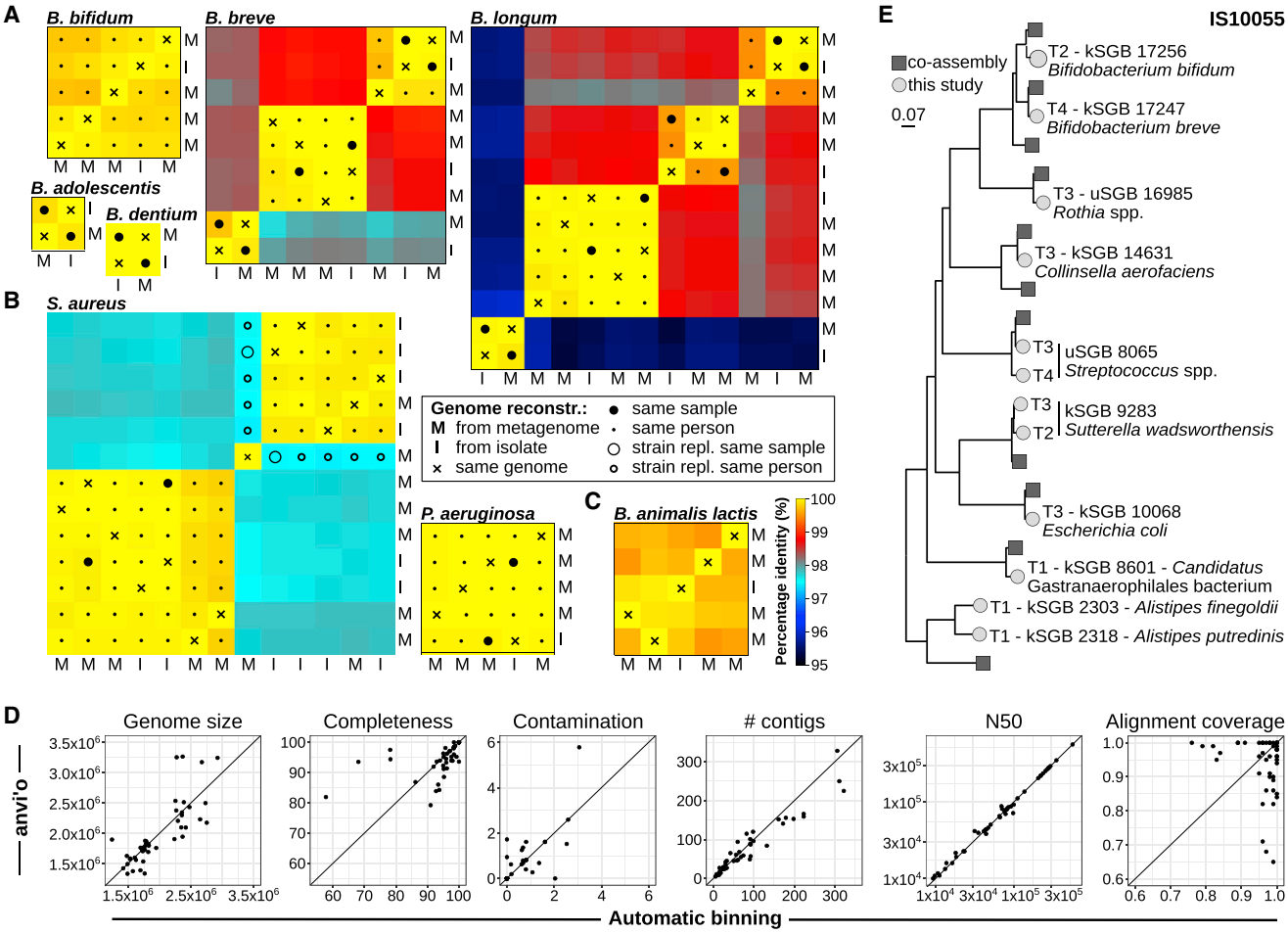


Figure 7. Quality of the Single-Sample Assembled Genomes against Multiple Alternative Genome Reconstruction Approaches

(A) Percentage identity between genomes from isolates (I) and genomes we reconstructed from metagenomes (M) for five *Bifidobacterium* species from the FerrettiP_2018 dataset (Ferretti et al., 2018). We mark isolates and metagenomes coming from the same specimen (big filled circles) and coming from specimens of the same mother-infant pair (small filled circles). In all cases, our automatic pipeline reconstructs genomes from metagenomes that are almost identical to the genomes of the expected isolated strains.

(B) The strains of *S. aureus* and *P. aeruginosa* isolated from three patients are almost perfectly matching the genomes reconstructed from sputum metagenomes sequenced at multiple time points. In the only case in which a *S. aureus* genome from a metagenome is not matching the strain isolated from a previous time point in the same patient, we verified with MLST typing that a clinical event of strain-replacement from ST45 to ST273 occurred.

(C) In the dataset by Nielsen et al. (2014), we successfully recover at >99.5% identity the strain of a *B. animalis* subspecies lactis present in a commercial probiotic product that was consumed by the enrolled subjects, even if the probiotic strain was at low relative abundance in the stool microbiome (<0.3% on average [Nielsen et al., 2014]).

(D) Comparison of the 46 manually curated genomes (using anvi'o) with automatically assembled (using metaSPAdes) and binned (using MetaBAT2) genomes.

(E) Example comparison between the set of single-sample assembled genomes and co-assembled genomes for a time series (n = 5) of gut metagenomes from a newborn. Several genomes reconstructed with the two approaches have the same phylogenetic placement, with single-sample assembly retrieving the same (or a very closely related) genome at multiple time points, and both methods retrieving some unique genomes. This is an example of the comprehensive comparison performed in the STAR Methods and reported in Table S2 and Figure S7B.

Even within the current data collection, a variety of results remain to be explored. Part of the metagenomic reads that could not be mapped against our extended bacterial and archaeal resource are likely coming from viral and eukaryotic genomes. For example, we found substantial amounts of viruses (>0.5% relative read depth in 101 samples for bacteriophages never found as prophages in reference bacterial genomes), of the intestinal eukaryotic parasite *Blastocystis* (>0.5% in 158 samples), and of the skin fungus *Malassezia* (>0.5% in 297 samples). Considering that *de novo* discovery of non-bacterial genomes is very challenging and should receive more attention in the future, eukaryotic microorganisms and viruses may thus account for some of the remaining unmappable sequences in these data (Figure 2). These results help to pinpoint microbes unique to a particular population, environment, or exposure, and most importantly, future work may then be able to more easily capture

specific strains or microbial molecular mechanisms that are causal in microbiome-associated human health conditions.

STAR★METHODS

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

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Overview of the approach
  - Meta-analyzed publicly available metagenomic datasets
  - Enrollment of participants from non-Westernized populations from Madagascar and Ethiopia
  - Sample collection of non-Westernized cohorts
  - DNA extraction and sequencing
  - Description of the non-Westernized cohorts
  - Isolate genomes and available metagenomic assemblies used as references
  - Metagenomic assembly and contig binning
  - Quality control of metagenomic assemblies
  - Validation of the pipeline for genome reconstruction from metagenomics using isolate sequencing and manually curated genomes
  - Evaluation of single-sample assemblies against co-assembly and co-binning methods
  - Grouping of metagenomic assemblies into species-level genome bins
  - Reconstruction of the human-microbiome phylogenetic structure
  - Quantification of the fraction of reads that can be mapped against SGBs
  - Pangenome, phylogenetic, and functional analysis of kSGB and uSGBs
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.cell.2019.01.001.

ACKNOWLEDGMENTS

We thank the MAHERY team, particularly Anjaranirina Evelyn Jean Gasta, Hervet Randriamady, and Miadana Vonona Arisoa, and the GeNaPi Project Team, particularly Mari Olcina, Lourdes Larruy, Carla Muñoz, and Cristina Alcántara. We thank Levi Waldron and all the members of the Segata and Huttenhower laboratories for fruitful discussions, the students of the Computational Microbial Genomics master course at University of Trento for help with the manual curation of assemblies, and the HPC and NGS facilities at University of Trento. This work was supported by EU-H2020 (DiMeTrack-707345) to E.P.; by NIH NHGRI (R01HG005220), NIDDK (R24DK110499), NIDDK (U54DE023798), CMIT (6935956) to C.H.; and by ERC (MetaPG-716575), MIUR (RBFR13EWWI), EU-FP7 (PCIG13-GA-2013-618833), CARITRO (2013.0239), and LEO Pharma Foundation to N.S. Madagascar data and sample collection was supported by the Rockefeller Foundation to C.D.G.

AUTHOR CONTRIBUTIONS

Conception and design, E.P. and N.S.; Methodology and analysis; E.P., F. Asnicar, S.M., M.Z., N.K., F.B., C.Q., and N.S.; Sample collection, processing, and data generation, F. Armanini, P.G., M.C.C., B.L.R., C.D., X.C.M., C.D.G., and C.H.; Public data collection and curation, E.P., N.K., F.B., P.M., and A.T.; Data interpretation, E.P., F. Asnicar, S.M., M.Z., C.H., and N.S.; Manuscript preparation: E.P., F. Asnicar, S.M., M.Z., N.K., C.H., and N.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 29, 2018
Revised: November 15, 2018
Accepted: December 28, 2018
Published: January 17, 2019

REFERENCES

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nat. Methods *11*, 1144–1146.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ *3*, e1029.

Asnicar, F., Manara, S., Zolfo, M., Truong, D.T., Scholz, M., Armanini, F., Ferretti, P., Gorfer, V., Pedrotti, A., Tett, A., et al. (2017). Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. mSystems 2. https://doi.org/10.1128/mSystems.00164-16.

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., et al. (2015). Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. Cell Host Microbe *17*, 852.

Beghini, F., Pasolli, E., Truong, T.D., Putignani, L., Cacciò, S.M., and Segata, N. (2017). Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. ISME J. *11*, 2848–2863.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al.; Genome Standards Consortium (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731.

Brinig, M.M., Lepp, P.W., Ouverney, C.C., Armitage, G.C., and Relman, D.A. (2003). Prevalence of bacteria of division TM7 in human subgingival plaque and their association with disease. Appl. Environ. Microbiol. *69*, 1687–1694.

Brister, J.R., Ako-Adjei, D., Bao, Y., and Blinkova, O. (2015). NCBI viral genomes resource. Nucleic Acids Res. *43*, D571–D577.

Brito, I.L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S.D., Jenkins, A.P., Naisilisili, W., Tamminen, M., Smillie, C.S., Wortman, J.R., et al. (2016). Mobile genes in the human microbiome are structured from global to individual scales. Nature *535*, 435–439.

Brooks, B., Olm, M.R., Firek, B.A., Baker, R., Thomas, B.C., Morowitz, M.J., and Banfield, J.F. (2017). Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. Nat. Commun. *8*, 1814.

Brown, B.L., Watson, M., Minot, S.S., Rivera, M.C., and Franklin, R.B. (2017). MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience *6*, 1–10.

Browne, H.P., Forster, S.C., Anonye, B.O., Kumar, N., Neville, B.A., Stares, M.D., Goulding, D., and Lawley, T.D. (2016). Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature *533*, 543–546.

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

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. *42*, D633–D642.

Costea, P.I., Coelho, L.P., Sunagawa, S., Munch, R., Huerta-Cepas, J., Forslund, K., Hildebrand, F., Kushugulova, A., Zeller, G., and Bork, P. (2017). Subspecies in the global human gut microbiome. Mol. Syst. Biol. *13*, 960.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc. Natl. Acad. Sci. USA *107*, 14691–14696.

Eddy, S.R. (2011). Accelerated Profile HMM Searchese1002195. PLoS Comput Biol *7*.

Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ *3*, e1319.

Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D.T., Manara, S., Zolfo, M., et al. (2018). Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe *24*, 133–145.e5.

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

Franzosa, E.A., McIver, L.J., Rahnavard, G., Thompson, L.R., Schirmer, M., Weingart, G., Schwarzberg Lipson, L., Knight, R., Caporaso, G., Segata, N., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. Nat. Methods *15*, 962–968.

Golden, C.D., Anjaranirina, E.J.G., Fernald, L.C.H., Hartl, D.L., Kremen, C., Milner, D.A., Jr., Ralalason, D.H., Ramihantaniarivo, H., Randriamady, H., Rice, B.L., et al. (2017). Cohort Profile: The Madagascar Health and Environmental Research (MAHERY) study in north-eastern Madagascar. Int. J. Epidemiol. *46*, 1747–1748d.

