Title page

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

Authors: 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, P.O. Box 400, 2418 Elverum, Norway

[3]Genetic Analysis AS, Kabelgaten 8, 0580 Oslo, Norway

Pranvera Hiseni – ph@genetic-analysis.com; pranvera.hiseni@nmbu.no

Knut Rudi – knut.rudi@nmbu.no

Robert C. Wilson - robert.wilson@inn.no

Finn Terje Hegge - fth@genetic-analysis.com

Lars Snipen - lars.snipen@nmbu.no

Corresponding author: Pranvera Hiseni, Department of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O. Box 5003, 1432 Aas, Norway

Abstract

Background

A major bottleneck in the use of metagenome sequencing for human gut microbiome studies has been the lack of a comprehensive genome collection to be used as a reference database. Several recent efforts have been made to re-construct genomes from human gut metagenome data, resulting in a huge increase in the number of relevant genomes. In this work, we aimed to create a collection of the most prevalent healthy human gut prokaryotic genomes, to be used as a reference database, including both MAGs from the human gut and ordinary RefSeq genomes.

Results

We screened > 5,700 healthy human gut metagenomes for the containment of > 490,000 publicly available prokaryotic genomes sourced from RefSeq and the recently announced UHGG collection. This resulted in a pool of > 379,000 genomes that were subsequently scored and ranked based on their prevalence in the healthy human metagenomes. The genomes were then clustered at subspecies resolution, and cluster representatives were retained to comprise the HumGut collection. Using the Kraken2 software for classification, we find superior performance in the assignment of metagenomic reads, classifying on average 94.5% of the reads in a metagenome, as opposed to 86% with UHGG and 44% when using standard Kraken2 database. HumGut, half the size of standard Kraken2 database and directly comparable to the UHGG size, outperforms them both.

Conclusions

The HumGut collection contains > 30,000 genomes clustered at subspecies resolution and ranked by human gut prevalence. We demonstrate how metagenomes from IBD-patients map equally well to this collection, indicating this reference is relevant also for studies well outside the metagenome reference set used to obtain HumGut. We believe this is a valuable resource in a field in dire need of method standardization. All data and metadata, as well as helpful code, are available at http://arken.nmbu.no/~larssn/humgut/.

Keywords

Human gut microbiome, genome collection, database

Introduction

Significant efforts have been undertaken to characterize the human gut microbiome, both by microbe isolation and DNA sequencing[1]. A major contribution has also been made by de novo-assembled genomes (Metagenome-Assembled Genomes – MAGs), facilitated by the latest advances in bioinformatics tools[2-6]. As a wrap, a Unified Human Gastrointestinal Genome (UHGG) collection comprised of > 200,000 non-redundant reference genomes was recently announced[7], marking a major milestone in this field.

These studies have laid a solid foundation, identifying a vast variety of genomes encountered in human guts. However, none of them addresses the global prevalence of genomes within healthy people, i.e. providing information about their frequency of occurrence. This knowledge is essential for setting up a collection of human gut-associated prokaryotic genomes that reflects the worldwide *healthy human gut microbiome*. It is especially important for building custom databases intended to be used for comparative studies in human gastrointestinal microbiome research.

Regionally, studies have shown that the intestinal flora is greatly shaped by the environment[8] and that its composition can be linked to a range of diseases and disorders[9-12], thus we are now at a stage where gut microbiota therapeutic interventions are being introduced[13, 14]. However, the lack of a global reference for the intestinal flora in healthy humans represents a bottleneck[15]. This impedes both the understanding of gut microbiota on a worldwide scale and the introduction of large-scale intervention strategies.

The aim of this work was to create a single, comprehensive genome collection of gut microbes associated with healthy humans, called HumGut, as a universal reference for all human gut

microbiota studies. We utilized the UHGG collection, mentioned above, along with the NCBI RefSeq genomes. The strategy of building HumGut is outlined in **Figure 1**.

HumGut genomes are ranked by their containment in healthy human gut metagenomes collected worldwide. The most commonly encountered genomes (i.e. top-ranked on the list) were selected as taxa representatives during dereplication, securing thus a list of those most relevant.

While it may seem like a relatively simple concept, this work has only become possible with the recent development of bioinformatics tools that allow the swift screening of publicly available human gut metagenomes for the containment of the ever-growing pool of prokaryotic genomes.

Results

Reference metagenomes

More than 5,700 gut metagenome samples collected from healthy people of various ages worldwide were downloaded. These belonged to 72 different BioProjects. To avoid the bias of containing groups of highly similar samples, we computed the MASH distance between metagenomes within each BioProject, then clustered samples with ≥ 95% sequence identity. From each cluster we only kept the medoid sample, resulting thus in a collection of 3,534 healthy human gut metagenomes (**Figure 2a**).

On average, samples within each project shared a 90% sequence identity (D = 0.1), indicating a relatively high degree of similarity between one another. There were some outliers, however. Some infant samples (10 belonging to PRJNA473126 project and 1 to PRJEB6456), 10 samples from a project studying the human gut microbiome of vegetarians, vegans, and omnivores (PRJNA421881), and a sample from a study focusing on microbiome diversity among Cheyenne & Arapaho of Oklahoma (PRJNA299502), showed the highest dissimilarity with at least one other sample from the same project (D = 1) (**Figure 2b**).

92 We wanted to see if samples clustered based on their continent of origin (**Figure 2c)**. To do so, we
93 computed the average linkage hierarchical clustering of BioProjects. The distance between two
94 BioProjects is the mean pairwise distance between all their samples. Here, we also included a
95 BioProject containing primate gut metagenome samples (n = 95) as an outgroup against which all
96 human BioProjects were compared. The lowest and highest observed average MASH distances (*D* =
97 0.05, and *D* = 0.14, respectively) were between two sets of projects stemming from separate
98 continents each, one from Europe and the other from North America. These observations, together
99 with the mixed distribution of BioProjects in the cluster dendrogram, suggested that the clustering
100 of samples did not heavily depend on continent-of-origin. The primate samples were markedly
101 separated from the rest of the tree, showing an average distance of 0.22 from all other BioProjects.

102 From genomes to HumGut collection

103 The majority of genomes stemming from the UHGG collection (99%) and 48% of RefSeq genomes
104 qualified for inclusion in HumGut, resulting thus in a total collection of 381,779 genomes (**Figure 3a**).
105 The qualified genomes were contained within at least one reference metagenome. We inferred the
106 containment by computing sequence identity between genomes and metagenomes using MASH
107 screen, and considered a genome as *contained* when identity was ≥ 0.95.

108 By applying a rarefaction, we found that the number of new genomes saturated after screening for
109 ca. 1,000 metagenomes, indicating that with > 3,500 metagenomes very few new genomes will be
110 added if screening even more metagenomes from the same population (supplementary material,
111 **Figure S1**).

112 The most prevalent genomes, i.e., the genomes contained in most metagenomes, belonged to the
113 genus *Bacteroides*, led by *B. vulgatus* (also known as *Phocaeicola vulgatus*), found in more than 70%
114 of samples. It's worth noting that the UHGG collection contained no genome with this species name.
115 The genomes are named as *Bacteroides dorei* instead. We presume that is related to an earlier GTDB

116 database release used for genome taxonomic classifications by Almeida et al (GTDB-Tk v0.3.1;
117 database release 04-RS89)[7]. In the current version of GTDB the species *Phocaeicola vulgatus* is listed.

118 We performed clustering of genomes based on sequence similarity using the top-ranked genome as
119 a cluster centroid for each iteration. We initially applied an ANI threshold of 97.5% to compile a
120 HumGut collection of highest resolution (HumGut_97.5). This collection resulted in 30,691 genomes
121 with ≥ 50% genome completeness and ≤ 5% contamination. They were all given a GTDB-Tk
122 taxonomic annotation [16] as well as an NCBI taxonomy assignment.

123 These genomes were subsequently clustered further to form a coarser collection at 95% identity, the
124 HumGut_95 with 5,170 genomes. This corresponds roughly to species resolution [17].

125 Looking into genome sources, we found that 9% of HumGut_95 clusters were RefSeq-only genomes
126 (**Figure 3c**). These genomes, 756 in total, clustered into 460 HumGut_95 clusters, belonged to 125
127 different genera. Most of the genomes (299 in total) belonged to various *Streptococcus* species.

