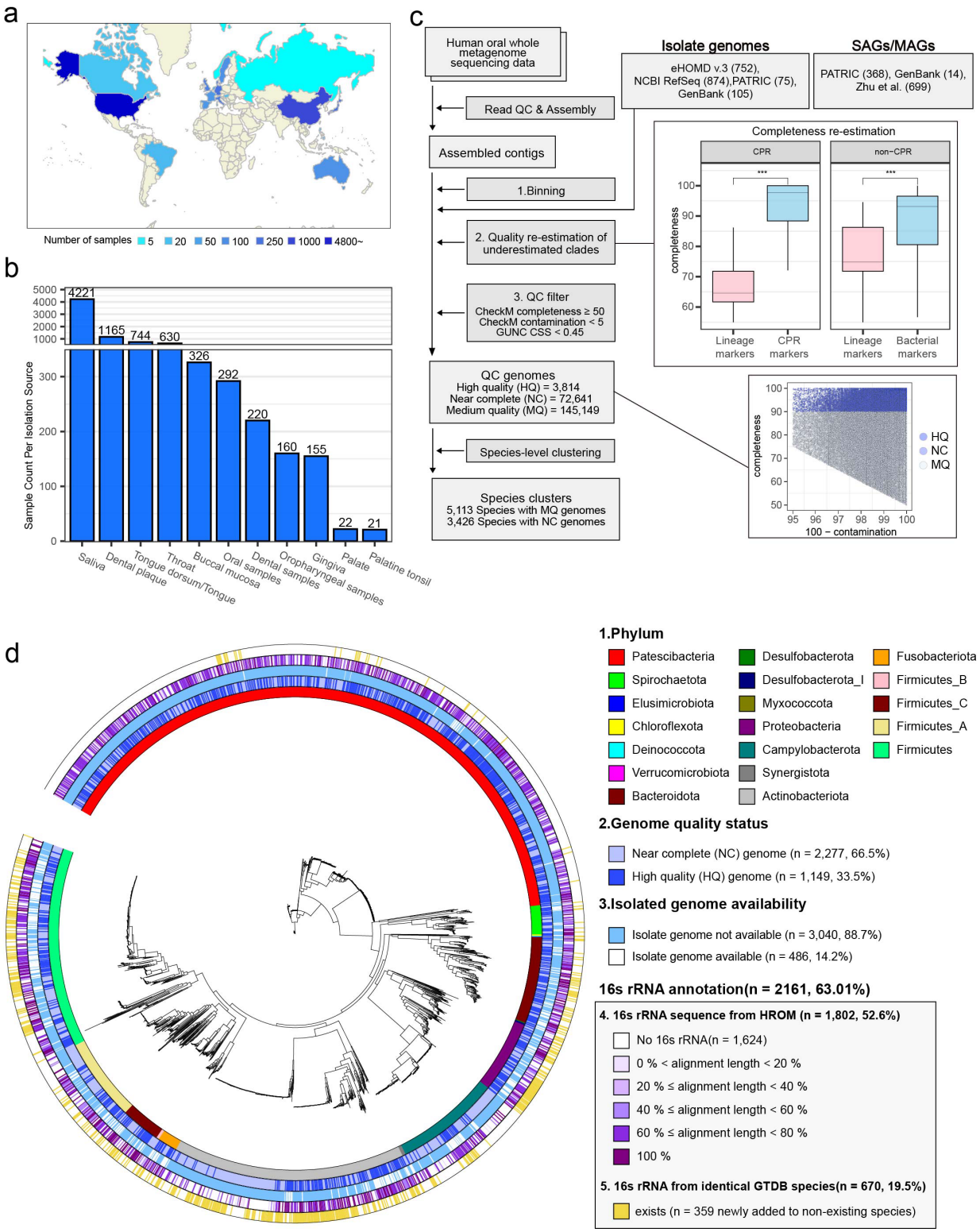


Fig. 1 | Overview of the human reference oral microbiome (HROM) construction and phylogenetic diversity. a, Geographic distribution of 7,956 whole metagenome sequencing (WMS) datasets from human oral cavities,

representing samples collected across various countries. **b**, Distribution of WMS datasets across diverse isolation sources within the oral cavity, highlighting the number of samples for each source. **c**, Bioinformatics workflow for constructing the HROM catalog, including read quality control (QC), assembly, genome binning, re-evaluation of completeness and contamination, and species-level clustering. Completeness differences in re-evaluated genomes are shown, with statistical significance assessed using a two-tailed Mann-Whitney U test (***$P < 0.001$). Boxplot elements: center line (median), box edges (25th and 75th percentiles), whiskers (1.5× interquartile range). **d**, Phylogenetic tree of 3,424 near-complete, species-level bacterial genomes in HROM. Tracks around the tree indicate (1) phylum-level classification, (2) genome quality status, (3) isolate genome availability, and (4) completeness of 16S rRNA annotations. Genome quality categories are defined in the workflow, and 16S rRNA annotations include genomes from HROM and external references.

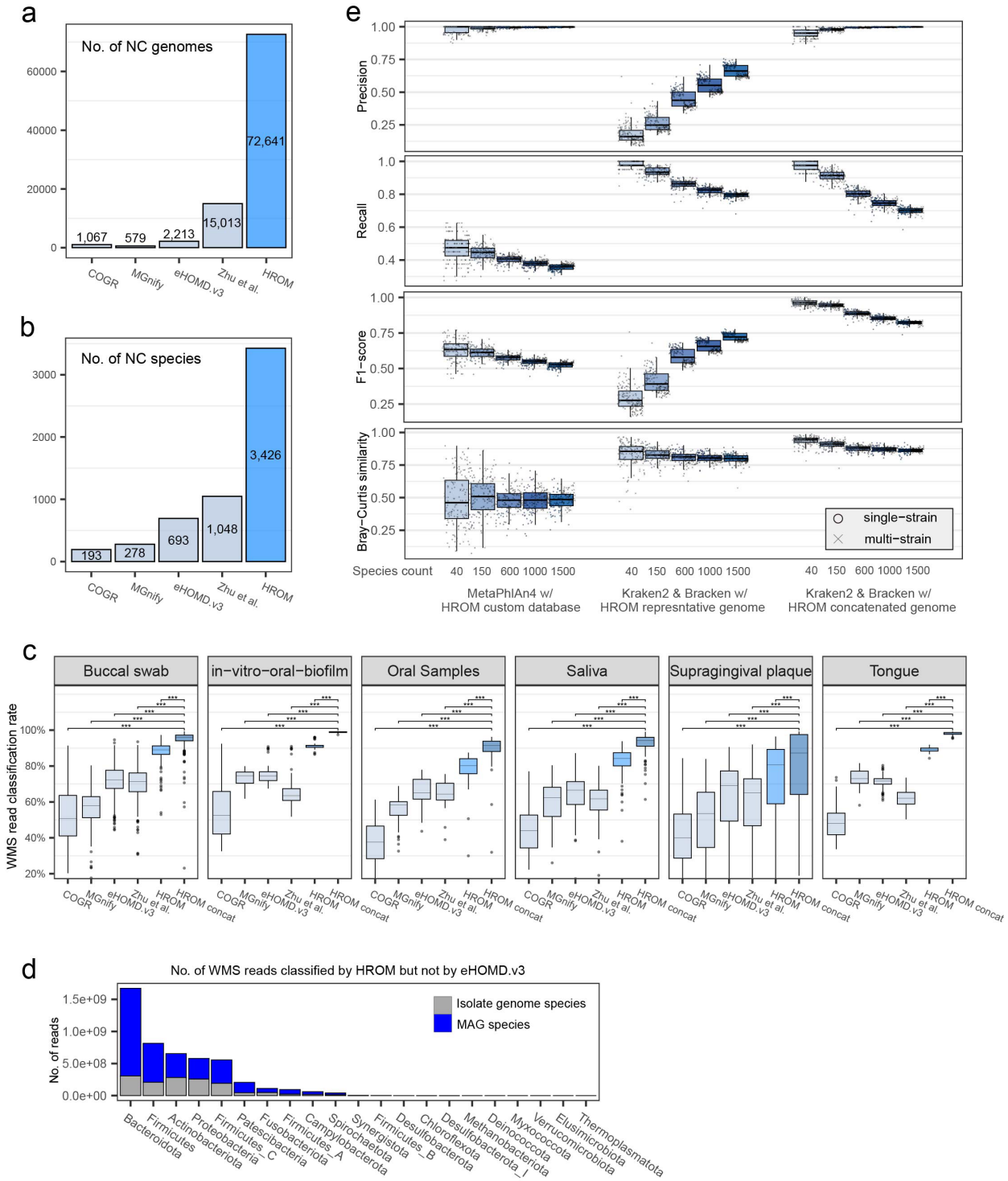


Fig. 2 | Comparative analysis and Benchmarking of the HROM catalog. **a**, Total number of near-complete (NC) genomes in the HROM catalog compared with other oral genome catalogs, including COGR, MGnify, eHOMD v3, and Zhu et al. For the Zhu et al. catalog, only representative genomes were publicly available, therefore, the total genome count is based on the reported statistics. **b**, Number of NC species identified in the HROM catalog compared to other genome catalogs. For Zhu et al., only representative genomes were publicly available; thus, we present the count of non-chimeric, NC representative genomes. **c**, Classification rates of WMS reads for different oral sample types (buccal swab, in-vitro oral biofilm, oral samples, saliva, supragingival plaque, and tongue) using various genome

1381 catalogs. For consistency, only non-chimeric and NC genomes were used for classification. Except for 'HROM concat,'
1382 which represents the concatenated HROM database, only representative genomes were utilized. HROM achieves the
1383 highest classification rates across all sample types. Statistical significance was assessed using a two-tailed Mann-
1384 Whitney U test (*** $P < 0.001$). **d**, Total number of WMS reads from a validation dataset (PRJNA917836, buccal
1385 mucosa) that were classified by HROM but not by eHOMD v3, broken down by isolate genome species and MAG
1386 species, across different taxonomic phyla. **e**, Benchmarking performance of species-level taxonomic profiling methods
1387 using simulated datasets of varying species complexity (40, 150, 600, 1,000, 1,500 species). Metrics include precision,
1388 recall, F1-score, and Bray-Curtis similarity. Results are shown for MetaPhlAn4 with the HROM custom database, and
1389 Kraken2/Bracken with both HROM representative and concatenated genome databases, under single-strain and multi-
1390 strain scenarios.

