nature Accelerated Article Preview

ARTICLE
doi:10.1038/s41586-019-1058-x

Novel insights from uncultivated genomes of the global human gut microbiome

Stephen Nayfach, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, Nikos Kyrpides

Cite this article as: Nayfach, S. et al. Novel insights from uncultivated genomes of the global human gut microbiome. *Nature* https://doi.org/10.1038/s41586-019-1058-x (2019).

Competing interests K.S.P. is on the advisory boards of uBiome and Phylagen.

Received: 8 August 2018; Accepted: 6 March 2019;
Accelerated Article Preview Published online 13 March 2019.

ARTICLE

OPEN

https://doi.org/10.1038/s41586-019-1058-x

Novel insights from uncultivated genomes of the global human gut microbiome

Stephen Nayfach[1,2]*, Zhou Jason Shi[3,4], Rekha Seshadri[1,2], Katherine S. Pollard[3,4,5] & Nikos Kyrpides[1,2]*

Largely due to challenges cultivating microbes under laboratory conditions, the genome sequence of many species in the human gut microbiome remains unknown. To address this problem, we reconstructed 60,664 prokaryotic draft genomes from 3,810 faecal metagenomes from geographically and phenotypically diverse human subjects. These genomes provide reference points for 2,058 previously unknown species-level operational taxonomic units (OTUs), representing a 50% increase in the phylogenetic diversity of sequenced gut bacteria. On average, new OTUs comprise 33% of richness and 28% of species abundance per individual and are enriched in humans from rural populations. A meta-analysis of clinical gut microbiome studies pinpointed numerous disease associations for new OTUs, which have the potential to improve predictive models. Finally, our analysis revealed that uncultured gut species have undergone genome reduction with loss of certain biosynthetic pathways, which may offer clues for improving cultivation strategies in the future.

The gut microbiome plays a myriad of important roles in human health and disease[1]. Microbial reference genomes are essential resources for understanding the functional role of specific organisms in the microbiome and for quantifying their abundance from metagenomes[2]. However, an estimated 40-50% of human gut species lack a reference genome[3,4]. While significant efforts have been made to culture and sequence members of the gut microbiome[5–7], many microorganisms have not been grown under laboratory conditions to date and still lack a sequenced genome, despite being prevalent in humans[8].

Recent advances in experimental technologies have begun to close this gap. Browne et. al[6] and Lagier et. al[7] used microbial culturomics to isolate and sequence hundreds of previously uncultured organisms in the human gut, while other studies have performed single-cell genome sequencing[9]. In contrast to experimental approaches, metagenome binning is a computational approach to obtain genomes directly from samples without isolation or culturing. Sequencing reads are first assembled into contigs, which are then binned into metagenome-assembled genomes (MAGs) based on nucleotide frequency, abundance, and/or co-variation of abundance across a group of samples[10]. This process is performed either for individual metagenomes[11] or multiple co-assembled metagenomes[12]. MAGs are subsequently evaluated for various indicators of genome quality, including estimated completeness and contamination, the presence of marker genes, and overall contiguity[13–15].

MAGs were first assembled from a low-complexity acid mine drainage community[16], but with advances in sequencing technology and computational methods, MAGs have now been recovered from a myriad of environments including the global ocean[17], cow rumen[12], aquifer systems[18], and others[11]. These uncultured genomes have expanded the tree of life, revealing novel lineages in diverse environments, and new biology[11,19]. Despite the growing number of publicly-available human gut metagenomes, there has not been any large-scale assembly of MAGs from the gut microbiome. Nielsen et al.[20] were the first to recover MAGs from gut metagenomes and similar concepts have been developed and applied to other individual studies[21]. We hypothesized that human gut MAGs systematically recovered from public metagenomes could significantly increase the diversity of species with a sequenced genome and shed light on the biology of uncultivated organisms in the gut microbiome.

Reconstructing genomes from global gut metagenomes

To recover genomes for novel human gut lineages, we performed metagenomic assembly and binning on 3,810 globally-distributed samples from phenotypically and demographically diverse human subjects using a pipeline developed for this study (Fig. 1a,b, Supplementary Tables 1–5). MAG quality was improved further using a pipeline we developed to identify and remove of incorrectly binned contigs (Fig. 1C, Extended Data Fig. 1, Supplementary Table 6–7, and Methods). We performed single-sample assembly and binning, rather than co-assembly, in order to preserve strain variation between human hosts and because co-assembly was not computationally feasible for our large dataset. Based on a subset of samples, our pipeline produced 1.8x more non-redundant high-quality MAGs compared to co-assembly and 3.3x more than a previously study[20] which utilized abundance co-variation across samples (Extended Data Fig. 2).

Our pipeline yielded 60,664 MAGs that met or exceeded the medium-quality MIMAG standard[14] which we refer to as the Global Human Gut MAG (HGM) dataset (Fig. 1b and Supplementary Table 8). The MAGs form 43,737 clusters at an average nucleotide identity (ANI) threshold of 99%, indicating that most are unique. The vast majority of MAGs displayed >98% DNA identity within the same species and <98% identity between species at individual marker-genes, suggesting they are not chimeric (Extended Data Fig. 3g–l). A subset of 24,345 high-quality MAGs were estimated to be near-complete, with minimal contamination, high contiguity, and were of similar length as isolate genomes of the same species (Fig. 1b,d, Extended Data Fig. 3f). Only 14.5% of these were classified as high-quality by the MIMAG standard, largely due to the absence of a full complement of rRNA genes, which are challenging to assemble from metagenomes[22] and often absent from otherwise near-complete MAGs[11].

Despite the large number of recovered genomes, we identified several challenges to recovering MAGs from human gut metagenomes. First, by mapping reads back to each MAG and quantifying

[1]U. S. Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [2]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA. [3]Gladstone Institutes, San Francisco, California, USA. [4]Chan-Zuckerberg Biohub, San Francisco, California, USA. [5]Institute for Human Genetics, Institute for Computational Health Sciences, Quantitative Biology Institute, and Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA. *e-mail: snayfach@lbl.gov; nckyrpides@lbl.gov

single-nucleotide polymorphisms (SNPs), we confirmed that strain diversity results in highly fragmented MAGs (Fig. 1e)[15]. Second, we found that reliably assembling a MAG required at least 10-20x read-depth (Fig. 1f), indicating that MAGs were only assembled for the most abundant taxa in each community[23]. MAG assembly was particularly challenging for certain phyla, like Bacteroidetes (Fig. 1f) and for metagenomes with high community diversity (Extended Data Fig. 4a–b). Despite these challenges, our results indicate that thousands of partial and near-complete genomes can be reconstructed from individual human gut metagenomes using standard pipelines for assembly and binning.

MAGs represent thousands of unknown species

To explore whether the HGM dataset represented novel taxa, we clustered the 60,664 MAGs plus 145,917 non-redundant reference genomes into species-level operational taxonomic units (OTUs) on the basis of 95% average nucleotide identity (ANI; Fig. 2a, Supplementary Table 9–10). While the species concept for prokaryotes is controversial[24], our operational definition is commonly used[3,4] and considered one gold standard[25]. We found our species-level OTUs were consistent with taxonomic annotations from other databases and were robust to genome incompleteness and contamination (Extended Data Fig. 3a,b,c and Extended Data Fig. 5).

Our procedure yielded a total of 23,790 species-level OTUs with 4,558 from the human gut microbiome (Fig. 2a,b, Extended Data Fig. 6a,b, and Supplementary Table 10). We formed the Integrated Gut Genomes Database (IGGdb) with the 156,478 genomes that comprise the human gut OTUs, which includes 2,058 new OTUs comprised exclusively of 10,368 MAGs (Fig. 2d). Supporting their novelty, 96% of new OTUs were not classified at the species level based on the Genome Taxonomy Database[26] (GTDB; Supplementary Table 10) and 69% of new OTUs had <90% ANI to any OTU containing a reference genome.

A significant number of MAGs were not taxonomically classified at or above the genus rank (N=3,215, Supplementary Table 10). To identify the novel clades represented by these MAGs, we constructed a phylogeny of all MAGs and reference genomes and clustered them based on rank-specific phylogenetic distance cutoffs (Fig. 2c, Extended Data Fig. 3d,e). This revealed 360 new genus-level OTUs, 15 new family-level OTUs, and 2 new order-level OTUs (Fig. 2d). A collector's curve revealed saturation of OTUs at or above the genus rank, but not yet for species (Extended Data Fig. 6c). Together, MAGs from new OTUs represented 70.9% of the total phylogenetic diversity (PD) of sequenced gut Bacteria and a 50.0% increase compared to reference genomes alone (Fig. 2e).

Novel OTUs were broadly distributed across taxonomic groups (Fig. 3), with hotspots of new diversity in the Firmicutes orders Lachnospirales and Oscillospirales. Nearly 400 novel OTUs were discovered within the Bacteroidetes despite challenges of assembling this phylum (Fig. 1f and Fig. 3). In contrast, nearly no new OTUs were found for Archaea even though MAGs were easily assembled (Fig. 1f), suggesting that most human gut Archaea already have a sequenced genome. Several large clades within Cyanobacteria and Clostridia were not represented by any high-quality genome, which may be explained by genome reduction or unknown factors that interfere with genome assembly (Extended Data Fig. 7a). Overall, these results indicate that the HGM dataset has greatly expanded the genomic diversity of bacteria across the tree of life in the human gut.

Distribution of new species in the human population

While a number of tools exist for metagenomic taxonomic profiling, none contain the MAGs from this study. To address this problem, we developed IGGsearch, which utilizes a similar strategy as MetaPhlAn[27] to rapidly estimate the abundance of all 23,790 species-OTUs by aligning metagenomic reads to a database of single-copy, species-specific genes identified from MAGs and reference genomes (Supplementary Fig. 1, Methods, and Software Availability). Based on benchmark datasets, we found that IGGsearch accurately quantifies OTU abundance and presence-absence (Supplementary Fig. 2 and Supplementary Tables 11–12).

Using IGGsearch profiling, we found that the novel species-OTUs accounted for 33.4% of richness and 27.7% of relative abundance per sample from healthy individuals (Extended Data Fig. 4c), were similarly abundant in metagenomes not used for assembly or binning (Supplementary Table 13), and were commonly detected in samples where no MAG was recovered (Extended Data Fig. 4e). New species-OTUs were particularly abundant in healthy adults from rural populations (Tanzania, Peru, Mongolia, Fiji, and El Salvador) but were surprisingly rare in infants from Europe and the United States (Fig. 2f and Extended Data Fig. 4d). Communities with high diversity were enriched for new OTUs, but no difference was observed between healthy and diseased individuals (Extended Data Fig. 4f, Supplementary Table 13–14). Together, these results reveal that the novel uncultured OTUs discovered in this study comprise a significant fraction of the healthy human gut microbiome and are more common in non-western populations.

Association of gut species with human diseases

Human gut microbiota have been linked to a myriad of diseases and associations with the microbiome can be leveraged for understanding disease etiology, for clinical diagnosis, or for building predictive models[1,21]. We hypothesized IGGsearch could identify novel associations with human diseases among the 2,058 new species-OTUs discovered in this study. To address this question, we performed metagenome-wide association of species-OTUs from the IGGdb versus disease status for ten different clinical microbiome studies, including six that were not used for MAG recovery (Supplementary Tables S15–16 and Methods).

Overall, we identified 2,283 species-disease associations at a false discovery rate (FDR <1%) that included an even balance of case-enriched and control-enriched OTUs (Extended Data Table 1). Nearly 40% of disease associations corresponded to novel OTUs, including many of the most significant associations (Fig. 4). For example, the most significant association for ankylosing spondylitis (an inflammatory arthritis affecting the spine and large joints) was a new species in the Negativicutes class (OTU-14148, adjusted p value $= 5.3 \times 10^{-28}$), which was strongly depleted in patients relative to healthy controls and 8-orders of magnitude more significant than the most strongly associated species with a reference genome.

To contextualize these results, we estimated microbial species abundances in the same datasets using three other commonly used tools - MIDAS[4], mOTU[3], and MetaPhlAn2[27] - along with reference databases distributed with each tool. After applying the same statistical procedure to each set of abundance profiles, we identified 716, 404, and 326 disease associations, respectively (Extended Data Table 1), which is nearly 5-fold fewer compared to IGGsearch. Additionally, we used abundance data from each tool to build Random Forest machine learning models to predict disease status. We found that IGGsearch abundance profiles yielded the most predictive model (or equivalent) for eight of the ten diseases, with significant improvements for colorectal cancer, cardiovascular disease, type-2 diabetes, and rheumatoid arthritis (Extended Data Table 1). More work is needed to understand how associated species relate to disease etiology and whether these results replicate in other human populations.

Genome reduction of uncultured gut bacteria

Previous MAG studies of environmental communities have uncovered large uncultured lineages with unusual genomic properties, including reduced genomes, slow replication rates, and the absence of conserved genes[19,28]. Surprisingly, we found that the human gut also harbors a number of large lineages that are exclusively represented by MAGs (Fig. 2d and Extended Data Fig. 6b). To elucidate biological properties of these organisms, we performed a comparative genomic analysis between cultured and uncultured species-OTUs from the gut (Supplementary Table 17 and Methods).

Strikingly, uncultured OTUs tended to have significantly reduced genomes, which was consistent across all major phyla and classes tested, including Actinobacteria, Bacilli, Clostridia, Bacteroidetes, and Proteobacteria (Fig. 5a). While previous studies have identified human gut taxa with reduced genomes, including TM7[29] and Melainabacteria[30], this is the first time this pattern has been reported at this scale. Other genomic features, including estimated replication rates, coding density, and GC content, did not consistently differ between cultured and uncultured OTUs (Extended Data Fig. 8 and Supplementary Table 18).

Given the overall pattern of genome reduction, we used phylogenetic logistic regression to identify functions that were commonly missing from uncultured OTUs (Methods). Overall, we found 1,492 KEGG orthology groups (KOs; 21.5% of total) that significantly differed between groups at an FDR of <1%, most of which were depleted from uncultivated OTUs (Fig. 5c). These patterns were consistent between MAGs and isolate genomes of the same species and were not affected by the database used for functional annotation (Extended Data Fig. 9a,b). Among our top hits, we found functions related to maintenance of osmotic pressure and protection against oxidative stress (Extended Data Fig. 9c), which may indicate that uncultivated bacteria are less viable after transfer to culture media or are more sensitive to oxygen exposure outside of the host[6].

The above patterns were best exemplified by RF39, which is an uncultivated order within the class Bacilli with a highly reduced genome and numerous auxotrophies (Fig. 5c). Remarkably little has been published regarding this enigmatic group even though RF39 has been detected in previous MAG studies[11,20] and was found in a large proportion of metagenomes analyzed in our study (Fig. 3). Numerous highly conserved metabolic pathways were entirely missing from nearly all RF39 genomes, including those for biosynthesis of fatty acids (FAs) and several amino acids and vitamins. The complete loss of the FA biosynthesis pathway was striking because FAs are integral components of cellular membranes and considered a housekeeping capacity of cells. These organisms may incorporate exogenous FAs into membrane phospholipids using a recently described mechanism in Firmicutes[31].