128 HumGut genome clusters

129 Not all species-level clusters were equally diverse, that is not all of them encompassed a similar
130 number of HumGut_97.5 clusters. The majority of HumGut_95 clusters (3,009 out of 5,100)
131 consisted of a single HumGut_97.5 cluster. On the other hand, the most diverse HumGut_95 cluster
132 was one built of 533 different HumGut_97.5 clusters, all named as *Agathobacter rectalis* with GTDB
133 taxonomy (*[Eubacterium] rectale ATCC 33656* with NCBI). It was followed by a group of 495
134 subspecies-level clusters consisting of various *Collinsella*-related species names, and a HumGut_95
135 cluster comprised of 400 different HumGut_97.5 clusters, all GTDB-named as *UBA11524*
136 *sp000437595*, and NCBI-named as *Faecalibacterium sp. CAG:74*.

137 Regarding taxonomy, many genomes were not given species names by GTDB, rather they were
138 named after the genus, family, order, or class they belong to. Similarly, the NCBI taxonomic
139 annotations for many genomes resulted in ambiguous names not specific to species, such as for

140 example *uncultured bacterium* or *Firmicutes bacterium*. This contributed greatly in a discrepancy
141 between the total number of species-level clusters (5,170 clusters in HumGut_95) and the total
142 number of distinct cluster names (3,310 GTDB names, 1,716 NCBI names).

143 There were also many species-level clusters that shared the same species name. This was especially
144 the case with various *Collinsella* clusters, where 81 different GTDB *Collinsella* species gave name to 7
145 different clusters each, on average. Comparably, 19 NCBI *Collinsella* species were seen in 44
146 different clusters on average.

147 Classifying the metagenome reads

148 We used the HumGut collection at both resolutions, in addition to the UHGG (species-level
149 collection, containing 4,644 genomes) and the standard Kraken2 database, to classify the
150 metagenomic reads from the 3,534 downloaded samples. On average, there were 56% unclassified
151 reads when using the standard Kraken2 database, while the average dropped substantially when any
152 one of the HumGut or the UHGG collection was utilized (UHGG = 14.1%, Humgut_95 = 11.7%, and
153 HumGut_97.5 = 5.4%, **Figure 4a**).

154 In comparison to the UHGG, both HumGut collections performed better. HumGut_95, a collection of
155 species-level representatives – much like the UHGG collection – classified on average 2.3% more
156 reads than the latter. With HumGut_97.5 as a custom database this increased by 8.7%, marking a
157 significant increase in recognized reads, with an obvious potential for improved classification
158 accuracy.

159 Both HumGut k-mer databases were smaller than the standard Kraken2 database of k-mers,
160 necessitating reduced computer memory to perform the analyses. The lowest memory was required
161 by the HumGut_95 database (Standard = 42.1 GB, UHGG = 20.9, HumGut_95 = 15.9 GB,
162 HumGut_97.5 = 24.9).

163 Analysis of an additional 963 gut metagenome samples (collected from people suffering from IBD),
164 not part of the reference set, showed similar results regarding the number of classified reads: 42.3%
165 unclassified reads on average when the Standard database was used, dropping to 6.2% with
166 HumGut_97.5 usage (**Figure 4b**).

167 Taxa abundances

168 We used the KrakenUniq as a means of identifying false positive classifications, and removing them
169 from the Kraken2 reports. We then used the Bracken software on the modified Kraken2 results, to
170 re-estimate species abundance in the classified human gut metagenomes. These tasks were
171 performed using HumGut_97.5 and GTDB taxonomy.

172 The results showed that, on average, healthy adults contained 202 species, people diagnosed with
173 IBD, 145, and infants, 79 species. The overall species number distribution is presented in **Figure 5a**.

174 In total, 52 species were found present in > 70% of healthy adult samples, led by *Agathobacter*
175 *rectalis, Blautia_A sp900066165, Bacteroides uniformis, KLE1615 sp900066985, Agathobaculum*
176 *butyriciproducens* and *Fusicatenibacter saccharivorans,* discovered in > 90% of healthy adult
177 samples, representing a core community of healthy adult human gut microbiota (**Figure 5b**). A
178 complete hierarchical linkage of samples, computed based on the abundance of these top 52
179 prevalent species, showed that African and South American (coming exclusively from Peru)
180 metagenomes were more distant from the rest, while two species were not encountered at all in
181 South American samples (*Alistipes onderdonkii* and *Lawsonibacter assacharolyticus*). In addition,
182 these samples clustered more distantly from the others on a PCA plot (built based on the readcounts
183 from all species), as depicted on **Figure 5c.** The PCA loadings show that *Prevotella* species were more
184 abundant in South American and African samples. In contrast, the *Alistipes* and *Bacteroides* species
185 and lay on the opposite side of the plot, indicating a negative correlation to the former.

186 Infant samples separated from the adult samples as well. They are represented with crosses instead
187 of dots on the PCA ordination plot, positioned on the leftmost part of the graph along PC1 axis. The

188 loading plot shows that *Escherichia coli* species exercise the largest effect on samples positioned
189 there. The most prevalent bacterium in infants was *Bifidobacterium longum* (68%), followed by *E.*
190 *coli* (64%).

191 *Bacteroides vulgatus*, which, after screening the metagenomes using the MASH screen software, was
192 the species of the top scoring genome, was no longer the most prevalent species among healthy
193 human guts when classifying with all HumGut genomes. This was due to a lower diversity among *B.*
194 *vulgatus* genomes, compared to *Agathobacter rectalis.* The genomes belonging to the former
195 grouped into 2 species-level clusters (D = 0.05), while the latter resulted in 16 such groups. It is
196 worth noting that we found the top *B. vulgatus* genome present in 2,536 healthy samples using
197 MASH screen, and we found this species present in 2,537 healthy samples using Kraken2-
198 KrakenUniq-Bracken classification tools. These almost identical results, obtained by two entirely
199 different approaches, increase confidence in the trustworthiness of these findings.

200 We also investigated the prevalence of species that only had RefSeq as a genome source in our
201 collection. *Streptococcus sanguinis* was found present in 73% of all samples (healthy infants and
202 adults, and IBD), followed by *Flavonifractor sp002161085*, *Escherichia sp005843885*, *Streptococcus*
203 *sp001587175*, *Pauljensenia sp000466265*, *Flavonifractor sp002161215*, *Actinomyces naeslundii*,
204 *Raoultella terrigena*, and *Mediterraneibacter sp900120155* (found in 10% - 36% of samples).

205 Discussion

206 The HumGut collection contains 30,691 genomes (HumGut_97.5), with a subset of 5,170 genomes
207 clustered at 95% sequence identity (HumGut_95). The criterion for including a genome in HumGut is
208 its prevalence in healthy human gut metagenomes.

209 Both HumGut versions showed superior performance in terms of assigned reads compared to the
210 standard Kraken2 database, while demanding far less computational resources, as presented in
211 **Figure 4**. We consider this to be a strong argument in favor of HumGut's comprehensiveness and

212 utility. Classifying a record-high proportion of classified reads per sample, HumGut aids the accuracy
213 of taxonomic classification, which in turn facilitates a next-generation exploration of the human gut
214 microbiome.

215 The vast majority of UHGG genomes qualified for inclusion in HumGut, as shown in **Figure 3a**.
216 However, in comparison to the UHGG collection, HumGut holds the advantage of containing more
217 relevant human gut prokaryotic genomes in its pool, reflected by the additional RefSeq genomes
218 that showed no sequence similarity with the qualified UHGG genomes, forming separate clusters of
219 95% sequence identity (**Figure 3b**). An example of its utility is the discovery of *Streptococcus*
220 *sanguinis* in > 70% of all metagenome samples, which would otherwise be impossible using the
221 UHGG collection as a custom Kraken2 database. Also, the identification of one of the most prevalent
222 species in human guts, *Bacteroides vulgatus*, would have been mistaken for *Bacteroides dorei*.

