A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome

Alexandre Almeida[1,2,*], Stephen Nayfach[3,4], Miguel Boland[1], Francesco Strozzi[5], Martin Beracochea[1], Zhou Jason Shi[6,7], Katherine S. Pollard[6,7,8,9,10,11], Donovan H. Parks[12], Philip Hugenholtz[12], Nicola Segata[13], Nikos C. Kyrpides[3,4] and Robert D. Finn[1,*]

[1]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, UK; [2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK; [3]U. S. Department of Energy Joint Genome Institute, Walnut Creek, California, USA; [4]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA; [5]Enterome Bioscience, Paris, France; [6]Gladstone Institutes, San Francisco, CA, USA; [7]Chan-Zuckerberg Biohub, San Francisco, CA, USA; [8]Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA; [9]Institute for Computational Health Sciences, University of California San Francisco, San Francisco, CA, USA; [10]Quantitative Biology Institute, University of California San Francisco, San Francisco, CA, USA; [11]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA; [12]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Queensland, Australia; [13]CIBIO Department, University of Trento, Trento, Italy.

*Corresponding authors

Alexandre Almeida: aalmeida@ebi.ac.uk

Robert D. Finn: rdf@ebi.ac.uk

Abstract

Comprehensive reference data is essential for accurate taxonomic and functional characterization of the human gut microbiome. Here we present the Unified Human Gastrointestinal Genome (UHGG) collection, a resource combining 286,997 genomes representing 4,644 prokaryotic species from the human gut. These genomes contain over 625 million protein sequences used to generate the Unified Human Gastrointestinal Protein (UHGP) catalogue, a collection that more than doubles the number of gut protein clusters over the Integrated Gene Catalogue. We find that a large portion of the human gut microbiome remains to be fully explored, with over 70% of the UHGG species lacking cultured representatives, and 40% of the UHGP missing meaningful functional annotations. Intra-species genomic variation analyses revealed a large reservoir of accessory genes and single-nucleotide variants, many of which were specific to individual human populations. These freely available genomic resources should greatly facilitate investigations into the human gut microbiome.

Main

The human gut microbiome has been implicated in important phenotypes related to human health and disease[1,2]. However, incomplete reference data that are missing microbial diversity[3] hamper our understanding of the roles of individual microbiome species, their interactions and functions. Hence, establishing a comprehensive collection of microbial reference genomes and genes is an important step for accurate characterization of the taxonomic and functional repertoire of the intestinal microbial ecosystem.

The Human Microbiome Project (HMP)[4] was a pioneering initiative to enrich our knowledge of human-associated microbiota diversity. Hundreds of genomes from bacterial species with no sequenced representatives were obtained as part of this project, allowing their use for the first time in reference-based metagenomic studies. The Integrated Gene Catalogue (IGC)[5] was subsequently created, combining the sequence data available from the HMP and the Metagenomics of the Human Intestinal Tract (MetaHIT)[6] consortium. This gene catalogue has been applied successfully to the study of microbiome associations in different clinical contexts[7], revealing microbial composition signatures linked to type 2 diabetes[8], obesity[9] and other diseases[10]. But, as the IGC comprises genes with no direct link to their genome of origin, it lacks contextual data to perform a high-resolution taxonomic classification, establish genetic linkage and deduce complete functional pathways on a genomic basis.

Culturing studies have continued to unveil new insights into the biology of our gut communities[11,12] and are essential for applications in research and biotechnology. However, the advent of high-throughput sequencing and new metagenomic analysis methods — namely involving genome assembly and binning — has transformed our understanding of the microbiome composition both in humans and other environments[13–15]. Metagenomic analyses

are able to capture substantial microbial diversity not easily accessible by cultivation by directly analysing the sample genetic material without the need for culturing, though biases do exist[16]. Recent studies have massively expanded the known species repertoire of the human gut, making available unprecedented numbers of new cultured and uncultured genomes[16–20]. Two culturing efforts isolated and sequenced over 500 human gut-associated bacterial genomes each[18,20], while three independent studies[16,17,19] reconstructed 60,000–150,000 microbial metagenome-assembled genomes (MAGs) from public human microbiome data, most of which belong to species lacking cultured representatives. Combining these individual efforts and establishing a unified non-redundant dataset of human gut genomes is essential for driving future microbiome studies. To accomplish this, we compiled and analysed 286,997 genomes and 625,251,941 genes from human gut microbiome datasets to generate the Unified Human Gastrointestinal Genome (UHGG) and Protein (UHGP) catalogues, the most comprehensive sequence resources of the human gut microbiome established to date.

Results

The UHGG represents over 280,000 human gut microbial genomes

We first gathered all prokaryotic isolate genomes and MAGs from the human gut microbiome (publicly available as of March 2019). We compiled the isolate genomes from the Human Gastrointestinal Bacteria Culture Collection (HBC)[18], the Culturable Genome Reference (CGR)[20], as well as cultured human gut genomes available in the NCBI[21], PATRIC[22] and IMG[23] repositories which include genomes from several other large studies[11,12,24]. In addition, we included all of the gut MAGs generated in Pasolli, et al.[19] (CIBIO), Almeida, et al.[17] (EBI) and Nayfach, et al.[16] (HGM). To standardize the genome quality across all sets, we used thresholds of >50% genome completeness and <5% contamination, combined with an estimated quality score (completeness – 5 × contamination) >50. Final numbers of genomes

matching these criteria were: 734 (HBC), 1,519 (CGR), 651 (NCBI), 7,744 (PATRIC/IMG), 137,474 (CIBIO), 87,386 (EBI) and 51,489 (HGM), resulting in a total of 286,997 genome sequences (Fig. 1a and Supplementary Table 1). Genomes were recovered in samples from a total of 31 countries across six continents (Africa, Asia, Europe, North America, South America and Oceania), but the majority originated from samples collected in China, Denmark, Spain and the United States (Fig. 1b).

To determine how many species were included in this gut reference collection, we clustered all 286,997 genomes using a multi-step distance-based approach (see 'Methods') with an average nucleotide identity (ANI) threshold of 95% over at least a 30% alignment fraction[25]. The clustering procedure resulted in a total of 4,644 inferred prokaryotic species (4,616 bacterial and 28 archaeal, Supplementary Table 2). We found the species clustering results to be highly consistent with those previously obtained[16,17,19] (Supplementary Table 3). The best quality genome from each species cluster was selected as its representative on the basis of genome completeness, contamination and assembly N50 (with isolate genomes always given preference over MAGs), and the final set was used to generate the Unified Human Gastrointestinal Genome (UHGG) catalogue (Fig. 1c). Out of the 4,644 species-level genomes, 3,207 were >90% complete (interquartile range, IQR = 87.2–98.8%) and <5% contaminated (IQR = 0.0–1.34%), with 573 of those having the 5S, 16S and 23S rRNA genes together with at least 18 of the standard tRNAs (Supplementary Fig. 1). These 573 genomes satisfy the "high quality" criteria set for MAGs by the Genomic Standards Consortium[26]. Thereafter, we classified each species representative using the Genome Taxonomy Database[27] Toolkit (Supplementary Fig. 2), a standardized taxonomic framework based on a concatenated protein phylogeny representing >140,000 public prokaryote genomes, fully resolved to the species level (see 'Methods' for details on the taxonomy nomenclature used). However, over 60% of

the gut genomes could not be assigned to an existing species, confirming the majority of the UHGG species lack representation in current reference databases.

Comparison of species recovered in individual studies

We investigated how many of the 4,644 gut species were found in the different study collections in order to determine their level of overlap and reproducibility, as well as the ratio between cultured and uncultured species (Fig. 2a). The largest intersection was between the collections of MAGs, with the same 1,081 species detected independently in the CIBIO, EBI and HGM datasets, but not in any of the cultured genome studies. By restricting the analysis to genomes recovered from 1,554 samples common to all three MAG studies, we found that 93-97% of species from each set were detected in at least one other MAG collection, and 79-86% across all three (Supplementary Fig. 3a). Similar level of species overlap was observed when comparing studies on a per-sample basis (Supplementary Fig. 3b). Further, conspecific genomes recovered from the same samples across different studies shared a median ANI and aligned fraction of 99.9% and 92.1%, respectively (Supplementary Fig. 3c). These results emphasize the reproducibility of the different assembly and binning methods used in the large-scale studies of human gut MAGs[16,17,19]. Importantly, rarefaction analysis indicates the number of uncultured species detected has not reached a saturation point, meaning additional species remain to be discovered (Fig. 2b). However, these most likely represent rarer members of the human gut microbiome, as the number of species is closer to saturating when only considering those with at least two conspecific genomes.

We also investigated the intersection between the three large culture-based datasets: the HBC, CGR and the NCBI (which contains gut genomes from the Human Microbiome Project, HMP[4]). Unlike the MAGs, the majority of cultured species were unique within a single

collection (486/698; 70%), with only 70 (10%) being common to all three collections (Supplementary Fig. 3d). This may be due to varied geographical sampling between the collections (Asia, Europe and North America) or highlight the stochastic nature of culture-based studies.

By calculating the number of genomes contained within each cultured and uncultured species, we found that species containing isolate genomes represented the largest clusters, while those exclusively encompassing MAGs tended to be the rarest, as discussed previously[16,17,19]. For example, only two of the 25 largest bacterial clusters were exclusively represented by MAGs (Fig. 2c), with 1,212 uncultured species represented by a single genome (80% of which originated from samples only analysed in one of the MAG studies; Supplementary Fig. 4). The bacterial species most represented in our collection were *Agathobacter rectalis* (recently reclassified from *Eubacterium rectale*[28]), Escherichia coli D and *Bacteroides uniformis* (Fig. 2c, Supplementary Fig. 5 and Supplementary Table 2), whereas the most frequently recovered archaeal species was Methanobrevibacter A smithii, with 608 genomes found across all six continents (Supplementary Fig. 6). The largest species clusters displayed similarly high levels of geographical distribution, indicating the most highly represented species were not restricted to individual locations (Fig. 2c and Supplementary Fig. 5b).

Most gut microbial species still lack isolate genomes

We found that 3,750 (81%) of the species in the UHGG did not have a representative in any of the human gut culture databases. To extend the search to isolate genomes from other environments or lacking information on their isolation source, we compared the UHGG catalogue to all NCBI RefSeq isolate genomes. We identified an additional set of 438 species

164 closely matching cultured genomes, leaving 3,312 (71%) of UHGG species as uncultured
165 (Supplementary Table 2).
166
167 The phylogenetic distribution of the 4,616 bacterial (Fig. 3a) and the 28 archaeal species
168 (Supplementary Fig. 6) revealed that uncultured species exclusively represented 66% and 31%
169 of the phylogenetic diversity of Bacteria and Archaea, respectively, with several phyla lacking
170 cultured representatives (Fig. 3b). The four largest monophyletic groups lacking cultured
171 genomes were the 4C28d-15 order (167 species, recently proposed as the novel order
172 Comantemales ord. nov.[29]; Fig. 3c), order RF39 (139 species), family CAG-272 (88 species),
173 and order Gastranaerophilales (67 species). While none have been successfully cultured,
174 several have been described in the literature, including RF39[16] and Gastranaerophilales
175 (previously classified as a lineage in the Melainabacteria[30]) which are characterized by highly
176 reduced genomes with numerous auxotrophies. This analysis suggests that, despite recent
177 culture-based studies[11,12,18,20], much of the diversity in the gut microbiome remains uncultured,
178 including several large and prevalent clades.
179
180 **The UHGP expands the protein universe in the human gut microbiome**
181 Metagenomic approaches have the ability to leverage gene content information not only for
182 more precise taxonomic analysis, but to also predict the functional capacity of individual
183 species of interest compared to marker gene-based methods (e.g. relying solely on the 16S
184 rRNA gene or a limited number of diagnostic genes). We built the Unified Human
185 Gastrointestinal Protein (UHGP) catalogue with a total of 625,251,941 full-length protein
186 sequences predicted from the 286,997 genomes here analysed. These were clustered at 50%
187 (UHGP-50), 90% (UHGP-90), 95% (UHGP-90) and 100% (UHGP-100) amino acid identity,
188 generating between 5 to 171 million protein clusters (Fig. 1c and Fig. 4a).

