HumGut: A comprehensive Human Gut prokaryotic genomes collection filtered by metagenome data

Pranvera Hiseni*[1,3], Knut Rudi[1,2], Robert C. Wilson[2], Finn Terje Hegge[3], Lars Snipen[1]

[1]Department of Chemistry, Biotechnology & Food Sciences, Norwegian University of Life Sciences, P.O. Box 5003, 1432 Aas, Norway

[2]Inland Norway University of Applied Sciences, Hamar, Norway

[3]Genetic Analysis AS, Oslo, Norway

Abstract

A major challenge with human gut microbiome studies is the lack of a publicly accessible human gut genome collection that is verifiably complete. We aimed to create Humgut, a comprehensive collection of healthy human gut prokaryotic genomes, to be used as a reference for worldwide human gut microbiome studies. We screened >2,300 healthy human gut metagenomes for the containment of >486,000 publicly available prokaryotic genomes. The contained genomes were then scored, ranked, and clustered based on their sequence identity, only to keep representative genomes per cluster, resulting thus in the creation of HumGut. Superior performance in the taxonomic assignment of metagenomic reads, classifying 97% of reads on average, is a benchmark advantage of HumGut. Re-analyses of healthy gut samples using HumGut revealed that >90% contained a core set of 129 bacterial species and that, on average, the guts of healthy people contain around 1,000 bacterial species. The HumGut collection will continuously be updated as the list of publicly available genomes and metagenomes expand. Our approach can also be extended to disease-associated genomes and metagenomes, in addition to other species. The comprehensive, yet slim HumGut database streamlines analyses while significantly improving taxonomic assignments in a field in dire need of method standardization and effectivity.

Introduction

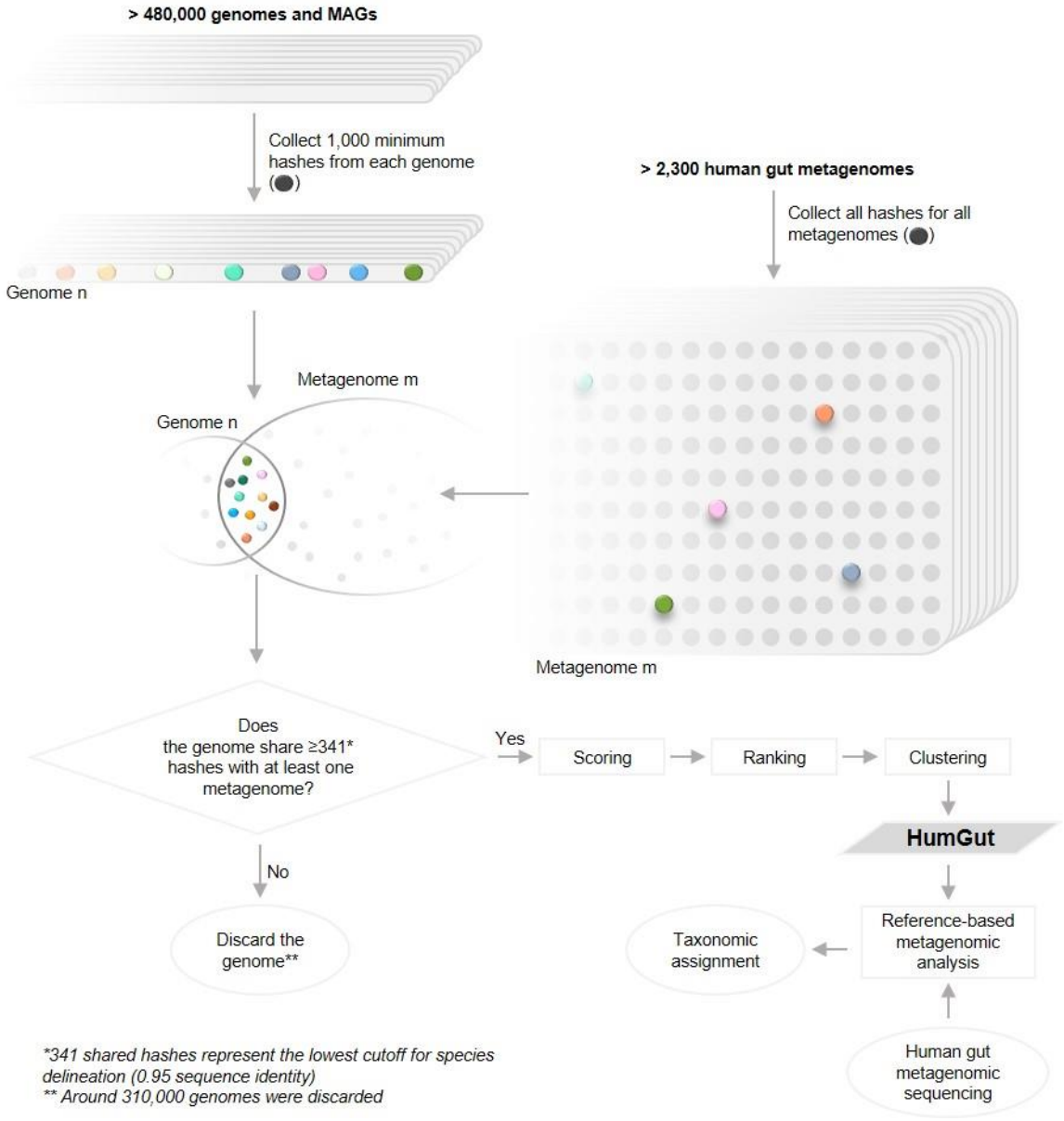
Major efforts have been undertaken to characterize the human gut microbiome, both by microbial isolation and sequencing [1]. Also, a significant contribution was made by *de novo*-assembled genomes (Metagenome-Assembled Genomes – MAGs), facilitated by recent advances in bioinformatics [2-6]. No studies, however, have addressed the actual containment of the available genomes and MAGs within a comprehensive set of representative human gut metagenomes, neither has the redundancy across the genomes/MAGs been evaluated. This knowledge is essential for establishing a complete collection of human gut-associated bacteria.

The comprehensive data set of microbial genetic information collected from the human gut is too large, rendering it inaccessible to most labs. The number of human gut metagenome BioProjects deposited in the Sequence Read Archive (SRA) database has grown enormously over the past few years. As of 2020, NCBI holds data from more than 1,400 individual such projects conducted worldwide, consisting of nearly 230,000 samples, comprised of more than 150 Tbases of sequence. Furthermore, the number of prokaryotic genomes deposited in GenBank has exceeded 550,000, marking an increase of more than 3-fold in 2019 alone. Therefore, there is a clear need to systemize the gut microbiota data on a global scale.

Regionally, gut microbiome studies have shown that gut microbiota can be linked to a range of diseases and disorders [7-10], and we are now at a stage where gut microbiota therapeutic interventions are being introduced [11,12]. However, the lack of a global reference for the gut microbiota in healthy humans represents a bottleneck. This limits both the understanding of gut microbiota on a worldwide scale and the introduction of large-scale intervention strategies.

We aimed to create a single, comprehensive genome collection of gut microbes associated with healthy humans, the HumGut, as a reference collection for all human gut microbiota studies globally. The

47 HumGut strategy is outlined in **Figure 1**. We show that using HumGut as a reference database makes
48 vast improvements to read assignment in human gut metagenomes by kraken2 [13]. Our results suggest
49 that HumGut, despite its relatively small size, is an outstanding representation of microbial genomes
50 present in the guts of healthy humans. The application of HumGut also reveals the list of species that we
51 consider to be the most prevalent and abundant bacteria in healthy human intestines globally.