223 HumGut sets were built after ranking the genomes based on their prevalence among metagenomes
224 and using the top-ranked ones as cluster representatives. This has ensured that the collections only
225 contain genomes highly relevant to healthy human guts worldwide. Comparing the HumGut_95
226 collection to the UHGG collection (same resolution) shows that more metagenomic reads are
227 classified for the former. Additionally, its set of unique k-mers is 24% smaller in size than the UHGG.
228 This indicates the UHGG contains a higher genomic diversity, requiring memory which is not really
229 needed for successful read classification. These are rare genomes found in the occasional human gut
230 metagenome, but with low prevalence.

231 HumGut can serve as a global reference for bacteria inhabiting the gut of healthy humans,
232 highlighting its importance for future gut microbiome studies and is available for download
233 (http://arken.nmbu.no/~larssn/humgut/).

234 Our analysis showed that the diversity of gut samples across the world is not profoundly affected by
235 geography (**Figure 2**); therefore, having a global genome collection like HumGut is reasonable. We
236 found 50 bacterial species present in more than 70% of the samples, regardless of the country of

237 origin. This group of species, led by *Agathobacter rectalis,* represents what we think is the core
238 human gut bacterial community (**Figure 5b**). We must, however, cautiously refer to *A. rectalis* as the
239 most prevalent/abundant species found in human gut samples. That because we found this species
240 to be highly diverse in sequence identity. In our collection, we have 16 different species-level
241 clusters, and more than 530 subspecies-level clusters with this name.

242 We discovered that, on average, healthy adults contain around 60 bacterial species more than IBD
243 subjects, and around 120 species more than healthy infants (**Figure 5a**). A low microbiome
244 complexity among the latter two groups is well documented in literature[18-22].

245 Although we found a great homogeneity of top prevalent species among healthy adults, our analysis
246 showed that samples originating from Africa and South America were rich in *Prevotella* and poor in
247 *Bacteroides*, which made them cluster in our principal component analysis, as depicted in **Figure 5c**.
248 A *Prevotella* – *Bacteroides* antagonism and their correlation to lifestyle and diet have long been
249 described in literature[23, 24]. Our results are, therefore, consistent with these findings.

250 We have demonstrated that HumGut is useful in research that goes beyond studying healthy
251 subjects, exemplified by the equally high number of classified metagenomic reads collected from IBD
252 subjects.

253 A challenge that remains is the nomenclature of species in our genome collection. There is a
254 profound inconsistency between the total number of species-level clusters and the total number of
255 names annotating them (a ratio of 1.5:1 with GDTB-based annotation, and 3:1 with NCBI names). We
256 believe that as long as not all names reflect species individuality, it will be difficult to truly explore
257 the composition differences between various cohorts, in addition to posing a challenge in studies
258 linking functions to species. On our website, we have prepared files for building a custom Kraken2
259 database where all HumGut clusters also have been given artificial 'taxonomy IDs,' making it possible
260 to classify to clusters instead of taxa. We note that the decision regarding which version the HumGut

261 collection to employ depends on users' computational resources as well as the level of taxonomic
262 resolution required.

263 As future work, we will also extend our approach to more disease-associated genomes and
264 metagenomes, in addition to screening for gut genomes that will expectedly be published in the
265 future.

266 Conclusion

267 We believe that by using HumGut as a reference, the scientific community will be one step closer to
268 method standardization sorely needed in the field of human gut microbiome analysis, and that the
269 discovery of potential microbiome markers will be facilitated with higher certainty in less time and
270 computational resources.

271 Methods

272 Human gut reference metagenomes

273 A set of publicly available human gut metagenome samples were collected and used for ranking all
274 genomes in the search for human gut relevant ones. A text search for all human gut microbiome
275 samples at the Sequence Read Archive (NCBI/SRA, https://www.ncbi.nlm.nih.gov/sra) was
276 performed. The list of hits was manually curated, keeping only samples with > 1,000,000 reads and
277 annotated as healthy individuals. NCBI/BioProject accessions of these projects were used to locate
278 the same data in the European Nucleotide Archive (EMBL-EBI/ENA, https://www.ebi.ac.uk/ena),
279 from which all samples were downloaded as compressed fastq-files, using the Aspera download
280 system (https://www.ibm.com/products/aspera). This resulted in 5,737 healthy metagenomes
281 (samples) covering 74 BioProjects. For many BioProjects, some samples tended to be very similar to
282 each other, presumably due to samples collected from individuals sharing the same lifestyle,
283 geographical sub-population, genetics, or other factors that may affect the human gut microbiome.
284 To avoid too much bias in the direction of such heavily sampled sub-populations, samples from the

285 same BioProject were clustered. From each metagenome sample, a MinHash sketch of 10,000 k-
286 mers was computed using the MASH software[25], discarding singleton k-mers (21-mers). Based on
287 these sketches the MASH distances between all pairs of samples were calculated. A MASH distance
288 close to zero means two samples are very similar, sharing most of their k-mers. Next, hierarchical
289 clustering with complete linkage was computed, and samples were partitioned at a 0.05 distance
290 threshold, resulting in clusters with 'diameters' no larger than this chosen threshold. The medoid
291 sample from each cluster, i.e., the one with the minimum sum of distances to all members of the
292 cluster, was retained as the reference sample representing its cluster. This resulted in the set of
293 3,534 healthy metagenome samples. Below we refer to this collection as *MetHealthy*.

294 The same procedure was utilized to collect another set of metagenomes from patients diagnosed
295 with Irritable Bowel Disease (Ulcerative Colitis, or Crohn's Disease). From initially 2,064
296 metagenomes the clustering resulted in a collection of 963 metagenomes covering 14 BioProjects.
297 This is the *MetIBD* collection. Finally, a set of 95 samples containing gut metagenome data from
298 primates was collected and used as an outgroup in a comparison of the human gut metagenomes.
299 The metagenomes' metadata is included in the **supplementary Table 1**.

300 Genome collections

301 The main source was the recently published Unified Human Gut Genomes (UHGG) collection,
302 containing 286,997 genomes exclusively related to human guts:
303 http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-
304 gut/v1.0/all_genomes/. The other source was NCBI/Genome, the RefSeq repository at
305 ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/ and
306 ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/archaea/. At the time of writing, ~204,000 genomes
307 were downloaded from this site.

308 Metadata about the genomes considered and qualified for HumGut are presented on supplementary
309 **Table 2**.

310 ## Genome ranking

311 Only metagenomes collected from healthy individuals, *MetHealthy*, were used in this step. For all
312 genomes, the MASH software was again used to compute sketches of 1,000 k-mers, including
313 singletons[26]. The MASH screen compares the sketched genome hashes to all hashes of a
314 metagenome, and, based on the shared number of them, estimates the genome sequence identity *I*
315 to the metagenome. Given that $I = 0.95$ (95% identity) is regarded as a species delineation for whole-
316 genome comparisons[17], it was used as a soft threshold to determine if a genome was contained in a
317 metagenome. Genomes meeting this threshold for at least one of the *MetHealthy* metagenomes
318 were qualified for further processing. Then the average *I*-value across all *MetHealthy* metagenomes
319 was computed for each genome, and this prevalence-score was used to rank them. The genome with
320 the highest prevalence-score was considered the most prevalent among the *MetHealthy* samples,
321 and thereby the best candidate to be found in any healthy human gut. This resulted in a list of
322 genomes ranked by their prevalence in healthy human guts.

323 ## Genome clustering

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

329 The clustering of the genomes was performed in two steps, like the procedure used in the dRep
330 software[27], but in a greedy way based on the ranking of the genomes. The huge number of genomes
331 (hundreds of thousands) made it extremely computationally expensive to compute all-versus-all
332 distances. The greedy algorithm starts by using the top ranked genome as a cluster centroid, and
333 then assigns all other genomes to the same cluster if they are within a chosen distance *D* from this

334 centroid. Next, these clustered genomes are removed from the list, and the procedure is repeated,
335 always using the top ranked genome as centroid.

336 The whole-genome distance between the centroid and all other genomes was computed by the
337 fastANI software[17]. However, despite its name, these computations are slow in comparison to the
338 ones obtained by the MASH software. The latter is, however, less accurate, especially for
339 fragmented genomes. Thus, we used MASH-distances to make a first filtering of genomes for each
340 centroid, only computing fastANI distances for those who were close enough to have a reasonable
341 chance of belonging to the same cluster. For a given fastANI distance threshold *D* we first used a
342 MASH distance threshold $D_{mash} >> D$ to reduce the search space. In supplementary material, **Figure**
343 **S2,** we show some results guiding the choice of D_{mash} for a given *D*.