Discussion

Here we illustrated that it is possible to use large-scale metagenomic assembly and binning to recover thousands of genomes for previously unknown members of the human gut microbiome. We generated the IGGdb and IGGsearch as resources to drive further discoveries in human microbiome science. During review of this manuscript several studies were published that generated many new human gut genomes from metagenomes[32,33] and cultivated isolates[34,35]. In the future, these new genomes could be integrated with the IGGdb to provide an updated catalog of genomes from the gut microbiome.

While we recovered thousands of MAGs, we also identified several challenges, including low species abundance, high strain diversity, and low recovery rates for certain phyla, like Bacteroidetes. Future efforts to recover MAGs from the gut microbiome may benefit from alternative approaches that target these hard to assemble organisms. Likewise, we found that adults from non-western countries were a major source of novel diversity, which indicates that future metagenome studies should focus on human populations outside of Europe, the United States, and China.

One of the most surprising results from our study was that the majority of microbial diversity in the human gut is not currently represented by cultured isolates, which are important for numerous applications in basic research and biotechnology. In the future, MAGs from this study could be used to improve culture conditions or identify novel growth factors for uncultured human gut species. For example, menaquinone and fatty acids have been shown to promote the growth of uncultured bacteria[36,37] and both pathways were missing from many uncultured OTUs from this study (Supplementary Table 19). Further, we found that uncultivated bacteria have undergone significant genome reduction, which may be an adaptive process resulting from utilization of public goods as outlined in the Black Queen hypothesis[38], although more work is needed to explore this question.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1058-x.

Received: 8 August 2018; Accepted: 6 March 2019;
Published online 13 March 2019.

1. Lynch, S.V. and O. Pedersen, The Human Intestinal Microbiome in Health and Disease. *N Engl J Med*, 2016. **375**(24): p. 2369-2379.
2. Kyrpides, N.C., et al., Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol*, 2014. **12**(8): p. e1001920.
3. Sunagawa, S., et al., Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods*, 2013. **10**(12): p. 1196-9.
4. Nayfach, S., et al., An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res*, 2016. **26**(11): p. 1612-1625.
5. Human Microbiome Jumpstart Reference Strains, C., et al., A catalog of reference genomes from the human microbiome. *Science*, 2010. **328**(5981): p. 994-9.
6. Browne, H.P., et al., Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature*, 2016. **533**(7604): p. 543-6.
7. Lagier, J.C., et al., Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat Microbiol*, 2016. **1**: p. 16203.
8. Fodor, A.A., et al., The "most wanted" taxa from the human microbiome for whole genome sequencing. *PLoS One*, 2012. **7**(7): p. e41294.
9. Brito, I.L., et al., Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 2016. **535**(7612): p. 435-439.
10. Alneberg, J., et al., Binning metagenomic contigs by coverage and composition. *Nat Methods*, 2014. **11**(11): p. 1144-6.
11. Parks, D.H., et al., Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol*, 2017. **2**(11): p. 1533-1542.
12. Stewart, R.D., et al., Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun*, 2018. **9**(1): p. 870.
13. 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.
14. 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*, 2017. **35**(8): p. 725-731.
15. Sczyrba, A., et al., Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat Methods*, 2017. **14**(11): p. 1063-1071.
16. Tyson, G.W., et al., Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 2004. **428**(6978): p. 37-43.
17. Tully, B.J., E.D. Graham, and J.F.Heidelberg, The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data*, 2018. **5**: p. 170203.
18. Anantharaman, K., et al., Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun*, 2016. **7**: p. 13219.
19. Brown, C.T., et al., Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, 2015. **523**(7559): p. 208-11.
20. Nielsen, H.B., et al., Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol*, 2014. **32**(8): p. 822-8.
21. Qin, J., et al., A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 2012. **490**(7418): p. 55-60.
22. Yuan, C., et al., Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics*, 2015. **31**(12): p. i35-43.
23. Luo, C., et al., Individual genome assembly from complex community short-read metagenomic datasets. *ISME J*, 2012. **6**(4): p. 898-901.
24. Rossello-Mora, R. and R. Amann, The species concept for prokaryotes. *FEMS Microbiol Rev*, 2001. **25**(1): p. 39-67.
25. Richter, M. and R. Rossello-Mora, Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A*, 2009. **106**(45): p. 19126-31.
26. Parks, D.H., et al., A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*, 2018.
27. Truong, D.T., et al., MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods*, 2015. **12**(10): p. 902-3.
28. Brown, C.T., et al., Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol*, 2016. **34**(12): p. 1256-1263.
29. Podar, M., et al., Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl Environ Microbiol*, 2007. **73**(10): p. 3205-14.
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*, 2013. **2**: p. e01102.
31. Cronan, J.E., A new pathway of exogenous fatty acid incorporation proceeds by a classical phosphoryl transfer reaction. *Mol Microbiol*, 2014. **92**(2): p. 217-21.

32. Almeida, A., et al., A new genomic blueprint of the human gut microbiota. *Nature*, 2019.
33. 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.
34. Forster, S.C., et al., A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol*, 2019. **37**(2): p. 186-192.
35. Zou, Y., et al., 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol*, 2019. **37**(2): p. 179-185.
36. Fenn, K., et al., Quinones are growth factors for the human gut microbiota. *Microbiome*, 2017. **5**(1): p. 161.
37. Hazlewood, G. and R.M. Dawson, Characteristics of a lipolytic and fatty acid-requiring Butyrivibrio sp. isolated from the ovine rumen. *J Gen Microbiol*, 1979. **112**(1): p. 15-27.
38. Morris, J.J., R.E. Lenski, and E.R. Zinser, The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio*, 2012. **3**(2).

Acknowledgements The authors would like to thank Sean Jungbluth, Elhanan Borenstein, Peter Turnbaugh, Michael Fishbach, and Jordan Bisanz for feedback and suggestions, and Patrick Bradley for advice on phylogenetic regression. The work conducted by the US Department of Energy Joint Genome Institute (JGI), a US Department of Energy Office of Science User Facility, is supported under contract no. DE-AC02-05CH11231. This work was also supported by funding from the NSF (grant #DMS- 1563159), the Chan-Zuckerberg Biohub, and the Gladstone Institutes.

Reviewer information *Nature* thanks Jack Gilbert, Phil Hugenholtz and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.N. conceived of the project, designed experiments, analyzed data, made figures, wrote software, and drafted the manuscript. J.S. contributed code for machine learning and tested software. R.S. contributed to analysis of MAGs from uncultured lineages. K.S.P. provided feedback, computational resources, and funding. N.K supervised the project, provided feedback, and drafted the manuscript. All authors read, edited, and reviewed the manuscript.

Competing interests K.S.P. is on the advisory boards of uBiome and Phylagen.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1058-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1058-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.N. or N.K.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

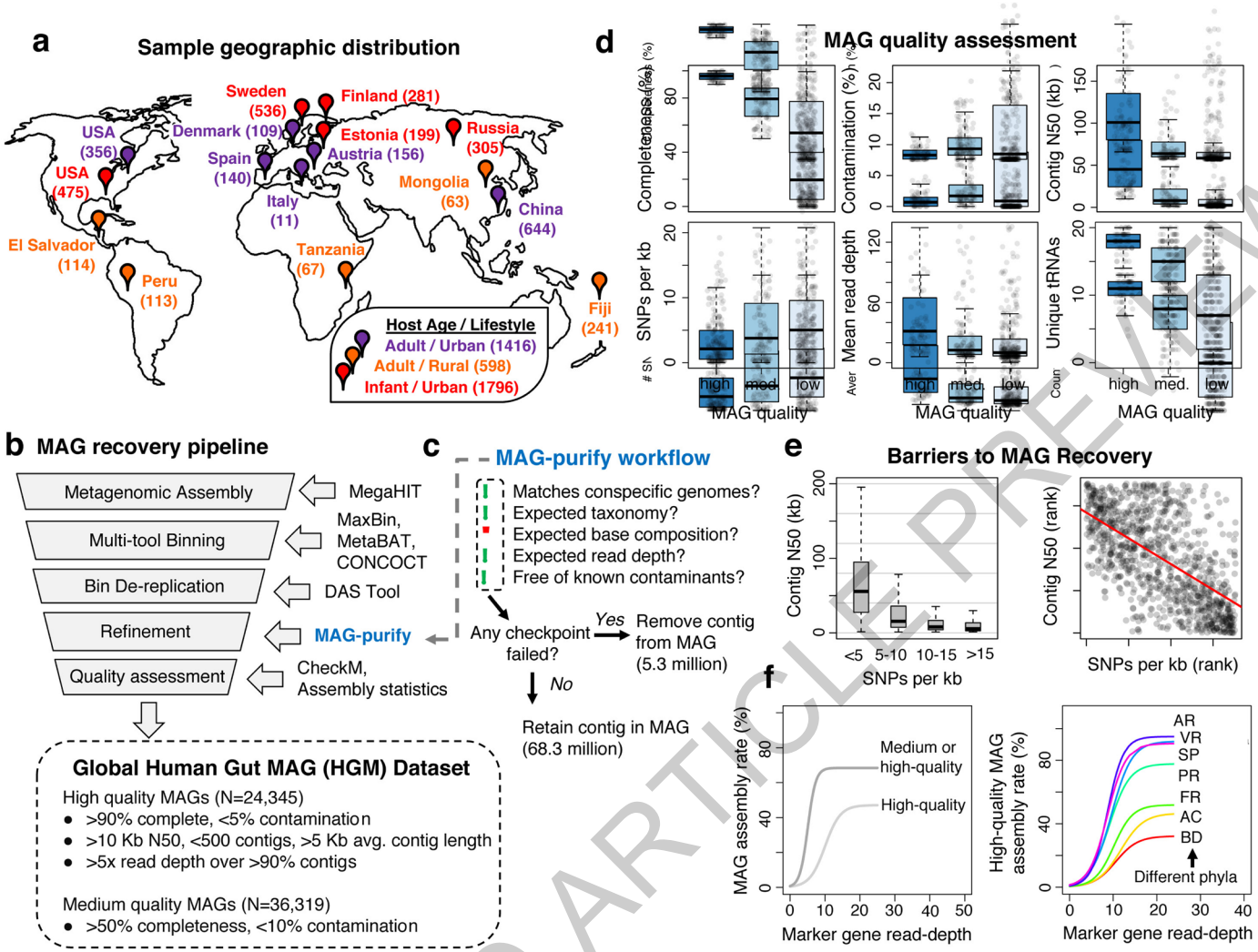


Fig. 1 | Recovery of genomes from globally-distributed gut metagenomes. A) Geographic distribution of metagenomes. Sample sizes are indicated in parenthesis and pin color indicates the majority age group and lifestyle (infants are ≤3 years old; adults are ≥18 years old). Several locations represent multiple studies, while several studies were conducted in multiple locations. **B**) Computational pipeline for assembling MAGs. **C**) Pipeline for identifying and removing incorrectly binned contigs. **D**) Quality metrics across low (N=101,651), medium (N=36,319), and high quality (N=24,345) MAGs. **E**) SNPs were called for MAGs with sufficient read-depth (N=17,671) and compared with N50. Red line is from a Spearman correlation (ρ=-0.61). **F**) At least 10-20x depth is required to assemble a MAG, but assembly rates vary between taxa (AR=Archaea; VR=Verrucomicrobia; SP=Spirochaetes; PR=Proteobacteria; FR=Firmicutes; AC=Actinobacteria; BD=Bacteroidetes). Sequencing read-depth was estimated using IGGsearch (see Methods) and curves fit using logistic regression. For box plots, the middle line denotes median; box denotes interquartile range (IQR); and whiskers denote 1.5x IQR.

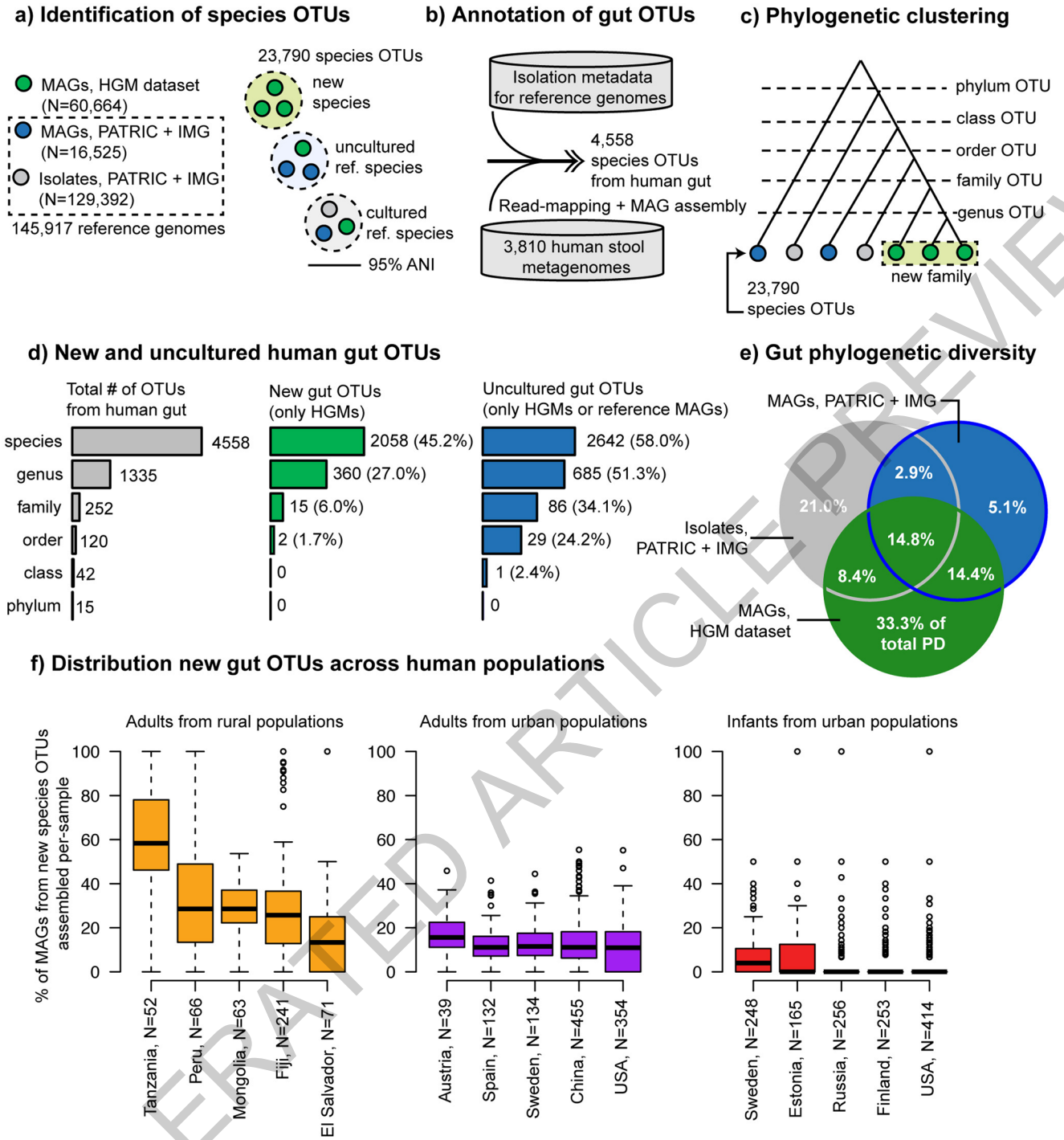


Fig. 2 | Human gut MAGs expand the genomic diversity of the gut microbiome. **A**) Reference genomes were clustered with MAGs at 95% ANI. **B**) Human gut OTUs were identified based on isolation metadata, read-mapping, or assembly of a gut MAG. **C**) All OTUs were further clustered into higher-ranking groups. **D**) A significant fraction of gut OTUs are represented exclusively by MAGs. **E**) Pie chart indicating the percentage of bacterial phylogenetic diversity (PD) in the gut covered by different genome sets. **F**) Distribution of new OTUs across healthy human populations. Only countries with at least 20 samples are shown. For box plots, the middle line denotes median; box denotes IQR; and whiskers denote 1.5x IQR.