Goslee, S.C., and Urban, D.L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. J. Stat. Softw. *22*. https://doi.org/10.18637/jss.v022.i07.

Gossling, J., and Moore, W.E.C. (1975). Gemmiger formicilis, n.gen., n.sp., an Anaerobic Budding Bacterium from Intestines. Int. J. Syst. Evol. Microbiol. *25*, 202–207.

He, X., McLean, J.S., Edlund, A., Yooseph, S., Hall, A.P., Liu, S.-Y., Dorrestein, P.C., Esquenazi, E., Hunter, R.C., Cheng, G., et al. (2015). Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc. Natl. Acad. Sci. USA *112*, 244–249.

Herlemann, D.P.R., Geissinger, O., and Brune, A. (2007). The termite group I phylum is highly diverse and widespread in the environment. Appl. Environ. Microbiol. *73*, 6682–6685.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. *44* (D1), D286–D293.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol. Biol. Evol. *34*, 2115–2122.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

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

Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ *3*, e1165.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., and Snyder, M. (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat. Biotechnol. *34*, 64–69.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

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

Liu, W., Zhang, J., Wu, C., Cai, S., Huang, W., Chen, J., Xi, X., Liang, Z., Hou, Q., Zhou, B., et al. (2016). Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. Sci. Rep. *6*, 34826.

Manara, S., Pasolli, E., Dolce, D., Ravenni, N., Campana, S., Armanini, F., Asnicar, F., Mengoni, A., Galli, L., Montagnani, C., et al. (2018). Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of Staphylococcus aureus strains in a paediatric hospital. Genome Med. *10*, 82.

Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., et al. (2003). CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Red *31*, 383–387.

Marcobal, A., Barboza, M., Sonnenburg, E.D., Pudlo, N., Martens, E.C., Desai, P., Lebrilla, C.B., Weimer, B.C., Mills, D.A., German, J.B., and Sonnenburg, J.L. (2011). Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. Cell Host Microbe *10*, 507–514.

Merino, E., Jensen, R.A., and Yanofsky, C. (2008). Evolution of bacterial trp operons and their regulation. Curr. Opin. Microbiol. *11*, 78–86.

Meyer, F., Hofmann, P., Belmann, P., Garrido-Oter, R., Fritz, A., Sczyrba, A., and McHardy, A.C. (2018). AMBER: Assessment of Metagenome BinnERs. GigaScience 7.. https://doi.org/10.1093/gigascience/giy069.

Morotomi, M., Nagai, F., Watanabe, Y., and Tanaka, R. (2010). Succinatimonas hippei gen. nov., sp. nov., isolated from human faeces. Int. J. Syst. Evol. Microbiol. *60*, 1788–1793.

NCBI Resource Coordinators (2013). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *41*, D8–D20.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

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

Nikolenko, S.I., Korobeynikov, A.I., and Alekseyev, M.A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics *14* (*Suppl 1*), S7.

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

Obregon-Tito, A.J., Tito, R.Y., Metcalf, J., Sankaranarayanan, K., Clemente, J.C., Ursell, L.K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P.M., et al. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. Nat. Commun. *6*, 6505.

Oh, J., Byrd, A.L., Deming, C., Conlan, S., Kong, H.H., and Segre, J.A.; NISC Comparative Sequencing Program (2014). Biogeography and individuality shape function in the human skin metagenome. Nature *514*, 59–64.

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

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132.

Oyama, L.B., Girdwood, S.E., Cookson, A.R., Fernandez-Fuentes, N., Privé, F., Vallin, H.E., Wilkinson, T.J., Golyshin, P.N., Golyshina, O.V., Mikut, R., et al. (2017). The rumen microbiome: an underexplored resource for novel antimicrobial discovery. NPJ Biofilms Microbiomes *3*, 33.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics *31*, 3691–3693.

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

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. *2*, 1533–1542.

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D.T., Beghini, F., Malik, F., Ramos, M., Dowd, J.B., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. Nat. Methods *14*, 1023–1024.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.

Pritchard, L., Glover, R.H., Humphris, S., Elphinstone, J.G., and Toth, I.K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal. Methods *8*, 12–24.

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

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017a). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 833–844.

Quince, C., Delmont, T.O., Raguideau, S., Alneberg, J., Darling, A.E., Collins, G., and Eren, A.M. (2017b). DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol. *18*, 181.

Rampelli, S., Schnorr, S.L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A.N., Henry, A.G., and Candela, M. (2015). Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr. Biol. *25*, 1682–1693.

Raveh-Sadka, T., Thomas, B.C., Singh, A., Firek, B., Brooks, B., Castelle, C.J., Sharon, I., Baker, R., Good, M., Morowitz, M.J., and Banfield, J.F. (2015). Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. eLife *4*, e05477.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature *499*, 431–437.

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

Segata, N., Haake, S.K., Mannon, P., Lemon, K.P., Waldron, L., Gevers, D., Huttenhower, C., and Izard, J. (2012a). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol. *13*, R42.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012b). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods *9*, 811–814.

Segata, N., Börnigen, D., Morgan, X.C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. *4*, 2304.

Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., and Banfield, J.F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. *23*, 111–120.

Smits, S.A., Leach, J., Sonnenburg, E.D., Gonzalez, C.G., Lichtman, J.S., Reid, G., Knight, R., Manjurano, A., Changalucha, J., Elias, J.E., et al. (2017). Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science *357*, 802–806.

Solden, L., Lloyd, K., and Wrighton, K. (2016). The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr. Opin. Microbiol. *31*, 217–226.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

Tett, A., Pasolli, E., Farina, S., Truong, D.T., Asnicar, F., Zolfo, M., Beghini, F., Armanini, F., Jousson, O., De Sanctis, V., et al. (2017). Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. NPJ Biofilms Microbiomes *3*, 14.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45* (D1), D158–D169.

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

Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. *27*, 626–638.

Uritskiy, G.V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome *6*, 158.

van der Walt, A.J., van Goethem, M.W., Ramond, J.-B., Makhalanyane, T.P., Reva, O., and Cowan, D.A. (2017). Assembling metagenomes, one community at a time. BMC Genomics *18*, 521.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. *73*, 5261–5267.

Xie, G., Bonner, C.A., Brettin, T., Gottardo, R., Keyhani, N.O., and Jensen, R.A. (2003). Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in Xylella species and in heterocystous cyanobacteria. Genome Biol. *4*, R14.

Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Stool samples from Madagascar cohort	Golden et al., 2017	N/A
Stool samples from Ethiopian cohort	This paper	N/A
Critical Commercial Assays		
PowerSoil DNA Isolation Kit	MoBio Laboratories Carlsbad, USA	Catalog No. 12888-50
NexteraXT DNA Library Preparation Kit	Illumina, California, USA	FC-131-1096
Deposited Data		
Raw sequencing data (Madagascar cohort)	This paper	NCBI-SRA BioProject: PRJNA485056
Raw sequencing data (Ethiopian cohort)	This paper	NCBI-SRA BioProject: PRJNA504891
Data for all genomes	This paper	http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html
Representative genome for *Ca.* Cibiobacter qucibialis	This paper	DDBJ/ENA/GenBank accession SAUS00000000
Software and Algorithms		
metaSPAdes (version 3.10.1)	Nurk et al., 2017	https://github.com/ablab/spades/releases
MEGAHIT (version 1.1.1)	Li et al., 2015	https://github.com/voutcn/megahit
MetaBAT2 (version 2.12.1)	Kang et al., 2015	https://bitbucket.org/berkeleylab/metabat
CheckM (version 1.0.7)	Parks et al., 2015	https://github.com/Ecogenomics/CheckM
CMSeq (version 1.0.0)	This study	https://bitbucket.org/CibioCM/cmseq
Mash (version 2.0)	Ondov et al., 2016	https://github.com/marbl/Mash
MetaPhlAn2 (version 2.0)	Segata et al., 2012b; Truong et al., 2015	https://bitbucket.org/biobakery/metaphlan2
HUMANn2 (version 0.7.1)	Franzosa et al., 2018	https://bitbucket.org/biobakery/humann2/
Bowtie2 (version 2.2.9)	Langmead and Salzberg, 2012	https://github.com/BenLangmead/bowtie2
Prodigal (version 2.6.3)		https://github.com/hyattpd/Prodigal
Pyani (version 0.2.6)	Pritchard et al., 2016	https://github.com/widdowquinn/pyani
StrainPhlAn (version 2.0.0)	Truong et al., 2017	https://bitbucket.org/biobakery/metaphlan2
Anvi'o (version 4)	Eren et al., 2015	https://github.com/merenlab/anvio
BWA (version 0.7.17)	Li and Durbin, 2009	https://github.com/lh3/bwa
CONCOCT (version 0.5.0)	Alneberg et al., 2014	https://github.com/BinPro/CONCOCT
RPSBlast	Marchler-Bauer et al., 2003	ftp://ftp.ncbi.nih.gov/blast/executables/
PhyloPhlAn (version dev, 0.25)	Segata et al., 2013	https://bitbucket.org/nsegata/phylophlan
Diamond (version 0.9.9.110)	Buchfink et al., 2015	https://github.com/bbuchfink/diamond
mafft (version 7.310)	Katoh and Standley, 2013	https://github.com/The-Bioinformatics-Group/Albiorix/wiki/mafft
trimal (version 1.2rev59)	Capella-Gutiérrez et al., 2009	https://github.com/scapella/trimal
RAxML (version 8.1.15)	Stamatakis, 2014	https://github.com/stamatak/standard-RAxML
IQ-TREE (version 1.6.6)	Nguyen et al., 2015	https://github.com/Cibiv/IQ-TREE
Roary (version 3.8)	Page et al., 2015	https://github.com/sanger-pathogens/Roary
blastn (version 2.6.0+)	Altschul et al., 1990	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast
FastTree (version 2.1.9)	Price et al., 2010	https://github.com/PavelTorgashov/FastTree
ecodist R package	Goslee and Urban, 2007	https://github.com/cran/ecodist
GraPhlAn (version 1.1.3)	Asnicar et al., 2015	https://bitbucket.org/nsegata/graphlan/
FigTree (version 1.4.3)	N/A	http://tree.bio.ed.ac.uk/software/figtree/
Prokka (version 1.12)	Seemann, 2014	https://github.com/tseemann/prokka
EggNOG mapper (version 1.0.3)	Huerta-Cepas et al., 2017	https://github.com/jhcepas/eggnog-mapper
HMM	Eddy, 2011	https://github.com/guyz/HMM

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Barrnap (version 0.9)	N/A	https://github.com/tseemann/barrnap
RDP (version 2.11)	Cole et al., 2014; Wang et al., 2007	https://github.com/rdpstaff/classifier
Other		
curatedMetagenomicData	Pasolli et al., 2017	https://github.com/waldronlab/curatedMetagenomicData
UniProt	The UniProt Consortium, 2017	https://github.com/ebi-uniprot
NCBI GenBank database	NCBI Resource Coordinators, 2013	https://www.ncbi.nlm.nih.gov/genbank/
RefSeq (viral genomes and plasmids)	Brister et al., 2015; O'Leary et al., 2016	https://www.ncbi.nlm.nih.gov/refseq/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources, reagents, and software should be directed to and will be fulfilled by the Lead Contact, Nicola Segata (nicola.segata@unitn.it).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects enrolled in our study are adults from the Madagascar and Ethiopian non-Westernized cohorts described in the Methods below. Ethical approvals were given by the Madagascar Ministry of Health and the Office for the Protection of Human Subjects at the Harvard T.H. Chan School of Public Health, protocol #22826 for the Madagascar cohort, and by the Research Ethics Committee of the Valencia University (reference number: H1484811493170) and also by the Ethics Committee of the Consejo Superior de Investigaciones Cientìficas (Madrid, Spain), number 058/2018, for the Ethiopia cohort. Informed consent was obtained for all individuals.

METHOD DETAILS

Overview of the approach

Our approach to reconstruct bacterial and archaeal genomes from the human microbiome (Figure 6A) exploits metagenomic single-sample assembly, contig binning, and species-level inter-sample genome grouping at the scale of the many thousands of metagenomes now available in public repositories.

