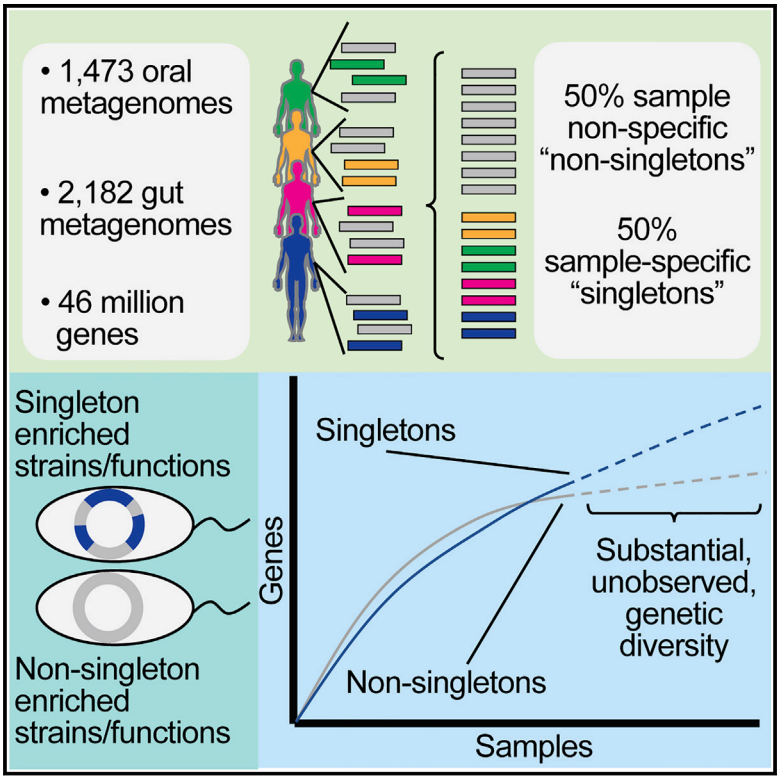
Resource

Cell Host & Microbe

The Landscape of Genetic Content in the Gut and Oral Human Microbiome

Graphical Abstract



Highlights

- Cross-study meta-analysis of metagenomes covering 3,655 samples from two body sites
- Meta-analysis uncovers staggering microbial gene diversity
- 50% of all genes in a metagenomic sample are individual-specific or "singletons"
- Individual's microbiomes can be fingerprinted via rare microbial strains

Authors

Braden T. Tierney, Zhen Yang, Jacob M. Luber, ..., Eleanor Mehlenbacher, Chirag J. Patel, Aleksandar D. Kostic

Correspondence

Chirag_Patel@hms.harvard.edu (C.J.P.), Aleksandar.Kostic@joslin.harvard.edu (A.D.K.)

In Brief

Tierney et al. presents a meta-analysis of metagenomes covering 3,655 samples from two body sites. They identify 45,666,334 non-redundant genes in the human oral and gut microbiome, and half of every person's microbial gene content is completely unique. These rare genes, denotes singletons, predominantly arise from extremely rare microbial strains.

Tierney et al., 2019, Cell Host & Microbe *26*, 283–295
August 14, 2019
https://doi.org/10.1016/j.chom.2019.07.008

CellPress

Cell Host & Microbe
Resource

The Landscape of Genetic Content in the Gut and Oral Human Microbiome

Braden T. Tierney,[1,2,3,4] Zhen Yang,[1,2,3,5] Jacob M. Luber,[1,2,3,4] Marc Beaudin,[1,2,3,6] Marsha C. Wibowo,[1,2,3] Christina Baek,[7] Eleanor Mehlenbacher,[8] Chirag J. Patel,[4,*] and Aleksandar D. Kostic[1,2,3,*]
[1]Section on Pathophysiology and Molecular Pharmacology, Joslin Diabetes Center, Boston, MA, USA
[2]Section on Islet Cell and Regenerative Biology, Joslin Diabetes Center, Boston, MA, USA
[3]Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA, USA
[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[5]Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada
[6]Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada
[7]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA
[8]Boston, MA
*Correspondence: Chirag_Patel@hms.harvard.edu (C.J.P.), Aleksandar.Kostic@joslin.harvard.edu (A.D.K.)
https://doi.org/10.1016/j.chom.2019.07.008

SUMMARY

Despite substantial interest in the species diversity of the human microbiome and its role in disease, the scale of its genetic diversity, which is fundamental to deciphering human-microbe interactions, has not been quantified. Here, we conducted a cross-study meta-analysis of metagenomes from two human body niches, the mouth and gut, covering 3,655 samples from 13 studies. We found staggering genetic heterogeneity in the dataset, identifying a total of 45,666,334 non-redundant genes (23,961,508 oral and 22,254,436 gut) at the 95% identity level. Fifty percent of all genes were "singletons," or unique to a single metagenomic sample. Singletons were enriched for different functions (compared with non-singletons) and arose from sub-population-specific microbial strains. Overall, these results provide potential bases for the unexplained heterogeneity observed in microbiome-derived human phenotypes. One the basis of these data, we built a resource, which can be accessed at https://microbial-genes.bio.

INTRODUCTION

Recent studies have made great strides in deepening our understanding of the strain-level diversity within the human gut microbiome, and 150,000 and 92,143 distinct microbial strains in two large meta-analyses, respectively, have been identified since the beginning of 2019 alone (Almeida et al., 2019; Pasolli et al., 2019). Additionally, others have demonstrated the importance of minute gene-level variation across strains in human health and disease (Zeevi et al., 2019). However, the implications of these discoveries for the overall microbial gene content of the human microbiota remains unexplored. The field still does not have a grasp on the scope of the microbiome's genetic content—in the gut and otherwise—a question crucial for understanding microbial function in the context of host disease (Sandoval-Motta et al., 2017).

The total number of distinct genetic elements within all prokaryotes is currently unknown, and theoretical estimates start at one billion genes (Wolf et al., 2016); (Lapierre and Gogarten, 2009) and range to maxima defined by permutations of nucleotide arrangements or thermodynamic stability in the context of protein folding (Lapierre and Gogarten, 2009). Specifically, in the human microbiome, most metagenomic analyses and methods that consider genes focus on core gene families (Lloyd-Price et al., 2017; Truong et al., 2015), where a core gene is defined as being present once, not a paralog, and more similar to its orthologs than any other gene in any other species (Young et al., 2006; Tettelin et al., 2005). Others have addressed metagenomic gene content by producing "gene catalogs," the set of all genes identified via assembly across a large number of samples. Within the human gut microbiome, up to 10 million non-redundant genes have been identified by major sequencing consortiums using *de novo* approaches (Dusko Ehrlich and The MetaHIT Consortium, 2011; Forster et al., 2016; Li et al., 2014; Nielsen et al., 2014; Qin et al., 2010). These efforts have been almost exclusively associated with the gut microbiome, are relatively limited in terms of sample sizes, and do not focus on the overall rarity of genes across a population.

Moreover, there is a need to link our understanding of metagenomics back to that of traditional microbial genetics. Microbial genetic elements can be grouped into "pan-genomes," which describe the set of all genes found in all strains of a particular species (Tettelin et al., 2005). The size of a pan-genome is most influenced by its effective population size and ability to migrate to new niches (McInerney et al., 2017). However, other intermittently present genes contribute significantly to the size and function of the pan-genome. In newly sequenced prokaryotic isolate genomes, up to a third of these genes have no detectable homologs in other species (Daubin and Ochman, 2004; Yin and Fischer, 2006). These "ORFans" are distinct from all open reading frames (ORFs) in the genome and are hypothesized to be neutral to selection pressure (i.e., ORFans are replaced at

Table 1. Table of Definitions Used in the Paper

Term	Definition
metagenome	total genomic potential of a microbial community (in this work, we use this term interchangeably with "sample" and "metagenomic sample")
singleton gene	a gene detected in only one metagenomic sample across a defined collection of samples
non-singleton gene	a gene detected in more than one metagenomic sample across a defined collection of samples
ORFan gene	genes that have no detectable homologs in other species and are distinct from all open reading frames (ORFs) in the genome
universe of genes	the set of all non-redundant genetic elements across all communities of organisms in a given niche
gene rarefaction curve	a curve tracking the accumulation of new genes as samples are incrementally added
gene discovery curve	the derivative of the rarefaction curve (It estimates the rate at which new genes are added to the catalog when samples are added incrementally, and it can be used to estimate the size and burden of sampling of the universe of genes.)
singleton fraction curve	a curve estimating the fraction of a gene catalog that consists of singletons versus non-singletons as samples are added incrementally (It is used to estimate the total number of samples that would be required for all singletons to be seen twice and thus no longer be singletons.)
mixture contig	a contig from *de novo* assembly consisting of both singletons and non-singletons
singleton contig	a contig from *de novo* assembly consisting of only singletons
non-singleton contig	a contig from *de novo* assembly consisting of only non-singletons

the natural rate of DNA uptake, recombination, and loss) (Wolf et al., 2016). With an increasing emphasis in the field on the importance of strain-level variation in the gut microbiome (Zhao et al., 2019), there is a need to identify the contribution of ORFan-like genes to overall metagenome gene content. We hypothesized, especially given the recent discoveries of massive strain diversity in the gut, that these genes would increase variation in gene content of the human microbiome.

Here, we sought to build a multi-body site microbiome gene catalog as a publicly available resource for the scientific community. We further aimed to use this catalog to identify and taxonomically and functionally document the metagenomic analogs of ORFan genes. Then, with ORFans in mind, we attempted to determine the scale of sequencing that would be required to sufficiently sample the total genomic content—the universe of genes—of each niche, therefore building a "complete" gene catalog of the human microbiome.

RESULTS

A Pan-microbiome Genetic Database

Like prior gene catalog analyses, we utilized a *de novo* approach (as, by design, reference-based approaches only detect genes present in a reference database) to construct non-redundant microbiome gene catalogs from publicly available short read data. We aggregated 2,183 samples from 6 gut microbiome studies. For the oral microbiome dataset, we retrieved 1,473 oral microbiome metagenomic samples from 7 studies, a cohort ~2× larger than the largest consortium effort to study this niche (Lloyd-Price et al., 2017). For a table of definitions used in this paper, please see Table 1.

We performed a meta-analysis of this aggregated metagenomic data, *de novo* assembling each metagenome (Figures 1A–1D; Table S1). This analysis uncovered a universe of prokaryotic genes massive in scale. Extending existing approaches (Li et al., 2014; Nielsen et al., 2014; Qin et al., 2012), we initially defined a unique gene as being distinct from all other ORFs at the 95% identity level. Overall, we predicted 157,241,550 ORFs from the assembled oral data, compared with 136,672,846 from the gut data. Clustering at the 95% identity threshold, the initial oral and gut catalogs contained 23,961,508 and 22,254,436 consensus genes, respectively. When these oral and gut catalogs were clustered together at 95% identity, the resultant, non-redundant catalog had 45,666,334 genes, given that at this percent-identity cutoff 549,610 ORFs overlapped (Figure 2A).

Using this final catalog, which is replete with functional and taxonomic annotations, we built a publicly available and searchable PostgresQL database with an associated front-end that contains summary data (i.e., gene counts per body site, average gene length, number of genes in each consensus gene cluster, etc.) as well as information on our pipelines (Figure 1E). Our database has 2,418 different gene EC Numbers (Bairoch, 2000), 222,308 unique gene annotations, and 15,746 NCBI taxonomies annotated within it. We additionally report consensus gene sequences and the number of genes in each 95% identity cluster. Finally, we also have made available for download MetaPhlAn2 (Truong et al., 2015) output for each sample and all of the gene catalogs generated in the latter sections of this study.

The Oral and Gut Microbiomes Contain Vast and Individual-Specific Genetic Content

We explored the reasons behind the substantial size of these gene catalogs. We hypothesized this effect was driven by the metagenomic equivalent of ORFan genes. As such, we sought to determine the frequency of occurrence of each gene on a sample-by-sample basis. Some genes assembled in multiple metagenomic samples (non-singletons), whereas other genes were found in exactly one sequencing sample (singletons). The oral gene catalog contained 11,891,670 (49.6%) singletons and 12,069,838 (50.4%) non-singletons, whereas the gut gene catalog contained 12,621,933 (56.7%) singletons and 9,632,503