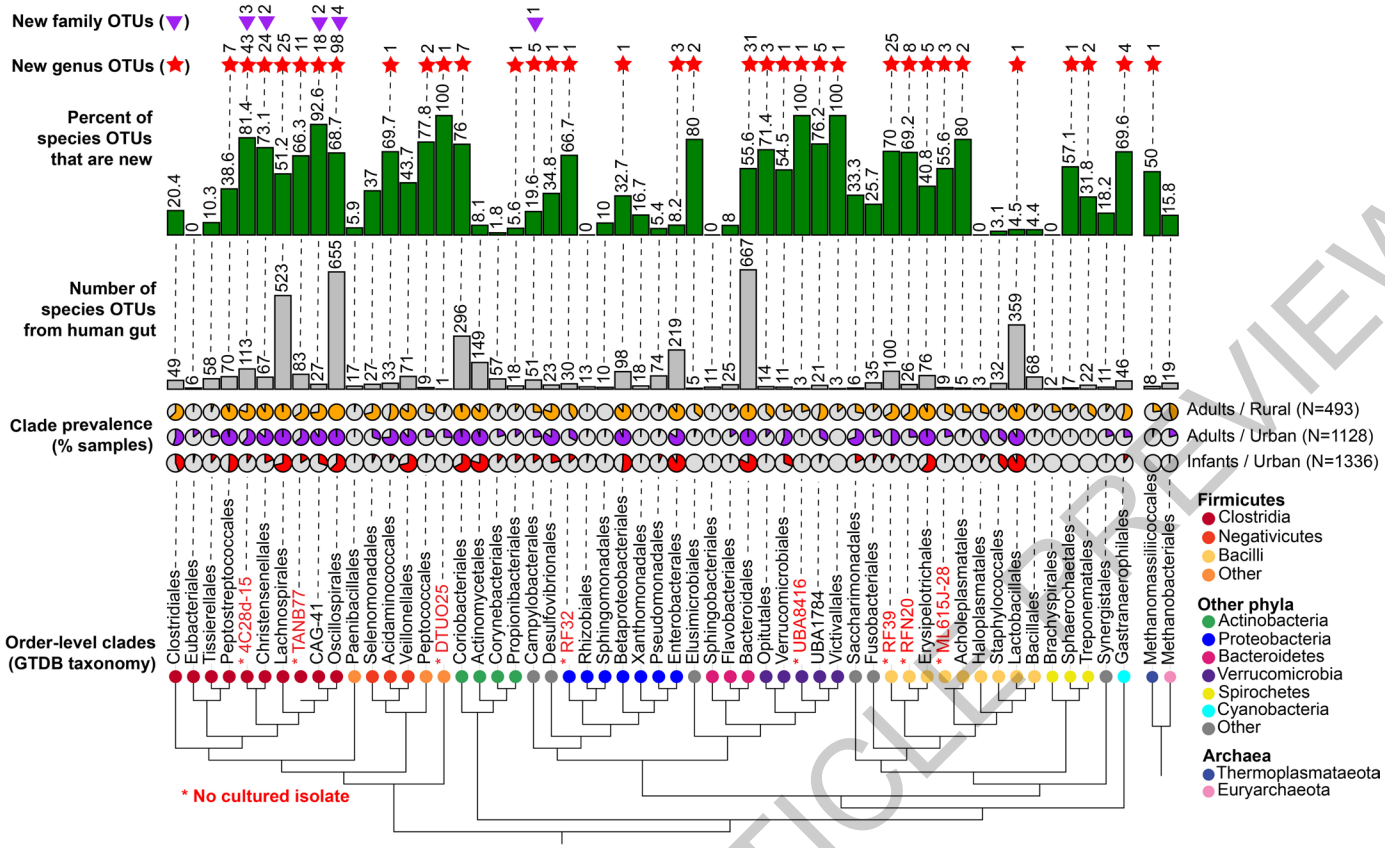


Fig. 3 | New gut species are broadly distributed across taxonomic groups. The figure indicates order-level clades with ≥10 human gut species-OTUs or detected in ≥10% of metagenomes from healthy individuals. Taxonomic labels are based on the GTDB. Red labels indicate orders represented exclusively by MAGs (current or previous studies). Pie charts indicate the prevalence of orders across metagenomes from healthy individuals. Gray bars indicate the number of gut species-OTUs per order, and the green bars indicate the percent of those OTUs that are new. Red stars and purple triangles indicate the number of new genus-level and family-level OTUs, respectively.

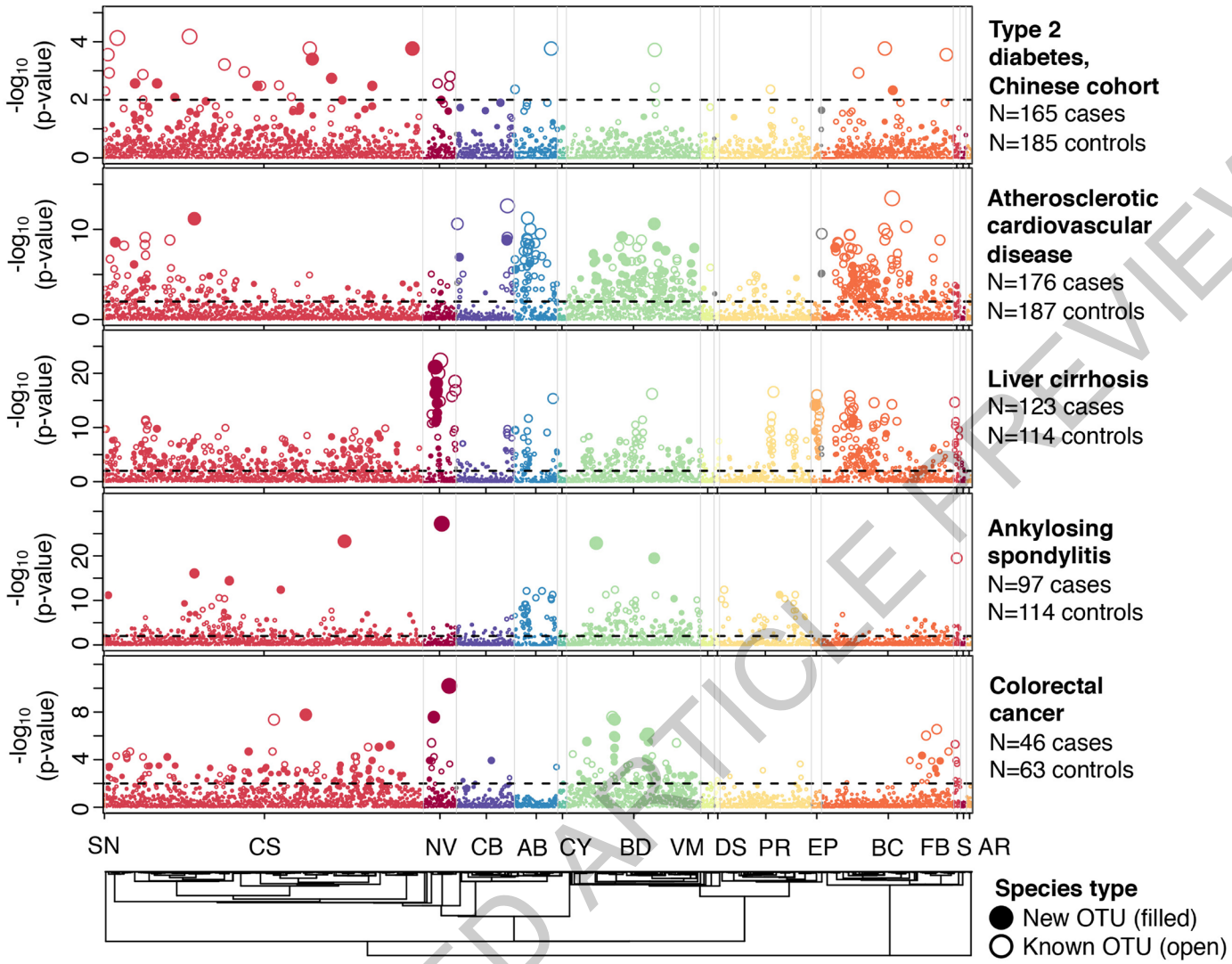


Fig. 4 | Metagenome-wide association of gut OTUs with human diseases. The Manhattan plot shows the phylogenetic distribution of species-disease associations for different metagenomic studies. Each point is one species-OTU and point height indicates the *p* value from a two-sided Wilcoxon rank-sum test of estimated species abundance between diseased and healthy individuals after correction for multiple hypothesis tests. The dotted line indicates a false discovery rate (FDR) of 1%. The plot shows results for five diseases with greater than 10 species-disease associations. Species are ordered according to their phylogeny, which is displayed at the bottom (SN: Synergistetes, CS: Clostridia, NV: Negativicutes, CB: Coriobacteriia, AB: Actinobacteria, CY: Cyanobacteria, BD: Bacteroidetes, VM: Verrucomicrobia, DS: Desulfobacteraeota, PR: Proteobacteria, EP: Epsilonbacteraeota, BC: Bacilli, FB: Fusobacteria, S: Spirochaetes, AR: Archaea).

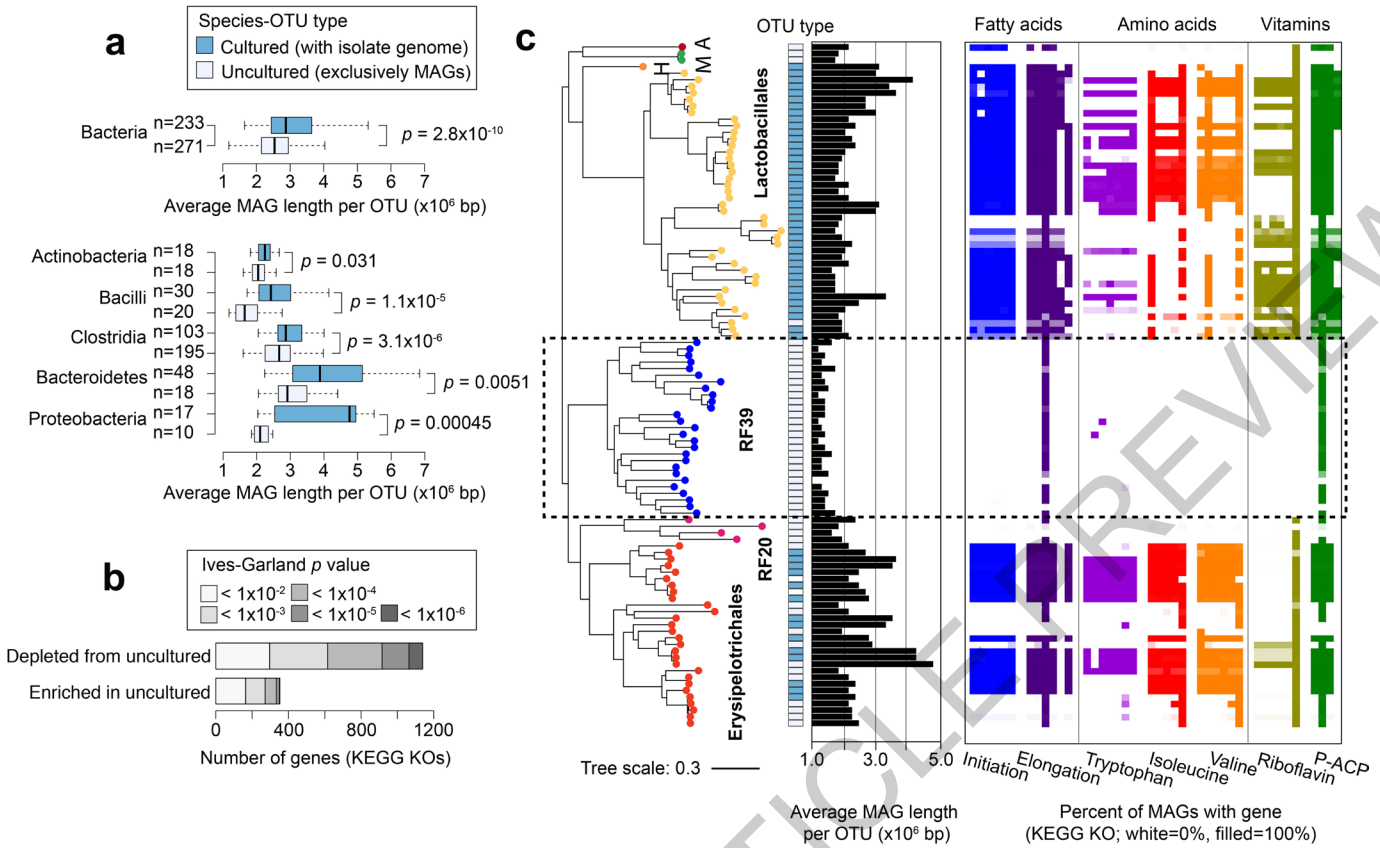


Fig. 5 | Uncultured OTUs have reduced genomes and are missing common biological functions. A) Comparison of genome size between cultivated and uncultivated species-OTUs after correction for incompleteness and contamination. Middle line of boxplots denotes median; box denotes IQR; and whiskers denote 1.5× IQR. **B**) Genes from the KEGG database were compared between 233 cultivated and 271 uncultivated species-level OTUs using phylogenetic logistic regression. Most significant genes are depleted from uncultivated species. **C**) Phylogenetic tree of species-OTUs from Bacilli that were detected in >1% of gut metagenomes. Tip labels and colors indicate order-level clades from the GTDB (A=Acholeplasmatales; M=ML615J-28; H= Haloplasmatales). RF39 has a highly reduced genome with numerous metabolic auxotrophies.

METHODS

Publicly available human gut metagenomes. We downloaded 11,523 sequencing runs for publicly available human gut metagenomes from the NCBI SRA[39]. These data correspond to 3,810 samples, 15 studies[9,21,40–52], and >181 billion sequencing reads with an average length of 100 bp (Supplementary Tables 1–2). Sequencing metadata was obtained from the SRAdb relational database[53] and host metadata was obtained from either the NCBI BioSample database[54] or from supplementary datasets linked to publications (Supplementary Table 3). No metadata was available online or upon request from the Fiji cohort[9]; these individuals were treated as healthy adults from a rural population.