In brief, we first collected and curated a metagenomic resource comprising a total of 9,428 metagenomes (from public resources and samples sequenced in this study, see below) and then applied metagenomic assembly - metaSPAdes (Nurk et al., 2017) or MEGAHIT (Li et al., 2015) - to each sample separately. Each metagenomic assembly was then quality controlled for minimum length and the 204M contigs were subjected to sample-specific contig binning based on tetranucleotide frequency and contig abundance using MetaBAT2 (Kang et al., 2015) resulting in over 345,000 putative genome bins (Figure 6A). Genome bins were then strictly quality controlled to identify reconstructed genomes with quality at least comparable with the typical quality of isolate genome sequencing. By controlling genome completeness and contamination using CheckM (Parks et al., 2015) and strain heterogeneity with the CMSeq pipeline described below, we identified 70,178 high-quality genomes and 84,545 additional MQ genomes (Figure 6A).

The 154,723 reconstructed genomes and the 80,990 reference genomes retrieved from public repositories (see below) were then clustered based on whole-genome nucleotide similarity estimation using Mash (Ondov et al., 2016). The cutoff on the hierarchical clustering was tuned based on the intra- and inter-species diversity of the confidently taxonomically labeled subset of the 80,990 reference genomes resulting in species-level genome bins (SGBs) spanning ~5% genetic diversity, as independently proposed elsewhere (Jain et al., 2018). Overall we obtained 16,332 SGBs that were further divided in known SGBs (kSGB) that contain at least one reference genome, unknown SGBs (uSGBs) without any reference genomes, and non-human SGBs containing only reference genomes and no genomes reconstructed from our assembly of the human microbiome (Figure 6A). The kSGBs were then taxonomically labeled with the species label (if available) of the reference genome(s) present in the bin, whereas uSGBs were assigned the phylum of their closest reference genome, and to a genus-level and family-level annotation when possible.

Meta-analyzed publicly available metagenomic datasets

We collected publically available metagenomic samples from 46 different studies, totaling 9,316 metagenomes and 4.1e11 Illumina reads. Overall, the samples cover 31 countries: USA (1,431 samples), China (1,342), Israel (956), Sweden (600) and Denmark (580) are the 5 most represented. The metagenomes were sampled from 5 major body sites: 7,783 samples from the gut (stool samples), 783 from the oral cavity, 503 from the skin (including 93 samples from anterior nares), 88 from the vagina, and 9 from maternal milk (excluded for visualization from the figures). Samples from adults (19 to 65 years of age) account for 6,615 samples, but all age categories are covered with 1,098 newborns (< 1 year of age), 465 children (age ≥1 year and <12 years), 216 school-age individuals

(age ≥12 and <19 years), and 876 from adults and seniors (age ≥19 and >65 years; merged with the class "adult" in Figure 1). Despite manual curation efforts, 46 samples from public repositories used here still miss the metadata for age category. All these and other manually-curated metadata fields are available in Table S1 and are included in the *curatedMetagenomicData* package (Pasolli et al., 2017) together with all the taxonomic (Segata et al., 2012b; Truong et al., 2015) and functional potential profiles (Franzosa et al., 2018) of the microbial species with available reference genomes. To cross-validate the results on the raw-reads mappability, we also retrieved 384 additional metagenomes not used to reconstruct the SGBs. Specifically, we considered 303 Westernized gut metagenomes, 52 Westernized oral metagenomes and 29 non-Westernized oral metagenomes as reported in Table S1.

Enrollment of participants from non-Westernized populations from Madagascar and Ethiopia

We enrolled, sampled, and sequenced the gut microbiome of individuals from the Madagascar Health and Environmental Research (MAHERY) study cohort that was set up in 2004 in a remote rainforest region in north-eastern Madagascar to study the impact of environmental change on human health (Golden et al., 2017). The cohort includes local people (Betsimisaraka and Tsimihety ethnicity) whose diet relies heavily on self-grown rice and wild plants and meats. Samples were collected between January 2013 and May 2014 from two subsistence communities (A and B) adjacent to the Makira Natural Park, approximately 10 km away from each other. A subset of the households in the two communities were randomly selected to be enrolled in the study (95 households out of 160 in Community A and 57 households out of 157 in Community B), for a total of 719 individuals < 74 years old. Enrolled people were subjected to clinical visits and questionnaires about dietary intake, and were asked to collect biological samples (fingernails, blood, faeces) to assess health and nutritional status. The samples considered in this study were collected from a total of 112 healthy volunteers (54 females and 58 males, Table S1). The gut microbiome of five female individuals were also sampled from a previously established cohort in Gimbichu (Ethiopia, Oromia Region).

Sample collection of non-Westernized cohorts

Faecal samples from the Madagascar cohort were self-collected in sterile polypropylene screw cap collection tubes (Sarstedt) after defecation on the waxy side of a banana leaf, and returned to the local research team within five hours of collection (Golden et al., 2017). Three ml of 97% ethanol were added to stabilize samples before storing them at −23°C within 14 days of collection. Samples were then shipped on dry ice to the USA to be stored at −80°C. Faecal samples from Ethiopian individuals were collected in REAL MiniSystem "Total - fix" (Durviz S.L., Valencia, Spain) and kept frozen at −80°C.

DNA extraction and sequencing

DNA was extracted with the PowerSoil DNA Isolation Kit (MoBio Laboratories) after pre-heating to 65°C for 10 min and to 95°C for 10 min (HMP Consortium, 2012). Libraries were prepared with the NexteraXT DNA Library Preparation Kit (Illumina) and sequenced on the HiSeq2500 machine (Illumina). The metadata for this cohort are available in Table S1 and are included in the *curatedMetagenomicData* package together with the taxonomic and functional potential profiles of the species with available reference genomes. We sequenced the 117 samples for a total of 593.9 Gb (5.3 Gb average per sample after quality control, 3.87 Gb standard deviation, Table S1). The raw reads were submitted to the NCBI-SRA archive and are available under the BioProjects PRJNA485056 (Madagascar cohort) and PRJNA504891 (Ethiopian cohort).

Description of the non-Westernized cohorts

Westernization and urbanization are complex processes that occurred during the last few centuries involving profound lifestyle changes compared to populations prior to the modern era. These changes include increased hygiene and sanitized environments, introduction and large availability of antibiotics and other drugs, switch toward a high-calorie high-fat dietary regimes and toward processed sterilized food, enhanced exposure to xenobiotics and pollutants, reduced contact with wildlife and domesticated animals, and transition from autarchic food production systems to a controlled food chain in a global economy. All these factors are thought to have dramatic effects on the human microbiome that co-evolved with our our body for hundred thousands of years in non-Westernized conditions. In this work, we adopt the terms "Westernized" and "non-Westernized" as umbrella terms to depict populations that differ by at least the majority of the above factors even though this definition comprises very heterogeneous populations.

In addition to the sequenced Madagascar cohort (above), 480 additional samples were annotated as "non-Westernized" from a total of 5 studies spanning 4 populations. These were a traditional Fijian population (Brito et al., 2016) (172 stool samples and 140 saliva samples), the hunter-gatherer Hadza population (Tanzania) from two different studies (Rampelli et al., 2015; Smits et al., 2017) (67 stool samples in total), the traditional agro-pastoral Mongolian population (Liu et al., 2016) (65 stool samples), and a Peruvian rural community (Obregon-Tito et al., 2015) (36 stool samples). With the Madagascar cohort, this work thus considers a total of 592 non-Westernized compared with 8,836 Westernized samples.

Isolate genomes and available metagenomic assemblies used as references

We considered the whole set of 17,607 microbial species (16,959 bacteria, 648 archaea) available as of March 2018 in the UniProt portal (The UniProt Consortium, 2017) for which at least one proteome (the set of coding sequences associated with the genome) is available. Quality control performed by UniProt to retain the proteomes and the associated genomes include the availability of a set of

annotated coding sequences and the check that the number of coding sequences is statistically consistent with the one of proteomes of neighboring species. We then considered all the available annotated genomes for these species and downloaded them from the NCBI GenBank database (NCBI Resource Coordinators, 2013) obtaining a total of 80,853 genomes. This large genome set comprises both complete (12%) and draft (88%) genomes, and it is the largest set of microbial isolate genomes with taxonomic assignments and quality-controlled sequences available as of March 2018. Draft genomes include also metagenomic species that are explicitly labeled with the ''MAG'' abbreviation (n = 37) and co-abundance gene groups metagenomic assemblies (CAGs, n = 377) (Nielsen et al., 2014). We further added this genome set to the 137 isolate genomes collected in (Browne et al., 2016) for a total of 80,990 considered as reference genomes. We refer to this set of 80,990 as ''isolate genomes'' for brevity, but they also comprise previous metagenomic assembly as mentioned above. To further expand the set of reference genomes we also considered all the 159,803 assemblies available in NCBI as of September 2018.

Metagenomic assembly and contig binning

Each of the 9,428 samples were processed with the standard quality-control employed by metaSPAdes (Nurk et al., 2017) which includes the read corrector BayesHammer (Nikolenko et al., 2013) and then independently subjected to *de-novo* metagenomic assembly through metaSPAdes (Nurk et al., 2017) (version 3.10.1; default parameters), which exhibited the best accuracies in recent comparisons among metagenomic assemblers (Forouzan et al., 2018; van der Walt et al., 2017). Samples that failed to be processed due to memory requirements (>1Tb of RAM), and samples with only unpaired reads, were assembled through MEGAHIT (Li et al., 2015) (version 1.1.1; default parameters). An extended comparison between metaSPAdes and MEGAHIT assemblers across all the datasets considered in this study confirmed that metaSPAdes performs consistently better especially in recovering long contigs (Figure S7A). Contigs shorter than 1,000 nt were discarded from further processing. This resulted in 2.04e8 different contigs for a total length of 8.67e11 nt. Reads were mapped to contigs using Bowtie2 (Langmead and Salzberg, 2012) (version 2.2.9; option '--very-sensitive-local') and the mapping output was used for contig binning through MetaBAT2 (Kang et al., 2015) (version 2.12.1; option '-m 1500'), which showed good performance in comparison with other binning methods (Meyer et al., 2018). MetaBAT2 achieved the best performances among single-sample binning tools also in the evaluation performed in the Metawrap paper (Uritskiy et al., 2018), a recent tool for multiple binning. The multiple binning approach looks promising, although lack of independent validation and high computational requirements make it infeasible to be used in the large-scale scenario exploited in this paper at this stage. The procedure of binning through MetaBAT2 generated 345,654 bins (i.e., putative genomes) for a total length of 6.55e11 nt indicating that 75% of the assembled contigs were grouped into bins.

The relative abundance of each reconstructed genome in the 9,428 metagenomes was calculated from the alignments of the raw reads against the assemblies of the same sample (performed using BowTie2 as reported above). This avoids spurious read assignments (i.e., reads mapping sufficiently well against more than one genome in the same or different species). Indeed, as a direct consequence of the assembly-based approach, it is very rare (< 0.01%) that a read can be assigned to more than one contig assembled from a metagenome containing the read itself. Thus, the relative genome abundance in each sample was defined as the number of reads aligning to each contig of the genome normalized by the total number of reads in the sample. Only primary alignments with alignment length $\geq$50 nt and edit-distance with respect to the contig $\leq$2 nt were considered. Abundances at SGB level in each sample were computed as the sum of the abundances of the reconstructed strains belonging to the same SGB.

Quality control of metagenomic assemblies

Putative genomes were subjected to quality control to generate the final set of reconstructed draft genomes. Three main measures were taken into account: i) completeness; ii) contamination; and iii) strain heterogeneity. Completeness and contamination were estimated using CheckM (Parks et al., 2015) (version 1.0.7; lineage specific workflow), while strain heterogeneity was estimated through a strategy we developed to identify assemblies resulting from strain mixtures even when the strains were very closely related. Following this procedure, reads were mapped against the reconstructed genomes from the same sample using Bowtie2 (Langmead and Salzberg, 2012) (version 2.2.9; option '--very-sensitive-local') and dominant and non-dominant alleles were determined over all protein coding nucleotides. We only considered base calls with a PHRED quality score of at least 30 and only those positions with a coverage of at least 10x. We considered a position as non-polymorphic if the dominant allele frequency was >80%. In order to calculate the polymorphic rate, we then considered only polymorphic positions corresponding to non-synonymous mutations. Validation experiments performed by mixing simulated metagenomic sequencing (with Illumina error models) of 5 randomly selected pairs of strains from each of the the 10 *Bacteroides* species of Figure 4 at decreasing dominant strain frequency (and thus higher nucleotide-level heterogeneity) confirmed that this approach reflects indeed the expected level of strain mixture. The strain heterogeneity estimation tool is available at https://bitbucket.org/CibioCM/cmseq.