To determine how comprehensive the UHGP was when compared to existing human gut gene catalogues, we combined the UHGP-90 (n = 13,910,025 protein clusters) together with the Integrated Gene Catalogue[5], a collection of 9.9 million genes from 1,267 gut metagenome assemblies, which we grouped into 7,063,981 protein clusters at 90% protein identity (referred to as IGC-90). Nearly all samples used to generate the IGC were also included in the UHGP catalogue (except for 59 transcriptome datasets), but the latter was generated from a larger and more geographically diverse metagenomic dataset (including samples from Africa, South America and Oceania). The UHGP-90 and IGC-90 resulted in a combined set of 15.2 million protein clusters, with an overlap of 5.8 million sequences (Fig. 4b). This revealed that 81% of the IGC is represented in the UHGP catalogue, with the missing 19% likely representing fragments of prokaryotic genomes <50% complete, viral or eukaryotic sequences, plasmids or other sequences not binned into MAGs. Most notably though, the UHGP provided an increase of 115% coverage of the gut microbiome protein space over the IGC. As the UHGP was generated from individual genomes and not from their original unbinned metagenome assemblies, our catalogue also has the advantage of providing a direct link between each gene cluster and its genome of origin. This ultimately allows combining individual genes with their genomic context for an integrated study of the gut microbiome.

Functional capacity of the human gut microbiota

We used the eggNOG[31], InterPro[32], COG[33] and KEGG[34] annotation schemes to capture the full breadth of functions within the UHGP. However, we found that 42.6% of the UHGP-100 was poorly characterized, as 28.1% lacked a match to any database and a further 14.5% only had a match to a COG with no known function (Fig. 4c). Based on the distribution of COG functions, the most highly represented categories were related to amino acid transport and metabolism, cell wall/membrane/envelope biogenesis and transcription.

We further leveraged the set of 625 million proteins derived from the human gut genomes to explore the functional diversity within each of the UHGG species. Protein sequences from all conspecific genomes were clustered at a 90% amino acid identity to generate a pan-genome for each species. Analysis of the functional capacity of the UHGG species pan-genomes identified a total of 363 KEGG modules encoded by at least one species (Supplementary Fig. 7a and Supplementary Table 4). Most conserved modules were related to ribosomal structure, glycolysis, inosine monophosphate biosynthesis, gluconeogenesis, and the shikimate pathway — all representing essential bacterial functions. However, we found that for certain phyla such as Myxococcota, Bdellovibrionota, Thermoplasmatota, Patescibacteria and Verrucomicrobiota, a substantial proportion of the species pan-genomes remained poorly characterized (Supplementary Fig. 7b). At the same time, species belonging to the clades Fibrobacterota, Bacteroidota, Firmicutes I, Verrucomicrobiota and Patescibacteria had the highest proportion of genes encoding carbohydrate-active enzymes (CAZy; Supplementary Fig. 7b). As most of these lineages are largely represented by uncultured species (Fig. 3b), this suggests the gut microbiota may harbour many species with important metabolic activities yet to be cultured and functionally characterized under laboratory conditions.

Patterns of intra-species genomic diversity

With the protein annotations and pan-genomes inferred for each of the UHGG species, we explored their intra-species core and accessory gene repertoire. Only near-complete genomes (≥90% completeness) and species with at least 10 independent conspecific genomes were analysed. The overall pattern of gene frequency within each of the 781 species here considered showed a distinctive bimodal distribution (Supplementary Fig. 8), with most genes classified as either core or rare (i.e. present in ≥90% or <10% of conspecific genomes). We analysed the pan-genome size per species in relation to the number of conspecific genomes to look for

239 differences in intra-species gene richness. We observed distinct patterns across different gut
240 phyla, with species from various Firmicutes clades showing the highest rates of gene gain (Fig.
241 5a). There was a wide variation in the proportion of core genes between species even among
242 clades with more than 1,000 genomes (Fig. 5b), with a median core genome proportion
243 (percentage of core genes out of all genes in the representative genome) estimated at 66% (IQR
244 = 59.6–73.9%).
245
246 To distinguish the functions encoded in the core and accessory genes, we analysed their
247 associated annotations. Core genes were well covered, with a median of 96%, 94%, 92% and
248 69% of the genes assigned with an eggNOG, InterPro, COG and KEGG annotation,
249 respectively (Fig. 5c). However, the accessory genes had a significantly higher proportion of
250 unknown functions (P <0.001), with a median of 21% of the genes (IQR = 16.7–27.3%) lacking
251 a match in any of the databases considered. Thereafter, we investigated the functions encoded
252 by the core and accessory genes on the basis of the COG functional categories. Genes classified
253 as core were significantly associated (adjusted P <0.001) with key metabolic functions
254 involved in nucleotide, amino acid and lipid metabolism, as well as other housekeeping
255 functions (e.g. related to translation and ribosomal structure, Fig. 5d). In contrast, accessory
256 genes had a much greater proportion of COGs without a known function, and of genes involved
257 in replication and recombination which are typically found in mobile genetic elements (MGEs,
258 Fig. 5d). A significant number of accessory genes were related to defence mechanisms, which
259 encompass not only general mechanisms of antimicrobial resistance (AMR) such as ABC
260 transporter efflux pumps, but also targeted systems towards invading MGEs (e.g. CRISPR-Cas
261 and restriction modification systems against bacteriophages). These results highlight the
262 potential of this resource to better understand the dynamics of chromosomally encoded AMR

263 within the gut and decipher to what extent the microbiome may be a source of both known and
264 novel resistance mechanisms.
265
266 We next investigated intra-species single nucleotide variants (SNVs) within the UHGG
267 species. We generated a catalogue consisting of 249,435,699 SNVs from 2,489 species with
268 three or more conspecific genomes (Fig. 6a). For context, a previously published catalogue
269 contained 10.3 million single nucleotide polymorphisms from 101 gut microbiome species[35].
270 Of note, more than 85% of these SNVs were exclusively detected in MAGs, whereas only 2.2%
271 were exclusive to isolate genomes (Fig. 6b). We found the overall pairwise SNV density
272 between MAGs to be higher than that observed between isolate genomes (Fig. 6c). Next, we
273 assigned the detected SNVs to the continent of origin of each genome and observed that 36%
274 of the SNVs were continent exclusive. Notably, genomes with a European origin contributed
275 to the most exclusive SNVs (Fig. 6d). However, genomes from Africa contributed over three
276 times more variation on average than European or North American genomes. Pairwise SNV
277 analysis also supported a higher cross-continent SNV density, especially between genomes
278 from Africa and Europe (Fig. 6e). Our results suggest there is a high strain variability between
279 continents and that a considerable level of diversity remains to be discovered, especially from
280 underrepresented regions such as Africa, South America and Oceania.
281
282 **Resource implementation**
283 Both the UHGG and UHGP catalogues are available as part of a new genome layer within the
284 MGnify[36] website, where summary statistics of each species cluster and their functional
285 annotations can be interactively explored and downloaded (see 'Data availability' section for
286 more details). We have also generated a BItsliced Genomic Signature Index (BIGSI)[37] of the
287 UHGG, which will allow users to interactively screen for the presence of small sequence

fragments (<5 kb) in this collection. As new genomes from the human gut microbiome are generated and made publicly available, we plan to periodically update the resource with newly discovered species or by replacing uncultured reference genomes with better quality versions.

Discussion

We have generated a unified sequence catalogue representing 286,997 genomes and over 625 million protein sequences of the human gut microbiome. Of the 4,644 species contained in the UHGG, 71% lack a cultured representative, meaning the majority of microbial diversity in the catalogue remains to be experimentally characterized. During preparation of our manuscript, a new collection of almost 4,000 cultured genomes from 106 gut species was released[38], which will be incorporated in future versions of the resource. As 96% of these genomes were reported to have a species representative in the culture collections here included, we do not anticipate this dataset to provide a substantial increase in the number of species discovered. Nevertheless, our analyses suggest additional uncultured species from the human gut microbiome are yet to be discovered, highlighting the importance and need for culture-based studies. Furthermore, given the sampling bias towards populations from China, Europe and the United States, we expect that many underrepresented regions still contain substantial uncultured diversity.

By comparing recently published large datasets of uncultured genomes[16,17,19], we were able to assess the reproducibility of the results from each study. We show that despite the different assembly, binning and refinement procedures employed in the three studies, almost all of the same species and near-identical strains were recovered independently when using a consistent sample set. These results further increase confidence in the use of metagenome-assembled genomes for the characterization of uncultured microbial diversity.

With the establishment of this massive sequence catalogue, it is evident that a large portion of the species and functional diversity within the human gut microbiome remains uncharacterized. Moreover, our knowledge of the intra-species diversity of many species is still limited due to the presence of a small number of conspecific genomes. Having this combined resource can help guide future studies and prioritize targets for further experimental validation. Using the UHGG or UHGP, the community can now screen for the prevalence and abundance of species/genes in a large panel of intestinal samples and in specific clinical contexts. By pinpointing particular taxonomic groups with biomedical relevance, more targeted approaches could be developed to improve our understanding of their role in the human gut. The functional predictions generated for the species pan-genomes could also be leveraged to develop new culturing strategies for isolation of candidate species. Target-enrichment methods such as single-cell[39] and/or bait-capture hybridization[40] approaches could also be applied. Being able to enrich for specific groups of interest, even without culturing, could allow recovery of better-quality versions of MAGs and improve the analysis derived from genome sequence data alone. Given the large uncultured diversity still remaining in the human gut microbiome, having a high-quality catalogue of all currently known species substantially enhances the resolution and accuracy of metagenome-based studies. Therefore, the presented genome and protein catalogue represents a key step towards a hypothesis-driven, mechanistic understanding of the human gut microbiome.

References

1. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
2. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
3. Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome research. *BMC Biol.* **17**, 48 (2019).
4. Human Microbiome Project Consortium, T. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
5. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
6. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
7. Nayfach, S., Fischbach, M. A. & Pollard, K. S. MetaQuery: a web server for rapid annotation and quantitative analysis of specific genes in the human gut microbiome. *Bioinformatics* **31**, 3368–70 (2015).
8. Wu, H. *et al.* Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat. Med.* **23**, 850–858 (2017).
9. Liu, R. *et al.* Gut microbiome and serum metabolome alterations in obesity and after weight-loss intervention. *Nat. Med.* **23**, 859–868 (2017).
10. Armour, C. R., Nayfach, S., Pollard, K. S. & Sharpton, T. J. A metagenomic meta-analysis reveals functional signatures of health and disease in the human gut microbiome. *mSystems* **4**, e00332-18 (2019).
11. Browne, H. P. *et al.* Culturing of ‘unculturable’ human microbiota reveals novel taxa

358 and extensive sporulation. *Nature* **533**, 543–546 (2016).

359 12. Lagier, J.-C. *et al.* Culture of previously uncultured members of the human gut
360 microbiota by culturomics. *Nat. Microbiol.* **1**, 16203 (2016).

361 13. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes
362 substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).

363 14. Stewart, R. D. *et al.* Compendium of 4,941 rumen metagenome-assembled genomes for
364 rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961
365 (2019).

366 15. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected
367 biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).

368 16. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. New insights from
369 uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510
370 (2019).

371 17. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**,
372 499–504 (2019).