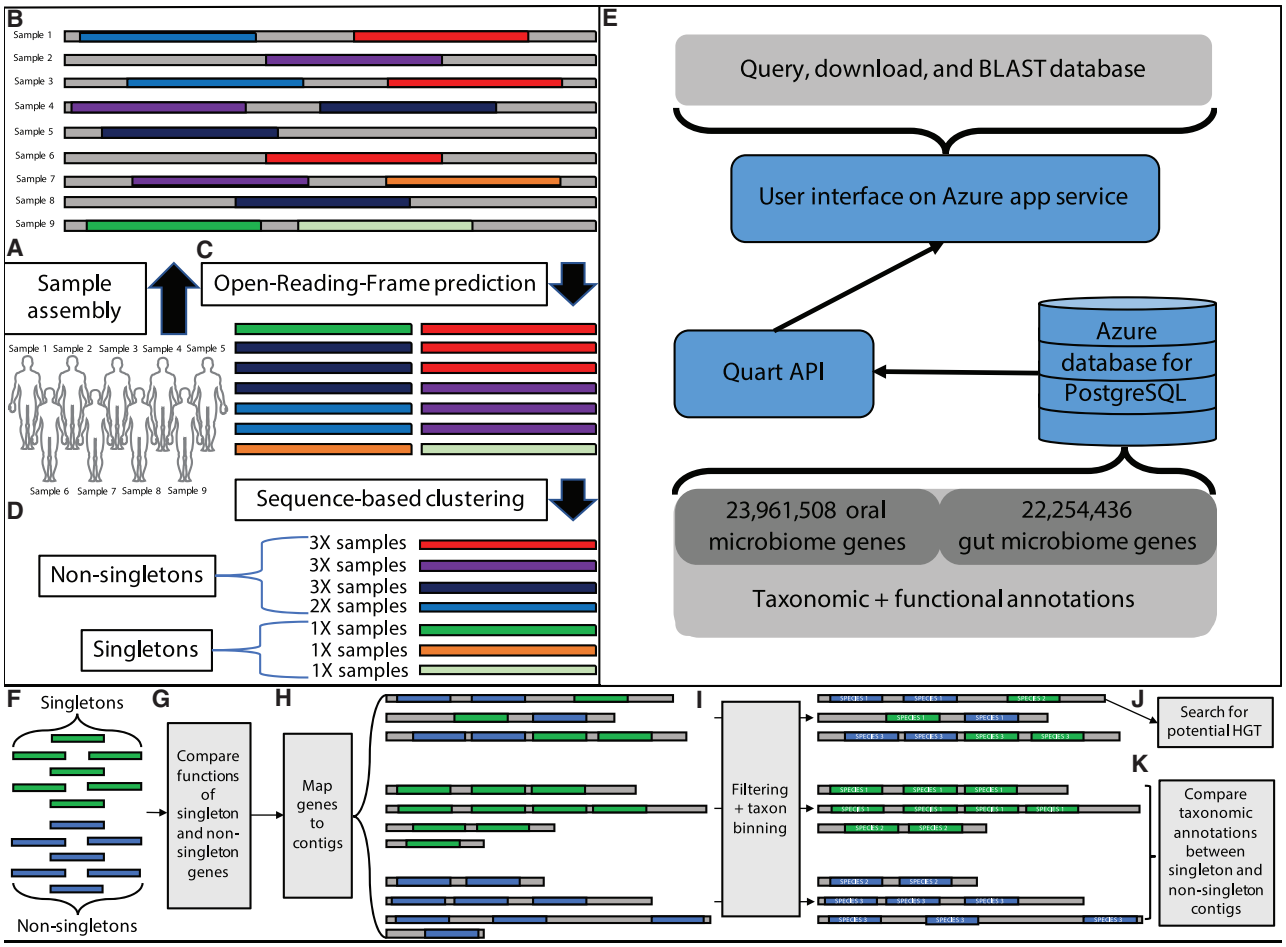


Figure 1. Meta-analysis of the Oral and Gut Microbiomes

(A and B) We aggregated publicly available oral and gut short read data and assembled it into contigs (in this example, each contig comes from a single sample).

(C) Gene open-reading-frames (ORFs) are identified on assembled contigs.

(D) ORFs are clustered at 95% identity to identify a non-redundant gene catalog.

(E) Database content, description of backend, description of user interface (UI).

(F–K) Downstream singleton analytical pipeline. In (F), we identify singletons and non-singletons in our dataset and in (G) compare their functional annotations. In (H), we then map genes to contigs, which we grouped into 3 categories: singleton-contigs (those consisting of only singletons), non-singleton contigs (those consisting of only non-singletons), and mixture contigs (those consisting of both singletons and non-singletons). In (I), we filter short contigs and bin the remainder according to the taxonomic classification of their gene content. We then attempted to identify the source of singletons as either (J) horizontal gene transfer (HGT) and/or (K) rare, singleton-rich microbial strains.

(43.2%) non-singletons (Figure 2B). On average, 2.9% of the genes in each sample were singletons (standard deviation μ3.5%).

We carried out substantial analysis on synthetic and real data with different assemblers and parameters to determine if singleton genes were artifacts of our analytic pipeline or false positive or short or low coverage genes. We found that singletons had modest associations with false positive genes, low coverage genes or contigs, short genes or contigs, or particular assemblers or assembly parameters compared with non-singletons. (Table S2; Figures S1–S3). We additionally sought to determine whether prior gene catalog analyses contained singletons and found that the Metahit Integrated Gene Catalog (Li et al., 2014) contained is 46% singletons (out of a total of 9.9 million genes) (Figure S3F). Second, we tested whether singleton identification could be explained by low depth of sequencing. If that were universally true, singletons could be present in many samples just below the threshold of detection by assembly. We were unable to identify a strong correlation (Spearman correlation: 0.22, $p < .05$) between total read count and singleton gene count within a sample (Figures S3G–S3J), implying depth alone is not driving singleton presence. Finally, to confirm whether the parameters for our choice of assembler, MEGAHIT, was supported by the literature, we reviewed every study (n = 99, 67 of which we had access to and were not dissertations or books) currently citing the MEGAHIT publication and determined similar projects used the same assembly settings (Table S3).

We next relaxed the gene catalog clustering identity threshold to determine if ORFans (singletons) were artifacts of high percent

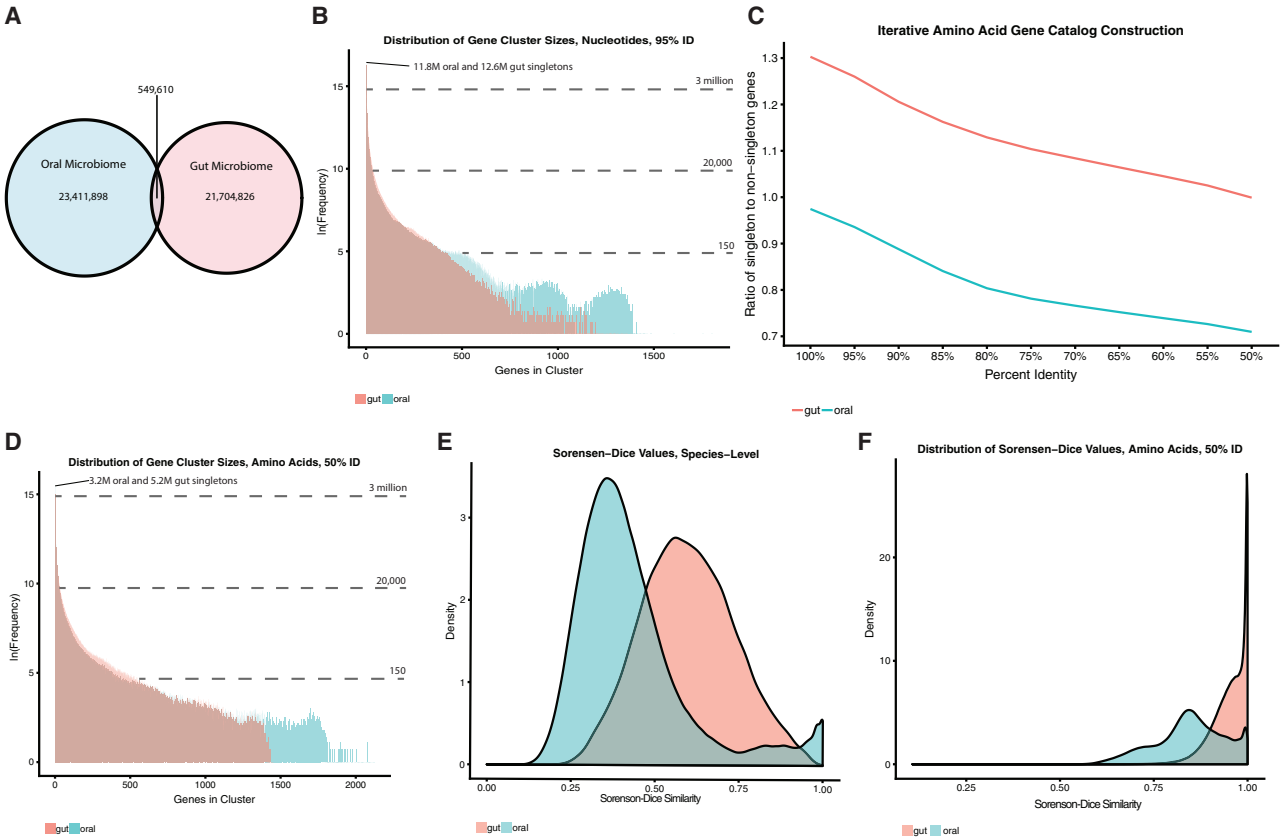


Figure 2. The Genetic Diversity of the Oral and Gut Microbiomes

(A) The overlap in genetic content (95% identity level) between the oral and gut microbiomes.

(B) Distribution of ORF cluster sizes at 95% identity in our oral (blue) and gut (red) gene catalogs.

(C) Iterative clustering of our amino acid gene catalogs.

(D) Distribution of gene cluster sizes for amino acid gene catalogs generated at the 50% identity level.

(E) Sorensen-Dice index measuring dissimilarity in gene content between all pairs of individuals.

(F) Sorensen-Dice dissimilarity of individuals in terms of MetaPhlAn2-derived species content.

identities (Figure 2C). To circumvent computational limitations of clustering in nucleic acid space, we first translated the nucleic acid catalog to amino acids and lowered the clustering threshold from 100% amino acid identity to the limit of reasonable computational feasibility, 50% identity. Although the catalog size shrank with the lower identity thresholds as expected, the fraction of singleton genes in the catalog remained approximately constant, particularly at lower percent identities, reflecting that the high proportion of singleton genes was not influenced by clustering thresholds. At 50% identity, the oral microbiome gene catalog contained 7,842,539 consensus genes, 3,255,115 (41.5%) of which were singletons (compared with 10,465,169 genes, 49.9% singletons, in the gut) (Figure 2D).

Notably, although the oral gene catalog was larger at 95% nucleotide identity and contained more singletons, it was smaller than the gut catalog (and contained fewer singletons) at 50% identity, implying overall lesser overall sequence variation at low percent identities in the former than the latter. For the remainder of this manuscript, singleton, and non-singleton genes will refer to those generated at the 50% clustering level, unless otherwise specified.

We next sought to determine whether subjects (human hosts) with similar reference-based species content, which we identified using MetaPhlAn2, also had similar genetic content. We found this not to be the case. Using Sorensen-Dice dissimilarity (where 0 is identical and 1 is most dissimilar), we found that the human microbiome exhibits more inter-individual similarity of overall species content (mean Sorensen-Dice oral = 0.43, mean Sorensen-Dice gut = 0.60) (Figure 2E) versus that of genes (mean Sorensen-Dice oral = 0.85, mean Sorensen-Dice gut = 0.95) (Figure 2F). Moreover, we found that most samples were equally dissimilar from each other, and the presence of singletons could not be explained by a few completely distinct samples in our dataset. Lastly, while genetic content varied between samples, singleton genes were evenly distributed throughout the sample population (Figures S3G and S3H; Table S1).

Singletons Are Functionally and Taxonomically Distinct from Non-singletons

Further, we collapsed each gene annotated with EC numbers (Bairoch, 2000) from Prokka into Minpath (Ye and Doak, 2009) annotations. Overall, 12.8% of singletons in the mouth and

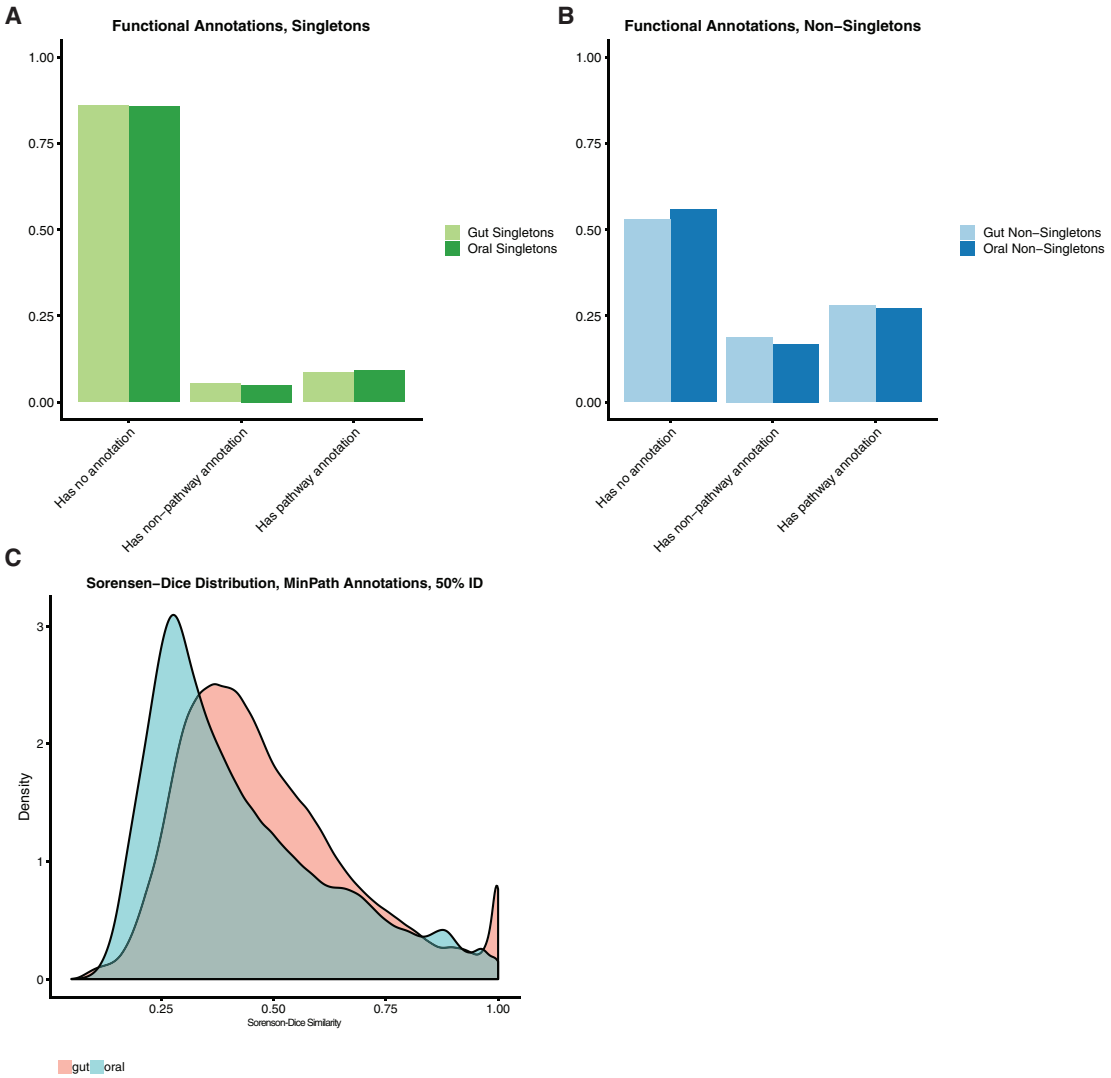


Figure 3. The Known and Unknown Functional Diversity of the Oral and Gut Microbiomes

(A and B) Fractions of singletons (A) and non-singletons (B) functionally annotated in the oral and gut microbiomes. Genes labeled with pathway annotations were used in the Minpath analyses.