1391

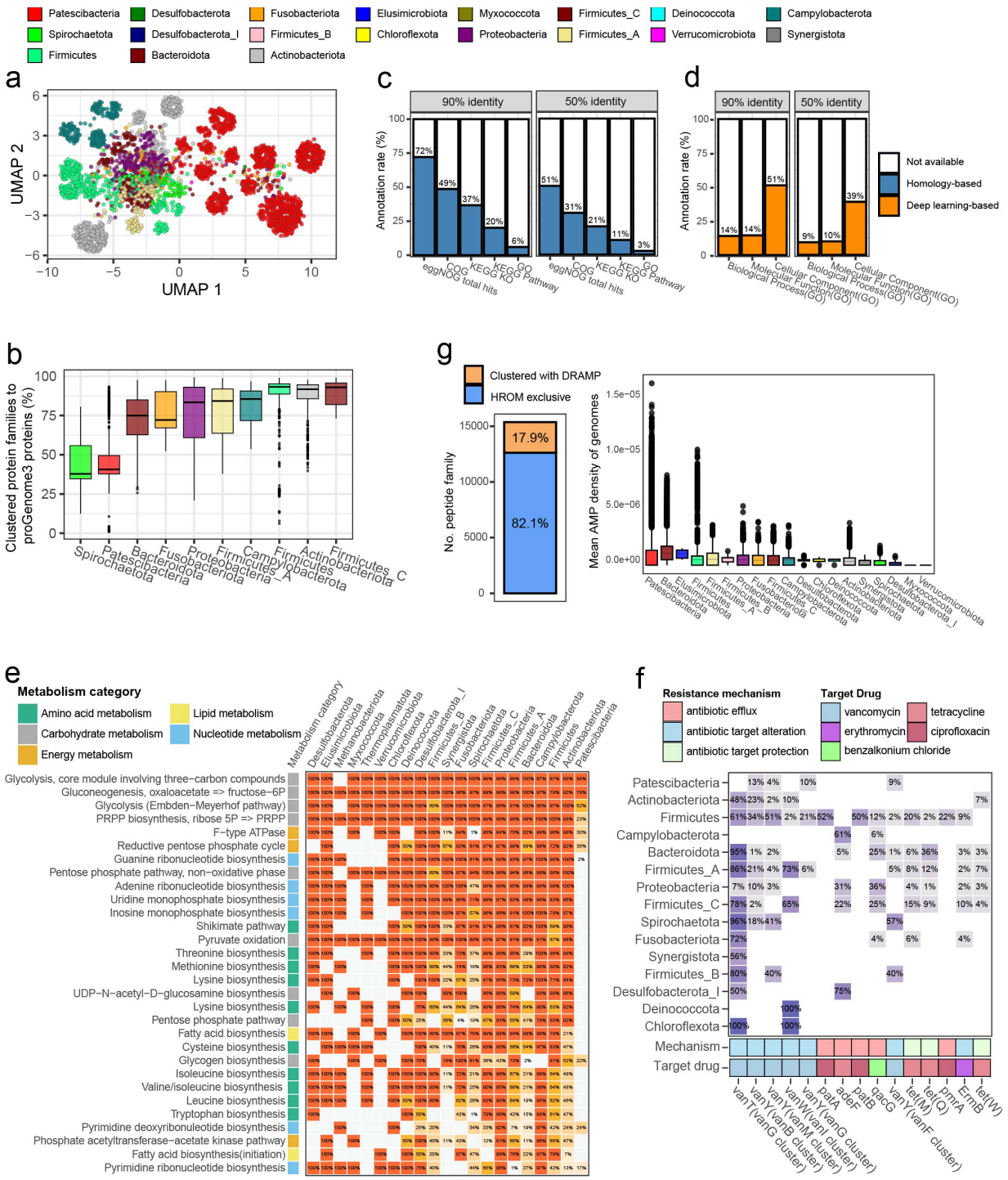
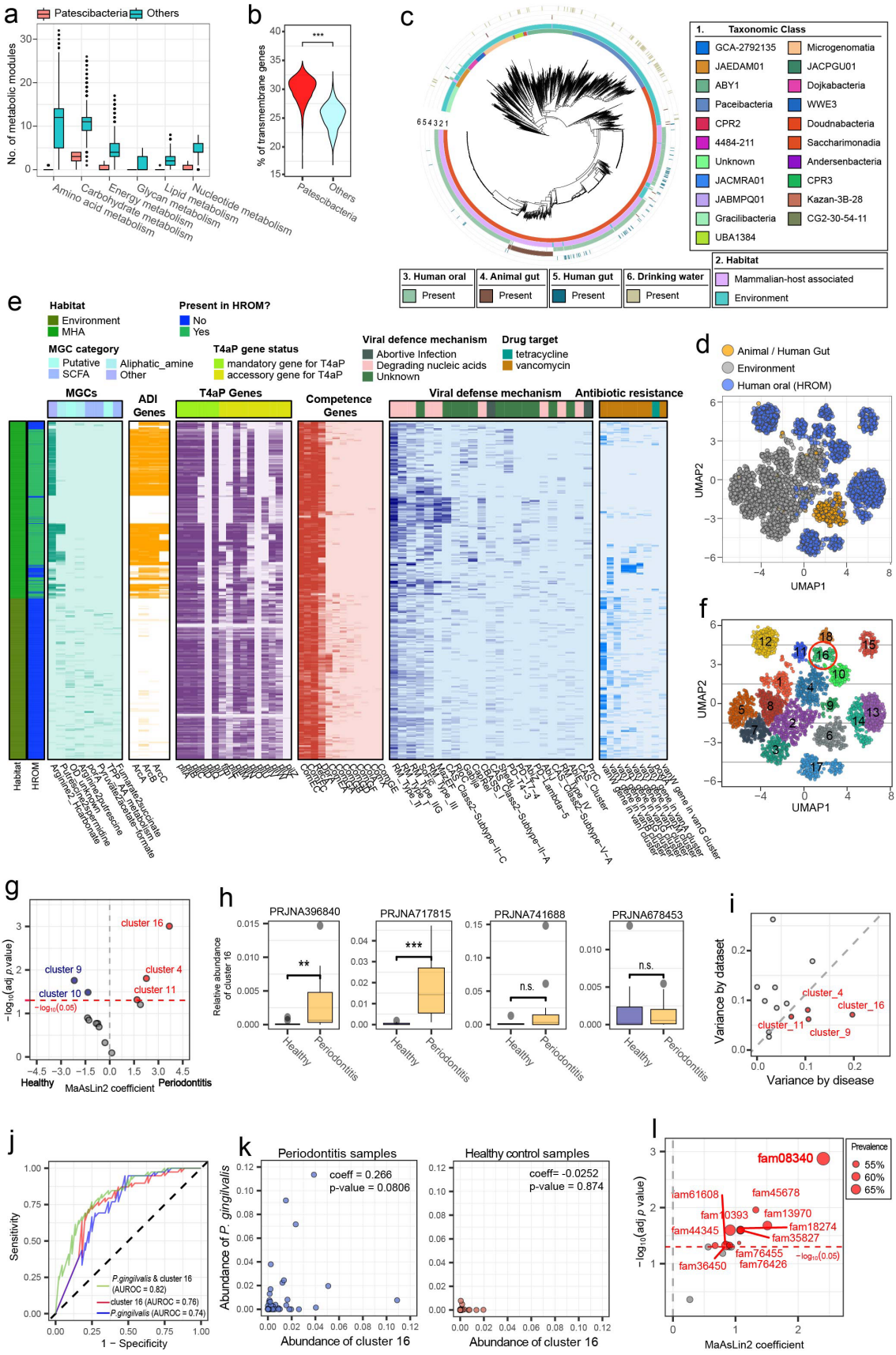


Fig. 3 | Functional and metabolic landscape of the HROM. a, UMAP visualization of protein family profiles across bacterial genomes in HROM, based on the Jaccard distance. Proteins sequences from representative genomes were clustered into protein families at 50% sequence identity. **b**, Distribution of the percentages of clustered HROM protein families with proGenomes3 proteins at 50% sequence identity across major phyla. **c**, Annotation rates of HROM protein families at 90% and 50% sequence identity for Cluster of Orthologs Groups (COG), KEGG Ortholog (KO), KEGG pathway, and Gene Ontology (GO) terms using a homology-based method, eggNOG-mapper. **d**, Annotation rates of HROM protein families at 90% and 50% sequence identity for GO biological processes, molecular functions, and cellular components terms using a deep learning-based method, DeepGOPlus. **e**, Heatmap of the top 20 most

1401 abundant KEGG metabolic modules across HROM phyla, categorized into amino acid metabolism, carbohydrate
1402 metabolism, energy metabolism, lipid metabolism, and nucleotide metabolism. Numbers indicate the prevalence of
1403 each module within phyla. A metabolic module was considered present if it was identified in at least one conspecific
1404 genome within the same species. **f**, Heatmap of the top 15 most abundant antibiotic resistance genes with their
1405 mechanisms and target drugs across HROM phyla. The proportion of genomes encoding resistance mechanisms such
1406 as efflux, target alteration, and target protection is shown for each phylum and drug type. **g**, Summary statistics of
1407 antimicrobial peptides (AMPs) found in HROM. Left: Proportion of HROM AMPs clustered with those from Database
1408 of Antimicrobial Peptides (DRAMP) compared to HROM-exclusive AMPs. Right: Boxplot showing the mean AMP
1409 density of genomes across phyla, highlighting the prevalence of AMPs within each taxonomic group. (b, g) Boxplot
1410 elements: center line (median), box edges (25th and 75th percentiles), whiskers (1.5× interquartile range).