344 A distance threshold of D = 0.05 is regarded as a rough estimate of a species, i.e. all genomes within
345 a species are within this fastANI distance from each other[16, 17]. This threshold was also used to arrive
346 at the 4,644 genomes extracted from the UHGG collection and presented at the MGnify website.
347 However, given shotgun data, a larger resolution should be possible, at least for some taxa. For this
348 reason, we started out with a threshold D = 0.025, i.e. half the "species radius." An even higher
349 resolution was tested (*D = 0.01*), but the computational burden increases vastly as we approach
350 100% identity between genomes. It is also our experience that genomes more than ~98% identical
351 are very difficult to separate, given today's sequencing technologies[28]. However, the genomes found
352 at D = 0.025 (HumGut_97.5) were also again clustered at D = 0.05 (HumGut_95) giving two
353 resolutions of the genome collection.

354 The taxonomic annotation of the genomes was performed with GTDB toolkit (GTDB-Tk, release 05-
355 RS95, https://github.com/Ecogenomics/GTDBTk)[16], but in the genome metadata tables we provide
356 on our website, we made efforts to list the corresponding NCBI Taxonomy names and ID's for all
357 genomes.

358 All UHGG genomes were already checked for completeness and contamination[7]. The completeness
359 and contamination of RefSeq genomes was performed using CheckM
360 (https://ecogenomics.github.io/CheckM/)[29]. The handful genomes not having > 50% completeness
361 and < 5% contamination were discarded.

362 Metagenome classifications

363 The Kraken2 software was used for classifying reads from the metagenome samples[30]. To see the
364 effects of using a different database, the standard Kraken2-database was compared by custom
365 databases built from the 4,644 UHGG genomes at the MGnify website as well as our HumGut
366 collections. In all custom databases, the standard Kraken2 library for the human genome was also
367 included, since host contamination is quite normal in metagenome data. All classifications were
368 performed using default settings in Kraken2.

369 Since Kraken2, like most other software for taxonomic classification, uses the Lowest Common
370 Ancestor (LCA) approach, many reads are assigned to ranks high up in the taxonomy. The Bracken
371 software[31] has been designed to re-estimate the abundances at some fixed rank, by distributing
372 reads from higher ranks into the lower rank, based on conditional probabilities estimated from the
373 database content. A Bracken database (100-mers) was created for HumGut_97.5 database and used
374 to re-estimate all abundances at the species rank.

375 If counting all listed taxa, regardless of low readcounts, the Kraken2 is known to produce many false
376 positives[32], i.e. list taxa as present when they are in fact not. The KrakenUniq software has been
377 developed to handle this problem[32]. We ran it to classify the metagenome reads for both healthy
378 and IBD metagenomes. The overall results from both Kraken2 and KrakenUniq tools were similar,
379 but KrakenUniq also reports the number of unique k-mers in each genome covered by the reads. On
380 the other hand, only Kraken2 reports are compatible for running Bracken software. Since we were
381 interested in both - that is finding the true positive identifications, and their estimated abundances -
382 we combined the two approaches. For each sample, we found from the KrakenUniq report a k-mer

count threshold, following the authors recommendations (2,000 unique k-mers per 1,000,000 sequencing reads depth)[32]. Taxa falling below this threshold were given zero read counts in the corresponding modified Kraken2 reports. We then ran Bracken on these modified Kraken2 reports.

A Principal Component Analysis was conducted on the matrix of species readcounts for all metagenome samples, after the following transformation: a pseudo-count of 10 was added to all species before using Aitchison's centered log-ratio transform[33, 34] to remove the unit-sum constraint otherwise affecting a PCA of such data.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The HumGut genome collection and all associated metadata can be found at http://arken.nmbu.no/~larssn/humgut/. This also includes files and recipes for building Kraken2 databases using these genomes.

Competing interests

Both PH and FTH 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.

Funding

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

Acknowledgements

Not applicable

Author contributions

LS conceived the study. LS and PH worked out the technical aspects of the paper. All authors discussed and interpreted the results. PH wrote the article with equal input from all authors.

Literature

1. Methé, B.A., et al., *A framework for human microbiome research.* Nature, 2012. **486**(7402): p. 215-221.
2. Forster, S.C., et al., *A human gut bacterial genome and culture collection for improved metagenomic analyses.* Nature Biotechnology, 2019. **37**(2): p. 186-192.
3. Almeida, A., et al., *A new genomic blueprint of the human gut microbiota.* Nature, 2019. **568**(7753): p. 499-504.
4. Zou, Y., et al., *1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses.* Nature Biotechnology, 2019. **37**(2): p. 179-185.
5. Nayfach, S., et al., *New insights from uncultivated genomes of the global human gut microbiome.* Nature, 2019. **568**(7753): p. 505-510.
6. Pasolli, E., et al., *Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.* Cell, 2019. **176**(3): p. 649-662.e20.
7. Almeida, A., et al., *A unified catalog of 204,938 reference genomes from the human gut microbiome.* Nature Biotechnology, 2020.
8. Rothschild, D., et al., *Environment dominates over host genetics in shaping human gut microbiota.* Nature, 2018. **555**(7695): p. 210-215.
9. Lozupone, C.A., et al., *Diversity, stability and resilience of the human gut microbiota.* Nature, 2012. **489**(7415): p. 220-230.
10. Shin, N.-R., T.W. Whon, and J.-W. Bae, *Proteobacteria: microbial signature of dysbiosis in gut microbiota.* Trends in Biotechnology, 2015. **33**(9): p. 496-503.
11. Halfvarson, J., et al., *Dynamics of the human gut microbiome in inflammatory bowel disease.* Nature Microbiology, 2017. **2**(5): p. 17004.
12. Rajilić–Stojanović, M., et al., *Global and Deep Molecular Analysis of Microbiota Signatures in Fecal Samples From Patients With Irritable Bowel Syndrome.* Gastroenterology, 2011. **141**(5): p. 1792-1801.
13. Cotillard, A., et al., *Dietary intervention impact on gut microbial gene richness.* Nature, 2013. **500**(7464): p. 585-588.
14. Wallace, T.C., et al., *Human gut microbiota and its relationship to health and disease.* Nutrition Reviews, 2011. **69**(7): p. 392-403.
15. McBurney, M.I., et al., *Establishing What Constitutes a Healthy Human Gut Microbiome: State of the Science, Regulatory Considerations, and Future Directions.* The Journal of nutrition, 2019. **149**(11): p. 1882-1895.
16. Chaumeil, P.-A., et al., *GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.* Bioinformatics, 2019. **36**(6): p. 1925-1927.
17. Jain, C., et al., *High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.* Nature communications, 2018. **9**(1): p. 5114.

18. Manichanh, C., et al., *Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach.* 2006. **55**(2): p. 205-211.
19. Manichanh, C., et al., *The gut microbiota in IBD.* Nature Reviews Gastroenterology & Hepatology, 2012. **9**(10): p. 599-608.
20. Matsuoka, K. and T. Kanai, *The gut microbiota and inflammatory bowel disease.* Seminars in Immunopathology, 2015. **37**(1): p. 47-55.
21. Rodríguez, J.M., et al., *The composition of the gut microbiota throughout life, with an emphasis on early life.* Microbial Ecology in Health and Disease, 2015. **26**(1): p. 26050.
22. Moore, R.E. and S.D. Townsend, *Temporal development of the infant gut microbiome.* 2019. **9**(9): p. 190128.
23. Gorvitovskaia, A., S.P. Holmes, and S.M. Huse, *Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle.* Microbiome, 2016. **4**(1): p. 15.
24. 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, 2019. **43**(1): p. 149-157.
25. Ondov, B.D., et al., *Mash: fast genome and metagenome distance estimation using MinHash.* Genome biology, 2016. **17**(1): p. 132.
26. Ondov, B.D., et al., *Mash Screen: High-throughput sequence containment estimation for genome discovery.* BioRxiv, 2019: p. 557314.
27. Olm, M.R., et al., *dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication.* The ISME Journal, 2017. **11**(12): p. 2864-2868.
28. Snipen, L., et al., *Reduced Metagenome Sequencing for strain-resolution taxonomic profiles.* Microbiome, 2021.
29. Parks, D.H., et al., *CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.* Genome Res, 2015. **25**(7): p. 1043-55.
30. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome biology, 2019. **20**(1): p. 257.
31. Lu, J., et al., *Bracken: estimating species abundance in metagenomics data.* PeerJ Computer Science, 2017. **3**: p. e104.
32. Breitwieser, F.P., D.N. Baker, and S.L. Salzberg, *KrakenUniq: confident and fast metagenomics classification using unique k-mer counts.* Genome Biology, 2018. **19**(1): p. 198.
33. Weiss, S., et al., *Normalization and microbial differential abundance strategies depend upon data characteristics.* Microbiome, 2017. **5**(1): p. 27.
34. Aitchison, J., *The statistical analysis of compositional data.* Journal of the Royal Statistical Society: Series B (Methodological), 1982. **44**(2): p. 139-160.