Based on these quality estimated and on recent guidelines (Bowers et al., 2017), we selected as medium-quality (MQ) genomes those having completeness >50% and contamination <5% resulting in a total of 154,723 microbial genomes. Stricter quality control reduced the set of near-complete, high-quality (HQ) genomes to 70,178 with completeness >90% and no evidence of strong intra-sample strain heterogeneity (<0.5% polymorphic positions). The strain heterogeneity threshold removed 3,653 reconstructed genomes (5.2%) of otherwise HQ genomes, and we verified that these genomes tended to have higher CheckM contamination (although always below the recommended 5% threshold) with a median of 0.74% against 1.56% ($p < 1e-50$). This provides an additional indication that the CMSeq heterogeneity score helps in controlling strain mixtures and contaminations.

We evaluated the presence of plasmids and viruses within reference genomes and reconstructed SGBs by mapping the 13,924 plasmids and 10,529 viruses in RefSeq against the 80,990 reference genomes and the 154,723 genomes in the SGBs with BLAST (Altschul et al., 1990). We filtered alignments shorter than 500 nucleotides and with less than 80% identity. A plasmid or virus was considered to be present if at least 50% of its sequence was covered by any genome or SGBs in our catalog. We found that 37% of the fully sequenced plasmids in the RefSeq repository were represented in the reconstructed genomes (95% in the available reference genomes). The 16S rRNA sequences in the SGB genomes were searched with Barrnap 0.9 (default parameters). The 16S rRNA taxonomy (Table S4) was inferred with RDP rRNA classifier version 2.11 (Cole et al., 2014; Wang et al., 2007) (default parameters), only on predicted rRNA sequences longer than 500 nucleotides. We set RDP's minimum confidence threshold to call for each taxonomic level at 75%. Although we confirmed that the 16S rRNA gene is challenging to be recovered by metagenomic assembly (it was recovered in only 7.43% of the reconstructed genomes), the search for the most 400 conserved coding genes from PhyloPhlAn (Segata et al., 2013) in the reconstructed genomes and isolate sequencing available for the 9 largest SGBs and the 10 *Bacteroides* SGBs of Figure 4, confirmed that cross-species conservation of genes is not an issue for metagenomic assembly. Metagenomically reconstructed genomes recovered more PhyloPhlAn markers in 10 cases and less markers in 9 cases, and all comparisons were within 5% average differences.

Validation of the pipeline for genome reconstruction from metagenomics using isolate sequencing and manually curated genomes

Genomes reconstructed from metagenomes were compared with the ones of isolates obtained from the same sample, or from samples obtained from the same individual at earlier or later time points (Figures 7A–7C; Table S2). We compared 18 isolates with 36 genomes reconstructed from metagenomes from 8 different bacterial species. Compared samples included sputum from cystic fibrosis patients (Manara et al., 2018), stool and breast milk samples from mother-infant pairs (Ferretti et al., 2018), and feces of adults consuming fermented milk product containing a probiotic strain (Nielsen et al., 2014).

Comparison between the genome reconstructed from the automatic pipeline and the one from isolate was done by computing the average nucleotide identity (ANI) and the corresponding alignment coverage using the pyani tool (Pritchard et al., 2016) (version 0.2.6; option '-m ANIb'). Results showed that in all cases the genomes reconstructed from metagenomes with our automatic pipeline were almost identical to the genome of the expected isolated strains. For the only case in which this was not true (*S. aureus* isolate MF093 and paired metagenome CM_cf__CF_FIFC009SS_t3M17__bin.3), we verified with MLST typing (both from assembled and unassembled reads) and with StrainPhlAn (Truong et al., 2017) that a clinical event of strain-replacement from ST45 to ST273 occurred.

A similar analysis was conducted to compare the genomes reconstructed using our fully automated pipeline with the ones obtained through manual curation using anvi'o (Eren et al., 2015) (Figure 7D; Table S2). Manually curated genomes were generated starting from the same set of unbinned contigs. A total of 50 metagenomes from the database considered in this study were randomly selected and assigned to six groups of students that were previously trained for the task of manual curation of contig binning by guided execution and discussion of the available anvi'o tutorials followed by curation of several example metagenomes common to all groups. Each group was asked to bin contigs for the strain with the highest reconstruction quality in the sample. This resulted in 46 manually-curated reconstructed genomes. Our automatic procedure recovered a genome closely matching (>99.5% whole genome genetic identity) the manually-curated one in all 46 cases. The comparison between genomes was done by computing the ANI score through the pyani tool and the results are reported in Figure 7D and Table S2.

Evaluation of single-sample assemblies against co-assembly and co-binning methods

In order to provide a comparison to the single sample strategy employed here, we co-assembled and co-binned a subset of the data where multiple samples from the same individual were available. Samples were taken from two studies: the already described investigation of the microbiome of newborns and of their mothers (Ferretti et al., 2018), and a study considering fecal microbiome time series for adults (Costea et al., 2017). From the first, we selected 22 infants for which at least 3 fecal samples taken during the first four months post-partum were available (maximum 5, median 4). We also co-assembled 21 fecal samples from the mothers from the same study to provide a comparison against cross-sectional co-assembly. Somewhat longer fecal time series were available from the second study, from which we selected four individuals with a number of time points between eight and ten (Costea et al., 2017). This gave us a total of 26 longitudinal time series from the same individual and one cross-sectional study (21 individuals) each of which we co-assembled using MEGAHIT (Li et al., 2015) with default parameters except for the kmer-list set to (21,31,..,99). The assembled contigs were then cut into 10kbp fragments and the reads from each sample within the time series (or mother in the cross-sectional study) were then mapped back onto the contig fragments using BWA and a per sample depth of coverage was calculated (Li and Durbin, 2009). The contig fragments were then clustered using the CONCOCT algorithm (default parameters) which combines both tetramer composition and coverage in a Gaussian mixture model after a PCA based dimensionality reduction (Alneberg et al., 2014). Following clustering, a consensus cluster assignment across fragments was given to each contig to assign clusters based on the original co-assembly.

We called ORFs on the co-assembled contigs and assigned COGs (Tatusov et al., 2003) using RPSBlast. The same procedure was applied to the genomes reconstructed by single sample assembly from the same set of samples used in each co-assembly. We then selected only those reconstructed genomes from both studies that possessed more than 75% of a panel of 36 single copy core genes in single copy (Alneberg et al., 2014). To remove redundancy across reconstructed genomes from the single sample clustering

(i.e., same genome reconstructed at multiple time points from the same individual), and to determine the intersection of genomes between the two approaches, we then performed a hierarchical average linkage clustering of all the genomes from both methods and clustered at 1% nucleotide identity on the core gene panel. The results of such procedure are given in Table S2. We then also evaluated the genomes obtained by the co-assembly by computing CheckM completeness, CheckM contamination, and CMSeq heterogeneity as described for the single-assembly reconstructed genomes. Co-assembled genomes were then assigned to the HQ or MQ category with the same thresholds used for the single-assembly reconstructed genomes. The number of HQ and MQ genomes obtained with the two approaches was then compared, and additional genome quality metrics such as genome length, N50, completeness estimate, and contamination estimates were considered. For the genomes obtained by single-sample assembly, the grouping into SGBs was used to compare the number of distinct species obtained compared to the co-assembly approach. This second set of evaluations is also reported in Table S2.

For the short infant time series, the increase in number of genomes obtained by co-assembling and co-clustering was typically modest after collapsing closely related strains from single-genome assembly (median increase of 3% for the 36-core gene based evaluation - Table S2, 6.87% for the CheckM-based evaluation with thresholds for HQ genomes - Table S2). Without removal of closely related strains, single-genome assembly recovered more genomes (12% HQ genomes, 50% MQ genomes) because the same strains (or closely related ones) were recovered at multiple time points (Figures 7E and S7B).

The improvement for the co-assembly approach was more clear from the second study where at least eight time points were available (median increase 31% - Table S2). Across all the considered individuals there was a weak correlation between increase in the number of reconstructed genomes obtained from co-clustering and sample number ($p = 0.08$). We conclude that co-assembling and co-binning of gut metagenomes requires a moderate number of samples (more than 5) to achieve substantial improvements. The co-assembly of mothers yielded an increase of 3% in the number of HQ genomes (after merging single-sample assemblies into 99% identity genome bins) when using the 36 single-copy genes for quality control (Table S2), and a decrease from 124 to 88 HQ SGB-grouped genomes when using >90% CheckM completeness and < 5% CheckM contamination thresholds (Table S2). Other genome quality statistics were very similar between the two approaches with however the co-assembly method showing slightly more contamination (1.7% against 0.9% for HQ genomes, Table S2). Overall, this suggests that large scale co-assembly may at best offer limited improvement in terms of overall recovered diversity.

It is of note that the co-assembly approach can reconstruct only one bin per species or subspecies (Figures 7E and S7B) and on a large cross-sectional database such as the one considered in this study, this would effectively be a composite population-level genome incorporating both variation in single-nucleotide variants on core genes and variation in accessory genes. It is possible to resolve this variation on co-assemblies via single-nucleotide variant calling (Quince et al., 2017b; Truong et al., 2017) and when this is followed by deconvolution of haplotypes across samples as employed in the DESMAN pipeline (Quince et al., 2017b) this does allow the reconstruction of whole-genome haplotypes and assignment of accessory genes to specific strains. However, when most species are present in a single dominant strain as it is the case in the human microbiome (Truong et al., 2017), directly assembling strains from individual samples is a more straightforward strategy that both avoids the deconvolution step and uncertainties associated with variant calling from mapped reads. It is therefore more suitable for the very large scale analyses considered here where the aim is to generate a small number of HQ strains from each sample to provide the most comprehensive picture of overall diversity in the human gut.

The general conclusion of this comparison is thus that co-assembly and co-binning approaches would be useful for retrieving substantially more genomes in relatively long (>5) subject-specific time series, whereas the potential advantage of retrieving more low-abundance species in a cross-sectional co-assembly is overcome by the disadvantage of having to use more complex approaches such as DESMAN to resolve the strain variation. That is perhaps more appropriate where the aim is to extract as much information as possible from a single study rather than to produce a single comprehensive high fidelity strain catalog. Because time series comprising more than 5 samples from the same subject and body site are very rare in the available cohorts (only 70 individuals - i.e., 1.0% - in our database), co-assembly is not considered in the present work as it would not provide advantages.

Grouping of metagenomic assemblies into species-level genome bins

The 154,723 reconstructed genomes, in addition to the 80,990 reference genomes, were organized into species-level genome bins (SGBs). We applied an all-versus-all genetic distance quantification (nucleotide identity) on the total of 235,713 genomes using Mash (Ondov et al., 2016) (version 2.0; option "-s 1e4" for sketching) followed by hierarchical clustering with average linkage (using the fastcluster Python library).

The cutoff on the resulting dendrogram to define species-level genome bins (SGBs) was selected based on the intra- and inter-species diversity of the confidently taxonomically labeled subset of the 80,990 reference genomes. Microbial species labels for the genomes were inferred from the taxonomic label provided by NCBI GenBank in association with the genomes, and excluding all genomes containing ambiguous terms in the species name (i.e., "_sp," "archaeon," "bacterium," or "candidatus"). This resulted in a total of 61,198 genomes spanning 5,494 named species.

With this labeling, the optimal dendrogram cutoff threshold to defined species-level genome bins (SGBs) was then chosen by taking into account two competing criteria (Figure S2C): i) minimization of the over-clustering error (Figure S2C-**i**) to avoid that genomes from the same species fall into different SGBs; ii) minimization of the under-clustering error (Figure S2C-**ii**) to prevent that genomes from different species fall into the same SGB. The two criteria were computed across all available species and cutoff choice,

normalized by the total number of available genomes, and summed up to get the value to minimize. Results showed a minimization of the error for a threshold equal to 0.05 (Figure S2C-iii), which was thus adopted to cut the dendrogram and generate SGBs spanning ~5% genetic diversity. A similar 5% genetic diversity range to define species boundaries was independently proposed and validated elsewhere (Jain et al., 2018), thus serving as a reasonable compromise despite the wide diversity of genomic similarities within existing defined species. This threshold was also confirmed by considering only prevalent species (>10 genomes) representing more studied and validated species, and by subsampling to a maximum of 10 genomes per species in order to avoid biases due to the different number of available genomes in existing species.