1411


a
Patescibacteria
Others
No. of metabolic modules
Amino acid metabolism
Carbohydrate metabolism
Energy metabolism
Glycan metabolism
Lipid metabolism
Nucleotide metabolism
b
% of transmembrane genes
Patescibacteria
Others
c
1. Taxonomic Class
GCA-2792135
JAEDAM01
ABY1
Paceibacteria
CPR2
4484-211
Unknown
JACMRA01
JABMPQ01
Gracilibacteria
UBA1384
Microgenomatia
JACPGU01
Dojkabacteria
WWE3
Doudnabacteria
Saccharimonadia
Andersenbacteria
CPR3
Kazan-3B-28
CG2-30-54-11
2. Habitat
Mammalian-host associated
Environment
3. Human oral
4. Animal gut
5. Human gut
6. Drinking water
Present
e
Habitat
Environment
MHA
Present in HROM?
No
Yes
MGC category
Putative
SCFA
Aliphatic_amine
Other
T4aP gene status
mandatory gene for T4aP
accessory gene for T4aP
Viral defence mechanism
Abortive Infection
Degrading nucleic acids
Unknown
Drug target
tetracycline
vancomycin
MGCs
ADI Genes
T4aP Genes
Competence Genes
Viral defense mechanism
Antibiotic resistance
d
Animal / Human Gut
Environment
Human oral (HROM)
UMAP1
UMAP2
f
UMAP1
UMAP2
g
cluster 16
cluster 9
cluster 4
cluster 10
cluster 11
Healthy
MaAsLin2 coefficient
Periodontitis
h
PRJNA396840
PRJNA717815
PRJNA741688
PRJNA678453
Relative abundance of cluster 16
Healthy
Periodontitis
n.s.
i
Variance by dataset
Variance by disease
cluster_4
cluster_16
cluster_11
cluster_9
j
Sensitivity
1 − Specificity
P.gingivalis & cluster 16 (AUROC = 0.82)
cluster 16 (AUROC = 0.76)
P.gingivalis (AUROC = 0.74)
k
Periodontitis samples
Healthy control samples
coeff = 0.266
p-value = 0.0806
coeff= -0.0252
p-value = 0.874
Abundance of P. gingivalis
Abundance of cluster 16
l
fam08340
fam61608
fam45678
fam10393
fam13970
fam44345
fam18274
fam35827
fam76455
fam36450
fam76426
Prevalence
55%
60%
65%
MaAsLin2 coefficient

1412

Fig. 4 | Functional, phylogenetic, and disease-associated analysis of oral Patescibacteria. a, Comparison of the number of metabolic modules in Patescibacteria versus other phyla, categorized into amino acid metabolism, carbohydrate metabolism, energy metabolism, glycan metabolism, lipid metabolism, and nucleotide metabolism. **b**, Violin plot showing the percentage of transmembrane genes in Patescibacteria compared to other phyla. Statistical significance was assessed using a two-tailed Mann-Whitney U test (*** $P < 0.001$). **c**, Phylogenetic tree of 2,963 Patescibacteria species, with outer tracks indicating taxonomic class, habitat, and presence in various environments, including human oral cavity, gut, animal gut, and drinking water. **d**, UMAP visualization of protein family profiles across 2,963 Patescibacteria species representative genomes based on the Jaccard distance, showing clustering by environmental sources (e.g., human oral cavity and gut, animal gut, environment). **e**, Heatmap showing the distribution of metabolic gene clusters (MGCs), ADI genes, T4aP genes, competence genes, viral defense mechanisms, and antibiotic resistance genes across 2,963 Patescibacteria species. Colors represent habitat types and gene presence. **f**, UMAP visualization of Patescibacteria species clusters based on protein family profiles similarity. Each cluster is numbered and color-coded, with cluster 16 highlighted by a red circle. **g**, Identification of differentially enriched Patescibacteria clusters in periodontitis samples using MaAsLin2. Clusters meeting the adjusted p-value threshold ($P \leq 0.05$, Bonferroni-Hochberg correction) are highlighted in red, with the horizontal red line indicating the significance threshold. Cluster 16 is strongly associated with periodontitis. **h**, Relative abundance of cluster 16 across four independent periodontitis datasets, showing significant enrichment in two out of four periodontitis samples compared to healthy controls. Statistical significance was assessed using a two-tailed Mann-Whitney U test (*** $P < 0.001$, ** $P < 0.01$, n.s. = not significant). **i**, Variance explained by dataset (y-axis) versus disease label (x-axis) for different clusters of Patescibacteria species in periodontitis-associated datasets, with clusters showing greater disease-associated variance highlighted in red. **j**, Receiver operating characteristic (ROC) curve for periodontitis classification based on the batch-corrected relative abundance of *P. gingivalis* (blue), cluster 16 (red), and their combined abundance (green). The area under the ROC curve (AUROC) scores are indicated. **k**, Comparison of the batch-corrected relative abundance of cluster 16 and *P. gingivalis* in healthy and periodontitis samples. The Pearson correlation coefficient and *p*-value are indicated. **l**, Enriched protein families in cluster 16 associated with periodontitis. The size of each dot represents prevalence, and significant families are labeled with MaAsLin2 coefficients and adjusted *p*-values threshold ($P \leq 0.05$, Bonferroni-Hochberg correction) are highlighted in red, with the horizontal red line indicating the significance threshold. (a, h) Boxplot elements: center line (median), box edges (25th and 75th percentiles), whiskers (1.5× interquartile range).

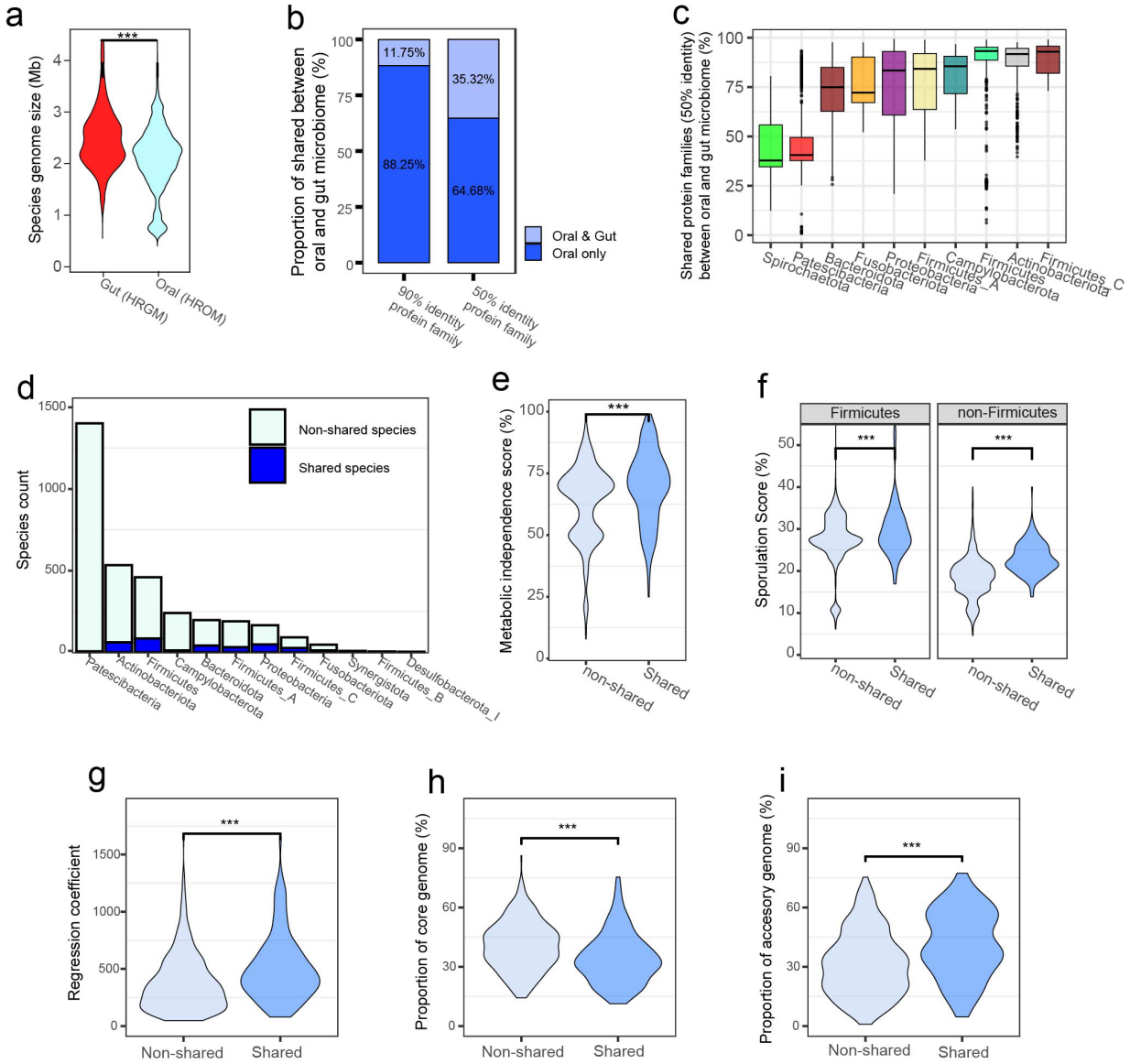


Fig. 5 | Functional analysis of oral-gut shared bacterial species. a, Violin plot comparing the genome size of species in the gut (HRGM2) versus the oral cavity (HROM). Statistical significance was assessed using a two-tailed Mann-Whitney U test (*** $P < 0.001$). **b**, Proportion of shared protein families between oral and gut microbiomes at 90% and 50% sequence identity, showing the percentages of protein families exclusive to the oral microbiome and those shared with the gut microbiome. **c**, Boxplot of the proportion of shared protein families (50% sequence identity) between oral and gut microbiomes across major phyla. Boxplot elements: center line (median), box edges (25th and 75th percentiles), whiskers (1.5× interquartile range). **d**, Species count for oral-gut shared and non-shared species across major phyla. Shared species are defined as those found in both microbiomes. **e**, Violin plot comparing the metabolic independence score (%) between oral-gut shared and non-shared species. Patescibacteria were excluded due to their low metabolic independence and minimal oral-gut sharing to prevent bias. Statistical significance was assessed using a two-tailed Mann-Whitney U test (*** $P < 0.001$). **f**, Violin plot comparing sporulation scores (%) between oral-gut shared and non-shared species, grouped by Firmicutes and non-Firmicutes, with each species' score determined by the maximum value among its conspecific genomes. Statistical significance was assessed using a two-tailed Mann-Whitney U test (*** $P < 0.001$). **g-i**, Violin plots comparing the regression coefficient of pangenome openness, with higher coefficients indicating a more open pangenome (**g**), the proportion of core genome (%) (**h**), and the proportion of accessory genome (%) (**i**) between oral-gut shared and non-shared species (based on 125 shared and

389 non-shared species). Statistical significance for these panels was assessed using a two-tailed Mann-Whitney U test (*** $P < 0.001$).