Metagenome assembly and binning. We co-assembled the 11,523 sequencing runs for each of the 3,810 biological samples using MegaHIT v1.1.1[55] with default parameters. This resulted in 333,661,332 contigs longer than 200 bp, totaling 453.5 × 10^9 bp, with an average N50 of 12,460 bp (Supplementary Table 2). Human gut MAGs were generated per-sample using three different tools with default options: MaxBin v2.2.4[56], MetaBAT v2.12.1[57], and CONCOCT v0.4.0[10], which all utilize a combination of sequence composition and coverage information. DAS Tool v1.1.0[58] with option '-score_threshold 0' was used to integrate MAGs produced by the three tools. Contigs shorter than 1 Kbp were discarded. This process resulted in 152,591 MAGs longer than 100 Kbp, which totaled 73,632,219 contigs (22% of total assembled) and 310.7 × 10^9 bp (69% of total assembled). All MAGs were screened for contamination against the human genome (build 38) and phiX genome with BLASTN v2.6.0[59].

Refinement of MAGs based on alignment of contigs between conspecific genomes. To refine MAGs from the HGM dataset, we performed pairwise alignment of contigs between MAGs and other closely-related, near-complete MAGs and reference genomes (Supplementary Table 6). Our logic was that strains of the same species should share homology between most contigs, and contigs failing this condition (i.e. present in one genome but absent in the other) likely represent contamination. For each input MAG, we used Mash v2.0[60] to find at least five closely-related, near-complete genomes in the database (>95% estimated completeness, <5% estimated contamination, Mash distance ≤0.05, p value ≤0.001) and then used BLASTN to align contigs between each MAG and all target genomes. Contigs in the MAG that failed to align at ≥70% nucleotide identity over ≥25% length to any of the closely related genomes were flagged for removal.

Refinement of MAGs based on taxonomic annotation of contigs. We identified and removed taxonomically discordant contigs from MAGs using two complementary approaches (Supplementary Table 6). The first approach performs taxonomic annotation based on universal single-copy marker genes. HMMs for marker gene families were downloaded from the PhyEco database[61] and searched against MAGs with HMMER v3.1b2[62]. A subset of 100 gene families were used for Archaea and 88 for Bacteria. Marker genes found in MAGs were then aligned against a reference database of taxonomically annotated marker genes from reference genomes using BLASTP. For each gene, we transferred the taxonomy of the best hit in the reference database at the appropriate rank based on % amino acid identity cutoffs specific to each gene family at each rank. We then taxonomically annotated each MAG based on the consensus taxonomy of marker genes at the lowest rank such that >70% of marker genes were annotated. Contigs were flagged for removal if they (1) contained a taxonomically discordant marker gene, and (2) lacked a concordant one. The second approach for taxonomic refinement is similar to the first, except that 855,764 clade-specific prokaryotic marker genes from the MetaPhlAn 2[27] database were used for taxonomic annotation after excluding "pseudo markers" that are not perfectly unique to a clade.

Refinement of MAGs based on outlier nucleotide composition and sequencing read-depth. Using an approach similar to Parks et al.[11], we identified and removed contigs from MAGs with either (1) outlier GC content, (2) outlier tetranucleotide frequency (TNF), or (3) outlier sequencing read-depth (Supplementary Table 6). We used principal component analysis to reduce the TNF dimensionality down to the first principal component (TNF PC1). For each MAG, we then measured the average GC content, average TNF PC1, and average sequencing read-depth. Contigs were flagged for removal if they deviated from these averages beyond cutoffs selected to minimize reduction in completeness (Supplementary Table 6).

Validation of MAG refinement pipeline. We simulated 1,000 human gut MAGs to validate our overall MAG refinement strategy (Supplementary Table 7). Each simulated MAG contained two genomes: one "host" genome, representing the target genome, and one "donor" genome, representing the contaminating genome. All 102 genomes used in simulations were isolated from the human gut, and were estimated to have >95% completeness, <1% contamination, and <25 contigs. MAGs were simulated with completeness (mean=61.9%), contamination (mean=10.0%), and N50 (mean= 35.8 Kbp) based on randomly sampled MAGs from the HGM dataset. MAGs were dropped in cases where contamination exceeded completeness, and thus the host genome was in the minority. The refinement pipeline was applied to each simulated MAG, and to evaluate the pipeline, we quantified the overall reduction in completeness and contamination (Extended Data Fig. 2a,b).

Application of refinement strategies to the HGM dataset. We applied each of the refinement approaches described above to the MAGs (Supplementary Table 6 and Extended Data Fig. 2c). In rare cases, these approaches may erroneously flag a large proportion of a MAG. To avoid this, we only applied a particular approach to a MAG if it resulted in ≤25% reduction in total length. Altogether, the five approaches removed 5,251,859 contigs (7.13% of total) and 20,821.2 Mb (6.70% of total) from the MAGs. After removing potential contaminants, we were left with 152,279 MAGs with a total length ≥100 Kbp and 10,036 individual contigs longer than 100 Kb that were either unbinned or removed during decontamination. These long contigs were included with other MAGs bringing the total number to 162,315.

Quality assessment of MAGs. CheckM v1.0.7[13] was used to estimate completeness and contamination of the 162,315 recovered MAGs (Supplementary Table 5), which is based on the copy-number of lineage-specific single-copy genes. Additional statistics were obtained for each genome, including: the contig N50, number of contigs, average contig length, contig read-depth, and number of tRNA and rRNA genes. tRNAs were identified using tRNAscan-SE v.1.3.1[63] and rRNA genes using Barrnap v0.9-dev[64] with options '–reject 0.01 –evalue 1e-3'. We identified 60,664 MAGs which met the MIMAG medium-quality criteria of ≥50% complete with ≤10% contamination[14]. For analyses requiring near-complete genomes, we used a subset of 24,345 high-quality MAGs that were ≥ 90% complete, ≤5% contaminated, with an N50 ≥10 Kb, an average contig length ≥5 kb, ≤500 contigs, and ≥90% of contigs with ≥5x read-depth.

Estimation of SNP density. Read mapping and SNP calling were performed to assess the genetic diversity of each MAG (Supplementary Table 5). Bowtie 2 v2.3.4[65] was used to construct a database of MAGs for each sample and align metagenomic reads. Reads with low mapping and sequence quality were discarded (quality scores <20 and <30, respectively) and we counted the occurrence of nucleotides with quality ≥30 across each MAG. To compare SNPs between MAGs sequenced to different depth, we down sampled each MAG to 40 mapped reads per site. MAGs with at least 200,000 sites of ≥40x depth were retained for analysis. A SNP was called if at least 2 reads matched the alternative allele at a genomic site. SNP density was calculated as the number of SNPs per kilobase.

Reference genomes used for comparison. We downloaded 201,102 publicly available bacterial and archaeal reference genomes from IMG (N=61,713) and PATRIC (N=139,389) on Jan 16, 2018. These included genomes from two human gut culturomics studies[6,7] and 16,525 previously published MAGs, including a previous MAG study from the human gut[20] and nearly 8,000 MAGs assembled from SRA metagenomes[11]. To remove redundancy within and between databases, we used Mash[60] with default parameters to cluster genomes with a Mash distance of 0.0, which are expected to be identical. This resulted in 153,900 non-redundant reference genomes, of which 127,419 were classified as high-quality, 18,498 as medium-quality, and another 7,983 as low-quality (Supplementary Table 9).

Species-level clustering of reference genomes and MAGs. Using an approach similar to Olm et al.[66], we clustered the 60,664 MAGs and 145,917 reference genomes meeting or exceeding the MIMAG medium-quality standard into species-level operational-taxonomic-units (OTUs) on the basis of 95% whole-genome ANI (Supplementary Table 10). We first performed single-linkage clustering of genomes based on a Mash ANI of 99%, resulting in 79,675 clusters that can be confidently assigned to the same species-OTU. Mash is extremely fast, although it can underestimate ANI for incomplete genomes[66]. To address this, we used the ANIcalculator v1.0[67] to compute gANI between the 99% identity clusters and required that at least 20% of genes were aligned. The 20% cutoff was chosen to minimize the negative impact of incomplete genomes and avoid formation of spurious OTUs (Extended Data Fig. 5a). To increase computational efficiency, we only calculated gANI for genome-pairs with >90% Mash ANI. Genomes were clustered into OTUs using average-linkage hierarchical clustering with a 95% gANI cutoff using the package MC-UPGMA v1.0.0[68], yielding 23,790 OTUs.

All OTUs were taxonomically annotated using the tool GTDBTk v0.0.6 (release 80, github.com/Ecogenomics/GtdbTk), which produces standardized taxonomic labels based on the Genome Taxonomy Database[26]. Additionally, we constructed pan-genomes based on clustering all genes within each OTU using VSEARCH v2.4.3[69] with 90% DNA identity and 50% alignment cutoffs (maximum 500 genomes per OTU). Human gut OTUs were identified from the set of 23,790 OTUs on the basis of (1) containing a MAG from the HGM dataset, (2) being detected by IGGsearch (see section "Development of IGGsearch for metagenomic profiling of species-OTUs") in at least one of 3,810 metagenomes used for MAG recovery (see below), or (3) containing a genome isolated from the human gut (Supplementary Table 10 and Extended Data Fig. 6a,b). A total of 4,558 species-OTUs were annotated as human-gut based on the combination of the three criteria.

Phylogenetic analysis of MAGs and reference genomes. We constructed phylogenetic trees of MAGs and reference genomes using concatenated alignments of conserved, single-copy marker gene families from the PhyEco database[61] for Bacteria (N=88 genes) and Archaea (N=100 genes). Individual marker genes were identified using HMMER v3.1b2 with gene-family-specific bit-score cut-

offs. For computational efficiency, genomes were collapsed down to species-OTUs that were represented as individual leaves in the phylogenetic tree. To reduce the effect of contamination, taxonomically discordant marker genes were removed, as previously described in "Refinement of MAGs based on taxonomic annotation of contigs". FAMSA v1.2.5[70] was used to construct protein-based multiple sequence alignments for each gene family. Columns with >15% gaps were removed, alignments were concatenated together, and sequences with >70% gaps were removed (N=39). FastTree2 v.2.1.10[71] was used to build a maximum likelihood phylogeny for Bacteria and Archaea with default options. All trees were visualized using iTOL v3[72]. To quantify the gain in phylogenetic diversity (PD) from the HGM dataset, we computed the total branch length of two subtrees: a tree of all 4,558 gut OTUs (PD_{Gut}) and a tree of 2,500 gut OTUs with reference genomes (PD_{RefGut}). The percent gain in phylogenetic diversity was computed as: $100 \times (PD_{Gut} - PD_{RefGut}) / PD_{RefGut}$. To identify OTUs for higher-ranking groups, we performed average-linkage hierarchical clustering of phylogenetic distances, which was implemented in R (Supplementary Table 10). Rank-specific cutoffs were identified by maximizing similarity to the Genome Taxonomy Database for reference genomes (Extended Data Fig. 3d,e).

Development of IGGsearch for metagenomic profiling of species-OTUs. Using an approach similar to MetaPhlAn 2[27], we developed an accurate and efficient tool for quantifying the abundance of species-OTUs from unassembled metagenomes. First, we identified marker genes for each OTU (Supplementary Fig. 1a). Up to 300 genes from the pan-genome of each OTU were selected with the maximum intra-OTU frequency and minimum inter-OTU frequency. The intra-OTU frequency was computed as the fraction of genomes within an OTU where a gene was found at 90% DNA identity. The inter-OTU frequency was determined based on DNA alignments (using HS-BLASTN v0.0.5[73]) between each gene and the pan-genomes of other OTUs, and accounts for: (1) the number of other pan-genomes where the gene is found, (2) the frequency of the gene in each pan-genome, and (3) the % identity of each alignment. For computational reasons, genes were first aligned within each phylum, and only the 300 top scoring candidates per OTU were subsequently checked for uniqueness between phyla. A total of 6,198,663 marker genes were identified for 21,790 OTUs.

A large number of OTUs contained just a single genome, making it difficult to accurately predict conserved genes. To refine our marker gene set, we utilized abundance co-variation information, which is a common strategy for binning genetic regions from the same species and has been applied previously[3,10,20,21,57]. Specifically, we performed read-mapping of the 3,810 metagenomic samples to the database of 6,198,663 marker genes using Bowtie 2 v2.3.4 and quantified the read-depth of each gene in each sample. We used average linkage clustering to group genes from each OTU into co-variance groups on the basis of Pearson correlations of read-depth across samples (Supplementary Fig. 1b). After applying a correlation threshold of 0.90, we selected the largest cluster of genes for the final marker gene set. This procedure removed 55,132 genes for 1,402 OTUs that were present in ≥10 samples with ≥1x coverage.

IGGsearch is a command-line tool that uses Bowtie 2 to map metagenomic reads to the database of marker-genes and quantify species-OTUs. Read alignments are removed with low % identity (minimum=95%), alignment coverage (minimum=70% of read), and base quality (minimum=20). For each metagenomic sample, OTU relative abundance is estimated by taking the average read-depth across marker genes and normalizing these values to 1.0 across all OTUs. Species presence is determined based on the % of marker genes with at least one mapped read.

The sensitivity and specificity of IGGsearch was evaluated on two benchmark datasets. First, we benchmarked IGGsearch on the CAMI challenge dataset (https://data.cami-challenge.org/participate; Supplementary Tables S11–12 and Supplementary Fig. 2a). Second, we benchmarked IGGsearch on simulated gut metagenomes that contained between 500K and 50M paired-end reads, read length of 100 bp, Illumina-style sequencing error, and one genome from each of 100 randomly selected gut species-OTUs (Supplementary Fig. 2b). Based on these benchmarks, we called OTUs present when at least 15% of their marker genes were detected, which gave a good balance between sensitivity and specificity.

Metagenome-wide association of species abundance with disease. We used IGGsearch species profiles to identify species-OTUs associated with disease for ten previously published studies, including: colorectal cancer[43], type 2 diabetes[21,44], rheumatoid arthritis[42], Parkinson's disease[74], atherosclerotic cardiovascular disease[75], ankylosing spondylitis[76], non-alcoholic fatty liver disease[77], liver cirrhosis[78], and obesity[79] (Extended Data Table 1 and Supplementary Tables 15–16). To identify species-disease associations, we compared species relative abundances for the 4,558 human gut species-OTUs between cases and healthy controls using the Wilcox rank-sum test. Non-gut OTUs were excluded to reduce the impact of multiple hypothesis testing. For each disease, *p* values were corrected for multiple hypothesis tests using the Benjamini-Hochberg procedure. We performed the same statistical procedure using species profiles from three other tools, including: MIDAS v1.3.0[4], MetaPhlAn2 v2.7.7[27], and mOTU v1.1.1[3]. All tools were run with default parameters and the distributed reference data. To prevent confounding signals due to disease treatment, we excluded 100 individuals taking drugs that affect microbiome composition, including metformin in T2D patients[21,44], acarbose, atorvastatin, fondaparinux, and metoprolol in ACVD patients[75], and antirheumatic drugs in rheumatoid arthritis patients[42].