52

Figure 1. HumGut overview. HumGut represents a collection of genomes and MAGs contained in 2311 healthy human gut metagenomes. To be considered as contained, a genome shared at least 341 hashes with one of the metagenomes. The qualified genomes were scored based on the sum of shared hashes >340 across all the metagenomes. Next, they were ranked based on their scores: the higher the score, the higher the position on the list. Subsequently, the genomes were clustered based on MASH distance (D). The top-ranked genome formed a cluster centroid. Various clusters were formed applying different D thresholds (0.00 – 0.05). The use of HumGut as a reference set helps the process of taxonomic assignments by drastically reducing the number of unclassified human gut metagenomic reads.

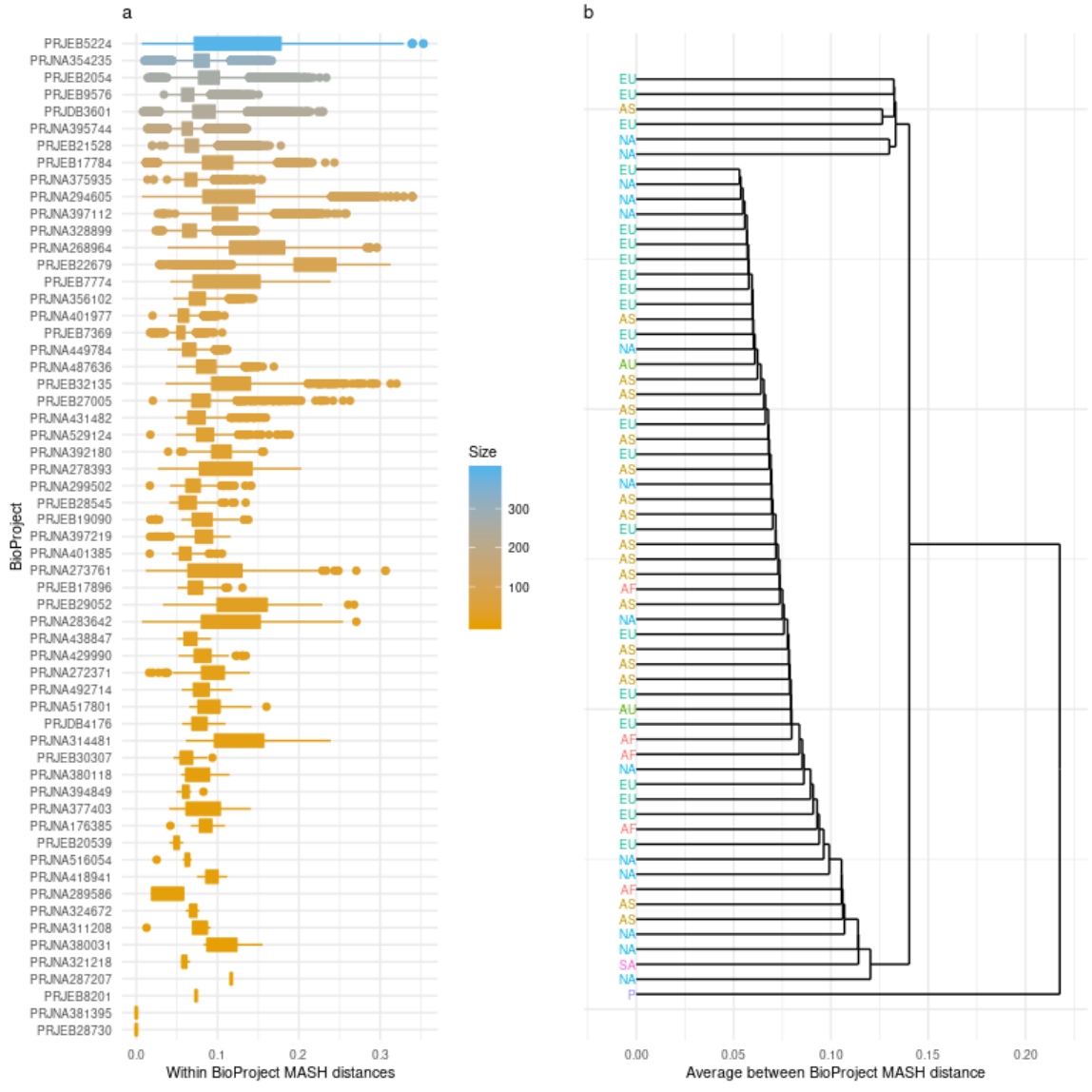
Results

Reference metagenomes

We downloaded >3,000 gut metagenome samples collected from healthy people worldwide. These belonged to 58 different BioProjects. We calculated MASH distances between samples within each BioProject to assess the diversity between them. The results showed that, on average, samples shared a 91% sequence identity (D = 0.09), indicating a high degree of similarity between one another. The sequence identity for the two most distant samples was 65% (D = 0.35) (**Figure 2a**).

We wanted to see if samples clustered based on their continent of origin (**Figure 2b)**. To do so, we computed the average linkage hierarchical clustering of BioProjects. The distance between two BioProjects is the mean pairwise distance between all their samples. Here, we also included a BioProject containing primate gut metagenome samples (n = 95), as an outgroup against which all human BioProjects were compared. The lowest observed average MASH distance (D = 0.06) was between two projects stemming from separate continents, one from Europe and the other from North America, while two most distant projects were both of European origin (D = 0.14). These observations, together with the mixed distribution of BioProjects in the cluster dendrogram, suggested that the clustering of samples did not heavily depend on continent-of-origin. The primate samples were markedly separated from the rest of the tree, showing an average distance of 0.22 from all other BioProjects.

77 After clustering at 0.05 MASH distance, a parameter value intended to keep only one sample in cases
78 where more were highly similar, we ended up with 2,311 metagenome samples covering all 58
79 BioProjects.



80

81 *Figure 2. Sample MASH distances within and between BioProjects.* ***a.*** *Boxplots illustrating the distribution of MASH distances*
82 *between samples within each BioProject. The BioProject accession is used as a label, and the color gradient indicates the size,*
83 *i.e., the number of samples in each.* ***b****. Average linkage hierarchical clustering of 58 BioProjects. Labels indicate the continent of*
84 *origin: EU – Europe, AS – Asia, NA – North America, AS – Australia, AF – Africa, SA – South America, and P stands for Primates.*

Except for the single primate BioProject (BioSample), each BioProject is listed in colored font according to the continent from which it originates. No severe clustering of samples based on origin is detected.

From genomes to HumGut collection

From 489,710 genomes in total, 163,693 qualified for inclusion in HumGut. The qualified genomes were at least 95% contained within at least one reference metagenome (inferred by >340 shared hashes). The most prevalent genomes, i.e., the genomes contained in most metagenomes, belonged to genus *Bacteroides*, led by *B. vulgatus*.

We checked the fraction of the recently published cultivated human gut bacteria genomes and MAGs that contributed to HumGut (**Figure 3)**. Some genomes exhibited a high score (horizontal axis), but the vast majority of them achieved rather low scores. This was especially evident for the MAGs in the SGB and IGG collections. We also checked the genomes of non-human-gut-bacteria, which, as expected, resulted in low scores in the MASH screen.

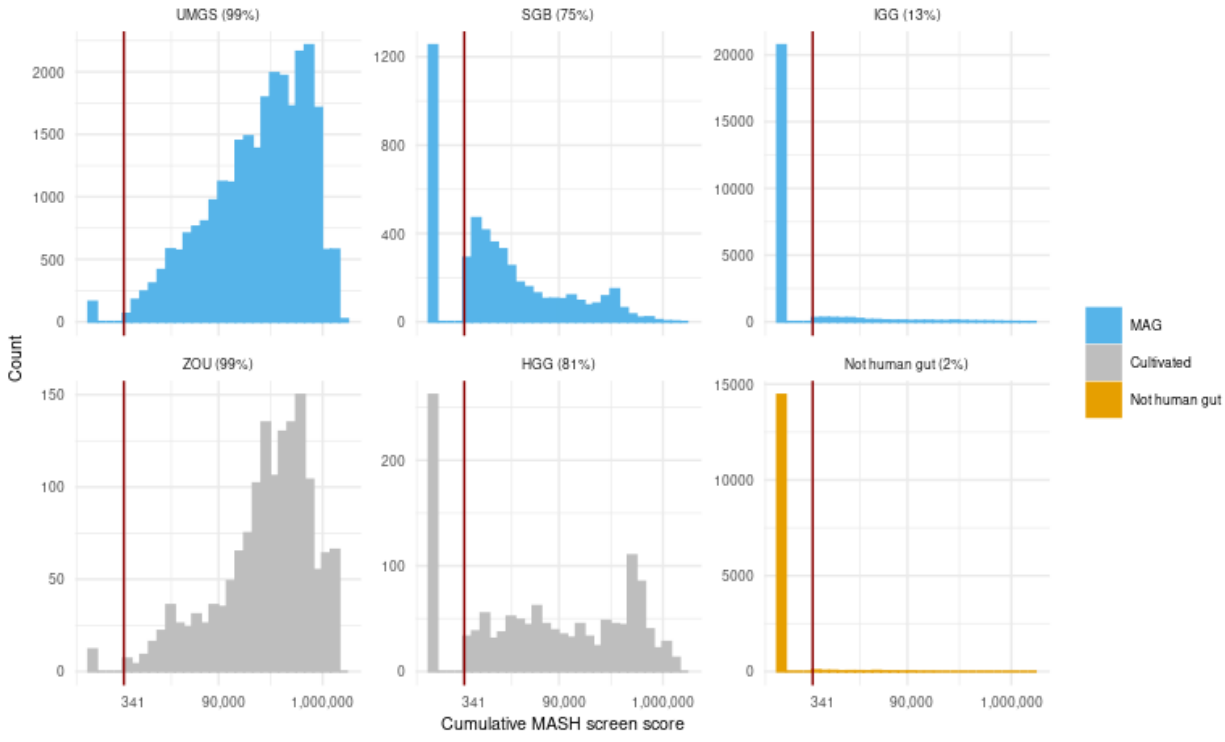


Figure 3. Cumulative MASH screen score histogram for newly published human gut bacterial sequences (shown in blue and grey), in addition to genomes described to not reside in the gut environment (shown in gold). The percentages in the titles represent the fraction of genomes scoring a k-mer threshold of at least 341 in at least one metagenome. The X-axis, which was square-root-transformed, shows the summed number of >340 shared k-mers. Y-axis indicates the number of genomes per each set. The red line on each panel represents the threshold that separates the genomes that did not get qualified for HumGut (on the left) from the ones qualified (on the right).

The contribution of qualified genomes to HumGut is presented in the supplementary material (**Figure S1**).

We performed clustering of genomes based on sequence similarity (MASH distance), using the top-ranked genome as a cluster centroid. By applying various MASH distance (*D*) thresholds, we created different subsets of HumGut collections (**Table 1**). Only cluster centroids were used to build the collections.

Table 1. The number of genome clusters at different levels of MASH distance thresholds

MASH distance threshold (D)	*Number of genome clusters*	*Number of unique Taxonomy IDs*
0.00	163,693	16,016
0.01	35,485	4,299
0.02	18,085	2,562
0.03	9,662	1,790
0.04	6,382	1,404
0.05	4,779	1,201

Classifying the metagenome reads

We used our six HumGut collections, in addition to the standard kraken2 database, to classify the metagenomic reads from the 2,311 downloaded samples. On average, there were 50.1% unclassified reads when using the standard kraken2 database, while the average dropped substantially when any

one of the HumGut collections was used (**Figure 4a**). On average, only 3.23 % of the reads remained unclassified when HumGut_00 was utilized, marking a significant increase in recognized reads, with an obvious potential for improved classification accuracy. In addition, HumGut k-mer database sizes were smaller than the standard kraken2 database of k-mers, reflecting a lower computer memory needed to perform the analysis (Standard = 39 GB, HumGut_05 = 19 GB).

Analysis of additional 100 gut metagenome samples, not part of the reference set, showed similar results regarding the number of recognized reads: 39.5% unclassified reads on average when Standard database was used, 2.1% with HumGut_00 usage (**Figure 4b**).

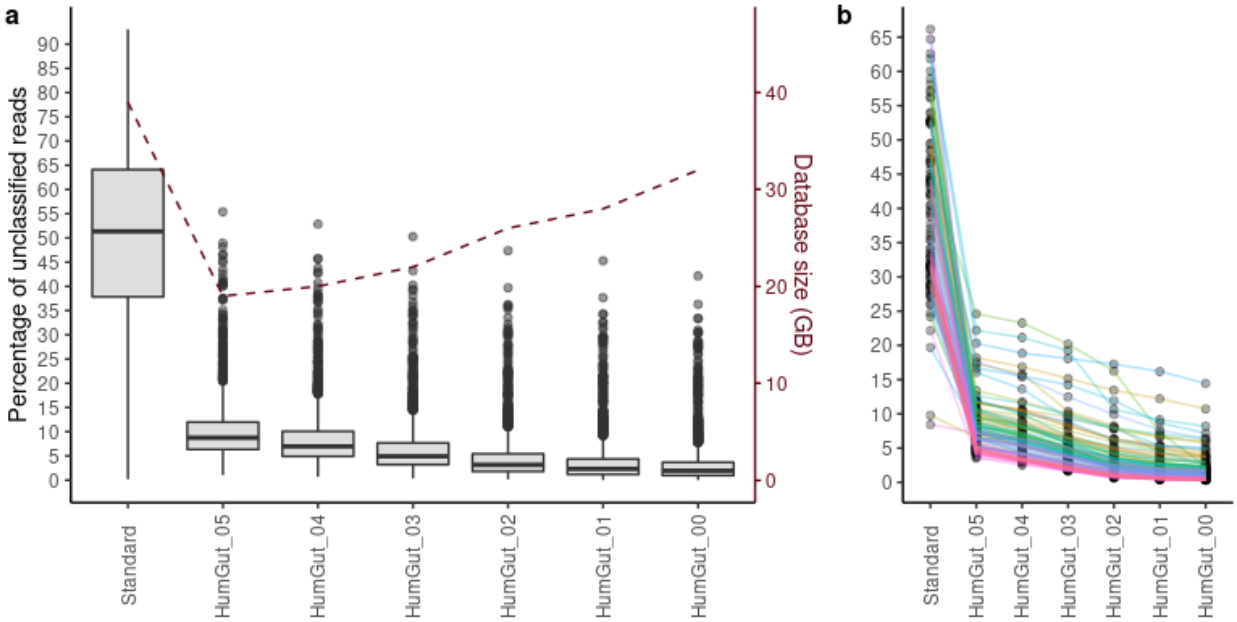


Figure 4. The performance of HumGut versions in comparison to the standard kraken2 database. a. Boxplot showing the distribution of unclassified reads for the 2,311 analyzed reference metagenome samples. The dashed line represents the k-mer database sizes. Every database version includes standard human, viral, and fungal sequences, in addition to database-specific (sub)sets of bacteria, and the difference in size is only due to differences in the latter. b. Classification of an additional 100 human gut metagenomes, not part of the reference set. Each dot represents a sample. The results for the same sample analyzed with different database versions are connected with a line of the same color.

Taxa abundances

We used the bracken software, and the kraken2 results, to re-estimate species abundance in the 2,311 classified human gut metagenomes. This task was performed using the HumGut_01 version as a trade-off between required computer memory and the resulting numbers of unclassified reads.

We noted that the most abundant species was the *uncultured Clostridiales bacterium* (NCBI taxonomy ID 172733), present in 99% of the samples with a 9.82 % average abundance. We also noted that 41 of 100 top species were annotated as "uncultured," i.e., MAGS (Supplementary material, **Figure S2**). Our previous clustering results indicated that many of these MAGS, represented by the same taxonomy ID, belonged to several hundred different clusters at $D = 0.05$ threshold (representing species delineation). This suggested that although they shared names and taxonomy IDs, they could, in fact, represent several hundred different species. To ensure inclusion of results reflecting only true abundance /prevalence of a single species, all "uncultured" species were therefore excluded from the bracken results.

We compiled a list of top remaining species. We found that there were 129 species present in more than 90% of the samples, suggesting that they represent a core community of healthy human gut microbiota. Unsurprisingly, the list was capped by *Bacteroides vulgatus* with 3.21 % average abundance, followed by *Bacteroides uniformis* at 2.42 %. All abundances were computed as readcount per genome megabase, reflecting cell abundances rather than the amount of DNA from each taxon in a sample.

There was a high correlation between the core species average abundances based on their continent of origin. A high correlation, as presented in **Figure 5**, was primarily observed between samples coming from Europe (n = 879), Asia (n = 840), and North America (n = 344) (Pearson $R > 0.9$, P-value < 0.05), showing that the core community is highly stable and geography-independent. The weakest linear relationship was observed between samples originating from Africa (n = 167) and North America (R =

153 0.79). We did not include samples from Australia (n = 20) and South America (n = 61) because of their
154 small sample sizes.

155 A list of top species found in > 80% of infants is presented in the supplementary material (**Figure S3**).

156 Data on the participation of non-bacterial reads and the distribution of reads at the phylum level are
157 presented in the supplementary material (**Figures S4, S5**).

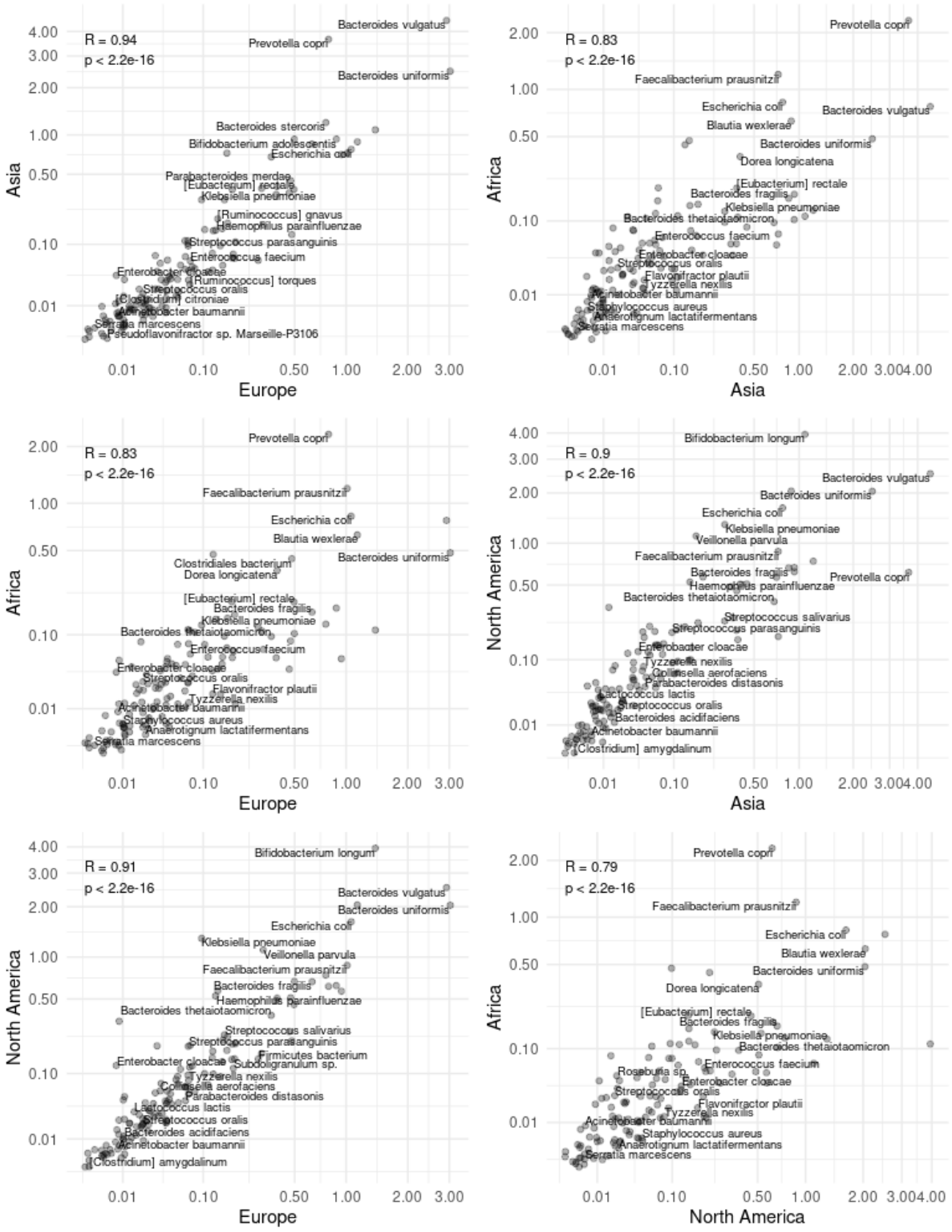


Figure 5. Average abundance scatterplots for the most prevalent healthy human gut bacterial species, 129 in total, found present in > 90% of samples worldwide. Each dot represents a bacterial species, and axes show their average abundance in respective

continents (depicted in axes' titles). Samples originating from Europe, Asia, and North America showed a high Pearson's correlation (R > 0.9). Both axes for all panels were square-root-transformed to aid visualization; however, the axes' ticks reflect the actual abundances.

We went further to calculate the number of reported species per sample by first rarefying the number of reads to the lowest depth found in our samples. We found that when MAGs were included, on average, there were 1,195 species per sample. The range of the number of species was from 108 to 2,250. The lowest extreme was found mainly in infants. When MAGs were removed, the average number of species dropped to 999, while the range of species was from 86 to 1,999.

Again, we wanted to check if the 2,311 metagenomes clustered together based on geography, using species abundances for comparison. We generated PCA plots based on the species' read counts, as shown in **Figure 6**. We found that African and South American samples (exclusively represented by samples originating from Peru) showed closer positioning in the PCA ordination plot (left panel) compared to samples from other continents. As expected, the clustering showed a gradient when considering the sampled person's age, i.e., infants showed a distinctly different composition. Samples from Europe, Asia, Australia, and North America did not form distinct clusters or gradients. The PCA loadings (right panel) show that *Prevotella* species were more abundant in Peruvian and African samples. In contrast, the *Bacteroides* species lay on the opposite side of the plot, indicating a negative correlation to the former. Infant samples were more abundant in *Escherichia* species.

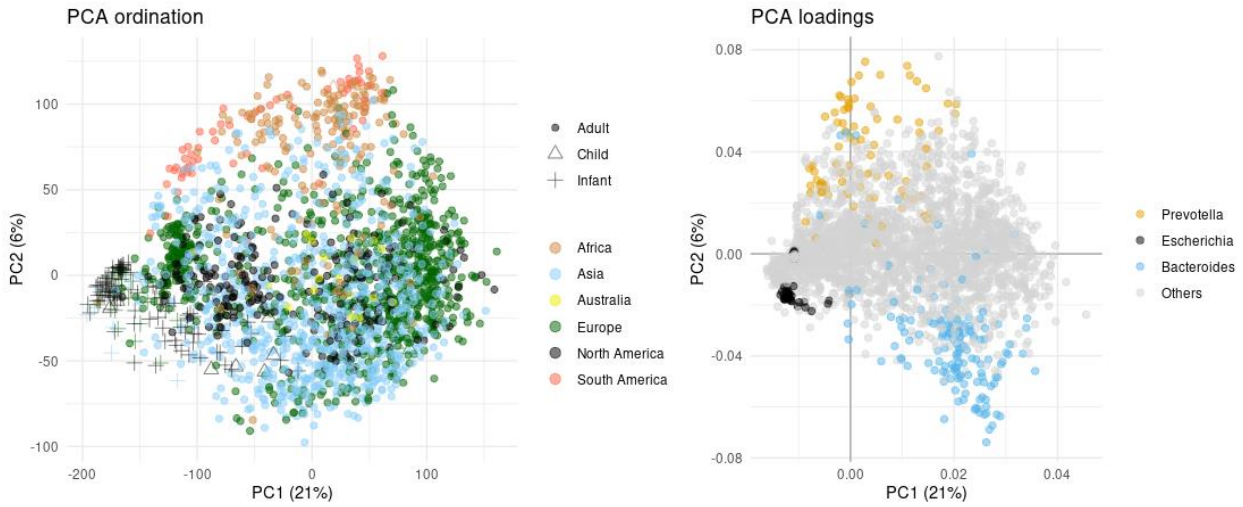


Figure 6. PCA score plot and loading plot. The left panel shows a PCA ordination plot, where each dot marks a metagenome sample. Color and marker indicate the continent and age group. The right panel shows the corresponding PCA loadings, where each dot is a species. The colors reflect some highlighted genera.

Discussion

All HumGut versions showed superior performance in terms of assigned reads compared to the standard kraken2 database, while demanding far less computational resources. We consider this to be a strong argument in favor of HumGut's comprehensiveness and utility. Classifying a record-high proportion of reads per sample, HumGut aids the accuracy of taxonomic classification, which in turn facilitates a next-generation exploration of the human gut microbiome.

To the best of our knowledge, HumGut is the only validated, publicly available genome collection that can serve as a global reference for bacteria inhabiting the gut of healthy humans, highlighting its importance for future gut microbiome studies (available for download in http://arken.nmbu.no/~larssn/humgut/index.htm).

Our analysis showed that the diversity of gut samples across the world is not profoundly affected by geography; therefore, having a global genome collection like HumGut is not only necessary, but

reasonable. We found 129 bacterial species present in more than 90% of the samples, regardless of the country of origin. This group of species, led by *Bacteroides vulgatus,* represents what we think is the core human gut bacterial community.

Having *B. vulgatus* head the list was not surprising for us, considering that top-scoring genomes in our collection belonged to this species as well. *B. vulgatus* has commonly been found in human guts [14]; however, its global prevalence, or the global prevalence of any bacterial species for that matter, was previously impossible to establish. We believe that by revealing the core human gut microbiome and the average number of species per individual (ca. 1,000), we cast light onto a crucial aspect of human health that can serve as a pillar for future diagnostic interventions.

Although samples shared hundreds of species regardless of their continent of origin, our analysis showed that samples originating from Africa and South America were rich in *Prevotella* species and poor in *Bacteroides*, which made them cluster in our principal component analysis. A *Prevotella* – *Bacteroides* antagonism and their correlation to lifestyle and diet have long been described in literature[15,16]. Our results are, therefore, consistent with these findings.

It is essential to state that these results do not consider species represented by MAGs. We decided not to report their abundance after observing that such genomes formed several hundred different clusters at the D = 0.05 threshold level, i.e., they had been assigned the same taxonomic species ID but scattered into hundreds of different clusters with 95% sequence identity. We consider these genomes to be an essential component of our collection. However, we believe that their current NCBI taxonomy IDs must reflect their individuality before we can include them in the characterization of gut microbiome taxonomy and abundance analysis. This will especially be important for our future work of linking functions to clusters based on the genomes they harbor.

Not all recently published human gut genomes and MAGs (retrieved specifically from human gut samples) qualified for HumGut inclusion, and many of those included were encountered in a limited number of metagenomes. This seemed to be the case, especially for many MAGs published by Nayfach et al. (2019) and Pasolli et al. (2019). This may be due to several reasons, but one is, of course, that re-constructing genomes from short-read metagenome data is still a difficult task, risking the generation of poor-quality MAGs. Another contributing factor may be genomes representing unique microorganisms found in a limited number of individuals throughout the world. This raises the question of whether it is sensible to solely depend on locally re-constructed MAGs when it comes to comparing the microbiome composition of healthy individuals against diseased ones. We believe that using a unified and stable HumGut collection as a reference will lead to more reproducible science.

We note that the decision regarding which version the HumGut collection to employ depends on users' computational resources as well as the level of taxonomic resolution required. As mentioned above, we found a substantial genomic diversity in genomes assigned to the same taxonomy ID. We also saw many cases of the opposite, where even tight clusters of highly similar genomes sometimes come with many different taxonomy IDs. This suggests that using the highest resolution, with more than 160,000 genomes and 16,000 taxonomy IDs, is probably a waste of effort for most applications. On our website, we have prepared files for building a custom kraken2 database where all HumGut clusters also have been given artificial 'taxonomy IDs,' making it possible to classify to clusters instead of taxa. HumGut will continuously be updated as more genomes, and human gut metagenomes will become available for the public. As future work, we will also extend our approach to disease-associated genomes and metagenomes, in addition to other species found in human guts.

In conclusion, we believe that by using HumGut as a reference, the scientific community will be one step closer to method standardization sorely needed in the field of human gut microbiome analysis, and that

240 the discovery of potential microbiome markers will be facilitated with higher certainty in less time and
241 computational resources.

242 Methods

243 Human gut reference metagenomes

244 A set of publicly available human gut metagenome samples was collected first. These were used for
245 ranking all genomes in our search for human gut relevant genomes.

246 A text search for all human gut microbiome samples at the Sequence Read Archive (NCBI/SRA,
247 https://www.ncbi.nlm.nih.gov/sra) was performed. The list of hits was manually curated, keeping only
248 samples annotated as healthy individuals. NCBI/BioProject accessions of these projects were used to
249 locate the same data in the European Nucleotide Archive (EMBL-EBI/ENA, https://www.ebi.ac.uk/ena),
250 from which all samples were downloaded as compressed fastq-files, using the Aspera download system
251 (https://www.ibm.com/products/aspera). This resulted in 3,654 metagenome runs (samples) covering
252 58 BioProjects. This collection contained more than 90 billion read pairs, covering human guts from all
253 continents. In addition, 95 samples containing gut metagenome data from primates were also
254 downloaded, only used as an outgroup for the comparison of the human gut samples.

255 A subset of this collection was used as a reference group of samples. For many BioProjects, some
256 samples tended to be very similar to each other. We presume this was due to persons sampled being
257 from the same geographical sub-population, sharing genetics, lifestyle, etc., factors that may affect the
258 human gut microbiome. To avoid too much bias in the direction of such heavily sampled sub-
259 populations, samples were initially clustered, then, from each cluster, one member was selected for our
260 reference group.

From each metagenome sample, a MinHash sketch of 10,000 21-mers was computed using the MASH software [17]. Singletons were discarded. Next, the MASH distances between all pairs of samples were calculated based on these sketches. A MASH distance close to 0 means two metagenomes are very similar, sharing many 21-mers. Next, hierarchical clustering with complete linkage was computed, and samples partitioned at a selected distance threshold. This means the resulting clusters have a 'diameter' no larger than this chosen threshold. The medoid sample from each cluster, i.e., the one with the minimum sum of distances to all members of the cluster, was retained as the reference sample representing its cluster.

Genome collections

The primary source of bacterial genomes was the NCBI/Genome, the GenBank repository at ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/. At the time of writing, >427,000 genomes were downloaded from this site. In addition, recently published genomes, specifically obtained from human guts, were collected (**Table 2**). From the Metagenome-Assembled Genomes (MAGs), only the highest quality subsets, as annotated by the authors, were downloaded. In total, >486,000 genomes were considered.

Table 2. **Recently published genomes explicitly retrieved from the human gut**

Genome set	Number of genomes	Type	References
UMGS	27058	MAG	Almeida et al. (2019)[3]
IGG	23793	MAG	Nayfach et al. (2019)[5]
SGB	4933	MAG	Pasolli et al. (2019)[6]
ZOU	1523	cultivated	Zou et al. (2019)[4]
HGG	1354	cultivated	Forster et al (2019)[2]

Genome ranking

For all genomes, again, the MASH software was used to compute sketches of 1,000 21-mers, including singletons[18]. Based on the genome-sketches, the number of shared hashes (*w*) between a genome and each reference metagenome was computed using MASH screen [18]. A high number of shared hashes between a genome and a metagenome sample means many 21-mers from the genome are also found in the metagenome sample. This indicates that the genome, or some close relative, is present in the sample.

The MASH screen compares the sketched genome hashes to all hashes of the metagenome, and if a genome has identity *I* to a genome in the metagenome, the binomial model means we expect to observe *w* shared hashes according to the equation

$$w = s \cdot I^k$$

where *s* is the sketch size (1,000), and *k* the length of the k-mers (21 in our case). Thus, for *I = 0.95,* we get *w = 340.56*, and we used the value *w = 341* as a lower threshold for considering a genome as present in a metagenome, given that identity *0.95* is regarded as a species delineation for whole-genome comparisons [19]. All *w*-values meeting this threshold were summed for each genome, resulting in a genome score, which was then used to rank them. The genome with the highest score was considered the most prevalent among the reference samples, and thereby the best candidate to be found in any human gut.

Even if a genome is absent from the reference metagenomes, its *w*-value will not, in general, be 0, since some 21-mers will overlap by chance. To investigate this, a list of 126 genera reported by many 16S-based studies to be found in the human gut were compiled. These represented seven different phyla (*Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Proteobacteria*, *Verrucomicrobia*, *Fusobacteria,* and *Synergistetes*). From our GenBank-collection, and using the NCBI/Taxonomy database, all genomes from

all the other phyla (excluding these seven phyla) were collected. There were 8,290 such genomes in total, which we expected to be absent from the human gut, or at least present at very low abundance, thereby producing a low *w*-value. For each of these genomes, we also computed the shared hashes with the reference samples as described above.

Genome clustering

The genomes were clustered from the ranked list of all genomes. Many genomes were very similar, some even identical. Due to errors introduced in sequencing and genome assembly, it made sense to group genomes and use one member from each group as a representative genome. Even without any technical errors, a lower meaningful resolution in terms of whole genome differences was expected, i.e., genomes differing in only a small fraction of their bases should be considered identical.

Again, the MASH software was used, and 1,000 21-mer sketches were computed for each genome, and the MASH distance between genomes was computed. The genomes were then grouped by the following greedy algorithm: Starting at the top of the ranked list, the first genome formed a cluster centroid and was removed from the list. Then, all other genomes with MASH distance below a given threshold to this centroid were assigned to this cluster and removed from the list. This was repeated for the remaining members of the list until all genomes were clustered. The centroid genomes formed the human gut genome collection. Using different distance thresholds produced various genome collections, i.e., using a threshold *D* will create clusters where no two centroids are closer than distance *D* from each other. Thresholds of 0.00, 0.01, 0.02, 0.03, 0.04 and 0.05 were used, each threshold giving a genome collection at gradually lower resolution.

Classifications

The kraken2 software was used for classifying reads from the metagenome samples. To see the effects of using a different database, the standard kraken2-database was used first. Next, custom databases

using the resolutions 0.00 up to 0.05 of the HumGut genome collection (see above) were made. In these databases, the standard libraries for the human genome, viruses, fungi, and vectors available from the kraken2 website were also included. Thus, only the prokaryotes (archaea and bacteria) were replaced with our HumGut genomes. All classifications were performed using default settings in kraken2.

Since kraken2, like most other software for taxonomic classification, uses the Lowest Common Ancestor (LCA) approach, many reads are assigned to ranks high up in the taxonomy. The bracken software [20] has been designed to re-estimate the abundances at some fixed rank, by distributing reads from higher ranks into the lower rank, based on conditional probabilities estimated from the database content. For each kraken2 database (standard and the six HumGut versions) a bracken database was also created and used to re-estimate all abundances at the species rank.

A Principal Component Analysis was conducted on the matrix of species readcounts for all metagenome samples, after the following transformation: All sample readcounts were rarefied to the lowest readcount (853,741), and a pseudo-count of 1 was added to all species before using Aitchison's centered log-ratio transform [21,22] to remove the unit-sum constraint otherwise affecting a PCA of such data.

Acknowledgment

This work was financially supported by Norway Research Council through R&D project grant no 283783 and 248792.

Author contributions

L.S. conceived the study. L.S. and P.H. worked out the technical aspects of the paper. All authors discussed and interpreted the results. P.H. wrote the article with equal input from all authors.

Competing interests statement

Both P.H. and F.T.H. are employed at Genetic Analysis AS, but all authors agree this fact does not represent a conflict of interest in the context of our manuscript.

Literature

1 A framework for human microbiome research. *Nature* **486**, 215-221, doi:10.1038/nature11209 (2012).

2 Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology* **37**, 186-192, doi:10.1038/s41587-018-0009-7 (2019).

3 Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499-504, doi:10.1038/s41586-019-0965-1 (2019).

4 Zou, Y. *et al.* 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology* **37**, 179-185, doi:10.1038/s41587-018-0008-8 (2019).

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

6 Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e620, doi:10.1016/j.cell.2019.01.001 (2019).

7 Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220-230, doi:10.1038/nature11550 (2012).

8 Shin, N.-R., Whon, T. W. & Bae, J.-W. Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends in Biotechnology* **33**, 496-503, doi:https://doi.org/10.1016/j.tibtech.2015.06.011 (2015).

9 Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology* **2**, 17004, doi:10.1038/nmicrobiol.2017.4 (2017).

10 Rajilić–Stojanović, M. *et al.* Global and Deep Molecular Analysis of Microbiota Signatures in Fecal Samples From Patients With Irritable Bowel Syndrome. *Gastroenterology* **141**, 1792-1801, doi:https://doi.org/10.1053/j.gastro.2011.07.043 (2011).

11 Cotillard, A. *et al.* Dietary intervention impact on gut microbial gene richness. *Nature* **500**, 585-588, doi:10.1038/nature12480 (2013).

12 Wallace, T. C. *et al.* Human gut microbiota and its relationship to health and disease. *Nutrition Reviews* **69**, 392-403, doi:10.1111/j.1753-4887.2011.00402.x (2011).

13 Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome biology* **20**, 257 (2019).

14 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65, doi:10.1038/nature08821 (2010).

15 Gorvitovskaia, A., Holmes, S. P. & Huse, S. M. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. *Microbiome* **4**, 15, doi:10.1186/s40168-016-0160-7 (2016).

16 Hjorth, M. F. *et al.* Prevotella-to-Bacteroides ratio predicts body weight and fat loss success on 24-week diets varying in macronutrient composition and dietary fiber: results from a post-hoc analysis. *International Journal of Obesity* **43**, 149-157, doi:10.1038/s41366-018-0093-2 (2019).

17 Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology* **17**, 132 (2016).

18 Ondov, B. D. *et al.* Mash Screen: High-throughput sequence containment estimation for genome discovery. *BioRxiv*, 557314 (2019).

19 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature communications* **9**, 5114 (2018).

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

21 Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27, doi:10.1186/s40168-017-0237-y (2017).

22 Aitchison, J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44**, 139-160 (1982).

Supplement

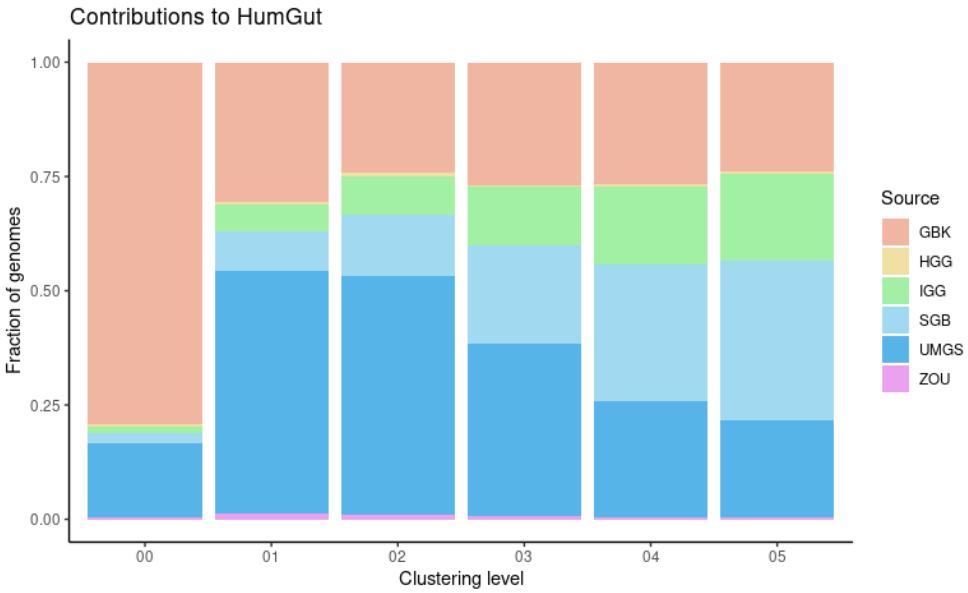


Figure S1. Participation of different genome sources in different clustering levels. The main source for clusters with D = 0.00 was GenBank, reflecting thusly the high proportion of genomes downloaded from there, rather than from other databases. As soon as clustering at 99% identity was performed (D = 0.01), MAGs and newly published cultivated genomes emerged, suggesting that they were different from one another in terms of sequence identity. Clustering at 95% (0.05) identity showed that the participation of GenBank genomes in the 4,779 genome groups was almost the same with the other MAG sources (~ 24%). Cultivated genomes – published by Zou et al (2019) and Forster et al (2019) – had a representation of ~1% each, IGG MAGs (Nayfach et al, 2019) comprised ~19%, ~21% of genomes originated from UMG MAGs (Almeida et al, 2019), while the biggest share belonged to SGB MAGs with ~35% (Passoli et al, 2019).

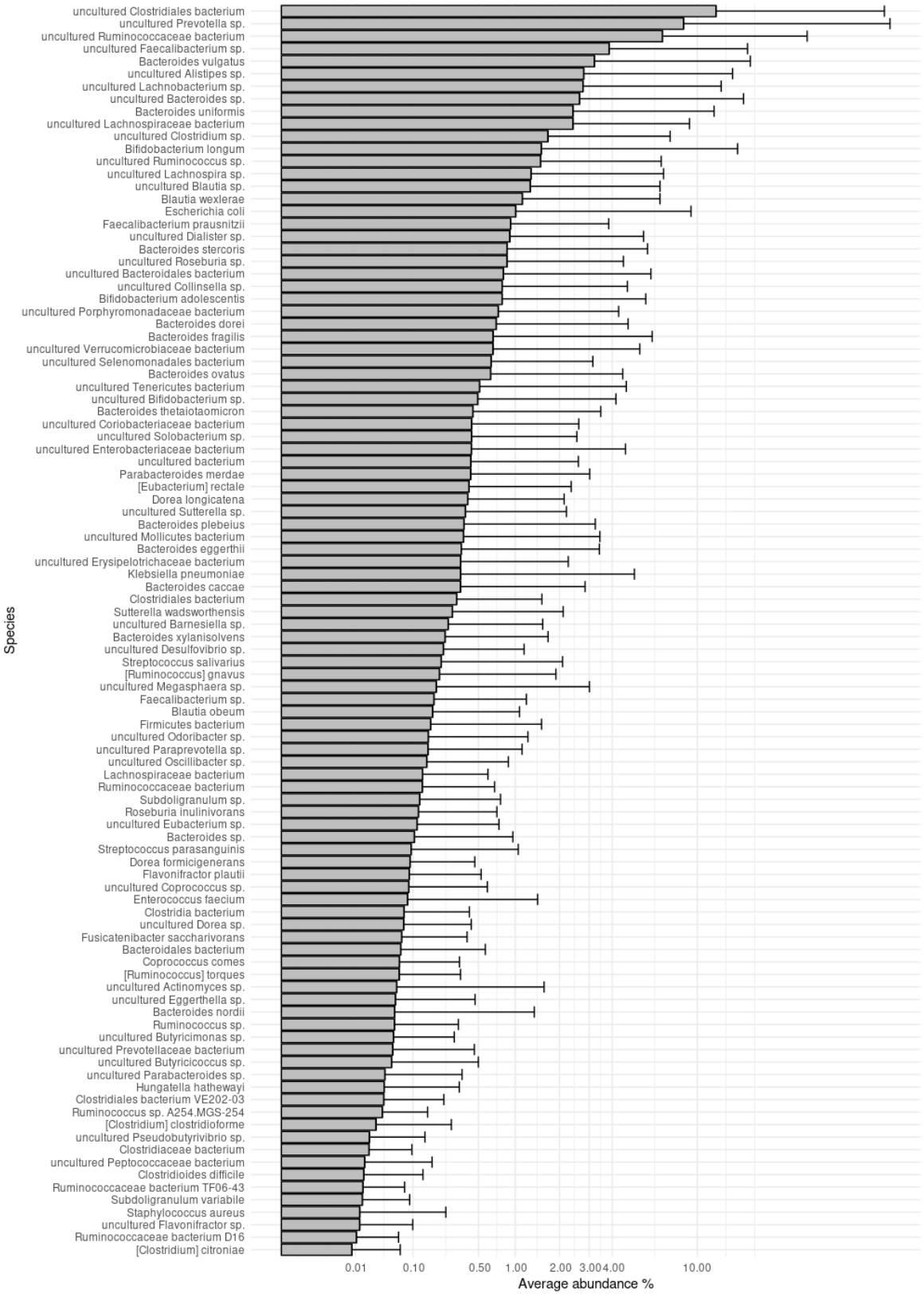


Figure S2. Top 100 abundant bacterial species in healthy human guts (MAGs included)

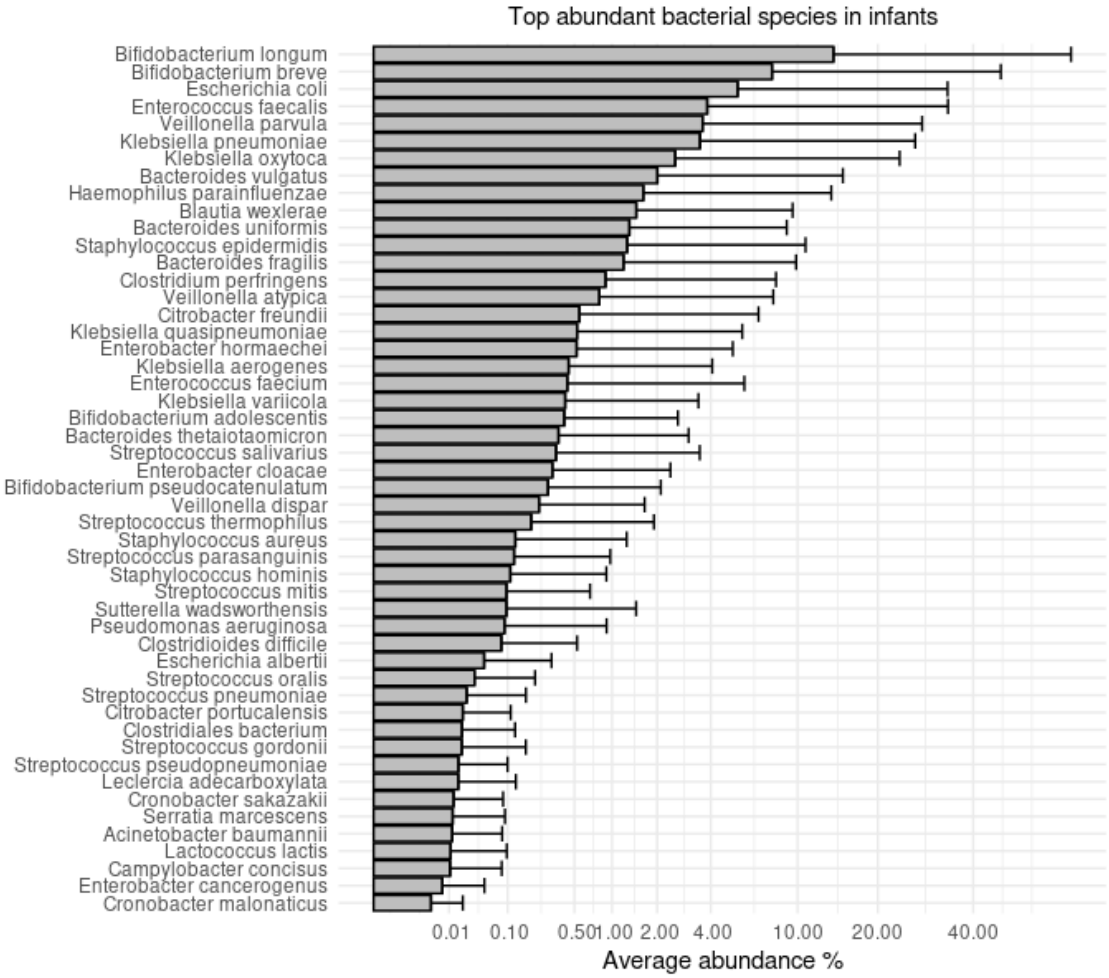
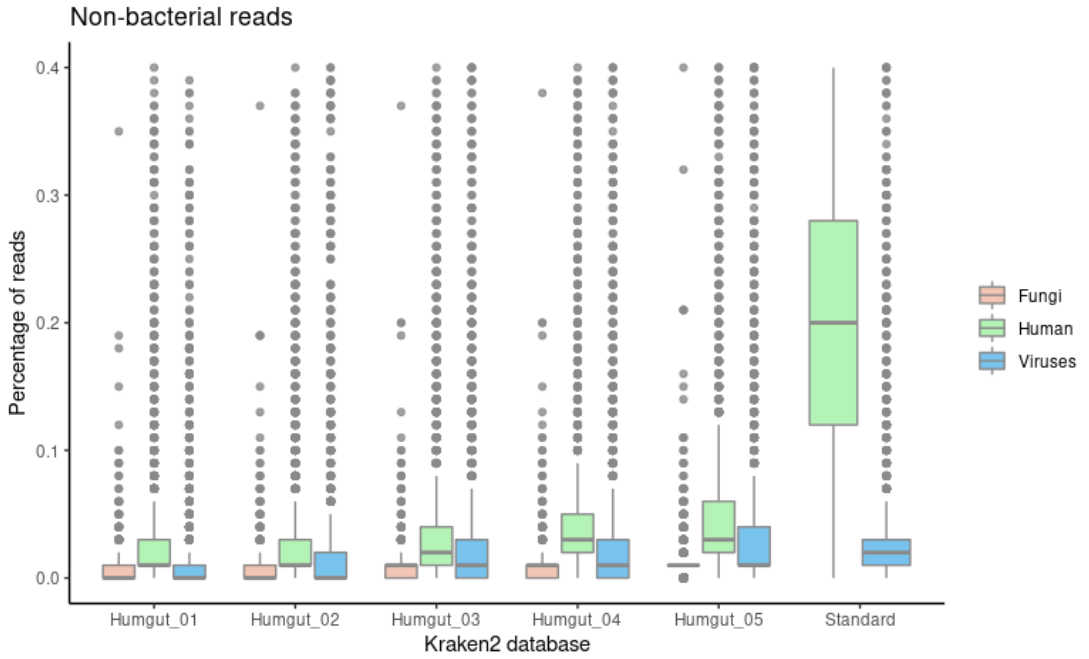


Figure S3. Top 50 abundant bacterial species that were encountered in > 80% of infants.

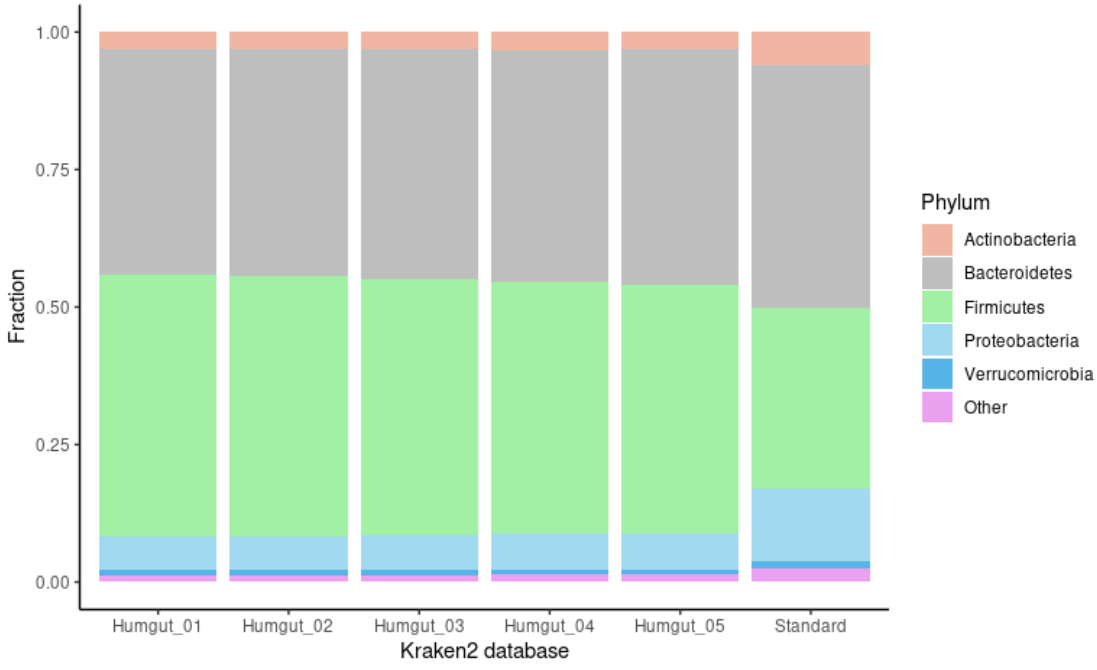


444

445 *Figure S4. Boxplot of reads classified as non-bacterial. The mean percentage of reads classified as non-bacterial is lower than*

446 *0.05% for HumGut, while reads classified as human were nearly 0.2% when the standard kraken2 database was used. Fungi were*

447 *not part of the standard database.*



448

449 *Figure S5. Bar charts showing the relative abundance of reads classified in the phylum level. All HumGut versions were very*

450 *consistent in their results while kraken2 standard database reported different proportions for these phyla. Relatively higher*

451 *fraction of reads were classified as Actinobacteria and Proteobacteria when samples were analyzed with the standard database.*