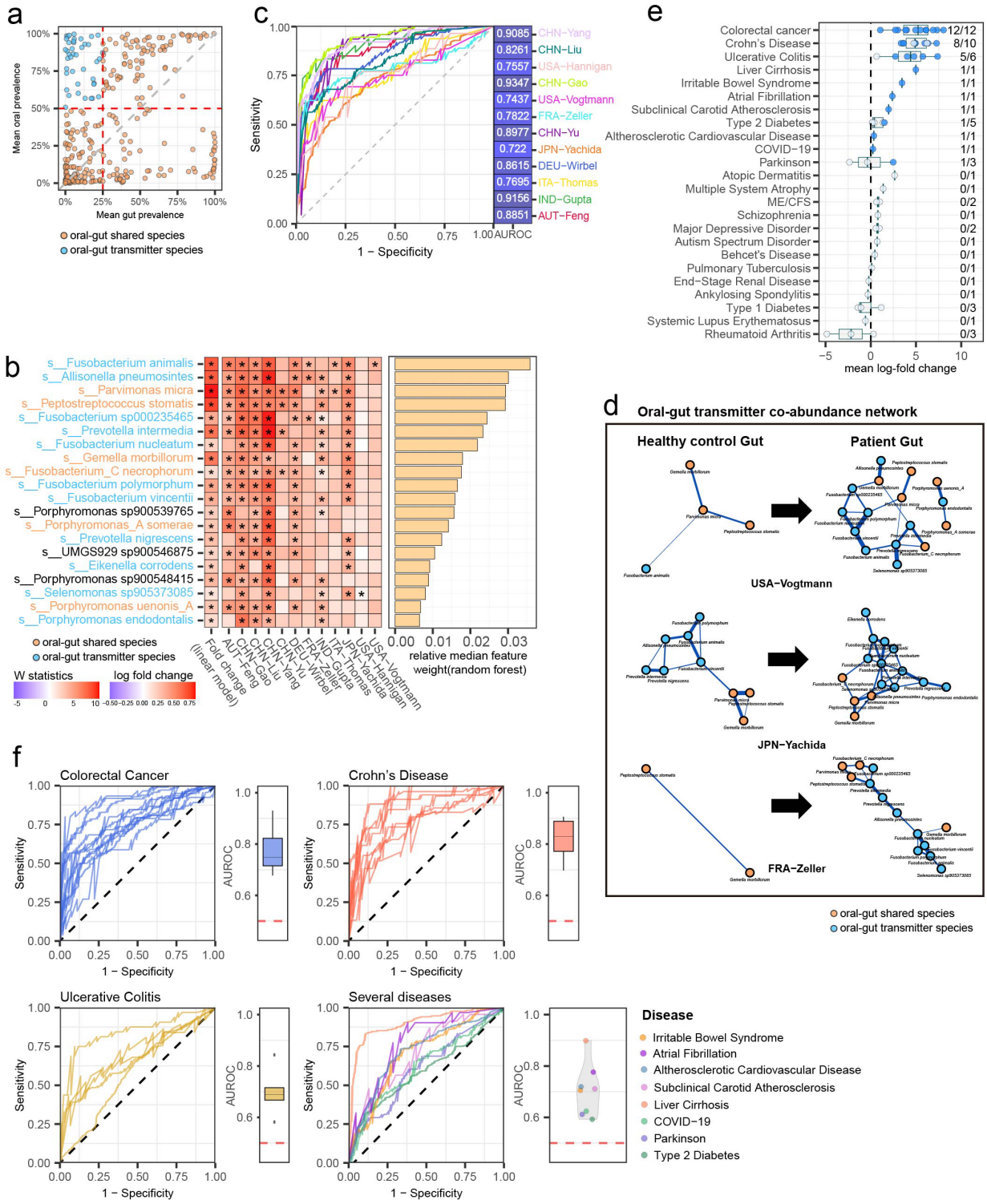
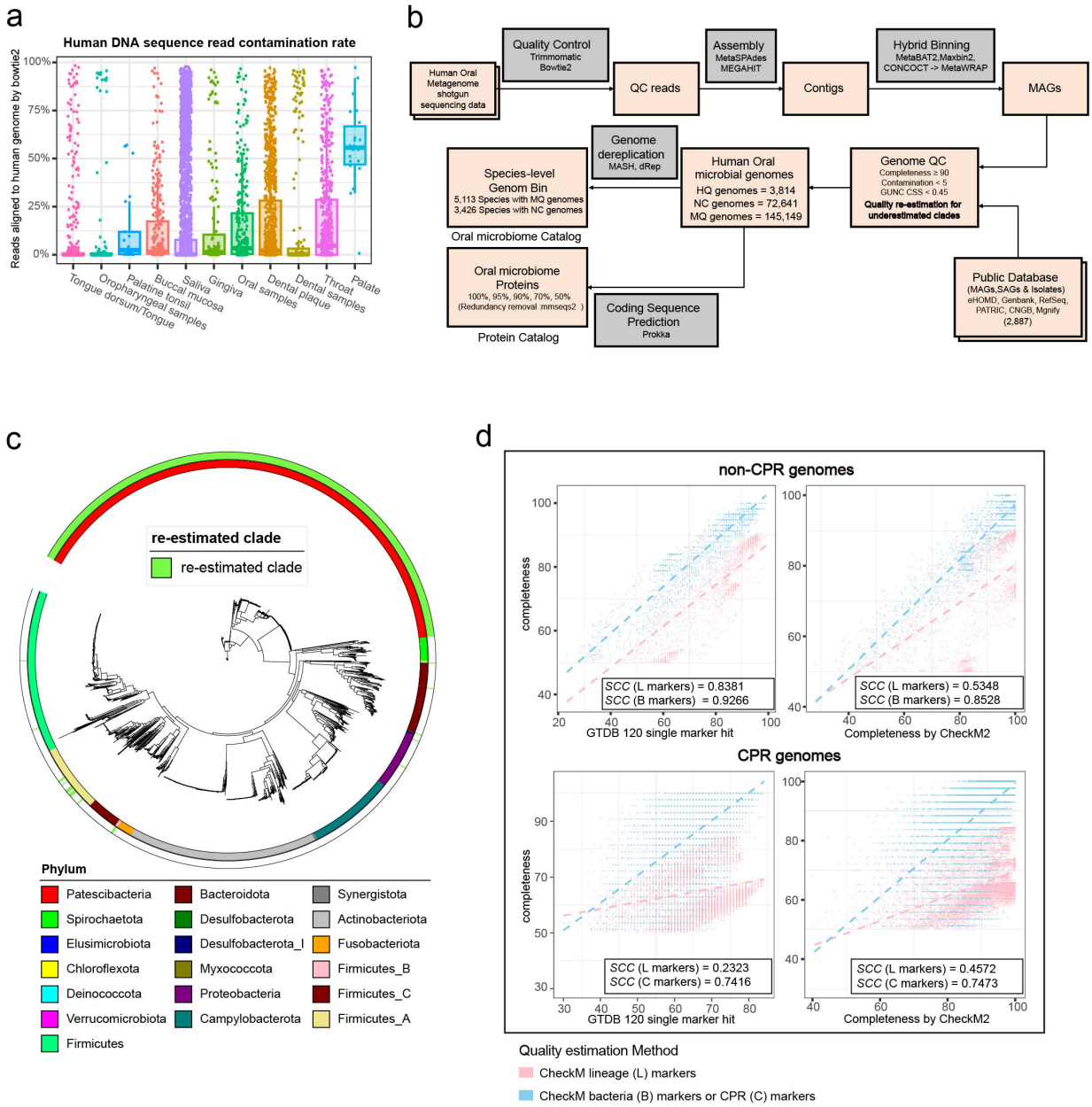