Machine learning models for disease prediction. We constructed Random Forest (RF) models to predict disease state from species abundance profiles generated with IGGsearch, MIDAS, mOTU, and MetaPhlAn2 (Extended Data Table 1). For IGGsearch, we included all 23,790 species OTUs and allowed the RF to choose the most predictive OTUs. RF models were implemented in the scikit-learn package v0.19.1[80] and were optimized for each of the four tools for each of the 10 diseases. Specifically, we tested 1,000 random combinations of parameter values for 1) the number of trees in the forest, 2) the number of features to consider at each split, 3) the maximum number of levels in each tree, 4) the minimum number of samples to split a node, 5) the minimum number of samples at each leaf, and 6) whether to use bootstrapping during model training. To avoid overfitting, each model was evaluated using 10-fold cross-validation and the combination of parameters yielding the best receiver operating curve (ROC) area under the curve (AUC) was selected. To obtain robust estimates of model performance, all models were re-run 100 times and ROC AUC values were averaged across runs.

Identifying genomic features and auxotrophies of uncultured gut bacteria. We selected a subset of 504 human gut species-OTUs from Bacteria for comparative genomic analysis between cultured and uncultured organisms (Supplementary Table 17). OTUs with <5% prevalence in human gut metagenomes were excluded since rare organisms may be amenable to cultivation but not yet sampled. Uncultivated OTUs were defined as those containing only MAGs (either from the current study or previous studies, N=271) and cultivated OTUs as those containing at least one isolate genome (N=233). We based all comparative analysis between OTUs using 24,345 high-quality MAGs from the HGM dataset, which was done (1) to avoid biases resulting from a comparison of MAGs to isolate genomes (which differ in assembly quality) and (2) to avoid issues arising from low completeness among MAGs in the medium-quality tier.

We compared several broad genomic features between groups, including: estimated genome size, GC content, coding density, and estimated replication rate. Estimated genome size was corrected for completeness and contamination using: $\hat{G} = G * 100 \,/\, \hat{C} - (G * \hat{T} \,/\, 100)$, where $\hat{G}$ is the estimated genome size of a MAG, G is the observed genome size, $\hat{C}$ is the estimated percent completeness, and $\hat{T}$ is the estimated percent contamination. Replication rate was estimated with iRep v1.10[28] for MAGs with >5x read-depth, which is based on differences in sequencing depth between the origin and terminus of replication. Genomic features were averaged across all high-quality MAGs for each OTU and then compared between OTUs using the Wilcoxon rank-sum test (Supplementary Table 18).

To identify potential auxotrophies, we compared the prevalence of genes, modules, and pathways from the KEGG database (release 77.1)[81] between cultivated and uncultivated OTUs. Proteins from high-quality MAGs were annotated based on amino acid alignments to KEGG using LAST v828[82] and assigned to the KEGG orthology group (KO) with lowest E-value <1e-5. Next, we computed the fraction of MAGs per OTU containing each KO and compared these values between OTUs using the Ives-Garland test implemented in the phylolm R package v.2.6[83]. The Ives-Garland test performs logistic regression while controlling for differences in phylogeny between groups and has been previously applied to microbiome data[84]. This analysis was repeated for modules and pathways from the KEGG database. *P* values were corrected for multiple hypothesis tests using the Benjamini-Hochberg procedure (Supplementary Table 19). This same analysis was also performed for functions from the TIGRFAM database (release 15.0)[85] (Extended Data Fig. 9a).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

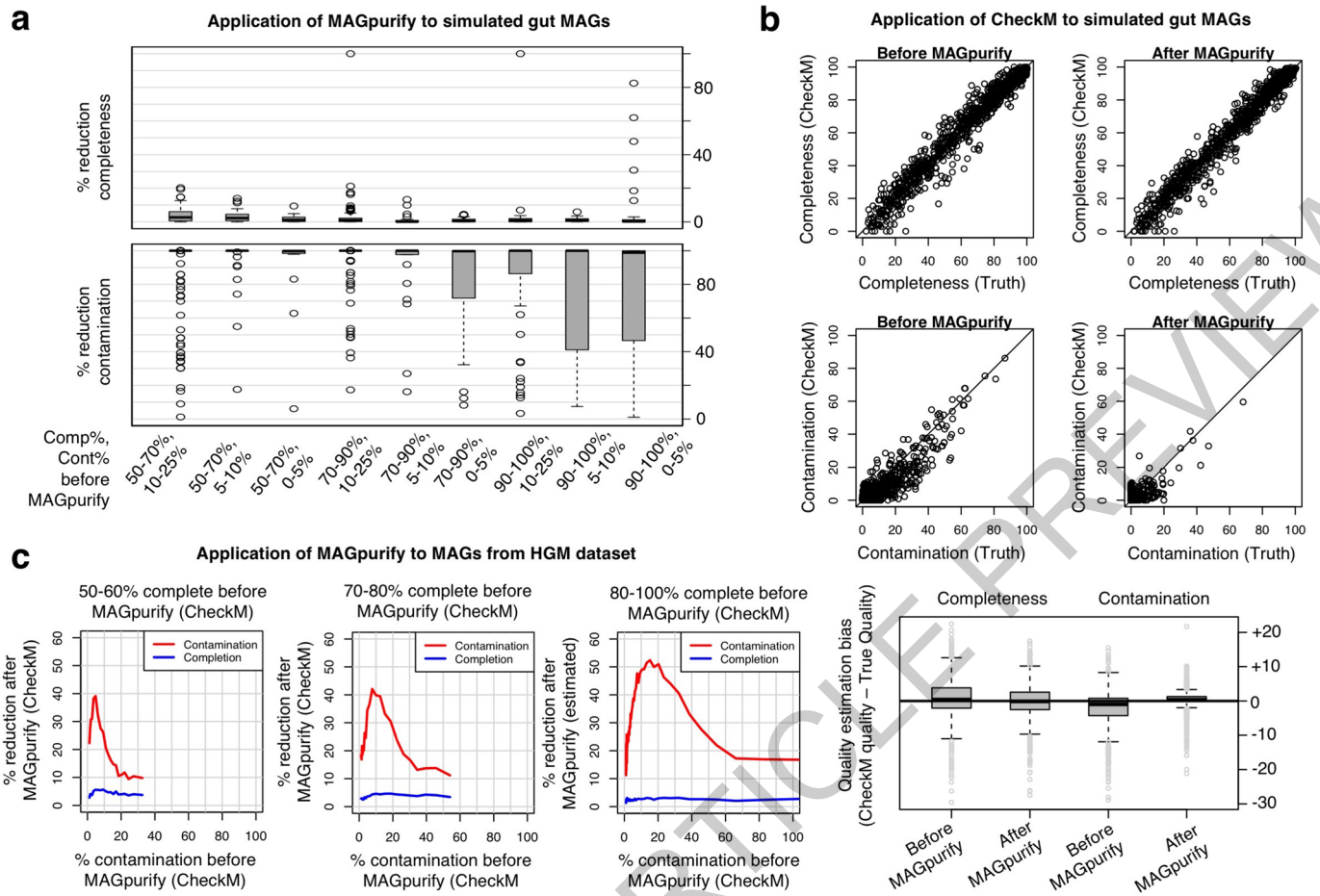
Code availability. IGGsearch and the database of conserved species-specific marker genes are freely available online at https://github.com/snayfach/IGGsearch. The code for removing contamination from genome bins, MAGpurify, is available at https://github.com/snayfach/MAGpurify.

Data availability

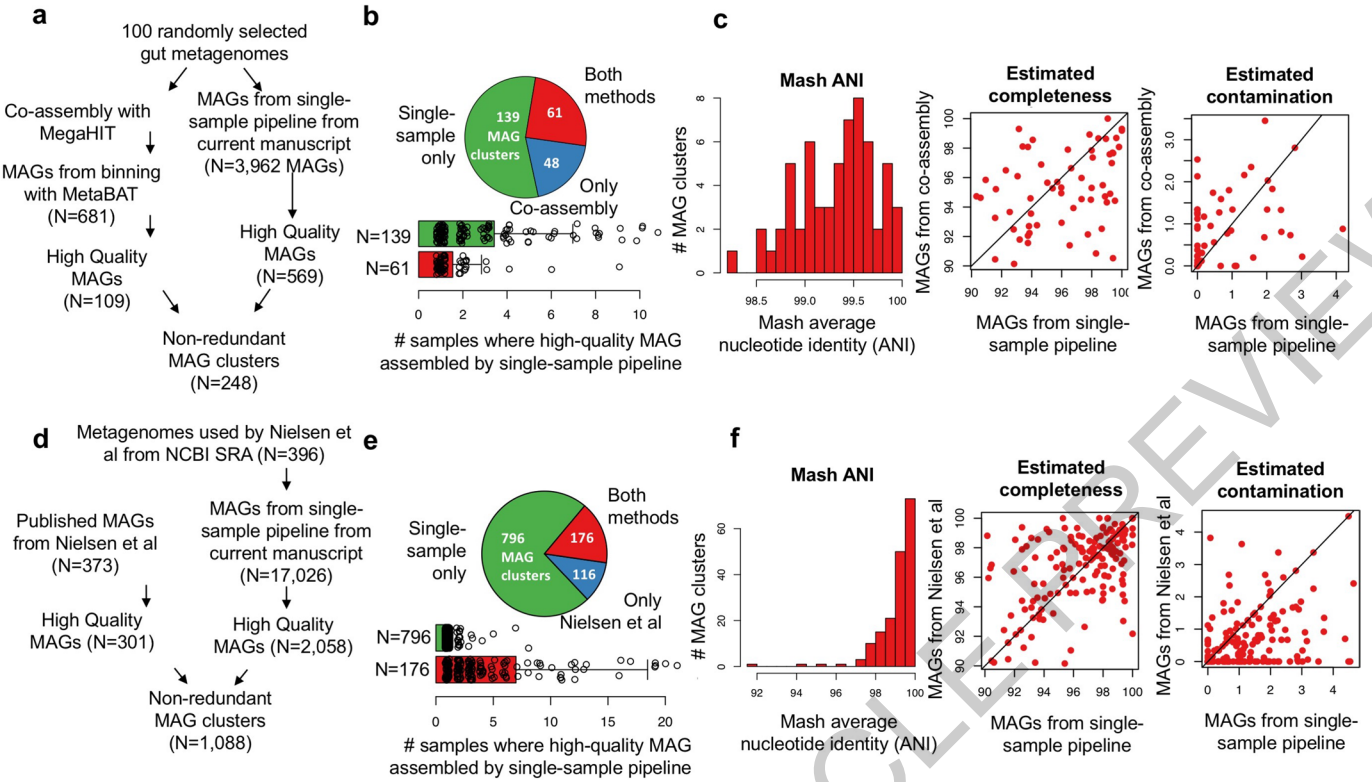
Representative MAGs for the 2,058 new species have been deposited in the European Nucleotide Archive (ENA) under accession PRJEB31003 (Supplementary Table 20). The entire HGM dataset, phylogenomic trees, and related metadata is freely available at https://github.com/snayfach/IGGdb.

39. Leinonen, R., et al., The sequence read archive. *Nucleic Acids Res*, 2011. **39**(Database issue): p. D19-21.
40. Consortium, H., Structure, function and diversity of the healthy human microbiome. *Nature*, 2012. **486**(7402): p. 207-14.
41. Li, J., et al., An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol*, 2014. **32**(8): p. 834-41.