The resulting SGBs were further refined in order to prevent that same-species genomes were split into multiple SGBs due to inaccurate estimation of Mash for incomplete draft genomes. First i) a representative genome was selected for each SGB (Table S4). This was done by ranking genomes based on five metrics: completeness (in decreasing order), contamination (increasing), coverage (decreasing), strain heterogeneity (increasing), N50 (decreasing). The representative genome was selected as the one minimizing the sum of the five ranks. Then ii) the closest SGB was identified for each SGB based on the distances among representatives and iii) a more accurate ANI score was computed between each pair of close SGBs using the pyani tool (Pritchard et al., 2016) (version 0.2.6; option '-m ANIb'). Finally, iv) pairs of SGBs having an ANI score >95% were merged into a single SGB and v) the process was iterated until no more mergings were obtained. This merging operation reduced the number of SGBs of 3% resulting in a total of 16,332 distinct SGBs.

The obtained SGBs can be subdivided into 3 main groups (Figure 1B): i) the set of 1,134 known SGBs (kSGBs) that contain at least one reconstructed and one reference genome (the ''unknownness'' score for an SGB represents the number of reconstructed genomes with respect to the total number of genomes belonging to the SGB); ii) the set of 3,796 unknown SGBs (uSGBs) that contain at least one reconstructed genome, but no reference genomes from isolate sequencing or publicly available metagenomic assemblies; (iii) the set of 11,402 non-human SGBs, which contain at least one reference genome, but no reconstructed genomes. Results reported in the manuscript that involved computation of the ANI score for a number of genomes minor than 100 were done using pyani (Pritchard et al., 2016), while in the other cases we relied on the Mash (Ondov et al., 2016) estimates.

The kSGBs were taxonomically labeled with the species label associated with the reference genome(s) present in the bin, considering the most common species label if multiple reference genomes with different assigned species are present (Table S4). For uSGBs, no reference genomes were present in the species-level bins by definition, and we thus provided an assignment at higher taxonomic level. The same procedure used to find the optimal genomic divergence cutoff to define SGBs described above and in Figures S2A–S2C was adopted to define genus-level and family-level genomic divergence. Results showed a minimization of the error for a threshold equal to 0.15 and 0.30 for genus-level and family-level bins, respectively (Figure S2D), which was thus adopted to generate genus-level genome bins (GGBs) and family-level genome bins (FGBs). Although we are not proposing to modify the underlying taxonomy based on GGBs and FGBs, this additional clustering allowed us to give confident genus-level assignments to the 1,472 uSGBs falling in a GGB and a family label to 1,383 additional uSGBs falling in a FGB. Higher taxonomic levels are challenging to recapitulate by whole-genome clustering because of limitations in whole-genome nucleotide similarity quantification at large phylogenetic divergences, and we thus decided to maintain the remaining 941 unlabeled uSGB taxonomically unassigned. Nevertheless, for each SGB we report the full taxonomy of the closest matching genome and the whole-genetic distant from it to provide a genomic context for all SGBs (Table S4). The information about the closest labeled genome for the 941 uSGBs not assigned to a GGB or a FGB is used to assign them a phylum-level taxonomy in the text and in the figures. Finally, a taxonomic estimation based on 16S rRNA sequences was provided for 135 of these 941 uSGBs following the procedure described above in the section ''Quality control of metagenomic assemblies''.

The set of 159,803 genomes available at the NCBI as of September 2018 was also considered to verify that our set of reconstructed genomes adds a substantial amount of unknown diversity. Indeed, we found that there were only 644 genomes that belong to uSGBs (1.9% of the set of reconstructed genomes in our uSGBs) using the same 5% whole-genome nucleotide divergence threshold described above. These 644 genomes, along with future updates, are added to the final resource available for download and we will continue integrating our resource with additional metagenomic assemblies and reference genomes that become available.

Reconstruction of the human-microbiome phylogenetic structure

The phylogenetic analyses were performed with PhyloPhlAn (Segata et al., 2013) using the ''dev'' branch of the repository as of end of June 2018 (https://bitbucket.org/nsegata/phylophlan/overview).

The phylogeny in Figure 1A was built using the 400 universal PhyloPhlAn markers with the following options: ''--diversity high --accurate --min_num_markers 80.'' For the internal steps the following tools with their set of parameters were used:

- diamond (version v0.9.9.110, (Buchfink et al., 2015)) with parameters: ''blastx --quiet --threads 1 --outfmt 6 --more-sensitive --id 50 --max-hsps 35 -k 0'' and with parameters: ''blastp --quiet --threads 1 --outfmt 6 --more-sensitive --id 50 --max-hsps 35 -k 0'';
- mafft (version v7.310, (Katoh and Standley, 2013)) with the ''--anysymbol'' option;
- trimal (version 1.2rev59, (Capella-Gutiérrez et al., 2009)) with the ''-gappyout'' option;
- RAxML (version 8.1.15, (Stamatakis, 2014)) with parameters: ''-m PROTCATLG -p 1989.''

The phylogeny in Figure S3A was built using the 400 PhyloPhlAn markers with the following parameters: "--diversity high --fast --min_num_markers 80" and the set of external tools with the same options used for the phylogeny in Figure 1A described above, except for the phylogeny reconstruction step. In this case the phylogeny has been inferred using IQ-TREE (version 1.6.6, (Nguyen et al., 2015)) with the following parameters: "-nt AUTO -m LG."

The phylogenies in Figures 3C, S3B, S3C, S5, and S7B were built using their corresponding set of cores genes at 95% as identified by Roary (Page et al., 2015) and with the following parameters in PhyloPhlAn: "--diversity low --fast --min_num_markers <50% of the number of core genes identified>--min_num_entries <90% of the number of input genomes>." The external tools used by PhyloPhlAn and their corresponding parameters were:

- blastn (version 2.6.0+, (Altschul et al., 1990)) with parameters: "-outfmt 6 -max_target_seqs 1000000";
- mafft (version v7.310, (Katoh and Standley, 2013)) using the "L-INS-i" algorithm and with parameters: "--anysymbol --auto";
- trimal (version 1.2rev59, (Capella-Gutiérrez et al., 2009)) with the "-gappyout" option;
- FastTree (version 2.1.9, (Price et al., 2010)) with "-mlacc 2 -slownni -spr 4 -fastest -mlnni 4 -no2nd -gtr -nt" options;
- RAxML (version 8.1.15, (Stamatakis, 2014)) with parameters: "-p 1989 -m GTRCAT -t <phylogenetic tree computed by FastTree>."

The non-metric multidimensional scaling plots in Figures 4C and S6A were computed on pairwise genetic distances between core gene alignments produced by Roary using the nmds function in the ecodist R package (Goslee and Urban, 2007)

The phylogenetic trees in Figures 1A, 3C and S3A were generated using GraPhlAn (version 1.1.3, (Asnicar et al., 2015)) and the phylogenies in Figures 3A, S3B, S3C, S5, and S7B were generated using FigTree (version 1.4.3, http://tree.bio.ed.ac.uk/software/figtree/).

Quantification of the fraction of reads that can be mapped against SGBs

To assess the proportion of reads that could be mapped against the previously available set of genomes and the genomes we reconstructed here from metagenomes, we built four collections of sequences belonging to: a) the set of 12,563 genomes representing the kSGBs from the 80,990 reference genomes, by selecting one representative genome (the longest) for each kSGB; b) the residual set of 68,427 reference genomes for all the kSGBs; c) the set of 4,930 reconstructed genomes that are representatives for each SGB (Table S4); d) the residual set of 149,793 reconstructed genomes in all the SGBs. Additionally, we retrieved and indexed nine reference genomes for *Blastocystis* spp. (Beghini et al., 2017); 39 *Malassezia* spp. genomes from the NCBI-Assembly database (accessed in March 2018) and 18 assemblies from (Tett et al., 2017); and 13,924 plasmids and 10,529 viruses from RefSeq (release 90 (O'Leary et al., 2016)). To parallelize the downstream analysis and keep reasonably small the index files, 379 Bowtie2 (Langmead and Salzberg, 2012) databases were built. We then subsampled all the 9,428 samples used in this study to 1%, because of the very high computational requirement of the mapping (~1,100 CPU hours for each sample would be required for the mapping of full metagenomes). The raw reads were filtered to remove short reads (length lower than 70 bp) and low-quality reads (mean sequencing quality < 20). We mapped each sample against the human genome using Bowtie2 (in end-to-end mode, hg19 index) to remove human DNA contamination and samples harboring more than 10% human reads were excluded. We excluded duplicated samples present in multiple studies, and samples that, after the quality-filtering, had no remaining reads. The reads from the remaining 8,908 samples were then mapped against the 379 Bowtie2 indexes in end-to-end mode. We applied the same procedure to the 389 additional cross-validation samples (384 publicly available, see above, and 5 sequenced gut metagenomes from Ethiopia). The resulting maping files were filtered to remove alignments with an alignment score (AS:i tag) lower than −20 to exclude spurious alignments that could influence the mappability assessment. For each sample, we computed the fraction of reads confidently mapping to each set of indexes and counted them according to the following criteria: **i)** reads aligning to at least one representative reference genome; **ii)** reads not aligning to i) and aligning to at least one other reference genome; **iii)** reads not aligning to i) and ii) and aligning against one of the 4,930 SGBs representatives; and **iv)** reads aligning only against one of the residual 149,793 reconstructed genomes. We followed the same incremental strategy to determine the fraction of residual reads mapping to micro-eukaryotes (*Blastocystis* spp., *Malassezia* spp.), plasmids and viruses. We reported in Figures 2A–2B and in Figure S4 the percentage of reads in each of these four categories (representative reference genomes, other reference genomes, representative SGBs and non-representative SGBs) with respect to the number of HQ non-human reads in each sample.

Pangenome, phylogenetic, and functional analysis of kSGB and uSGBs

We used Prokka (version 1.12, (Seemann, 2014), with default parameters) for annotating the reference and the reconstructed genomes of the 10 *Bacteroides* kSGBs. The annotated genomes were then processed with Roary (version 3.8, (Page et al., 2015) with "-e -z -g 1000000" params) for the pangenome analysis and to identify the set of core genes. The core genes (at 95% gene family clustering identity threshold) identified by Roary were then used as a database in PhyloPhlAn for phylogenetic analyses. Functional annotation was performed using EggNOG mapper (version 1.0.3, (Huerta-Cepas et al., 2017)) based on EggNOG orthology data (Huerta-Cepas et al., 2016), and the sequence searches were performed using HMM. For the functional profiles shown in Figures 3E and 5E, we used the Brite Hierarchy from KEGG to screen metabolic related pathways and KOs among all the KOs annotated by EggNOG. We employed the same EggNOG pipeline to functionally annotate all the 4,930 representative of the SGBs (Table