373 18. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved
374 metagenomic analyses. *Nat. Biotechnol.* **37**, 186–192 (2019).

375 19. Pasolli, E. *et al.* Extensive unexplored human microbiome diversity revealed by over
376 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**,
377 649-662.e20 (2019).

378 20. Zou, Y. *et al.* 1,520 reference genomes from cultivated human gut bacteria enable
379 functional microbiome analyses. *Nat. Biotechnol.* **37**, 179–185 (2019).

380 21. Kitts, P. A. *et al.* Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids*
381 *Res.* **44**, D73–D80 (2016).

382 22. Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial bioinformatics database

and analysis resource center. *Nucleic Acids Res.* **45**, D535–D542 (2017).

23. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).

24. Human Microbiome Jumpstart Reference Strains Consortium, T. H. M. J. R. S. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328**, 994–9 (2010).

25. 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. *Nat. Commun.* **9**, 5114 (2018).

26. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

27. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996 (2018).

28. Rosero, J. A. *et al.* Reclassification of *Eubacterium rectale* (Hauduroy et al. 1937) Prévot 1938 in a new genus *Agathobacter* gen. nov. as *Agathobacter rectalis* comb. nov., and description of *Agathobacter ruminis* sp. nov., isolated from the rumen con. *Int. J. Syst. Evol. Microbiol.* **66**, 768–773 (2016).

29. Hildebrand, F. *et al.* Antibiotics-induced monodominance of a novel gut bacterial order. *Gut* (2019). doi:10.1136/gutjnl-2018-317715

30. Di Rienzi, S. C. *et al.* The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *Elife* **2**, (2013).

31. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

32. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

33. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–D269 (2015).

34. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

35. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).

36. Mitchell, A. L. *et al.* EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.* **46**, D726–D735 (2017).

37. Bradley, P., den Bakker, H. C., Rocha, E. P. C., McVean, G. & Iqbal, Z. Ultrafast search of all deposited bacterial and viral genomic data. *Nat. Biotechnol.* **37**, 152–159 (2019).

38. Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* 1–11 (2019). doi:10.1038/s41591-019-0559-3

39. Xu, Y. & Zhao, F. Single-cell metagenomics: challenges and applications. *Protein Cell* **9**, 501–510 (2018).

40. Noyes, N. R. *et al.* Enrichment allows identification of diverse, rare elements in metagenomic resistome-virulome sequencing. *Microbiome* **5**, 142 (2017).

41. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

433 42. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments.
434 *Bioinformatics* **25**, 1335–1337 (2009).
435 43. Kalvari, I. *et al.* Rfam 13.0: shifting to a genome-centric resource for non-coding RNA
436 families. *Nucleic Acids Res.* **46**, D335–D342 (2018).
437 44. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer
438 RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–64 (1997).
439 45. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate
440 genomic comparisons that enables improved genome recovery from metagenomes
441 through de-replication. *ISME J.* **11**, 2864–2868 (2017).
442 46. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**,
443 2825–2830 (2011).
444 47. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.*
445 **5**, R12 (2004).
446 48. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new
447 developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
448 49. Turner, I., Garimella, K. V, Iqbal, Z. & McVean, G. Integrating long-range connectivity
449 information into de Bruijn graphs. *Bioinformatics* **34**, 2556–2565 (2018).
450 50. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–
451 2069 (2014).
452 51. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
453 identification. *BMC Bioinformatics* **11**, 119 (2010).
454 52. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis.
455 *Bioinformatics* **31**, 3691–3693 (2015).
456 53. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat.*
457 *Commun.* **9**, 2542 (2018).

458 54. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology
459 assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
460 55. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The
461 carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–
462 D495 (2014).
463 56. Torchiano, M. Effsize - a package for efficient effect size computation. (2016).
464 doi:10.5281/ZENODO.1480624
465 57. Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V. & Egozcue, J. J. It's all relative: analyzing
466 microbiome data as compositions. *Ann. Epidemiol.* **26**, 322–329 (2016).
467

Methods

Genome collection

We compiled all the prokaryotic genomes publicly available as of March 2019 that have been sampled from the human gut. To retrieve isolate genomes, we surveyed the IMG[23], NCBI[21] and PATRIC[22] databases for genome sequences annotated as having been isolated from the human gastrointestinal tract. We complemented this set with bacterial genomes belonging to two recent culturomics collections: the Human Gastrointestinal Bacteria Culture Collection (HBC)[18] and the Culturable Genome Reference (CGR)[20]. To avoid including duplicated entries due to redundancy between reference databases, we combined genomes obtained from the PATRIC and IMG repositories, and added only those without an identical genome in the sets extracted from NCBI, HBC and CGR. Metagenome-assembled genomes (MAGs, i.e. uncultured genomes) were obtained from Pasolli, et al.[19] (CIBIO), Almeida, et al.[17] (EBI) and Nayfach, et al.[16] (HGM). For the CIBIO set, only those genomes retrieved from samples collected from the intestinal tract were used. Metadata for each genome was retrieved using the API of the various public repositories and combined with that available in each of the original studies.

Assessing genome quality

Genome quality (completeness and contamination) was estimated with CheckM v1.0.11[41] using the 'lineage_wf' workflow to select only those that passed the following criteria: >50% genome completeness, <5% contamination and an estimated quality score (completeness – 5 × contamination) >50. We also searched for the presence of ribosomal RNAs in each genome with the 'cmsearch' function of INFERNAL[42] (options '-Z 1000 --hmmonly --cut_ga --noali –tblout') against the Rfam[43] covariance models for the 5S, 16S and 23S rRNAs. tRNAs of the

492 standard 20 amino acids were identified with tRNAScan-SE[44] with options '-A -Q' for archaeal
493 species and '-B -Q' for those belonging to bacterial lineages.
494
495 **Species clustering**
496 We clustered the total set of 286,997 genomes at an estimated species level (average nucleotide
497 identity, ANI ≥95%[25]) using dRep v2.2.4[45] with the following options: '-pa 0.9 -sa 0.95 -nc
498 0.30 -cm larger'. Because of the computational burden of clustering together the entire genome
499 set, we employed an iterative approach where random chunks of 50,000 genomes were
500 clustered independently. The selected representatives from each chunk were combined and
501 subsequently clustered, reducing the final computational load. To ensure the best quality
502 genome was selected as the species representative in each iteration, a score was calculated for
503 each genome based on the following formula:
504 $$\text{Score} = \text{CMP} - 5 \times \text{CNT} + 0.5 \times \log(\text{N50})$$
505 where CMP represents the completeness level, CNT the estimated contamination and N50 the
506 assembly contiguity characterized by the minimum contig size in which half of the total
507 genome sequence is contained. The genome with the highest score was chosen as the species
508 representative, with cultured genomes prioritized over uncultured genomes (i.e. if a MAG had
509 a higher score than an isolate genome, the latter would still be chosen as the representative).
510
511 **Evaluating methods reproducibility**
512 The species clusters inferred here were compared with those previously generated in the human
513 gut MAG studies[16,17,19] from a common set of genomes. Similarity between species clusterings
514 was estimated using the Adjusted Rand Index (ARI) computed in the Scikit-learn python
515 package[46]. This metric considers both the number of clusters and cluster membership to
516 compute a similarity score ranging from 0 to 1.

Conspecific genomes recovered in the same metagenomic samples but in different studies were compared with FastANI v1.1[25] with default parameters to obtain both the maximum aligned fraction and ANI for each pairwise comparison.

Inferring cultured status

To determine their cultured status, the UHGG species representatives were searched against NCBI RefSeq release 93 after excluding uncultured genomes (i.e. metagenome-assembled or single-cell amplified genomes). Genome alignments were performed in two stages: (1) Mash v2.1 was used as an initial screen (using the function 'mash dist') to identify the most similar RefSeq genome to each of the UHGG species, and (2) 'dnadiff' from MUMmer v4.0.0beta2[47] was subsequently used to compute whole genome ANI between the genome pairs. A species was considered to have been cultured if (1) it contained a cultured gut genome from the UHGG catalogue, or (2) if it matched an isolate RefSeq genome with at least 95% ANI over at least 30% of the genome length.

Calculating number of conspecific genomes

For an accurate assessment of the number of non-redundant genomes belonging to each species, we de-replicated all conspecific genomes at a 99.9% ANI threshold using dRep with options '-pa 0.999 –SkipSecondary'. Furthermore, the frequency of each species was only counted once per sample to avoid cases where the same genome was recovered multiple times because of overlapping samples between the three MAG studies.

Estimating geographical diversity

A geographical diversity index was estimated to assess how widely distributed each species was. We calculated the Shannon diversity index on the proportion of samples each species was

542 found per continent. This metric combines both richness and evenness, so the level of estimated
543 diversity is highest in species found across all continents at a similar proportion.
544
545 **Phylogenetic analyses**
546 Taxonomic annotation of each species representative was performed with the Genome
547 Taxonomy Database[27] Toolkit (GTDB-Tk) v0.3.1
548 (https://github.com/Ecogenomics/GTDBTk) (database release 04-RS89) using the
549 'classify_wf' function and default parameters. To use consistent species boundaries between
550 the genome clustering and taxonomic classification procedures, genomes were assigned at the
551 species level if the ANI to the closest GTDB-Tk species representative genome was ≥95% and
552 the alignment fraction ≥30%. In this taxonomy scheme, genera and species names with an
553 alphabetic suffix indicate taxa that are polyphyletic or needed to be subdivided based on
554 taxonomic rank normalization according to the current GTDB reference tree. The lineage
555 containing the type strain retains the unsuffixed (valid) name and all other lineages are given
556 alphabetic suffixes, indicating they are placeholder names that need to be replaced in due
557 course. Taxon names above the rank of genus appended with an alphabetic suffix indicate
558 groups that are not monophyletic in the GTDB reference tree, but for which there exists
559 alternative evidence that they are monophyletic groups. We also generated NCBI taxonomy
560 annotations for each species-level genome based on its placement in the GTDB tree, using the
561 'gtdb_to_ncbi_majority_vote.py' script available in the GTDB-Tk repository
562 (https://github.com/Ecogenomics/GTDBTk/tree/stable/scripts).
563
564 Maximum-likelihood trees were generated *de novo* using the protein sequence alignments
565 produced by the GTDB-Tk: we used IQ-TREE v1.6.11 to build a phylogenetic tree of the 4,616
566 bacterial and 28 archaeal species. The best-fit model was automatically selected by

567 'ModelFinder' on the basis of the Bayesian Information Criterion (BIC) score. The LG+F+R10
568 model was chosen for building the bacterial tree, while the LG+F+R4 model was used for the
569 archaeal phylogeny. Trees were visualized and annotated with the Interactive Tree Of Life
570 (iTOL) v4.4.2[48]. Phylogenetic diversity (PD) was estimated by the sum of branch lengths, with
571 the amount that was exclusive to uncultured species calculated as $PD_{total} - PD_{cultured}$. Uncultured
572 monophyletic groups were defined as nodes in the tree containing child leaves exclusively
573 comprised of uncultured genomes.
574
575 **BIGSI construction**
576 A BItsliced Genomic Signature Index (BIGSI)[37] was generated for all species-level genomes
577 with BIGSI v0.3.8. First, *k*-mers of size 31 were extracted from each genome with McCortex
578 v1.0.1[49] ('mccortex31 build -k 31'). Thereafter, Bloom filters were built for each *k*-mer set
579 using 'bigsi bloom' and inserted into the BIGSI index with 'bigsi build'. BIGSI config
580 parameters *h* (number of hash functions applied to each *k*-mer) and *m* (Bloom filter's length in
581 bits) were set at 1 and 28,000,000, respectively. A final API layer for querying the index was
582 built using hug (http://www.hug.rest/) and hosted on the MGnify[36] website:
583 https://www.ebi.ac.uk/metagenomics/genomes.
584
585 **Pan-genome analysis and functional annotation**
586 Protein coding sequences (CDS) for each of the 286,997 genomes were predicted and annotated
587 with Prokka v1.13.3[50], using Prodigal v2.6.3[51] with options '-c' (predict proteins with closed
588 ends only), '-m' (prevent genes from being built across stretches of sequences marked as Ns)
589 and '-p single' (single mode for genome assemblies containing one single species). Pan-
590 genome analyses were carried out using Roary v3.12.0[52]. We set a minimum amino acid
591 identity for a positive match at 90% ('-i 90'), a core gene defined at 90% presence ('-cd 90')