(C) Sorensen-dice dissimilarity of individuals in terms of overall pathway content.

12.9% in the gut were functionally annotated by Prokka, compared with 36.7% of oral non-singletons and 34.6% of gut non-singletons (Figures 3A and 3B). While we were limited by relatively scant functional annotation information, we sought to test, using Sorensen-Dice dissimilarity, whether individual samples had, on average, the same pathways. We found this as well not to be the case (mean oral = 0.43, mean gut = 0.29) (Figure S4A).

We sought to taxonomically and functionally characterize the singletons that remained in the 50% identity amino acid catalog. We compared the enrichment of functional annotations across singleton and non-singleton genes (Figures 1F, 1G, and 4; Table S4). We found non-singletons and singletons to have little overlap in their functional diversity. In the top 50 most enriched Minpath classes for gut and oral non-singletons, 27 overlapped, whereas only 9 of the top 50 oral and gut singleton enriched pathways did. Overall, non-singletons were enriched for primary metabolic processes, such as the Citric Acid Cycle and amino acid biosynthesis, whereas singletons were enriched for a wide range of diverse biosynthesis and degradation pathways.

In addition to a subset of singletons arising from genes with divergence greater than 50% identity, we hypothesized that singletons might arise from (1) horizontal gene transfer (HGT) or (2) extremely rare microbial strains, or some combination of the two. To test these hypotheses, we mapped genes back to their original contigs and classified contigs that arose exclusively from non-singletons (73M) (i.e., only containing non-singletons), exclusively from singletons (2.5M), and contigs arising from both (1M) (Figure 1H). 78.7% of singletons and 90.1% of non-singletons could be taxonomically annotated using NCBI's refseq database (Figures S4B and S4C). We grouped contigs (Figure 1I) by using these gene-level annotations and searched the

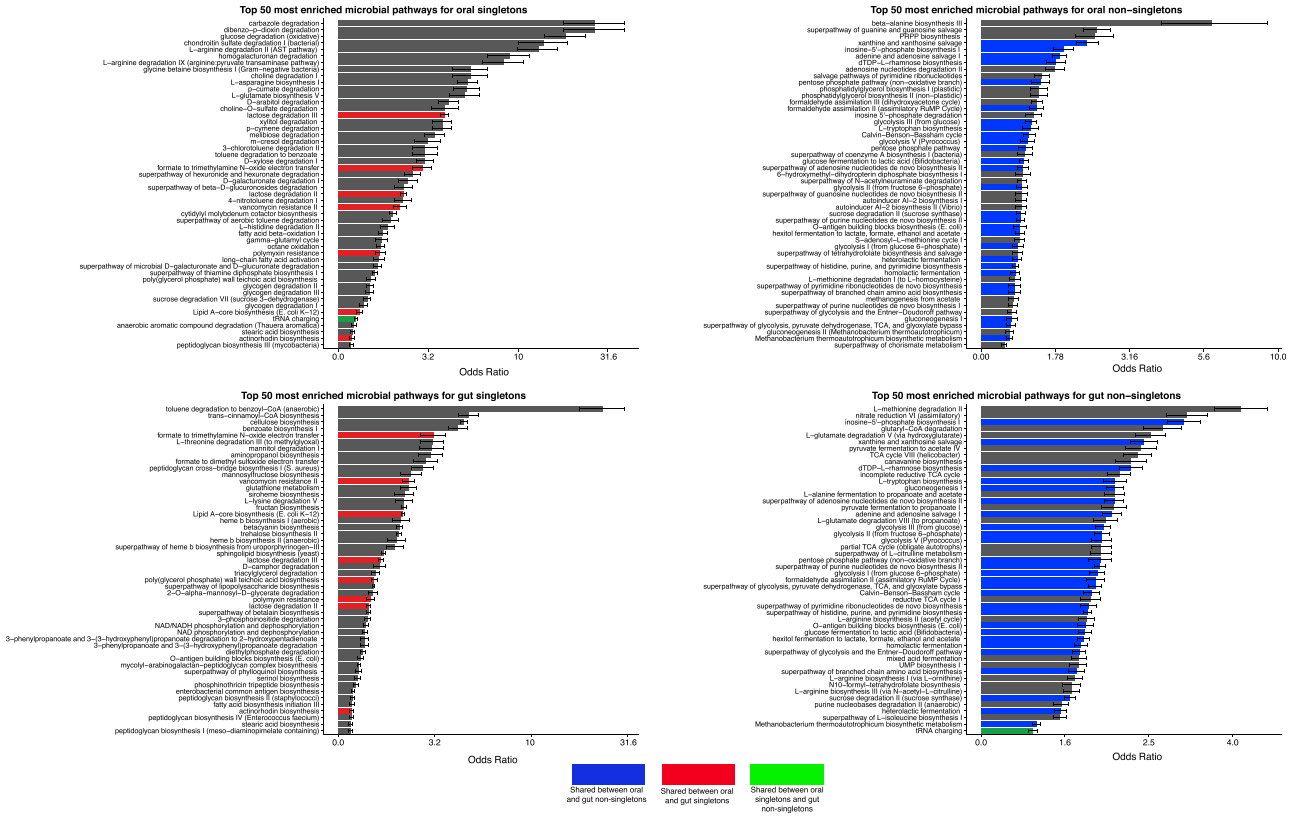


Figure 4. Enrichment of Functions in Gut and Oral Niches for Singletons and Non-Singletons

Here, we display the top 50 most enriched pathways for oral singletons (A), oral non-singletons (B), gut singletons (C), and gut non-singletons (D). Bars represent odds ratios from a Fisher's Exact Test and include 95% confidence intervals. Blue bars are pathways enriched in both oral and gut non-singletons, red bars are pathways enriched in both oral and gut singletons, and the green bar is a pathway enriched in both oral singletons and gut non-singletons.

resulting groups for evidence of horizontal gene transfer and taxonomic variation between singleton and non-singleton contigs (Figures 1J and 1K).

To test hypothesis (1) and screen for potential HGT, we searched for contigs consisting of both non-singletons and singletons where the non-singletons were annotated as coming from one species or genus and singletons were annotated to a different species or genus. We found that HGT did not contribute substantially to singleton presence. The genes on the contigs that were a mixture of singletons and non-singletons tended to emerge from the same species or genus. Only 8,557 (0.8%) of all mixture contigs in the oral microbiome contained potential cross-genus HGT. In the gut, there were 33,224 of these cross-genus, mixture contigs, a total of 1.8%.

Singletons Arise from Rare, Sub-population Specific Bacterial Strains

In testing our second hypothesis (highly uncommon microbial strains as the source of singletons), we identified differences in the taxa from which singleton-contigs and non-singleton-contigs originated. For each taxa, the singleton and non-singleton counts were in some cases modest, and we observed some rare taxa had more singleton than non-singleton contigs. The Pearson correlation between singleton-contig and non-singleton-contig counts for each taxa was 0.27 in the oral microbiome and 0.34 in the gut (Table S5; Figures S4D–S4L). We sought to identify whether the bias toward particular taxonomies was being driven by singletons arising from shorter contigs or contigs with fewer ORFs. We found this not to be the case (Figure S5; Table S6). In total, we found that contigs with greater than one gene mapped to 2,071 and 2,476 species-level taxonomic annotations in the oral and gut microbiomes, respectively. Of these, 1,155 (55%) species in the mouth and 1,648 (67%) in the gut had more singleton than non-singleton contigs. We refer to these contigs as arising from "rare strains," and from their presence concluded that hypothesis (2) was more likely than hypothesis (1).

Having found that singletons were enriched in different taxa than non-singletons, we sought to test whether singleton-only contigs came from sub-population-specific strains. The alternative would be that species that contained singleton contigs were evenly distributed across the population. To test this, we compared the number of samples in which singleton and non-singleton contigs with given taxonomic annotations appeared. On average, in both the oral and gut microbiomes (Figure 5A), we found singletons-contig-derived taxonomic annotations in fewer samples (oral_mean = 6.7, gut_mean = 8.3, Wilcoxon test $p < .05$) than non-singletons (oral_mean = 22.0, gut_mean = 25.0, Wilcoxon test $p < .05$), demonstrating that singleton-enriched taxa are uncommon with respect to the entire population. We further tested to see whether even

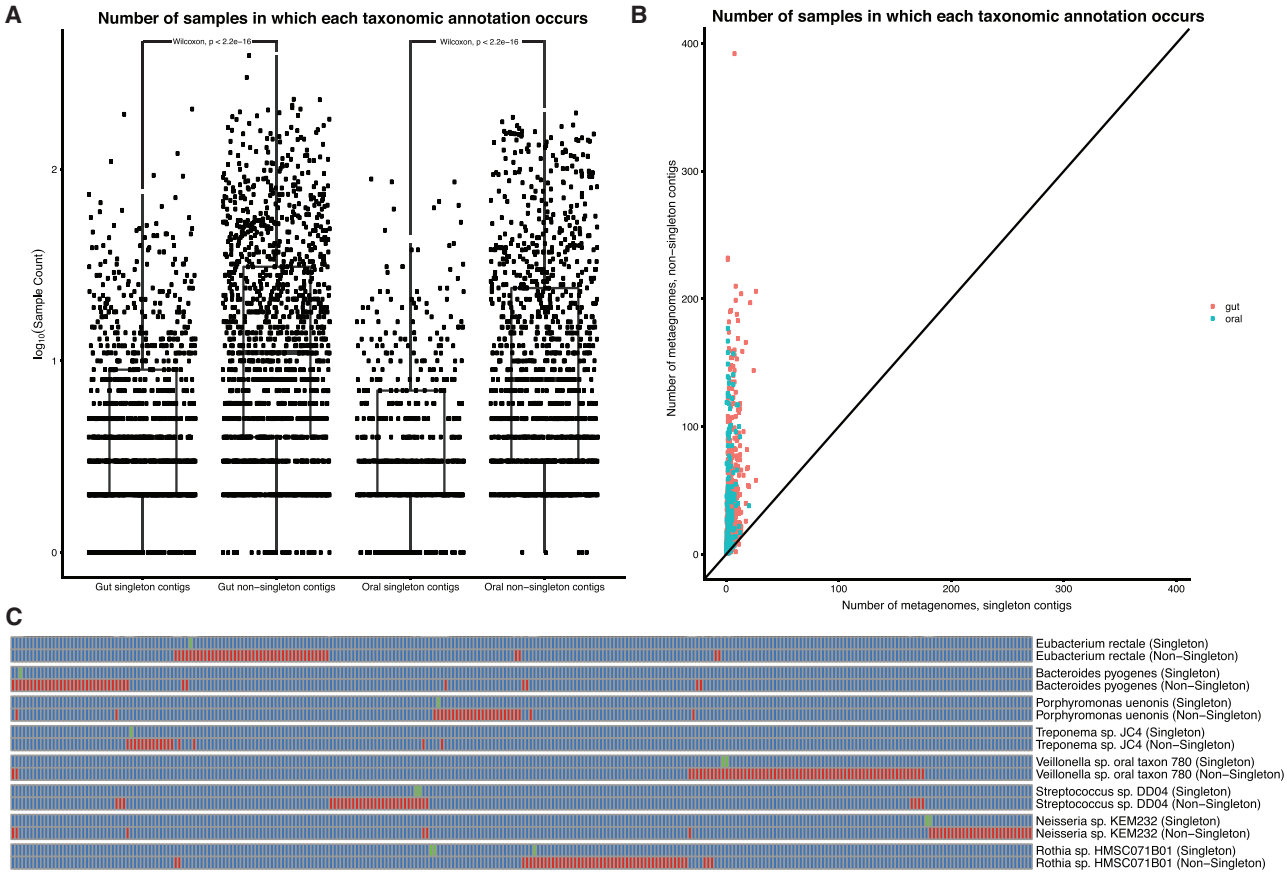


Figure 5. Singleton Taxa as Sub-population-Specific, Rare Strains

(A) Counts of taxonomic annotations for singleton and non-singleton contigs in the oral and gut microbiomes.

(B) Number of metagenomes singleton contigs and non-singleton contigs are present in for different taxonomies. Each point represents a different taxonomic annotation.

(C) Examples of strain-specific "fingerprints." Each pair of rows corresponds to singleton and non-singleton contigs containing at least two genes that were binned into the same taxonomic annotation. Columns are different metagenomic samples (each corresponding to a different individual). Green boxes correspond to singleton contigs. Red boxes correspond to non-singleton contigs.

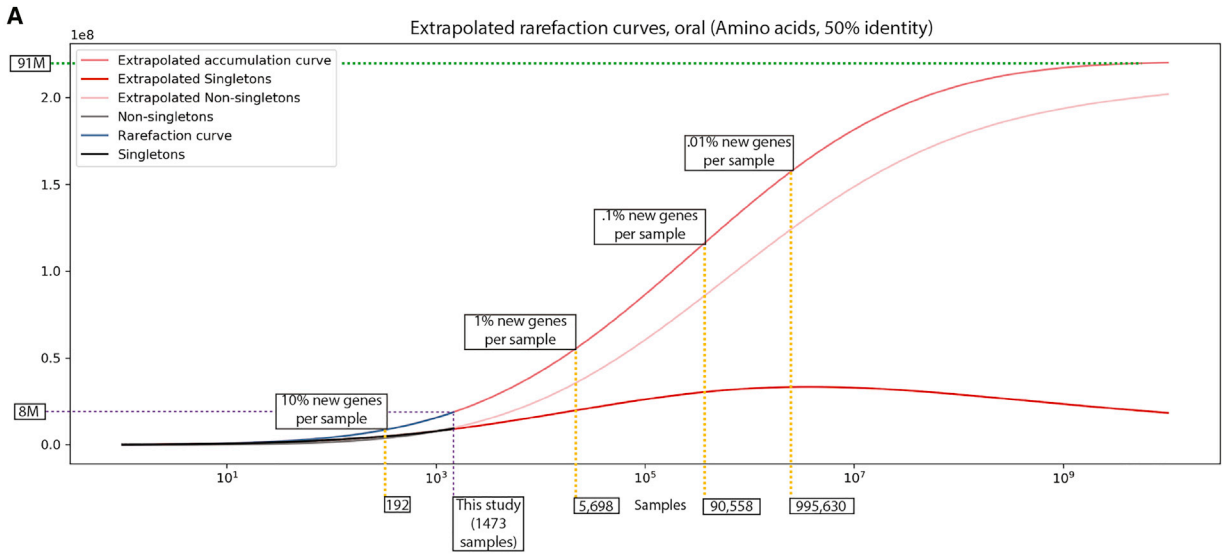
when singleton and non-singleton contigs mapped to the same taxonomies, singleton-rich strains still arose from specific individuals or sub-populations (Figure 5B). We found this to be the case; for example, 28 singleton and 42 non-singleton contigs map to *Eubacterium rectale*; however the singleton contigs come from 1 individual, whereas the non-singleton contigs are from 39 different individuals (Figure 5C).

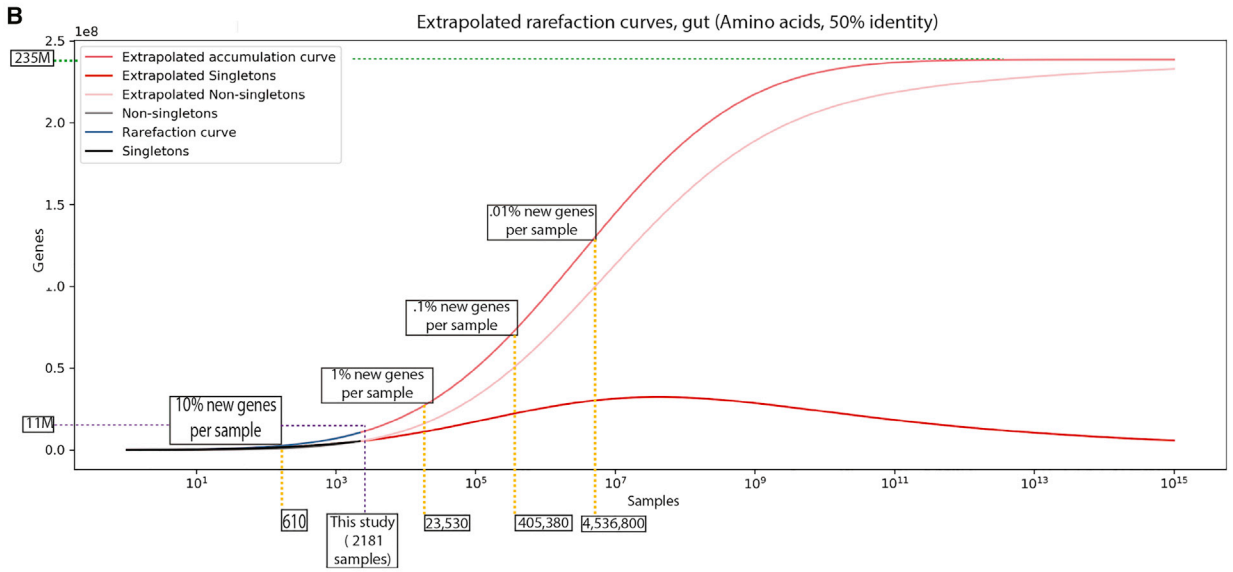
Estimating the Burden of Sequencing the Human Microbiome

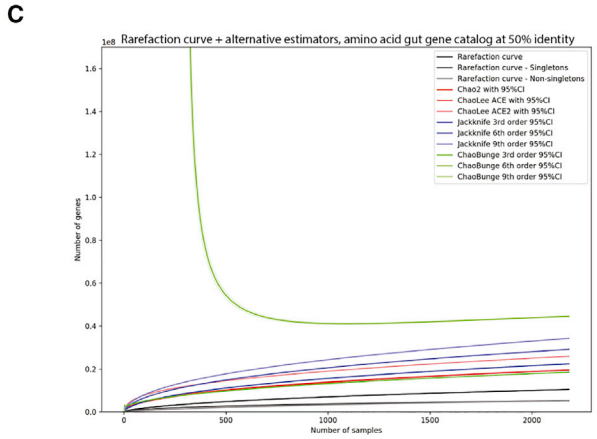
Given the size and heterogeneity of our gene catalog, we sought to identify the amount of sequencing that would be required to capture the entire "universe of genes" in the oral and gut microbiomes. We used 10 rarefaction methods to estimate the rate of accumulation of unique genes at the 50% identity level (Figures 6 and S6). We found imprecise estimates of the total gene content of the human microbiome. We claim this catalog is sampling between 8%–72% and 4%–50% of the total potential genetic richness of the gut and oral microbiomes, respectively. Assuming a constant rate of singleton accumulation with sampling, we estimate that to achieve a point where only 1% of genes per sample sequenced had not been seen before, we would need to sequence on the order of 5,698 samples in the oral microbiome and 23,530 samples in the gut. However, given the high variation associated with our extrapolations (which ranged from on the order of 40 million to 200 million) in Figure S17, we emphasize that these estimates could be off by up to an order of magnitude. Furthermore, they are dependent on parameters that are challenging to optimize, such as gene sequence percent identity threshold. Therefore, we can only conclude that gene variation within the human microbiome is vast and deeply uncertain in scope despite the relatively large sample sizes of our meta-analysis.

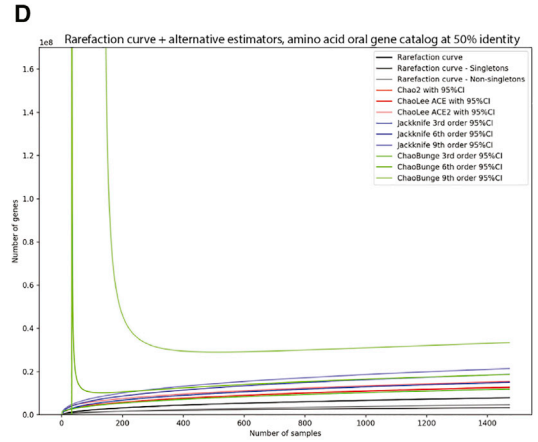
DISCUSSION

We have built a large microbiome-gene database that incorporates multiple human body sites clustered at a range of percent









(legend on next page)

identities. We built this resource with a focus on the variation of genomic content across the human microbiome, identifying an order of magnitude more genes in both the oral and gut microbiomes than ever before. We also identified singletons, which we propose are the metagenomic equivalent of ORFan genes. On analysis of our catalog, we find that the genetic richness in the human microbiome has been underestimated and undersampled, though estimating the degree and uncertainty of undersampling was nontrivial, despite this large collection of data.

In line with other recently published work (Almeida et al., 2019; Pasolli et al., 2019), our results indicate substantial strain-level diversity. For context, consider the following: suppose the average prokaryote has 5,000 genes (Land et al., 2015) and that 90% of genetic content is shared between genomes of a single species (Zhu et al., 2015). To explain the size of the 95% identity oral gene catalog (24 million genes), each of the 2,000 species we identified would require on the order of 20 sub-species/strains. If we were to only consider the 788 species identified by reference-based methods, each species would require on the order of 50 strains. Finally, outside of only showing diversity in strains, we were additionally able to show that strains rich in singletons can act as microbial fingerprints, tending to be unique to sub-populations within this dataset and in some cases even individuals.

Questions remain regarding best analytic practices for *de novo* metagenomic studies and, in the future, meta-analyses of metagenomic studies. Reference-genome-based approaches are superior to gene catalog analyses in terms of computational feasibility and interpretability; however, given the lack of observed correlation between taxonomy and genetic content, databases derived from primarily cultured isolates might lack many functionally important genes. As such, the successful biological interpretation of metagenomic findings is contingent upon building resources and databases with microbial genetic diversity in mind, considering both ORFans and otherwise.

We found that singleton genes are enriched in functionality for a variety of unrelated metabolic functions compared with non-singletons, which were enriched in more conserved bacterial processes. However, functions encoded by singletons are not irrelevant. We identified a number of pathways (e.g., antibiotic resistance and cell wall biosynthesis), that might affect both the structure of the microbiome and host health. The limited overlap in top enriched singleton functions between the oral and gut, compared with non-singletons, implies that singletons encode more niche-specific functions than non-singletons. As such, given the functional variety encoded within singleton genes, we propose that singletons form an evolutionary organ within the microbiome, one that can be leveraged by microbes to adapt readily to environmental conditions. It is possible that, analogously to recent work done in the field of human genetics (Wainschtein et al., 2019), ORFan genes might explain a large portion of the currently unexplained variation in microbiome-associated human disease states (Sandoval-Motta et al., 2017). Recent work has demonstrated that this might be the case. For example, sub-population specific and intransient strains are associated with human disease and colonization (Zeevi et al., 2019).

The definition of a singleton gene states that it was only observed in a single sample. Therefore, a gene could feasibly still be present in other samples in such low abundance that it cannot be identified via assembly. Although computing relative abundance is fraught with the challenge of spurious alignments, especially in the case of low abundance genes, it could partially address this issue and is a reasonable future direction for this work. However, given the low correlation identified between sample depth and singleton presence (Figure S3), we posit that forces other than undersampling are, at least in part, driving singleton presence in our data.

Overall, we have built a resource intended for studying gene-level variation across multiple human body sites and samples. We also showed that the gene landscape of the human microbiome is immense and that its heterogeneity across people is staggering. Moreover, we have quantified the need to increase sequencing efforts to fully explore both the oral and gut niches, as well as other body sites. Our findings imply that an order of magnitude more sequencing data (than currently exists) is necessary to sufficiently sample (with only 1% of genes being novel per metagenome) human microbiome sequence diversity and function at even the 50% identity level. That being said, clearly this estimation is immensely challenging because of both the variation in available modeling methods as well as the difficulty of sequence-identity-based microbial gene definitions (i.e., 95% versus 50% would yield vastly different results). It is also worth noting that despite large samples sizes, our cohorts are geographically constrained, with most of our data coming from European and American subjects. As such, future estimates human microbiome gene content will likely be further improved by capturing even greater geographic heterogeneity.

These results make a comprehensive genetic understanding of the human microbiome, or even a compilation of its non-redundant gene catalog, seem very challenging. However, with greater focus on *de novo* assembled genes, we can avoid oversimplified analytical approaches, such as those based exclusively on taxonomy. Additionally, using extrapolated data, we see a clear need to increase sample sizes in metagenomic studies to the orders of tens of thousands if we are to adequately sequence the ''genome'' of the human microbiome. Incorporation of these large-scale gene level analyses into currently existing technologies can add a genetic context to the meaning of microbial species, allowing for more meaningful studies rooted in microbial genetics. We hope the scientific community will be able to use the set of resources provided here to deepen the field's understanding

Figure 6. Extrapolating the Gene Content of the Human Microbiome

(A and B) Extrapolation of the universe of genes using curves fit to our oral microbiome data (A) and gut microbiome data (B). Yellow dashed lines demarcate sampling required to observe certain percentages of new singletons per sample. Purple dashed line marks size of this study. Green dashed line is the asymptotic number of genes in the oral microbiome.

(C and D) Alternative, more conservative extrapolation methods for estimating total gene content in the oral/gut niches.

of the relationship between taxonomy and microbial genetic variation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- METHOD DETAILS
  - Overview of the approach
  - Synthetic data benchmarking of gene catalog pipeline
  - Relevant code: synthetic_data_benchmarking/download_homd_data.py
  - Relevant code, for example parameters: synthetic_data_benchmarking/art_parameters_example.sh
  - Relevant code: synthetic_data_benchmarking/run_synthetic_data_modeling.py.
  - Relevant code: synthetic_data_benchmarking/run_synthetic_data_modeling.py.
  - Identification of false positive genes
  - Relevant code: synthetic_data_benchmarking/run_synthetic_data_modeling.py.
  - Relevant code: synthetic_data_benchmarking/run_synthetic_data_modeling.py, synthetic_data_benchmarking/compute_gene_contig_coverage.sh.
  - Relevant code: statistical_analysis_and_figures/summary_data_analysis.R
  - Relevant code: statistical_analysis_and_figures/gene_by_gene_synthetic_analysis.R
  - Relevant code: synthetic_data_benchmarking/run_synthetic_cdhit_analysis.py
  - Relevant code: statistical_analysis_and_figures/gene_by_gene_synthetic_singleton_analysis.R
  - Relevant code: statistical_analysis_and_figures/gene_by_gene_singleton_analysis_real_data_oral.R
  - Relevant code: statistical_analysis_and_figures/gene_by_gene_metaspades_megahit_real.R
  - Relevant code: gene_catalog_construction/iterative_cdhit.sh, gene_catalog_construction/parse_iterative_cdhit.py
  - Relevant code: gene_catalog_construction/sorensen.cpp
  - Relevant code: statistical_analysis_and_figures/orfleton_figures_both.Rmd
  - Relevant code: gene_catalog_construction/full_contig_parsing_and_singleton_hunting_pipeline.py
  - Relevant code: contig_analysis/build_contig_database.py, contig_analysis/bin_contigs_species.py
  - Relevant code: contig_analysis/bin_contigs_species.py
  - Construction of disaggregated sample-based rarefaction curves
  - Creation of the gene discovery curve from the rarefaction curve
  - Determining a fitness function for the gene discovery curve
  - Determining the marginal sample s that yields a maximum number or percentage of new genes
  - Relevant code: extrapolation/RollingSpeciesEstimator.r, extrapolation/species_estimator.py
  - Relevant code: extrapolation/RollingSpeciesEstimator.r, extrapolation/species_estimator.py
  - Relevant code: extrapolation/RollingSpeciesEstimator.r, extrapolation/species_estimator.py
  - Relevant code: extrapolation/RollingSpeciesEstimator.r, extrapolation/species_estimator.py
  - Figure generation
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.chom.2019.07.008.

ACKNOWLEDGMENTS

We thank Debora Marks for her advice at the outset of our project, as well as Microsoft Azure and Amazon Web Services for providing compute resources for this work. This research was additionally supported by the National Institutes of Health NIEHS (T32 DK110919, R00ES23504, and R21ES205052), the National Science Foundation (1636870), the National Institute of Allergy and Infectious Diseases (R01AI127250), the American Diabetes Association (ADA) Pathway to Stop Diabetes Initiator Award 1-17-INI-13, and a Smith Family Foundation Award for Excellence in Biomedical Research.

AUTHOR CONTRIBUTIONS

B.T.T., A.D.K, C.J.P, and J.M.L conceived the project. B.T.T., C.B., and Z.Y. aggregated the initial studies. B.T.T., with the assistance of Z.Y., M.C.W, and J.M.L carried out the assemblies and gene calling and singleton analysis. M.B. and B.T.T. carried out the gene content extrapolation. E.D.M constructed the web interface and queryable database.

DECLARATION OF INTERESTS

The authors have no competing interests to declare.

Received: January 14, 2019
Revised: May 1, 2019
Accepted: June 19, 2019
Published: August 14, 2019

SUPPORTING CITATIONS

The following references appear in the Supplemental Information: Andrei et al., 2019; Ayling et al., 2018; Bredon et al., 2018; Cabanás et al., 2018; Carlos et al., 2018; Chen et al., 2018; Cheng, 2018; Delgado et al., 2019; Delsuc et al., 2018; Dong et al., 2017; Flota, n.d.; Georganas et al., 2018; Gerner et al., 2018; Graham et al., 2017; Graham et al., 2018; Han et al., 2018; Hannigan et al., 2018a, 2018b; Huang et al., 2018; Jackman et al., 2017; Kleiner et al., 2017; Kroeger et al., 2018; Kusy et al., 2018; Learman et al., 2019; Li et al., 2018; Martin et al., 2019; Maus et al., 2018; Mizzi et al., 2017; Nurk et al., 2017; O'Leary et al., 2016; Pärnänen et al., 2016; Patin et al., 2018; Pedron et al., 2019; Rebollar et al., 2018; Rengasamy, 2018; Rengasamy et al., 2017; Roux et al., 2017; Royalty and Steen, 2018; Schulz et al., 2018; Shiller et al., 2017; Souvorov et al., 2018; Steven and Kuske, 2018; Sutton et al., 2019; Titus Brown et al., 2019; Tschitschko et al., 2018; Tully et al., 2017; Tully et al., 2018; Tyagi et al., 2019; Vasconcellos et al., 2019; Verbruggen et al., 2017; Wang et al., 2018a, 2018b; Ward et al., 2017a, 2017b; Ward et al., 2018a, 2018b; Xing et al.,

2017; Younge et al., 2018; Zaikova et al., 2019; Zhou et al., 2018; Zhou et al., 2019; Zinke et al., 2019

REFERENCES

Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504.

Andrei, A.-Ş., Salcher, M.M., Mehrshad, M., Rychtecký, P., Znachor, P., and Ghai, R. (2019). Niche-directed evolution modulates genome architecture in freshwater Planctomycetes. ISME J. *13*, 1056–1071.

Ayling, M., Clark, M.D., and Leggett, R.M. (2018). "New Approaches for Assembly of Short-Read Metagenomic Data." e27332v1 (PeerJ Preprints). https://doi.org/10.7287/peerj.preprints.27332v1.

Bairoch, A. (2000). The ENZYME database in 2000. Nucleic Acids Res. *28*, 304–305.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. *29*, 1165–1188.

Bredon, M., Dittmer, J., Noël, C., Moumen, B., and Bouchon, D. (2018). Lignocellulose degradation at the holobiont level: teamwork in a keystone soil invertebrate. Microbiome *6*, 162.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Cabanás, G.-L., Carmen, D.R.-R., Legarda, G., Pizarro-Tobías, P., Valverde-Corredor, A., Triviño, J.C., Roca, A., and Mercado-Blanco, J. (2018). Bacillales members from the Olive Rhizosphere are effective biological control agents against the defoliating pathotype of Verticillium Dahliae. Collection FAO: Agriculture *8*, 90.

Carlos, C., Fan, H., and Currie, C.R. (2018). Substrate shift reveals roles for members of bacterial consortia in degradation of plant cell wall polymers. Front. Microbiol. *9*, 364.

Chen, Z., DeSalle, R., Schiffman, M., Herrero, R., Wood, C.E., Ruiz, J.C., Clifford, G.M., Chan, P.K.S., and Burk, R.D. (2018). Niche adaptation and viral transmission of human papillomaviruses from archaic hominins to modern humans. PLoS Pathog. *14*, e1007352.

Cheng, Z. (2018). "PPAD, *Porphyromonas Gingivalis* and the subgingival microbiome in periodontitis and autoantibody-positive individuals at risk of rheumatoid arthritis." Phd, University of Leeds. http://etheses.whiterose.ac.uk/22247/.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Daubin, V., and Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res. *14*, 1036–1042.

Delgado, B., Bach, A., Guasch, I., González, C., Elcoso, G., Pryce, J.E., and Gonzalez-Recio, O. (2019). Whole rumen metagenome sequencing allows classifying and predicting feed efficiency and intake levels in cattle. Sci. Rep. *9*, 11.

Delsuc, F., Kuch, M., Gibb, G.C., Hughes, J., Szpak, P., Southon, J., Enk, J., Duggan, A.T., and Poinar, H.N. (2018). Resolving the phylogenetic position of Darwin's extinct ground sloth (*Mylodon darwinii*) using mitogenomic and nuclear exon data. Proc. Biol. Sci. *285*, 20180214.

Dong, B., Yi, Y., Liang, L., and Shi, Q. (2017). High throughput identification of antimicrobial peptides from fish gastrointestinal microbiota. Toxins (Basel) *9*, E266.

Dusko Ehrlich, S.; The MetaHIT Consortium (2011). "MetaHIT: the european union project on metagenomics of the human intestinal tract". In Metagenomics of the Human Body (Springer), pp. 307–316.

Flota, J.J.M. n.d. "CONSULTING SERVICES REPORT." The-Alien-Project.com. https://www.the-alien-project.com/wp-content/uploads/2018/12/ABRAXAS-EN.pdf.

Forouzan, E., Shariati, P., Mousavi Maleki, M.S., Karkhane, A.A., and Yakhchali, B. (2018). Practical evaluation of 11 de novo assemblers in metagenome assembly. J. Microbiol. Methods *151*, 99–105.

Forster, S.C., Browne, H.P., Kumar, N., Hunt, M., Denise, H., Mitchell, A., Finn, R.D., and Lawley, T.D. (2016). HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. Nucleic Acids Res. *44* (D1), D604–D609.

Georganas, E., Egan, R., Hofmeyr, S., Goltsman, E., Arndt, B., Tritt, A., Buluc, A., Oliker, L., and Yelick, K. 2018. "Extreme Scale De Novo Metagenome Assembly." *arXiv [cs.DC]*. arXiv. http://arxiv.org/abs/1809.07014.

Gerner, S.M., Rattei, T., and Graf, A.B. (2018). Assessment of urban microbiome assemblies with the help of targeted *in silico* gold standards. Biol. Direct *13*, 22.

Graham, E., Heidelberg, J.F., and Tully, B. 2017. "Undocumented potential for primary productivity in a globally-distributed bacterial photoautotroph." bioRxiv. https://www.biorxiv.org/content/10.1101/140715v2.abstract.

Graham, E.B., Crump, A.R., Kennedy, D.W., Arntzen, E., Fansler, S., Purvine, S.O., Nicora, C.D., Nelson, W., Tfaily, M.M., and Stegen, J.C. (2018). Multi 'omics comparison reveals metabolome biochemistry, not microbiome composition or gene expression, corresponds to elevated biogeochemical function in the hyporheic zone. Sci. Total Environ. *642*, 742–753.

Han, M., Yang, P., Zhong, C., and Ning, K. (2018). The human gut virome in hypertension. Front. Microbiol. *9*, 3150.

Hannigan, G.D., Duhaime, M.B., Koutra, D., and Schloss, P.D. (2018a). Biogeography and environmental conditions shape bacteriophage-bacteria networks across the human microbiome. PLoS Comput. Biol. *14*, e1006099.

Hannigan, G.D., Duhaime, M.B., Ruffin, M.T., 4th, Koumpouras, C.C., and Schloss, P.D. (2018b). Diagnostic potential and interactive dynamics of the colorectal cancer virome. MBio *9*, e02248-18. https://doi.org/10.1128/mBio.02248-18.

Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics *28*, 593–594.

Huang, P., Zhang, Y., Xiao, K., Jiang, F., Wang, H., Tang, D., Liu, D., Liu, B., Liu, Y., He, X., et al. (2018). The chicken gut metagenome and the modulatory effects of plant-derived benzylisoquinoline alkaloids. Microbiome *6*, 211.

Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.

Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A., Jahesh, G., Khan, H., Coombe, L., Warren, R.L., and Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. *27*, 768–777.

Kleiner, M., Thorson, E., Sharp, C.E., Dong, X., Liu, D., Li, C., and Strous, M. (2017). Assessing species biomass contributions in microbial communities via metaproteomics. Nat. Commun. *8*, 1558.

Kroeger, M.E., Delmont, T.O., Eren, A.M., Meyer, K.M., Guo, J., Khan, K., Rodrigues, J.L.M., Bohannan, B.J.M., Tringe, S.G., Borges, C.D., et al. (2018). New biological insights into how deforestation in amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. Front. Microbiol. *9*, 1635.

Kusy, D., Motyka, M., Bocek, M., Vogler, A.P., and Bocak, L. (2018). Genome sequences identify three families of Coleoptera as morphologically derived click beetles (Elateridae). Sci. Rep. *8*, 17084.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., et al. (2015). Insights from 20 years of bacterial genome sequencing. Funct. Integr. Genomics *15*, 141–161.

Lapierre, P., and Gogarten, J.P. (2009). Estimating the size of the bacterial pan-genome. Trends Genet. *25*, 107–110.

Learman, D.R., Ahmad, Z., Brookshier, A., Henson, M.W., Hewitt, V., Lis, A., Morrison, C., Robinson, A., Todaro, E., Wologo, E., et al. (2019). Comparative genomics of 16 *Microbacterium* spp. that tolerate multiple heavy metals and antibiotics. PeerJ *6*, e6258.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., et al.; MetaHIT Consortium; MetaHIT

Consortium (2014). An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. *32*, 834–841.

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

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., and Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods *102*, 3–11.

Li, H.-Y., Wang, H., Wang, H.-T., Xin, P.-Y., Xu, X.-H., Ma, Y., Liu, W.-P., Teng, C.Y., Jiang, C.L., Lou, L.P., et al. (2018). The chemodiversity of paddy soil dissolved organic matter correlates with microbial community at continental scales. Microbiome *6*, 187.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A.B., Brady, A., Creasy, H.H., McCracken, C., Giglio, M.G., et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. Nature *550*, 61–66.

Luber, J.M., Tierney, B.T., Cofer, E.M., Patel, C.J., and Kostic, A.D. (2017). Aether: leveraging linear programming for optimal cloud computing in genomics. Bioinformatics (December). https://doi.org/10.1093/bioinformatics/btx787.

Martin, R.M., Moniruzzaman, M., Mucci, N.C., Willis, A., Woodhouse, J.N., Xian, Y., Xiao, C., Brussaard, C.P.D., and Wilhelm, S.W. (2019). Cylindrospermopsis raciborskii Virus and host: genomic characterization and ecological relevance. Environ. Microbiol. *21*, 1942–1956.

Maus, I., Rumming, M., Bergmann, I., Heeg, K., Pohl, M., Nettmann, E., Jaenicke, S., Blom, J., Pühler, A., Schlüter, A., et al. (2018). Characterization of *Bathyarchaeota* genomes assembled from metagenomes of biofilms residing in mesophilic and thermophilic biogas reactors. Biotechnol. Biofuels *11*, 167.

McInerney, J.O., McNally, A., and O'Connell, M.J. (2017). Why prokaryotes have pangenomes. Nat. Microbiol. *2*, 17040.

Mizzi, J.E., Lounsberry, Z.T., Brown, C.T., and Sacks, B.N. (2017). Draft genome of tule elk *Cervus canadensis nannodes*. F1000Res. *6*, 1691.

Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier, E., et al.; MetaHIT Consortium; MetaHIT Consortium (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. *32*, 822–828.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res. *27*, 824–834.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44* (D1), D733–D745.

Pärnänen, K., Karkman, A., Tamminen, M., Lyra, C., Hultman, J., Paulin, L., and Virta, M. (2016). Evaluating the mobility potential of antibiotic resistance genes in environmental resistomes without metagenomics. Sci. Rep. *6*, 35790.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649–662.e20.

Patin, N.V., Pratte, Z.A., Regensburger, M., Hall, E., Gilde, K., Dove, A.D.M., and Stewart, F.J. (2018). Microbiome Dynamics in a Large Artificial Seawater Aquarium. Appl. Environ. Microbiol. *84*, e00179-18, https://doi.org/10.1128/AEM.00179-18.

Pedron, R., Esposito, A., Bianconi, I., Pasolli, E., Tett, A., Asnicar, F., Cristofolini, M., Segata, N., and Jousson, O. (2019). Genomic and metagenomic insights into the microbial community of a thermal spring. Microbiome *7*, 8.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al.; MetaHIT Consortium (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60.

Rebollar, E.A., Gutiérrez-Preciado, A., Noecker, C., Eng, A., Hughey, M.C., Medina, D., Walke, J.B., Borenstein, E., Jensen, R.V., Belden, L.K., and Harris, R.N. (2018). The Skin Microbiome of the Neotropical Frog *Craugastor fitzingeri*: Inferring Potential Bacterial-Host-Pathogen Interactions From Metagenomic Data. Front. Microbiol. *9*, 466.

Rengasamy, V. (2018). Engineering High Performance Workflows for End-to-End Acceleration of Genomic Applications (The Pennsylvania State University).

Rengasamy, V., Medvedev, P., and Madduri, K. (2017). "Parallel and Memory-Efficient Preprocessing for Metagenome Assembly". In 2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), pp. 283–292. ieeexplore.ieee.org.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., and Sullivan, M.B. (2017). Benchmarking viromics: an *in silico* evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ *5*, e3817.

Royalty, T., and Steen, A.D. (2018). Simulation-Based Approaches to Characterize Metagenome Coverage as a Function of Sequencing Effort and Microbial Community Structure. bioRxiv. https://doi.org/10.1101/356840.

Sandoval-Motta, S., Aldana, M., Martínez-Romero, E., and Frank, A. (2017). The Human Microbiome and the Missing Heritability Problem. Front. Genet. *8*, 80.

Schulz, F., Alteio, L., Goudeau, D., Ryan, E.M., Yu, F.B., Malmstrom, R.R., Blanchard, J., and Woyke, T. (2018). Hidden diversity of soil giant viruses. Nat. Commun. *9*, 4881.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

Shiller, A.M., Chan, E.W., Joung, D.J., Redmond, M.C., and Kessler, J.D. (2017). Light rare earth element depletion during Deepwater Horizon blowout methanotrophy. Sci. Rep. *7*, 10389.

Sørensen, T. (1948). {A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and Its Application to Analyses of the Vegetation on Danish Commons}. Biol. Skr. *5*, 1–34.

Souvorov, A., Agarwala, R., and Lipman, D.J. (2018). SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. *19*, 153.

Steven, B., and Kuske, C.R. (2018). Resuscitation of intact and disturbed biological soil crusts in response to a wetting event characterized by metatranscriptomic sequencing. Frontiers in Microbiology https://www.osti.gov/servlets/purl/1479950.

Sutton, T.D.S., Clooney, A.G., Ryan, F.J., Ross, R.P., and Hill, C. (2019). Choice of assembly software has a critical impact on virome characterisation. Microbiome *7*, 12.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA *102*, 13950–13955.

Titus Brown, C., Moritz, D., O'Brien, M.P., Reidl, F., Reiter, T., and Sullivan, B.D. (2019). "Exploring Neighborhoods in Large Metagenome Assembly Graphs Reveals Hidden Sequence Diversity". bioRxiv. https://doi.org/10.1101/462788.

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods *12*, 902–903.

Tschitschko, B., Erdmann, S., DeMaere, M.Z., Roux, S., Panwar, P., Allen, M.A., Williams, T.J., Brazendale, S., Hancock, A.M., Eloe-Fadrosh, E.A., and

Cavicchioli, R. (2018). Genomic variation and biogeography of Antarctic haloarchaea. Microbiome *6*, 113.

Tully, B.J., Sachdeva, R., Graham, E.D., and Heidelberg, J.F. (2017). 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. PeerJ *5*, e3558.

Tully, B.J., Graham, E.D., and Heidelberg, J.F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci. Data *5*, 170203.

Tyagi, A., Singh, B., Billekallu Thammegowda, N.K., and Singh, N.K. (2019). Shotgun metagenomics offers novel insights into taxonomic compositions, metabolic pathways and antibiotic resistance genes in fish gut microbiome. Arch. Microbiol. *201*, 295–303.

Ugland, K.I., Gray, J.S., and Ellingsen, K.E. (2003). The Species-Accumulation Curve and Estimation of Species Richness. J. Anim. Ecol. *72*, 888–897.

Vasconcellos, A.F., Silva, J.M.F., de Oliveira, A.S., Prado, P.S., Nagata, T., and Resende, R.O. (2019). Genome sequences of chikungunya virus isolates circulating in midwestern Brazil. Arch. Virol. *164*, 1205–1208.

Verbruggen, H., Marcelino, V.R., Guiry, M.D., Cremen, M.C.M., and Jackson, C.J. (2017). Phylogenetic position of the coral symbiont Ostreobium (Ulvophyceae) inferred from chloroplast genome data. J. Phycol. *53*, 790–803.

Wainschtein, Pierrick, Jain, Deepti P., Yengo, Loic, Zheng, Zhili, TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, Adrienne Cupples, L., et al. (2019). Recovery of Trait Heritability from Whole Genome Sequence Data. bioRxiv. https://doi.org/10.1101/588020.

Wang, J., Tang, L., Zhou, H., Zhou, J., Glenn, T.C., Shen, C.-L., and Wang, J.-S. (2018a). Long-term treatment with green tea polyphenols modifies the gut microbiome of female sprague-dawley rats. J. Nutr. Biochem. *56*, 55–64.

Wang, X., Xiong, X., Cao, W., Zhang, C., Werren, J., and Wang, X. (2018b). "Genome Assembly of the A-Group Wolbachia in Nasonia Oneida and Phylogenomic Analysis of Wolbachia Strains Revealed Genome Evolution and Lateral Gene Transfer." *bioRxiv*. https://www.biorxiv.org/content/10.1101/508408v1.abstract.

Ward, L.M., McGlynn, S.E., and Fischer, W.W. (2017a). Draft Genome Sequence of *Chloracidobacterium* sp. CP2_5A, a Phototrophic Member of the Phylum *Acidobacteria* Recovered from a Japanese Hot Spring. Genome Announc. *5*, e00821-17. https://doi.org/10.1128/genomeA.00821-17.

Ward, L.M., McGlynn, S.E., and Fischer, W.W. (2017b). Draft Genome Sequences of a Novel Lineage of *Armatimonadetes* Recovered from Japanese Hot Springs. Genome Announc. *5*, e00820-17. https://doi.org/10.1128/genomeA.00820-17.

Ward, L.M., Hemp, J., Shih, P.M., McGlynn, S.E., and Fischer, W.W. (2018a). Evolution of Phototrophy in the Chloroflexi Phylum Driven by Horizontal Gene Transfer. Front. Microbiol. *9*, 260.

Ward, L.M., McGlynn, S.E., and Fischer, W.W. (2018b). Draft Genome Sequences of Two Basal Members of the *Anaerolineae* Class of *Chloroflexi* from a Sulfidic Hot Spring. Genome Announc. *6*, e00570-18. https://doi.org/10.1128/genomeA.00570-18.

Wolf, Y.I., Makarova, K.S., Lobkovsky, A.E., and Koonin, E.V. (2016). Two fundamentally different classes of microbial genes. Nat. Microbiol. *2*, 16208.

Xing, X., Liu, J.S., and Zhong, W. (2017). MetaGen: reference-free learning with multiple metagenomic samples. Genome Biol. *18*, 187.

Ye, Y., and Doak, T.G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Comput. Biol. *5*, e1000465.

Yin, Y., and Fischer, D. (2006). On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. BMC Evol. Biol. *6*, 63.

Young, J.P., Crossman, L.C., Johnston, A.W., Thomson, N.R., Ghazoui, Z.F., Hull, K.H., Wexler, M., Curson, A.R., Todd, J.D., Poole, P.S., et al. (2006). The genome of Rhizobium leguminosarum has recognizable core and accessory components. Genome Biol. *7*, R34.

Younge, N.E., Araújo-Pérez, F., Brandon, D., and Seed, P.C. (2018). Early-life skin microbiota in hospitalized preterm and full-term infants. Microbiome *6*, 98.

Zaikova, E., Goerlitz, D.S., Tighe, S.W., Wagner, N.Y., Bai, Y., Hall, B.L., Bevilacqua, J.G., Weng, M.M., Samuels-Fair, M.D., and Johnson, S.S. (2019). Antarctic Relic Microbial Mat Community Revealed by Metagenomics and Metatranscriptomics. Front. Ecol. Evol. *7*, 1.

Zeevi, D., Korem, T., Godneva, A., Bar, N., Kurilshikov, A., Lotan-Pompan, M., Weinberger, A., Fu, J., Wijmenga, C., Zhernakova, A., and Segal, E. (2019). Structural variation in the gut microbiome associates with host health. Nature *568*, 43–48.

Zhao, S., Lieberman, T.D., Poyet, M., Kauffman, K.M., Gibbons, S.M., Groussin, M., Xavier, R.J., and Alm, E.J. (2019). Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host Microbe *25*, 656–667.

Zhou, Y., Xu, Z.Z., He, Y., Yang, Y., Liu, L., Lin, Q., Nie, Y., Li, M., Zhi, F., Liu, S., et al. (2018). Gut Microbiota Offers Universal Biomarkers across Ethnicity in Inflammatory Bowel Disease Diagnosis and Infliximab Response Prediction. mSystems *3*, e00188-17. https://doi.org/10.1128/mSystems.00188-17.

Zhou, R., Zeng, S., Hou, D., Liu, J., Weng, S., He, J., and Huang, Z. (2019). Occurrence of human pathogenic bacteria carrying antibiotic resistance genes revealed by metagenomic approach: A case study from an aquatic environment. J. Environ. Sci. (China) *80*, 248–256.

Zhu, A., Sunagawa, S., Mende, D.R., and Bork, P. (2015). Inter-individual differences in the gene content of human gut bacterial species. Genome Biol. *16*, 82.

Zinke, L.A., Glombitza, C., Bird, J.T., Røy, H., Jørgensen, B.B., Lloyd, K.G., Amend, J.P., and Reese, B.K. (2019). Microbial Organic Matter Degradation Potential in Baltic Sea Sediments Is Influenced by Depositional Conditions and *In Situ* Geochemistry. Appl. Environ. Microbiol. *85*, e02164-18. https://doi.org/10.1128/AEM.02164-18.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
MEGAHIT (v1.1.2)	Li et al. 2015	https://github.com/voutcn/megahit
metaSPAdes (v3.13.0)	Nurk et al. 2017	https://github.com/ablab/spades
Prokka (v1.12)	Seemann 2014	https://github.com/tseemann/prokka
CD-HIT (v4.6.8)	Li and Godzik 2006	https://github.com/weizhongli/cdhit
Diamond (v0.9.24)	Buchfink et al. 2015	https://github.com/bbuchfink/diamond
Art (MountRainier)	Huang et al. 2012	https://www.niehs.nih.gov/research/resources/software/biostatistics/art/index.cfm
Data links and publicly available resource information	Table S1	
Database of results		https://microbial-genes.bio

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources, software, and data sharing should be directed to and will be fulfilled by the Lead Contact, Aleksandar Kostic (Aleksandar.Kostic@joslin.harvard.edu).

METHOD DETAILS

Overview of the approach

We aggregated 3,655 publically available oral and gut microbiome metagenomes used *de novo* assembly, Open-Reading-Frame (ORF) calling, and sequence-based clustering via CD-HIT (Li and Godzik, 2006) to identify a set of 45,666,334 non-redundant genes within them. We found 23,961,508 and 22,254,436 non-redundant genes in the oral and gut cavities, respectively. To enable access to our data by the broader research community, we built a public-facing interface, queryable PostgresQL database, and data repository hosted on Figshare.

To validate our gene-calling pipeline, we performed extensive analysis on synthetic and real data. Synthetic read data were generated with Art (Huang et al., 2012) from complete oral microbe isolate genomes downloaded from the Human Oral Microbiome Database. We assembled our synthetic metagenomes with metaSPAdes (Forouzan et al., 2018) and MEGAHIT running a variety of settings (Li et al., 2015). We called ORF's and checked the false discovery rate of each assembler as well as the correlation with genes of different prevalence (i.e., incidence in multiple samples or just one) with coverage and length of genes/contigs.

We quantified the distribution of genes across samples within our dataset, undertaking an in-depth analysis as to the number, frequency, taxonomic, and functional classifications of these genes. We quantified taxonomy by alignment to NCBI's NR database with Diamond that had been indexed with NCBI's taxon mapping files (available at ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/). Functional classifications were carried out as part of our ORF calling process, which leverages information from UniProt, Pfam, TIGRFAMs, and NCBI's RefSeq to classify genes *ab initio*. We were able to identify 2,418 discrete pathway ECiDs (out of 222,308 unique gene annotations) as well as confidently map to 15,746 microbial NCBI ID's. We have additionally made these results available as part of our resource, which can be searched by gene name, gene annotation, or taxonomy, if need be.

Further, we clustered our gene catalog at a variety of percent identities (down to 50% amino acid identity) to study the rate of clustering of different genes. We have made these available as downloadable links in our dataset as well.

We provide example scripts for each phase in our pipeline at https://github.com/kosticlab/universe_of_genes_scripts.

Synthetic data benchmarking of gene catalog pipeline

An outline of our methodological pipeline can be viewed in Figure S1.

We attempted to address if singletons are likely to be false positive genes. Due to the gene-level focus in our data we do not analyze misassemblies, as contigs are an intermediate stage in our analysis. We felt it more prudent to search for predictors of success in our primary endpoints: genes. Our confidence in this analytic decision was further increased by our literature-review-based analysis yielding that MEGAHIT and metaSPAdes have been shown in other publications to yield equivalent numbers of misassemblies when compared head-to-head (see "*Literature search for comparisons between MegaHit and other assemblers, including metaSPAdes*" in the Star Methods).

This in mind, we carried out an extensive analysis of the performance of different assemblers at different levels of coverage on synthetic metagenomic data in order to answer four questions:

1) Is there a best assembler in terms of false discovery rate and singleton discovery at varying coverages?
2) Are low coverage contigs/genes usually false positives?
3) Are short contigs/genes usually false positives?
4) Can we identify optimal quality filtering parameters such that we minimize false positive genes and maximize true positive genes– a minimum contig length/coverage or gene length/coverage?

Aggregation of complete microbial genomes

We downloaded all 467 complete, circularized genomes, as well as their corresponding Open-Reading-Frame predictions, from the Human Oral Microbiome Database (www.homd.org).

Relevant code: synthetic_data_benchmarking/download_homd_data.py

Construction of synthetic read data

We ran Art (Huang et al., 2012) to create synthetic read data at 1X coverage for each genome (parameters: art_illumina -ss HS25 -sam -i input_genome -p -l 150 -f 1 -m 200 -s 10 -o paired_dat, where input_genome is a fasta file containing the complete genome assembly). These are the recommended parameters in Art's README for generating synthetic Illumina sequencing data.

Relevant code, for example parameters: synthetic_data_benchmarking/art_parameters_example.sh

Construction of synthetic metagenomes

We found from our MetaPhlAn2 output, on average, there were 95 species per sample in our oral microbiome data. As such, we randomly picked 95 of the 467 genomes to be combined into a synthetic metagenome at varying levels of coverage. We randomly selected a value X (where $X > 0$) between a specific coverage range for each of the 95 metagenomes. If the value were greater than 1, we combined X copies of the 1X coverage synthetic read file for that genome into the metagenome. If it were less than 1, we subsampled that fraction of reads from each of the fastq files using seqtk (parameters: -s X)

Relevant code: synthetic_data_benchmarking/run_synthetic_data_modeling.py.

Assembly and gene calling parameters

We assembled with the following parameters:

1. MEGAHIT (parameter descriptions taken from http://www.metagenomics.wiki/tools/assembly/megahit):
meta '–min-count 2–k-list 21,41,61,81,99'
(generic metagenomes, default)
meta-sensitive '–min-count 2–k-list 21,31,41,51,61,71,81,91,99'
(more sensitive but slower)
meta-large '–min-count 2–k-list 27,37,47,57,67,77,87'
(large & complex metagenomes, like soil)
2. metaSPAdes (used default parameters):
metaSPAdes.py −1 synthetic_metagenome_1.fq.gz −2 synthetic_metagenome_2.fq.gz–only-assembler -o output
3. Prokka
prokka–outdir prokka_output–addgenes–metagenome–cpus 0–mincontiglen 1 assembly_output

Relevant code: synthetic_data_benchmarking/run_synthetic_data_modeling.py.

Coverage ranges

We ran our pipeline at three coverage ranges. Each of the 95 organisms in each metagenome was added to said metagenome at a level of coverage within a specific range.In the recent Pasolli et al. used a minimum coverage cutoff of 10X for genome extraction from metagenomic data, they – S s such, we chose to test coverage ranges centered around this value. We performed 10 iterations (i.e., generated 10 synthetic metagenomes) for each range. The ranges we chose were low coverage (0-1X), low-medium coverage (0-10X) and medium-high coverage (10-20X).

Identification of false positive genes

We identified false positive predicted genes by aligning our predicted genes back to the Open-Reading-Frames found in the 95 complete genomes we initially put into the synthetic metagenome. We aligned with Diamond (Buchfink et al., 2015) (additional parameters:–max-target-seqs 1–id 0.95). Genes in the predicted gene set that did not align to the "ground truth" genes were marked as false positives.

Relevant code: synthetic_data_benchmarking/run_synthetic_data_modeling.py.

Computing coverage for each gene/contig

We computed coverage of each gene/contig by aligning raw reads back to the predicted gene/contig output files, respectively, using BBMap (parameters: bbmap/bbwrap.sh ref = $reference in = $f1 in2 = $f2 out = output kfilter = 22 subfilter = 15 maxindel = 80). We based these parameters based on those recommended in the MEGAHIT wiki for computing contig coverage (https://github.com/voutcn/megahit/wiki/An-example-of-real-assembly). We computed average coverage per contig/gene, as well as average percent of contig/gene covered.

Relevant code: synthetic_data_benchmarking/run_synthetic_data_modeling.py, synthetic_data_benchmarking/compute_gene_contig_coverage.sh.

Summary-level analysis

We carried out linear regression and correlational analyses on summary- level data, which consisted of averaged statistics across all the genes/contigs . We computed average false discovery rate for a given iteration/coverage level, average contig/gene coverage (total percent covered as well as fold coverage), and average contig/gene length. Using base-R's glm function, we ran the regression *False discovery rate* $\sim$ *Assembler type*, where Assembler Type is a categorical variable consisting of the four different assembly parameters we used. Further, we used the stat_comp function from the ggpubr package to compute correlations false discovery rate and gene/contig fold coverage/length.

Relevant code: statistical_analysis_and_figures/summary_data_analysis.R

Gene-by-gene false positive analysis

We used base-R's glm function to run the following two logistic regressions, using contig-level and gene-level summary statistics, respectively:

1. $\textit{False positive gene} \sim \textit{Gene length}(\textit{per}1sd) + \textit{Geneavg fold coverage}\ (\textit{per}1sd) + \textit{Assembler type} + \textit{Genome coverage range}$
2. $\textit{False positive gene} \sim \textit{Contig length}\ (\textit{per}\ 1sd) + \textit{Contig avg fold coverage}\ (\textit{per}\ 1sd) + \textit{Assembler type} + \textit{Genome coverage range}$

We computed two different regressions to avoid including highly correlated variables (gene length/coverage and contig length/coverage) in the same model.

a. false_positive = a given gene is a false positive (1) or a true positive (0)
b. fold_coverage_contig = the fold coverage for the contig a particular gene arose from
c. fold_coverage_gene = the fold of coverage for a gen
d. length_contig = the length of a contig a particular gene arose from
e. length_gene = the length of a gene
f. assembler_type = which of the 4 assembly parameters were used (megahit large, megahit sensitive, megahit default, metaSPAdes)
g. coverage_range = which coverage range a given gene came from (0 to 1, 0 to 10, 10 to 20)

We plotted the distributions of gene/contig length/coverage by false positive/singleton status in Figure S2. Judging from the relationships displayed in these distributions, we hypothesized that shorter genes would have a higher probability of being a false positive.

Total number of false positive genes assembled by MEGAHIT default was 514,446 (15.3%). Gene length ranged from 61 to 2,448 bases with a mean of 267.2 and a median of 243.0. Contig length ranged from 64 to 5,905 bases with a mean of 480.4 and a median of 430.4. Gene average fold coverage ranged from 0 to 3722.385 reads with a mean of 23.458 and a median of 19.609. Contig average fold coverage ranged from 0 to 3646.935 reads with a mean of 23.185 and a median of 20.397. In order to aid in interpretability of our analysis, we normalized each of these variables by their standard deviations for each regression they were used in (gene length SD: 123.22 bases, contig length SD: 202.70 bases, gene average fold coverage SD: 21.16 reads, contig average fold coverage SD: 19.03 reads). By doing this, our odds ratios could be interpreted as change in odds for a gene being a false positive given a 1 standard deviation change in length/coverage.

Relevant code: statistical_analysis_and_figures/gene_by_gene_synthetic_analysis.R

Clustering and identification of singleton genes

We additionally clustered all the metagenomes within each assembler parameter/type group (so across all coverage ranges) into four separate non-redundant gene catalogs, so we could identify how singleton status of a given gene associated with coverage statistics and assembly method. To do so, we grouped all 10 iterations within a given coverage range and used CD-HIT to cluster the genes therein (parameters: cdhit/cd-hit -n 3 -i all_genes_for_cdhit -T 0 -M 0 -s 0.9 -aS 0.9 -c 0.5 -o cdhit_output_50perc) at the 50% identity level. We chose 50% identity to mimic the analysis that we had done in much of the paper.

Relevant code: synthetic_data_benchmarking/run_synthetic_cdhit_analysis.py

Gene-by-gene singleton analysis

We used base-R's glm function to run the following two logistic regressions, using contig-level and gene-level summary statistics, respectively:

1. *Singleton gene* ~ *False positive gene* + *Gene length* (*per* 1*sd*) + *Gene avg fold coverage* (*per* 1*sd*) + *Assembler type* + *Genome coverage range*
2. *Singleton gene* ~ *False positive gene* + *Contig length* (*per* 1*sd*) + *Contig avg fold coverage* (*per* 1*sd*) + *Assembler type* + *Genome coverage range*

The parameter definitions are the same as above with the addition of singleton_status, which refers to if, after clustering, a gene was a singleton (1) or a non-singleton (0). We computed area under the curve (AUC) estimates using using the roc function in the pROC package (Robin et al., 2011).

Relevant code: statistical_analysis_and_figures/gene_by_gene_synthetic_singleton_analysis.R

Benchmarking of gene catalog pipeline on real data

Modeling singleton gene status and oral microbiome contig coverage/length, gene length, and read counts

We used the following regressions to find associations between contig coverage/gene length/contig length/depth of sequencing and singleton status in our oral microbiome data. We computed AUC estimates using using the roc function in the pROC package (Robin et al., 2011).

a. *Singleton gene* ~ *Gene length* (*per* 1*sd*) + *Total reads* (*per* 1*sd*)
b. *Singleton gene* ~ *Contig length* (*per* 1*sd*) + *Contig avg fold coverage* (*per* 1*sd*) + *Total reads* (*per* 1*sd*)

One drawback of the synthetic data analysis is that due to small sample size compared to our actual study, the singleton gene fraction was higher than we would have expected (i.e., some non-singletons may have been classified as singletons). As such, we modeled our real data as well. In this case, we lack information on true/false positive genes, but we have larger sample sizes and a lower overall fraction of singleton genes.

We used bbMap once again to compute contig-by-contig coverage for each predicted element that was identified by PROKKA. For this analysis, we opted initially not to compute gene-by-gene coverage (or contig-by-contig coverage for the gut microbiome) due to the 1) additional time and monetary cost that would be required and 2) the similarity between the gene/contig results in the synthetic data analysis.

Given the similar distribution of singleton/non-singleton genes in association with our independent variables of interest Figure S3), we hypothesized our regressions would yield small effect sizes and minimal changes in the probability of a gene being a singleton compared to baseline. We found modest in effect size but statistically significant associations between singleton genes and coverage of the contig from which a gene came, gene length, and contig length (Table S2). The total number of singletons was 3,183,181 (2.0%). Contig length ranged from 200 to 1,041,740 base pairs with a mean of 1,343 and a median of 2,390. Contig coverage, in terms of average number of reads aligning to each base of a contig, ranged from 0 to 98,048.39 reads with a mean of 16.58 and a median of 5,910. Gene length ranged from 66 to 31,656 with a mean of 671 and a median of 513. As with our synthetic data analysis, in our regressions we normalized each continuous variable by its standard deviation (Gene length SD: 547.15 bases, Contig length SD: 40,924.06 bases, Contig average fold coverage SD: 59.10 reads, Total reads SD: 83,030,393 reads).

Relevant code: statistical_analysis_and_figures/gene_by_gene_singleton_analysis_real_data_oral.R

Comparison between MEGAHIT and metaSPAdes in identification of singleton genes

We ran our assembly, gene calling, and gene catalog construction pipeline on a subset of 10 randomly selected samples, computing gene-by-gene and contig-by-contig coverage (as above) and identifying singletons. We computed AUC estimates using using the roc function in the pROC package. (Robin et al., 2011) We ran the following regressions for our analysis:

a. *Singleton gene* ~*Assembler type*

Relevant code: statistical_analysis_and_figures/gene_by_gene_metaspades_megahit_real.R

Meta-analytic data collection

We identified 13 publications (Table S1) with shotgun sequencing metagenomic data taken from any human oral and gut microbiomes. We used 2,182 gut samples and 1,473 oral samples. We downloaded relevant study data from either the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), the metagenomics RAST server (MG-RAST), or the Human Oral Microbiome Database (HOMD).