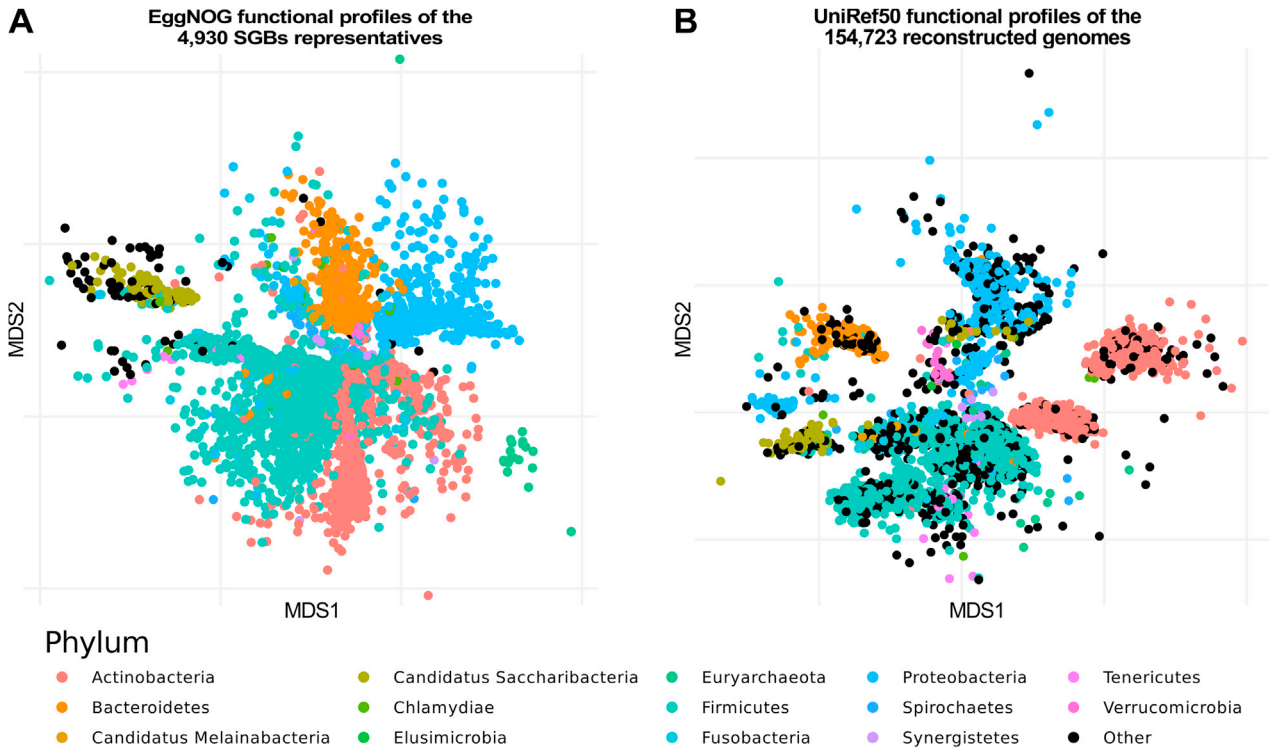
S4). Figure S1A shows, based on the presence/absence of the EggNOG ortholog to which a KEGG KO is associated, an ordination plot relating each of the 4,930 SGBs from the functional point of view. All the 154,723 reconstructed genomes were functionally annotated by mapping them against Uniref90 and Uniref50 using diamond (version v0.9.9.110). The UniRef-based functional profiles are shown in the ordination plot in the Figures S1B–S1D. All functional profiles (EggNOG-based and UniRef-based) are available for download at the supporting website (see Data and software availability).

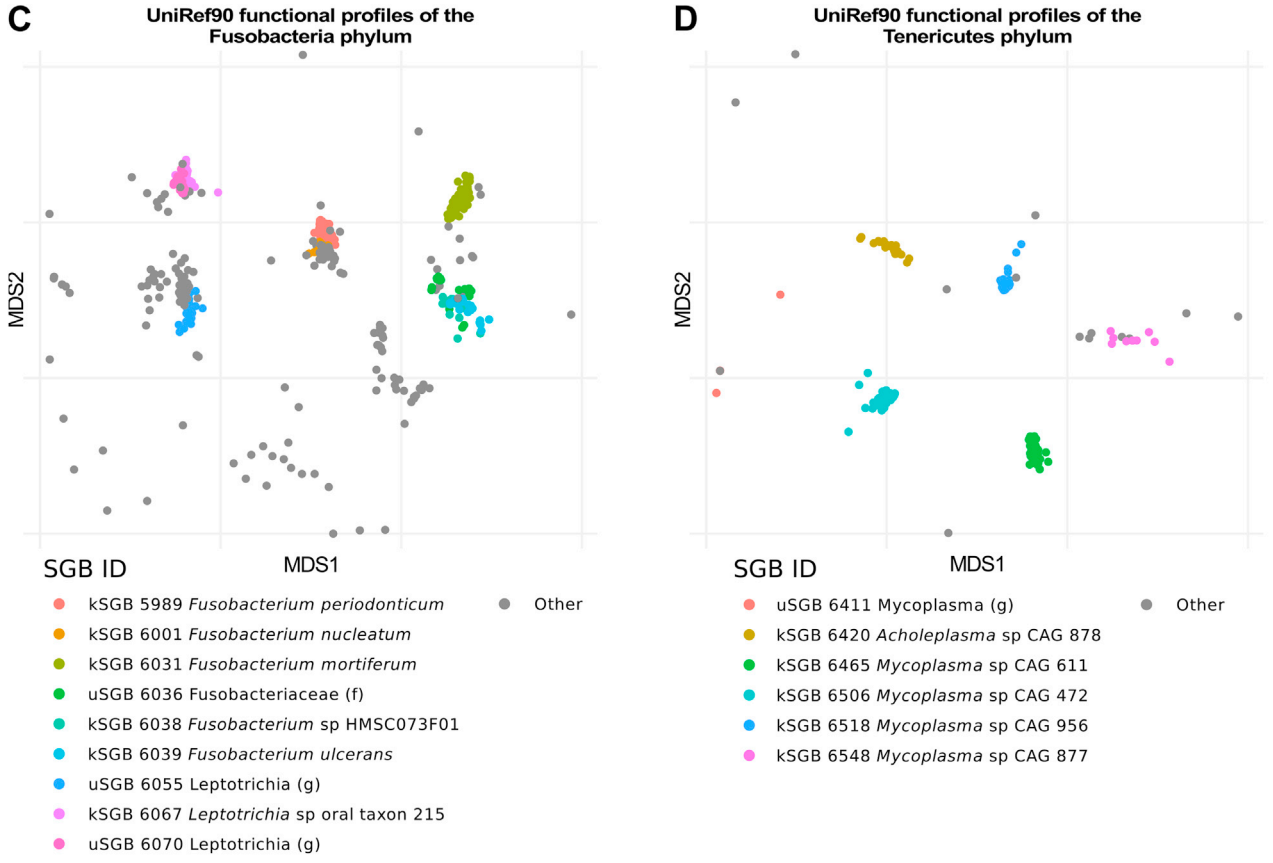
QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical significance was verified through Fisher's test, Mann-Whitney U-test, or Welch's t-test as reported in the text. Multiple hypothesis testing correction was done using the Bonferroni or the false discovery rate (FDR) method as also reported in the manuscript. All other computational and statistical analyses were performed with the open source software tools referenced in the STAR Methods along with the described procedures.

DATA AND SOFTWARE AVAILABILITY

All the recovered genomes, SGBs, and functional profiles (eggNOG- and UniProt-based) are available at http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html and at http://opendata.lifebit.ai/table/?project=SGB. The raw sequencing data for the sequenced datasets are available in NCBI-SRA under the BioProject: PRJNA485056 (Madagascar cohort) and PRJNA504891 (Ethiopia cohort). The proposed representative genome of "*Candidatus* Cibiobacter qucibialis" has been deposited at DDBJ/ENA/GenBank under the accession SAUS00000000, assembled from NCBI-SRA accession ERS1343406. The metadata for all the samples considered are available in *curatedMetagenomicData* (Pasolli et al., 2017) at http://waldronlab.io/curatedMetagenomicData/, and all the other considered genomes and metagenomes are publicly available in NCBI. We also included in the resource the list of 644 genomes that recently became available in NCBI and the link to their uSGBs. Assembled contigs are available at http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html, and software generated in this study is open source and available at https://bitbucket.org/CibioCM/cmseq/src/default/.

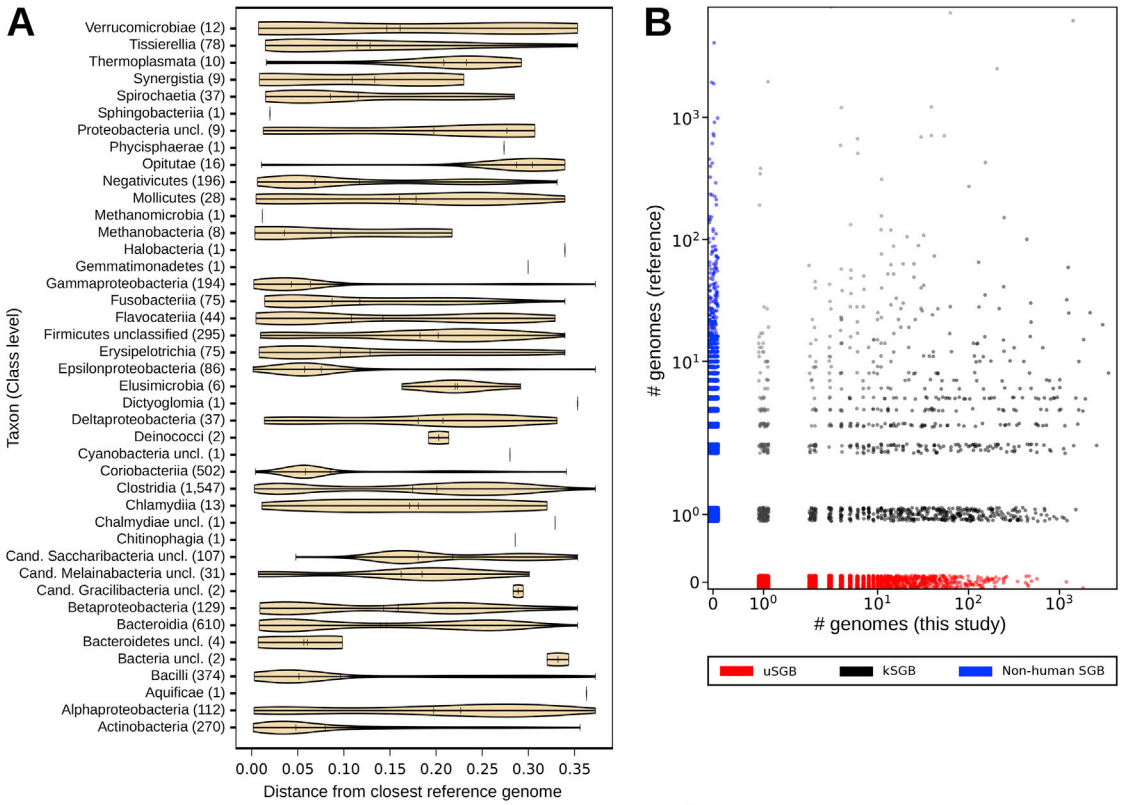




(legend on next page)

Figure S1. Overview of the Functional and Metabolic Annotations of the Representatives of the SGBs and of the Whole Set of 154,723 Reconstructed Genomes, Related to Figure 1

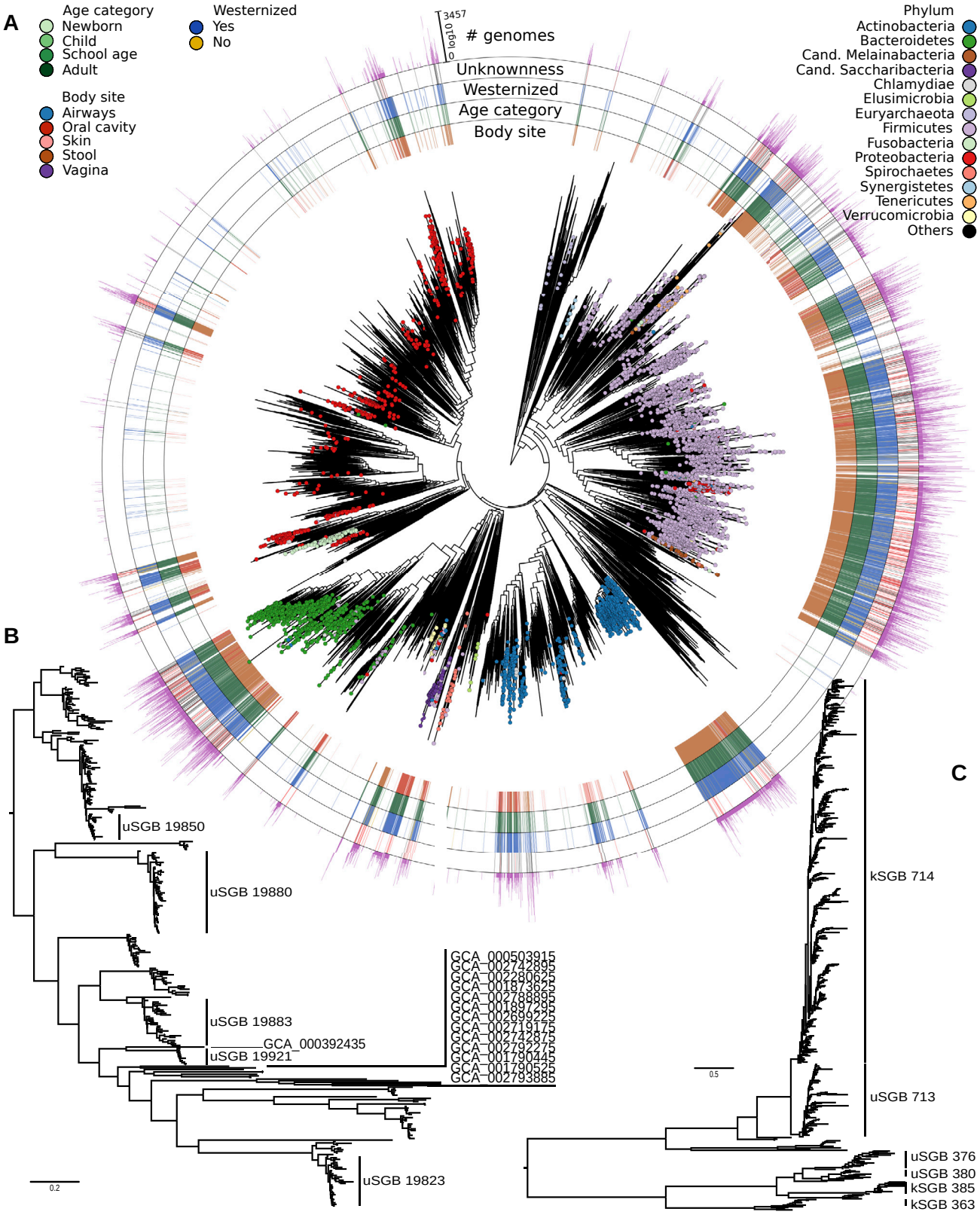
(A) Ordination plot of the KEGG gene families annotated using eggNOG (see STAR Methods) of the 4,930 SGBs' representatives, colored by the 14 most represented phyla. (B) Ordination plot of the UniRef50 gene families present in the 154,723 reconstructed genomes as annotated by mapping the genomes against both Uniref90 and Uniref50 (see STAR Methods). Ordination plots of the UniRef90 gene families for all the reconstructed genomes assigned to the (C) Fusobacteria and (D) Tenericutes phyla are also reported as examples of fine-grained functional differentiation.