592 and no paralog splitting ('-s'). A normalized pan-genome size was estimated by dividing the
593 total number of core and accessory genes by the number of genes contained in the species
594 representative genome.
595
596 The Unified Human Gastrointestinal Protein (UHGP) catalogue was generated from the
597 combined set of 625,251,941 CDS predicted. Protein clustering of the UHGP and the Integrated
598 Gene Catalogue (IGC)[5] was performed with the 'linclust' function of MMseqs2 v6-f5a1c[53]
599 with options: '--cov-mode 1 -c 0.8' (minimum coverage threshold of 80% length of the shortest
600 sequence) and '--kmer-per-seq 80' (number of *k*-mers selected per sequence, increased from
601 the default of 21 to improve clustering sensitivity). The '--min-seq-id' option was set at 1, 0.95,
602 0.9 and 0.5 to generate the catalogues at 100%, 95%, 90% and 50% protein identity,
603 respectively. We clustered the IGC solely at a 90% and 50% protein identity as it was originally
604 de-replicated at a 95% nucleotide identity[5]. Functional characterization of all protein sequences
605 was performed with eggNOG-mapper v2[54] (database v5.0[31]) and InterProScan v5.35-74.0[32].
606 COG[33], KEGG[34] and CAZy[55] annotations were derived from the eggNOG-mapper results.
607 Differences in annotation coverage and COG functional categories between the core and
608 accessory genes were evaluated with a two-tailed Wilcoxon rank-sum test in R v3.6.0 (function
609 'wilcox.test'). Expected *P* values were corrected for multiple testing with the Benjamini-
610 Hochberg method. Cohen's d effect sizes were estimated with the function 'cohen.d' from the
611 Effsize[56] R package. To accurately estimate the proportion of each KEGG module in the
612 species pan-genome, we used the compositional data analysis R package CoDaSeq[57]. Pseudo
613 counts for zero-count data were first imputed using a Bayesian-Multiplicative simple
614 replacement procedure implemented in the 'cmultRepl' function (method 'CZM'). Final counts
615 were thereby converted to centred log-ratios using the 'codaSeq.clr' function to account for the
616 compositional nature of the data and for differences in pan-genome size.

SNV analyses

A total of 2,489 species with at least three conspecific genomes were used for generating a catalogue of single nucleotide variants (SNVs). For each species, we mapped all conspecific genomes to the representative genome using the 'nucmer' program from MUMmer v4.0.0.beta2[47] and filtered alignments using the 'delta-filter' program with options '-q -r' to exclude chance- and repeat-induced alignments. Thereafter, we identified SNVs using the 'show-snps' program. Single base insertions and deletions were not counted as SNVs. Each SNV locus was included in the catalogue only when the alternate allele was detected in at least two conspecific genomes. The final SNV catalogue was generated by unifying the SNV coordinates on the basis of their position in the species representative genome. The SNV entries in the catalogue were characterized as genome type-specific or continent-specific based on whether the alternate allele could be found solely in genomes from a specific genome type or continent. The number of continent-specific SNVs was normalized by the number of genomes from the corresponding continent to estimate the contribution per genome to the continent-specific SNV discoveries.

Similar programs and parameters were used for the pairwise genome alignment, but in this case only near-complete genomes (≥90% completeness) and species with at least 10 independent conspecific genomes were considered. Due to the high computational demand, pairwise alignments of species encompassing more than 1,000 genomes were limited to the best-quality 1,000 genomes. A total of 29,283,684 pairwise genome alignments were performed between almost 113,000 genomes from 909 species. For each pairwise comparison, we estimated the total number of SNVs and the overall density as the number of SNVs per kb. In addition, the pairwise comparisons were organized based on the type and the continent origin of the genomes in the pair for further downstream analyses. A two-tailed Wilcoxon rank-sum test was used to

642 evaluate differences in SNV distributions. Resulting *P* values were corrected for multiple
643 testing with the Benjamini-Hochberg method.
644
645 **Data availability**
646 Genome assemblies of the UHGG have been deposited in the European Nucleotide Archive
647 under study accession ERP116715. The UHGG, UHGP and SNV catalogues are available in a
648 public FTP server (http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/)
649 alongside functional annotations and the pan-genome results. These data together with the
650 BIGSI search index of the UHGG can also be accessed interactively on the MGnify website:
651 https://www.ebi.ac.uk/metagenomics/genomes.
652

653 ## Acknowledgements
654 We would like to thank Phelim Bradley and Zamin Iqbal for their help in the BIGSI
655 implementation, and Dongying Wu for assistance in the identification of uncultured
656 monophyletic groups. Funding: European Molecular Biology Laboratory (EMBL);
657 Biotechnology and Biological Sciences Research Council [BB/N018354/1 and
658 BB/R015228/1]; European Research Council (project ERC- STG MetaPG-716575) to N.S.
659

660 ## Author contributions
661 A.A., S.N., N.K.C. and R.D.F. conceived the study. A.A. performed the genome clustering and
662 annotations, compared study sets, carried out the pan-genome analyses, built the BIGSI index
663 and drafted the manuscript. S.N. provided feedback, performed phylogenetic, rarefaction and
664 clustering analyses, as well as the comparison with RefSeq. M.Boland and M.Beracochea built
665 the resource implementation within the MGnify website. F.S. built the protein catalogue and
666 performed the comparison with the IGC. Z.J.S. generated the SNV catalogue and performed

667 related analyses according to genome type and geographic origin. K.S.P. provided feedback,
668 funding and contributed to the SNV analyses. D.H.P. and P.H. provided feedback and assisted
669 in the species taxonomic classification. N.S. provided feedback, funding and contributed to the
670 generation of the protein catalogue. N.K.C. and R.D.F. supervised the work, provided feedback
671 and funding. All authors read, edited and approved the final manuscript.
672

673 ## Competing interests

674 F.S. is an employee of Enterome SA. P.H. is a co-founder and Director of Microba Life
675 Sciences Limited. D.H.P. is a consultant to Microba Life Sciences Limited. R.D.F. is a
676 consultant to Microbiotica Pty Ltd.
677

Figures

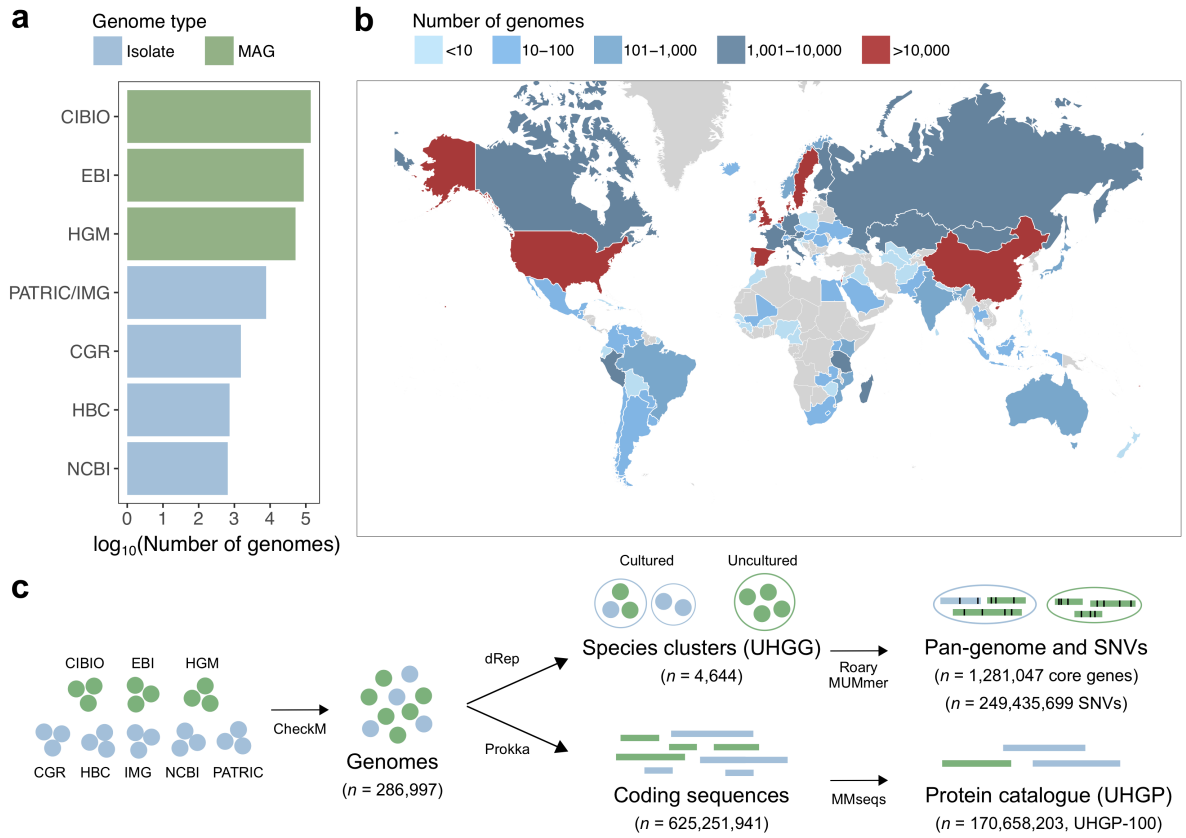


Figure 1. The unified sequence catalogue of the human gut microbiome.

a, Number of gut genomes per each study set used to generate the sequence catalogues, coloured according to whether they represent isolate genomes or metagenome-assembled genomes (MAGs). **b**, Geographic distribution of the number of genomes retrieved per country. **c**, Overview of the methods used to generate the genome (UHGG) and protein sequence (UHGP) catalogue. Genomes retrieved from public datasets were first quality-controlled by CheckM. Filtered genomes were clustered at an estimated species-level (95% average nucleotide identity) and their intra-species diversity was assessed (genes from conspecific genomes were clustered at a 90% protein identity). In parallel, a non-redundant protein catalogue was generated from all the coding sequences of the 286,997 genomes at 100% (UHGP-100, n = 170,658,203), 95% (UHGP-95, n = 20,240,320), 90% (UHGP-90, n = 13,910,025) and 50% (UHGP-50, n = 4,736,012) protein identity.