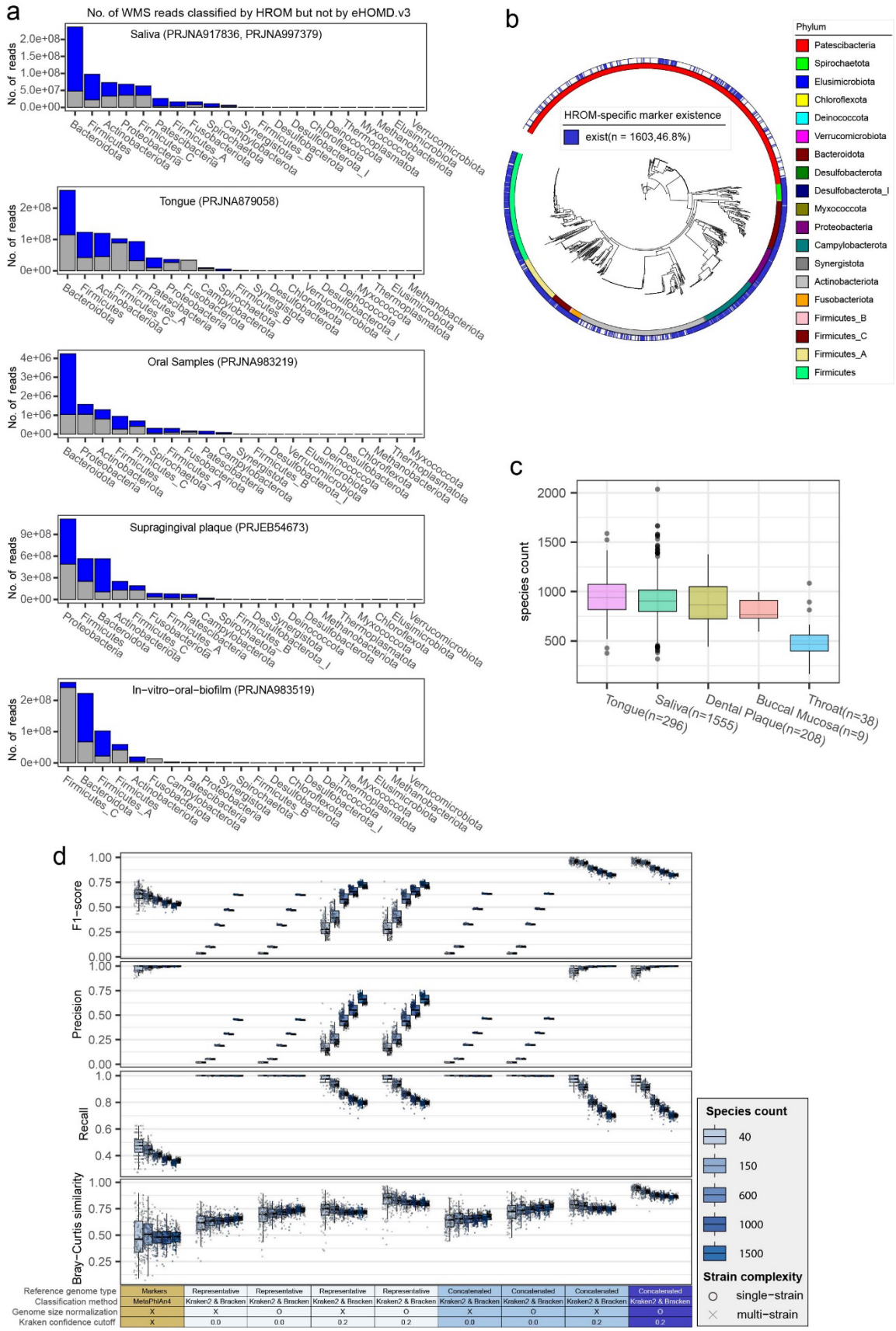
Fig. 6 | Disease association of oral-gut transmitter species. a, Scatterplot showing the mean prevalence of 330 oral-gut shared species (orange) across 9 healthy gut samples (x-axis) and 14 healthy oral samples (y-axis). Forty-two oral-gut transmitter species (blue) are defined as those present in more than 50% of healthy oral samples but less than 25% of healthy gut samples. The prevalence thresholds are indicated by red lines. **b**, Heatmap displaying the top 20 bacterial species identified by random forest models as important features for colorectal cancer (CRC) classification. The left heatmap shows log-fold changes derived from linear models using SIAMCAT, with statistical significances

determined by FDR-adjusted p-values (* adjusted $P \leq 0.05$). The central heatmap presents W statistics from ANCOM-BC analysis across 12 CRC cohorts, with stars denoting statistically significant associations (* adjusted $P \leq 0.05$) based on Holm-Bonferroni-corrected p-value. The bar chart (right) displays the relative median feature weight of each species extracted from random forest model. **c**, Receiver operating characteristic (ROC) curves and area under the ROC (AUROC) values for CRC prediction across 12 cohorts based on the sum of relative abundances of the 17 oral species among the top 20 key species for CRC classification. **d**, Co-abundance network of 17 oral species among the top 20 key species for CRC classification in healthy (left) and patient (right) gut microbiome samples across different cohorts (USA-Vogtmann, JPN-Yachida, FRA-Zeller). Orange and blue nodes represent oral-gut shared and oral-gut transmitter species, respectively. Thicker edges in patient networks indicate stronger co-abundance relationships. **e**, Mean log-fold change of sum of relative abundance of 42 oral-gut transmitter species in gut microbiomes associated with various systemic diseases, including colorectal cancer, Crohn's disease, ulcerative colitis, and cardiovascular diseases. Disease association is indicated by the fraction of cohorts showing a significant increase in abundance (two-tailed Mann-Whitney U test, $P \leq 0.05$). Numbers on the right side indicate the number of datasets with significant enrichment. ME/CFS: Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. **f**, ROC curves and AUROC values for disease classification using the sum of relative abundance of 42 oral-gut transmitter species for colorectal cancer, Crohn's disease, ulcerative colitis, and several other diseases. Boxplots display AUROC values across cohorts, with significant predictive performance for intestinal diseases and moderate performance for other diseases. (e-f) Boxplot elements: center line (median), box edges (25th and 75th percentiles), whiskers (1.5× interquartile range).



Extended Data Fig. 1 | Quality control and completeness re-estimation of oral microbial genomes. a, Boxplot showing the human DNA sequence contamination rates of whole metagenome sequencing datasets across various oral sample types. Contamination rates were estimated by aligning reads to the human reference genome, GRCh38, using Bowtie2. Boxplot elements: center line (median), box edges (25th and 75th percentiles), whiskers (1.5× interquartile range). **b**, Bioinformatics workflow for the construction of the Human Reference Oral Microbiome (HROM) catalog, including quality control (QC), genome assembly, binning, and completeness re-estimation for underestimated clades. The pipeline integrates publicly available databases and hybrid binning methods to produce species-level genome bins and a comprehensive protein catalog. **c**, Phylogenetic tree highlighting clades with re-estimated genome completeness. Green bars indicate clades where completeness was re-estimated due to underestimation using lineage-specific markers. Taxonomic assignments are color-coded by phylum. Most of the re-estimated genomes belong to phylum Patescibacteria. **d**, Scatterplots comparing genome completeness estimates across different methods for non-CPR (top row) and CPR (bottom row) genomes. Completeness was evaluated using CheckM with lineage-specific markers (L markers), universal bacterial markers (B markers), and CPR-specific markers (C markers). Spearman correlation

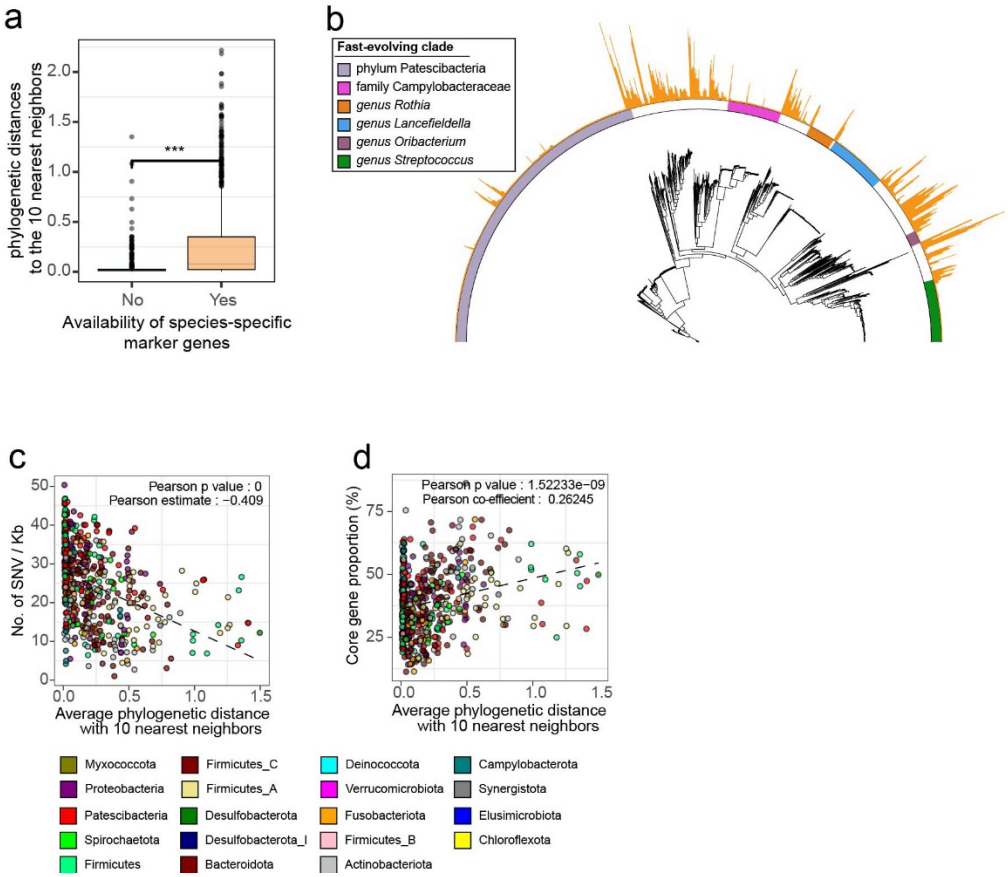
1503 coefficients (SCC) between the methods and GTDB 120 single marker hits or CheckM2 completeness are shown for
1504 each plot.
1505


a
No. of WMS reads classified by HROM but not by eHOMD.v3
Saliva (PRJNA917836, PRJNA997379)
Tongue (PRJNA879058)
Oral Samples (PRJNA983219)
Supragingival plaque (PRJEB54673)
In-vitro-oral-biofilm (PRJNA983519)
No. of reads
b
HROM-specific marker existence
exist(n = 1603,46.8%)
Phylum
Patescibacteria
Spirochaetota
Elusimicrobiota
Chloroflexota
Deinococcota
Verrucomicrobiota
Bacteroidota
Desulfobacterota
Desulfobacterota_I
Myxococcota
Proteobacteria
Campylobacterota
Synergistota
Actinobacteriota
Fusobacteriota
Firmicutes_B
Firmicutes_C
Firmicutes_A
Firmicutes
c
species count
Tongue(n=296)
Saliva(n=1555)
Dental Plaque(n=208)
Buccal Mucosa(n=9)
Throat(n=38)
d
F1-score
Precision
Recall
Bray-Curtis similarity
Species count
40
150
600
1000
1500
Strain complexity
single-strain
multi-strain
Reference genome type
Classification method
Genome size normalization
Kraken confidence cutoff
Markers
Representative
Concatenated
MetaPhlAn4
Kraken2 & Bracken