(legend on next page)

Figure S2. Overview of the Reconstructed SGBs and Criteria for SGB Definition and Taxonomic Assignment, Related to Figure 6

(A) Distribution of the distances of each reconstructed genome to the closest available isolate genomes, grouped by the class assigned to the matching isolate genomes. (B) The 4,930 identified species-level genome bins (SGBs) comprise a very variable fraction of already available genomes versus genomes we reconstructed from metagenomes. (C) Minimization criterion adopted to find the optimal cutoff in the hierarchical clustering of genomes to define SGBs. Two criteria are taken into account: minimization of the over-clustering error (C-i), and minimization of the under-clustering error (C-ii). Results showed a minimization of the error for a threshold equal to 0.05 (C-iii), which was thus adopted to discretize subtrees in the dendrogram and generate SGBs spanning ~5% genetic diversity. (D) The same minimization criterion reported in (C-iii) for species-level bins is also adopted to identify the genomic diversity for genus-level and family-level bins.



(legend on next page)

Figure S3. Phylogenetic Trees for All SGBs and Reference Genomes and Subtrees of Saccharibacteria and Archaea, Related to Figure 1

(A) Phylogenetic tree that includes the representatives of the SGBs presented in Figure 1A together with all the non-human bins (represented in white in the external rings), for a total of 16,332 genomes (15,299 after the internal quality control in PhyloPhlAn). (B) Phylogenetic tree of the 337 reconstructed genomes taxonomically assigned to the candidate phylum Saccharibacteria present in the 108 SGBs, including available reference genomes (publicly available reference genomes are labeled with the "GCA" prefix). (C) Phylogenetic tree of the 675 archaeal genomes reconstructed in this study. 487 genomes belong to the *Methanobrevibacter smithii* kSGB (ID 714).

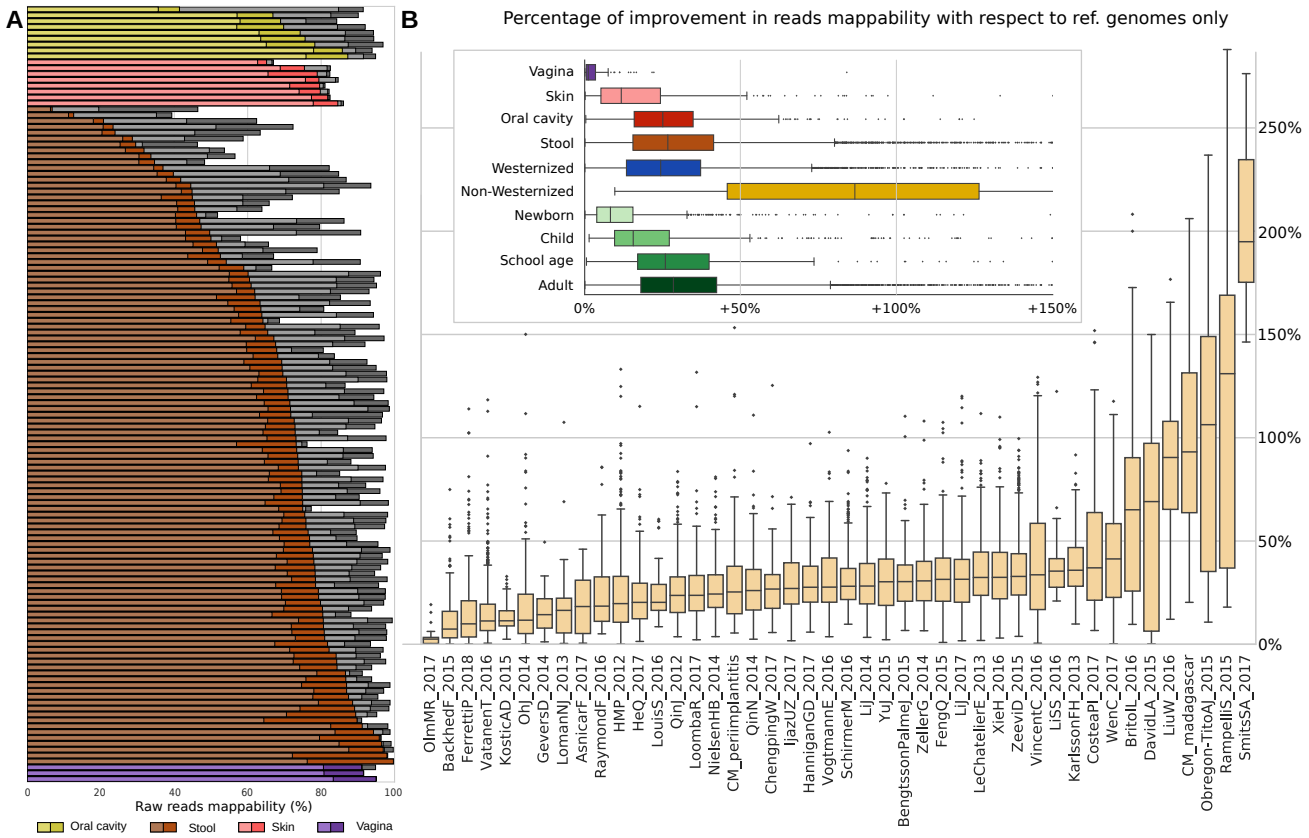


Figure S4. Improvement of Read Mappability Statistics by Considering the Set of Microbial Genomes We Assembled in This Work, Related to Figure 2

(A) Fraction of reads that can be mapped against different sets of genomes from isolate sequencing and the metagenomically reconstructed genomes. A subset of 132 full (i.e., not subsampled) metagenomes is shown (3 metagenomes randomly selected from each study). Samples are colored and grouped by body site. The colored part of the bar refers to the reads that can be mapped against a previously available reference genome, while the gray bars extend to highlight the total mappability we achieved using the 154,723 microbial genomes reconstructed in this study. (B) Percentage of increase in the mappability when using also the 154,723 reconstructed SGBs to map metagenomic reads. Boxplots represent values grouped by body site, lifestyle, age category (upper panel) and study (lower panel). The percentage of improvement is calculated with respect to the fraction of reads that could map using only and all the reference genomes. All the 9,428 metagenomes used in this study were mapped after being subsampled at 1% (see STAR Methods). Averaged statistics are reported in Figures 2A–2B.