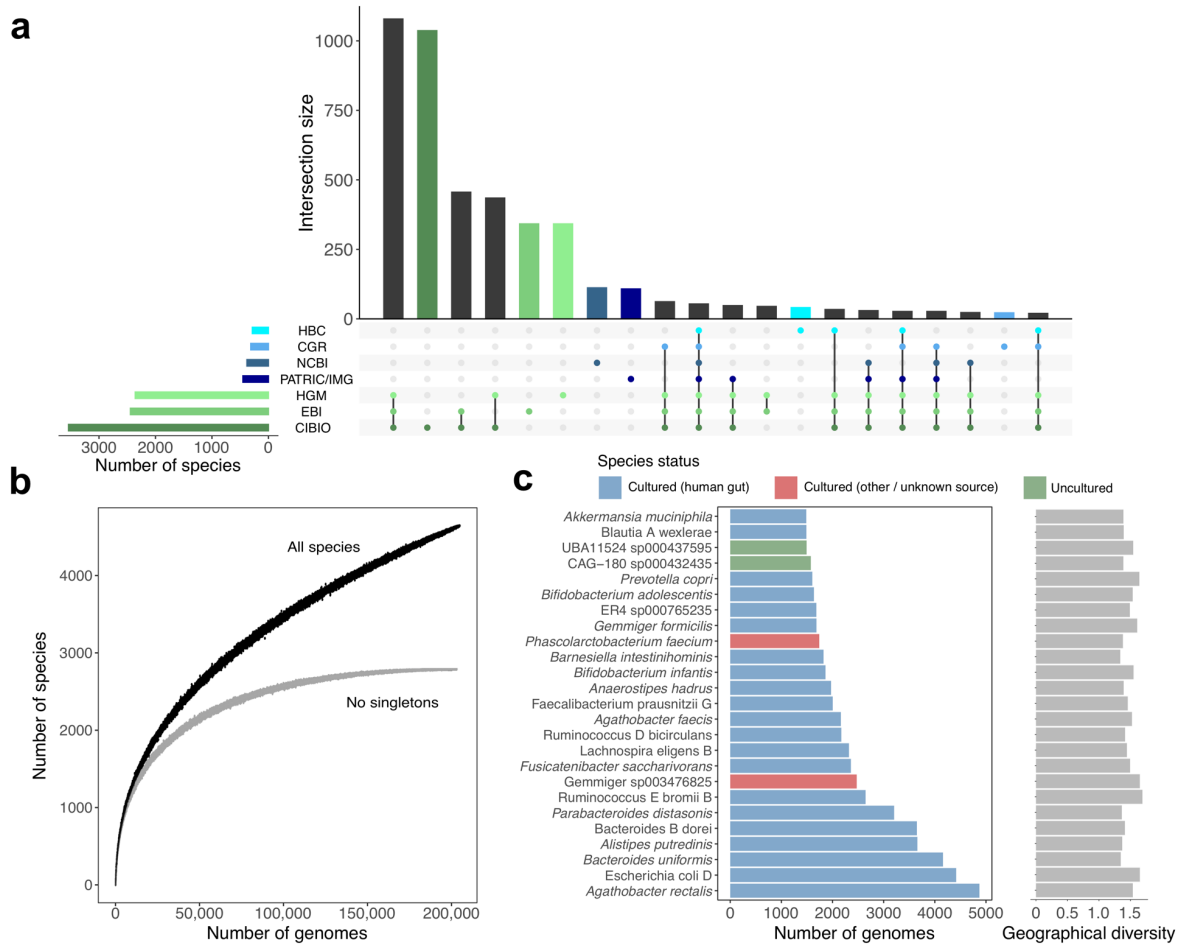


Figure 2. Intersection and frequency of species across studies.

a, Number of species found across the genome study sets here used, ordered by their level of overlap. Vertical bars represent the number of species shared between the study sets highlighted in the lower panel. **b**, Rarefaction curve of the number of species detected as a function of the number of non-redundant genomes analysed. Curves are depicted both for all the UHGG species, and after excluding singleton species (represented by only one genome). **c**, Number of non-redundant genomes detected per species (left) alongside the degree of geographical diversity (calculated with the Shannon diversity index, right).

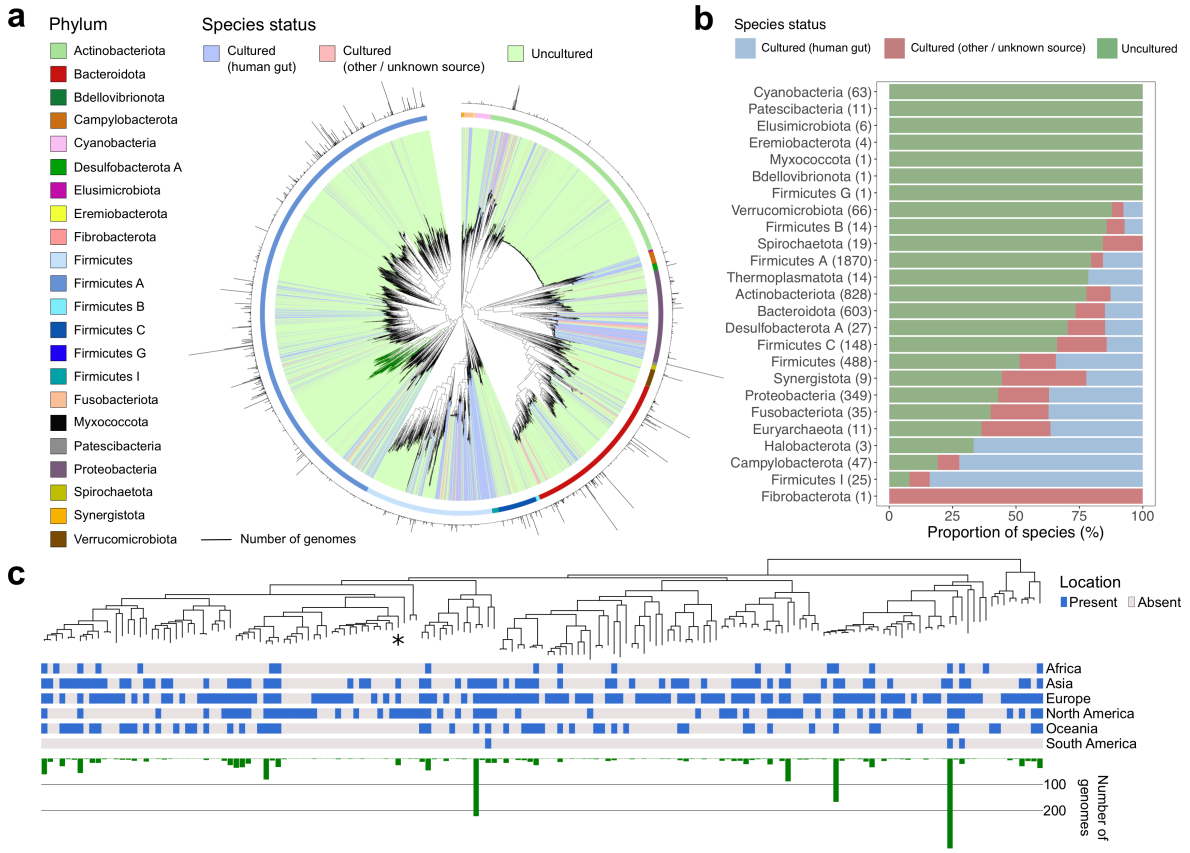


Figure 3. Uncultured species are predominant among human gut phyla.

a, Maximum-likelihood phylogenetic tree of the 4,616 bacterial species detected in the human gut. Clades are coloured by species cultured status with outer circles depicting the GTDB phylum annotation. Bar graphs in the outermost layer indicate the number of genomes from each species. The order Comantemales ord. nov. is highlighted with dark green branches. **b**, Proportion of species within the 25 prokaryotic phyla detected according to their cultured status. Numbers in brackets represent the total number of species in the corresponding phylum. **c**, Phylogenetic tree of species belonging to the order Comantemales ord. nov. (phylum Firmicutes A), the largest phylogenetic group exclusively represented by uncultured species. The geographic distribution of each species and the number of genomes recovered is represented below the tree. The species previously classified as *Candidatus* Borkfalki ceftriaxensis is indicated with an asterisk.

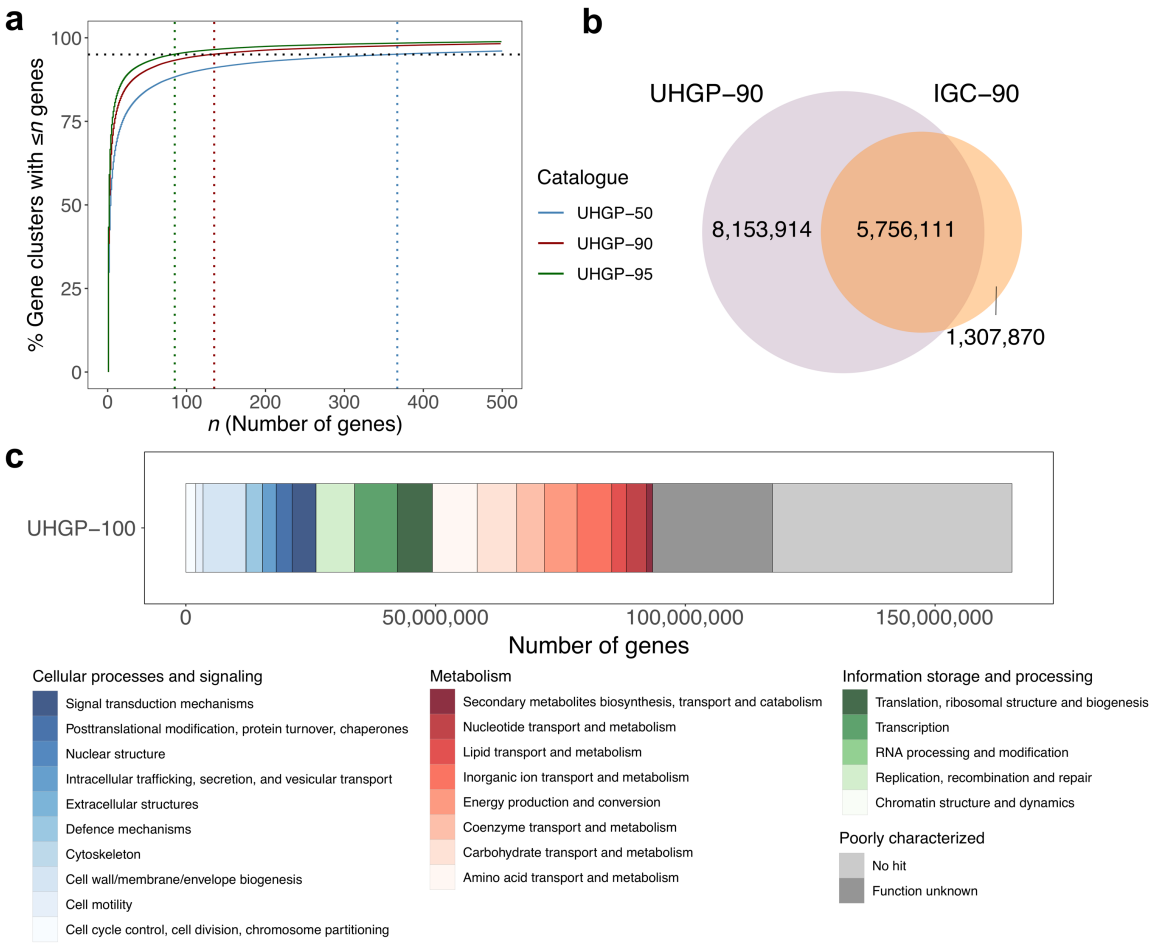


Figure 4. The UHGP improves coverage of the human gut protein landscape.

a, Cumulative distribution curve of the number and size of the gene clusters of the UHGP-95 (*n* = 20,240,320), UHGP-90 (*n* = 13,910,025) and UHGP-50 (*n* = 4,736,012). Dashed vertical lines indicate the cluster size below which 90% of the gene clusters can be found. **b,** Overlap between the UHGP (purple) and IGC (orange), both clustered at 90% amino acid identity. **c,** COG functional annotation results of the unified gastrointestinal protein catalogue clustered at 100% amino acid identity (UHGP-100).