Figure legend

Figure 1. HumGut overview. HumGut represents a collection of genomes and MAGs contained in 3,534 healthy human gut metagenomes. To be considered as contained, a genome shared at least 0.95 sequence identity with at least one of the metagenomes (inferred by the number of shared hashes). The qualified genomes were scored based on the average sequence identity 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 and fastANI distance (D). The top-ranked genome formed a cluster centroid. Around 30,600 clusters were formed applying a D=0.025-threshold. 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.

Figure 2. An outline of the metagenomes used in this study. ***a****. The geographical, age, and gender distribution of 3,534 metagenomes collected from healthy people.* ***b.*** *Boxplots illustrating the distribution of MASH distances between samples within each BioProject. The BioProject accession is used as a label, and the color gradient indicates the size, i.e., the number*

498 *of samples in each.* ***c****. Average linkage hierarchical clustering of 72 BioProjects containing healthy samples. BioProjects*
499 *containing samples from different continents are presented separately. Labels indicate the continent of origin: EU – Europe,*
500 *AS – Asia, NA – North America, AS – Australia, AF – Africa, SA – South America, and P stands for Primates. Except for the*
501 *single primate BioProject (BioSample), each BioProject is listed in colored font according to the continent from which it*
502 *originates. No severe clustering of samples based on origin is detected.*

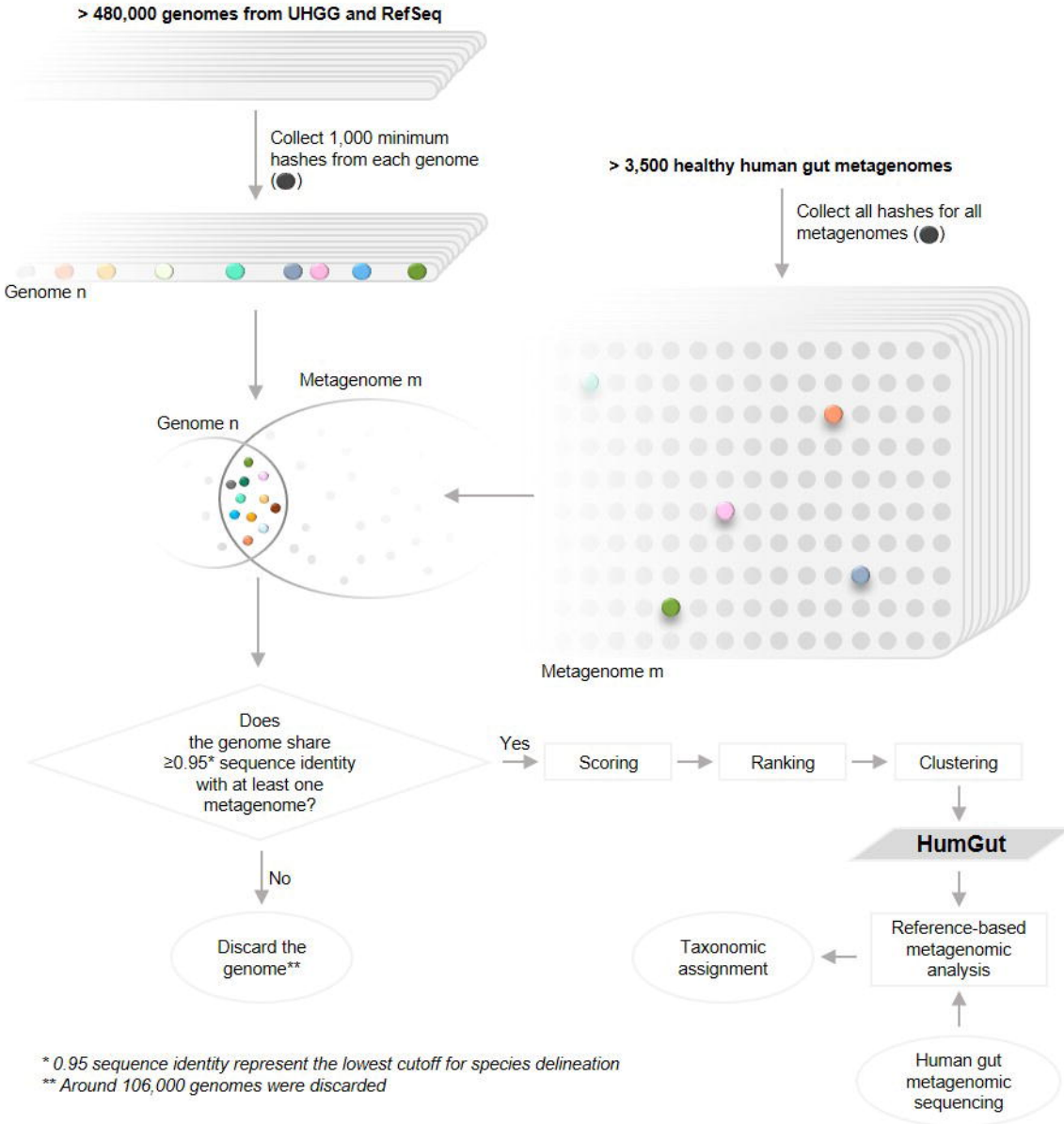
503 *Figure 3. An overview of the genomes used to build HumGut.* ***a****. The pie charts show the proportion of genomes from each*
504 *collection (UHGG above, RefSeq below) included in HumGut. To qualify for HumGut inclusion, genomes had to have at least*
505 *0.95 MASH screen identity with at least one healthy metagenome, as did most of the UHGG and half of the RefSeq*
506 *genomes. Histograms show the distribution of the mean identity shared between the qualified genomes and healthy*
507 *metagenomes. A high average identity means that the qualified genome has been found contained in most of the screened*
508 *samples* ***b****. The genome sources for HumGut clusters. The upper pie chart shows data for 30,691 clusters belonging to*
509 *HumGut_97.5 (genomes grouped based on 97.5% genome sequence identity), the bottom one presents data for 5,170*
510 *HumGut_95 clusters (95% sequence identity – species level threshold). The majority of clusters in both HumGut collections*
511 *are comprised of only UHGG genomes, while 6% and 9% of the clusters consist of only RefSeq genomes (HumGut_97.5, and*
512 *HumGut_95, respectively).*