Raw read filtering and quality control

If reads had not been trimmed or had human sequences filtered out in their respective studies, we used KneadData(https://bitbucket.org/biobakery/humann2/wiki/Home) to do so prior to assembly. This pipeline involves two primary steps. 1) Aligning raw reads back

to the human genome reference (GRCh37/hg19) to filter out human contaminants (settings:–very-sensitive). 2) Using Trimmomatic to remove adaptor contamination (settings: SLIDINGWINDOW 4:20, MINLEN 50).

Open-reading-frame prediction and initial functional annotation

We ran Prokka (Seemann, 2014) (version: 1.12, settings:–metagenome–addgenes–mincontiglen 1) on the raw contigs from our *de novo* assembly to predict genes.

Assembly, gene calling, and construction of non-redundant gene catalog

We assembled raw reads into contiguous sequences (or, contigs) using MEGAHIT (Li et al., 2016) V1.1.2 (parameters:–default–mem-flag 2). We removed contigs under 200 base pairs in length. We used Prokka (Seemann, 2014) V1.12 to annotate genes from the MEGAHIT output (settings:–cpus 0–addgenes–metagenome–mincontiglen 1). We used the default databases installed with Prokka (UniProt, Pfam, TIGRFAMs, and NCBI's RefSeq) for functional annotation. We then ran CD-HIT-EST (Li and Godzik, 2006) V4.6.8 with a 95% identity cutoff (–n 10–c 0.95 -aS 0.9 -S 0.9 -M 0 -T 0). We removed genes under 100 bases in length that did not align to any sequence NR reference database at 95% identity. For any other gene catalogs we made we either used CD-HIT or CD-HIT-EST with varying percent identity and word length (according to the instructions in the CD-HIT user's manual https://github.com/weizhongli/cdhit/blob/master/doc/cdhit-user-guide.pdf)

Iterative gene catalog construction

We translated our nucleic acid gene catalog into amino acids with Python's Biopython (Cock et al., 2009) package. We ran CD-HIT V4.6.8 with the same parameters as above on the translated gene catalog with progressively lower percent identities, starting at 100% and decreasing in increments of 5 down to 50%. For example, we fed the output of CD-HIT run at 100% identity into another CD-HIT run with the -c flag changed to 0.95, the output of which was run through CD-HIT again at -c 0.9, and so on.

Relevant code: gene_catalog_construction/iterative_cdhit.sh, gene_catalog_construction/parse_iterative_cdhit.py

Reference-based species identification

We ran MetaPhlAn2(Truong et al., 2015) V2.1.0 with the default settings to identify the species content in each sample. To create incidence data from the MetaPhlAn output, which we used to in our cross-sample dissimilarity calculations, we collapsed the raw output into a relative abundance matrix, where the columns were samples and the rows were species. We then created an incidence matrix by recoding non-zero cells as having values of 1.

Calculation of cross-sample dissimilarity

Similarity metrics were calculated using Sorenson-Dice (Sørensen, 1948) similarity, which is simply Bray-Curtis Dissimilarity applied to prevalence rather than abundance data. To speed up data processing, we used a custom, parallelized, c++ implementation.

Relevant code: gene_catalog_construction/sorensen.cpp

MinPath annotation

We ran MinPath(Ye and Doak, 2009) V1.4 (command: python ../MinPath1.2.py -any ecid_mapping -map ec2path -report ec.report -details ec.details) on the set of all EcID's captured in each gene catalog to identify a mapping between gene, EcID, and parsimonious pathway annotation.

Functional enrichment analysis

We identified pathways enriched in singletons or non-singletons for the gut and oral microbiomes using a Fisher's Exact test, where we compared the ratio of counts of singletons and non-singletons of any given pathway to the overall ratio of singletons to non-singletons across all populations. We adjusted for False-Discovery Rate using Benjamini-Yekutieli(Benjamini and Yekutieli, 2001) correction. For the plots in Figure 4, we reported the top 50 most enriched pathways in the gut microbiome and oral microbiome for singletons and non-singletons. For the plot in Figures S13–S14, we show the top 25 most enriched species/genera using the same methods.

Relevant code: statistical_analysis_and_figures/orfleton_figures_both.Rmd

Gene-level taxonomic annotation

We used Diamond's(Buchfink et al., 2015) taxonomic annotation configuration (which uses NCBI's taxon nodes and taxonmap files in conjunction with the Lowest Common Ancestor algorithm) to align against NCBI's RefSeq non-redundant protein database, which we downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz. After combining the separate files, we configured the diamond database with the command "diamond makedb–in nr.gz–db nr–taxonnodes nodes.dmp–taxonmap prot.accession2taxid.gz " We used Diamond's default cutoff, a minimum e-value of 0.0001, to identify confident hits.

Relevant code: gene_catalog_construction/full_contig_parsing_and_singleton_hunting_pipeline.py

Gene-based taxon contig binning

We mapped genes onto the contigs from which they originated, and we binned contigs into a particular taxonomic group if at least 75% of the genes on a given contig had the same taxonomic annotation. To increase our confidence in our annotations, we filtered out contigs with fewer than two genes on them, as well as those that were binned at a taxonomic level above genus (kingdom, phylum, class, order, family levels).

Relevant code: contig_analysis/build_contig_database.py, contig_analysis/bin_contigs_species.py

Identification of horizontal gene transfer

We tested of horizontal gene transfer was noticeably giving rise to singletons by examining in mixture-contigs, those consisting of both non-singletons and singletons. We searched for where the discordant species or genus taxonomic annotations of singletons(s) and non-singleton(s), excluding those annotations that were unable to be ascribed to any specific microbe. Out of concern of being biased by low resolution annotations, we only classified HGT as possibly having occurred when the taxonomic bins of the singleton(s) and the non-singleton(s) were of different genera.

Relevant code: contig_analysis/bin_contigs_species.py

Preparing data for functional and taxonomic enrichment scatterplots

In order to display enrichment within singletons and non-singletons (Figure S1B and S1F) while accounting for sample size, prior to plotting we normalized the counts of number of genes/contigs in particular pathway/taxa by dividing by the total number of singletons or non-singletons for a given niche. For example, if 10 singletons and 10 non-singletons aligned to a pathway and a total of 100 singletons and 10000 non-singletons were found to align to any pathway at all, we divided 10/100 and 10/10000 to get normalized values of 0.1 and 0.001, respectively, for said pathway.

Construction of disaggregated sample-based rarefaction curves

We create an R-by-S disaggregated sample-based rarefaction matrix D (where d_{rs} is the expected number of r-tons when s samples are drawn from a set of S samples), starting from a binary G-by-S incidence matrix W (where $w_{gs} = 1$ if the gene g was found in sample s). The rarefaction curve can be solved analytically using hypergeometric distributions, and depends only on the frequency in which each gene is found (Ugland et al., 2003).

First, we calculate the frequency y_g that gene g appears in all the samples, as the row sum of the incidence matrix W (1), then calculate the incidence frequency count q_k (2) where q_k is the number of times that genes appear k times in the sample S, and $I() = 1$ if its argument is true.

$$y_g = \sum_{s=1}^{S} w_{gs} \tag{1}$$

$$q_k = \sum_{g=1}^{G} I\left(y_g = = k\right) \tag{2}$$

Second, let $\mathcal{R}$ denote the maximum value of k where $q_k \neq 0$. Then, the expected number of r-tons accumulated after s random samples collected without replacement $d_{r,s}$ is calculated as follows:

$$d_{r,s} = \sum_{k=1}^{R} q_k h(s, S, r, k) \tag{3}$$

where the hypergeometric function (4), $h(s, S, r, k)$ returns the probability of drawing exactly r out of k possible units when sampling s times without replacement out of a set S. For example, suppose we have a collection of 20 samples. The probability of finding exactly 3 incidences in a 10-ton set if we choose at random 12 of the 20 samples, is $h(12, 20, 3, 10)$ and is calculated using binomial coefficients as follows:

$$h(s, S, r, k) = \frac{\binom{k}{r}\binom{S-k}{s-r}}{\binom{S}{s}} \tag{4}$$

Finally, the r-ton sample-based rarefaction curves are plotted from the r^{th} line of D using the Python matplotlib (Hunter, 2007) package. The aggregated sample-based rarefaction curve is the expected number of unique genes d_s^{agg} when s samples are collected without replacement:

$$d_s^{agg} = \sum_{r=1}^{R} d_{rs} \tag{5}$$

Creation of the gene discovery curve from the rarefaction curve

The gene discovery curve Q is the derivative of the rarefaction curve, where q_s is the number of new genes discovered on sample s:

$$q_s = d_{r,s}^{agg} - d_{r-1,s}^{agg} \tag{6}$$

Determining a fitness function for the gene discovery curve

The gene discovery curve q_s from (6) is used to extrapolate gene richness in the microbiome pangenome by polynomially fitting q_s.

We used the function curve_fit from the Python package scipy.optimize to fit the discovery curve to a function. Because the discovery curve is derived from a rarefaction curve, the chosen function must, on the positive x axis, be continuous, non-increasing, and convex. Further, as there are combinatorial limits on genes, the function must asymptotically reach 0. Finally, it must fit q_s with high-fidelity, especially at the right tail, in order to get the best estimator. As none of the usual eligible candidate functions (e.g., negative exponential, negative power curves) adequately fit the discovery curve at the right tail, it was necessary to select our own.

As the q_s appeared curved in logarithmic space, we applied a 2nd degree polynomial regression on logarithm-transformed data (i.e., fitting a curve after remapping the axes to and $x = loglog\ s$:

$$f(x,a,b,c) = ax^2 + bx + c \tag{7}$$

where *a, b, and c* are the regression parameters. Note that if $a = 0$, the fitting function becomes a power curve.

Determining the marginal sample s that yields a maximum number or percentage of new genes

We determined the marginal sample *s* required to contain less than a fraction of new genes z^{frac}. As *f(s, a, b, c)* is not trivial to invert in logarithmic space, we use the root function from scipy.optimize in Python to solve for *s*:

$$0 = e^{f(\log(s),a,b,c)} / q_1 - z^{frac} \tag{8}$$

Similarly, the same root finding algorithm is used to find the marginal sample *s* that yields less than z^{num} new genes, by solving:

$$0 = e^{f(log\ (s)\ ,a,b,c\ )} - z^{num} \tag{9}$$

Relevant code: extrapolation/RollingSpeciesEstimator.r, extrapolation/species_estimator.py

Gene richness estimation of oral/gut microbiome pangenomes

As the area under the gene discovery curve function is finite for a < 0 and $s > 0$, we integrate $\int_{0^+}^{s} e^{f(log\ (s),\ a,b,c\ )}$ to extrapolate the rarefaction curve for an arbitrary value of *s*, by using the scipy.integrate function from Python. The richness of oral and gut genes asymptotically reaches 91,439,476 and 238,585,237 genes, respectively, when *s* = infinity.

Relevant code: extrapolation/RollingSpeciesEstimator.r, extrapolation/species_estimator.py

Estimating the number of singletons in the extrapolated rarefaction curve

A function that extrapolates the number of singletons as a function of samples collected must meet certain properties: the function must be continuous for $s > 0$, represent a non-increasing fraction of the rarefaction curve, ideally be the same value as the rarefaction curve at $s = 1$, asymptotically reach zero, and fit $d_{1,s}$ with high-fidelity, especially at the right tail, in order to get the best estimator. As none of the functions that we attempted fit the above criteria, we decided to regress the fraction of singletons in logarithmic space to the fitting function (7) while setting $a = 0$, then multiply it with the extrapolated rarefaction curve.

Relevant code: extrapolation/RollingSpeciesEstimator.r, extrapolation/species_estimator.py

Gene richness estimation of oral microbiome pangenome via the Chao2 and Chao-Bunge estimator

From the disaggregated rarefaction curve *D*, we estimate the gene richness of the oral microbiome by using estimators available in the SPECIES R package. Rolling estimates using Chao2, Chao-Bunge, Jackknife, and Chao-Lee were produced as samples were collected.

Relevant code: extrapolation/RollingSpeciesEstimator.r, extrapolation/species_estimator.py

Cloud Computing

All analyses were carried out entirely in the cloud on a combination of Amazon Web Services (AWS) and Microsoft Azure resources. We ran our initial assemblies on AWS spot instances using *Aether(Luber and Tierney et al., 2017)* and stored the resulting data on Azure's cloud storage. We used Azure, Linux-based virtual machines running Ubuntu 16.04 for the remainder of our analyses.

Figure generation

All plotting, except for that done for the rarefaction curves, was done in R using the packages "ggplot2" and "cowplot" (https://cran.r-project.org/web/packages/cowplot/index.html). Rarefaction analysis and extrapolation was done using Python's "Matplotlib" (Hunter, 2007) package. Figures were assembled in Adobe Illustrator (https://www.adobe.com/products/illustrator.html).

QUANTIFICATION AND STATISTICAL ANALYSIS

We used Fisher's exact tests, linear, and logistic regression to quantify associations between various covariates across the manuscript. We controlled for multiple hypothesis testing with Benjamini-Yekutieli p value adjustment.

DATA AND CODE AVAILABILITY

Example scripts for each step of this analytical pipeline are publicly accessible at https://github.com/kosticlab/universe_of_genes_scripts. When relevant, each section of the methods section below refers to script in this repository used for that particular analysis. The post-assembly pipeline, which includes non-redundant gene catalog construction, gene catalog quality control, gene-level taxonomy mapping, iterative gene catalog construction, binary gene incidence matrix generation, and Sorenson-Dice dissimilarity calculation, is run by "**gene_catalog_construction/full_contig_parsing_and_singleton_hunting_pipeline.py**." We built our public facing database using Microsoft Azure's Database for PostgreSQL service. We built our website with a Flask API.

ADDITIONAL RESOURCES

We additionally present a database of our results at https://microbial-genes.bio.