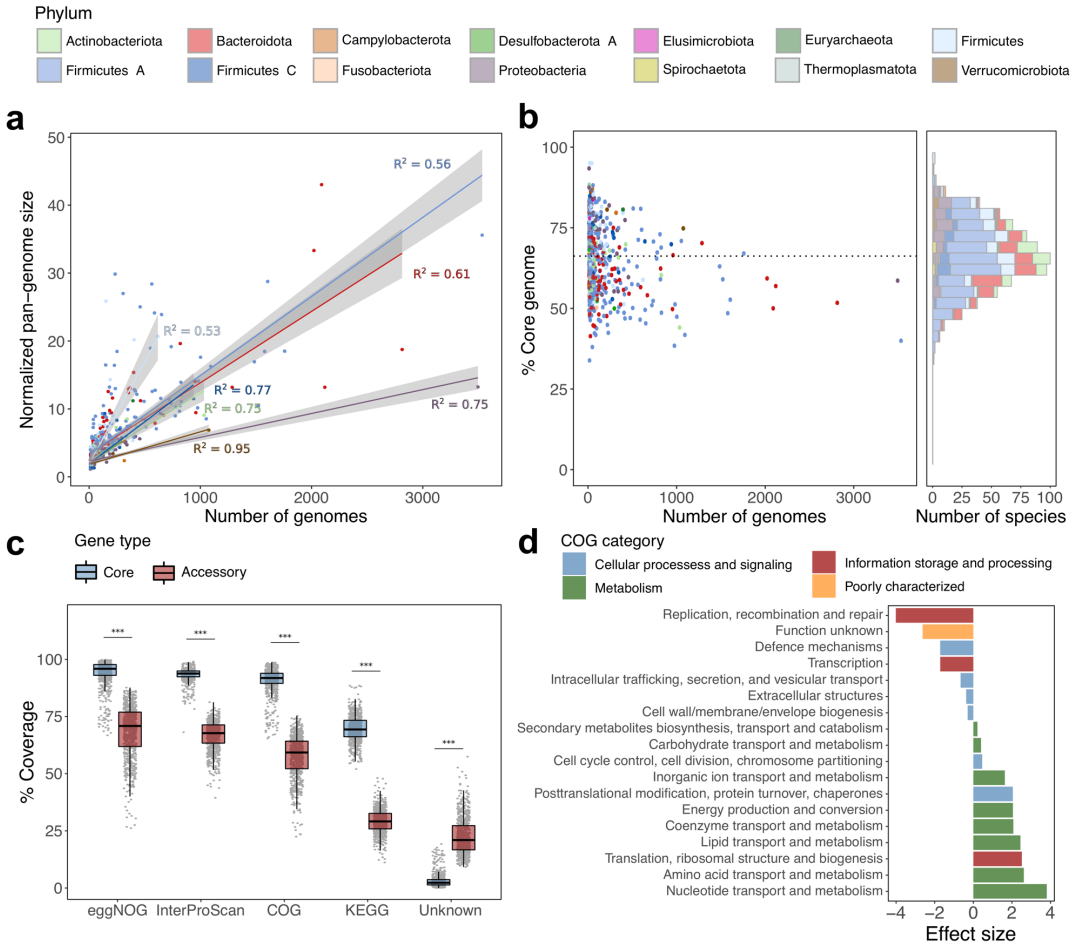


Figure 5. Pan-genome diversity patterns within the gut microbiome.

a, Normalized pan-genome size as a function of the number of conspecific genomes. Regression curves were generated per phylum, with the corresponding coefficients of determination indicated next to each curve. **b**, Fraction of each species core genome (proportion of core genes out of all genes in the representative genome) according to the number of conspecific genomes (left) and as a histogram (right), coloured by phylum. Horizontal dashed line represents the median value across all species. **c**, Proportion of core and accessory genes from each species that was classified with various annotation schemes, alongside the percentage of genes lacking any functional annotation. ***P <0.001 **d**, Comparison between the functional categories assigned to the core and accessory genes. Only those statistically significant (adjusted P <0.05) are shown. A positive effect size indicates overrepresentation in the core genes.

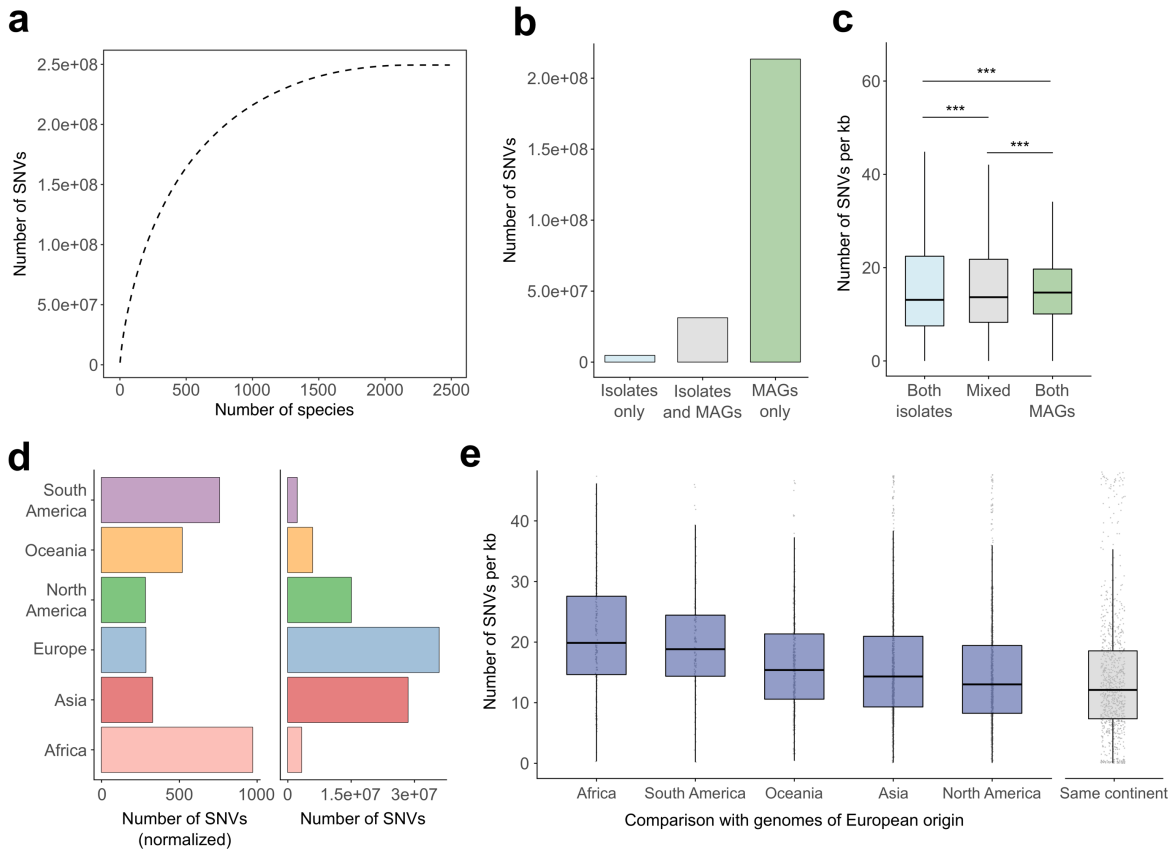
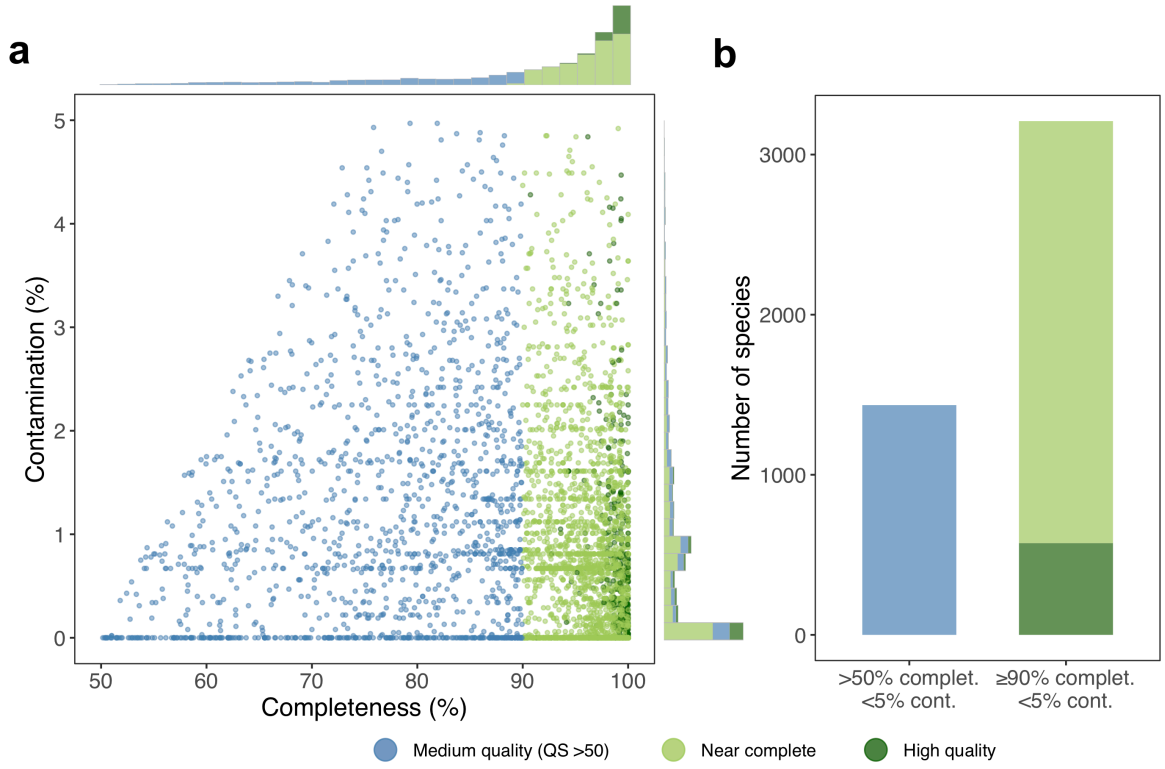


Figure 6. Analysis of intra-species single nucleotide variation.

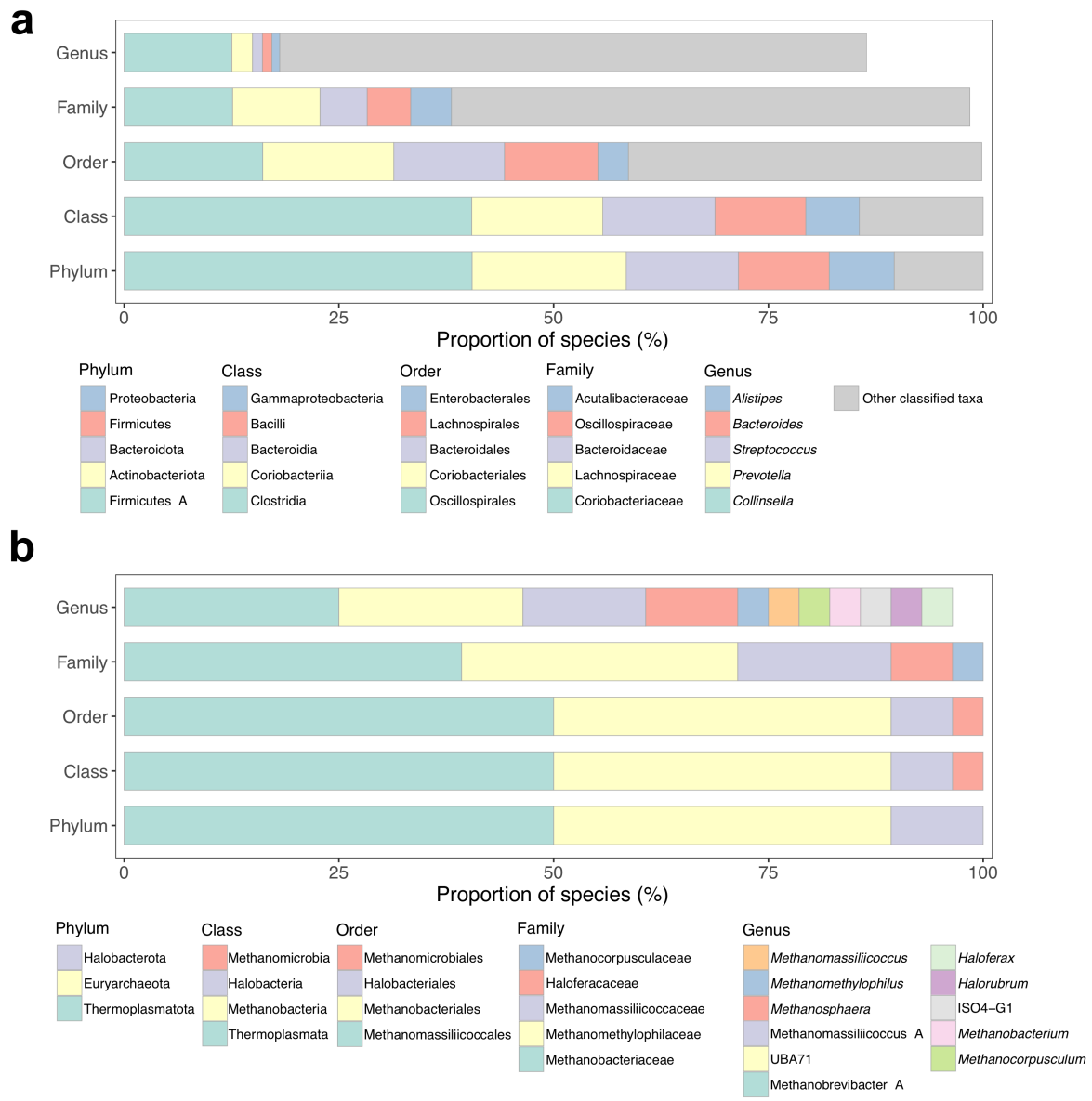
a, Total number of SNVs detected as a function of the number of species. The cumulative distribution was calculated after ordering the species by decreasing number of SNVs. **b**, Number of SNVs detected only in isolate genomes, MAGs, or in both. **c**, Pairwise SNV density analysis of genomes of the same or different type. ***$P < 0.001$ **d**, Right panel shows the number of SNVs exclusively detected in genomes from each continent. The left panel shows the number of exclusive SNVs normalized by the number of genomes per continent. **e**, Pairwise SNV density analysis between genomes from Europe, the largest genome subset, and other continents. The median SNV density was calculated per species and the distribution is shown for all species. Comparison of genomes recovered from the same continent was used as reference. The SNV density between genomes of the same continent is significantly lower (adjusted $P < 0.05$) to that calculated for genomes from different continents.

Supplementary Figures

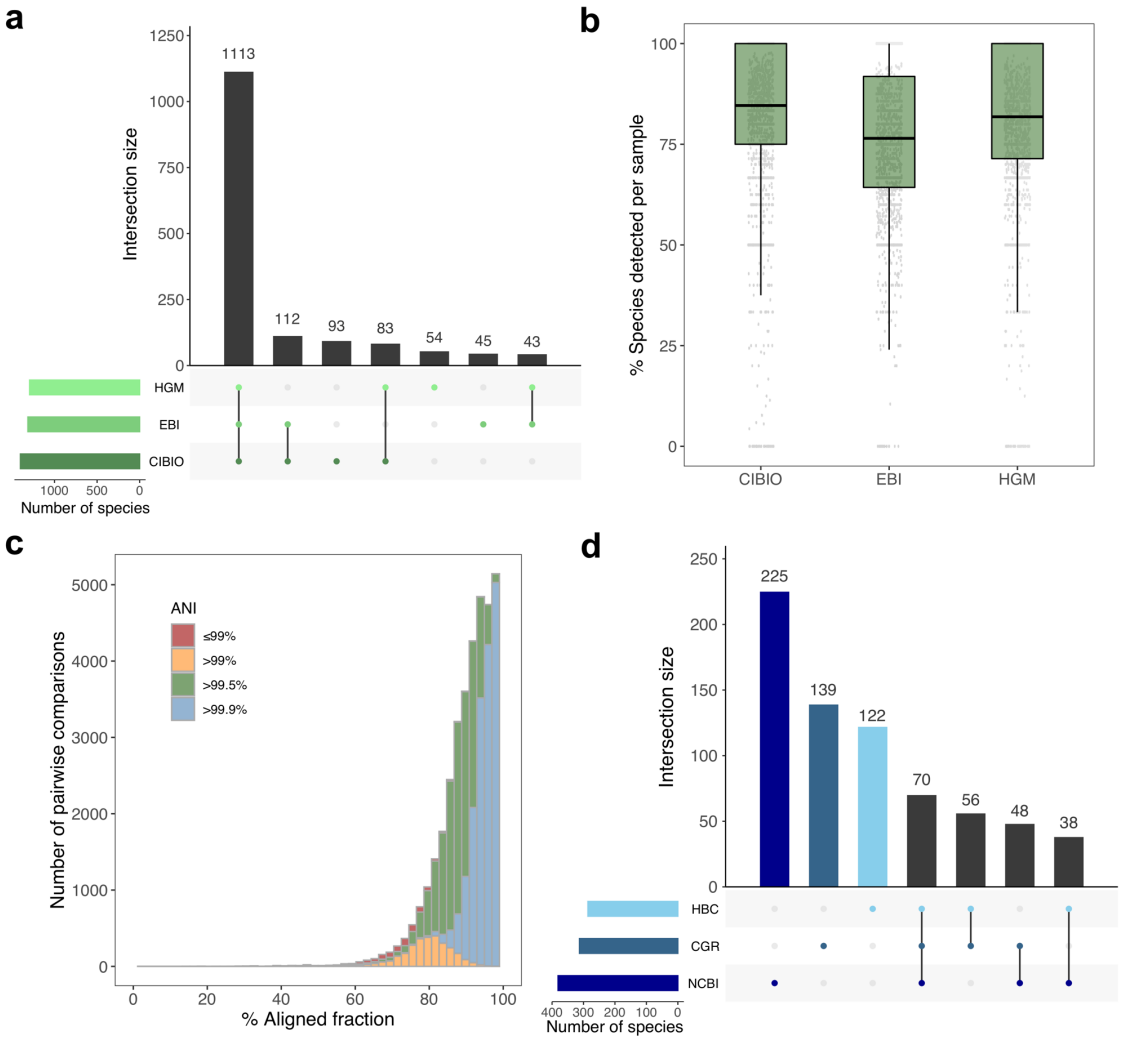


Supplementary Figure 1. Genome quality of species representatives.

a, Completeness and contamination scores for each of the 4,644 species representatives, coloured by their quality classification category. Medium quality: >50% completeness; near complete: ≥90% completeness; high-quality: >90% completeness, presence of 5S, 16S and 23S rRNA genes, as well as at least 18 tRNAs. All genomes have a quality score (QS = completeness – 5 × contamination) above 50. **b**, Number of species according to different completeness and contamination criteria.

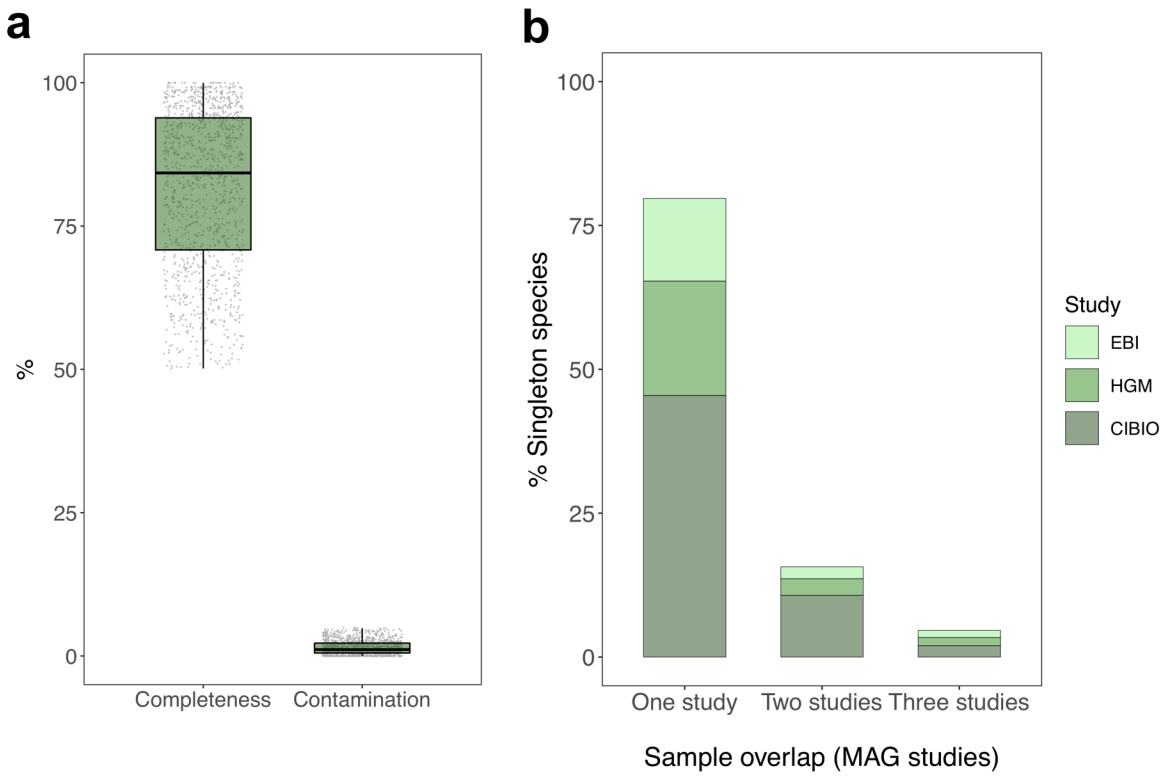


Supplementary Figure 2. Taxonomy composition of the bacterial and archaeal species. **a**, Taxonomic affiliation of the 4,616 bacterial species detected. Data is partitioned by taxonomic rank, with only the five most highly represented taxa per rank depicted in the legend. **b**, Taxonomic affiliation of the 28 archaeal species detected, partitioned by taxonomic rank.