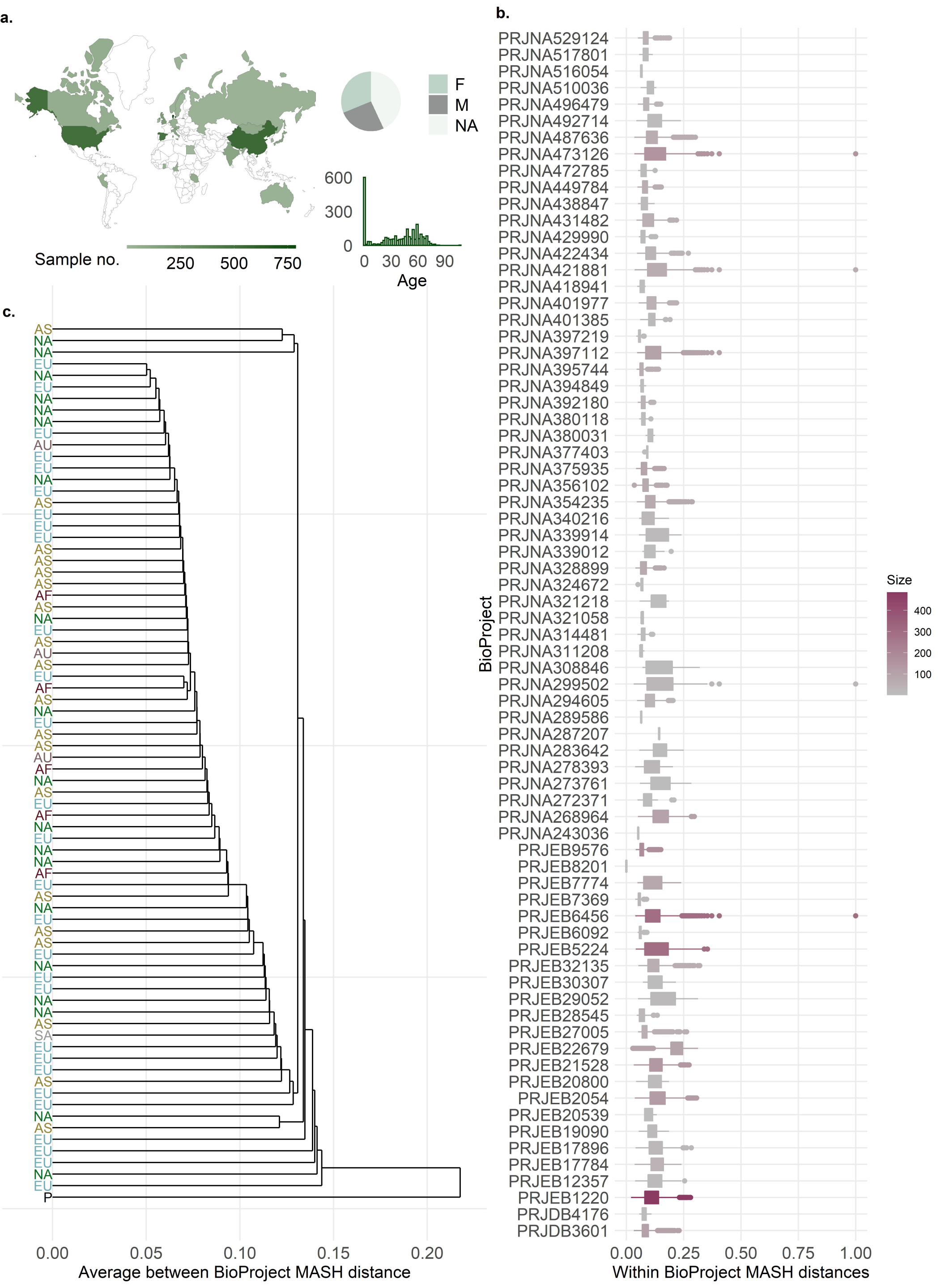
513 *Figure 4. The performance of HumGut versions in comparison to the standard Kraken2 database and UHGG collection.* ***a****.*
514 *The boxplot shows the distribution of unclassified reads for the 3,534 analyzed healthy reference metagenome samples. The*
515 *dashed line represents the k-mer database sizes (right y-axis). Every database version includes standard human genome*
516 *sequences, in addition to database-specific (sub)sets of bacteria and archaea, and the difference in size is only due to*
517 *differences in the latter.* ***b****. The classification of an additional 963 human gut metagenomes, not part of the reference set.*

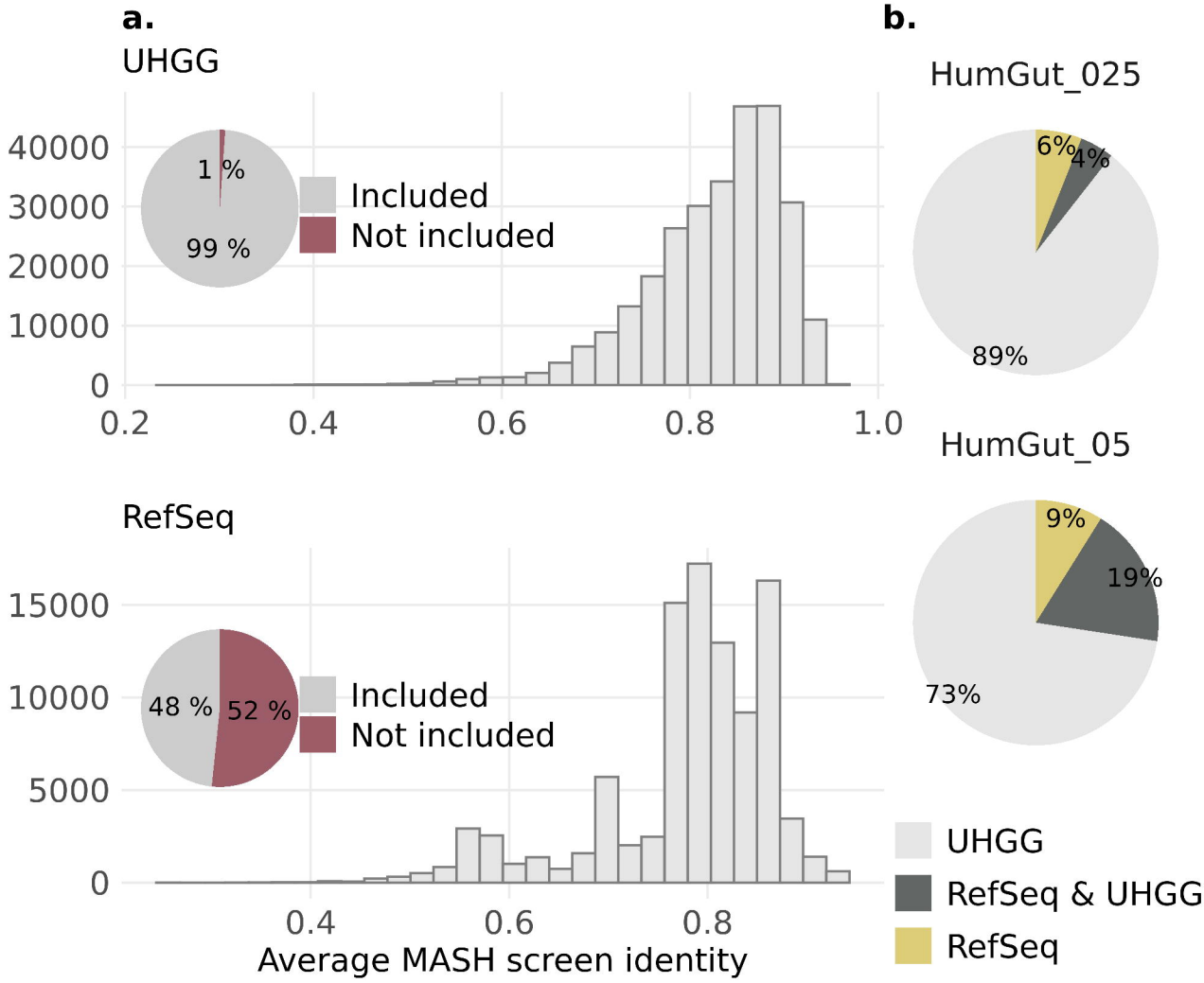
518 *Figure 5. Taxa analysis* ***a****. Species number distribution between three cohorts: healthy infants, healthy adults, and IBD*
519 *samples.* ***b****. Heatplot listing 52 bacterial species found in > 70% healthy adult samples. The tiles are colored based on the*
520 *average abundance of species in samples from the same continent (Y-axis). On top of the graph, a dendrogram shows the*
521 *complete linkage between samples from separate continents.* ***c****. PCA score plot and loading plot. Each dot on the PCA*
522 *ordination plot represents a metagenome sample. They are colored based on continent of origin. African and South*
523 *American samples separate more from samples coming from Asia, Australia, Europe, and North America. The superimposed*
524 *text represents some of the species (loadings) with the strongest effect along the PC2 axis.*