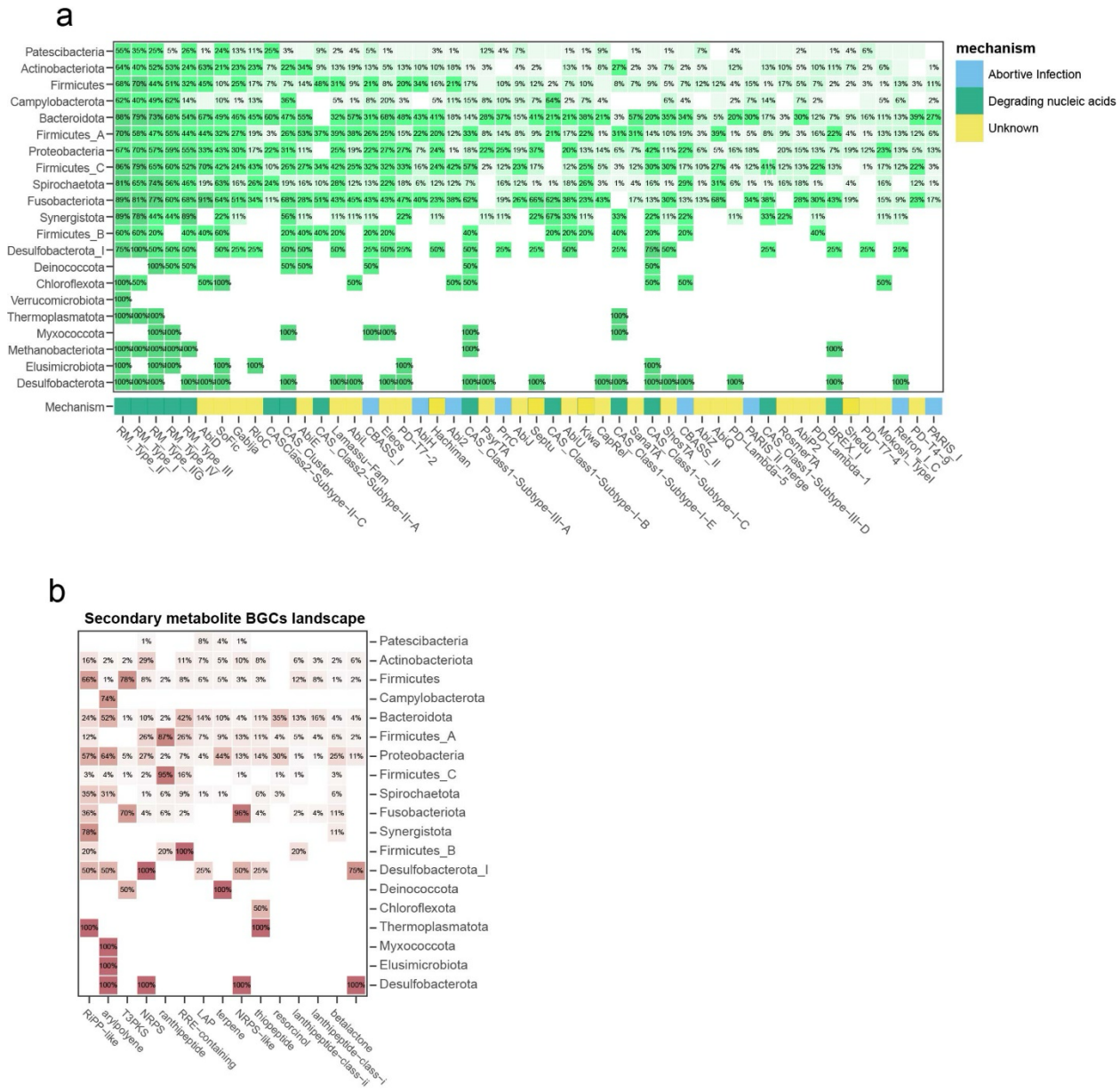
1506

1507 **Extended Data Fig. 2 | Benchmarking and expanded classification capability of the HROM catalog. a**, Number
1508 of whole metagenome sequencing (WMS) reads classified by HROM but not by eHOMD v3 across various oral
1509 sample types, including saliva, tongue, oral samples, supragingival plaque, and in-vitro oral biofilm. Taxonomic
1510 breakdown highlights the significant contribution of HROM in classifying previously unclassified reads across diverse
1511 phyla. **b**, Phylogenetic tree of HROM representative genomes with tracks indicating the availability of species-specific
1512 marker genes (blue), with 46.8% of genomes showing species-specific marker existence. Phyla are color-coded. **c**,
1513 Boxplot of species counts per sample across different oral sample sources, including tongue, saliva, dental, buccal
1514 plaque, and throat mucosa. Sample sizes for each source are indicated. **d**, Benchmarking of species profiling
1515 performance across varying levels of species and strain complexity using different classification methods
1516 (MetaPhlAn4 with a custom HROM species-specific marker database, Kraken2/Bracken with a custom HROM
1517 representative genome database, and Kraken2/Bracken with a custom HROM concatenated genome database). The
1518 evaluation considered variations with and without genome length normalization and a confidence score cutoff 0.2.
1519 Metrics include F1-score, precision, recall, and Bray-Curtis similarity, highlighting the superior performance of
1520 HROM in both single-strain and multi-strain scenarios. (c-d) Boxplot elements: center line (median), box edges (25th
1521 and 75th percentiles), whiskers (1.5× interquartile range).

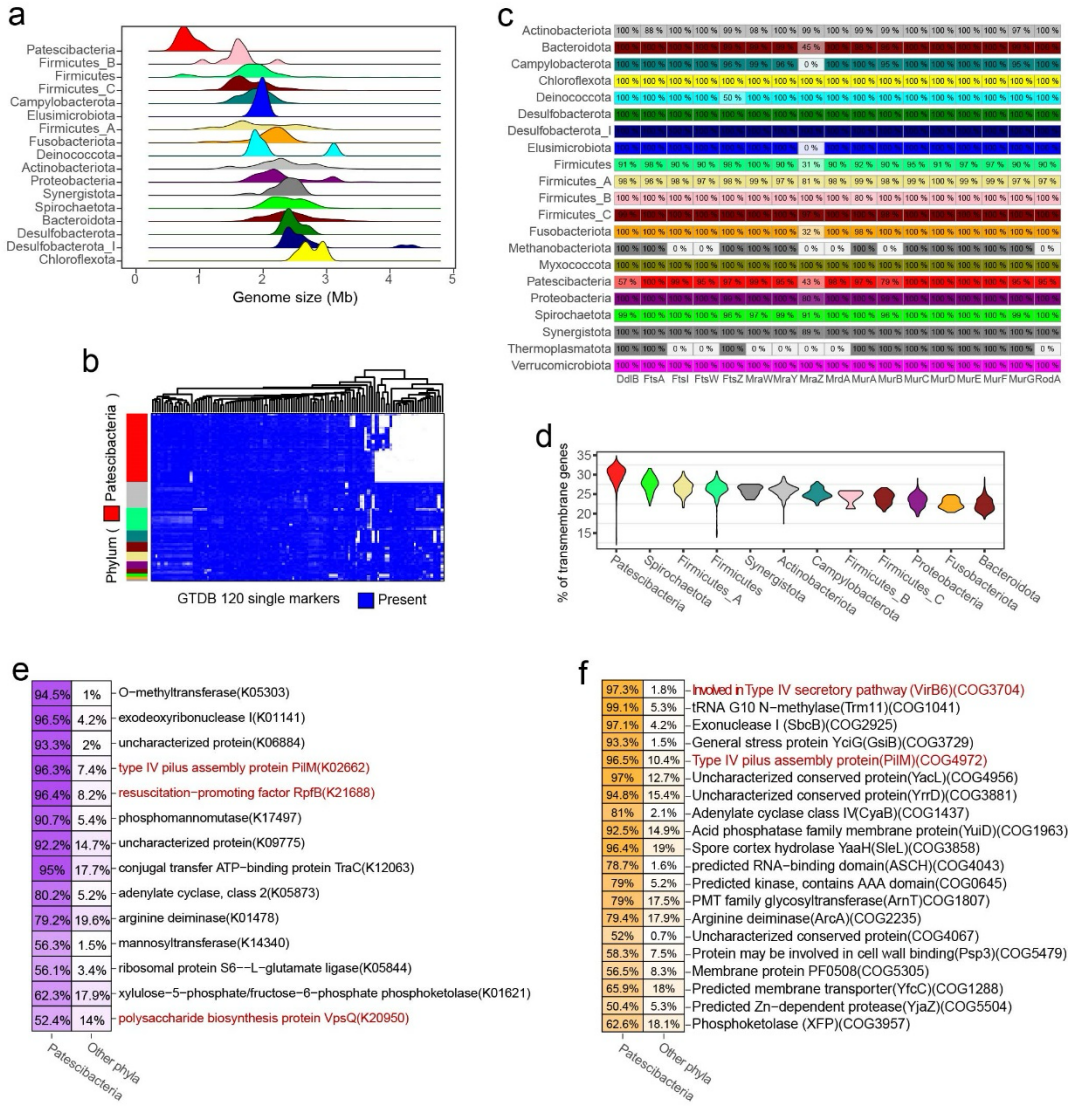
1522



Extended Data Fig. 3 | Phylogenetic and genomic features of fast-evolving oral microbial clades. a, Boxplot comparing the phylogenetic distances to the 10 nearest neighbors for species with (Yes) and without (No) species-specific marker genes. Species lacking specific markers exhibit smaller phylogenetic distances, indicating closely related neighbors. Boxplot elements: center line (median), box edges (25th and 75th percentiles), whiskers (1.5× interquartile range). Statistical significance was determined using a two-tailed Mann-Whitney U test (*** $P < 0.001$). **b**, Phylogenetic tree of oral microbiota highlighting six clades undergoing rapid speciation. Outer tracks indicate phylogenetic distances to the 10 nearest neighbors (orange bars) and taxonomic classification (color-coded by phylum). **c**, Scatterplot showing the relationship between single nucleotide variant (SNV) density (the number of SNVs per kb) and the average phylogenetic distance to the 10 nearest neighbors. **d**, Scatterplot showing the relationship between the proportion of core genes (%) and the average phylogenetic distance to the 10 nearest neighbors.



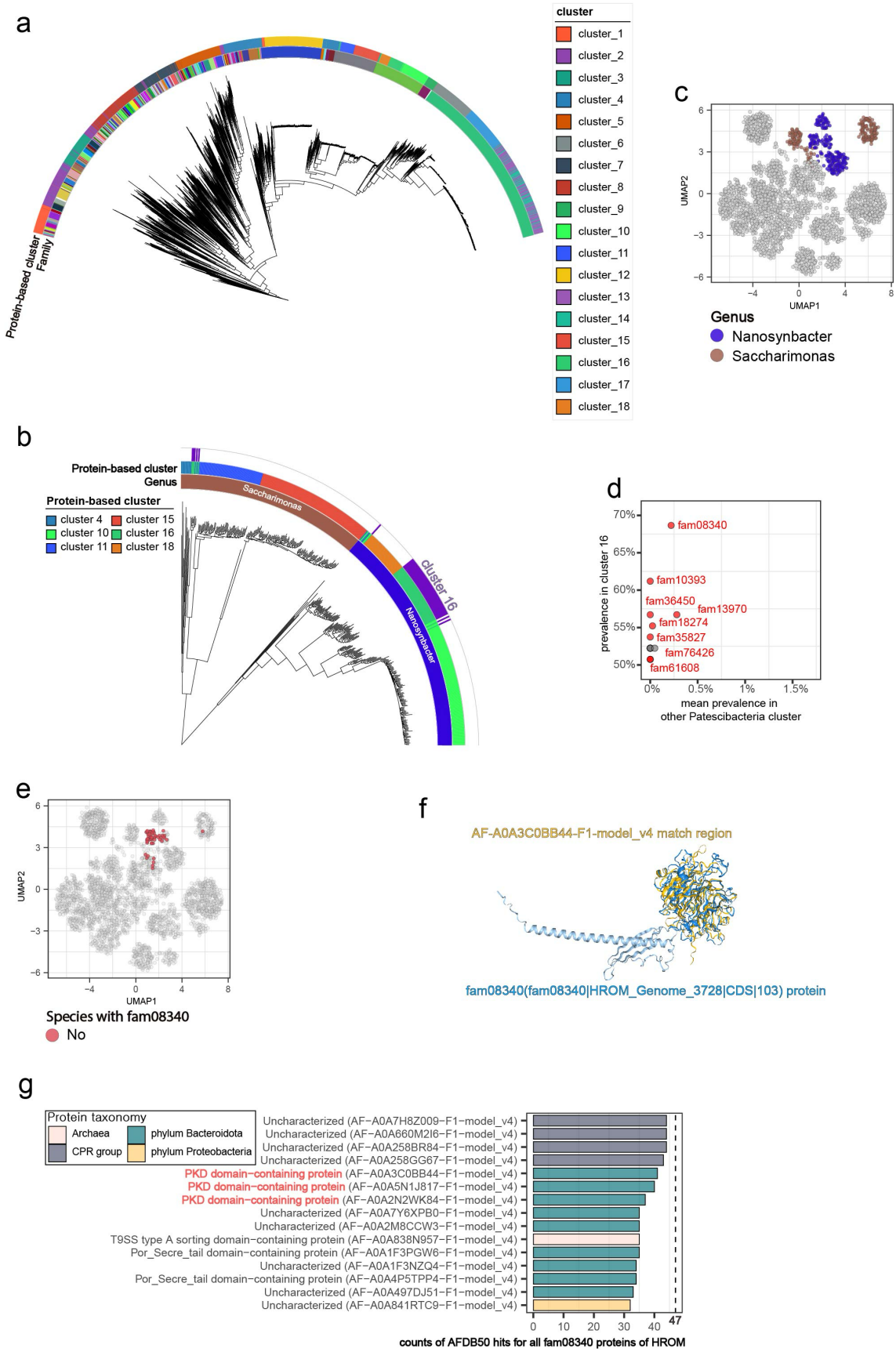
Extended Data Fig. 4 | Functional landscape of viral defense mechanisms and secondary metabolite BGCs in oral microbial phyla. a, Heatmap showing the distribution of viral defense mechanisms across oral microbial phyla. Rows represent phyla, and columns correspond to specific viral defense mechanisms. Prevalence percentages are displayed for each phylum and mechanism. Viral defense mechanism was considered present if it was identified in at least one conspecific genome within the same species. **b**, Heatmap representing the distribution of secondary metabolite biosynthetic gene clusters (BGCs) across major oral microbial phyla. Rows correspond to phyla, and columns indicate specific types of BGCs. The prevalence of each BGC type within phyla is shown as percentages, highlighting the diversity of secondary metabolite production capabilities among phyla. BGCs was considered present if it was identified in at least one conspecific genome within the same species.


a
Patescibacteria
Firmicutes_B
Firmicutes
Firmicutes_C
Campylobacterota
Elusimicrobiota
Firmicutes_A
Fusobacteriota
Deinococcota
Actinobacteriota
Proteobacteria
Synergistota
Spirochaetota
Bacteroidota
Desulfobacterota
Desulfobacterota_I
Chloroflexota
Genome size (Mb)
b
Phylum (Patescibacteria)
GTDB 120 single markers
Present
c
Actinobacteriota
Bacteroidota
Campylobacterota
Chloroflexota
Deinococcota
Desulfobacterota
Desulfobacterota_I
Elusimicrobiota
Firmicutes
Firmicutes_A
Firmicutes_B
Firmicutes_C
Fusobacteriota
Methanobacteriota
Myxococcota
Patescibacteria
Proteobacteria
Spirochaetota
Synergistota
Thermoplasmatota
Verrucomicrobiota
DdlB FtsA FtsI FtsW FtsZ MraW MraY MraZ MrdA MurA MurB MurC MurD MurE MurF MurG RodA
d
% of transmembrane genes
Patescibacteria
Spirochaetota
Firmicutes_A
Firmicutes
Synergistota
Actinobacteriota
Campylobacterota
Firmicutes_B
Firmicutes_C
Proteobacteria
Fusobacteriota
Bacteroidota
e
94.5% 1% O-methyltransferase(K05303)
96.5% 4.2% exodeoxyribonuclease I(K01141)
93.3% 2% uncharacterized protein(K06884)
96.3% 7.4% type IV pilus assembly protein PilM(K02662)
96.4% 8.2% resuscitation-promoting factor RpfB(K21688)
90.7% 5.4% phosphomannomutase(K17497)
92.2% 14.7% uncharacterized protein(K09775)
95% 17.7% conjugal transfer ATP-binding protein TraC(K12063)
80.2% 5.2% adenylate cyclase, class 2(K05873)
79.2% 19.6% arginine deiminase(K01478)
56.3% 1.5% mannosyltransferase(K14340)
56.1% 3.4% ribosomal protein S6--L-glutamate ligase(K05844)
62.3% 17.9% xylulose-5-phosphate/fructose-6-phosphate phosphoketolase(K01621)
52.4% 14% polysaccharide biosynthesis protein VpsQ(K20950)
Patescibacteria
Other phyla
f
97.3% 1.8% Involved in Type IV secretory pathway (VirB6)(COG3704)
99.1% 5.3% tRNA G10 N-methylase(Trm11)(COG1041)
97.1% 4.2% Exonuclease I (SbcB)(COG2925)
93.3% 1.5% General stress protein YciG(GsiB)(COG3729)
96.5% 10.4% Type IV pilus assembly protein(PilM)(COG4972)
97% 12.7% Uncharacterized conserved protein(YacL)(COG4956)
94.8% 15.4% Uncharacterized conserved protein(YrrD)(COG3881)
81% 2.1% Adenylate cyclase class IV(CyaB)(COG1437)
92.5% 14.9% Acid phosphatase family membrane protein(YuiD)(COG1963)
96.4% 19% Spore cortex hydrolase YaaH(SleL)(COG3858)
78.7% 1.6% predicted RNA-binding domain(ASCH)(COG4043)
79% 5.2% Predicted kinase, contains AAA domain(COG0645)
79% 17.5% PMT family glycosyltransferase(ArnT)COG1807)
79.4% 17.9% Arginine deiminase(ArcA)(COG2235)
52% 0.7% Uncharacterized conserved protein(COG4067)
58.3% 7.5% Protein may be involved in cell wall binding(Psp3)(COG5479)
56.5% 8.3% Membrane protein PF0508(COG5305)
65.9% 18% Predicted membrane transporter(YfcC)(COG1288)
50.4% 5.3% Predicted Zn-dependent protease(YjaZ)(COG5504)
62.6% 18.1% Phosphoketolase (XFP)(COG3957)
Patescibacteria
Other phyla
g
cation-transporting P-type ATPase E (K12952)
Exonuclease I (degrades ssDNA)(SbcB)(COG2925)
ABC-type uncharacterized transport system YhaQ(COG4152)
Uncharacterized conserved protein YacL(COG4956)
3'-5' exoribonuclease Orn(COG1949)
oligoribonuclease(K13288)
O-methyltransferase(K05303)
ABC-type Na+ efflux pump, permease component NatB(COG1668)
aminopeptidase N(K01256)
Putative cell wall-binding domain LytB(COG2247)
Glycerol uptake facilitator aquaporin GlpF(COG0580)
General stress protein YciG,GsiB(COG3729)
Uncharacterized conserved protein YrrD(COG3881)
uncharacterized protein(K06884)
resuscitation-promoting factor RpfB(K21688)
phosphomannomutase(K17497)
Malate/lactate dehydrogenase Mdh(COG0039)
Glycopeptide antibiotics resistance protein VanZ(COG4767)
Predicted ABC-type transport system YbbP(COG3127)
Carbamate kinase ArcC(COG0549)
L-lactate dehydrogenase ldh(K00016)
ornithine carbamoyltransferase(K00611)
Arginine deiminase ArcA(COG2235)
ABC-type polar amino acid transport system, ATPase component GlnQ(COG1126)
ABC-type amino acid transport system, permease component HisM(COG0765)
polar amino acid transport system permease protein(K02029)
membrane protein HdeD(K21908)
polar amino acid transport system ATP-binding protein(K02028)
polar amino acid transport system substrate-binding protein(K02030)
ABC-type amino acid transport system, periplasmic component HisJ(COG0834)
ASC-1 homology (ASCH) domain, predicted RNA-binding domain (ASCH)(COG4043)
Predicted kinase, contains AAA domain (COG0645)
Ornithine carbamoyltransferase ArgF(COG0078)
anaerobic ribonucleoside-triphosphate reductase activating protein nrdD(K04068)
Uncharacterized membrane protein(COG4905)
cetate kinase AckA(COG0282)
Predicted membrane transporter YfcC(COG1288)
Organic radical activating enzyme NrdG/QueE(COG0602)
Phosphoketolase XFP(COG3957)
Lipopolysaccharide biosynthesis regulator LapB(COG2956)
CYTH domain, found in class IV adenylate cyclase(COG2954)
Phosphotransacetylase(COG0280)
LGFP repeat-containing protein, may be involved in cell wall binding(Psp3)(COG5479)
ribosomal protein S6--L-glutamate ligase(K05844)
uracil-DNA glycosylase(K03648)
queuosine precursor transporter(K09125)
Gas vesicle protein YhaH (GvpP)(COG4980)
Predicted DNA-binding protein MmcQ(COG2315)
Environment
Prevalence (%)
MHA
h
***
No. of ABC Transporter
MHA
Environment