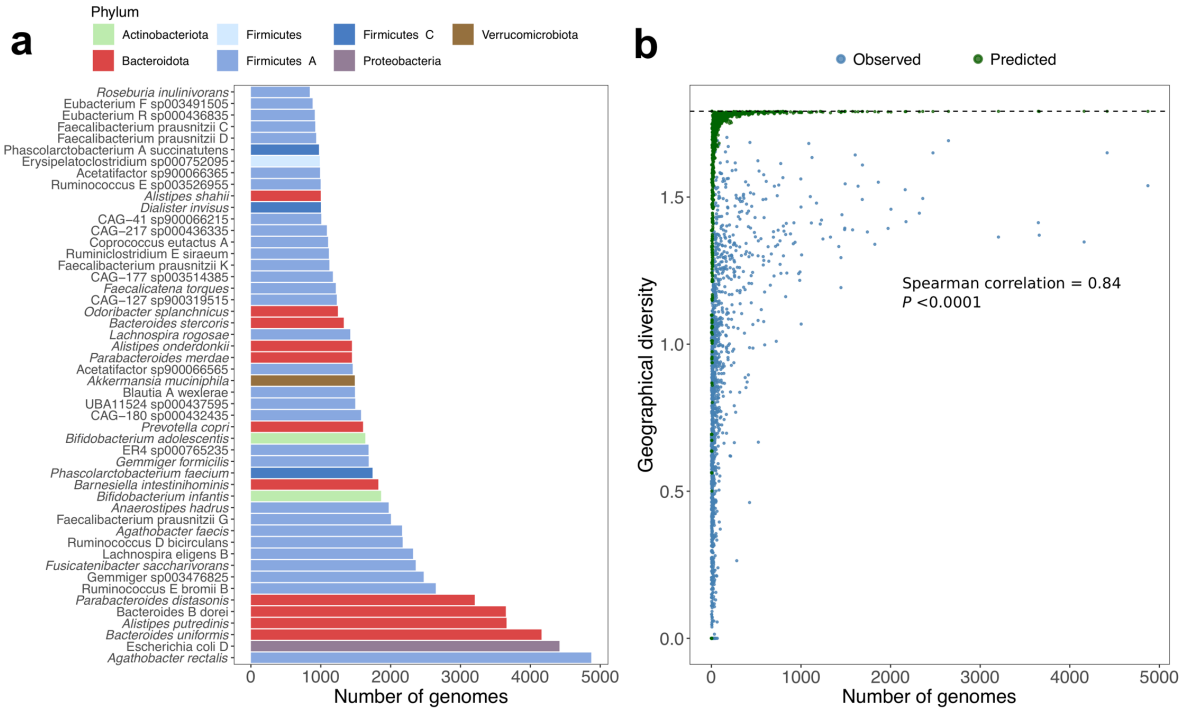


Supplementary Figure 3. Species overlap across study sets.

a, Number of species found across the three metagenome-assembled genome sets, ordered by their level of overlap. Only those genomes recovered from the 1,554 metagenomic samples used by all three studies were considered in this analysis. **b**, Distribution of the proportion of species recovered per sample in each study out of all species recovered across all three studies in the same samples. **c**, Estimated aligned fractions and average nucleotide identities (ANI) between conspecific genomes obtained in the same sample but in different MAG studies. **d**, Number of species identified in three culture-based studies and their degree of overlap. The NCBI study set consists mainly of genomes from the Human Microbiome Project (HMP).

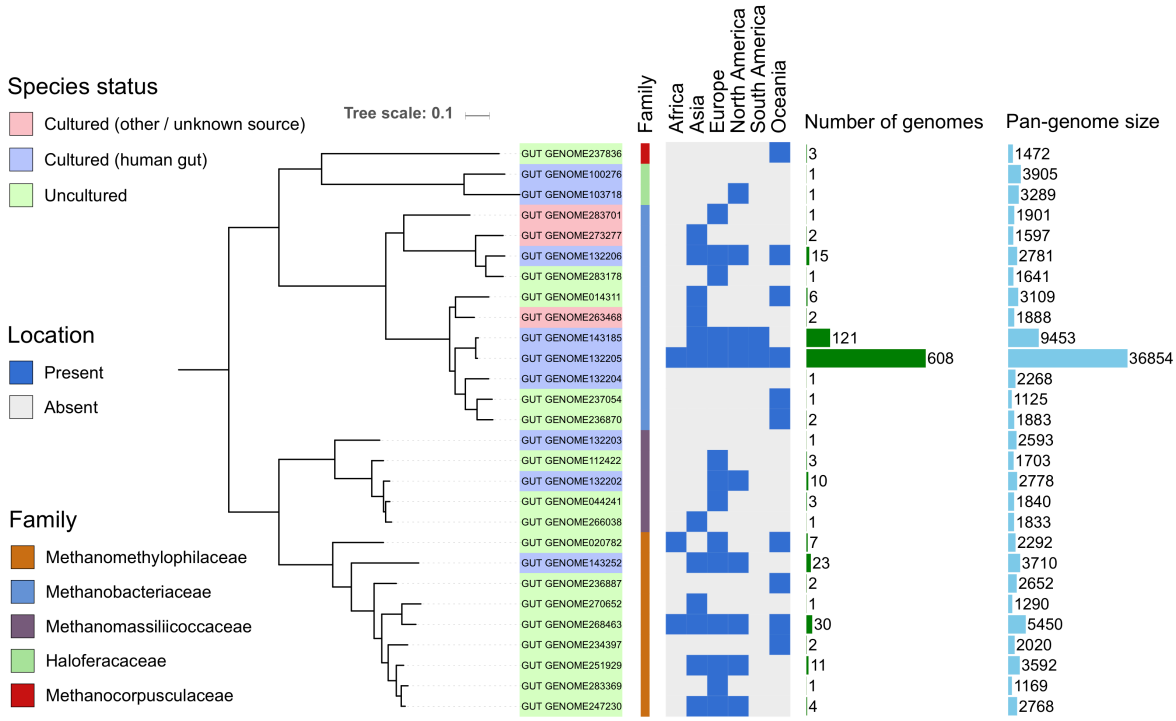


Supplementary Figure 4. Quality and sample origin of uncultured singleton species. **a**, Genome completeness and contamination estimates of the 1,212 uncultured species represented by a single genome. **b**, Proportion of the 1,212 singleton species, by study set, that originated from samples analysed in one, two or three of the MAG studies (CIBIO, EBI and HGM). The CIBIO study used metaSPAdes and MetaBAT 2 for assembling and binning sequencing runs previously merged by sample; the HGM study used MEGAHIT to assemble runs merged by sample and applied a combination of MaxBin 2, MetaBAT 2, CONCOCT and DAS Tool for binning and refinement; the EBI study used metaSPAdes and MetaBAT 2 for assembling and binning individual runs without merging by sample.



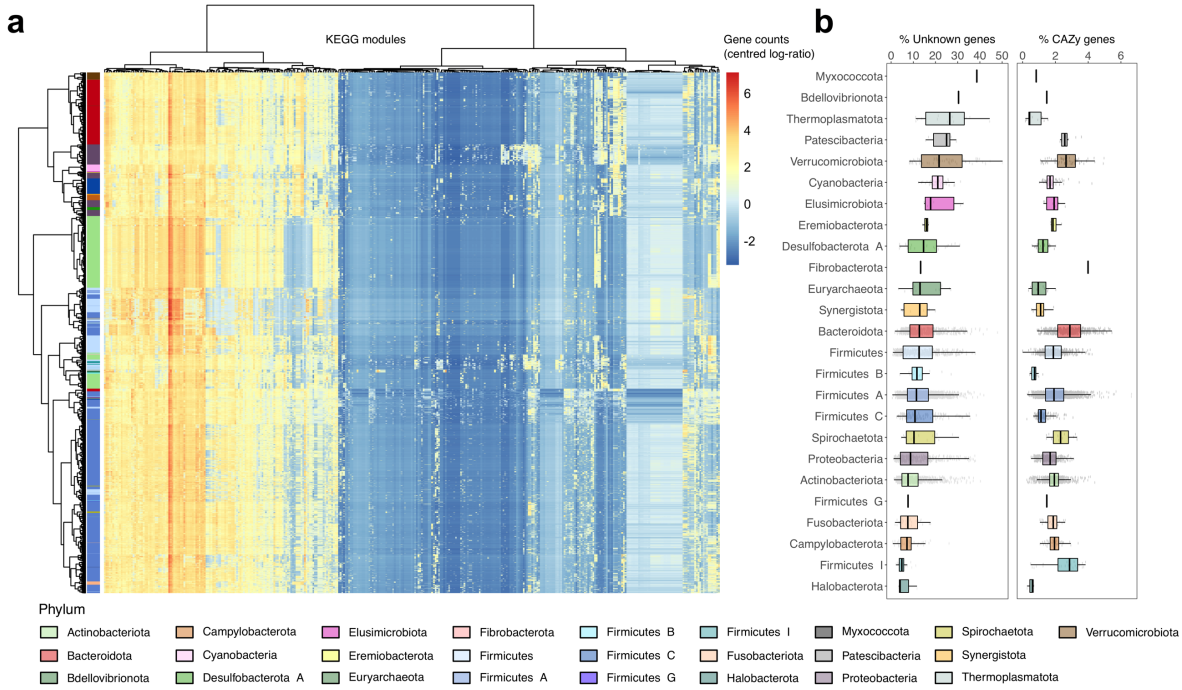
Supplementary Figure 5. Species frequency and geographical diversity.

a, Number of non-redundant genomes retrieved from the 50 most highly represented species in the UHGG. Each species is coloured by its assigned phylum according to the figure legend. **b**, Geographical diversity estimated using the Shannon index in relation to the number of non-redundant genomes from each species. The Spearman's rank correlation coefficient and *P* value are depicted in the graph. Predicted values represent the random geographical distribution of equivalent numbers of genomes observed for each species. Dashed horizontal line indicates the maximum theoretical value of geographical diversity corresponding to equal sample proportions across the six major continents (Africa, Asia, Europe, North America, South America and Oceania).



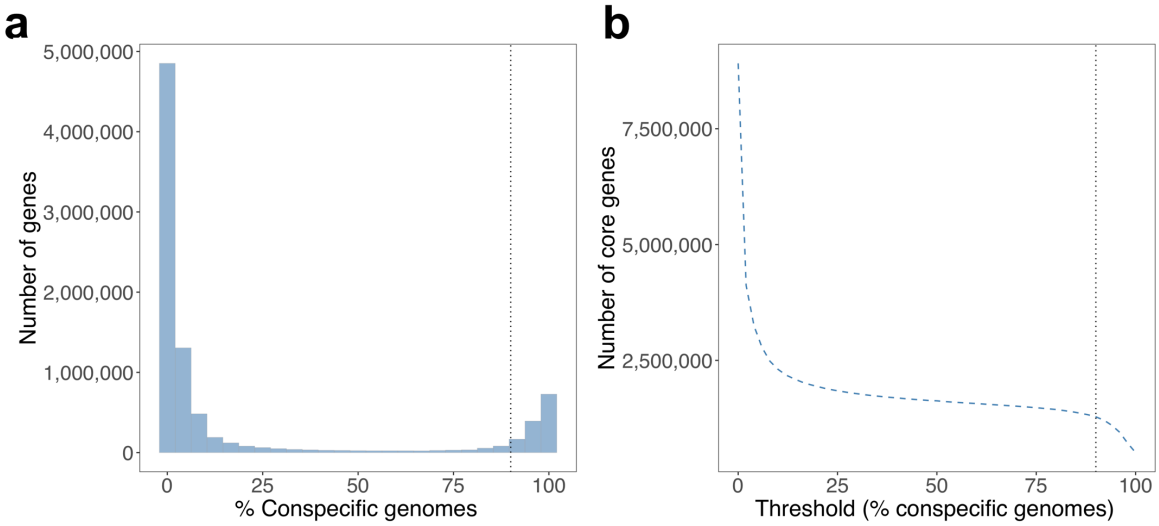
Supplementary Figure 6. Diversity of the gut archaeal species detected.

Phylogenetic tree of the 28 archaeal species detected in the human gut. Tips are labelled with the corresponding species representative code and coloured according to its cultured status. The taxonomic affiliation (family), geographical distribution, number of non-redundant genomes and total pan-genome size are represented next to the tree.



Supplementary Figure 7. Functional annotation of gut microbiome species.

a, Functional profiles of the UHGG species pan-genomes (rows) according to 363 KEGG modules (columns). Numbers of genes matching each module were normalized to centred log-ratios after imputing values with zero counts. Species are coloured according to phylum. KEGG modules and species were hierarchically clustered using the Ward's criterion method. **b**, Proportion of each species pan-genome, partitioned by phylum, without any assignment to the eggNOG, InterPro, COG or KEGG databases (left). Proportion of the pan-genome with a match to the carbohydrate-active enzymes (CAZy) database (right).



Supplementary Figure 8. Gene frequency distribution within the species-level clusters. **a**, Distribution of the number of genes found per fraction of conspecific genomes. Only near-complete genomes (≥90% completeness) were considered in the analysis. **b**, Number of core genes detected based on the threshold of genomes per species used to classify as core. Vertical dashed line represents the 90% threshold used in this study.