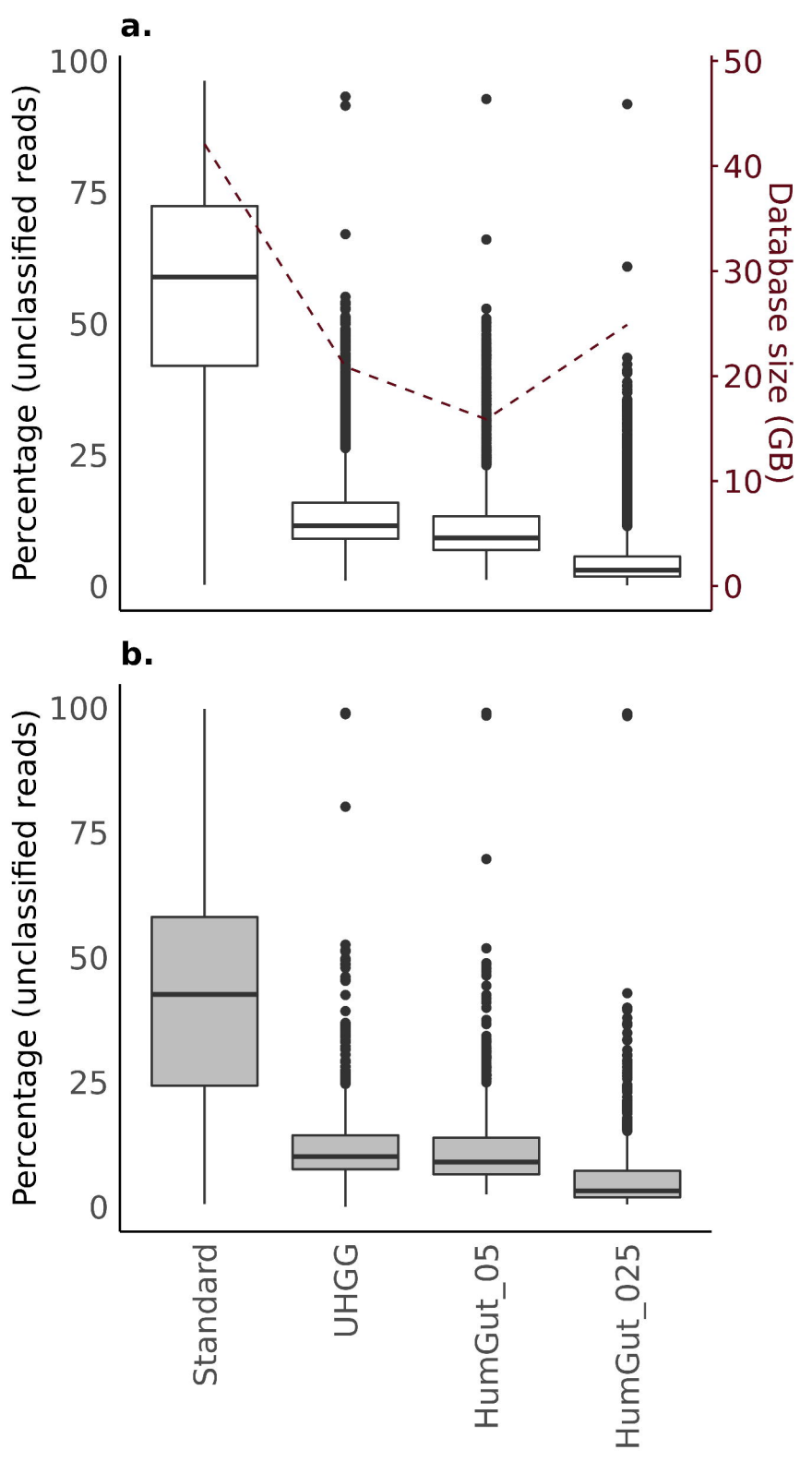
525 *Figure S1. Rarefaction curves for healthy (left panel) and IBD metagenomes (right panel), showing that the number of new*
526 *expected genomes flattens after screening ca. 1,000 metagenomes.*

527 *Figure S2. MASH and fastANI distances.* ***a.*** *A plot of ca. 20,000 genome distances computed with both fastANI (x-axis) and*
528 *MASH (y-axis). fastANI distances tend to be a little smaller than MASH distances, they however have a substantial variance.*
529 ***b.*** *The rationale behind using 0.08, and 0.1 MASH distance thresholds (vertical dashed lines) for HumGut clustering*
530 *algorithm. The vast majority of fastANI distances < 0.025 have a MASH distance < 0.08 and genomes with fastANI < 0.05*
531 *have a MASH distance < 0.1. When clustering, the distance between all genomes was first computed using MASH, then only*
532 *genomes with distances below the abovementioned thresholds were included to speed up fastANI computations.*

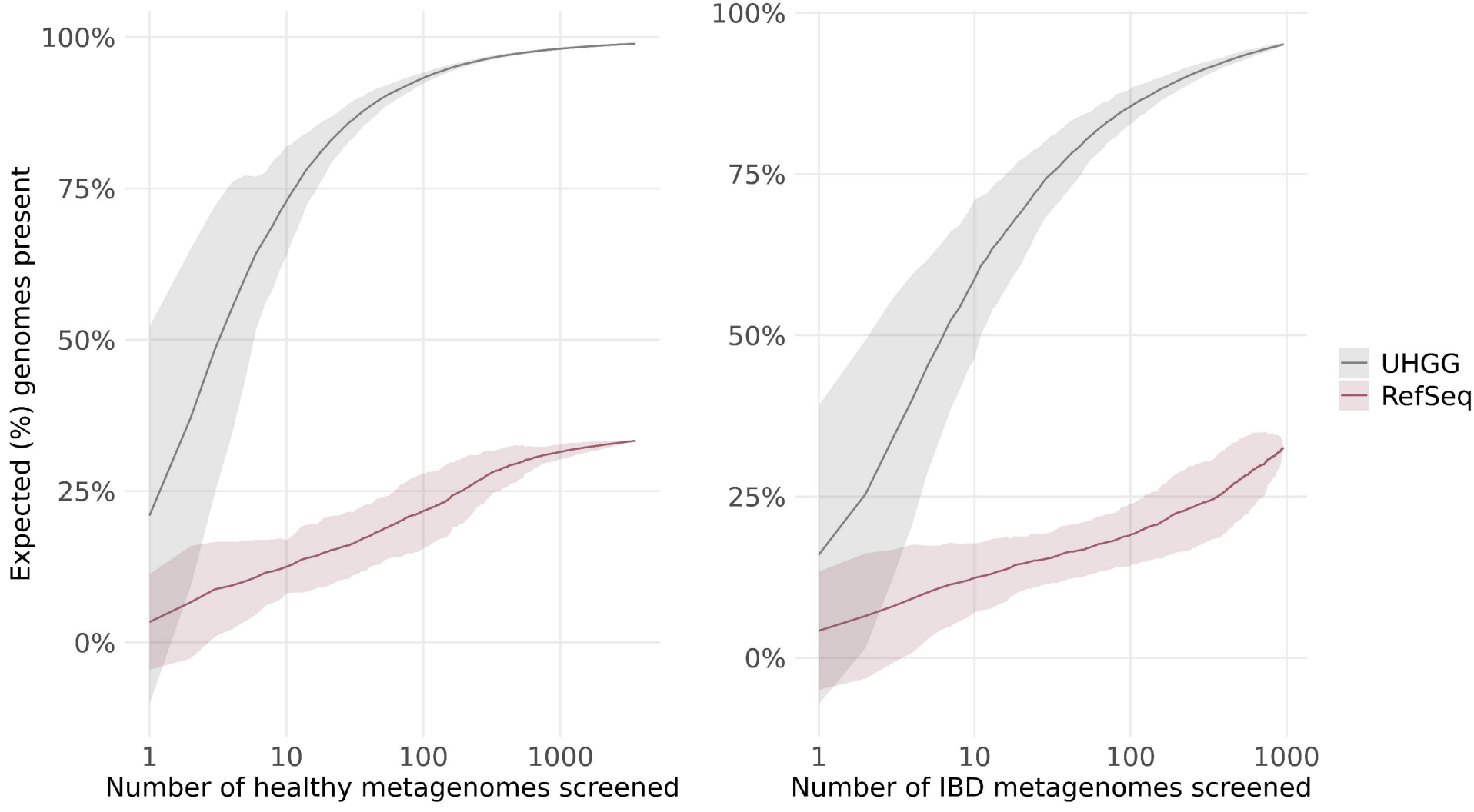

> 480,000 genomes from UHGG and RefSeq
Collect 1,000 minimum hashes from each genome (●)
Genome n
> 3,500 healthy human gut metagenomes
Collect all hashes for all metagenomes (●)
Metagenome m
Genome n
Metagenome m
Does the genome share ≥0.95* sequence identity with at least one metagenome?
Yes
Scoring
Ranking
Clustering
HumGut
No
Discard the genome**
Taxonomic assignment
Reference-based metagenomic analysis
Human gut metagenomic sequencing
* 0.95 sequence identity represent the lowest cutoff for species delineation
** Around 106,000 genomes were discarded