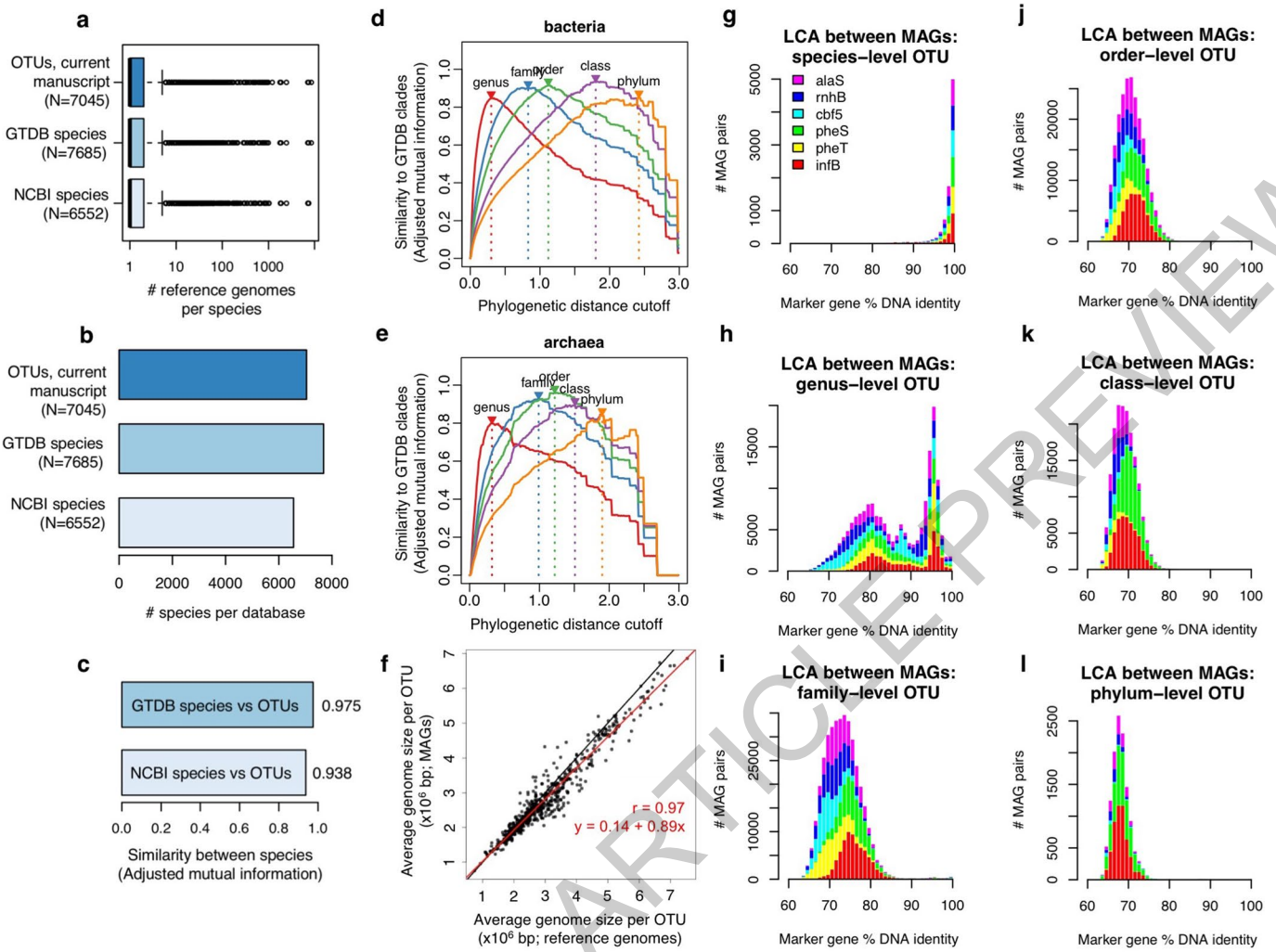
42. Zhang, X., et al., The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med*, 2015. **21**(8): p. 895-905.
43. Feng, Q., et al., Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun*, 2015. **6**: p. 6528.
44. Karlsson, F.H., et al., Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 2013. **498**(7452): p. 99-103.
45. Obregon-Tito, A.J., et al., Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat Commun*, 2015. **6**: p. 6505.
46. Rampelli, S., et al., Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr Biol*, 2015. **25**(13): p. 1682-93.
47. Vatanen, T., et al., Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell*, 2016. **165**(4): p. 842-53.
48. Backhed, F., et al., Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe*, 2015. **17**(5): p. 690-703.
49. Liu, W., et al., Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. *Sci Rep*, 2016. **6**: p. 34826.
50. Smits, S.A., et al., Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science*, 2017. **357**(6353): p. 802-806.
51. Pehrsson, E.C., et al., Interconnected microbiomes and resistomes in low-income human habitats. *Nature*, 2016. **533**(7602): p. 212-6.
52. *Infant Gut Metagenomics Initiative, Broad Institute (broadinstitute.org).*
53. Zhu, Y., et al., SRAdb: query and use public next-generation sequencing data from within R. *BMC bioinformatics*, 2013. **14**(19).
54. Barrett, T., et al., BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res*, 2012. **40**(Database issue): p. D57-63.
55. Li, D., et al., MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 2015. **31**(10): p. 1674-6.
56. Wu, Y.W., B.A. Simmons, and S.W. Singer, MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 2016. **32**(4): p. 605-7.
57. Kang, D.D., et al., MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 2015. **3**: p. e1165.
58. Sieber, C.M.K., et al., Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*, 2018. **3**(7): p. 836-843.
59. Altschul, S.F., et al., Basic local alignment search tool. *Journal of Molecular Biology*, 1990. **215**(3): p. 403-410.
60. Ondov, B.D., et al., Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*, 2016. **17**(1): p. 132.
61. Wu, D., G. Jospin, and J.A. Eisen, Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLoS One*, 2013. **8**(10): p. e77033.
62. Eddy, S.R., Accelerated Profile HMM Searches. *PLoS Comput Biol*, 2011. **7**(10): p. e1002195.
63. Lowe, T.M. and S.R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, 1997. **25**(5): p. 955-64.
64. Seemann, T., *barrnap 0.8 : rapid ribosomal RNA prediction*. 2013.
65. Langmead, B. and S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 2012. **9**(4): p. 357-9.
66. Olm, M.R., et al., dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*, 2017. **11**(12): p. 2864-2868.
67. Varghese, N.J., et al., Microbial species delineation using whole genome sequences. *Nucleic Acids Res*, 2015. **43**(14): p. 6761-71.
68. Loewenstein, Y., et al., Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics*, 2008. **24**(13): p. i41-9.
69. Rognes, T., et al., VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 2016. **4**: p. e2584.
70. Deorowicz, S., A. Debudaj-Grabysz, and A. Gudys, FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Sci Rep*, 2016. **6**: p. 33964.
71. Price, M.N., P.S. Dehal, and A.P.A. Arkin, FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 2010. **5**(3).
72. Letunic, I. and P. Bork, Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*, 2016. **44**(W1): p. W242-5.
73. Chen, Y., et al., High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Res*, 2015. **43**(16): p. 7762-8.
74. Bedarf, J.R., et al., Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naive Parkinson's disease patients. *Genome Med*, 2017. **9**(1): p. 39.
75. Jie, Z., et al., The gut microbiome in atherosclerotic cardiovascular disease. *Nat Commun*, 2017. **8**(1): p. 845.
76. Wen, C., et al., Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol*, 2017. **18**(1): p. 142.
77. Loomba, R., et al., Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. *Cell Metab*, 2017. **25**(5): p. 1054-1062 e5.
78. Qin, N., et al., Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 2014. **513**(7516): p. 59-64.
79. Le Chatelier, E., et al., Richness of human gut microbiome correlates with metabolic markers. *Nature*, 2013. **500**(7464): p. 541-6.
80. Pedregosa, F.a.V., G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011. **12**: p. 2825-2830.
81. Kanehisa, M. and S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 2000. **28**(1): p. 27-30.
82. Kielbasa, S.M., et al., Adaptive seeds tame genomic sequence comparison. *Genome Res*, 2011. **21**(3): p. 487-93.
83. Ho, L. and C. Ane, A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst Biol*, 2014. **63**(3): p. 397-408.
84. Bradley, P.H., S. Nayfach, and K.S. Pollard, Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. *PLoS Comput Biol*, 2018. **14**(8): p. e1006242.
85. Haft, D.H., et al., TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res*, 2013. **41**(Database issue): p. D387-95.



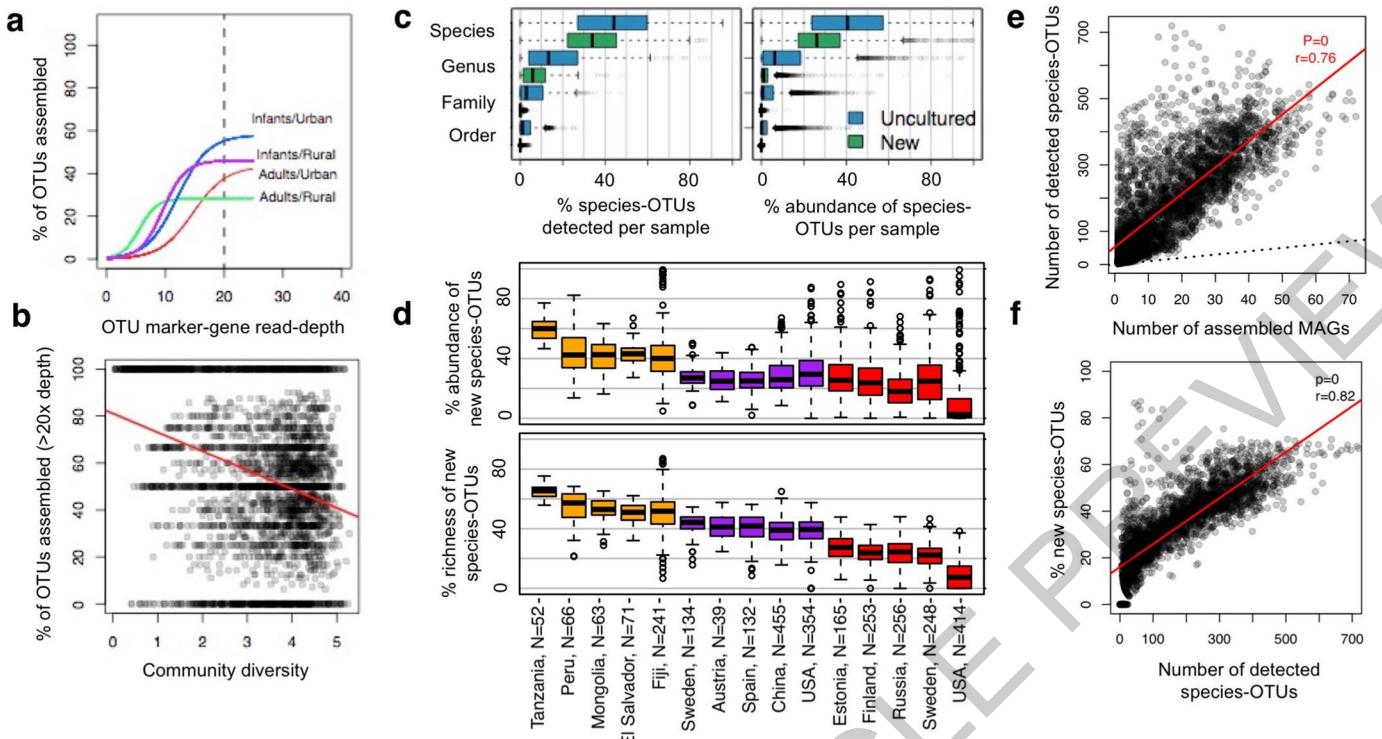
Extended Data Fig. 1 | MAGpurify removes contamination, maintains completeness, and does not result in biased estimates of genome quality. A–B) 1,000 human gut MAGs were simulated to validate the MAGpurify pipeline. Each MAG contained two genomes: one "host" genome, representing the target genome, and one "donor" genome, representing the contaminating genome (Supplementary Table 7). All 102 input genomes were isolated from the human gut, and were estimated to have >95% completeness, <1% contamination, and <25 contigs. MAGs were simulated with completeness, contamination, and N50 based on randomly sampled MAGs from the HGM dataset. 65 MAGs were dropped from the analysis where contamination exceeded completeness, and thus the host genome was in the minority. **A**) The boxplots indicate the percent reduction in completeness (top) and contamination (bottom) after applying MAGpurify. Regardless of initial quality, MAGpurify sensitively removed contamination for most MAGs while avoiding removal of the host genome. **B**) CheckM was applied to simulated MAGs before and after applying MAGpurify. (Top) The scatterplots show that true genome quality is correlated with the estimated genome quality before and after applying MAGpurify. Black lines indicate the line of equality. (Bottom) The distribution of differences between true and estimated quality is centered at zero, indicating no bias applying MAGpurify. **C**) MAGpurify was applied to all MAGs from the HGM dataset. The figure shows the reduction in CheckM quality estimates before and applying MAGpurify. Estimated quality improvement is greatest when completeness is between 90 to 100% and contamination is between 10 to 30%. In all box plots, middle line denotes median; box denotes interquartile range (IQR); and whiskers denote 1.5× IQR.



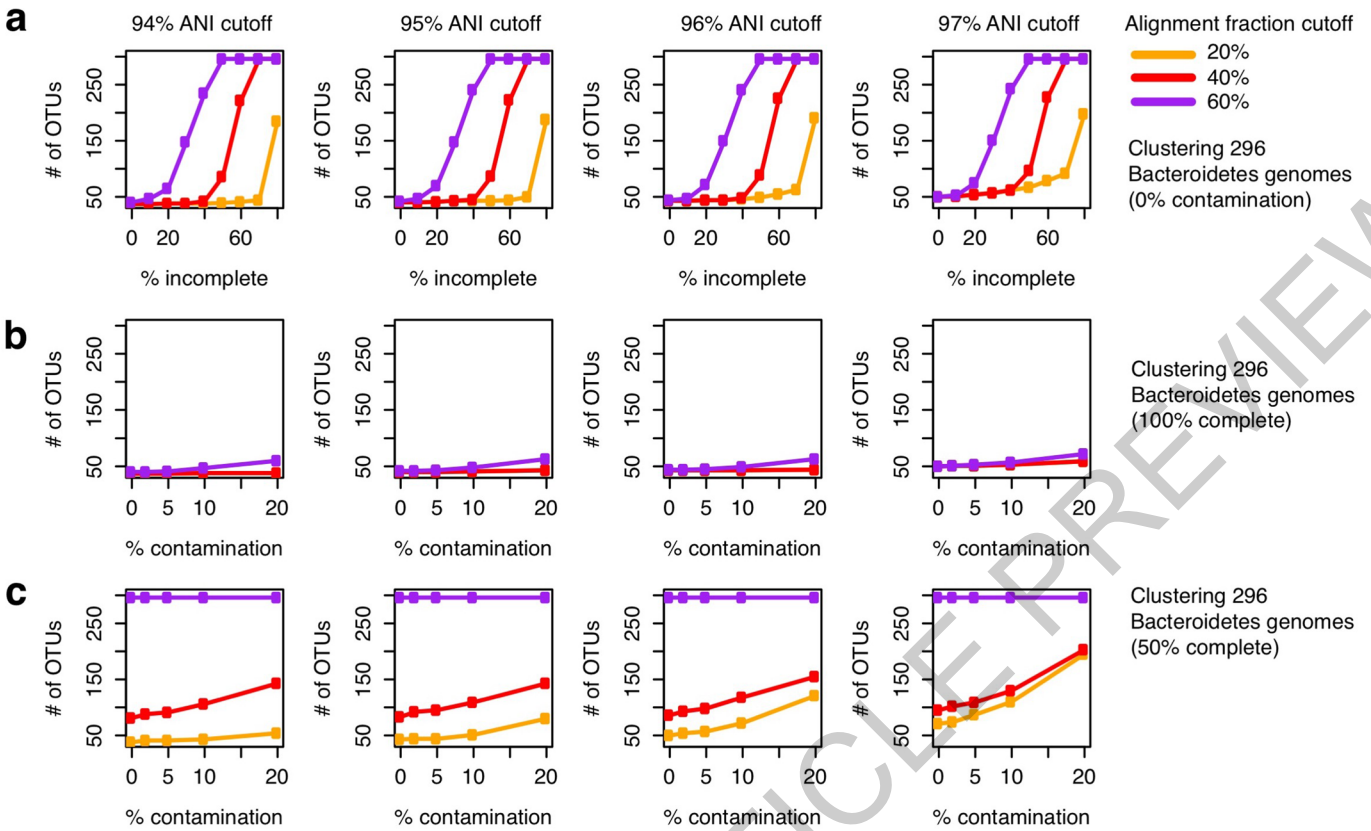
Extended Data Fig. 2 | Single-sample assembly and binning yields more non-redundant, high-quality MAGs compared to other approaches. **A**–**C**) Comparison to co-assembly and binning. **A**) 100 randomly selected human gut metagenomes were co-assembled with MegaHIT (v1.1.4, options: --k-min 27 --k-max 127 --k-step 10 --kmin-1pass --continue), taking 3,608 CPU hours. Reads from each sample were mapped back to the co-assembly to quantify the read-depth of each contig in each sample. This information was used as input to MetaBAT (v.2.12.1, default options) to generate MAGs. Other binning programs, including CONCOCT and MaxBin2 did not complete due to the large size of the assembly. MAGs from the single-sample pipeline were grouped with MAGs from the co-assembly using Mash at 90% ANI to form 248 clusters. **B**) A large fraction of clusters is exclusively represented by MAGs from the single-sample pipeline. These clusters tend to be found in multiple samples, which may interfere with co-assembly. For bar plots, the center bar indicates the mean; the error bar indicates the standard deviation; and all data points are overlaid. **C**) The MAGs recovered by both pipelines (N=61) have high ANI, indicating they are very similar genomes, and tend to have similar levels of estimated completeness and contamination, as determined by CheckM. Black lines indicate the line of equality. **D**–**F**) Comparison to co-abundance binning performed by Nielsen et al.[20]. **D**) MAGs from the single-sample pipeline were grouped with MAGs from Nielsen et al. using Mash at 90% ANI to form 1088 clusters. **E**) A large fraction of clusters is only represented by MAGs from the single-sample pipeline, which tend to be restricted to individual metagenomes, which may be explained by the fact the Nielsen method requires MAGs to be present in multiple samples to accurately quantify co-variation and bin contigs. For bar plots, the center bar indicates the mean; the error bar indicates the standard deviation; and all data points are overlaid. **F**) The MAGs recovered by both pipelines (N=176) have high ANI, indicating they are very similar genomes, and tend to have similar levels of estimated completeness and contamination, as determined by CheckM. Black lines indicate the line of equality.