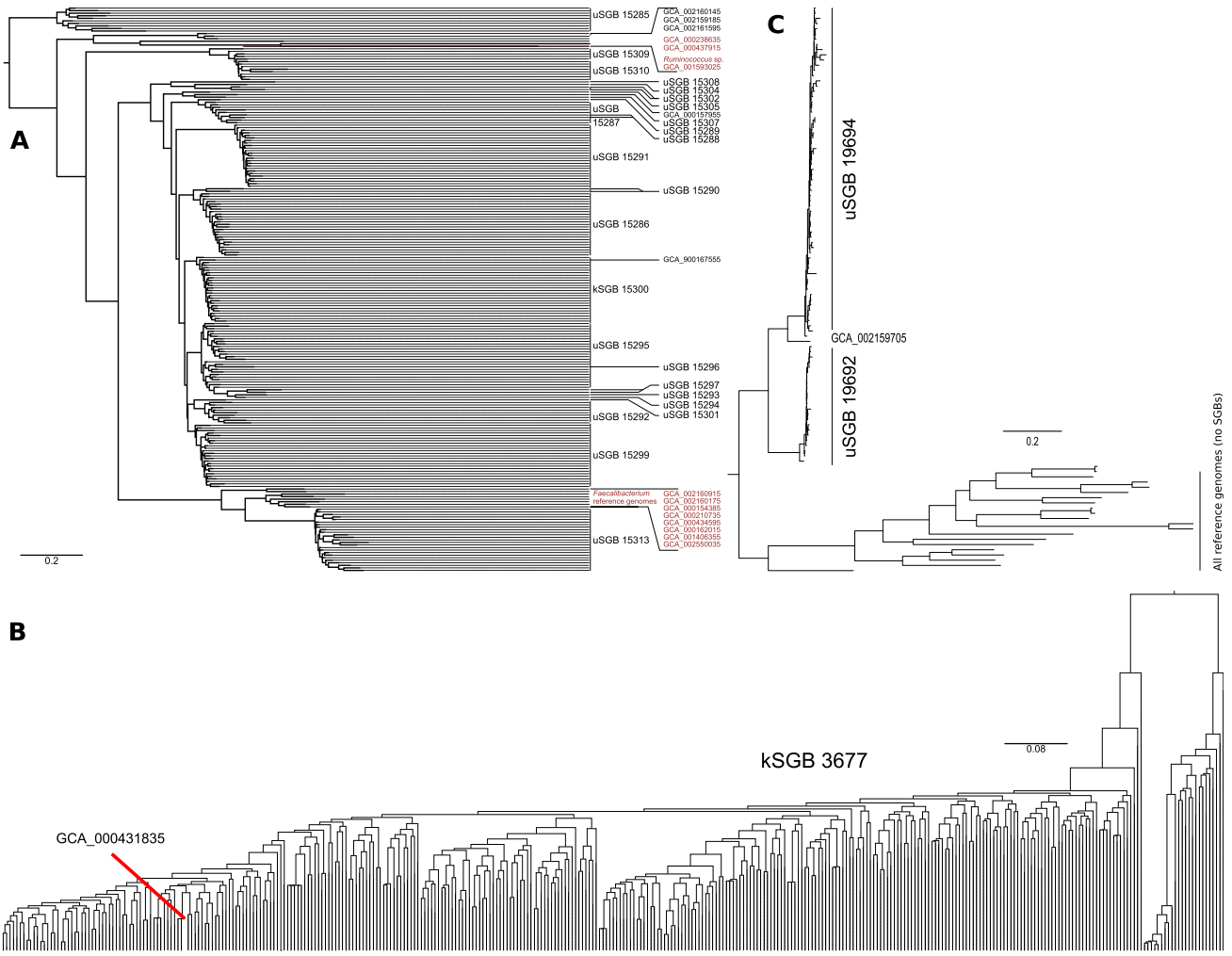
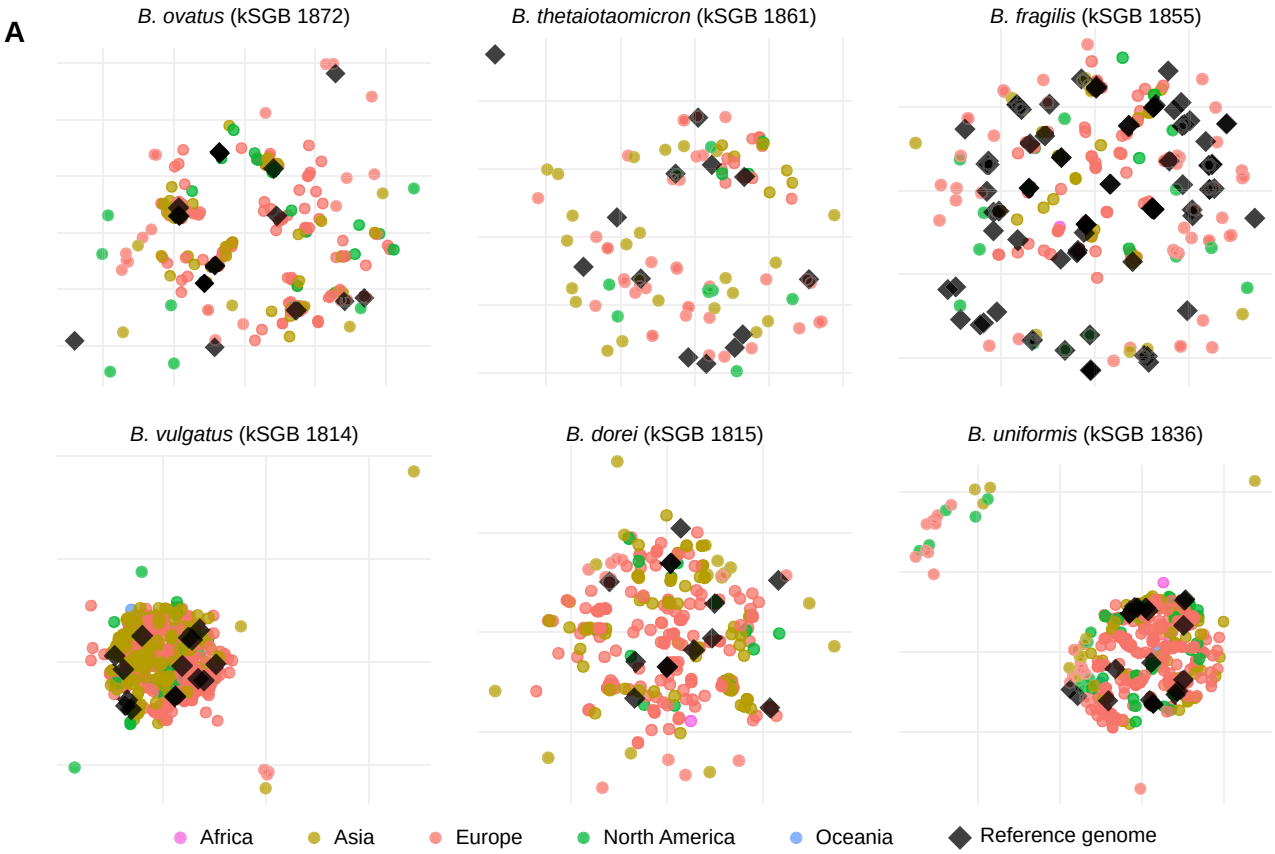
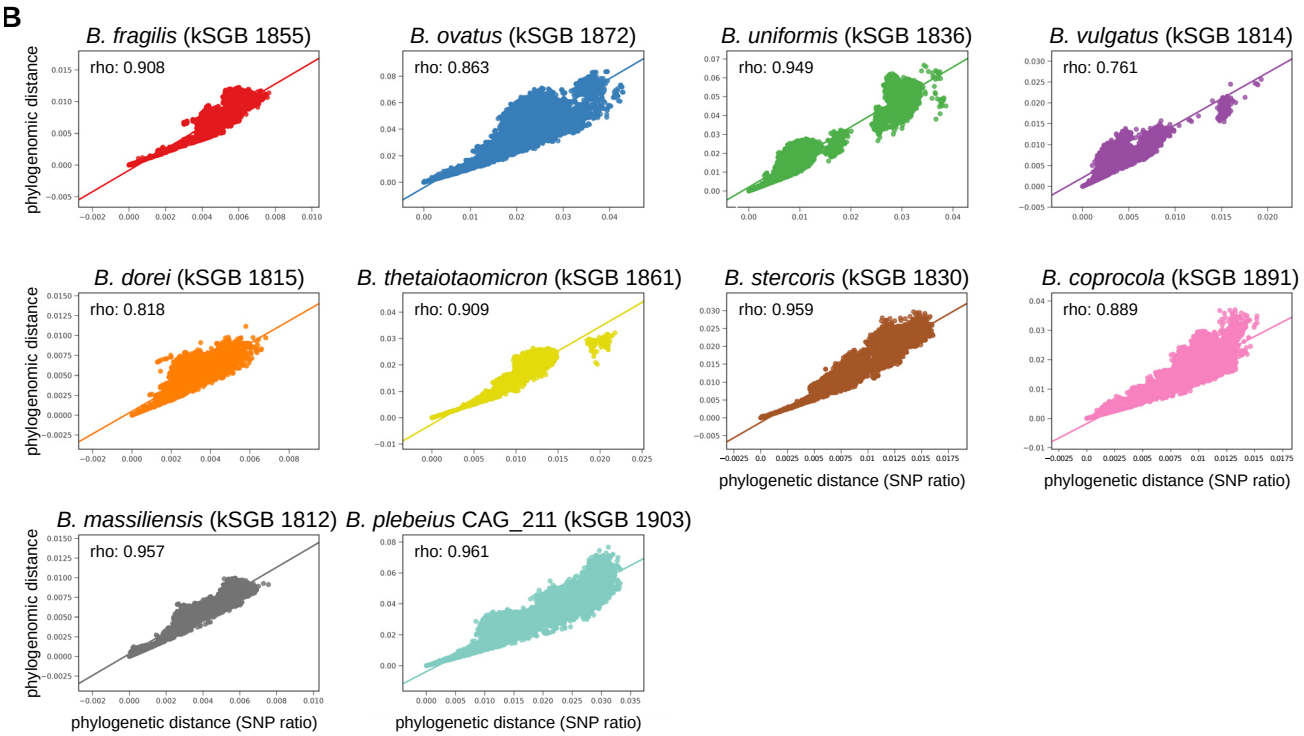


Figure S5. Phylogenetic Trees for SGBs Placed between *Ruminococcus* and *Faecalibacterium*, Succinatimonas kSGB (ID 3677), and Two Elusimicrobia uSGBs, Related to Figure 3 and 5

(A) Phylogenetic tree of SGBs placed between reference genomes for *Ruminococcus* and *Faecalibacterium* species in Figure 1A (highlighted in red), as already reported in Figure 3A but without collapsed branches and including the two reference genomes GCA_000238635 and GCA_000437915 (also highlighted), originally labeled as *Subdoligranulum* sp. 4_3_54A2FAA and *Subdoligranulum* sp. CAG:314, respectively. (B) Phylogenetic tree of the Succinatimonas kSGB (ID 3677) including the only available reference genome. (C) Phylogenetic tree of the two Elusimicrobia uSGBs enriched in non-Westernized populations and of all the available Elusimicrobia reference genomes.





(legend on next page)

Figure S6. Genetic Diversity and Correlation between Genetic and Functional Similarity for Bacteroides Species, Related to Figure 4

(A) MDSs on intra-SGB genetic distances for *Bacteroides* species not reported in Figure 4C. (B) Scatterplots for the ten most prevalent *Bacteroides* kSGBs showing the relation between pairs of genomes measured as branch length distance on the core-genome-based phylogenetic tree (x axis) and as branch length on the hierarchical clustering built on the presence and absence of pan-genes (phylogenomic distance, y axis).

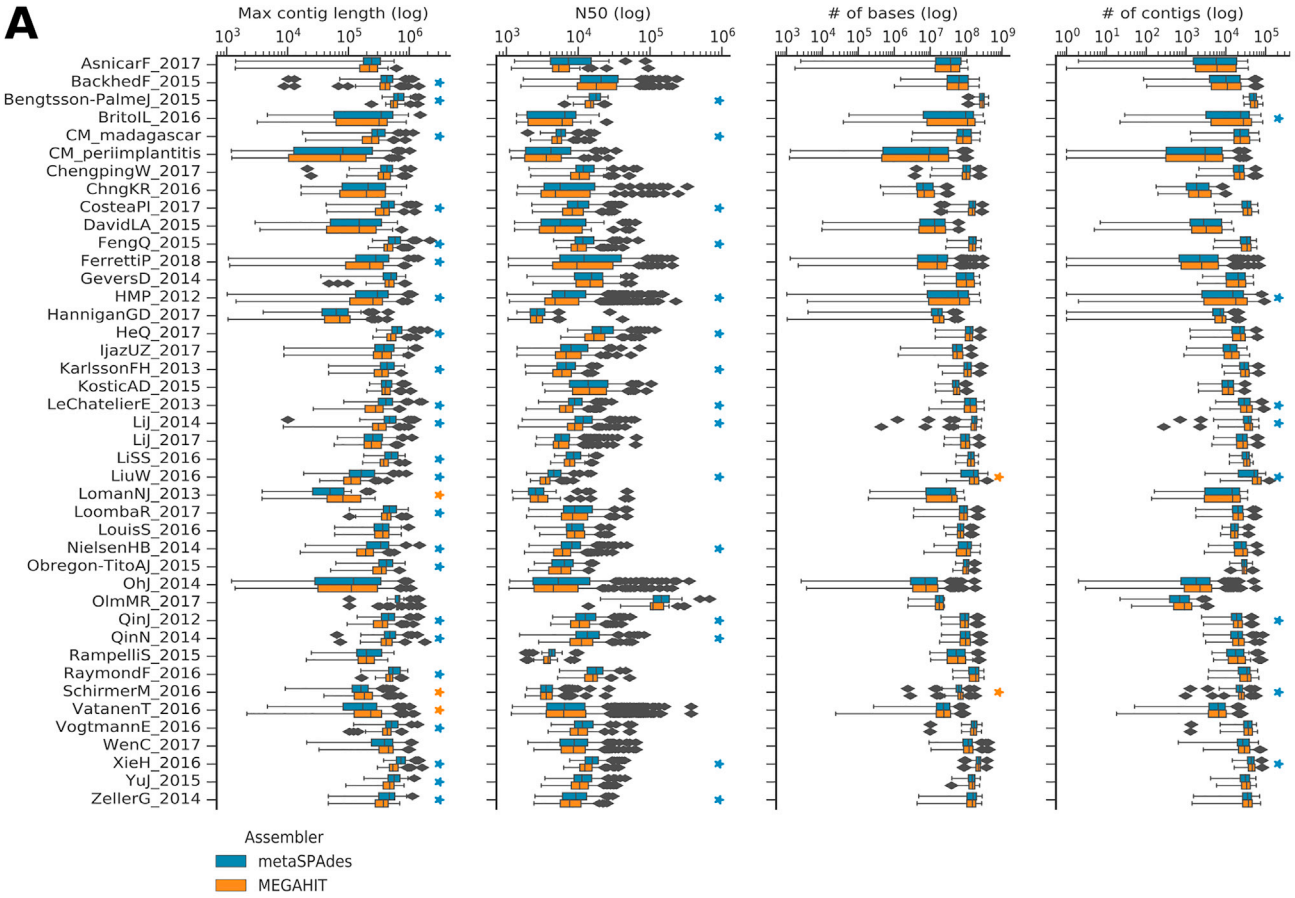


Figure S7. Comparison between MEGAHIt and metaSPAdes Assemblies and between Assembly and Co-assembly, Related to Figure 7

(A) Comparison between metaSPAdes and MEGAHIT assemblers across all the considered datasets confirms that metaSPAdes performs consistently better especially in recovering long contigs. Stars indicate statistically significance (Welch's t test, $p < 0.05$). (B) Phylogenetic tree built on the genomes of gut adult metagenomes from 25 women from the FerrettiP_2018 dataset showing comparison between the set of single-sample assembled genomes (in green) and co-assembled genomes (in red). Several genomes reconstructed with the two approaches have the same phylogenetic placement, with single-sample assembly retrieving a total of 605 genomes spanning 257 SGBs, while co-assembly retrieved 172 genomes.