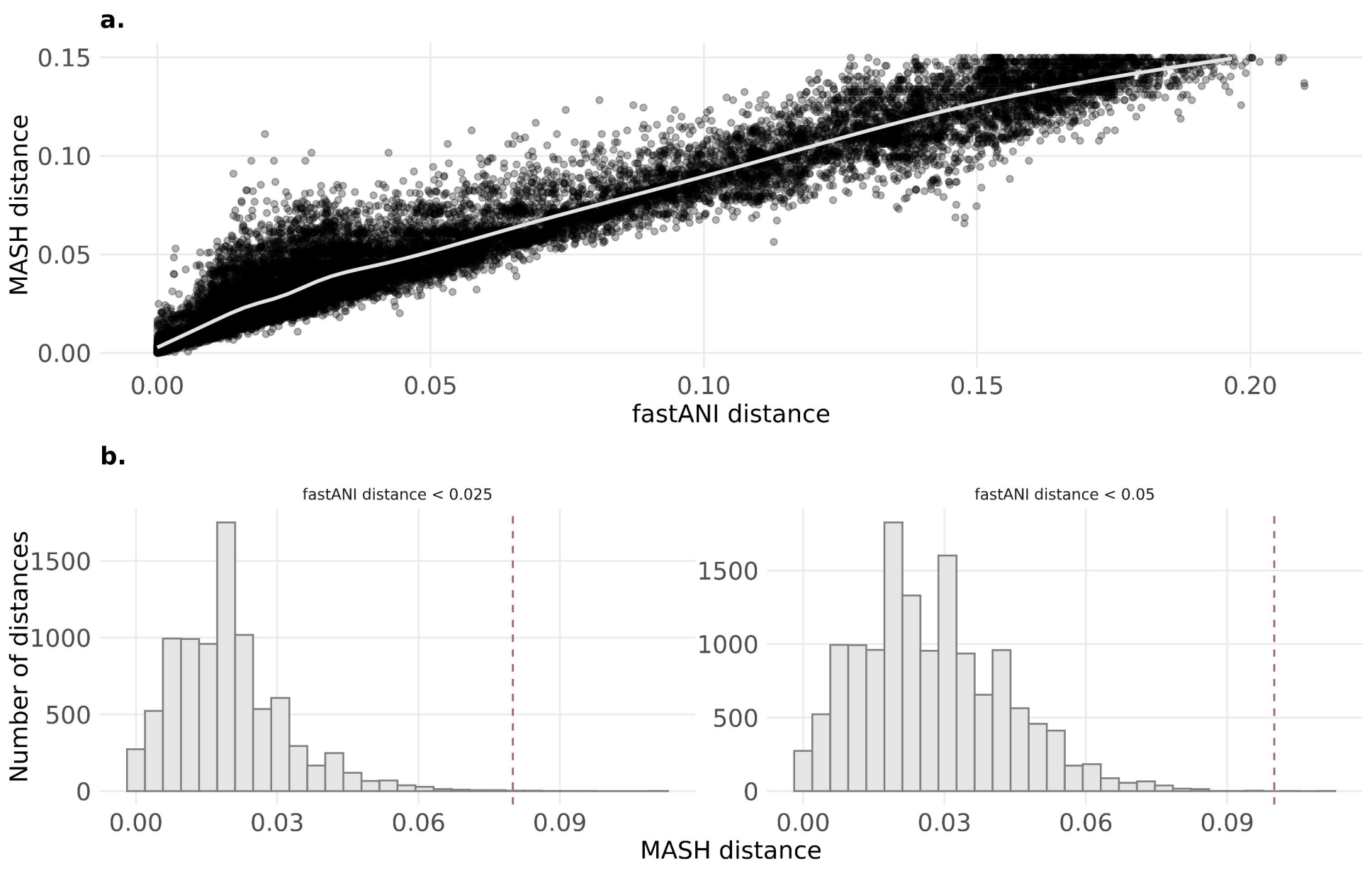

a.
F
M
NA
Sample no.
250
500
750
600
300
0
0
30
60
90
Age
b.
BioProject
Size
400
300
200
100
0.00
0.25
0.50
0.75
1.00
Within BioProject MASH distances
c.
0.00
0.05
0.10
0.15
0.20
Average between BioProject MASH distance


a.
UHGG
40000
30000
20000
10000
0
0.2
0.4
0.6
0.8
1.0
1 %
99 %
Included
Not included
RefSeq
15000
10000
5000
0
0.4
0.6
0.8
48 %
52 %
Included
Not included
Average MASH screen identity
b.
HumGut_025
6%
4%
89%
HumGut_05
9%
19%
73%
UHGG
RefSeq & UHGG
RefSeq


a.
Percentage (unclassified reads)
100
75
50
25
0
Database size (GB)
50
40
30
20
10
0
b.
Percentage (unclassified reads)
100
75
50
25
0
Standard
UHGG
HumGut_05
HumGut_025

a.
Count
Species number
Healthy
IBD
Infant
b.
Average abundance (%)
Agathobacter rectalis
Bacteroides uniformis
Phocaeicola vulgatus
Ruminococcus_E bromii_B
Alistipes putredinis
Faecalibacterium prausnitzii
Phocaeicola dorei
Faecalibacterium prausnitzii_C
Gemmiger qucibialis
Parabacteroides distasonis
Phocaeicola plebeius_A
Bacteroides ovatus
Faecalibacterium prausnitzii_G
Alistipes onderdonkii
Gemmiger formicilis
ER4 sp000765235
Mediterraneibacter faecis
Fusicatenibacter saccharivorans
Faecalibacterium prausnitzii_D
Bacteroides xylanisolvens
Anaerostipes hadrus
Acetatifactor sp900066365
Agathobacter faecis
Bacteroides thetaiotaomicron
Enterocloster sp000431375
Blautia_A wexlerae
Barnesiella intestinihominis
Odoribacter splanchnicus
Acetatifactor sp900066565
CAG-110 sp000434635
Blautia_A sp900066165
Bariatricus comes
Ruthenibacterium lactatiformans
Blautia_A obeum
KLE1615 sp900066985
Agathobaculum butyriciproducens
Lawsonibacter asaccharolyticus
Eisenbergiella sp900066775
Dorea formicigenerans
CAG-41 sp900066215
Roseburia inulinivorans
Alistipes shahii
CAG-83 sp000431575
Roseburia hominis
Bilophila wadsworthia
Clostridium_Q sp003024715
Erysipelatoclostridium sp000752095
Eubacterium_F sp003491505
Coprococcus_A catus
CAG-81 sp900066535
UBA7182 sp003480725
Enterocloster sp900541315
South America
Africa
Australia
North America
Asia
Europe
c.
PC2 (7%)
PC1 (13%)
Mediterraneibacter torques
Bacteroides ovatus
Lawsonibacter asaccharolytic
Alistipes onderdonkii
Bacteroides cellulosilyticus
Alistipes finegoldii
Alistipes obe
Bifidobacterium longum
Escherichia coli
Escherichia coli_D
Escherichia coli_C
CAG-110 sp900540635
CAG-110 sp900547405
Prevotella sp000434975
Prevotella copri_A
Africa, South America
Asia, Australia, Europe, North America
Infant
Not infant


Expected (%) genomes present
100%
75%
50%
25%
0%
1
10
100
1000
Number of healthy metagenomes screened
Number of IBD metagenomes screened
UHGG
RefSeq


a.
MASH distance
0.15
0.10
0.05
0.00
0.00
0.05
0.10
0.15
0.20
fastANI distance
b.
fastANI distance < 0.025
fastANI distance < 0.05
Number of distances
1500
1000
500
0
0.00
0.03
0.06
0.09
MASH distance