Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity.

Daniel R. Utter[1]*, Gary G. Borisy[2], A. Murat Eren[3,4], Colleen M. Cavanaugh[1]*, Jessica L. Mark Welch[3]*

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA; dutter@g.harvard.edu
[2]The Forsyth Institute, Cambridge, MA 02142, USA
[3]The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543
[4]Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA

*Address correspondence to jmarkwelch@mbl.edu or dutter@g.harvard.edu

Abstract

Background: The increasing availability of microbial genomes and environmental shotgun metagenomes provides unprecedented access to the genomic differences within related bacteria. The human oral microbiome with its diverse habitats and abundant, relatively well-characterized microbial inhabitants presents an opportunity to investigate bacterial population structures at an ecosystem scale.

Results: Here, we employ a metapangenomic approach that combines public genomes with Human Microbiome Project (HMP) metagenomes to study the diversity of microbial residents of three oral habitats: tongue dorsum, buccal mucosa, and supragingival plaque. For two exemplar taxa, *Haemophilus parainfluenzae* and the genus *Rothia*, metapangenomes revealed distinct genomic groups based on shared genome content. *H. parainfluenzae* genomes separated into three distinct subgroups with differential abundance between oral habitats. Functional enrichment analyses identified an operon encoding oxaloacetate decarboxylase as diagnostic for the tongue-abundant subgroup. For the genus *Rothia*, grouping by shared genome content recapitulated species-level taxonomy and habitat preferences. However, while most *R. mucilaginosa* were restricted to the tongue as expected, two genomes represented a cryptic population of *R. mucilaginosa* in many buccal mucosa samples. For both *H. parainfluenzae* and the genus *Rothia*, we identified limitations in the ability of cultivated organisms to represent populations in their native environment, as the environmental distribution of cultivar gene sequences ranged from absent to ubiquitous.

Conclusions: Our findings provide insights into population structure and biogeography in the mouth and form specific hypotheses about habitat adaptation. These results illustrate the power of combining metagenomes and pangenomes to investigate the ecology and evolution of bacteria across analytical scales.

Keywords: oral microbiome, metagenomes, pangenomes, *Rothia*, *Haemophilus parainfluenzae*, population structure, biogeography

Introduction

The human microbiome encompasses tremendous microbial diversity. The growing recognition of this diversity and its importance for human well-being has prompted detailed investigation into the identity and distribution patterns of bacteria throughout the human body (Human Microbiome Project Consortium, 2012). More recent studies have focused on finer-scale patterns, such as the role of host individuality in determining microbiome composition, the number and diversity of strains that can co-exist within a habitat, and the distribution of strains across body sites (Lloyd-Price et al., 2017; Pasolli et al., 2019; Tierney et al., 2019). However, the sheer numbers and genetic diversity of bacteria in even a simple real-world microbiome present significant challenges to study.

One approach to studying bacterial populations is metagenomics --the direct sequencing of the total DNA obtained from an environmental sample (Quince et al., 2017). By circumventing the need for cultivation, metagenomics can afford deeper and more accurate insights into the genetic diversity of naturally occurring microbial populations (Chen et al., 2019). The Human Microbiome Project (HMP; HMP Consortium, 2012; Lloyd-Price et al. 2017) sequenced metagenomes from hundreds of samples from sites all over the human body. However, the use of metagenomic methods alone can be limited by the challenges inherent in associating short reads back to a single organism without combining sequences from separate strains (Nielsen et al., 2014; Quince et al., 2017).

On the other hand, the genomic diversity and relatedness within a group of bacteria can be studied using pangenomes. The pangenome, the sum of all genes found across members of a given group, reveals the functional essence and diversity held within that group (Medini et al., 2005; Tettelin et al., 2005; Vernikos et al., 2015). Pangenomics can identify core and accessory genes (genes shared and not shared by all, respectively) within a group of related bacteria, as well as the relationships between different bacteria based on shared gene content. Notably, relating genomes by gene content allows a phylogenetically-naïve approach to compare genomes, so that any phylogenetic or ecological correlation that emerges from the comparison is informative (Delmont & Eren, 2018; Snel et al., 1999). Because concepts of species pose

challenges when working with bacteria, bacterial pangenomes may be generated at the genus or family level to illuminate gene sharing and the degree of relatedness within these larger groupings (Cornejo et al., 2013; Simon et al., 2017). However, the environmental distribution of groups and genes identified in the pangenome remain unidentified.

Combining pangenomes and metagenomes offers a novel perspective into the adaptation of microbial populations to different habitats. Pangenomes and metagenomes are complementary in that the pangenome reveals the genomic population structure of the different organisms, while the metagenome can reveal the environmental distribution of those same genes and genomes permit identification of biogeographic patterns. Particularly, metagenomes allow the estimation of a given gene sequence variant's abundance in an environmental population through short-read mapping, which avoids the limitations of both cultivation and assembly (Donati et al., 2016; Truong et al., 2017; Yassour et al., 2018). Thus, by using a set of well-characterized genomes from related strains (i.e., a pangenome of cultivars) as a reference set to recruit reads from metagenomes spanning a variety of habitats, the relative frequency of each gene sequence in naturally occurring populations can be quantified. The ability of short-read mapping algorithms to map related but non-identical reads can be exploited to use reference genomes as reference points to probe the composition and structure of wild populations (Denef 2019). This combination of metagenomes with pangenomes, referred to as 'metapangenomics' (Delmont & Eren, 2018; Dvořák et al., 2014) reveals the population-level results of habitat-specific filtering of the pangenomic gene pool.

The oral microbiome is an ideal system in which to investigate microbial population structures in a complex landscape due to its tractability and the large body of existing work. A curated database, the Human Oral Microbiome Database (HOMD; Dewhirst et al., 2010), has identified the bacteria that are resident as opposed to transient in the mouth; most of the organisms (70%; 527 of 755 taxa) have been brought into culture, and genomes are available for 90% of these (472 of 527 taxa). Further, the oral microbiome is composed of multiple distinct sub-habitats in close communication with each other. The tongue dorsum, buccal mucosa, tooth, and other surfaces within the mouth constitute distinct habitats, each with a characteristic

85 microbiome (Aas et al., 2005; Eren et al., 2014; Hall et al., 2017; Segata et al., 2012). Below the genus level, the microbiota at each site are surprisingly distinct and show a strong differential abundance between oral habitats (Eren et al., 2014) despite the ready communication of the habitats with each other through salivary flow and mastication. This differential abundance is likely the result of differential success of microbes due to specialization for specific habitats
90 within the mouth, an idea that has been recently reviewed as the "site-specialist hypothesis" (Mark Welch et al., 2019). However, some apparent "habitat generalist" species span the different oral habitats, such as *Haemophilus parainfluenzae* (Gammaproteobacteria: Pasteurellales) which can be found throughout the mouth (Eren et al., 2014; Segata et al., 2012). The Human Microbiome Project (HMP) sequenced hundreds of habitat-resolved oral
95 metagenomes from healthy subjects, allowing investigation of native bacterial populations across many different mouths (HMP Consortium, 2012; Lloyd-Price et al., 2017). Some population genetics-based analyses of HMP metagenomes have suggested that *H. parainfluenzae* is composed of separate habitat-associated subgroups (Costea et al., 2017; Lloyd-Price et al., 2017), but genomic characteristics of these groups remain unidentified.

100 Here, we investigate the genome content of bacteria from the oral cavity and how genes are distributed across populations at distinct sites within the mouth. We focused on two prevalent and abundant commensal oral taxa, the species *Haemophilus parainfluenzae* and the genus *Rothia*. As the basis for analysis of natural populations, we constructed pangenomes using genome sequences from cultured isolates. We then investigated the degree to which each
105 gene in the pangenome is represented in the healthy human mouth, using HMP metagenomic data. We found that genomes can be clustered into distinct, nested genomic groups that show differences in abundance between habitats. Our results suggest a framework where bacteria are structured into multiple cryptic subpopulations, some of which match observed habitat boundaries.

110 **Results**

Metapangenome workflow and the environmental core/accessory designation. A metapangenome provides an overview of how genes are distributed across reference genomes

and across metagenomes. A conceptual schematic for how isolate genomes and oral metagenomes are combined into a metapangenome is shown in Figure 1 and Additional file 1. Construction of the pangenome rests on the definition of gene clusters, groups of genes that are close to one another in sequence space at the amino acid level (Figure 1A parts 1, 2). The presence or absence of gene clusters in individual genomes can be displayed so that sets of homologous genes are easily identified that are shared by all genomes, shared by subsets of genomes, or unique to a particular genome (Figure 1A part 3). In parallel with the construction of the pangenome at the amino acid level, the distribution of each gene within the human mouth is assessed at the nucleotide level by mapping metagenomic reads against the entire collection of cultivar genomes (Figure 1B part 1). The result of mapping is a value for "depth of coverage," hereafter simply "coverage," the number of metagenomic short reads that were mapped to a given nucleotide in a given genome; the coverage value serves as an estimate of the abundance of the gene in the sample. The coverage for each gene in a genome, for each of a large number of samples, can then be shown as a circular bar chart (Fig. 1B part 2), with concentric rings showing the coverage recruited from each metagenomic sample.

Comparative abundance in natural habitats can then be assessed for each gene using a metric to determine whether genes are core or accessory to an environment, rather than to a set of genomes (Delmont & Eren 2018). This is accomplished by relating the abundance of a gene to the abundance of the genome that carries it, with respect to a set of environmental samples (Fig. 1B part 3). By this metric, a gene in an isolate genome is considered environmentally "core" if its median coverage, across all mapped metagenomes, is at least one-fourth the median coverage of the genome in which it resides. The gene is environmentally "accessory" if its coverage falls below this cutoff. The environmental core/accessory metric thus provides a way of assessing the contribution of the gene to population structure in the environment – provided that the pangenome adequately represents the nucleotide sequences of genes found in the population. If the survival of a microbial cell in the environment under study depends on having this gene in its genome, the gene should register as environmentally core, while if the gene were dispensable or required in only a subset of the cells of the population, the gene would register as environmentally accessory.

The metapangenome (Figure 1C) combines the pangenome with a summary of the mapping information. The outermost concentric ring of the metapangenome, here colored in red and blue, summarizes the environmental core/accessory metric across all genes in a gene cluster as a stacked bar chart with the heights normalized to the number of genes in that gene cluster. The scale of this outer ring thus changes from one part of the ring to another, as the number of genes per gene cluster ranges from one (as is the case between 10 o'clock and 12 noon on Fig. 1C) to three in the case of this example as shown between 3 o'clock and 8 o'clock on the figure. Thus, the metapangenome format summarizes the cultivar genome data in a visual format that emphasizes sets of shared or unique genes, and then summarizes the metagenomic data in the form of an environmental core/accessory metric for each of the genes in this pangenome, assessed across all the mapped metagenome samples.

The apparent generalist *H. parainfluenzae* is composed of multiple subgroups

The species *H. parainfluenzae* is an apparent oral generalist in that it is both abundant and prevalent at multiple sites within the human mouth (Lloyd-Price et al. 2017, Mark Welch et al. 2019); however, previous reports have suggested that genomically distinct sub-populations may exist within the mouth (Lloyd-Price et al. 2017, Costea et al. 2017). To investigate the genome structure of the global *H. parainfluenzae* population as represented by the sequenced cultivated strains, we downloaded thirty-three high-quality isolate genomes from NCBI RefSeq and constructed a pangenome (Figure 2, Additional file 2). Inspection of this pangenome shows a large core genome of 1,493 gene clusters, shown as the continuous black bars between 9 o'clock and 12 o'clock in Fig. 2. The dendrogram in the center of the figure organizes the gene clusters according to their presence/absence across genomes, and thereby visually separates the core genome from the accessory genome. The accessory genome consists of 923 gene clusters unique to a single isolate genome, shown on the figure between 3 and 5 o'clock, and 1,902 gene clusters shared by some but not all isolate genomes, shown between 5 o'clock and 9 o'clock on the figure.

When the genomes themselves are clustered according to the number and identity of gene clusters that they share, they segregate into three groups (Groups 1-3) that are distinguished by

170 shared blocks of gene clusters (Fig. 2). The dendrogram in the upper right of the figure (Fig. 2) shows the clustering topography, and the major branch points in this dendrogram separate the groups. As the number of gene clusters ranges from 1,773 to 2,071 per genome, the core of 1,493 gene clusters represents 72-84% of the gene clusters in each genome and the gene clusters found in only a single genome contribute up to an additional 5%. Thus, collectively, the
175 blocks of gene clusters that characterize the subgroups of *H. parainfluenzae* constitute a relatively small fraction of the genome.

Haemophilus parainfluenzae subgroups are habitat-specific

Mapping of metagenomic data onto the genomes shows that the groups defined by genome content have significantly different distributions among oral sites ($p < 0.001$, ADONIS test of
180 median HMP coverages). We applied the competitive recruitment approach to the billions of short reads sequenced by the Human Microbiome Project (HMP) (HMP Consortium, 2012; Lloyd-Price et al., 2017) for hundreds of healthy individuals for three different oral habitats (tongue dorsum, TD; buccal mucosa, BM; and supragingival dental plaque, SUPP). The mapping information is summarized in the heatmap shown in the upper right corner of Figure 2. For
185 each oral habitat, the median metagenome's coverage of each genome is displayed in the colored bar charts below the heatmap.

Comparison of the pangenome groups with HMP coverage data shows that the middle group of genomes, Group 2, is much more abundant in tongue dorsum than genomes in the other two groups (Figure 2 heatmap, median coverages). The median coverage for the nine genomes in
190 Group 2 is several-fold higher than the other twenty-four genomes (Figure 2). By contrast, dental plaque metagenomes map with higher coverage to the genomes in Group 3 (outermost group), whereas buccal mucosa metagenomes map with similar coverage to all three groups (Fig. 2). Thus, genomically-defined subgroups of *H. parainfluenzae* have differential abundance across oral habitats, as reflected in differing levels of coverage of these genomes by
195 metagenomic reads. Of these, the TD-abundant group of nine genomes appears most distinct.

Analysis of the phylogenetic relationships among *H. parainfluenzae* genomes, based on single-copy genes, revealed that groups defined by genome content differed from those defined by evolutionary relatedness at the strain level. We constructed a phylogeny based on nucleotide sequences from 139 bacterial genes previously identified as present in a single copy in most genomes (Campbell et al., 2013). This phylogeny placed the TD-associated genomes in separate clades of the *H. parainfluenzae* tree (Additional file 3A). Two additional methods of assessing similarity, using 16S rRNA gene sequences and whole-genome kmer comparisons, provided little phylogenetic signal but indicated substantial nucleotide-level divergence among strains, respectively (Additional file 3B). Thus, these analyses suggest that tongue-associated *H. parainfluenzae* share genome content and ecology, but do not share a unique common ancestor to the exclusion of other *H. parainfluenzae* strains.

Genomic characteristics of the tongue-enriched *H. parainfluenzae* subgroup

Correspondence between genome content and environmental distribution raises the possibility that presence of certain genes in the genome may be responsible for the success of a particular strain in a given habitat within the mouth. Specifically, we asked whether any genes were particularly characteristic of the nine *H. parainfluenzae* strains with high abundance in TD (Fig. 2, middle group of genomes). Only a small set of genes were present in all genomes of the TD group of *H. parainfluenzae* and not in any of the other isolate genomes; these genes are marked by a dark blue wedge labeled “TD group core” on the figure. We carried out a functional enrichment analysis, as described in Shaiber et al. (2020), to compare the prevalence of predicted functions among TD genomes to their prevalence among non-TD genomes revealed by the metapangenome. This analysis identified exactly three functions in three gene clusters altogether encoding the three subunits of a sodium-dependent oxaloacetate decarboxylase enzyme exclusive to the TD group (Additional file 4). This enzyme converts oxaloacetate to pyruvate while translocating two sodium ions from the cytoplasm to the periplasm, providing a shunt to gluconeogenesis while establishing a potentially useful Na+ gradient (Dimroth & Halpert 1984, Dimroth et al. 2001). No complete oxaloacetate decarboxylase operon was detected in any of the other 24 *H. parainfluenzae* genomes

(Supplemental Text). No other functions or gene clusters evidenced this universal presence in
225 the TD-associated group but the complete absence from the other *H. parainfluenzae* genomes. Thus, this operon stands out as a strong candidate for further experimental investigation into the source of selective advantage for the group of TD-abundant genomes in the tongue dorsum habitat.

Many *H. parainfluenzae* core gene clusters contain high proportions of gene sequences scored
230 as environmentally accessory, particularly in TD (Figure 2, the spikes of red in the 'Environmental core/accessory' layer; Additional file 2). This result likely stems from differences in nucleotide-level sequence divergence from gene to gene within the population. These core gene clusters do contain sequences that are environmentally core to TD, i.e., the proportion of environmentally accessory sequences in these gene clusters is never 100% (Figure 2). Thus, the
235 traits represented by these core gene clusters are not missing from *H. parainfluenzae* living in the mouth. Further, metagenomic mapping can clearly distinguish between the genomes overall at the nucleotide level, as shown by the differential coverage results by habitat (Figure 2 heatmap and median coverage bar chart). The differential abundance among some core genes' sequence variants thus suggests population-level differentiation between different oral
240 habitats. As the pangenome contains proportionally fewer TD-representative genomes, the environmentally accessory gene sequences (red spikes) are higher in TD than in BM or SUPP. Although the metapangenome can identify gene sequences that are depleted in TD, it cannot discriminate between neutral and adaptive reasons for their differential abundance. Regardless, sequences for many *H. parainfluenzae* core genes are differentially present in
245 certain habitats and may contain signatures of distinct subpopulations.

Pangenomic analysis of oral members of the genus *Rothia*

Having decomposed the species *H. parainfluenzae* into discrete habitat-resolved subpopulations, we applied the same method of analysis to a genus, *Rothia*, that is composed of multiple habitat-specialized species. An advantage of constructing pangenomes at the genus
250 level is that the genus-level core genes, as well as core and accessory genes of the individual species and strain groups, become readily identifiable. To assess the similarity of genome

content among oral species as well as strains within the genus *Rothia*, we downloaded sixty-nine high-quality *Rothia* genomes from NCBI and constructed a genus-level pangenome (Figure 3, Additional file 5). One immediately evident feature of the oral *Rothia* genus pangenome is that the individual genomes segregate based on gene content into three major groups, each of which shares a substantial set of genes that are absent from the others (Fig. 3). Taxonomic designations provided by NCBI (depicted by coloring the genome layers) show that these groups correspond to the three recognized human oral *Rothia* species. A large set of gene clusters (n = 1,131) were present in all of the *Rothia* genomes and represent the genus-level core genome. Given that each *Rothia* genome contains between 1,693 and 2,252 gene clusters, the genus core represents a significant portion of any given *Rothia* genome. Other sets of gene clusters were characteristic of and exclusive to *R. mucilaginosa* (n = 206), *R. dentocariosa* (n = 266), or *R. aeria* (n = 454). Taking the species-level core genes into account, the three *Rothia* species were identified unambiguously by their conserved gene content, with 77%, 77%, and 81% of the median *R. mucilaginosa*, *R. dentocariosa*, and *R. aeria* genomes, respectively, occupied by genus- or species-level core genes. The remaining ~20% of each genome represents accessory genes that were present in one or more genomes but not in all genomes of the species. Thus, for the genus *Rothia*, pangenomic analysis recapitulated species designations.

Genomes within a single group also form subgroups. The major group of *R. mucilaginosa* (Rm) genomes can be subdivided into two subgroups defined by 39 gene clusters exclusive to the larger subgroup (grey line 'Rm1' in Fig. 3) and 38 gene clusters exclusive to the smaller subgroup (grey line 'Rm2' in Fig. 3). Genomes deposited in NCBI with only a genus-level designation (i.e., *Rothia* sp.; black layers in Figure 3) also fell into each *R. mucilaginosa* subgroup, increasing confidence in the discreteness of the subgroups. To investigate whether these gene clusters were localized to a single region, as could result from, e.g., a phage insertion, or whether they were scattered through the genome, we reordered the gene clusters (columns) to follow the genome order in an arbitrary *R. mucilaginosa* group 1 genome (Additional file 6). The gene clusters present only in group 1 do not localize to a single region but are scattered throughout the chromosome (Additional file 6), suggesting that the

differentiation between the groups is not the result of a single recent chance event and may be instead the result of ecological and evolutionary pressures. Thus, *R. mucilaginosa* is comprised of at least two cryptic subgroups.

Metagenomic mapping reveals habitat preferences of genome groups

285 Metagenomic mapping to the *Rothia* genomes demonstrated that the different genomic groups occupy different environments within the mouth. We carried out competitive mapping to the set of *Rothia* spp. cultivars using the same HMP metagenomic datasets as above. The resulting abundance information is summarized in the coverage heatmap and bar charts in Figure 3 which show that two of the *Rothia* species (*R. aeria* and *R. dentocariosa*) were most abundant
290 in SUPP samples. The third species, *R. mucilaginosa*, was highly abundant in TD and for the most part displayed only negligible coverage from SUPP and BM samples. Two outliers were also apparent – one *R. dentocariosa* genome that recruited high coverage from BM and TD metagenomes, and two genomes in the *R. mucilaginosa* group that received high coverage from approximately one-third of BM metagenomes.

295 Whereas the heat map summarizes coverage information for each genome and metagenome as a single data point, the mapping analysis provides finer-grained information about the frequency of each gene sequence in natural populations by relating the abundance of each gene to the abundance of the genome that carries it. The three outer multicolored layers of Fig. 3 summarize the outcome of this analysis with the *Rothia* genus pangenome for SUPP, BM, and
300 TD samples. If a cultivar genome does not receive enough coverage in a habitat to be considered "detected", i.e. if more than half of a genome's nucleotides received no coverage in every metagenome from a habitat, then the result for genes from that genome is shown in gray rather than in color to indicate that the environmental core/accessory status could not be assessed.

305 Mapping results, as summarized by the environmental core/accessory metric, reinforced the conclusion from the coverage heatmap that the genomic groups corresponding to different named *Rothia* species occupied different habitats within the mouth. The *Rothia* genus core

genes were environmentally core in all habitats except where their surrounding genome was not detected – which occurs because many of the *R. mucilaginosa* genomes were undetectable in BM and SUPP, and many of the *R. dentocariosa* genomes were undetectable in TD. In contrast, species-specific core genes were only environmentally core to specific habitats. The core genes unique to *R. mucilaginosa* ('Rm' gene clusters, Figure 3) were environmentally core in TD but their parent genomes were not detected in SUPP and BM – with the exception of the two *R. mucilaginosa* genomes that were abundant in BM and one that passed the detection threshold in SUPP (thin purple and green lines in Fig. 3). Conversely, the core genes unique to *R. dentocariosa* ('Rd', Figure 3) were environmentally core in SUPP and BM but their parent genomes for the most part were not detected in TD. Thus, these two species show distinct and complementary habitat preferences. *R. aeria* genomes behaved differently: they were detected in SUPP, BM, and TD and while their unique core genes ('Ra' gene clusters, Figure 3) were environmentally core at all three sites, they attained significantly more coverage from SUPP than from TD or BM (Figure 3, "median coverages") reflecting a bias towards SUPP. Thus, the core genes of *Rothia* species can distinguish their distinct habitat ranges.

R. mucilaginosa genomes divide into subgroups

Subgroups can be distinguished within major species by presence and absence of sets of gene clusters, and mapping of metagenomic reads can be used to assess whether these within-species groups have similar distribution patterns in the sampled oral habitats. The large group of *R. mucilaginosa* genomes can be subdivided at the pangenome level into two subgroups that differ by a small set of core genes ('Rm1' and 'Rm2' in Figure 3). Their genome-scale abundance as assessed by mapping is similar, with both recruiting coverage primarily from TD metagenomes (Fig. 3 heat map) and detected primarily in TD (Fig. 3 environmental core/accessory layers). Similarly, at the finer, gene level of mapping resolution both the *R. mucilaginosa* group 1 core genes and the *R. mucilaginosa* group 2 core genes were environmentally core in TD. Further, both subgroups obtained high coverage from many HMP samples. Thus, these two *R. mucilaginosa* subgroups do not appear to be the result of a broad habitat shift such as from tongue to teeth or buccal mucosal sites. Instead they may represent

co-existing subpopulations, perhaps with distinct microhabitats within the TD community or between individual mouths.

We also detected evidence for habitat shifts between major habitats by a small number of genomes, in the form of outlier results in the mapping of human oral metagenomes to *Rothia* cultivar genomes. These outliers consist of the two *R. mucilaginosa* genomes that recruited high coverage from BM metagenomes, as well as *R. dentocariosa* OG2_1 that recruited high coverage from BM and TD metagenomes (heat map, Fig. 3). To investigate whether *R. dentocariosa* OG2_1 represents a real population in BM and TD, we investigated the evenness of coverage across OG2_1 genes. We discovered that the majority of the OG2_1 genome was evenly well-covered across SUPP metagenomes but largely undetected in TD except for a single, deeply covered gene, with BM having intermediate coverage and detection between that of SUPP and TD (Additional file 7). Thus, we conclude that *R. dentocariosa* OG2_1 best represents a SUPP-dwelling population like other *R. dentocariosa*. On the other hand, the two *R. mucilaginosa* genomes abundant in BM satisfied the detection metric in BM and plaque, and the genes shared only by these two genomes (labeled RmBM in Fig. 3) were environmentally core in BM. This distribution provides evidence that a bacterium similar to *R. mucilaginosa* is in buccal mucosa samples at high enough abundance to satisfy the detection metric, and that sequences from this bacterium find their closest match in these two *R. mucilaginosa* genomes in our set of sequenced cultivars.

Two *Rothia mucilaginosa* genomes represent a BM-adapted subpopulation

To assess how well the outlier *R. mucilaginosa* genomes with high coverage and detection in BM represent a true buccal mucosa *Rothia* community, we inspected the coverage of one of the two outlier genomes, *R.* sp. E04, in more detail at the gene level (Fig. 4). The BM and SUPP metagenomes covered the majority of this genome's genes relatively evenly, as expected for samples containing populations related to E04 (Fig. 4A). However, this pattern was not observed with TD metagenomes (Figure 4A, inner 30 rings); instead the coverage from TD metagenomes was dense in some regions and absent in others. Similar analysis of the other BM-abundant genome, *R.* sp. C03, revealed similar results (Additional file 8). This pattern

suggests the TD coverage results from spurious mapping of isolated regions of high similarity or
365 mobility, e.g. phage elements. Particularly, the intermittent TD coverage at both the gene level (Fig. 4A) and the nucleotide level (Fig. 4B) indicates these samples do not contain a population closely related to *R.* sp. E04. If the reads from TD metagenomes that do map originate from phylogenetically related but non-*Rothia* or non-*R. mucilaginosa* populations, their evolutionary distance should produce higher densities of single-nucleotide variants relative to the *R.* sp. E04
370 genome. This is observed in Figure 4B where the vertical black lines report the mean Shannon entropy of mapped variants; high levels of entropy indicate variability at that nucleotide position. Thus, inspection of gene-level coverage of outlier isolate genomes validates these genomes as representatives of an *R. mucilaginosa* population adapted to buccal mucosa rather than tongue.

375 The identification of *R.* sp. E04 as the closest match to the BM-dwelling *R. mucilaginosa* population allows investigation into the genomic characteristics of the BM populations. There are 22 gene clusters shared by both genomes abundant in BM but absent from other known *Rothia* isolates. Further, these 22 gene clusters uniquely shared by *R.* sp. C03 and E04 are scattered throughout the genome (Figure 4A genes 'RmBM', Figure 4A genes marked with black
380 tick marks, Additional file 6), suggesting that these two genomes are not related simply by a single large shared insertion event, but that their set of distinctive genes accumulated over time. While these 22 gene clusters do not contain any unique predicted functions, only a handful of them were environmentally accessory in TD but environmentally core elsewhere (bolded gene clusters in Figure 4C) and thus represent prominent candidates for future
385 experimental investigation of their potential role in adaptation of a TD-abundant taxon to the new habitat of BM. However, even in BM and SUPP samples, large contiguous portions of the E04 and C03 genomes recruited little to no coverage relative to the rest of the genome (e.g., filled red arrowhead in Fig. 4A). Thus, although the cultivated strains *R.* sp. E04 and C03 are the best match out of all 69 genomes and provide some insight into the BM-inhabiting *R.*
390 *mucilaginosa* population, the populations native to BM and possibly SUPP likely harbor many additional novel genes. In summary, gene-level mapping reveals features of the distribution of *Rothia* strains that suggest fine-tuned adaptation to each oral site.

Some gene sequences are scarce across all metagenomes while others are intermittently abundant

395 More broadly, visualizing the mapping at the gene scale reveals different patterns of abundance for genes found in a single cultivated genome. Gene-level mapping highlights that some of this genome's sequences for both core and accessory genes may be at low abundance in the sample relative to the *R.* sp. E04 genome across the numerous metagenomes sampled. Many of the environmentally accessory genes are also accessory in the pangenome (genes from accessory
400 gene clusters colored grey in the innermost ring, Figure 4A), but many genes identified as members of the core genome (bright pink; innermost ring) are also environmentally accessory in HMP metagenomes. However, many such genes scored as 'environmentally accessory' are not uniformly low abundance across all individuals. Rather, they are at or below the limit of detection in most samples (i.e., most mouths) but abundant in a few, (e.g., the gene indicated
405 by a filled black arrowhead in Figure 4A). Thus, the existence of a cultivar containing this sequence indicates that cultivation selected a strain with this particular sequence from relatively low prevalence, either stochastically or as a result of cultivation bias. On the other hand, some gene sequences are more or less uniformly rare across samples (e.g., the empty black arrowhead in Figure 4A). Such uniform rarity could reflect their restriction to a low-
410 abundance niche or selection against that sequence in the environment, yet a lineage with that gene survived the bottleneck of isolation. Altogether, these results illustrate how a given cultivar's gene sequences may not be equivalently representative of environmental populations.

Discussion

415 Metapangenomics provides a means of analyzing complex natural populations from a springboard of well-characterized isolate genomes. Starting from the solid foundation of virtually complete genome sequences from cultivated isolates, we constructed pangenomes and then used metagenomic read mapping to analyze the distribution of each gene in the pangenome in wild populations. Our metapangenomic analysis demonstrated that genomes

420 that nominally belong to the same species in fact comprise habitat-specific subgroups within *H. parainfluenzae* and within a species of the genus *Rothia*.

Our metapangenomic findings elucidate the population structure of *H. parainfluenzae* and the genus *Rothia*, revealing differential distribution of species across habitats within the mouth that are within millimeters of one another and are in continual communication via saliva. Such
425 biogeographic distributions have been suggested by prior studies based on cultivation as well as 16S rRNA gene surveys (Costea et al., 2017; Eren et al., 2014; Mark Welch et al., 2019; Wilbert et al., 2019); however, the metapangenomic mapping approach is more comprehensive than cultivation and relies not on a marker gene to define a population but on complete genomes to demonstrate unequivocally the presence of different sets of microbes in the different habitats
430 of supragingival plaque, tongue dorsum, and buccal mucosa.

The finding of habitat-specific subgroups shows that the buccal mucosa, in particular, is colonized by a previously unrecognized, distinctive microbiota. Among the three sampled oral habitats, dental plaque and the tongue dorsum are both characterized by extraordinarily dense, complex, and highly structured microbial communities (Zijnge et al., 2010; Holliday et al., 2015;
435 Mark Welch et al., 2016; Wilbert et al., 2019), whereas the buccal mucosa are more thinly colonized by a generally simpler set of microbes in which the genus *Streptococcus* comprises approximately half the cells. In this context, the *Rothia* spp. detected in 16S rRNA gene surveys on the buccal mucosa could be interpreted as microbes shed from the teeth and tongue and lodged transiently on the buccal mucosa or present in the sample as incidental contaminants
440 from saliva. Our finding of a distinctive mapping pattern in each of the three sampled oral sites rules out this interpretation and indicates that these microbes in fact represent a distinctive subpopulation adapted to the buccal mucosa niche.

Pangenomes that combine metagenome-assembled genomes (MAGs) with reference genomes can offer deeper insight into the gene pool and population structure of environmental microbes
445 (Reveillaud et al., 2019). Given the rapid increase of publicly available MAGs (Almeida et al., 2019; Nayfach et al., 2019; Pasolli et al., 2019), pangenomes of critical populations such as the BM-abundant *R. mucilaginosa* could be dramatically expanded. However, combining MAGs with

cultivar genomes poses some fundamental obstacles. Due to the inherent complexity of metagenome-sampled populations and assembly algorithms, a MAG is at best a consensus genome of closely-related, perhaps clonal, members of a population, as opposed to a cultivar genome from a single cell's clonal lineage that provides comparatively higher confidence that all genes co-occur in the same cell (Nielsen et al., 2014; Quince et al., 2017). While assembly and binning algorithms are improving rapidly, poorly refined MAGs can suffer from significant contamination issues (Chen et al., 2019), which can influence ecological and pangenomic insights (Shaiber & Eren, 2019). Hence, we focused only on high-quality cultivar genomes to benefit from higher confidence in their accuracy.

Niche adaptation at the genomic level

Metapangenomic analysis reveals the differential abundance of genes across habitats and thus presents an opportunity to ask whether the presence or absence of particular genes is key to abundance in a given environment and whether these genes may encode traits under differential selection pressures. Several recent studies have applied functional enrichment or depletion analyses to a pangenome to investigate adaptation to the particular habitats from which those genomes were obtained (Cornejo et al., 2013; Martino et al., 2016; Simon et al., 2017). One recent study used methods conceptually similar to ours, i.e., using metagenomic abundances of reference genes to detect ecologically different subgroups within a population, to identify genes important for determining which *Bacteroides* strains engrafted from human mothers to infants (Yassour et al., 2018). By summarizing metagenomic recruitment across the entire pangenome, we extend such investigations to evolutionary scales, allowing the detection of genomic subgroups with novel niche associations and enabling direct investigation of the frequencies of subgroup-specific genes among environmental populations. A limitation of our approach, however, is that it addresses only the presence and absence of coding sequences in the genome, and cannot identify regulatory regions, structural variants, or other more subtle genomic traces of potential significance for niche adaptation.

Our finding of distinctive subpopulations of *H. parainfluenzae* is consistent with previous studies that have reported that *H. parainfluenzae* may be divided into at least three biotypes

based on classical metabolic phenotyping (Kilian, 1976). Further, two recent studies employed population genetic approaches using metagenomes to detect *H. parainfluenzae* subpopulations with different habitat abundances (Costea et al., 2017; Lloyd-Price et al., 2017). These and our results point towards this apparent generalist population being structured into distinct subpopulations that represent the partitioning of the oral niche. Additionally, our metapangenomic approach identified the genomic subtypes and specific genes that are associated with the TD-abundant subpopulation, the oxaloacetate decarboxylase operon.

The environmental relevance of reference genomes

Ideally, reference genomes for ecological analysis should accurately represent their native environments. However, in practice, most reference genomes are obtained from organisms that have been isolated and subjected to repeated subculture under laboratory conditions. Many organisms are unable to grow under such conditions; those that grow may undergo genomic changes due to the imposition or relaxation of selection pressure under cultivation foreign to their native selective regime. Thus, it is important to evaluate the degree to which existing cultivar genomes serve as suitable references.

In general, although most core and some accessory genes of cultivars were well-represented in the oral environment, a few core and many accessory genes were surprisingly uncommon in the oral environment. Genomic core and accessory genes do not correspond precisely to environmentally core and accessory genes, respectively. Particularly, singleton accessory genes were environmentally accessory, relative to their surrounding genomes. This result is not unexpected, as the set of accessory genes may be very large in an open pangenome whereas the microbiome in any mouth consists of a finite number of strains. Nonetheless, it indicates that the features of any individual strain that are conferred by such accessory genes will be unrepresentative of a community in the mouth. The fact that no single cultivar contains all environmentally-abundant genes exposes the limitations of cultivars for ecological analysis. Cultivars are clearly sufficient to reveal habitat restrictions of individual species (e.g., Figure 3) (Costea et al., 2017; Kraal et al., 2014), but their discordance with bacterial populations in their

native habitats raises questions about the adequacy of any individual strain to serve as a model for understanding microbial community and host-microbe interactions.

We take no position on the species concept

Darwin, recognizing the difficulty in discriminating 'doubtful species' (Darwin, 1859) declined to discuss the various definitions that had been given to the term species in his day, and indeed hoped that science would 'be freed from the vain search for the undiscovered and undiscoverable essence of the term species.' He noted that the amount of difference needed to confer the rank of species was at times 'quite indefinite', that 'no clear line of demarcation' could be drawn between species and sub-species, and therefore the term species was one 'arbitrarily given for the sake of convenience' adding, '…to discuss whether they are rightly called species or varieties, before any definition of these terms has been generally accepted, is vainly to beat the air.'

Metapangenomics does not alter Darwin's general conclusion; rather, it confirms and refines it at the genomic level. Our results indicate that for some taxa, such as the genus *Rothia*, pangenome analysis broadly supports the currently recognized species definitions, while for other taxa such as *H. parainfluenzae* habitat-associated subpopulations are detected that may or may not warrant species-level recognition. Even grouping whole genome sequences by hierarchic clustering based on gene composition results in 'no clear line of demarcation.' Rather, we observed a spectrum of genome cluster relatedness.

Conclusions

In conclusion, our results reveal the detailed association between the environmental distribution and genomic diversity of oral bacterial populations. These patterns reveal that seeming generalist species are composed of cryptic subpopulations and that potentially only a small number of genes are associated with each subpopulation. More broadly, diversification and adaptation to fully exploit available ecological niches is observed at many levels, from recognized species distinguished by many genes down to closely related subpopulations.

Methods

530 Metapangenomes were prepared using publicly-available genomes annotated as belonging to the genus *Rothia*, a gram-positive oral facultative anaerobe in the phylum Actinobacteria; and genomes annotated as the species *Haemophilus parainfluenzae*, a facultative gram-negative anaerobe in the phylum Proteobacteria. A flowchart linking the major methods and analyses is provided in Additional file 1, and a detailed narrative methods with reproducible code is
535 available in the Supplemental Methods.

Genomic and metagenomic data acquisition

Genomes were downloaded from NCBI RefSeq based on the associated names using the assembly summary report obtained from ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/. Sequence names were simplified to
540 contain only the strain identity and a unique contig id for all contigs, and then concatenated into a single FASTA file containing all genomes of interest. This file was used as the starting point for the anvi'o pangenomic analysis (Eren et al., 2015).

For the genus *Rothia*, 69 of the 73 genomes present in RefSeq were used. Of the four not used, *R. nasimurium* was discarded for not being recognized as an oral resident by the HOMD, while
545 *R.* sp. Olga and *R.* sp. ND6WE1A were discarded as non-oral isolates and also for aberrantly large unique gene contents (potentially contaminants); one *R. dentocariosa* genome was discarded for a slightly oversized genome and aberrant mapping profiles (i.e., a handful of genes strongly biased the coverage by recruiting excessive reads). For *Haemophilus parainfluenzae*, all 33 genomes in RefSeq were used for analysis. Metadata from NCBI available
550 for each strain is provided in Additional file 9.

Raw short-read metagenomic data from the Human Microbiome Project (HMP) (HMP Consortium, 2012; Lloyd-Price et al., 2017) were downloaded for tongue dorsum (TD, n= 188), buccal mucosa (BM, n= 169), and supragingival plaque (SUPP, n= 194) using the HMP data portal at https://portal.hmpdacc.org/.

555 ***Metapangenomic workflow***

The metapangenome was constructed for each taxon using anvi'o with methods modified from Delmont et al. 2018 (Additional file 1). Open reading frames (ORFs) were identified on the downloaded contigs using Prodigal (Hyatt et al., 2010) and converted into an anvi'o-compatible database using the command anvi-gen-contigs-db. ORFs were exported and functionally
560 annotated with InterProScan (version 5.30-69) using Pfam, TIGRFAM, ProDom, and SUPERFAMILY (Haft et al., 2001; Bru et al., 2005; Wilson et al., 2008; El-Gebali et al., 2019). Coverage of each taxon's genomes was determined in each HMP metagenome using bowtie2 (Langmead & Salzberg, 2012) with default parameters (--sensitive). Coverage of units encompassing multiple nucleotides, e.g., a gene or genome, is calculated from the per-
565 nucleotide bowtie2 coverages by dividing the sum of nucleotide coverages in that unit by the number of nucleotides, as is the standard method for reporting a unit's coverage. Short reads from each metagenome were mapped against all genomes for that taxon; thus, the bowtie2 matched each read to the best-matching genomic locus, randomly choosing between multiple loci if they were equally best. Thus, coverage at highly conserved regions is affected by the total
570 population abundance in that sample, while the coverage at variable loci reflects that particular sequence variant's abundance. These per-sample coverages were then incorporated into the anvi'o database, and per-ORF coverages and summary metrics (max, min, mean, median) were determined. Single-nucleotide variants (SNVs) were also called per nucleotide if that nucleotide was covered at least 10x.

575 To compute pangenomes we used the anvi'o workflow for pangenomics (see http://merenlab.org/p for a tutorial). Briefly, this workflow uses BLASTP (Altschul et al., 1990) to compute amino acid-level identity between all possible ORF pairs, from which removes weak matches by employing the --minbit criterion (we used the default value, 0.5), then uses The Markov Cluster Algorithm (MCL) (van Dongen & Abreu-Goodger, 2012) to group ORFs into
580 putatively homologous gene groups (gene clusters), and aligns amino acid sequences in each gene cluster using MUSCLE (Edgar, 2004) for interactive visualization. MCL uses a hyperparameter (inflation) to adjust the clustering sensitivity, i.e., the tendency to split clusters.

The *Haemophilus parainfluenzae* pangenome was clustered with an inflation parameter of 10 as the pangenome putatively reflects only a single species' diversity, while the *Rothia* genus pangenome was clustered with an inflation parameter of 6. To estimate the robustness, the inflation parameter was varied by up to ±2 but did not produce significantly different numbers of gene clusters or different core/accessory patterns (Supplementary Table 1). Groups of gene clusters (e.g. gene clusters labelled 'Genus core', 'Rm BM' in Figure 2) were manually selected based on criteria as explained in the results (e.g., found in all *Rothia* genomes). Genome layers were clustered based on gene cluster frequencies, also shown as a dendrogram above the genome layers (e.g., top right of Figure 2, Figure 3). Dendrogram branch length is fixed to an arbitrary constant.

Each gene's environmental core/accessory status in the environment (in this case, the HMP oral metagenomes) was determined for any genomes covered at least 1x over half the genome length in at least one sample. A gene was determined to be environmentally core if that gene had at least one-fourth the median coverage of the genome from which it came (Delmont & Eren, 2018) – that is, if the coverage of a gene recruited from its median metagenome was at least one fourth of the coverage of the genome recruited from its median metagenome. Conversely, a gene was environmentally accessory if it had less than one-fourth the median coverage of its parent genome. The environmental core/accessory metric is a property of the gene with respect to the entire set of mapped metagenomes, so that a gene may attract high coverage from a subset of metagenomes and still be classified as environmentally accessory, and may be absent from a subset of metagenomes and still be classified as environmentally core. This metric was calculated across all metagenomes in a given site, for each gene in each genome included in a metapangenome. If the genome was not confidently detected (less than half of the nucleotides were covered in all metagenomes), all genes were reported NA instead of environmentally core or accessory.

A phylogenomic tree for the *H. parainfluenzae* isolates was constructed in RAxML (Stamatakis 2014) with the GTRCAT model and autoMRE boostrapping option using a nucleotide alignment

610 of 139 concatenated single-copy core genes obtained with the anvi-get-sequences-for-hmm-hits program.

Single-genome mapping

Per-sample coverage of a single genome was determined from the results of the metapangenomic analysis. For each oral habitat, coverage data for each strain's genome was
615 extracted from the metapangenomic data using the command anvi-script-gen-distribution-of-genes-in-a-bin, with the same environmental detection threshold as above (0.25). Per-habitat analyses were combined using a custom wrapper script employing anvi'o functions, subsetting to show coverage for the 30 samples per oral habitat with the highest median coverages, while retaining the environmental core/accessory determination from all samples, not just the
620 selected 30 samples (Supplemental Methods).

For visualization of nucleotide-level coverage (Figure 4B), coverage was obtained using the anvi-get-split-coverages for the splits containing the gene(s) of interest and plotted with a custom R script (Supplemental Methods). SNV information from each metagenome was reported using anvi-gen-variability-profile command (variability information was recorded during the mapping
625 step described earlier) which outputs the Shannon entropies of each variable position. Higher Shannon entropy signifies more environmental variability while low entropy signifies less variability. The mean of the observed Shannon entropies from all reporting metagenomes was then plotted for each position by oral habitat.

Functional enrichment analyses

630 Functional enrichment analyses were carried out following the pipeline described in Shaiber et al (2020). The analysis centers around the anvi-get-enriched-functions-per-pan-group function, which de-replicates the predicted functions of each genome and then carries out a series of functional contrasts between specified groups of genomes. For the *H. parainfluenzae* analysis, the three groups displayed in Figure 2 were used as the groups, and both TIGRFAM functions
635 were used. The analysis creates a list of enriched and depleted functions by group, based on the prevalence of each function in genomes belonging that group vs. the prevalence of that

function in genomes outside that group. This table and the statistical outputs are presented for Tigrfam (Additional file 4).

Figures were generated with anvi'o (Eren et al., 2015) and cleaned for publication in Inkscape
640 (https://inkscape.org/).

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: The raw data used in this study are publicly available at NIH GenBank and RefSeq (https://www.ncbi.nlm.nih.gov/genome/) for genomes (specific genome accessions listed in Additional file 9) and HMP metagenomes from https://portal.hmpdacc.org/. Analyzed data in the form of anvi'o databases can be found at https://doi.org/10.6084/m9.figshare.11591763.v1. A reproducible methods document (Supplemental Methods) provides the code used for the analyses.

Competing interests: The authors declare that they have no competing interests.

Funding: Support was provided to D.R.U, C.M.C, and G.G.B from Harvard Catalyst | The Harvard Clinical and Translational Science Center (National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health Award UL1 TR001102 and financial contributions from Harvard University and its affiliated academic health care centers). Additional support was provided from NIH Grant DE022586 to G.G.B. Additional support was provided by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745303 to D.R.U. Additional support was provided to D.R.U by Harvard University's Department of Organismic and Evolutionary Biology program.

Author contributions: D.R.U., G.G.B, A.M.E, C.M.C., and J.L.M.W. designed research; D.R.U, G.G.B, and J.L.M.W. performed research and analyzed data; D.R.U, G.G.B, A.M.E, C.M.C., and J.L.M.W. wrote the paper.

Acknowledgements: We thank Floyd Dewhirst for helpful discussions. We thank the many researchers who deposited the publicly-available genomes we used. The content is solely those of the authors and does not necessarily reflect the views of Harvard Catalyst, Harvard University and its affiliated academic health care centers, the National Science Foundation, or the National Institutes of Health.

References:

Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I., & Dewhirst, F. E. (2005). Defining the normal bacterial flora of the oral cavity. *Journal of Clinical Microbiology*, *43*(11), 5721–5732. http://doi.org/10.1128/JCM.43.11.5721-5732.2005

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., et al. (2019). A new genomic blueprint of the human gut microbiota. *Nature*, *568*(7753), 499–504. http://doi.org/10.1038/s41586-019-0965-1

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. http://doi.org/10.1016/S0022-2836(05)80360-2

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Research*, *33*. http://doi.org/10.1093/nar/gki034

Campbell, J. H., O'Donoghue, P., Campbell, A. G., Schwientek, P., Sczyrba, A., Woyke, T., et al. (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proceedings of the National Academy of Sciences*, *110*(14), 5540–5545. http://doi.org/10.1073/pnas.1303090110

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., & Banfield, J. F. (2019). Accurate and Complete Genomes from Metagenomes. *bioRxiv*, *1*, 808410. http://doi.org/10.1101/808410

Cornejo, O. E., Lefébure, T., Bitar, P. D. P., Lang, P., Richards, V. P., Eilertson, K., et al. (2013). Evolutionary and population genomics of the cavity causing bacteria Streptococcus mutans. *Molecular Biology and Evolution*, *30*(4), 881–893. http://doi.org/10.1093/molbev/mss278

Costea, P. I., Coelho, L. P., Sunagawa, S., Munch, R., Huerta-Cepas, J., Forslund, K., et al. (2017). Subspecies in the global human gut microbiome. *Molecular Systems Biology*, *13*(12), 960. http://doi.org/10.15252/msb.20177589

Darwin, C. (1859). *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life.* London: John Murray, Albemarle St.

Delmont, T. O., & Eren, A. M. (2018). Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ*, *6*, e4320. http://doi.org/10.7717/peerj.4320

Denef, V. J. (2019). Peering into the Genetic Makeup of Natural Microbial Populations using Metagenomics. In *Population Genomics: Microorganisms* (pp. 49-75). New York, NY: Springer.

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., et al. (2010). The human oral microbiome. *Journal of Bacteriology*, *192*(19), 5002–5017. http://doi.org/10.1128/JB.00542-10

Dimroth, P., Hilpert, W. (1984). Carboxylation of pyruvate and acetyl coenzyme A by reversal of the sodium pumps oxaloacetate decarboxylase and methylmalonyl-CoA decarboxylase. *Biochemistry 23*(22), 5360-5366. http://doi.org/10.1021/bi00317a039

Dimroth, P., Jockel, P., Schmid, M. (2001). Coupling mechanism of the oxaloacetate decarboxylase Na+ pump. *Biochimica et Biophysica Acta Bioenergetics, 1505*(1), 1-14.

Donati, C., Zolfo, M., Albanese, D., Tin Truong, D., Asnicar, F., Iebba, V., et al. (2016). Uncovering oral Neisseria tropism and persistence using metagenomic sequencing. *Nature Microbiology*, *1*(7), 16070. http://doi.org/10.1038/nmicrobiol.2016.70

Dvořák, P., Casamatta, D. A., Poulíčková, A., Hašler, P., Ondřej, V., & Sanges, R. (2014). Synechococcus: 3 billion years of global dominance. *Molecular Ecology*, *23*(22), 5538–5551. http://doi.org/10.1111/mec.12948

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. http://doi.org/10.1093/nar/gkh340

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. http://doi.org/10.1093/nar/gky995

Eren, A. M., Borisy, G. G., Huse, S. M., & Mark Welch, J. L. (2014). Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(28), E2875–84. http://doi.org/10.1073/pnas.1409644111

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, *3*(358), e1319. http://doi.org/10.7717/peerj.1319

Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T., & White, O. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research*, *29*(1), 41–43.

Hall, M. W., Singh, N., Ng, K. F., Lam, D. K., Goldberg, M. B., Tenenbaum, H. C., et al. (2017). Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity. *Npj Biofilms and Microbiomes*, *3*(1), 2. http://doi.org/10.1038/s41522-016-0011-0

Holliday, R., Preshaw, P. M., Bowen, L., & Jakubovics, N. S. (2015). The ultrastructure of subgingival dental plaque, revealed by high-resolution field emission scanning electron microscopy. *BDJ Open*, *1*(1), 15003–6. http://doi.org/10.1038/bdjopen.2015.3

Human Microbiome Project Consortium. (2012). A framework for human microbiome research. *Nature*, *486*(7402), 215–221. http://doi.org/10.1038/nature11209

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. http://doi.org/10.1186/1471-2105-11-119

Kilian, M. (1976). A taxonomic study of the genus Haemophilus, with the proposal of a new species. *Journal of General Microbiology*, *93*(1), 9–62. http://doi.org/10.1099/00221287-93-1-9

Kraal, L., Abubucker, S., Kota, K., Fischbach, M. A., & Mitreva, M. (2014). The Prevalence of Species and Strains in the Human Microbiome: A Resource for Experimental Efforts. *PloS One*, *9*(5), e97279. http://doi.org/10.1371/journal.pone.0097279

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. http://doi.org/10.1038/nmeth.1923

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*, *550*(7674), 61. http://doi.org/10.1038/nature23889

Mark Welch, J. L., Dewhirst, F. E., & Borisy, G. G. (2019). Biogeography of the Oral Microbiome: The Site-Specialist Hypothesis. *Annual Review of Microbiology*, *73*(1), 335–358. http://doi.org/10.1146/annurev-micro-090817-062503

Mark Welch, J. L., Rossetti, B. J., Rieken, C. W., Dewhirst, F. E., & Borisy, G. G. (2016). Biogeography of a human oral microbiome at the micron scale. *Proceedings of the National Academy of Sciences*, *113*(6), E791–E800. http://doi.org/10.1073/pnas.1522149113

Martino, M. E., Bayjanov, J. R., Caffrey, B. E., Wels, M., Joncour, P., Hughes, S., et al. (2016). Nomadic lifestyle of Lactobacillus plantarum revealed by comparative genomics of 54 strains isolated from different habitats. *Environmental Microbiology*, *18*(12), 4974–4989. http://doi.org/10.1111/1462-2920.13455

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, *15*(6), 589–594. http://doi.org/10.1016/j.gde.2005.09.006

Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., & Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature*, *568*(7753), 505–510. http://doi.org/10.1038/s41586-019-1058-x

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, *32*(8), 822–828. http://doi.org/10.1038/nbt.2939

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, *176*(3), 649–662.e20. http://doi.org/10.1016/j.cell.2019.01.001

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833–844. http://doi.org/10.1038/nbt.3935

Reveillaud, J., Bordenstein, S. R., Cruaud, C., Shaiber, A., Esen, Ö. C., Weill, M., et al. (2019). The Wolbachia mobilome in Culex pipiens includes a putative plasmid. *Nature Communications*, *10*(1), 1051–11. http://doi.org/10.1038/s41467-019-08973-w

Segata, N., Haake, S. K., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., et al. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biology*, *13*(6), 1. http://doi.org/10.1186/gb-2012-13-6-r42

Shaiber, A., Eren, A. M. (2019). Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories. *mBio*, *10*(3), 208. http://doi.org/10.1128/mBio.00725-19

Shaiber, A., Willis, A. D., Delmont, T. O., Roux, S., Chen, L.-X., Schmid, A. C., Yousef, M., Watson, A. R., Lolans, K., Esen, O. C., Lee, S. TM., Downey, N., Morrison, H. G., Dewhirst, F., Mark Welch, J. L., Eren, A. M. (2020). Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *bioRxiv*, 2020.04.29.069278; doi: https://doi.org/10.1101/2020.04.29.069278

Simon, M., Scheuner, C., Meier-Kolthoff, J. P., Brinkhoff, T., Wagner-Döbler, I., Ulbrich, M., et al. (2017). Phylogenomics of Rhodobacteraceae reveals evolutionary adaptation to marine and non-marine habitats. *Isme J*, *11*(6), 1483–1499. http://doi.org/10.1038/ismej.2016.198

Snel, B., Bork, P., & Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nature Genetics*, *21*(1), 108–110. http://doi.org/10.1038/5052

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. http://doi.org/10.1093/bioinformatics/btu033

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955. http://doi.org/10.1073/pnas.0506758102

Tierney, B. T., Yang, Z., Luber, J. M., Beaudin, M., Wibowo, M. C., Baek, C., et al. (2019). The Landscape of Genetic Content in the Gut and Oral Human Microbiome. *Cell Host & Microbe*, *26*(2), 283–295.e8. http://doi.org/10.1016/j.chom.2019.07.008

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., & Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research*, *27*(4), 626–638. http://doi.org/10.1101/gr.216242.116

van Dongen, S., & Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. *Methods in Molecular Biology (Clifton, N.J.)*, *804*(Chapter 15), 281–295. http://doi.org/10.1007/978-1-61779-361-5_15

Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, *23*, 148–154. http://doi.org/10.1016/j.mib.2014.11.016

Wilbert, S. A., Mark Welch, J. L., & Borisy, G. G. (2019). Spatial Ecology of the Human Tongue Dorsum Microbiome. *Cell Reports*. http://doi.org/10.2139/ssrn.3438369

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., et al. (2008). SUPERFAMILY—comparative genomics, datamining and sophisticated visualisation. *Nucleic Acids Research*, *37*, D380–D386. http://doi.org/10.1093/nar/gkn762

Yassour, M., Jason, E., Hogstrom, L. J., Arthur, T. D., Tripathi, S., Siljander, H., et al. (2018). Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host & Microbe*, *24*(1), 146–154.e4. http://doi.org/10.1016/j.chom.2018.06.007

Zijnge, V., van Leeuwen, M. B. M., Degener, J. E., Abbas, F., Thurnheer, T., Gmür, R., & Harmsen, H. J. M. (2010). Oral Biofilm Architecture on Natural Teeth. *PloS One*, *5*(2), e9321. http://doi.org/10.1371/journal.pone.0009321

Figure Legends

Figure 1: Metapangenomic workflow (A) Pangenome construction. (1) All putative protein-coding gene sequences (colored block arrows) are extracted from each bacterial genome (colored bacilli above genes) to be included in the pangenome and (2) clustered into homologous gene clusters via blastp results grouped by the Markov Clustering Algorithm (sequence variants cartoonized as shades of the same color). (3) These gene clusters become the central dendrogram of the pangenome. Note that the gene clusters are organized by occurrence in genomes, not based on the order found in a particular genome. The detection of each gene cluster in a genome is visualized by filling in to indicate the presence or absence of each gene cluster across the genomes. The genomes are ordered by a dendrogram (top right) based on each genome's gene cluster content. **(B)** Metagenomic mapping. (1) The exact genomes as above in (A) are used as a reference onto which reads from metagenomes are mapped. Gene-level coverage for each gene is then calculated. (2) These coverages are plotted for all genes from a given genome to show that genome's gene-level representation in those samples. (3) Environmental representation is evaluated for each gene to decide whether that gene is environmentally core (gene's median coverage > 0.25 of the genome's median coverage across metagenomes from that environment) or environmentally accessory (gene's median coverage < 0.25 of the genome's median coverage). **(C)** Metapangenome construction. The environmental representation from (B) is then summarized for all genes and then overlaid onto the pangenome created in (A) – the inner layers show the genomic representation of the pangenome, and the outer layer shows the environmental representation of the pangenome. This outer layer summarizes the fraction of genes in each gene cluster that were environmentally core or accessory in those metagenomes (callout).

Figure 2: Metapangenomic analysis of *Haemophilus parainfluenza* reveals hidden diversity and habitat-specific subgroups. The inner radial dendrogram shows the 4,318 gene clusters in the pangenome, clustered by presence/absence across genomes. The 33 genomes of *H. parainfluenzae* strains are plotted on the innermost 33 layers (black 270˚ arcs), spaced to reflect discernable groups based on genomic composition. Gene clusters within a given genome are filled in with black; gene clusters not present remain unfilled. Genomes are ordered by gene cluster frequency (top right dendrogram), with radial spacing added between major groups to improve visibility. The outermost three layers show the proportion of genes within each gene cluster determined to be environmental accessory (red) or core in HMP metagenomes from TD (blue), BM (violet), and SUPP (green), from inside to outside, respectively. If a genome was not well detected (<0.5 of nucleotides covered by all metagenomes) all its genes were NA (grey) instead of environmentally core or accessory. Extending off the pangenome above 3 o'clock are bar charts of relevant information for each genome, with the y-axis limits in parentheses. Above the genome content summaries, each genome's median coverage across all TD, BM, and SUPP metagenomes is shown in the colored bar graph. Per-sample coverage of each genome is shown in the heatmap above, where each row represents a different sample and cell color intensity reflecting the coverage normalized to the maximum value of that row (black = 0, bright = maximum; colors as before for each site).

Figure 3: Genomes in the *Rothia* genus metapangenome are organized by gene content into groups that reveal associations to specific habitats. (A) Tips on the inner radial dendrogram, setting the angular axis, correspond to gene clusters organized by presence/absence across

875 genomes. The angular distance thus holds the entirety of the *Rothia* pangenome, 6,160 distinct gene clusters. The inner 69 layers (270˚ arcs) represent genomes, colored by NCBI's taxonomic assignments, and organized by their gene cluster frequencies (top right vertical dendrogram). Each genome's gene cluster content is displayed by filling in cells (gene clusters) for genomes in which that gene is present. Gaps in radial spacing of layers delineate major groups determined
880 by inspection of the pangenome and dendrogram. Groups of gene clusters are annotated with text and a grey arc – "Genus core" are gene clusters core to the genus *Rothia*; "Rm", *R. mucilaginosa*; "Rm1", *R. mucilaginosa* subgroup 1; "Ra + Rd"; both *R. aeria* and *R. dentocariosa*; "Rd", *R. dentocariosa*; "Rm2", *R. mucilaginosa* subgroup 2; "Ra", *R. aeria*; "Rm BM", BM-abundant *R. mucilaginosa*. The outermost colored three layers show the proportion of genes
885 within each gene cluster deemed environmental accessory (red) or core for HMP metagenomes from tongue dorsum (blue), buccal mucosa (violet), and supragingival plaque (green). If a genome was not well detected (<0.5 of nucleotides covered by all metagenomes) all its genes were NA (grey) instead of environmentally core or accessory. Above the genome content summaries, each genome's median coverage across all TD, BM, and SUPP metagenomes is
890 shown in the colored bar graph. Per-sample coverage of each genome is shown in the heatmap above, where each row represents a different sample and cell color intensity reflects the coverage normalized to the maximum value of that row (black = 0, bright = maximum; colors as before for each site).

895 **Figure 4. Gene-scale metapangenomic analysis suggests candidate gene-level drivers of habitat adaptation. A)** Gene-level coverage of *Rothia* sp. E04. Units along the angular axis are *R.* sp. E04 genes, arranged in order found in *R.* sp. E04 with contigs joined arbitrarily. The innermost ring labels whether each gene was shared with all of *R. mucilaginosa* group 1 (pink), only between the BM-enriched strains *R.* spp. E04 and C03 (black; also shown with black lines
900 outside figure), or otherwise (light pink). The innermost 30 layers show coverage of each gene for 30 TD metagenomes with the highest coverage; middle 30, BM metagenomes; outer 30, SUPP metagenomes. Each layer's y-axis shows coverage by an individual sample, with y-axes scaled independently for each layer. The three outermost layers show whether genes were determined as environmental accessory (red) or core in TD (blue), BM (violet), or SUPP (green).
905 Arrowheads show examples of gene abundance patterns: uniformly low-to-absent coverage across metagenomes (empty black) vs stochastically abundant but typically environmentally accessory (filled black). **B)** Nucleotide-level coverage for a 20kb contiguous stretch of *R.* sp. E04's genome that includes a candidate gene driver of the BM adaptation, GC_00004770. This stretch is shown by the labelled black arc in (A). Each trace shows a single sample's coverage,
910 colored according to its oral site. Black bars show the mean Shannon entropy for variant sites covered at least 10x. Grey boxes above the SUPP traces mark genes, with GC_00004770 highlighted in black. **C)** Table of the 22 gene clusters unique to *R.* sp. E04 and *R.* sp. C03 (also marked with black ticks in panel A). The columns labelled "Environmental Core/Accessory" show the fraction of genes in each gene cluster that are core (colored according to that
915 habitat's color) or environmentally accessory (red). The corresponding Pfam function is listed for gene clusters for which a function could be predicted. The gene clusters environmentally core in BM and SUPP but not in TD are bolded.

920

Additional file 1. Flowchart of key methods and bioinformatic analyses performed. Boxes represent datasets (color coded by category / filetype), and arrows show the programs used to connect or transform the data. The shaded portion on the right (starting with "Individual metagenomes") was performed for each oral site independently (i.e. once each for tongue
925 dorsum, buccal mucosa, and supragingival plaque), and then the habitat-specific metapangenomes were combined onto a single pangenome, as described in the Methods.

Additional file 2. Summary of *H. parainfluenzae* gene clusters. Each row in the table describes a different gene, listing the genome from which it came, the gene cluster to which it belongs, its predicted function, and other summary information.

930 **Additional file 3. Comparison of genome relatedness by gene content with phylogenomics, 16S, and sourmash. A)** Phylogenomic tree based on 139 concatenated single-copy core genes. Tip names colored in red correspond to those of Group 2 in Figure 2. **B)** Pangenome is arranged as in Figure 2, but the yellow heatmap shows 16S % similarity and the red heatmap shows sourmash similarity. Each heatmap's order from bottom to top matches the order from left to
935 right.

Additional file 4. Functional enrichment results for *H. parainfluenzae*. Each row reports the enrichment of a function in a group or groups of *H. parainfluenzae* genomes, according to the groups labelled in Figure 2.

Additional file 5. Summary of *Rothia* gene clusters. Each row in the table describes a different
940 gene, listing the genome from which it came, the gene cluster to which it belongs, its predicted function, and other summary information.

Additional file 6. Syntenic arrangement of *Rothia mucilaginosa* genomes relative to a *R. mucilaginosa* subgroup 1 genome (*R.* sp. E04). Gene clusters from *R. mucilaginosa* genomes are arranged in syntenic order according to *R.* sp. E04 (red arrow); gene clusters not found in *R.*
945 sp. E04 are omitted. Red arrows above and below mark the 22 gene clusters uniquely shared by both R. sp E04 and R. sp C03 (blue arrow). The order and spacing of layers is identical to that Figure 2 but linearized.

Additional file 7. Coverage and detection of genes from outlier genome *R. dentocariosa* OG2_1. Each unit around the circle represents a different OG2_1 gene, while each ring shows
950 data for a single metagenome. The left column shows the coverage of each gene across all metagenomes from SUPP (top), BM (middle), and TD (bottom), while the right column shows the detection (fraction of gene's nucleotides receiving any coverage at all) for each gene.

Additional file 8. *R.* sp. C03 gene-level recruitment of HMP metagenomes. Layout as in Figure 4A but for *R.* sp. C03. Genes shared with *R.* sp. E04 are marked to also show that the genes
955 unique to the BM-enriched subgroup are also well distributed throughout the genome.

Additional file 9. Accession information for genomes used in in this study. Genomes are listed in the same order as in Figures 2 and 3, from inside to outside.

960 **Supplementary Table 1: Number of gene clusters per pangenome with varying MCL inflation factors.**

	***Rothia* genus**			***H. parainfluenzae***		
MCL Inflation factor	4	6*	8	6	8	10*
Num. gene clusters	5994	6,127	6,234	4257	4303	4318
% change from *	-2.17%	0%	1.7%	-1.41%	0.35%	0%

* MCL inflation factor selected for use in all subsequent analyses presented

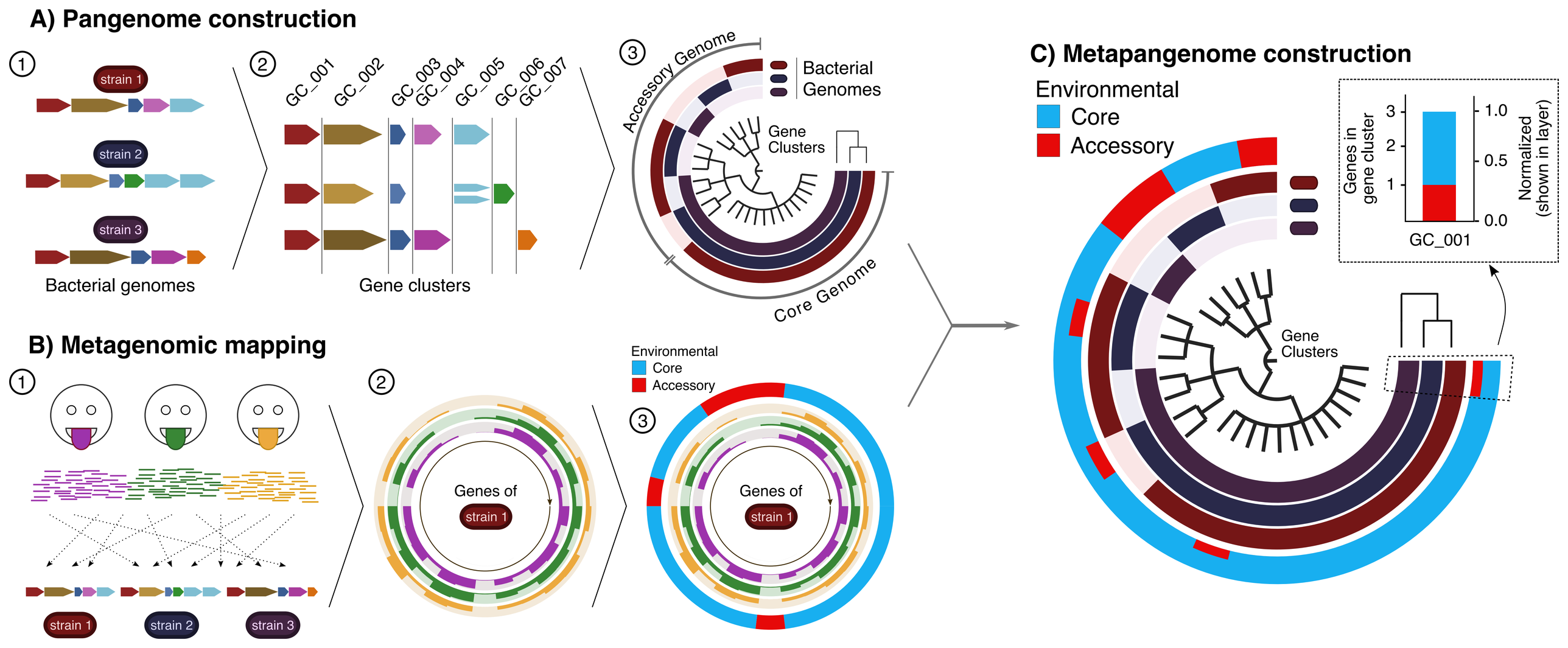
Supplemental Text – Detecting and defining homology in a pangenome

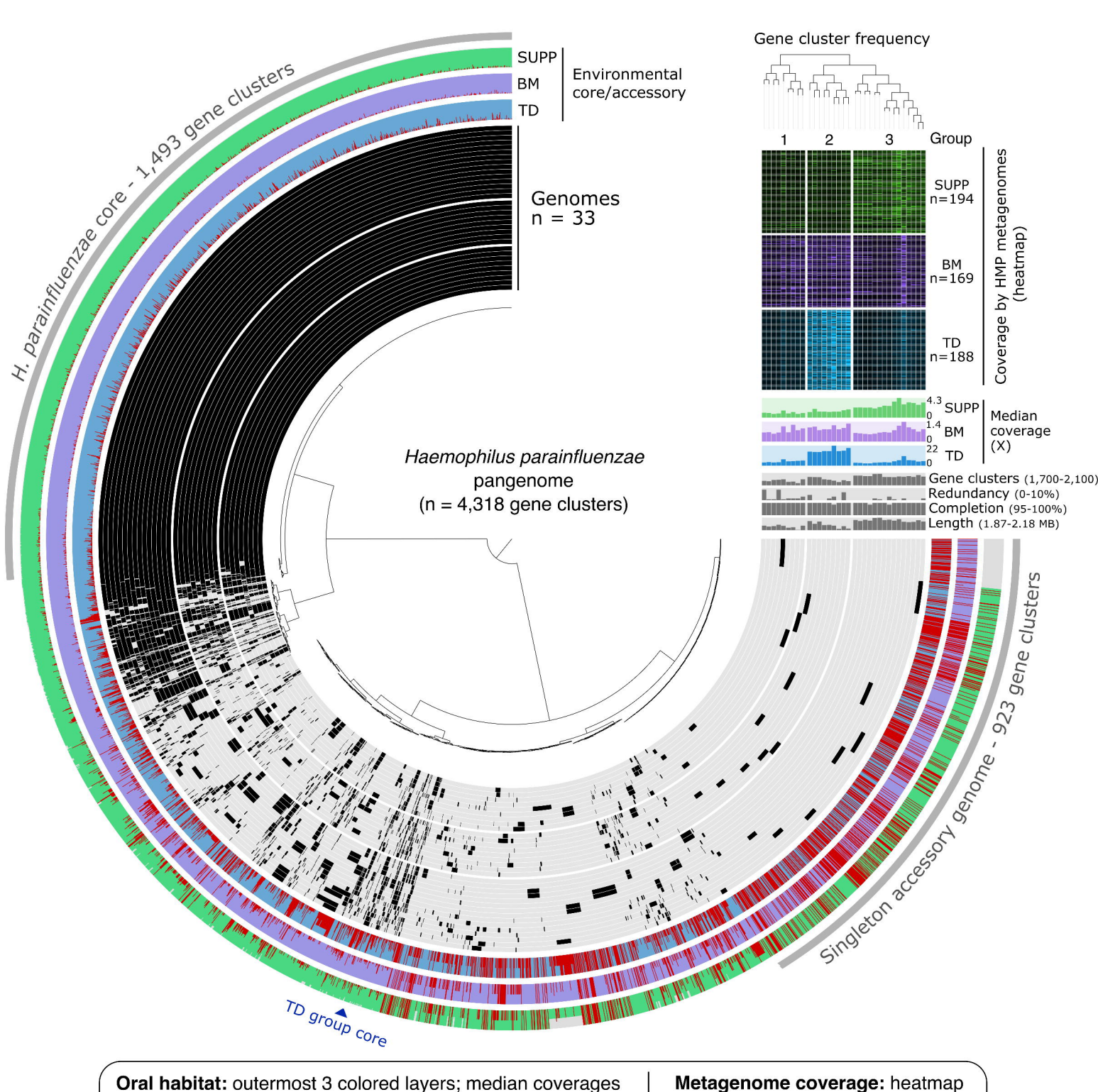
A crucial element underlying construction of a pangenome is being able to identify and group homologous genes. As described in the methods, the approach used was to compare amino acid sequences for all gene pairs, prune weak hits, and then cluster the resultant network with MCL. Given that any given gene in a genome can evolve at a different rate, no absolute or universal threshold can be derived. Thus, a threshold for defining homologous clusters of genes must be operationally chosen based on the scope of the question at hand and interpreted with the expectation that while the majority of resultant gene clusters may reflect the intended level of homology, some gene clusters will be over-split while other gene clusters will be under-split. Thus, differentiating between a legitimately unique gene cluster vs an over-split gene cluster is of importance when doing comparative genomics.

We employed a variety of methods to investigate both the robustness and validity of our homology definitions. To gauge robustness, as described in the Methods, identical pangenomes were run but the inflation parameter of the MCL algorithm was varied, which adjusts the sensitivity of the clustering algorithm to clusters detected in AA-similarity network. Despite varying inflation parameters, the resultant gene clusterings were qualitatively the same and quantitatively similar. On the other hand, universally proving the validity of all gene clusters is a challenging task. Based on the predicted functions of gene clusters belonging to the genus-level core (Additional files 2 and 5), the gene clustering behaved as expected. Additionally, for the specific gene clusters focused on in this manuscript, we used two additional tests to validate the predicted homology – functional annotation and sequence alignment.

For validating a gene cluster's independence from other gene clusters, our functional annotation test consisted of comparing that gene cluster's predicted function to those in other groups. For example, three gene clusters that were shared exclusively to *Rothia* sp. strains E04 and C03 (the two enriched in buccal mucosa) had functional annotation. However, other gene clusters with identical predicted functions were found in other *Rothia* genomes, and so these three gene clusters were discounted as potential hitchhikers. The sequence divergence within those gene clusters may confer differential fitness between habitats, however, we do not feel confident enough to put forward that hypothesis given this data.

Additionally, the validity of a gene cluster can be investigated by inspecting the alignment of its constitutive amino acid sequences. This allows one to investigate whether a gene cluster's representation in a genome is truly legitimate. For instance, in our investigation of the TD-preferring *Haemophilus parainfluenzae*, we noticed that the TD-preferring strains were characterized by the three gene clusters encoding the three subunits of oxaloacetate dehydrogenase. However, a single non-TD strain also contained one of the three gene clusters. By inspecting the sequence alignment, we discovered that the sequence was highly aberrant relative to the other sequences, likely reflecting mis-assignment to this gene cluster, and so provided further evidence that the TD-preferring genomes were set apart by their exclusive possession of this operon.


A) Pangenome construction
1
strain 1
strain 2
strain 3
Bacterial genomes
2
GC_001
GC_002
GC_003
GC_004
GC_005
GC_006
GC_007
Gene clusters
3
Accessory Genome
Bacterial Genomes
Gene Clusters
Core Genome
B) Metagenomic mapping
1
strain 1
strain 2
strain 3
2
Genes of
strain 1
3
Environmental
Core
Accessory
Genes of
strain 1
C) Metapangenome construction
Environmental
Core
Accessory
Gene Clusters
Genes in gene cluster
Normalized (shown in layer)
GC_001



Oral habitat: outermost 3 colored layers; median coverages

TD BM SUPP

Environmental core/accessory: outermost 3 layers

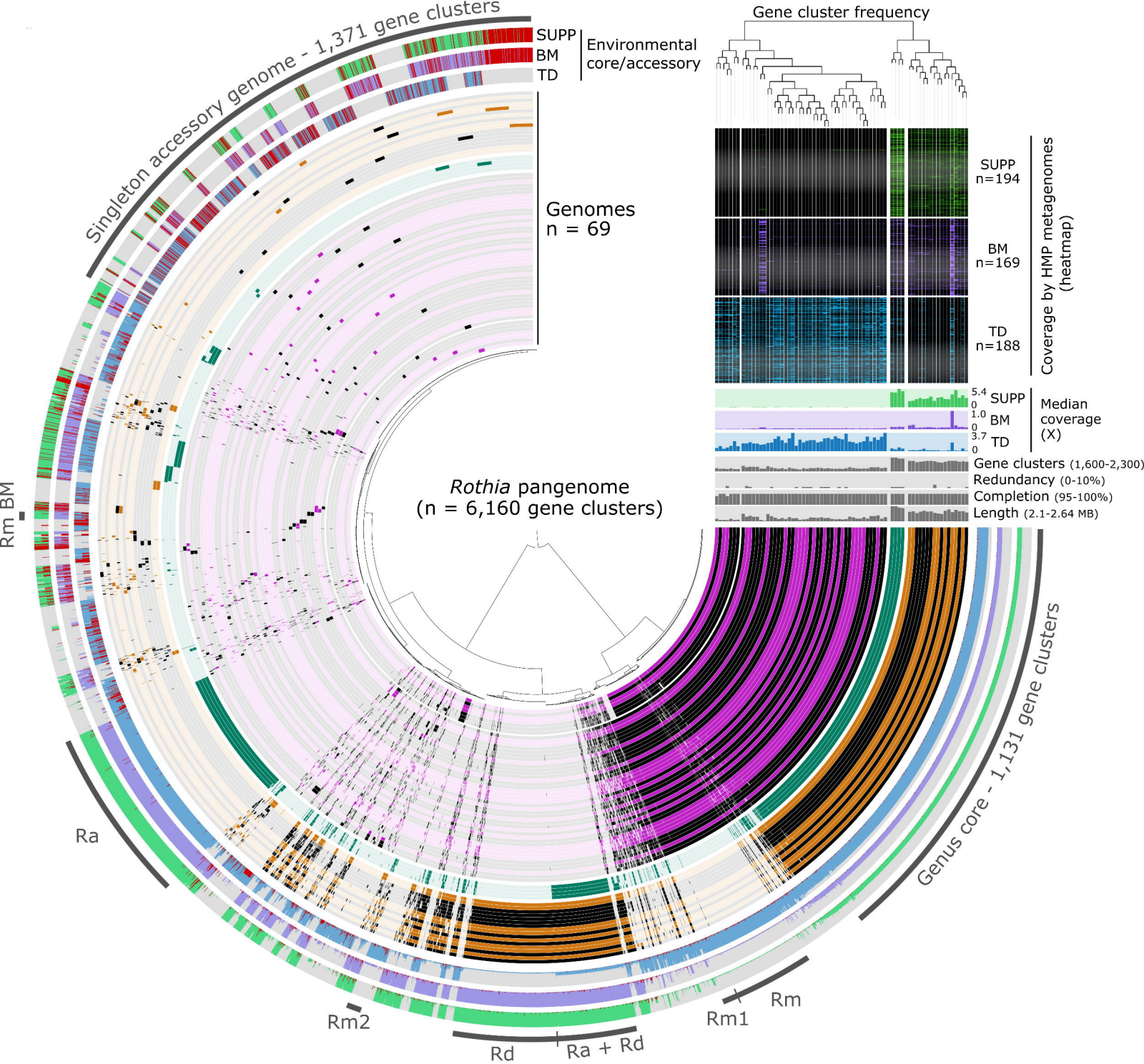
Core (habitat colors) Accessory NA

Metagenome coverage: heatmap

TD Max 0

BM Max 0

SUPP Max 0



Genome taxonomy: inner 69 layers

R. mucilaginosa *R. aeria* *R. dentocariosa* *R. sp.*

Oral habitat: outermost 3 colored layers; median coverages

TD BM SUPP

Environmental core/accessory: outermost 3 layers

Core (habitat colors) Accessory NA

Metagenome coverage: heatmap

TD Max 0

BM Max 0

SUPP Max 0

A)

Region in B)

R. sp. E04

Genes shared with:

All *R. muc.* group 1

Only C03 (BM group)

Others + singletons

TD BM SUPP

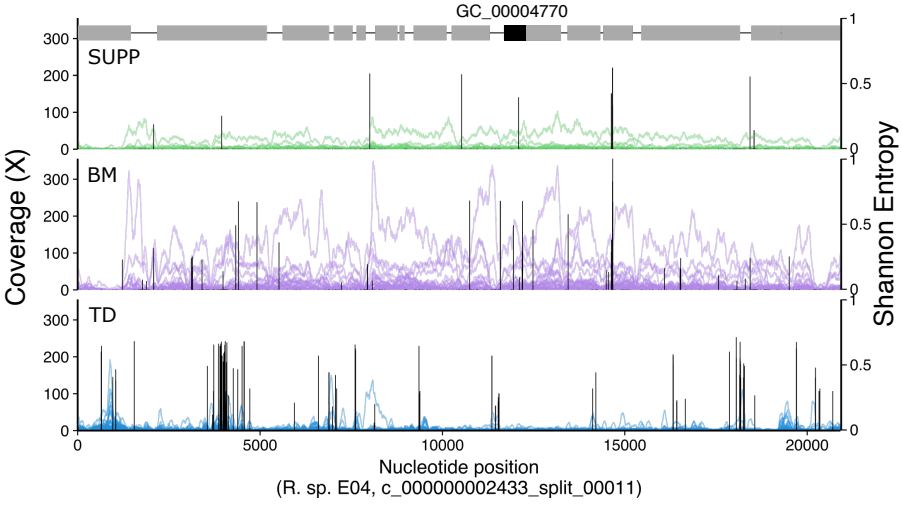
Environmental

Accessory

Core

TD BM SUPP

B)



C)