1546

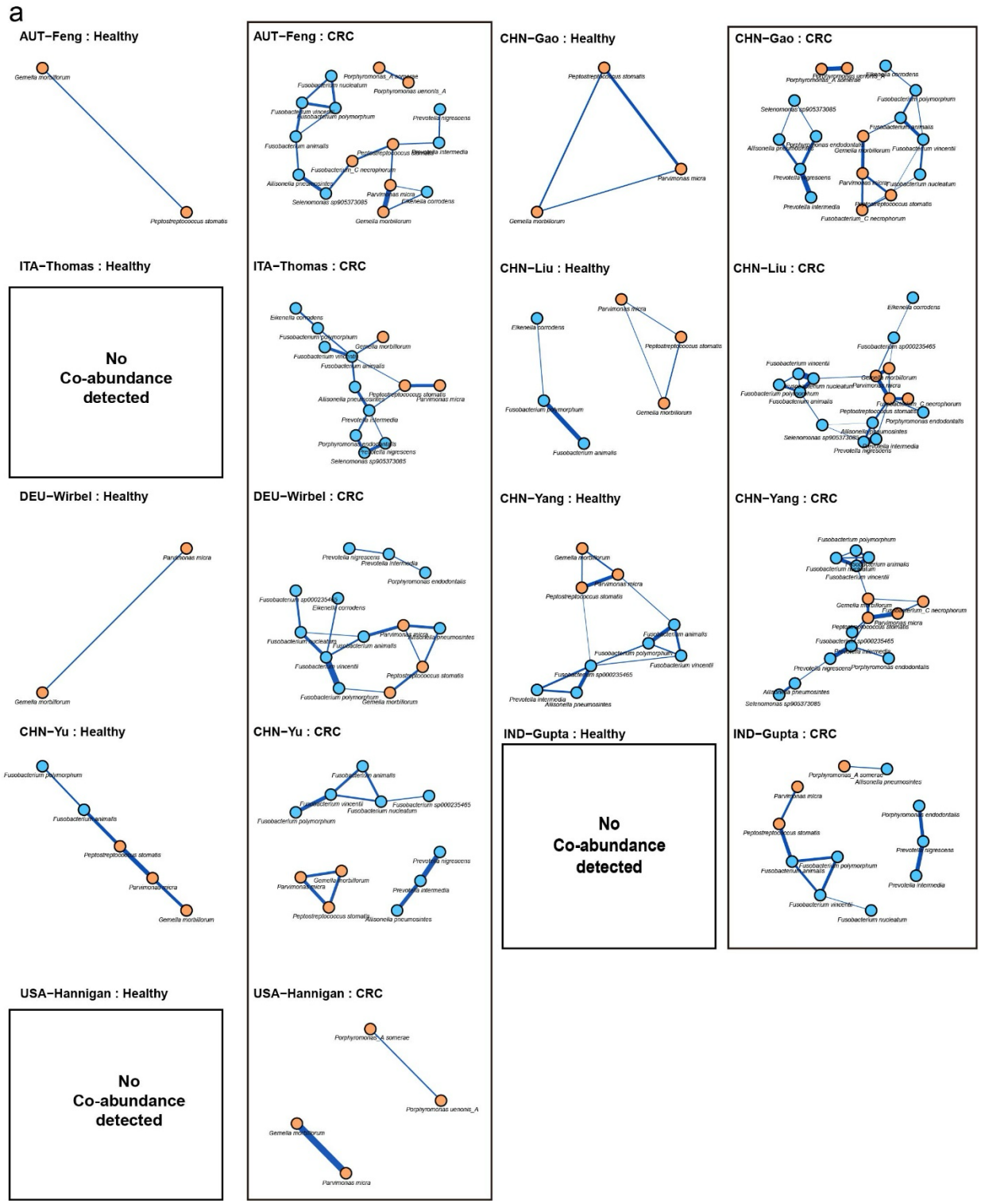
1547 **Extended Data Figure 5 | Comparative genomic and functional analyses of Patescibacteria. a**, Ridge plot showing
1548 HROM representative genome size distribution across major phyla, with Patescibacteria exhibiting the smallest
1549 genome sizes. **b**, Heatmap of GTDB 120 single marker gene presence across 3,426 HROM representative species
1550 genomes. The left color strip indicates the phylum of each species, while the column represents the GTDB 120 single
1551 markers. A red strip identifies the phylum Patescibacteria. **c**, Heatmap showing the presence of division and cell wall
1552 (*dcw*) genes across phyla. Each row corresponds to a phylum, and columns represent specific *dcw* genes. The
1553 percentage inside each box denotes the prevalence of corresponding gene for each phylum. **d**, Violin plot displaying
1554 the proportion of transmembrane genes across phyla, with Patescibacteria showing higher distribution compared to
1555 other phyla. **e-f**, Table listing the top enriched KEGG orthologs (**e**) and COG terms (**f**) in Patescibacteria genomes
1556 compared to other phyla, with percentages indicating prevalence within Patescibacteria and the relative enrichment
1557 compared to other phyla. **g**, Lollipop chart showing the prevalence of functional terms in Patescibacteria genomes,
1558 comparing mammalian-host-associated (MHA) and environment species. Blue points indicate prevalence in
1559 environment species, while violet points indicate prevalence in MHA genomes. **h,** Violin plot comparing the number
1560 of ABC transporter protein in MHA versus environment Patescibacteria genomes, with MHA genomes showing
1561 significantly more transporters. Statistical significance was assessed using a two-tailed Mann-Whitney U test (*** P
1562 < 0.001).

1563


a
Protein-based cluster
Family
cluster
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
cluster_6
cluster_7
cluster_8
cluster_9
cluster_10
cluster_11
cluster_12
cluster_13
cluster_14
cluster_15
cluster_16
cluster_17
cluster_18
c
UMAP2
UMAP1
Genus
Nanosynbacter
Saccharimonas
b
Protein-based cluster
Genus
Protein-based cluster
cluster 4
cluster 10
cluster 11
cluster 15
cluster 16
cluster 18
Saccharimonas
cluster 16
Nanosynbacter
d
prevalence in cluster 16
mean prevalence in
other Patescibacteria cluster
fam08340
fam10393
fam36450
fam13970
fam18274
fam35827
fam76426
fam61608
e
UMAP2
UMAP1
Species with fam08340
No
f
AF-A0A3C0BB44-F1-model_v4 match region
fam08340(fam08340|HROM_Genome_3728|CDS|103) protein
g
Protein taxonomy
Archaea
CPR group
phylum Bacteroidota
phylum Proteobacteria
Uncharacterized (AF−A0A7H8Z009−F1−model_v4)
Uncharacterized (AF−A0A660M2I6−F1−model_v4)
Uncharacterized (AF−A0A258BR84−F1−model_v4)
Uncharacterized (AF−A0A258GG67−F1−model_v4)
PKD domain−containing protein (AF−A0A3C0BB44−F1−model_v4)
PKD domain−containing protein (AF−A0A5N1J817−F1−model_v4)
PKD domain−containing protein (AF−A0A2N2WK84−F1−model_v4)
Uncharacterized (AF−A0A7Y6XPB0−F1−model_v4)
Uncharacterized (AF−A0A2M8CCW3−F1−model_v4)
T9SS type A sorting domain−containing protein (AF−A0A838N957−F1−model_v4)
Por_Secre_tail domain−containing protein (AF−A0A1F3PGW6−F1−model_v4)
Uncharacterized (AF−A0A1F3NZQ4−F1−model_v4)
Por_Secre_tail domain−containing protein (AF−A0A4P5TPP4−F1−model_v4)
Uncharacterized (AF−A0A497DJ51−F1−model_v4)
Uncharacterized (AF−A0A841RTC9−F1−model_v4)
counts of AFDB50 hits for all fam08340 proteins of HROM

1564

Extended Data Figure 6 | Phylogenetic and functional insights into Patescibacteria clusters. a, Phylogenetic tree highlighting clusters of 2,963 Patescibacteria species based on the occurrence profile of protein families at 50% sequence identity. The first strip indicates the Family, while the second strip represents clusters based on protein family presence/absence. **b**, Detailed phylogenetic view of 402 species from Patescibacteria clusters 4, 10, 11, 15, 16, and 18, with annotations indicating associated genera (*Saccharimonas* and *Nanosynbacter*). **c**, UMAP visualization of Patescibacteria species based on protein family profiles. Species belonging to the genera Nanosynbacter and Saccharimonas are color coded. **d**, Scatterplot displaying the prevalence of protein families in cluster 16 (y-axis) compared to its mean prevalence across other Patescibacteria clusters (x-axis). **e**, UMAP visualization showing species containing the enriched protein family fam08340 (red points). **f**, Predicted protein structure of a representative fam08340 protein using ESMFold, with highlighted functional match regions. **g**, Bar plot illustrating structural search hits using Foldseek. The bars represent the number of proteins with hits, while the vertical line indicates the total number of proteins in fam08340.



Extended Data Fig. 7 | Co-abundance networks of 17 oral species among top 20 important bacterial species for colorectal cancer (CRC) classification. a, Each panel represents the co-abundance relationships among the 17 oral species including oral-gut shared (orange) and oral-gut transmitter (blue) species from the top 20 important species for CRC classification. Edges indicate significant co-abundance relationships. Co-abundance was more frequently detected in CRC samples compared to healthy control samples, highlighting the altered microbiome interactions in

1584 CRC. Cohorts where no significant co-abundance relationships were detected in healthy samples are marked as "No
1585 Co-abundance Detected." This pattern emphasizes the association of oral bacterial species with CRC-associated gut
1586 microbiome dysbiosis.

1587