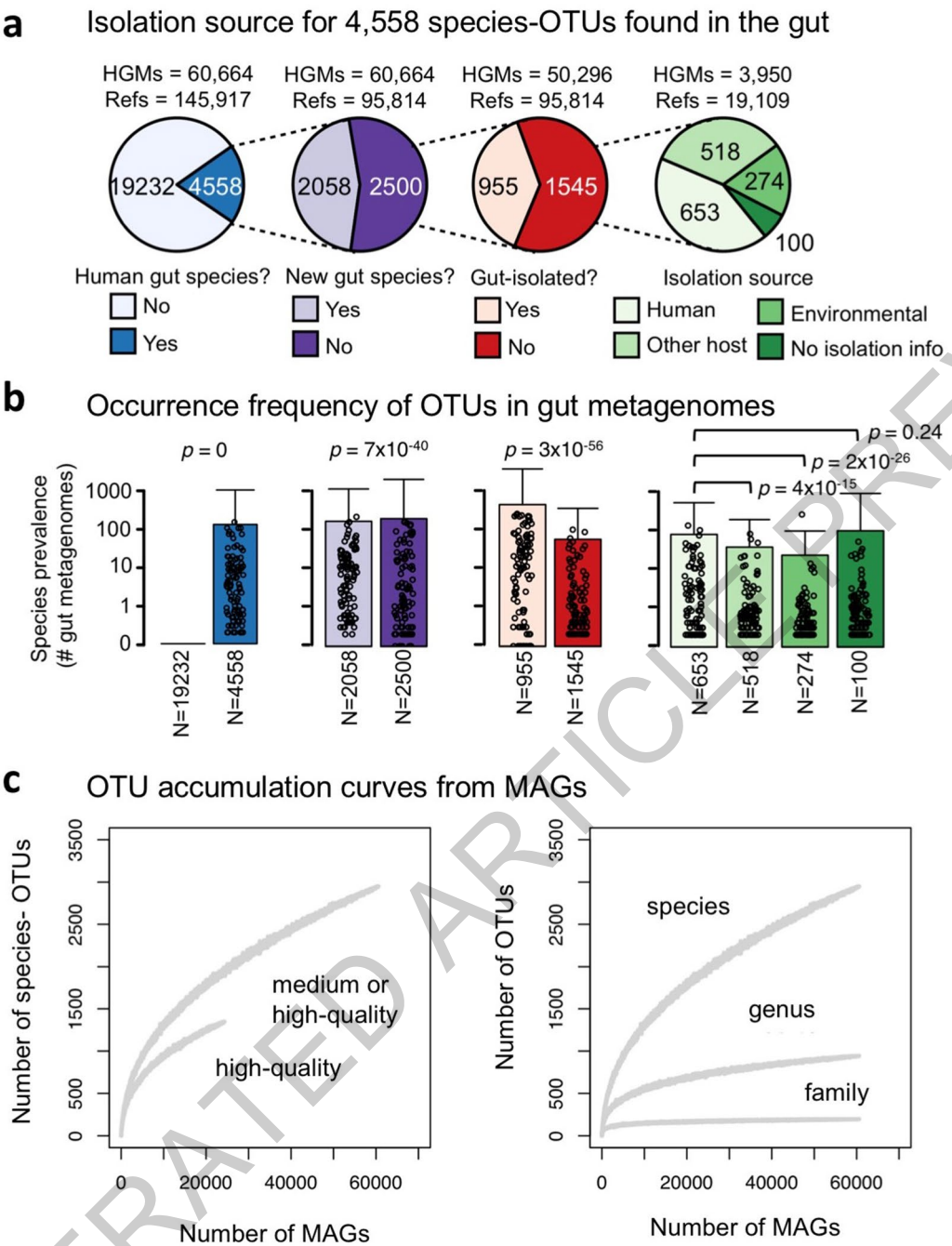
Extended Data Fig. 3 | Additional checks of MAG quality after clustering genomes into OTUs. A–C) MAGs and reference genomes were clustered into species-OTUs based on 95% ANI. As validation, OTUs were compared to the NCBI and Genome Taxonomy Database (GTDB) for 65,900 reference genomes with valid species names. **A**) Box plots of the number of genomes per species, where middle line denotes median; box denotes IQR; and whiskers denote 1.5× IQR. **B**) The number of species per database. **C**) Similarity between OTUs and other databases, as measured using the adjusted mutual information statistic (AMI). Species-OTUs are concordant with the NCBI and GTDB taxonomies. **D–E**) MAGs and reference genomes were further clustered into higher-ranking OTUs on the basis of phylogenetic distance cutoffs. Rank-specific cutoffs were identified that maximized similarity to the GTDB. **F**) As an additional indicator of completeness, genome sizes of high-quality MAGs and reference genomes from the same OTU were compared. Each point indicates one species-OTU (N=625). A positive slope of close to 1.0 indicates to systematic loss of gene content. **G–L**) As an additional check of contamination, six single-copy marker genes (alaS, rnhB, cbf5, pheS, pheT, infB) were aligned between MAGs using BLASTN. MAGs devoid of contamination should display high % identity from the same OTU, and low % identity between different OTUs. The 6 markers genes were selected on the basis of (1) present in >90% of high-quality MAGs and reference genomes at single copy, and (2) have species-level % DNA identity cutoffs <98%. Highly conserved genes may be similar between different OTUs and were not suitable for this analysis. For between-OTU comparisons we used one MAG for each of 2,962 species-OTUs. For within-OTU comparisons, we used two MAGs for each of 1,616 species-OTUs. The histograms indicate the distribution of DNA % identity between MAGs from (**G**) the same species-OTU (i.e. lowest common ancestor, LCA = species) and (**H–L**) between MAGs that are more distantly related (lowest common ancestor, LCA = genus, family, order, class, or phylum). The vast majority of genes from the same species-OTU display >98% identity while those from different OTUs display <98% identity.



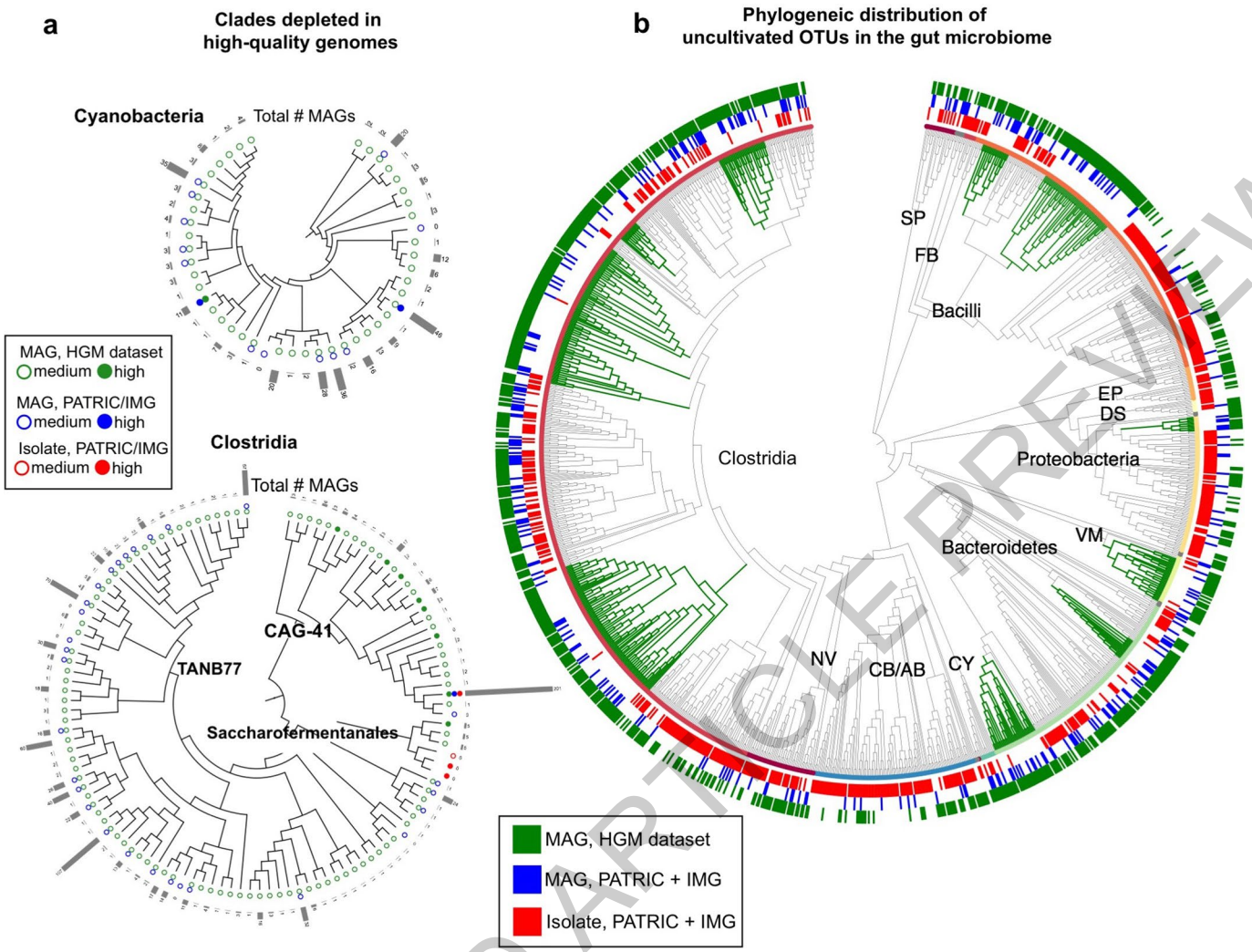
Extended Data Fig. 4 | Assembly and distribution of MAGs across human populations. IGGsearch was applied to 3,083 metagenomes from healthy individuals that were used for assembly and binning in order to estimate the abundance of human gut OTUs per sample. **A**–**B**) The overall assembly rate was computed at each read-depth, defined as the % of detected OTUs with an assembled MAG. **A**) Curves were fit using logistic regression. Conditioning on read-depth, MAGs are recovered more readily from an infant metagenome compared to an adult metagenome from a rural population. **B**) The x-axis indicates the Shannon diversity of each of the 3,810 metagenomic samples, and the y-axis indicates the MAG recovery rate for OTUs with >20x depth. MAGs are recovered less often from a high diversity community, even when read-depth is sufficiently high (Pearson's ρ=-0.31, p value=4.3x10^{-75}). **C**) Relative abundance and richness of new and uncultured OTUs at different taxonomic ranks across metagenomes from healthy individuals (n=3,083). **D**) The same data presented in (**C**), but only for new species-OTUs and conditioned by host population. Only populations with at least 30 metagenomes are shown. Orange box plots indicate samples from adults in rural countries, purple from adults in urban countries, and red from infants in urban countries. **C**–**D**) In box plots, middle line denotes median; box denotes interquartile range (IQR); and whiskers denote 1.5× IQR. **E**) IGGsearch sensitively detects the presence of species-OTUs in samples where no MAG was recovered. The x-axis indicated the number of MAGs assembled and the y-axis indicated the number of species-OTUs detected from IGGsearch profiling. Each point indicates one metagenomic sample (N=3,083). The red regression line is from a Pearson correlation. The vast majority of detected species are not assembled into a MAG. **F**) Species richness versus the relative % of new species-OTUs across metagenomic samples (n=3,083). Red regression line is from a Pearson correlation (rho=0.82, p value=0). New species-OTUs comprise a greater % of the community when diversity is high. This pattern was robust rarefying metagenomes to one million reads and using a prevalence-matched set of 1,000 new species and 1,000 known species (ρ=0.59, p value=0).



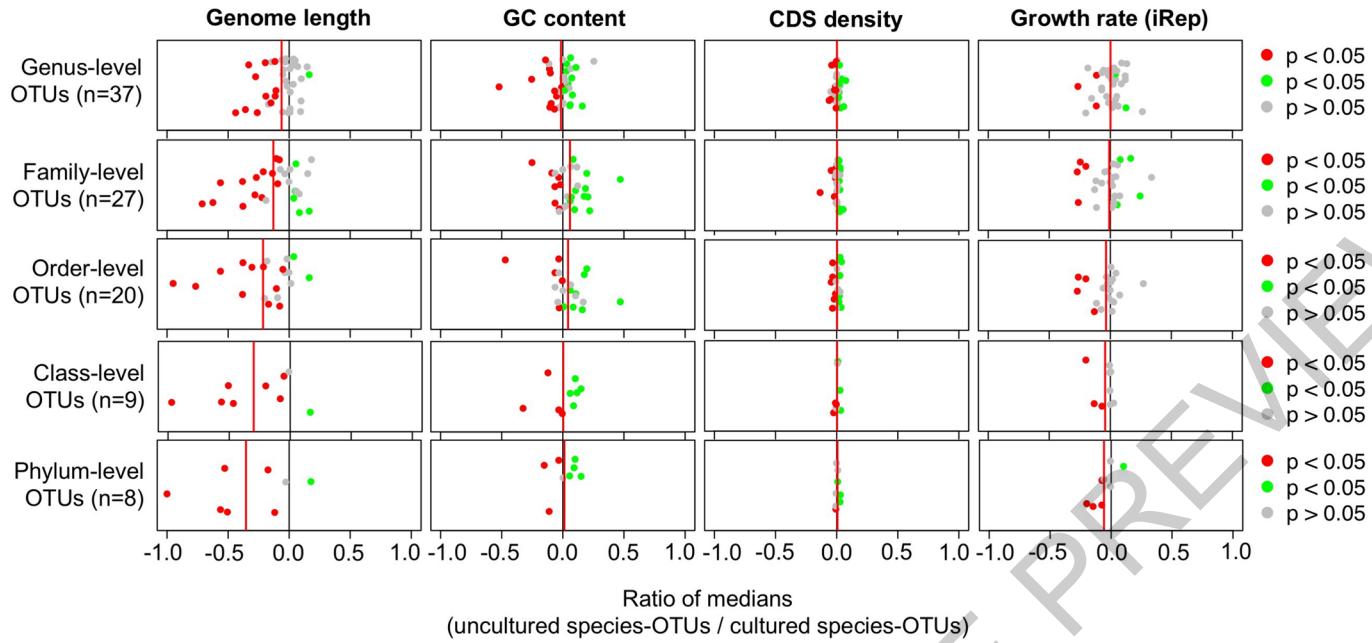
Extended Data Fig. 5 | Effect of completeness and contamination on identification of OTU from whole genomes. A-**C**) OTUs were identified for 296 *Bacteroides* genomes based on average-linkage clustering of whole-genome ANI using the ANIcalculator (v1.0). The ANI cutoffs used for forming OTUs are indicated by panel titles (94-97% ANI). The alignment fraction cutoffs (AF; 20-60%; defined as the required % of genome length aligned between genome pairs) is indicated by line color. In each panel, the vertical axis indicates the number of OTUs identified from genomes based on the ANI cutoff, AF cutoff, and the amount of missingness and/or contamination present in the 296 genomes. **A**) OTUs were identified for 296 *Bacteroidetes* genomes with up to 80% of genes randomly removed. The number of OTUs is inflated when genomes are incomplete and the alignment fraction is >20%. **B**) OTUs were identified for 296 *Bacteroidetes* genomes with up to 20% of genes from a different *Bacteroidetes* genome. The number of OTUs is not affected by contamination when genomes are complete. **C**) OTUs were identified for 296 *Bacteroidetes* genomes with 50% of genes randomly removed and up to 20% of genes from a different *Bacteroidetes* genome, representing a worst-case scenario. The number of OTUs is inflated by contamination when genomes are 50% complete. Using a lower ANI threshold (e.g. 94 or 95% vs 96 or 97%) reduces the negative impact of contamination. Based on these experiments, we chose an AF cutoff of 20% and ANI cutoff of 95% for identifying OTUs from MAGs and reference genomes in the current study.



