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 not only limitations in the ability of cultivated organisms to represent populations in their native environment, but also specifically which cultivar gene sequences were absent or 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 prompted a major effort to investigate the identity and distribution patterns of bacteria throughout the human body, the Human Microbiome Project (HMP 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 even when the organisms used for the pangenome were not isolated from the same samples whose metagenomes will be studied. A pangenome constructed from isolates collected at different times and around the world reveals the shared and variable gene content of the different organisms. The isolate genomes can then be used for short-read mapping from entirely unrelated metagenomic samples, revealing the environmental distribution of those same genes and genomes and permitting identification of biogeographic patterns. This short-read mapping 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 members of a species or genus (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) 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. Different surfaces in the mouth, such as the tongue dorsum, buccal mucosa, and teeth, constitute distinct habitats, each with a characteristic microbiome (Aas et al., 2005; Eren et al., 2014; Hall et al., 2017; Segata et al., 2012). These microbiomes are dominated by a few dozen taxa with high abundance and prevalence (Mark Welch et al. 2016, Utter et al. 2016, Mark Welch et al. 2019), most of which have cultivable representatives from which genomes have been sequenced (Dewhirst et al. 2010, homd.org), making the system unusually tractable relative to other natural microbiomes. The microbiomes that assemble in the different oral habitats are clearly related to one another – composed of many of the same genera, for example – but are largely composed of different species (Mark Welch et al. 2019). For example, the major oral genera *Actinomyces*, *Fusobacterium, Neisseria*, *Veillonella*, and *Rothia* occur throughout the mouth, but their individual species show strongly differential habitat distributions. Individual species within these genera typically have 95-100% prevalence across individuals and make up several percent of the community at one oral site, but show lower prevalence and two orders of magnitude lower abundance at other oral sites (Eren et al., 2014; Mark Welch et al., 2016; Mark Welch et al., 2019; Wilbert et al., 2020). The reproducibility of taxon distribution across individuals, despite the frequent communication of the habitats with one another via salivary flow, suggests that founder effects and other stochastic processes are unlikely to explain the differences in species-level distribution and that these differences likely arise from selection. However, some apparent "habitat generalist" species, such as *Haemophilus parainfluenzae*, *Streptococcus mitis*, and *Porphyromonas pasteri*, can be found throughout the mouth (Eren et al., 2014; Segata et al., 2012). Altogether, the mouth is colonized by well-characterized bacteria that build distinctive communities in the different oral habitats in the absence of dispersal barriers. This setting presents an opportunity to investigate the genomic characteristics underlying the differential success of closely related species in different habitats.

Here, we combine pangenomes and metagenomes to investigate how genes are distributed across populations at distinct sites within the mouth. We focused on two exemplar oral taxa with high prevalence (>95%; Segata et al. 2012, Eren et al. 2014, Mark Welch et al. 2016) and high abundance in the mouth: the species *Haemophilus parainfluenzae* and the genus *Rothia*. These two taxa represent the two oral biogeographic patterns, with *H. parainfluenzae* representing apparent habitat generalist species and the genus *Rothia* representing genera composed of habitat-specific species. Both *Rothia* spp. and *H. parainfluenzae* are sufficiently abundant – making up on average several percent of the microbiota at their sites of highest abundance – that metagenomic read recruitment to reference genomes can reliably sample their natural populations. As the basis for analysis of natural populations, we constructed pangenomes using genome sequences from previously cultured isolates. We then investigated the degree to which each gene in the pangenome is represented in populations from the healthy human mouth using metagenomic data from the Human Microbiome Project (HMP Consortium, 2012). 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.

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. Critically, this mapping of all samples to all genomes is naïve to any assumptions about which genomes occur in which habitats. The detection of a genome in a metagenome is operationally defined as the finding that at least half of the nucleotides in the genome are covered at least once. 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 a given fraction of the median coverage of the genome in which it resides. We used a fraction of one-fourth, following Delmont & Eren (2018). The gene is environmentally “accessory” if its coverage falls below this cutoff. This method normalizes gene coverage to the genome and so is robust to differences in sequencing depth across samples. The one-fourth threshold is arbitrary, but most genes in our samples were either completely covered (detected) in many metagenomes and were environmentally core, or recruited no coverage and were environmentally accessory (Additional File 2, Supplemental Methods). Thus, the specific value of the core/accessory cutoff has minimal effect on the identification of genes as environmentally core or accessory. The environmental core/accessory metric 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. These genomes were sequenced over 8 years at 9 institutions with listed isolation sources ranging from sputum to blood (Additional File 3), with many from an unspecified body site. Thus, we consider it likely that each study and institution sampled from independent donors. We constructed a pangenome from these 33 genomes (Figure 2, Additional File 4). Inspection of this pangenome (4,318 gene clusters in total) shows a large core genome encompassing 35% of the pangenome (N= 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 943 gene clusters (22% of the pangenome) unique to a single isolate genome, shown on the figure between 3 and 5 o'clock, and 1,882 gene clusters (44% of the pangenome) shared by some but not all isolate genomes, shown between 5 o'clock and 9 o'clock on the figure. Functionally, while the core and accessory genome contained representatives of most COG categories, compositional differences were apparent, mostly due to fewer genes of unknown function in the core genome and fewer conserved functions like translation in the accessory genome (Additional File 5AB, Supplemental Text).

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 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. Genome completeness was >99% and redundancy was <10% in all genomes (Figure 2, middle two grey bar charts), suggesting that the observed grouping is not based on the quality of the genome assemblies. As the number of gene clusters ranges from 1,773 to 2,071 per genome (Figure 2, top right grey bar chart), 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 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, permutational multivariate analysis of variance using Bray-Curtis dissimilarities, calculated using ADONIS in R; Anderson, 2001). 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 each oral habitat, the coverage of each genome by the median metagenome 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 the 188 tongue dorsum metagenomes than genomes in the other two groups (Figure 2 heatmap, median coverages). The heatmap in Figure 2 shows that each TD metagenome typically provided high coverage to several Group 2 genomes, although there was sample-to-sample variation in which genomes were most highly abundant. The median coverage bar plots show that reads from the median TD metagenome covered each of the nine genomes in Group 2 to an average depth of at least 15X, indicating that organisms similar to these strains are in high abundance on most people's tongues. Median coverage of the other twenty-four genomes by TD metagenomes is several-fold less (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 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 6A). 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 6B). Thus, these analyses suggest that genomes of *H. parainfluenzae* that are enriched in tongue metagenomes share similar gene content but do not form a monophyletic evolutionary group.

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

Correspondence between genome content and environmental distribution raises the possibility that the success of a particular strain in a given habitat within the mouth may rely on the presence of certain genes fixed by selection. 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 7). 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 had this universal presence in the TD-associated group but complete absence from the other *H. parainfluenzae* genomes. Aside from selection, an alternate explanation for the unique occurrence of this oxaloacetate operon in the TD-associated genomes could be shared evolutionary history, such as if these genomes were all isolated from the same subject. However, not only are the TD-associated genomes not monophyletic (Additional File 6A), they come from strains isolated from human sputum, the human toe, and the oropharynx of a rat and have sequences deposited by four different groups over 8 years (Additional File 3). Thus, the oxaloacetate 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 as environmentally accessory, particularly in TD (Figure 2, shown as spikes of red in the ‘Environmental core/accessory’ layer; Additional File 4). 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 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 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 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 (Mark Welch et al., 2019; Segata et al., 2012). An advantage of constructing pangenomes at the genus 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-seven high-quality *Rothia* genomes from NCBI. Of the genomes for which the isolation source of the strain was reported, most were from sputum or bronchoalveolar lavage (Additional File 3). From these 69 genomes we constructed a genus-level pangenome consisting of 5,992 gene clusters (Figure 3, Additional File 8).

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 over 200-400 gene clusters that are absent from the others (Figure 3). Taxonomic designations provided by NCBI (depicted by coloring the genome layers) show that these groups correspond to the three different recognized human oral *Rothia* species. A large set of gene clusters (n = 1,129, 19% of the pangenome) 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 half to two-thirds of any given *Rothia* genome. Other sets of gene clusters were characteristic of and exclusive to *R. mucilaginosa* (n = 207, 3% of the pangenome)*, R. dentocariosa* (n = 274, 5%), or *R. aeria* (n = 455, 8%). 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 9). The gene clusters present only in group 1 do not localize to a single region but are scattered throughout the chromosome (Additional File 9), 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 distributions of genome groups

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. As in Figure 2, the heatmap shows coverage data for hundreds of metagenomes (rows) collected from over a hundred different volunteers by the HMP. Two of the *Rothia* species (*R. aeria* and *R. dentocariosa*) were most abundant in SUPP samples, where the mean depth of coverage from the median SUPP metagenome was approximately 5X for the *R. aeria* genomes and 2 to 3X for most *R. dentocariosa* genomes (Figure 3). The third species, *R. mucilaginosa*, was highly abundant in TD and for the most part displayed only negligible coverage from SUPP and BM samples. Outliers were also apparent – two genomes in the *R. mucilaginosa* group received high coverage from approximately one-third of BM metagenomes.

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 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.

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 distributions. *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. Investigating the predicted functions core to each species also supported the observed differentiation of species (Supplemental Text, Additional File 5C).

***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 (heat map, Fig. 3). 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). In Figure 4A, each unit around the near-complete circle represents a different gene in the genome, and the 90 small tracks show each gene's coverage in the 30 metagenomes per habitat with the highest *R.* sp. E04 coverage (Supplemental Reproducible Workflow). The BM and SUPP metagenomes covered the majority of this genome's genes relatively evenly, evidenced by the taller and more dense bars in the purple (BM) and green (SUPP) metagenomes, 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 low to absent in most regions of the genome and only dense for a handful of genes. Similar analysis of the other BM-abundant genome, *R.* sp. C03, revealed similar results (Additional File 10). This pattern suggests the TD coverage results from spurious mapping of isolated regions of high similarity or 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 genome. This is observed in Figure 4B where the vertical black lines report the mean Shannon entropy of mapped nucleotide 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 more abundant in buccal mucosa rather than tongue.

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 tick marks, Additional File 9), 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). In other words, the four genes in bold in Figure 4C recruited less coverage than their surrounding genomes in the 188 TD metagenomes where other *R. mucilaginosa* were abundant (Figure 3 heatmap), yet these four genes had abundance similar to their surrounding genomes in the 169 BM and 198 SUPP metagenomes (Figure 4A,B). These four genes thus represent prominent candidates for future 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. 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

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 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 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-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

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 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 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., 2020); 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 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; Mark Welch et al., 2016; Wilbert et al., 2020), 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 from saliva. Our finding of a distinctive mapping pattern consistent across hundreds of

HMP metagenomes for 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 (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 et al. 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). Genomic biogeographic patterns also exist at the global scale, as shown by a recent study that identified differentiation in the motility and metabolic potential of European *E. rectale* populations relative to those from other continents (Karcher et al., 2020). Another 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 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. Indeed, recent studies have shed light on the magnitude of previously-unknown genes contained within the human microbiome (Pasolli et al., 2019; Tierney et al., 2019). 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.

A major benefit of our metapangenomic strategy is that it permits us to identify which genes and cultivars are most representative of the microbiota growing in a given natural habitat. We used metapangenomics to query each gene from the *Rothia* and *H. parainfluenzae* cultivar pangenomes across metagenomes from the human oral cavity and measure the abundance of each cultivar gene across environments. These data can serve as a resource to guide the selection of the most environmentally representative strains and gene sequences for future experiments.

Cultivars are a valuable starting point for a metapangenomic analysis because they provide a high-quality foundation for assessing the presence, absence, and precise nucleotide sequence of genes. The source from which a cultivar is isolated, however, is not necessarily indicative of its environmental distribution; this distribution is more suitably assessed using metagenomic data. The Baas-Becking hypothesis that “everything is everywhere, but the environment selects” (Baas-Becking, 1934) suggests that the isolation of a single cell does not necessarily imply the existence of a population. The mapping of metagenomic data to a cultivar genome, by contrast, does indicate the overall abundance of an isolate in a habitat (Kraal et al., 2014; Shaiber et al., 2020), and the depth of coverage provided by different samples can indicate that the location of highest abundance of a resident population may not be its original site of isolation. For example, the obligate bacterial symbiont TM7x was first isolated from a salivary sample in association with an *Actinomyces odontolyticus* strain (He et al., 2015). However, as saliva is a transient mixture of bacteria shed from other oral sites, the ultimate source of TM7x remained ambiguous until metagenomic mapping was used to identify dental plaque as its native habitat (Shaiber et al. 2020). Many of the genomes we used in this study came from strains isolated from sputum and non-oral sources such as blood, gallbladder, and skin (Additional File 3). Nonetheless these genomes proved to be valuable references to probe the oral distribution of populations related to these genomes using metagenomic mapping. Based on our mapping results that show the high prevalence and abundance of oral populations similar to the isolate genomes, we infer that the strains isolated from blood and other non-oral samples are migrants dispersed from resident oral populations.

Mapping metagenomic short reads onto reference genomes can be used to investigate the relative divergence between a sampled population and the reference genome (Simmons et al., 2008; Denef 2018). The specific patterns of single-nucleotide variants (SNVs) among even closely-related strains provide one of the most powerful ways to distinguish and track highly-related strains, e.g., from mothers to infants (Yassour et al., 2018). In this study, we compared the relative frequencies of SNVs between different habitats as a proxy for relatedness to infer which sites had populations that were most similar to the reference strain. However, we did not explicitly search for specific SNVs that were enriched in one habitat vs. another. Future studies of nucleotide and codon variants across habitats will reveal the importance of nucleotide- and amino-acid level changes for habitat specialization (Delmont et al., 2019).

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 to fully exploit available ecological niches is observed at many levels, from recognized species distinguished by many genes down to closely related subpopulations.

Methods

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 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 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*, 67 of the 73 genomes present in RefSeq were used. Genomes were inspected for potential errors and contamination which were identified based on expected genome size and gene count, fragmentary assemblies composed of short contigs, aberrantly high coverage of specific genes of unknown function (e.g., >1000x coverage for genes that are neither rRNA not mobile elements), and existing literature (Breitwieser et al., 2019). Of the six genomes not used, *R. nasimurium* was discarded for not being recognized as an oral resident by the HOMD, while *R*. sp. Olga and R. sp. ND6WE1A were discarded as non-oral isolates with aberrantly large unique gene contents (potentially contaminant genes). One *R*. *dentocariosa* genome was discarded for aberrant coverage and two *R. dentocariosa* genomes (OG2-1 and OG2-2) for containing potential contaminant genes based on aberrant coverage and for being identified as contaminated by Breitwieser et al. (2019). For *Haemophilus parainfluenzae*, all 33 genomes in RefSeq passed the contamination inspection and were used for analysis. Metadata from NCBI available for each strain is provided in Additional File 3.

Raw short-read metagenomic data from the Human Microbiome Project (HMP; HMP Consortium, 2012) 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/. These short-read data had undergone the HMP quality-control pipeline which includes trimming of low-quality bases and subsequently discarding of reads below 60bp (HMP Consortium, 2012).

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 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-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 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.

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 with default value 0.5 which requires that the bitscore of BLAST be at least half the maximum possible bitscore given the length of the sequences. The workflow then uses the Markov Cluster Algorithm (MCL) (van Dongen & Abreu-Goodger, 2012) to group ORFs into putatively homologous gene groups (gene clusters; Supplemental Text), and aligns amino acid sequences in each gene cluster using MUSCLE (Edgar, 2004) for interactive visualization. For display of the pangenome, the order of the genome layers was determined by clustering the genomes based on the frequency with which each gene cluster appeared in each genome, 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 was determined for any genomes covered at least 1x over half the genome length in at least one sample. A gene was classified as environmentally core with respect to an oral site if the median coverage of that gene by samples from that oral site was at least one-fourth the median coverage by those same samples of the genome from which it came (Delmont & Eren, 2018); otherwise the gene was classified as environmentally accessory. If the genome was not confidently detected (if less than half of the nucleotides in the genome were covered) in any metagenome, 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 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 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 selected 30 samples (Supplemental Methods).

For visualization of nucleotide-level coverage in 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 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

Functional enrichment analyses were carried out following the pipeline described in Shaiber et al (2020). The analysis uses 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 were used. The analysis identifies 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.

Figures were generated with anvi'o (Eren et al., 2015) and cleaned for publication in Inkscape (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 2) 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

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*(1), 32–46. http://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Baas-Becking, L. (1934). Geobiologie; of inleiding tot de milieukunde.

Breitwieser, F. P., Pertea, M., Zimin, A. V., & Salzberg, S. L. (2019). Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Research*, *29*(6), 954–960. http://doi.org/10.1101/gr.245373.118

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

Delmont, T. O., Kiefl, E., Kilinc, O., Esen, Ö. C., Uysal, I., Rappé, M. S., et al. (2019). Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife*, *8*, 403. http://doi.org/10.7554/eLife.46497

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

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

He, X., McLean, J. S., Edlund, A., Yooseph, S., Hall, A. P., Liu, S.-Y., et al. (2015). Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proceedings of the National Academy of Sciences*, *112*(1), 244–249. http://doi.org/10.1073/pnas.1419038112

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

Karcher, N., Pasolli, E., Asnicar, F., Huang, K. D., Tett, A., Manara, S., et al. (2020). Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. *Genome Biology*, *21*(1), 1–27. http://doi.org/10.1186/s13059-020-02042-y

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., Huttenhower, C., & Izard, J. (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., & Relman, D. A. (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., et al. (2020). Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *bioRxiv*, 2020.04.29.069278.

Simmons, S. L., Dibartolo, G., Denef, V. J., Goltsman, D. S. A., Thelen, M. P., & Banfield, J. F. (2008). Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS Biology*, *6*(7), e177. http://doi.org/10.1371/journal.pbio.0060177

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, 30,* 4003-4015. https://doi.org/10.1016/j.celrep.2020.02.097

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 reflects the coverage. Coverage is normalized to the maximum value of that sample (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. Tips on the inner radial dendrogram,

setting the angular axis, correspond to gene clusters organized by presence/absence across 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 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 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 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. Coverage is normalized to the maximum value of that sample (black = 0, bright = maximum; colors as before for each site).

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 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). 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, 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 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.

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 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. Gene detection in metagenomes is largely bimodal. For all metagenomes covering at least half the nucleotides in a genome, the detection (fraction of each gene receiving any coverage at all) of all genes that genome was counted. For **A)** *H. parainfluenzae* and **B)** *Rothia* spp., the number of metagenomes (y axis) providing each observed gene detection is plotted as a histogram. The genes were split into two categories (colors) – those determined to be environmentally accessory genes (EAG) or environmentally core genes (ECG) by having a median coverage less than or at least 0.25x the parent genome’s median coverage, respectively. **C)** and **D)** show the probability density function for *H. parainfluenzae* and *Rothia,* respectively, using the same gene detection data shown in **A** and **B**. Detection is shown on the x-axis, and the y-axis shows the probability of the metagenomes producing that detection. The distribution of detections for EAGs are shown in orange and ECG in blue.

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

Additional File 4. 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.

Additional File 5. Functional similarities in the pangenome. A) COG categories of the different *H. parainfluenzae* pangenome fractions (x-axis). The pangenome was apportioned into the core genome (gene clusters found in all genomes), the singleton accessory genome (gene clusters in exactly one genome), and the intermediate accessory genome (the remaining gene clusters). The height of each bar shows the number of COG-annotated gene clusters per pangenome portion, colored by COG category. Only gene clusters annotated with a single COG category, or none, were included. **B)** Enrichment of TIGRFAM functions in by group of *H. parainfluenzae* genomes detected in the pangenome (Figure 2). Each group or combination of groups is listed along the x axis. The y-axis is the count of TIGRFAM genes enriched by group, with each gene colored by its aggregate proportional enrichment. Aggregate enrichment was calculated for each TIGRFAM by subtracting the mean proportional occurrence of each function in the group(s) in which it was not enriched from the mean of its proportional occurrence in the group(s) in which the TIGRFAM was enriched. **C)** The same analysis as in **B** is shown but for species of the genus *Rothia*.

Additional File 6. 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 right.

Additional File 7. Functional enrichment results for *H. parainfluenzae*. Each row reports the enrichment of a TIGRFAM function in a group or groups of *H. parainfluenzae* genomes, according to the groups labelled in Figure 2.

Additional File 8. Summary of *Rothia* 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.

Additional File 9. 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.* 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 10. *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 unique to the BM-enriched subgroup are also well distributed throughout the genome.

Additional File 11. Comparison of varying MCL inflation factors on pangenome structure. Identical pangenomes were run but with varying MCL inflation factors. The left and right columns of pangenome plots show the *H. parainfluenzae* and *Rothia* pangenomes, respectively. The MCL inflation parameter used for each pangenome shown is listed next to the central dendrogram.

Additional File 12. Functional enrichment results for *Rothia* species. Each row reports the enrichment of a different TIGRFAM function in one or more *Rothia* species.

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

	Rothia genus			*H. parainfluenzae*		
MCL Inflation factor	4	6*	8	8	10*	12
Num. gene clusters	5,855	5,992	6,090	4,303	4,318	4,338
% change from *	-2.29%	0%	1.64%	0.35%	0%	0.46%

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

1015

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. Within a group of closely related genomes, amino acid sequences of homologous genes are likely to be largely conserved across genomes while non-homologous genes both within and across genomes are distinct. Thus, clusters for a species- or genus-level pangenome may be unambiguous. Nonetheless, ambiguous homology and errors in clustering may occur. We used two methods to investigate the overall robustness and validity of our homology definitions – first, determining the robustness of the pangenome to various amino acid similarity thresholds, and second, assessing the level of functional heterogeneity within our gene clusters.

Our pangenome construction approach compares amino acid sequences for all gene pairs, prunes weak hits, and resolves the network of hits with the Markov Cluster Algorithm (MCL) to determine gene clusters. MCL uses a hyperparameter, "inflation," to adjust the clustering sensitivity, i.e., the tendency to split clusters. To gauge robustness of the pangenome to the inflation parameter of the MCL algorithm, we varied the inflation parameters by ± 2. The resultant number of gene clusters was quantitatively similar (Supplementary Table 1), differing by <0.5% for *H. parainfluenzae* and <2.5% for *Rothia*, and the pangenome arrangement was qualitatively similar in that the overall pattern and relative size of the genus core (in the case of *Rothia*), species cores, and accessory genome remained nearly identical (Additional file 11).

Gene clusters are defined purely by amino acid sequence similarity. Although functional similarity is not part of the definition, nevertheless, intuitively one expects to produce gene clusters that are composed of genes with similar function. We assessed the validity of this expectation by assessing the fraction of gene clusters whose constituent genes were annotated with different COG functions. Heterogeneity of functional annotation within a gene cluster was rare in our data; for *H. parainfluenzae*, only 2.6% (75 out of 2892 gene clusters with predicted COG functions) of gene clusters had within-cluster functional heterogeneity, and *Rothia* was comparably low at 3.5% (96 of 2757 gene clusters with COG annotation).

For the specific gene clusters that we focused on in this manuscript, we used two additional tests to assess the internal consistency of the gene clusters – functional annotation and manually inspection of sequence alignments.

Gene clusters may be legitimately split according to amino acid sequence, and yet still represent homologous genes carrying out the same function. For gene clusters identified as unique to a group of interest, our functional annotation test consisted of comparing the predicted function of that gene cluster to functions of gene clusters characteristic of other groups. For example, of the gene clusters that were shared exclusively to *Rothia* sp. strains E04 and C03 (the two genomes enriched in buccal mucosa metagenomes), three had functional annotation. However, other gene clusters with identical predicted functions were found in other *Rothia* genomes; therefore we did not hypothesize that these three gene clusters conferred functions potentially important for differential survival in the buccal mucosa environment. 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 internal consistency of a gene cluster can be investigated by inspecting the alignment of its constituent amino acid sequences. An alignment of homologs should produce clearly conserved regions across the majority of the sequence with few gaps. For instance, in our investigation of the *Haemophilus parainfluenzae*, we noticed that the TD-abundant strains were characterized by three gene clusters encoding the three subunits of oxaloacetate dehydrogenase. A single non-TD strain (*Haemophilus parainfluenzae* C2004002729) also contained one of the three gene clusters. By inspecting the sequence alignment, which can be obtained from the aa_sequence column of Additional File 7 by searching for “oadA” or “GC_00001928” in Additional File 7, we discovered that the sequence from the non-TD genome was aberrant relative to the other sequences with numerous gaps and many mismatches. Based on the poor alignment, the inclusion of this non-TD gene sequence in the gene cluster likely reflects mis-assignment to this gene cluster. We therefore consider the oxaloacetate decarboxylase operon as exclusive to the genomes of TD-abundant strains.

***Functions of the core and accessory genome for* H. parainfluenzae *and* Rothia**

In addition to comparing differences between genomes based on gene content, we also investigated functional differences between core and accessory genes and between species of *Rothia* and strains of *H. parainfluenzae*.

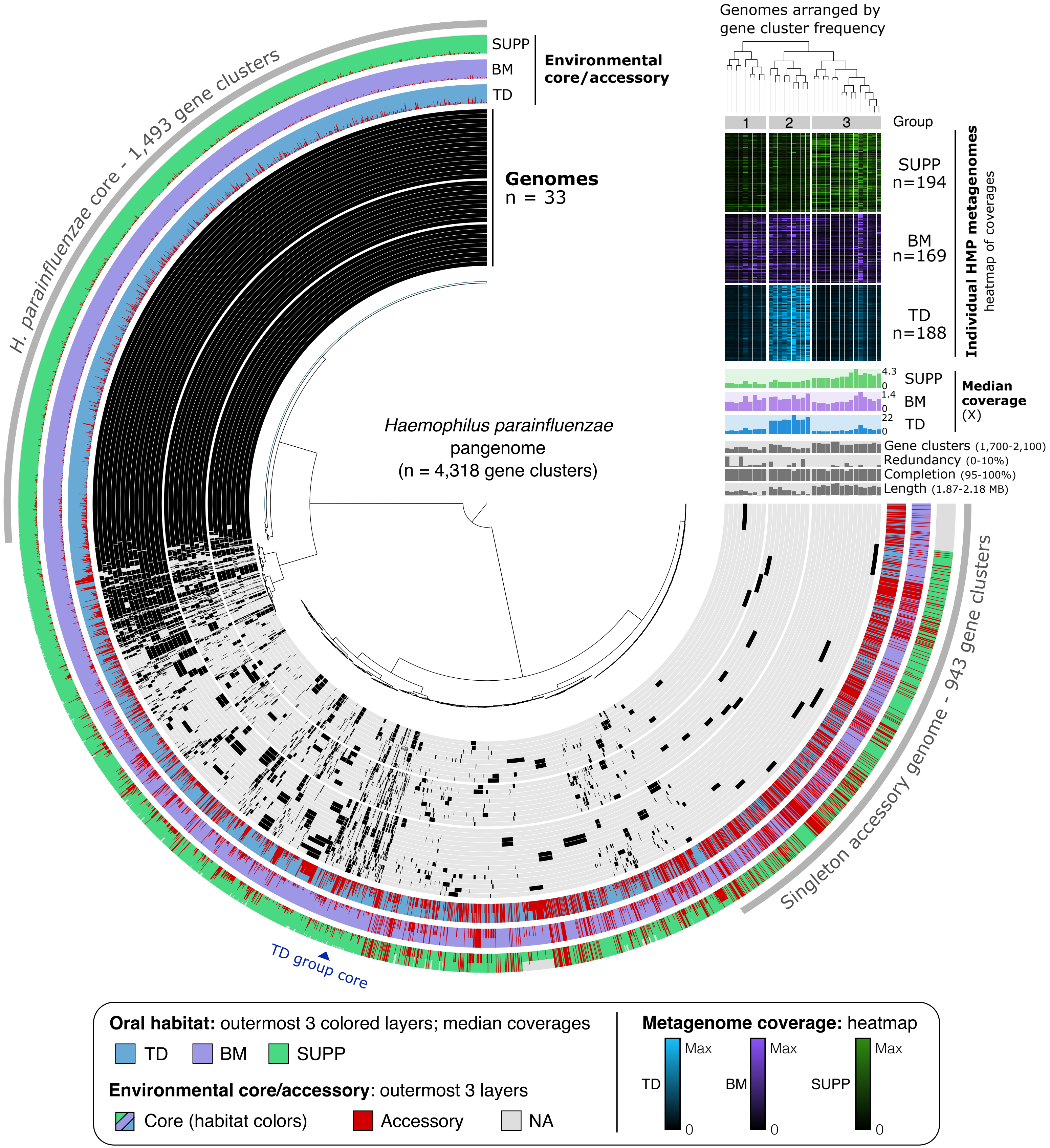
To investigate functional similarities and differences between core and accessory genes, we assessed the frequencies of each COG category in core, singleton accessory, and intermediate accessory genes as identified based on the pangenome. For simplicity we compared only genes assigned an single COG category and omitted genes that were assigned multiple COG categories. For *H. parainfluenzae*, the core consisted of gene clusters shared by all 33 genomes; the singleton accessory genome, gene clusters found in exactly one genome; and the intermediate accessory genome, gene clusters occurring in 2-32 genomes. Overall, each portion of the pangenome contained genes belonging to each COG category (Additional File 5A) but the frequencies differed. For example, genes involved in translation (J) and nucleotide metabolism (F) were both more numerous and proportionally more enriched in the core genome. On the other hand, defense mechanisms (V) and the mobilome (X) were more abundant in both the singleton and the intermediate accessory genome.

To investigate functional enrichment in one set of genomes compared to another, we recorded the proportion of genomes containing each TIGRFAM function. From this proportional data, the enrichment of each function in each group was determined using a logistic regression by the method of Shaiber et al. (2020). The full enrichment data is presented in Additional File 7 for each gene. To obtain a high-level view of which group(s) were more similar based on shared functions, we aggregated the enrichment scores by subtracting the mean proportional occurrence of each function in the group(s) in which it was not enriched from the mean of its proportional occurrence in the group(s) in which the TIGRFAM was enriched (Additional File 5B). For example, if a function was enriched in Groups 1 and 2 with a proportional occurrence of 1 and 0.8 in Groups 1 and 2 but also 0.1 in Group 3, the aggregate enrichment would be (0.8 + 1)/2 – 0.1 = 0.8. This aggregate enrichment of each function is shown in Additional File 5B. The three genes of the oxaloacetate operon unique to Group 2 stand out clearly, but more broadly the functional similarity between groups can be estimated. Group 2 and Group 3 share more genes with higher enrichment than do Group 1 and Group 2, or Group 1 and Group 3. This observation agrees with the arrangement of genomes based on gene cluster content shown as the dendrogram arranging genome layers in Figure 2, which places Group 2 sister to Group 3.

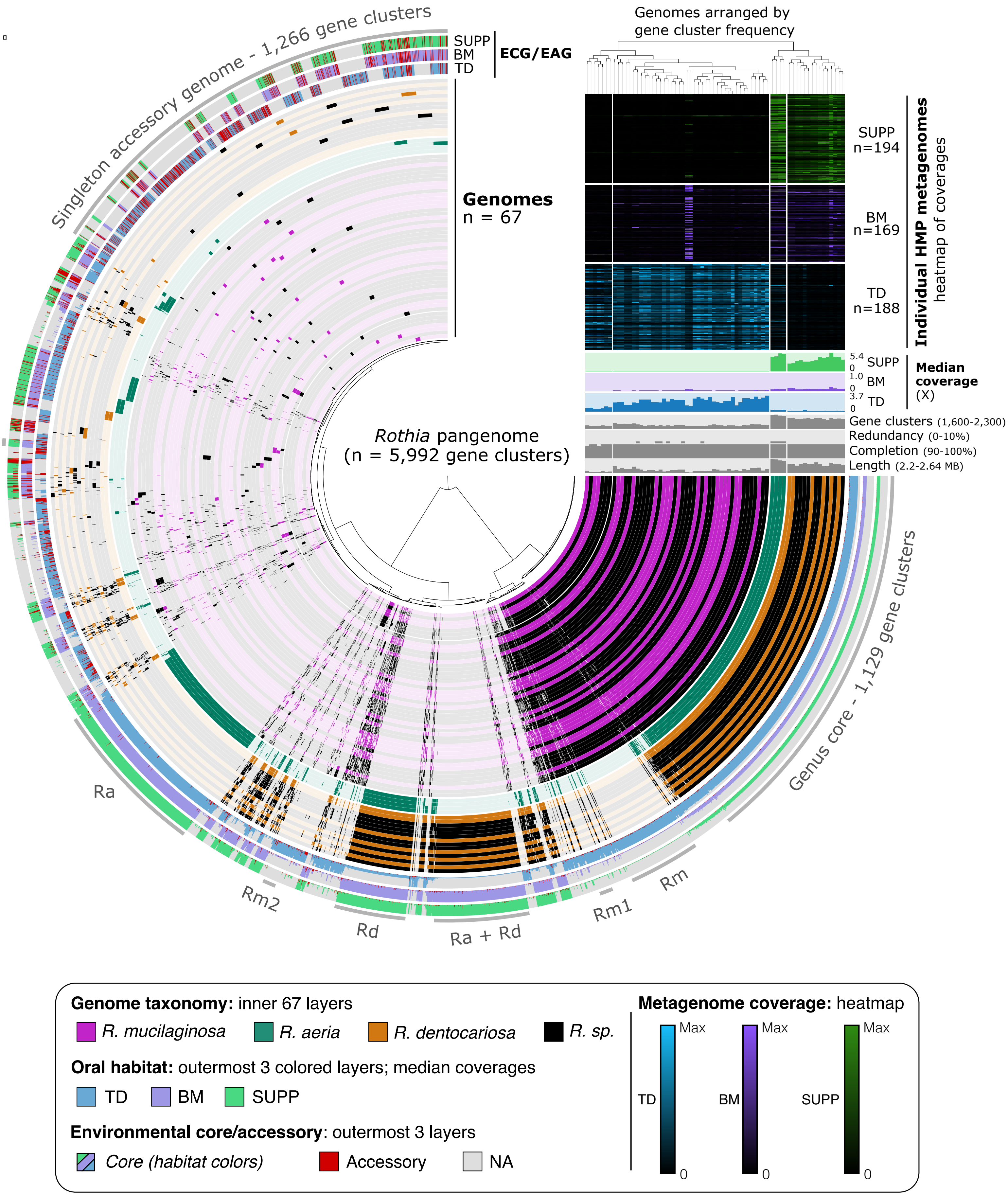
Functional enrichment analysis indicated that *Rothia* species with similar gene cluster content also contained similar functions. Predicted TIGRFAM functions were used to apply the same functional enrichment analysis as for *H. parainfluenzae*, but this time the groups were the three *Rothia* species (Additional File 5C, Additional File 12). Unlike the *H. parainfluenzae* analysis, the number of genomes per group varied much more substantially, with 48 *R. mucilaginosa*, 15 *R. dentocariosa*, and 4 *R. aeria* genomes. Yet, *R. dentocariosa* and *R. aeria* were still more functionally similar than either were to *R. mucilaginosa* based on aggregate enrichment scores (Additional File 5C), agreeing with the similarity of *R. dentocariosa* and *R. aeria* genomes based on gene cluster content (Figure 3 dendrogram).

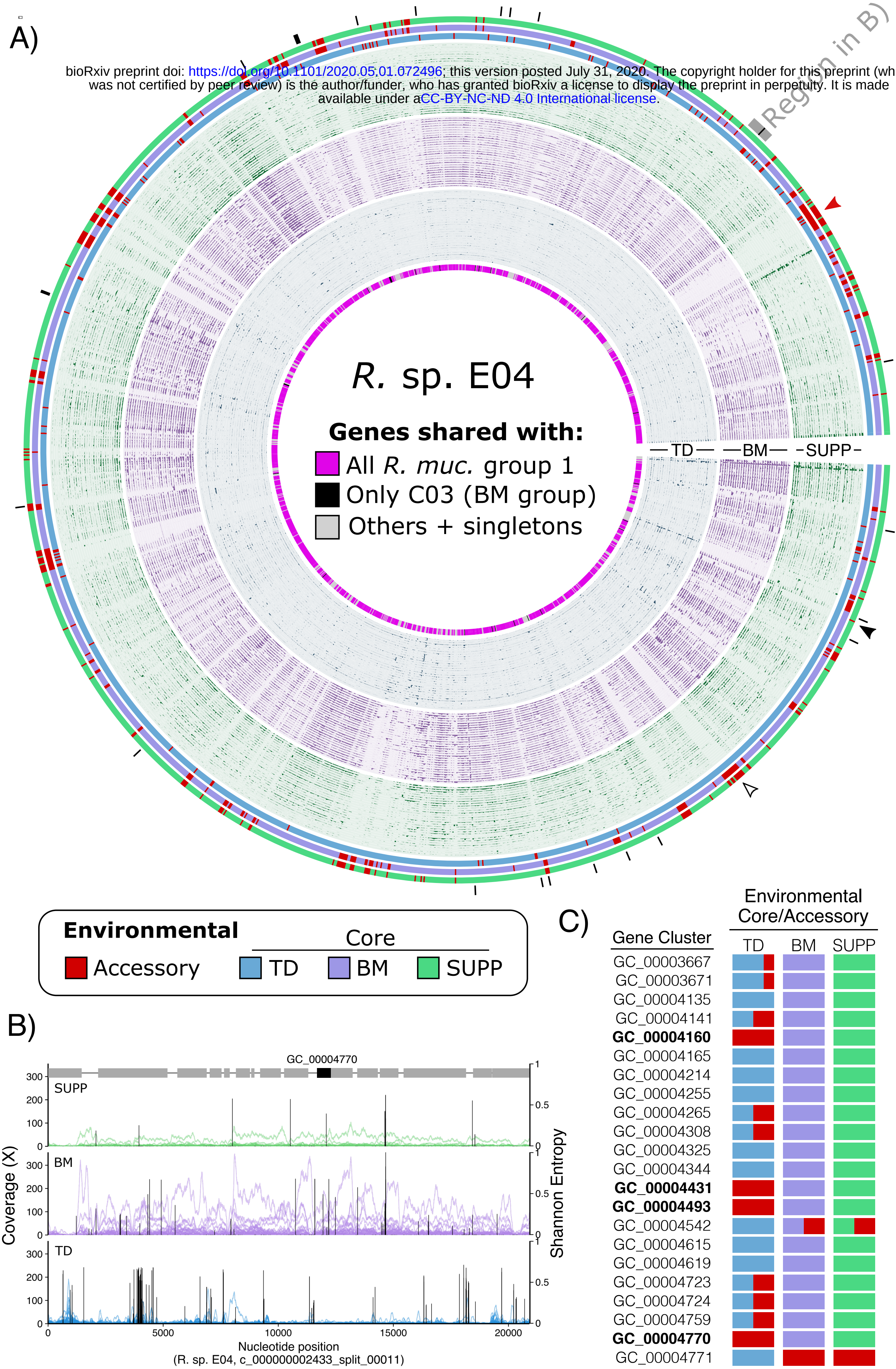
The functions enriched in each species also revealed possible sources of niche differentiation. Two functions were found in all 15 *R. dentocariosa* genomes but no other *Rothia* species, a PTS-system sucrose transporter component and a transcription repressor gene (Additional File 12). Further, of the 13 functions core to all *R. dentocariosa* and *R. aeria* genomes but absent from all *R. mucilaginosa* genomes, three were cytochrome related (Additional File 12). As both *R. dentocariosa* and *R. aeria* appear most abundant in plaque (Figure 3 heatmap), these cytochrome differences relative to *R. mucilaginosa* could potentially reflect selection by the different oxygen conditions of their respective microhabitats within tongue and plaque habitats.

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
Environmental
Core
Accessory
3
Genes of
strain 1
C) Metapangenome construction
Environmental
Core
Accessory
Gene Clusters
Genes in gene cluster
3
2
1
Normalized (shown in layer)
1.0
0.5
0.0
GC_001



H. parainfluenzae core - 1,493 gene clusters
SUPP
BM
TD
Environmental core/accessory
Genomes
n = 33
Haemophilus parainfluenzae pangenome (n = 4,318 gene clusters)
Genomes arranged by gene cluster frequency
1
2
3
Group
SUPP n=194
BM n=169
TD n=188
Individual HMP metagenomes
heatmap of coverages
4.3
0
SUPP
1.4
0
BM
22
0
TD
Median coverage (X)
Gene clusters (1,700-2,100)
Redundancy (0-10%)
Completion (95-100%)
Length (1.87-2.18 MB)
Singleton accessory genome - 943 gene clusters
TD group core
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
Max
TD
0
Max
BM
0
Max
SUPP
0


Singleton accessory genome - 1,266 gene clusters
SUPP
BM
TD
ECG/EAG
Genomes
n = 67
Genomes arranged by gene cluster frequency
SUPP n=194
BM n=169
TD n=188
Individual HMP metagenomes heatmap of coverages
5.4 0 SUPP
1.0 0 BM
3.7 0 TD
Median coverage (X)
Gene clusters (1,600-2,300)
Redundancy (0-10%)
Completion (90-100%)
Length (2.2-2.64 MB)
Rothia pangenome (n = 5,992 gene clusters)
Rm BM
Ra
Rm2
Rd
Ra + Rd
Rm1
Rm
Genus core - 1,129 gene clusters
Genome taxonomy: inner 67 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
Max
TD
0
Max
BM
0
Max
SUPP
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)
GC_00004770
SUPP
BM
TD
Coverage (X)
Shannon Entropy
0
100
200
300
0.5
1
0
5000
10000
15000
20000
Nucleotide position
(R. sp. E04, c_000000002433_split_00011)
C)
Environmental Core/Accessory
Gene Cluster
TD
BM
SUPP
GC_00003667
GC_00003671
GC_00004135
GC_00004141
GC_00004160
GC_00004165
GC_00004214
GC_00004255
GC_00004265
GC_00004308
GC_00004325
GC_00004344
GC_00004431
GC_00004493
GC_00004542
GC_00004615
GC_00004619
GC_00004723
GC_00004724
GC_00004759
GC_00004770
GC_00004771