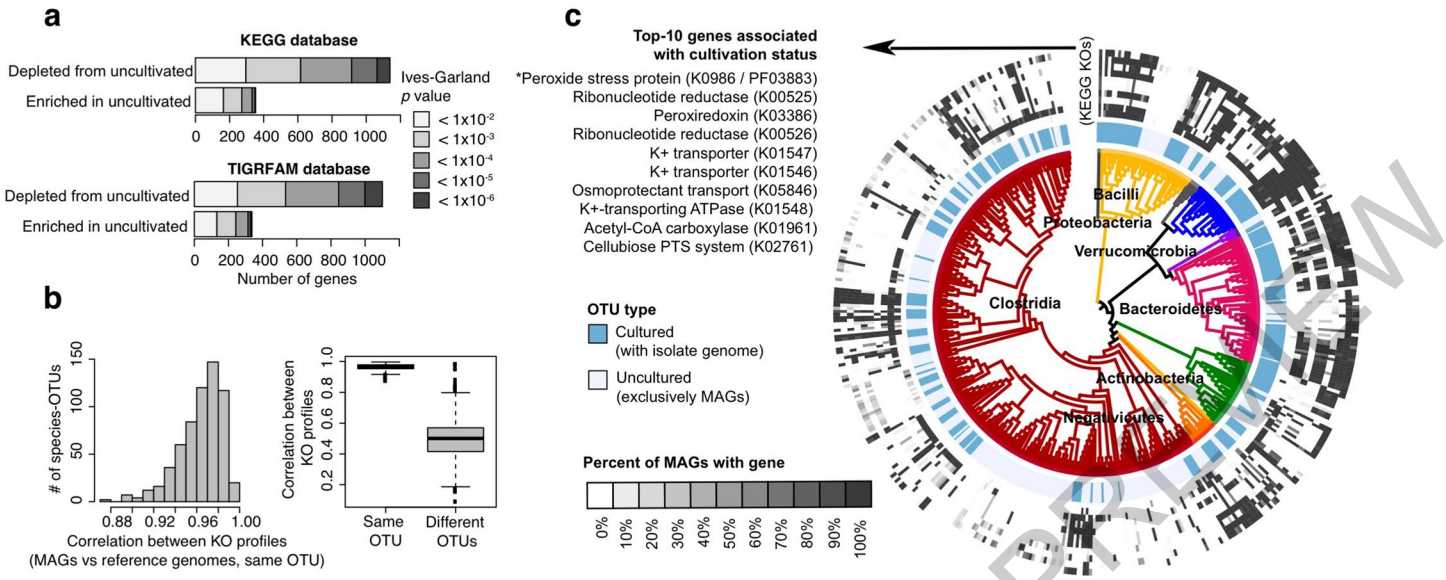
Extended Data Fig. 6 | Annotation and accumulation of human gut OTUs. A) Of the 23,790 species-OTUs identified from MAGs and reference genomes, 4,558 were classified as human gut based on (i) having a MAG from the HGM dataset, (ii) being detected in a human gut metagenome via read-mapping with IGGsearch, or (iii) containing a reference genome with metadata indicating isolation from a human stool sample. 2,058 of the 4,558 gut OTUs are represented exclusively by MAGs from the current study and are therefore new. Of the remaining 2,500 represented by reference genomes, only 955 contained a gut-isolated reference genome. The remaining 1,545 OTUs contain metadata indicating other isolation sources, including human, non-human, and environmental. For example, several gut species from non-host associated environments were isolated from human food products including milk, cheese, meat, and fermented foods. **B**) The occurrence frequency of all 4,558 gut OTUs was estimated across 3,810 human stool metagenomes using IGGsearch. For bar plots, the center bar indicates the mean; the error bar indicates the standard deviation; and 100 random data points are overlaid. *p* values are from two-sided Wilcoxon rank-sum tests. **C**) Accumulation curves indicating that discovery of genus and family-level OTUs from MAGs has saturated, but discovery of species-level OTUs. To make the plots, MAGs were randomly sampled without replacement and for each sample the number of unique OTUs was counted.



Extended Data Fig. 7 | Large lineages are depleted in high-quality genomes and isolate genomes. A) The trees indicate the phylogenetic distribution of species OTUs from the human gut for Cyanobacteria and a subclade within Clostridia. Each tip indicates one species OTU. Circles indicates whether a medium (open) or high-quality genome (closed) was recovered for MAGs from the HGM dataset (green), MAGs from PATRIC + IMG (blue), or isolate genomes from PATRIC + IMG (red). Diversity within these clades would have been missed without inclusion of medium-quality MAGs. **B**) The tree indicates the phylogenetic distribution of bacterial genus-level OTUs from the human gut (N=1,321 OTUs). The outer rings indicate whether an OTU contains: a MAG from the HGM dataset (green), a MAG from PATRIC + IMG (blue), or an isolate genome from PATRIC + IMG (red). Labels indicate phyla (NV=Negativicutes; CB/AB=Coriobacteriia/Actinobacteria; CY=Cyanobacteria; VM=Verrucomicrobia; DS=Desulfobacteraeota; EP=Epsilonbacteraeota; FB=Fusobacteria; SP=Spirochaetes). Large monophyletic clades depleted in isolate genomes are highlighted with green branches.



Extended Data Fig. 8 | Genome size (but not other features) consistently differs between MAGs from cultivated and uncultivated species-OTUs. Each column indicates one genomic feature (genome size, GC content, coding density, growth rate) that was compared between high-quality MAGs (N=24,345) from cultivated species OTUs (N=233) and MAGs from uncultivated species OTUs (N=271). To reduce redundancy, genomic features were averaged across all MAGs per species OTU. The value of each point in the figure indicates the $\log_2$ ratio of each genomic feature between uncultivated species OTUs and cultivated species OTUs. Each point indicates a single OTU at a higher taxonomic rank, with the rank indicated by row labels, and only higher-ranking OTUs with at least 10 cultivated and 10 uncultivated species-OTUs. Red and green points indicate if the distribution of a genomic feature was significantly different between groups based on a two-sided Wilcoxon rank-sum test after correction for multiple hypothesis tests (α=0.05). For example, a value of -1.0 at the phylum level for genome size indicates that: the genome size of MAGs within uncultivated species was 2x smaller than for cultivated species within a single phylum. Overall, MAGs from uncultivated species had consistently smaller genomes across taxonomic groups, regardless of the taxonomic rank, whereas other genomic features (GC content, coding density, and growth rate) did not consistently or systematically differ.



Extended Data Fig. 9 | Uncultivated OTUs are depleted in numerous functions, including genes for osmotic and oxidative stress. Genes from high-quality MAGs were functionally annotated based on the KEGG and TIGRFAM databases and the presence-absence of functions was averaged across MAGs per OTU. Functions were then compared between uncultivated OTUs (N=271) and cultivated OTUs (N=233) using the Ives-Garland phylogenetic logistic regression test and *p* values were corrected for multiple hypothesis tests using the Benjamini-Hochberg procedure. **A**) The number of genes associated with cultivation status does not depend on the database used for functional annotation. **B**) KEGG functional annotations were compared between high-quality MAGs and reference genomes from the same species-OTU (Left; N=665 OTUs) and between MAGs and reference genomes from different OTUs using Pearson correlation (Right; N=665 OTUs). MAGs and reference genomes have concordant functional annotations. In the box plots, the middle line denotes median; box denotes IQR; and whiskers denote 1.5× IQR. **C**) Phylogenetic tree of 271 uncultivated OTUs and 233 cultivated OTUs. The inner ring indicates whether an OTU is cultivated or not. The outer ring indicates the presence or absence genes from the KEGG database. The top 10 genes associated with cultivation status are shown. Of these, four are related to maintenance of osmotic pressure (K05846, K01547, K01546, K01548) and two (including the top hit) are related to oxidative stress (K0986, K03386). Note that the top hit, K0986, is listed as an uncharacterized protein in the KEGG database, but as a peroxide stress protein in the PFam database (PF03883). Organisms lacking these functions may have decreased viability during cultivation due to oxygen exposure and osmotic stress from growth in culture media.

Extended Data Table 1 | Metagenome-wide disease associations using IGGsearch and other tools

Dataset information			Number of disease associations (FDR < 0.01)					Minimum corrected two-sided p-value					ROC AUC from Random Forest classifier			
Disease	Citation	Sample Size (cases/ctrls)	IGGsearch ref only	IGGsearch new only	MIDAS	mOTU	MP2	IGGsearch ref only	IGGsearch new only	MIDAS	mOTU	MP2	IGGsearch	MIDAS	mOTU	MP2
Liver cirrhosis	Qin et al. 2014	123/114	**548**	325	367	227	149	3.8E-23	8.03E-22	**8.5E-24**	2.7E-23	2.6E-20	**0.95**	0.92	0.93	**0.95**
Atherosclerotic cardiovascular disease	Jie et al. 2017	176/187 (+)	**429**	197	222	104	93	**3.4E-14**	6.6E-12	6.9E-10	2.3E-12	1.4E-12	**0.90**	0.84	0.83	0.85
Ankylosing spondylitis	Wen et al. 2017	97/114	**304**	199	98	40	47	2.9E-20	**5.3E-28**	4.6E-11	1.4E-09	1.7E-10	**1.00**	0.98	0.97	0.98
*Colorectal cancer	Feng et al. 2015	46/63	112	**118**	14	20	15	2.8E-08	**6.4E-11**	4.8E-07	8.4E-07	2.4E-07	**0.95**	0.89	0.79	0.88
*Type II diabetes, Chinese cohort	Qin et al. 2012	165/185 (+)	**23**	9	11	11	19	**6.6E-05**	1.8E-04	8.0E-05	1.8E-04	1.6E-04	**0.79**	0.73	0.67	0.74
*Rheumatoid arthritis	Zhang et al. 2015	92/97 (+)	**11**	1	4	1	0	**2.6E-11**	0.0078	0.001	0.0086	0.040	**0.83**	0.78	0.68	0.73
Obesity	Le Chatelier et al. 2013	169/123	1	**4**	0	0	0	0.0060	**0.0035**	0.32	0.03	0.666	**0.68**	0.62	0.67	0.62
Non-alcoholic fatty liver disease	Loomba et al. 2017	14/72	1	1	0	1	**3**	0.0072	0.0072	0.0123	0.005	**0.0012**	**0.82**	0.65	0.71	**0.82**
Parkinson's disease	Bedarf et al. 2017	31/28	0	0	0	0	0	0.076	0.078	0.049	0.031	**0.030**	0.80	**0.83**	0.81	0.73
*Type II diabetes, Swedish cohort	Karlsson et al. 2013	33/43	0	0	0	0	0	0.320	0.320	**0.500**	0.22	0.135	0.56	0.59	0.56	**0.69**

IGGsearch and three other existing tools (MIDAS, mOTU, MP2=MetaPhlAn2) were used to estimate the abundance of species across samples spanning 10 studies. Two-sided Wilcoxon rank sum tests were used to identify differentially abundant species and *p* values were corrected for multiple hypothesis testing using the FDR procedure. For IGGsearch, disease associations are split into ref only (species-OTUs with reference genomes) and new only (species OTUs with only MAGs). Additionally, species profiles from all four tools were used to train Random Forest machine learning classifiers to predict disease state. Optimized models were identified by testing 1000 Random Forests with random combinations of model parameters and choosing the model with the greatest ROC AUC. To avoid overfitting, 10-fold cross-validation was performed. Reported AUC values are averages across 100 random forest runs. Bold text indicates the best performing tool for each disease; asterisks indicates studies used for MAG recovery; '+' indicates studies where a subset of cases was excluded due to medication for disease treatment.

nature research

Corresponding author(s): Stephen Nayfach, Katherine Pollard, Nikos Kyrpides

Last updated by author(s): Feb 27, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Assemblies were generated using MegaHIT v1.1.1. MAGs were generated using Maxbin v2.2.4, MetaBAT v2.12.1, CONCOCT 0.4.0, and DAS Tool v1.1.0 (option: '-score_threshold 0'.
Data analysis	MAGpurify v1.0 was used to refine MAGs. CheckM v1.0.7 was used to estimate MAG quality. BLASTN v.2.6.0 was used to remove contigs matching the human genome and phiX genome. HMMER v3.1b2 was used to identify marker genes. tRNAs and rRNAs were identified using tRNAscan-SE v.1.3.1 and Barrnap v0.9-dev (options: '–reject 0.01 –evalue 1e-3'), respectively. Bowtie 2 v2.3.4 was used for aligning reads to each MAG. Mash v2.0, ANIcalculator v1.0, and MC-UPGMA v1.0.0 were used to compute ANI and cluster genomes. GTDBTk v0.0.6 was used to taxonomically annotate MAGs. VSEARCH v2.4.3 (options: '-id 0.9, -target_cov 0.5 -query_cov 0.5) was used to construct pan-genomes. IGGsearch v1.0 was used to estimate the presence-absence and abundance of OTUs from metagenomes. FAMSA v1.2.5 was used for multiple sequence alignment. FastTree2 v.2.1.10 was used for tree building. HS-BLASTN v0.0.5 was used for performing alignment of pan-genome genes between species. IGGsearch was compared to MIDAS v1.3.0, MetaPhAn2 v2.7.7, and mOTU v1.1.1. Machine learning was performed using the scikit-learn package v0.19.1. iRep v1.10 was used to estimate replication rate. LAST v828 was used for alignment against the KEGG database. The phylolm R package v.2.6 was used for phylogenetic logistic regression.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Representative MAGs for the 2,058 new species have been deposited in the European Nucleotide Archive (ENA) under accession PRJEB31003. The entire data set and related metadata is freely available at https://github.com/snayfach/IGGdb.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Previous studies have found that a large proportion of species in the gut microbiome lack a sequenced genome. We addressed this problem by systematically recovering >60,000 draft genomes from nearly 4,000 metagenomes from phenotypically and geographically diverse human subjects.
Research sample	We downloaded 3,810 publicly available human fecal metagenome samples from the NCBI SRA spanning 15 studies.
Sampling strategy	Publicly available human gut metagenomes from major studies representing different geographic regions, lifestyles, age groups, and disease states.
Data collection	Downloaded from the NCBI sequence read archive
Timing and spatial scale	Data sets were selected to include samples from a wide range of ages (e.g. include both infants and adults), host lifestyles (e.g. urban, rural), host geography (e.g. United States, Denmark, Spain, Italy, Sweden, Finland, Estonia, Russia, Peru, El Salvador, Tanzania, Fiji, and China), and disease states (e.g. rheumatoid arthritis, diabetes, colorectal cancer, and autoimmunity)
Data exclusions	Several data sets from already well-samples regions (e.g. Europe and China) were excluded, which was pre-determined at the outset of the study.
Reproducibility	n/a
Randomization	n/a
Blinding	n/a

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging