Elevated rates of horizontal gene transfer in the industrialized human microbiome

Mathieu Groussin, Mathilde Poyet, Ainara Sistiaga, Sean Kearney, Katya Moniz, Mary Noel, Jeff Hooker, Sean Gibbons, Laure Segurel, Alain Froment, et al.

Title

Elevated rates of horizontal gene transfer in the industrialized human microbiome

Authors

Mathieu Groussin*[1,2,3,4], Mathilde Poyet*[1,2,3,4], Ainara Sistiaga[4,5,6], Sean M. Kearney[1,2], Katya Moniz[1,2,4], Mary Noel[4,7], Jeff Hooker[4,7], Sean M. Gibbons[4,8,9], Laure Segurel[4,10], Alain Froment[4,11], Rihlat Said Mohamed[12], Alain Fezeu[4,13], Vanessa A. Juimo[4,13], Sophie Lafosse[10], Francis E. Tabe[14], Catherine Girard[4,15,16], Deborah Iqaluk[4,17], Le Thanh Tu Nguyen[1,2,3,4], B. Jesse Shapiro[4,15], Jenni M. S. Lehtimäki[4,18,19], Lasse Ruokolainen[4,18], Pinja P. Kettunen[4,18], Tommi Vatanen[3,4,20], Shani Sigwazi[4,21], Audax Mabulla[4,22], Manuel Domínguez-Rodrigo[4,23,24], Yvonne A. Nartey[4,25], Adwoa Agyei-Nkansah[4,26], Amoako Duah[4,27], Yaw A. Awuku[4,28], Kenneth A. Valles[4,29], Shadrack O. Asibey[4,30], Mary Y. Afihene[4,31], Lewis Roberts[4,32], Amelie Plymoth[4,25], Charles A. Onyekwere[4,33], Roger E. Summons[4,5], Ramnik J. Xavier[3,4,34], Eric J. Alm[1,2,3,4].

* These authors equally contributed to this work.

Affiliations

1. Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
2. Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
3. The Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA
4. The Global Microbiome Conservancy, Massachusetts Institute of Technology, Cambridge, MA, 02142, USA
5. Department of Earth, Atmospheric and Planetary Science, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
6. Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark.
7. Chief Dull Knife College, Lame Deer, MT, 59043, USA
8. Institute for Systems Biology, Seattle, WA 98109, USA
9. eScience Institute, University of Washington, Seattle, WA 98195, USA
10. UMR7206 Eco-anthropologie, CNRS-MNHN-Univ Paris Diderot-Sorbonne, France
11. Institut de Recherche pour le Développement UMR 208, Muséum National d'Histoire Naturelle, Paris, France
12. SA MRC / Wits Developmental Pathways for Health Research Unit, Department of Paediatrics, School of Clinical Medicine, Faculty of Health Sciences, University of Witwatersrand, Johannesburg, South Africa
13. Institut de Recherche pour le Développement, Yaounde, Cameroon
14. Faculté de Médecine et des Sciences Biomédicales - Université Yaoundé 1, Cameroun
15. Université de Montréal, Département de sciences biologiques, C.P. 6128, succursale Centre-ville, Montreal, Quebec H3C 3J7, Canada
16. Centre d'études nordiques, Département de biochimie, de microbiologie et de bio-informatique, Université Laval, 1030 rue de la Médecine, Québec (QC) Canada G1V0A6
17. Resolute Bay, Nunavut X0A 0V0 Canada
18. Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental sciences, University of Helsinki, Finland
19. COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Ledreborg Alle 34, 2820, Gentofte, Denmark

20. The Liggins Institute, University of Auckland, Auckland, 1023, New Zealand
21. Tumaini University Makumira, Arusha, Tanzania
22. Archaeology Unit, University of Dar es Salaam, Dar es Salaam, Tanzania
23. Department of Prehistory, Complutense University, Madrid, Spain
24. Institute of Evolution in Africa, University of Alcalá de Henares, Madrid, Spain
25. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
26. Department of Medicine and Therapeutics, University of Ghana Medical School and Korle Bu Teaching Hospital, Accra, Ghana
27. Department of Medicine, St. Dominic Hospital, Akwatia, Ghana
28. Department of Internal Medicine and Therapeutics, School of Medical Sciences University of Cape Coast, Cape Coast, Ghana
29. Medical Scientist Training Program, Mayo Clinic, 200 First Street SW, Rochester, 55905, Minnesota, USA
30. Catholic University College, Sunyani, Ghana
31. Department of Medicine, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
32. Division of Gastroenterology and Hepatology, Mayo Clinic, 200 First Street SW, Rochester, 55905, Minnesota, USA
33. Department of Medicine, Lagos State University College of Medicine, Lagos, Nigeria
34. Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

Abstract

Industrialization has impacted the human gut ecosystem (e.g. through changes in diet or medical practices), resulting in altered microbiome composition and diversity. Whether bacterial genomes may also adapt to the industrialization of their host populations remains largely unexplored. Here, we investigate the extent to which the rates and targets of horizontal gene transfer (HGT) vary across thousands of bacterial strains from 15 human populations spanning a range of industrialization. We show that HGTs have accumulated in the microbiome over recent host generations, and that HGT occurs at high frequency within individuals. Comparison across human populations reveals that industrialized lifestyles are associated with higher HGT rates and that the functions of HGTs are related to the level of host industrialization. Our results suggest that gut bacteria continuously acquire new functionality based on host lifestyle and that high rates of HGT may be a recent development in human history linked to industrialization.

Keywords: human gut microbiome; industrialization; urbanization; lifestyle; horizontal gene transfer; bacterial genomics; host-microbe interactions; culturomics; antimicrobial resistance; virulence.

Introduction

Transitioning from non-industrialized to industrialized lifestyles is associated with changes in gut microbiome composition and decreased bacterial species diversity (Brewster et al., 2019; Hansen et al., 2019; McDonald et al., 2018; Pasolli et al., 2019; Schnorr et al., 2014; Sonnenburg and Sonnenburg, 2019b; Yatsunenko et al., 2012). While the precise causes of these changes are unknown, factors associated with the development of industrialized societies such as sanitation, the consumption of processed food, higher frequency of caesarean section, and increased use of antibiotics likely play key roles in remodeling the gut microbiome (Sonnenburg and Sonnenburg, 2019a). These perturbations in the gut ecosystem can occur shortly after individuals transition from non-industrialized to industrialized areas, and persist for years (Vangay et al., 2018), further confirming that lifestyle strongly influences the function of our gut microbiome. However, the effects of host and environmental factors associated with industrialized lifestyles on individual gut bacterial genomes are poorly characterized.

Bacteria can use horizontal gene transfer (HGT) to adapt rapidly to unstable environments through the acquisition of new functions. Mammalian gut bacteria have experienced frequent HGT events over millions of years of evolution (Hehemann et al., 2010; Smillie et al., 2011). Previous studies of specific bacterial species showed that HGT can occur and be conserved in the gut microbiome within a single individual (Coyne et al., 2014; Garud et al., 2019; Munck et al., 2020; Yaffe and Relman, 2019; Zhao et al., 2019; Zlitni et al., 2020), especially when there is strong selection for target functions such as antibiotic resistance (Forsberg et al., 2012; Lopatkin et al., 2017; Modi et al., 2013). Yet it remains unclear whether HGT can occur broadly enough to impact gut microbiome function over an individual's lifetime – such as in response to significant lifestyle changes – or whether microbiomes primarily acquire new functions through the acquisition of new strains. It was previously observed that individual bacterial strains can reside within a host microbiome for decades (Faith et al., 2013). So if the rate of gene transfer is sufficiently rapid, then a microbiome that is 'stable' in terms

of bacterial populations (Faith et al., 2013; Gibbons et al., 2017; Mehta et al., 2018) could nonetheless evolve in response to host-specific environmental perturbations through HGT, perhaps in response to changes in host lifestyle.

In a previous study (Smillie et al., 2011), we found high levels of HGTs in the human microbiome involving >500bp length sequences with greater than 99% similarity. Those results lacked the temporal resolution and the diversity in human populations necessary to address the questions of timescales and host lifestyle. Over short evolutionary timescales, the substitution rate of many bacterial species typically falls in the range of ~1 SNP/genome/year (Didelot et al., 2016; Drake, 1991; Duchêne et al., 2016; Zhao et al., 2019). Assuming this rough molecular clock approximation, and a genome size of 10^6 bp, the HGTs we detected using those criteria (>500bp, >99% similarity) were consistent with transfer events that occurred between 0 and 10,000 years ago (which corresponds to the time during which a 500bp sequence can accumulate a maximum of 1% sequence divergence, *i.e.* 5 SNPs). Variations in the molecular clock across species and genomic regions may shorten or expand this time interval. In any case, our previous results could not constrain the dates of HGT that occurred more recently than the rise of modern industrialization, dated to the 18-19th century (Vries and de Vries, 1994). To answer the question of whether commensal strains can frequently acquire new functionality through HGT within an individual, such that recent adaptations to industrialization are detectable in contemporary bacterial genomes, more precise estimates of the rate and extent of HGT are needed.

Existing reference isolate genomes (Browne et al., 2016; Faith et al., 2013; Forster et al., 2019; Goodman et al., 2011; Zou et al., 2019) originate almost exclusively from industrialized populations and, for the vast majority of strains, from different individuals, making investigation of within-person HGT impossible. Here, we present the Global Microbiome Conservancy (GMbC) isolate collection, composed of >4,000 cultured, isolated, and sequenced gut bacteria from diverse industrialized and non-industrialized populations,

including rich sets of strains from single individuals. We used these genomes to investigate the rate and patterns of gene transfers that occurred very recently in human history. We show that HGTs can occur at high and heterogeneous frequency within individuals, and we report elevated rates of gene transfer in industrialized populations.

Results

A diverse collection of bacterial isolate genomes from worldwide gut microbiomes

We cultured, isolated, and whole-genome sequenced 4,149 gut bacteria from 37 individuals from 14 distinct populations with different levels of industrialization (Fig 1A & B). Bacteria were isolated from stool samples under anaerobic conditions, using previously published protocols (Poyet et al., 2019). We combined these new genomes with a set of 3,632 isolate genomes that we recently generated from 11 urban American donors (Poyet et al., 2019), yielding a dataset of 7,781 isolate genomes. We then divided our cohort of 48 individuals according to two different parameters, which we defined as: “urban” vs. “rural” (based on local population density) (SEDAC Population Estimation Service, 2015), and “industrialized” vs. “non-industrialized” (based on the Human Development Index at the country level, HDI) (United Nations Development Program, 2020). For the purposes of this analysis, we used HDI as a proxy for industrialization because it reflects parameters that are relevant to health and the microbiome, *e.g.* the consumption of processed foods, rates of non-communicable diseases, sanitation infrastructure, and health expenditure (United Nations Development Program, 2020). This classification system yielded four groups of different lifestyles: rural non-industrialized populations from Tanzania, Cameroon and Ghana; urban non-industrialized populations from Ghana, Nigeria and Cameroon; rural industrialized populations from Canada, Finland and the USA; and urban industrialized groups from Finland and the USA; see Fig 1A & B, Supp. Fig. 1 & Supp. Table 1 for descriptions of population ethnicity, location, population density, HDI, subsistence strategy, and microbiome

composition. The non-industrialized rural cohort includes populations with diverse subsistence strategies, including hunter gatherers, pastoralists, fishermen, and farmers (Fig. 1B).

We grouped our 7,781 isolate genomes into species clusters based on genomic similarity, using the Mash distance as a proxy for Average Nucleotide Identity (see Methods). This identified 339 bacterial species across 6 phyla, grouping into 73 known and 88 unknown genera (Figure 1C & Supp. Tables 2 & 3 for culturing data and genome assembly statistics). We compared our genome collection to the Unified Human Gastrointestinal Genome (UHGG) database, which comprises the largest set of human gut bacterial genomes, with the vast majority being metagenome-assembled genomes from uncultivated bacterial species (Almeida et al., 2020). We measured genomic distances between our representative genomes and all UHGG representative genomes with Mash and looked at the number of species that have not been previously sequenced or cultured. We found that 13% of the species in our collection represent newly characterized species, and 41% represent newly cultivated species (Fig. 1D). We sampled a median of 93 isolate genomes and 17 species per individual, covering a wide range of within-people bacterial taxonomies and *in vivo* abundances (Fig. 1E and Supp. Table 4), providing within-person genomic and ecological diversity for high-resolution investigation of HGTs.

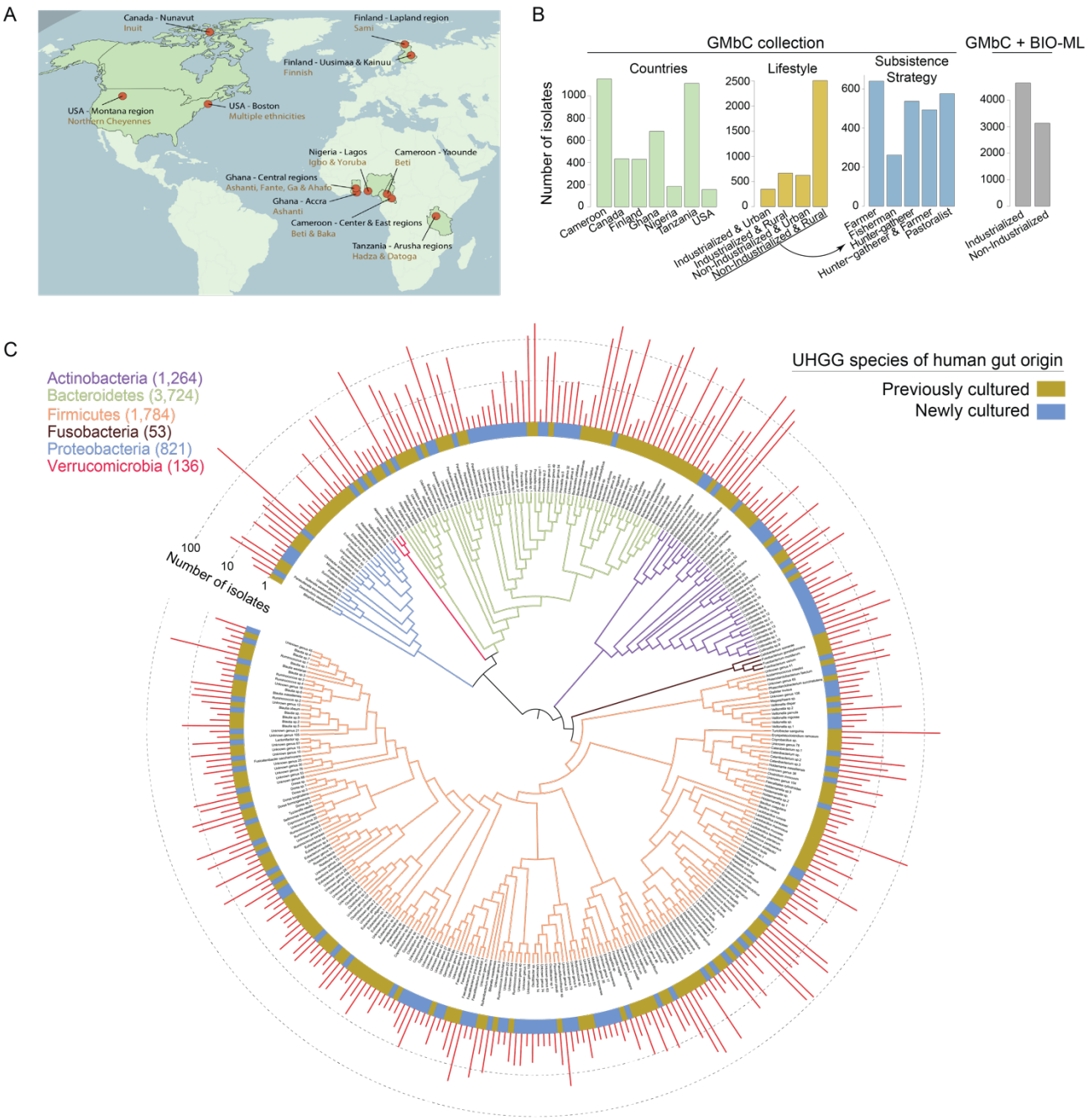
A
Canada - Nunavut
Inuit
Finland - Lapland region
Sami
Finland - Uusimaa & Kainuu
Finnish
USA - Montana region
Northern Cheyennes
USA - Boston
Multiple ethnicities
Nigeria - Lagos
Igbo & Yoruba
Cameroon - Yaounde
Beti
Ghana - Central regions
Ashanti, Fante, Ga & Ahafo
Ghana - Accra
Ashanti
Cameroon - Center & East regions
Beti & Baka
Tanzania - Arusha regions
Hadza & Datoga
B
GMbC collection
GMbC + BIO-ML
Countries
Lifestyle
Subsistence Strategy
Number of isolates
Cameroon
Canada
Finland
Ghana
Nigeria
Tanzania
USA
Industrialized & Urban
Industrialized & Rural
Non-Industrialized & Urban
Non-Industrialized & Rural
Farmer
Fisherman
Hunter-gatherer
Hunter-gatherer & Farmer
Pastoralist
Industrialized
Non-Industrialized
C
Actinobacteria (1,264)
Bacteroidetes (3,724)
Firmicutes (1,784)
Fusobacteria (53)
Proteobacteria (821)
Verrucomicrobia (136)
UHGG species of human gut origin
Previously cultured
Newly cultured
100
10
1
Number of isolates
D
UHGG genomes - all
UHGG genomes - cultivated & of human gut origin
Mash distance to closest UHGG genome
Bacterial species in our dataset (GMbC + BIO-ML)
E
ln vivo abundance (log) of species in our isolate collections
Phylum
Actinobacteria
Bacteroidetes
Firmicutes
Fusobacteria
Proteobacteria
Verrucomicrobia
Kraken2 OTUs
Absent
Present
GMbC
BIO-ML
Individuals

Figure 1 - Assembly of a geographically, phylogenetically and ecologically diverse collection of human gut bacterial isolate genomes.
(A) Samples were collected from 15 communities in the USA, Canada, Finland, Cameroon, Tanzania, Ghana, and Nigeria. Red dots show the geographic locations of sampling sites. Participants represented four different lifestyle categories: 14 urban industrialized (UI) individuals in the USA (Boston area – various ethnicities), eastern Finland (Kainuu - Finnish people), and southern Finland (Helsinki - Finnish people); 5 rural industrialized (RI) individuals in the USA (Montana - Northern Plain Tribes people), arctic Finland (Lapland - Sami people), and the Canadian arctic (Nunavut - Inuit people); 6 urban non-industrialized (UN) individuals in Cameroon (Yaounde - Beti people), Nigeria (Lagos - Igbo and Yoruba people), and in Ghana (Accra - Ga and Ahafo people); and 23 rural non-industrialized (RN) individuals in Cameroon (Ngoantet, Center region - Beti people; Mintom, East region - Baka people), in Tanzania (Arusha region - Hadza and Datoga peoples), and in Ghana (Ampenyi - Central region, Fante people; Barekuma, Ashanti region - Ashanti people). See Supplementary Table 1 for further information on the demographics and subsistence strategy of each individual and community recruited in this study (agriculturalists, hunter-gatherers, farmers, fishermen, etc). **(B)** Distribution of isolate genomes across countries, lifestyles and subsistence strategies. For investigating HGT, we completed the GMbC genome collection with the BIO-ML collection composed of bacterial genomes isolated from individuals living an industrialized and urban lifestyle in the USA (Boston area - mixture of ethnicities). **(C)** Phylogenetic tree of representative genomes of all 339 bacterial species in our isolate genome collections (GMbC + BIO-ML). The inner ring shows species which, prior to our work, did not have representative genomes among the cultured bacteria of human gut origin in the UHGG database (shown in blue). The outer ring shows the distribution of isolate genomes across all species in the GMbC+BIO-ML collection. The total number of isolate genomes per phylum is shown. **(D)** Genomic distance between each representative genome of the GMbC+BIO-ML collection and the closest representative genome of the UHGG database. Orange dots show results with all UHGG genomes, which includes metagenome-assembled genomes (MAGs). Green dots show comparisons only with genomes from cultivated bacteria of human gut origin. The red dash line shows the threshold (D=0.05) that is classically used to delineate bacterial species. **(E)** *In vivo* abundance of all species in the GMbC+BIO-ML collection, across all individuals. Individuals are colored by lifestyle category (UI in orange, RI in green, UN in blue and RN in purple). Abundances are represented on a log scale. Species that were not detected by metagenomic profiling with Kraken2 (species of low abundance or that have no close representatives in genome collections) are shown as dots; species detected by Kraken2 are shown as triangles. The *in vivo* abundance of each species in the isolate collection was calculated by mapping metagenomic reads against isolate genomes of each species sampled from each individual (see Methods).

Individual gut microbiomes harbor extensive recent HGTs

We first detected and quantified HGT events that occurred recently in human history. We screened all genomes for large blocks of 100% identical DNA that were shared between any pairs of genomes of different species, retaining blocks larger than 500bp (hereafter named "500bp+ HGTs") or larger than 10kb ("10kb+ HGTs"). HGT is the best explanation for these observations compared to vertical inheritance, as the expected number of mutations between highly conserved and vertically inherited ribosomal genes of different species far exceeds the threshold (0 SNP) used in our heuristic to retain candidate HGTs (Supp. Fig. 2A). 10kb+ HGTs that do not contain any mutation correspond to events that occurred between 0 and ~100

248 years ago: assuming a genome size in the order of 10^6 bp and molecular clock of 1
249 SNP/genome/year, it would take 1 year for a 10kb HGT to accumulate 10-2 SNPs, which
250 corresponds to taking 100 years to experience 1 SNP and to be filtered out from our analysis.
251 Thus, these 10kb+ HGTs likely occurred over the most recent 2-3 human generations,
252 including within the sampled individuals. In this study, we focus on transfers occurring between
253 bacterial species, ignoring within-species gene recombination events. We removed putative
254 contaminants from the analysis by filtering out HGTs with low relative sequencing coverage
255 (*i.e.* compared to the coverage of the two genomes under consideration; see Methods),
256 resulting in a set of HGTs with median relative coverage of 1.13 (Supplementary Figure 2B).
257 We found that 90% (7,031/7,781) and 53% (4,096/7,781) of our genomes are involved in at
258 least one 500bp+ HGT and one 10kb+ HGT, respectively (Fig 2A, Supp. Table 5), covering a
259 diversity of taxonomic groups (Fig. 2B). HGTs included genes that are involved in a variety of
260 cellular, metabolic and informational functions (Supp. Fig. 3A), with selfish element and
261 phage/conjugative transposon functions being enriched in the set of 500bp+ HGTs and 10kb+
262 HGTs, respectively (Supp. Fig. 3B). Many of the genes carried by within-person 10kb+ HGTs
263 segregate at high frequencies in bacterial populations within each host, suggesting potential
264 fixation (Supp. Fig. 4). However, the majority of transferred genes are found at low frequency,
265 reflecting their recent acquisition in the population (Supp. Fig. 4).
266 While HGTs were detected at the level of genomes, we computed HGT counts and
267 frequencies at the level of species. To measure the HGT count of a given pair of species in a
268 given pair of individuals, we counted the number of genome pairs that share at least one HGT,
269 and divided this number by the total number of genome pairs for this species pair in those
270 individuals to derive the HGT frequency (see Methods). We used this conservative approach
271 (Smillie et al., 2011) rather than considering the absolute number of distinct blast hits between
272 two genomes to avoid inflating estimates of HGT frequency, as poor assembly or genomic

processes such as transposition might result in splitting a single large HGT into many smaller apparent HGT events.

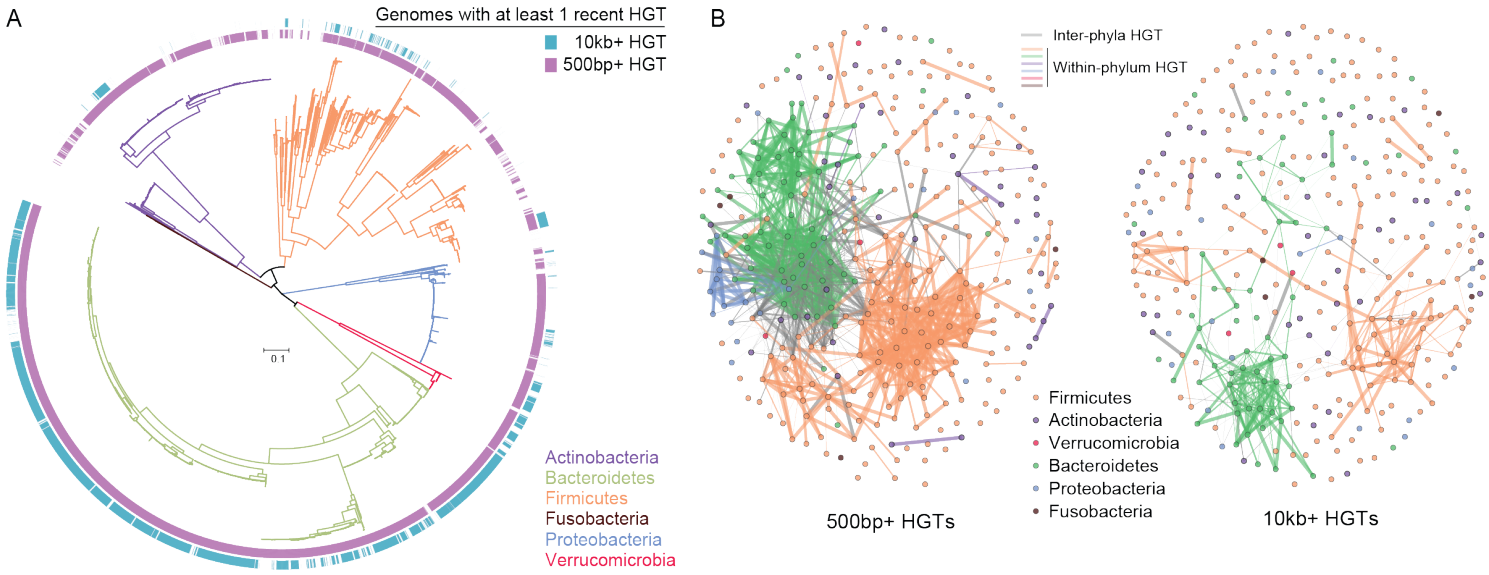

Figure 2 - Diverse human gut bacteria recently engaged in frequent HGT
A) Phylogenomic tree of the 7,781 human gut bacterial isolates that we analyzed in this study, which were sampled from 15 human populations. The tree has been reconstructed by maximum likelihood with a multiple sequence alignment of ribosomal protein coding-genes. Branches are colored by phylum and branch lengths are expressed in expected number of substitutions per site. The inner (purple) and outer (blue) rings show genomes in which at least 1 HGT larger than 500bp and 10kb was detected, respectively. **B)** Networks of within-person HGT frequency derived from 500bp+ (left) and 10kb+ (right) HGTs. Vertices represent bacterial species and are colored by phylum. Edge width is proportional to the average within-person HGT frequency between the two connected species. Colored edges show within-phylum HGTs, while grey edges represent between-phylum HGTs.

To test whether these HGTs occurred recently, we compared the frequency and count of 10kb+ HGTs observed between bacteria isolated from a single individual with that observed between the same bacteria from different individuals. We hypothesized that if transfers occur frequently within individual microbiomes, then we would observe higher levels of transfer between strains isolated from a single host. Alternatively, if transfers rarely occur, *i.e.* at rates slower than strain turnover, then we would observe similar levels of HGT between bacteria regardless of whether they were isolated from the same host. Importantly, both within-person and between-people HGTs include some background level of more ancient HGT (*e.g.* very slowly evolving genomic regions that are still 100% similar over the 10kb+ region) that do not result from direct sharing between two co-residing species in present microbiomes. Bacterial

298 species that share genes directly, however, will only be found in within-person comparisons.
299 The difference between the within-person and between-people HGTs reflects the very recent
300 HGTs that occurred within individuals, and thus can be quantified. We found that bacterial
301 species pairs sampled within individuals are more likely to share recently transferred DNA
302 than the same species pairs sampled from two different people: using a Poisson distribution,
303 we compared the observed count of HGT events for pairs of species sampled within individual
304 people to its expected value based on HGT frequencies of the same species pairs found
305 between people (Figure 3A, *p-value* < 2.2×10^{-308}). This comparison allows us to correct for
306 differences in the number of both genome and individual pairs being sampled between the
307 two categories (within-person vs. between-people) (See Methods, section "Statistical
308 analyses"). We also randomly downsampled our data to further control for the unequal
309 sampling of genomes across individual pairs (see Methods, section "Statistical analyses"),
310 which confirmed that observed HGT counts within individual people are higher than expected
311 HGT counts (100 random replicates, Welsh t-test, t=259.56, df=102.44, p-value = 3.3×10^{-146}).
312
313

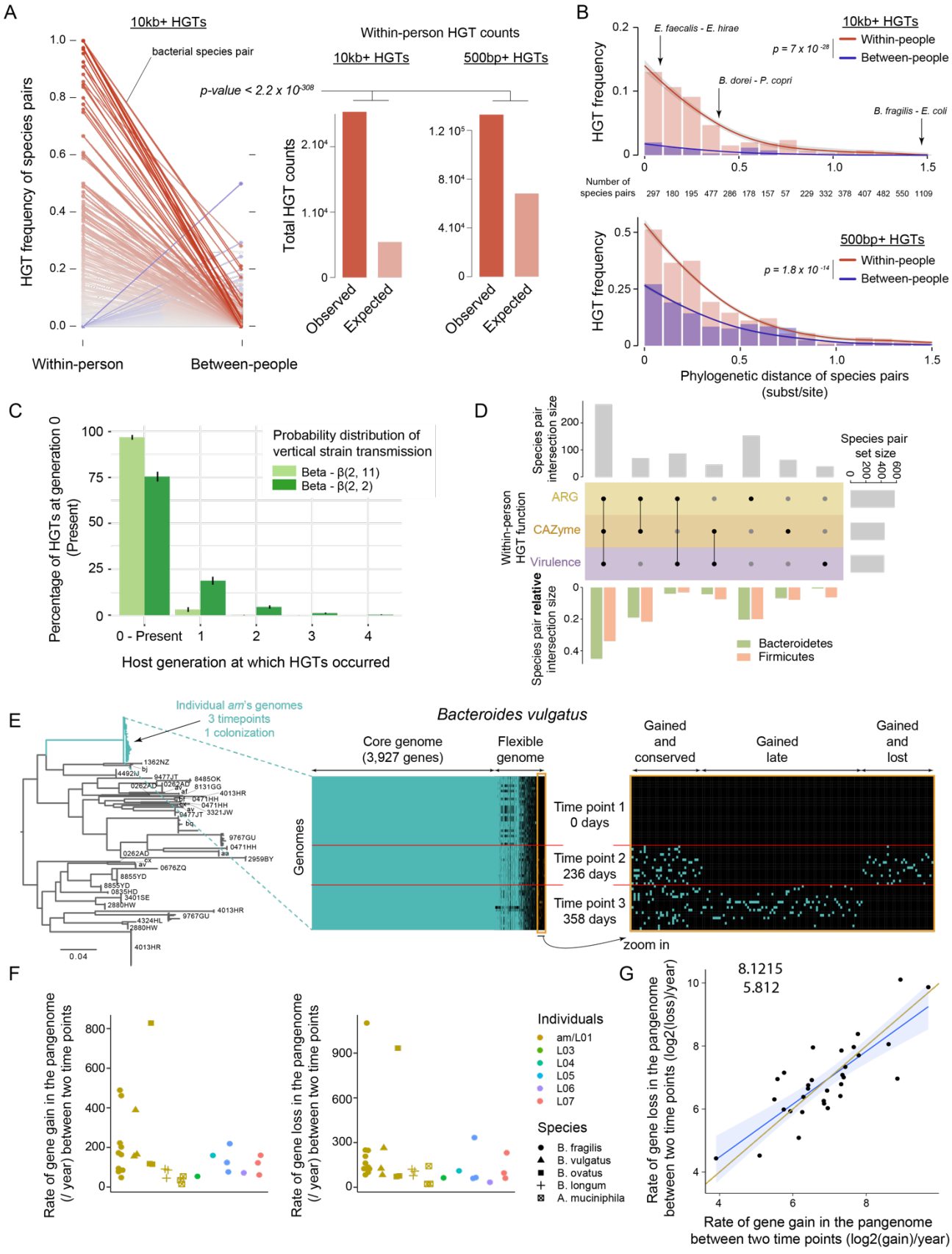

Figure 3 - HGTs accumulate rapidly within the gut microbiome of individual people

A) HGT frequencies within and between people were computed using the whole set of 7,781 genomes and were averaged across all within-person and between-people pairs, respectively. Each solid line represents a bacterial species pair sampled both within and between individuals, and connects HGT

frequencies in the two categories of pairs of individuals. The order in which lines are displayed is at random. **A null HGT frequency in either within-person or between-people categories means that no recent HGT was detected across all genome pairs.** Differences in HGT frequency are colored along a gradient from grey (no difference) to red (within-person HGT frequency is higher than between-people) or from grey to blue (between-people HGT frequency is higher than within-person), with darker colors representing greater differences. The two barplots show the observed total 10kb+ and 500bp+ HGTs for bacterial species pairs found within individuals (left bars), compared to their expected values (right bars), based on HGT frequencies of the same species pairs found between people. The p-value was calculated by comparing the observed to the expected HGT counts with a Poisson distribution. The number of species and genome pairs for each comparison and category are listed in Supp. Table 6. **B)** Association between within-person HGT frequency and phylogenetic distance of pairs of species, compared to between-people HGT frequency. The top and bottom panels show the associations for 10kb+ and 500bp+ HGTs, respectively. The HGT frequency is plotted using a LOESS regression. Phylogenetic distances were derived from the phylogenetic tree shown in Figure 2A and are expressed in number of amino acid substitutions per site in the multialignment of ribosomal proteins. Three species pairs are placed on the x-axis for illustration. HGT frequencies at distances lower than the smallest between-species distances (left part of the curves) are extrapolated. Bands represent confidence intervals calculated from the standard errors. Bars show the average HGT frequencies across all species pairs in each bin. The within-person HGT frequency is higher than between-people frequency across phylogenetic distance bins (Fisher's method to combine p-values - see Methods, section 'Statistical Analyses'). **C)** Host generation in which observed HGT0s occurred in our simulation. HGT0s correspond to HGTs detected in the microbiome in generation 0, at present time. Light green bars show results obtained when using a Beta probability distribution β(2, 11) the intergenerational transmission of bacterial species of mean ~0.16. Dark green bars show results obtained with a Beta distribution β(2, 2) of mean 0.5. **D)** 'Upset' plot showing the intersections between the sets of species pairs involved in within-person HGTs of ARG, CAZyme and virulence genes. Each row corresponds to a function set, and each column corresponds to an interaction configuration. Empty cells (light-gray) indicate that the set is not part of the intersection, and filled (black) cells show sets that participate in the intersection. Barplots on the top and right of the matrix layout show absolute counts of species pairs for each intersection and each set, respectively. Barplots in the bottom show the relative intersection sizes for Firmicutes and Bacteroides species pairs. **E)** Within-person acquisition of genes in *Bacteroides vulgatus* genomes over the course of 358 days. The core-SNP phylogenetic tree depicts the relationships between all *B. vulgatus* isolates sampled across all individuals in our dataset. The clade colored in blue shows the isolates that were longitudinally sampled from individual "*am"*. The IDs of the other individual hosts are shown next to each corresponding clade of isolates. The tree strongly suggests that *am's* isolates originate from one colonization event. Middle and right panels show gene presence/absence in all isolate genomes (rows), sorted by sampling times. The right panel is a zoom-in of the set of gene families (orange box in the middle panel) that were absent in the *B. vulgatus* pangenome at the first time point, but were later present within individual *am*. Supp. Fig. 7 shows similar figures for *B. ovatus*, *B. longum* and *A. muciniphila* in donor *am,* and all *B. fragilis* lineages in donors *am*, *L03*, *L04*, *L05*, *L06* and *L07*. Supp. Table 9 shows the numbers, IDs and sampling dates of all genomes. **F)** Within-person rates of gene gain (left) and loss (right) in the pangenome (expressed as number of events per year). In each individual, rates were calculated for all pairs of sampling timepoints. Rates were calculated using the set of gene families absent in the pangenome at the first timepoint, but present in the pangenome at the later timepoint. Rates of gene gain were heterogeneous, in contrast to SNP-accumulation rates, which are constant within individuals (the molecular clock hypothesis (Zhao et al., 2019)). We controlled for read coverage at the gene level to call for the presence/absence of a gene in a given genome (see Methods). G**)** Correlation between within-person rates of gene gain and gene loss in gut bacterial pangenomes. The blue line represents the linear regression between gene gain and loss. The yellow line shows the y=x line.

We next controlled for the effect of phylogeny on this result, as more closely-related species are more likely to engage in HGT (Smillie et al., 2011) and could be unevenly distributed between within-person and between-people categories. In our data, phylogenetic relatedness

strongly associates with 10kb+ HGT frequency (Generalized Linear Mixed Effects models (GLME), N (species pairs) = 3,667, Odds Ratio (OR) = 0.02, CI (95%) = 0.01 - 0.06; combined with a Likelihood-Ratio Test (LRT), $\chi^2 = 62.96$, *p-value* = 2.1×10^{-15}), but does not confound our result: the within-individual HGT is significantly higher than the between-people HGT across phylogenetic distance bins (Fisher's method; $\chi 2$ = 204.5 and p-value = 7x10-28 for 10kb+ HGTs; $\chi 2$ = 149.1 and p-value = 1.8x10-14 for 500bp+ HGTs) (Figure 3B). In addition, the higher levels of within-person HGTs are also observed when looking at the larger set of 500bp+ HGTs (Poisson distribution, p-value < 2.2×10^{-308}) (Figure 3A & Supp. Fig. 5J). We also investigated whether the higher within-person HGT that we observed at the aggregate level was present in individual populations as well. Performing our analyses for each of the sampled countries or ethnic groups containing more than 4 individuals separately, we found that this observation was replicated in each individual group (Supp. Fig. 5A-I). In addition, we controlled for the effect of the in vitro culturing of bacteria on the quantification of HGTs, as bacteria co-cultured on the same plate or in the presence of antibiotics could experience HGTs that do not reflect in vivo events. Comparing within-person HGTs for species pairs sampled both within and between culturing plates, or with and without antibiotics in the culturing media, we did not find any significant increase in HGT for genome pairs grown on the same plate (Poisson distribution, total observed within-plate HGT counts vs. total expected counts: p-value = 0.92; Paired Wilcoxon test, within-plate vs. between-plate HGT frequencies: p-value = 0.64) or in the presence of antibiotics (Poisson distribution, total observed with-antibiotic HGT counts vs. total expected counts: p-value = 1; Paired Wilcoxon test, with antibiotics vs. without antibiotics HGT frequencies: p-value = 0.35) (see Methods and Supp. Table 8 for all statistical comparisons).

The signal of HGT enrichment within individuals compared to its expected value suggests that a broad and diverse set of bacterial species very recently engaged in HGT, and that HGTs can rapidly accumulate in bacterial pangenomes. Strictly speaking, we cannot yet distinguish between individual transfers that occurred in the host of origin from those that may

have occurred in a host's parent or even grandparent. However, host intergenerational co-transmission of species involved in past HGTs must occur to observe ancient HGT events in today's microbiome. To be counted in our analyses, these HGTs must also not experience any mutation. We used a simulation approach to quantify the amount of HGTs in the host of origin (generation 0, sampled at present time) that would represent past HGT events originating from previous generations and that would not have experienced any mutation. Using estimates from our own data (see Methods), we fixed the number of species at each generation to 200 species, and the probability of engaging in HGT for any pair of species at each generation to 0.09. We used a previously published rate of mother-to-infant strain transmission, estimated to be about 16% (Ferretti et al., 2018), to fix the rate of intergenerational species transmission in our simulations. We also simulated data using a more extreme rate of 50% of species transmission across generations. We compared the use of a Uniform to a Beta distribution for estimating the probability of species vertical transmission from parental host to child, and compared results obtained with a mean transmission probability of either 16% (Ferretti et al., 2018) or 50%. We ran the simulation across 5 generations, performing 100 replicates, and identified the origin of HGTs observed in the last generation, at present time. We found that the number of HGTs rapidly decays across generations (Fig. 3C and Supp. Fig 6). In total, the amount of 100% similar HGTs observed at present generation that originate from ancient generations is about 3% with the 16% probability of vertical species transmission, and about 25% when considering the extreme probability of 50% species transmission. These results strongly suggest that the vast majority of HGTs being seen in within-person species comparisons occurred during the present generation, i.e. during the lifetime of each sampled individual.

We next investigated whether, within people, bacterial species engage in the transfer of gene functions that may impact bacterial metabolism or host physiology. To test this, we looked at within-person transferred genes involved in antibiotic resistance (ARG), carbohydrate degradation (CAZyme) and virulence. We chose these functions in part because

they seemed likely to reflect relevant selective pressures in the human host, and also because there exist well curated databases of annotated genes. We found hundreds of species pairs engaging in the transfer of at least one of these three functions, with the majority of species pairs exchanging multiple functions (Fig. 3D), an observation relevant to both Firmicutes and Bacteroides species pairs (Fig. 3D).

Bacterial species acquire genes at high and heterogenous frequency within individual people

Next, we hypothesized that if bacteria frequently acquire new genes within each person, their pangenomes should exhibit strong variations in gene content over time. To directly measure the rate of within-person gene acquisition, we analyzed the gene repertoires of isolate genomes that were longitudinally sampled over the course of ~6 to 18 months in two previous studies: 198 isolate genomes from five species (*Bacteroides fragilis*, *Bacteroides vulgatus*, *Bacteroides ovatus*, *Bifidobacterium longum* and *Akkermansia muciniphila*) sampled in one individual (Poyet et al., 2019), and 191 *Bacteroides fragilis* isolate genomes sampled in five additional people (Zhao et al., 2019) (Fig 3E, Supp. Fig 7 & Supp. Table 9). As strain replacement between time points can contribute to pangenome diversity, we used SNPs and phylogenetic reconstructions to restrict our quantification of the dynamics of gene repertoires to clades of closely related genomes that diversified within their host following initial colonization of the gut (see Methods, Fig 3E, Supp. Fig 7 and phylogenetic trees reconstructed in reference (Zhao et al., 2019)). We also controlled for differences in genome set sizes and genome coverage between time points (see Methods and Supp. Fig 7). To account for potential errors during the assembly process, we used the read coverage information at the individual gene level to derive the final gene presence/absence profile of a given genome (see Methods). We first quantified the rates at which new genes are gained in the pangenome of these five species between any two time points. For each species in a single individual, we

found that the rate of gene acquisition in the pangenome is heterogeneous over time (Fig. 3F), varying from tens to hundreds of gene gains per year. This suggests that gene transfers do not accumulate in a clock-like fashion, probably because one HGT event can include a single gene or a large plasmid. Our results further show that average rates of gene gain in the pangenome per year are heterogeneous across species : Bacteroides species acquire new genes in their pangenome at higher rates compared to *B. longum* and *A. muciniphila* (238 (+/- 132) genes/year for *B. vulgatus*, 353 (+/- 412) genes/year for *B. ovatus*, and 161 (+/- 124) genes/year for *B. fragilis* compared to 74 (+/- 25) genes/year for B. longum and 34 (+/- 20) genes/y for *A. muciniphila*) (Figure 3F & Supp. Table 10). These rates, which are directly estimated from longitudinal data, mirror those calculated from our cross-sectional inference in Figure 3A. Using the set of within-person HGTs, we calculated the average HGT frequency across all genome pairs involving either *B. vulgatus*, *B. ovatus*, *B. fragilis*, *B. longum* and *A. muciniphila*. We confirmed that Bacteroides species engage more frequently in HGT compared to *B. longum* and *A. muciniphila*, with average HGT frequencies equal to 2.2%, 2.3%, 0.85%, 0.04% and 0.06% for 10kb+ HGTs in *B. vulgatus, B. ovatus, B. fragilis, B. longum and A. muciniphila*, respectively, and 8.6%, 10.1%, 6.0%, 0.81% and 1.64% for 500bp+ HGTs, respectively. As expected, rates of gene gains are strongly correlated with rates of gene loss (Figure 3G; Spearman correlation, S = 1188, rho = 0.76, p-value = 2.3×10^{-6}), ultimately maintaining overall proteome sizes (Mira et al., 2001). Altogether, our results suggest that a variety of gene functions are horizontally exchanged in the gut microbiome of each host individual, and at rates that may be sufficiently high to reshape the functions of gut bacterial populations during an individual's lifetime.

HGT occurs at higher frequency in the gut microbiomes of industrialized populations

Having found that HGT occurs frequently within individuals, we next investigated the extent to which HGT rates and functions vary across human populations that have different

levels of industrialization. For this, we looked at the bacterial species pairs in our dataset that are shared by pairs of population groups along our gradient of industrialization and urbanization, which comprises four lifestyle categories (Figure 4A, Supp. Table 1 & 11). This approach allowed us to compare populations with both major and more modest differences in lifestyle. This analysis also restricts HGT comparisons to species pairs that are shared between two host populations. As a consequence, we used a more inclusive definition of HGT (the set of 500bp+ HGTs) for this analysis to make up for the loss of statistical power that resulted from comparing populations two at a time.

We found that species pairs sampled in the urban industrialized populations exchanged genes more frequently than when they occurred in the rural non-industrialized group. The number of observed HGTs found in species pairs of the urban industrialized group was compared to the expected number of events, based on the HGT frequency of the same species pairs in the rural non-industrialized populations, using a Poisson distribution, *p-value* $< 2.2 \times 10^{-308}$) (Figure 4B). These results hold whether averaging both within-person and between-people HGTs, or only within-person HGTs (Figure 4B). We also randomly downsampled the data to control for the unequal sampling of genomes across individual pairs (see Methods, section “Statistical analyses”), which confirmed that observed HGT counts in the urban industrialized group are higher than expected HGT counts (100 random replicates, Welsh t-test, t=225.04, df=154.8, p-value = 1.2×10^{-196}). To check whether these effects were driven by outlier individuals rather than population-level differences, we shuffled membership of individuals across groups – either by shuffling the lifestyles of individuals or pairs of individuals – and re-ran the analysis; the true urban industrialized cohort still had significantly higher rates of HGT than the randomly created groups (1,000 permutations each, *p-values* < 0.001, see Supp. Fig. 8). This effect also holds when restricting the analysis to each type of subsistence strategy (e.g. hunter-gatherer, pastoralist or farmer) within the rural non-industrialized cohort, which we compared individually to the urban industrialized group (Supp. Fig 9).

513 Along our lifestyle gradient (Figure 4A), we consistently found that HGTs are much
514 more frequent among the industrialized and/or urban populations across all pairwise group
515 comparisons (Figure 4C & Supp. Fig. 10). This effect was observed across different
516 comparison metrics, such as the average difference in HGT frequency, and the count and
517 proportion of species pairs with higher HGT frequency (Figure 4D-F).
518

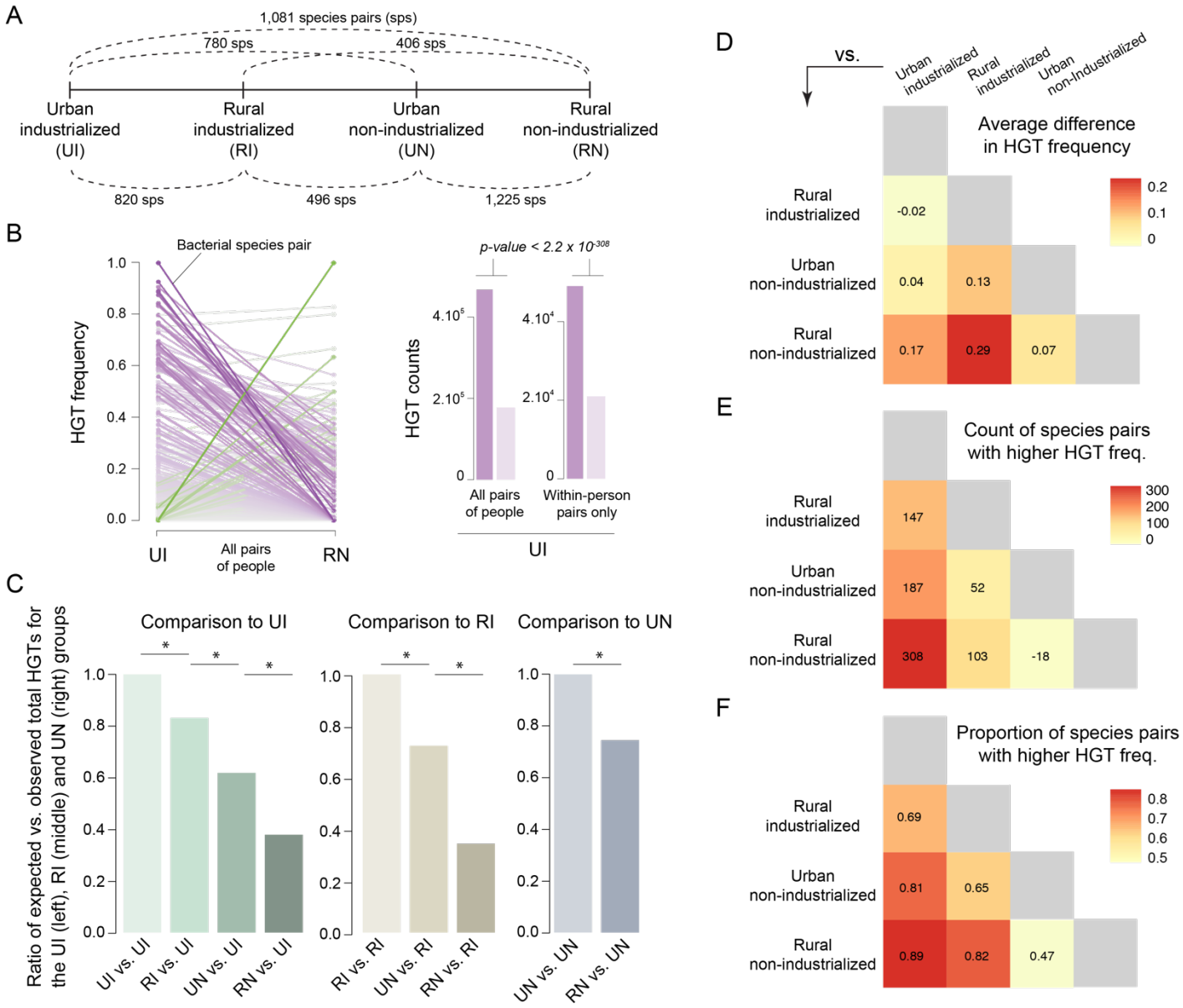

Figure 4 - Higher HGT frequency in the gut microbiomes of industrialized populations.
(A) We compared the HGT frequency of all species pairs shared between groups of populations with different lifestyles, along a gradient of industrialization and urbanization. Comparisons were performed for all group pairs, using all available species pairs for the two groups under comparison. The number of species pairs sampled for each pair of population groups is shown. For each given species pair in a group, the average HGT frequency was calculated, using all within-person and between-people pairs. See Supp. Table 1 for population groupings. **(B)** Comparison of HGT frequencies for pairs of species sampled in both the urban industrialized and rural non-industrialized groups, averaged across all pairs of people (both within-person and between-people HGTs). Each line of the paired line plot represents a species pair sampled in both groups, and a null HGT frequency for a given group means that no recent HGT was detected across all genome pairs. The order in which lines are displayed is at random. Differences in HGT frequency between the two groups are colored along a gradient from grey (no difference) to purple (HGT frequency is higher in the urban industrialized populations) or from grey to green (HGT frequency is higher in the rural non-industrialized populations), darker colors representing higher differences. The barplots show the observed total HGT of bacterial species pairs found in the urban industrialized populations (left bar), compared to their expected value (right bar) based on HGT frequencies of the same species pairs sampled in the rural non-industrialized group. The left barplot shows HGT counts when considering all pairs of people, and the one on the right shows HGT counts from within-person HGTs only. Observed and expected HGT counts were compared with a Poisson

distribution (*: p-value < $2.2x10^{-308}$) The number of species pairs and genomes for each comparison and category are listed in Supp. Table 11. **(C)** We compared HGT counts with all lifestyle pairs (panel (A)). For lifestyle pairs involving the urban industrialized group (UI, left barplot), we computed the observed total HGTs of bacterial species pairs sampled in both groups, and generated an expected total HGT value for the UI group. The ratios of observed vs. expected HGT counts for the UI group were computed for each lifestyle pair, and are shown relative to the UI group. We used the same approach for lifestyle pairs involving the rural industrialized group (RI, middle barplot) and the urban non-industrialized group (UN, right barplot). See Supp. Figure 10 for the comparison of all raw HGT counts. For each cohort pair, observed and expected HGT counts were compared with a Poisson distribution (*: p-value < $2.2x10^{-308}$). **(D)** Heatmap of the average difference in HGT frequency across all lifestyle pairs. Columns are compared against rows, with positive differences indicating higher HGT frequencies in lifestyles described in columns. **(E)** Heatmap of the difference in the absolute count of bacterial species pairs with higher HGT frequency, across all lifestyle pairs. Columns are compared against rows, with positive counts indicating a higher number of bacterial species pairs with higher HGT frequency in lifestyles described in columns. Species pairs with no HGT observed in neither category of lifestyle pairs were excluded from the counts. **(F)** Heatmap of the proportion of bacterial species pairs with higher HGT frequency, across all lifestyle pairs. Columns are compared against rows, with proportions higher than 50% indicating a higher proportion of bacterial species pairs with higher HGT frequency in lifestyles described in columns. Species pairs with no HGT observed in neither category of lifestyle pairs were excluded from the counts.

We then controlled for different microbial and ecological factors that could confound this effect of lifestyle on HGT frequencies, such as bacterial phylogeny, bacterial cell-wall architecture and, more importantly, differences of species abundances between cohorts. We hypothesized that pairs of highly abundant species in a given ecosystem would have a higher probability of gene exchange compared to pairs involving at least one low-abundance species, independent of their phylogenetic distance. This hypothesis has never been directly tested because datasets that paired in-depth genomic sampling with accurate abundance estimates did not yet exist. To test the abundance hypothesis, we generated metagenomic data for the stool samples from which we had cultured bacterial isolates, and calculated the average abundance of each bacterial species within each person by mapping metagenomic reads against the isolate genomes (see Methods and Fig. 1E). To test for the effect of cell-wall architecture, we used reference Gram staining data for each bacterial species as a proxy of cell wall architecture, in order to separate gram-positive monoderm bacteria (single cytoplasmic membrane and a thick peptidoglycan layer) from gram-negative diderm bacteria (two membranes surrounding a thin peptidoglycan layer). We used generalized linear mixed effects (GLME) models combined with likelihood-ratio tests (LRTs) on the complete dataset to measure the effect of host lifestyle on HGT frequencies while also accounting for the

aforementioned factors (see Methods). We confirmed a significant association between lifestyle and HGT frequency (N (species pairs) = 10,104; OR for the industrialized lifestyle = 1.99; CI (95%) = 1.96 - 2.03; LRT, χ^2 = 6629.4, *p-value* < 2.2×10^{-308}). We also found that species abundance is a strong determinant of HGT (N (species pairs) = 10,104; OR for lowly abundant species = 0.40; CI (95%) = 0.39 - 0.43; LRT, χ^2 = 3225.4, *p-value* < 2.2×10^{-308}) even after accounting for the effect of other factors in the GLME models (Fig. 5A). Abundant bacteria are more likely to engage in HGT with other abundant bacteria, which is consistent with the canonical mechanisms of HGT (e.g. conjugation, transformation, and transduction (Thomas and Nielsen, 2005)) which involve cell-to-cell contact or access to free DNA in the environment. In addition, we found that Gram-negative bacteria engage more frequently in HGTs than Gram-positive bacteria (N (species pairs) = 10,104; OR for Gram-negative bacteria = 9.2; CI (95%) = 6.6 - 12.8; LRT, χ^2 = 166.3, *p-value* = 4.7×10^{-38}, Figure 5B). This intriguing result motivates further investigation to understand the mechanisms driving increased rates of HGT between intestinal Gram-negative bacteria.

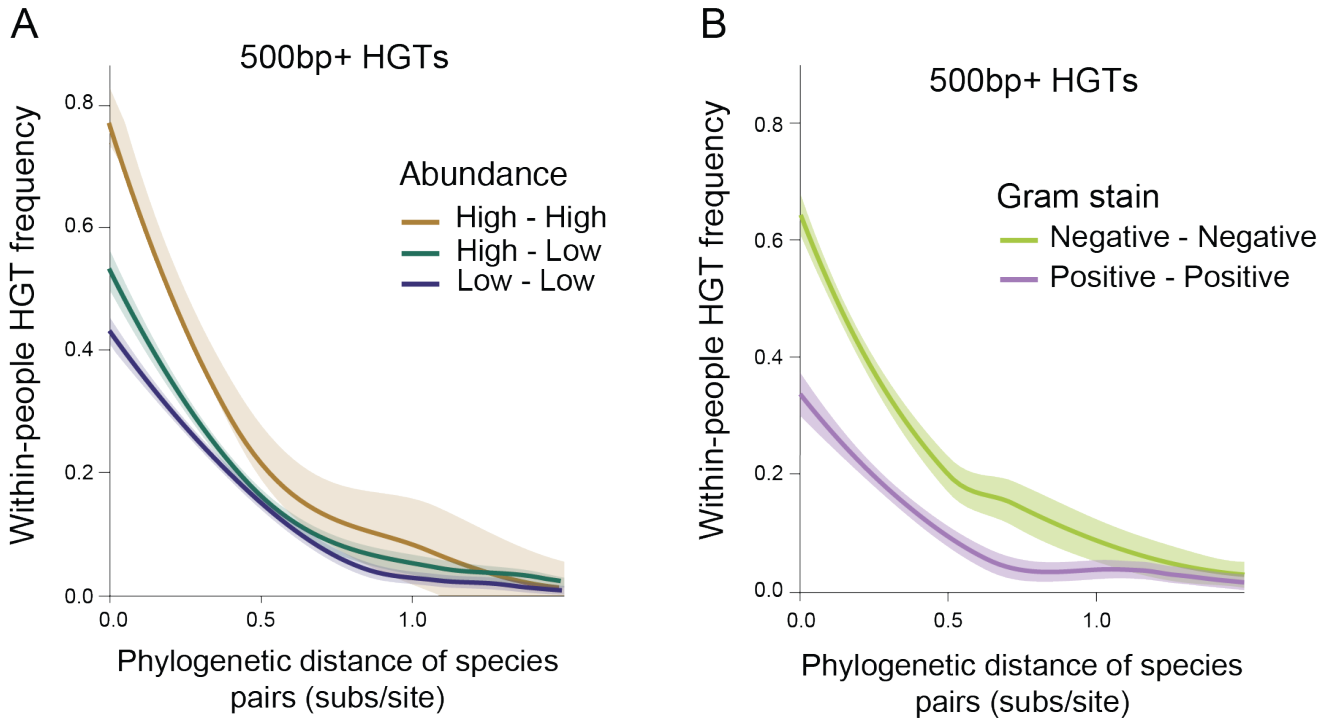

Figure 5 - Highly abundant bacteria and Gram-negative bacteria are associated with higher rates of HGT.
A) Contribution of bacterial species abundance to HGT frequency, measured with 500bp+ HGTs. The individual effect of abundance was measured with a GLME model (p-value < 2.2×10^{-308}, see Methods) and plotted using a LOESS regression. HGT frequency is plotted for different species abundance bins. Bacterial abundances were calculated for each species in each individual by mapping metagenomic reads against individual isolate genomes (see Methods). We used the distribution of bacterial abundances within individual people (Fig. 1E) to define a threshold of 1% relative abundance to separate highly and lowly abundant bacteria (see Methods). Our results hold using a 5% threshold to define high abundance (GLME, OR for lowly abundant species = 0.47; CI (95%) = 0.45 - 0.48; LRT, $\chi 2$ = 2668.1, p-value = 1.5×10^{-71}). **B)** Contribution of cell wall architecture on HGT frequency, measured with 500bp+ HGTs. The effect of cell wall architecture was measured with a GLME model and plotted with a LOESS regression. We used Gram staining as a proxy to call for monoderm or diderm bacteria (see Methods).

Functions of recent HGTs reflect host lifestyle

We reasoned that if HGT can rapidly occur in response to changes in host lifestyle, then the type of genes being transferred should reflect the selective pressures associated with different populations (Brito et al., 2016). We first compared the profile of HGTs across broadly defined functional categories using species pairs found across different lifestyles. We found significant differences in HGT functions, with the rural non-industrialized cohort having the

most different profile compared to other lifestyles (Figure 6, chi-square Goodness-of-fit test, *p-values* < 0.001).

We then focused on genes involved in key functions that likely differ across populations, such as antibiotic resistance, carbohydrate active enzyme (CAZyme), and virulence genes. We also looked at genes involved in the function of mobile genetic systems (such as phages, plasmids and transposons). We found that gut bacteria in industrialized populations tend to have higher rates of gene exchange for genes involved in plasmids and transposons (Supp. Fig. 12A, two-proportions Z-tests, corrected *p-values* < 0.001). This finding is consistent with the elevated rates of HGTs that we observed in the gut microbiomes of these individuals (Figure 4). In almost all comparisons, non-industrialized cohorts, who generally consume larger amounts of non-digestible fiber (Makki et al., 2018; Smits et al., 2017), harbored gut bacteria that exchanged CAZyme genes at higher frequencies than individuals living in industrialized and/or urban regions (Figure 6). High transfer rates of antibiotic resistance genes were also found in the gut microbiomes of both urban and rural non-industrialized populations, which correlates with the higher environmental prevalence of antibiotic resistance genes (ARGs) in low- and middle-income countries (Hendriksen et al., 2019; Pehrsson et al., 2016). This is further consistent with studies showing that antimicrobial resistance is increasing in livestock from low- and middle-income regions (Van Boeckel et al., 2019).

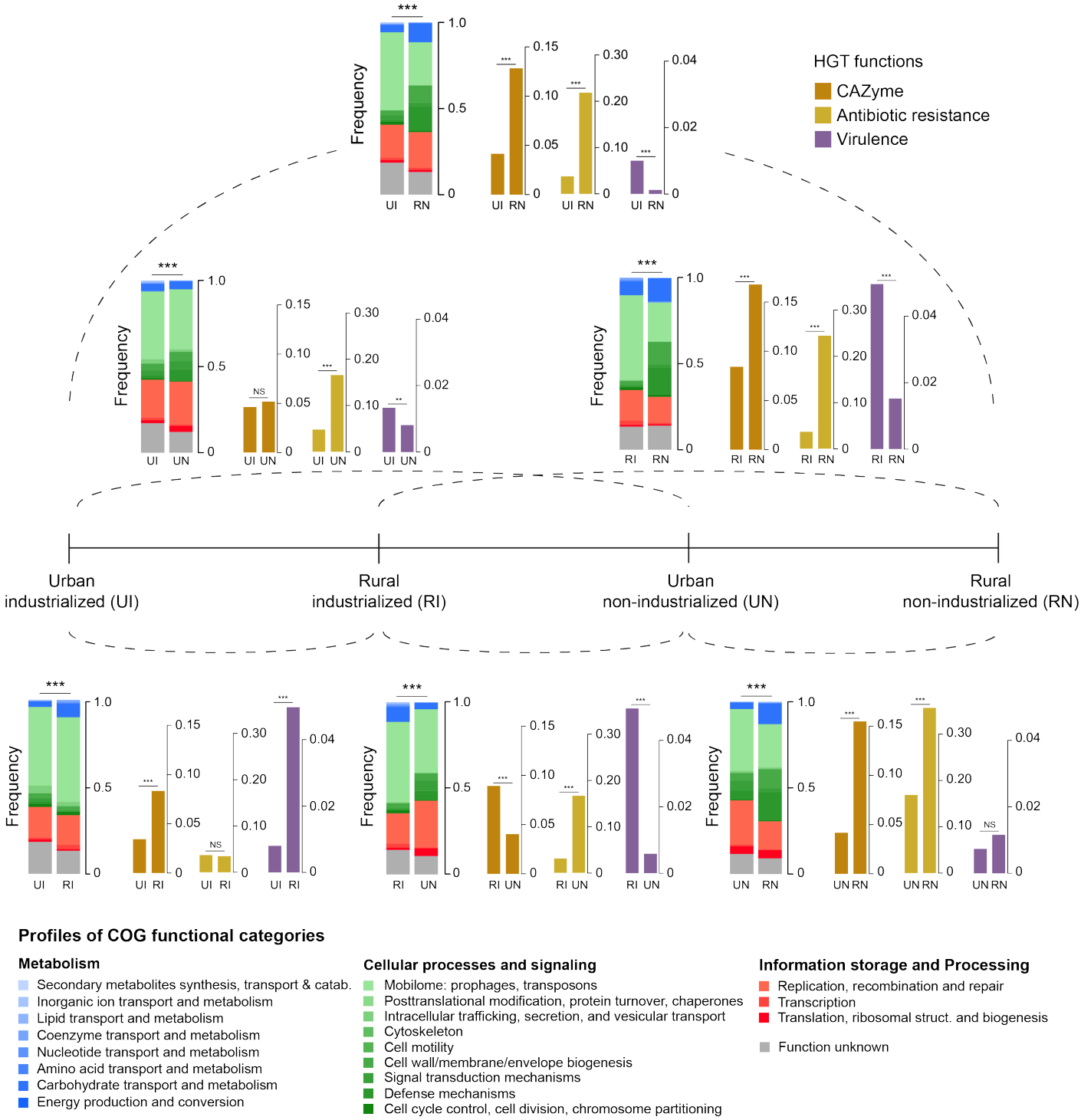

Figure 6 - Functions of recently transferred genes are associated with host lifestyle.
Genes within HGTs were annotated using a variety of reference gene function databases (see Methods) to compare functional profiles of transferred genes across our gradient of industrialization and urbanization. Profiles of COG functional categories were compared using chi-square Goodness-of-fit tests (***: p-values < 0.001) ; HGT frequencies for ARG, CAZyme, and virulence genes were compared for all lifestyle pairs using two-proportion Z-tests followed by Bonferroni correction for multiple tests (**: *p-values* < 0.01; ***: *p-values* < 0.001). For a given cohort pair of different lifestyles, functions were averaged across all pairs of individuals in each cohort. In addition, for any given cohort comparison, frequencies of HGT functions were calculated using only species pairs that were sampled in both cohorts. Because sets of co-sampled species change across pairwise cohort comparisons, the

functional HGT profile of a given cohort differs slightly from one cohort pair to another. However, these differences are non-significant (Levene's test for Homogeneity of Variance, *p-value* = 0.17, see Supp. Fig. 11), suggesting that our functional HGT profiles are not biased by differences in species sampling.

We found that the Datoga - Tanzanian pastoralists who primarily raise cattle and consume high levels of meat and dairy products from their animals - had the highest levels of ARG transfers (Supp. Fig. 12B). Like other pastoral farmers in northern Tanzania, they administer antibiotics to their herds (Caudell et al., 2017; Sieff, 1999). Our results suggest that these recent agricultural practices rapidly altered the fitness landscape in the guts of the Datoga people and have already impacted the patterns of gene transfers within their microbiomes. As the use of commercial antimicrobials is now widespread among pastoralist populations in developing countries, similar effects may occur in many populations worldwide with broader impact on the spread of antimicrobial resistance outside the clinic.

Discussion

This article reports the first large-scale genomic investigation of the effects of industrialization and urbanization on HGTs in the human gut microbiome. To accurately measure HGT frequency, we cultured and isolated gut bacteria to generate high quality genome assemblies. In addition, to identify the most recent HGTs and investigate effects of host lifestyle, we generated an extensive diversity of isolate genomes within individuals and between people, but also from diverse human populations. Taken together, our results suggest that HGT occurs frequently within the gut microbiome of each person, and is particularly rampant in industrialized populations. These results indicate that transitioning to industrialized (and urban) lifestyles resulted in an increase in gene transfers within the gut microbiome. We can speculate that increased population density and/or perturbations in the gut ecosystem associated with the consumption of processed foods and increased sanitation more frequently promote gene exchange in the gut microbiome. The overall elevated

679 frequency of HGTs in industrialized microbiomes could also indirectly result from the shifts in
680 microbiome composition that occur when transitioning to industrialized lifestyles (Vangay et
681 al., 2018), resulting in new assortments of species that frequently exchange genes. However,
682 our analyses captured an intrinsic response of bacterial genomes to industrialization, as our
683 HGT estimates were calculated for pairs of species that were present across different
684 lifestyles, in all pairs of population groups under comparison.

685 Our study has limitations. First, our sampling design did not allow us to quantify rates
686 of gene acquisition in non-industrialized individuals. Many non-industrialized populations have
687 seasonal variations in diet and social activities, which are reflected in seasonal variations in
688 microbiome compositions (Smits et al., 2017). It is likely that variations in these environmental
689 factors also impose varying selective pressures on gut bacteria. Investigating such effects on
690 the frequency and patterns of HGTs would greatly contribute to our understanding of how the
691 gut microbiome responds to lifestyle. Second, we did not examine the mechanisms by which
692 lifestyle-associated factors may drive increased HGT in the gut microbiome of industrialized
693 populations.

694 Microbiome perturbations that occur during adaptation to industrialization are
695 hypothesized to contribute both to the establishment of low-grade chronic intestinal
696 inflammation in healthy individuals and to the higher incidence of inflammation-associated
697 diseases of the industrialized world, such as inflammatory bowel disease (Sonnenburg and
698 Sonnenburg, 2019b). Inflamed environments drive changes in species composition by
699 favoring the bloom of oxygen-tolerant and pathogenic species that are particularly prone to
700 engage in HGT (Zeng et al., 2017), such as Enterobacteriaceae. In a mouse colitis model,
701 *Salmonella enterica* and *Escherichia coli* were previously shown to bloom and to engage in
702 HGT (Stecher et al., 2012). Further investigations are needed to illuminate how inflammation
703 could drive the increase in HGT in the industrialized microbiome.

704 Numerous studies have investigated how changes in diet and clinical practices, such
705 as fecal microbiota transplants (Li et al., 2016; Smillie et al., 2018), impact the composition of
706 the gut microbiome. But inferring mechanistic understanding from compositional changes is

707 difficult. Our study reveals that HGT within the gut microbiome reflects the unique selective
708 pressures of each human host. Thus, HGT patterns can be used to identify selective forces
709 acting within each individual and thereby to gain a more mechanistic understanding of these
710 events. Our results also show that whole genome sequencing data provides information on
711 individual microbiome function at a level of precision that popular approaches, such as 16S
712 amplicon and metagenomic sequencing, cannot achieve. Finally, the high rate of HGT in the
713 human gut may be a recent development in response to the industrialized lifestyle,
714 accompanied by changes in the function of genes being exchanged. We may not yet fully
715 appreciate the consequences of these potential shifts in HGT frequency and function for
716 human health.
717
718

STAR★Methods

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial Strains		
From GMbC individuals	This paper, see Supp Table 2	dbGaP Study ID: 38715 Accession: phs002235.v1.p1
From USA individuals of the Boston area	Poyet et al., 2019	NCBI BioProject PRJNA544527
Critical Commercial Assays		
DNeasy PowerSoil Kit	Qiagen	Cat No./ID: 12955-4
DNeasy UltraClean 96 Microbial Kit	Qiagen	Cat No./ID: 10196-4
Nextera® DNA Sample Preparation Kit	Illumina	Cat No./ID: FC-121-1031
Deposited Data		
Metagenomes and isolate genomes from GMbC individuals	This paper	dbGaP Study ID: 38715 Accession: phs002235.v1.p1
Metagenomes and isolate genomes from USA individuals	Poyet et al., 2019	NCBI BioProject PRJNA544527
Software and Algorithms		
cutadapt (version 1.12)	Martin, 2011	https://cutadapt.readthedocs.io/en/stable/
Trimmomatic (version 0.36)	Bolger et al., 2014	http://www.usadellab.org/cms/?page=trimmomatic
SPAdes (version .3.9.1)	Bankevich et al., 2012	https://github.com/ablab/spades
SSPACE (version 3.0)	Boetzer et al., 2011	https://github.com/nsoranzo/sspace_basic
GapFiller (version 1-10)	Nadalin et al., 2012	https://sourceforge.net/proje

		cts/gapfiller/
BBmap (version 37.68)		https://jgi.doe.gov/data-and-tools/bbtools/
Prokka (version 1.12)	Seemann, 2014	https://github.com/tseemann/prokka
CheckM (version 1.0.7)	Parks et al., 2015	https://github.com/Ecogenomics/CheckM/wiki
Mash (version 1.1.1)	Ondov et al., 2016	https://mash.readthedocs.io/en/latest/
micropan R package	Snipen and Liland, 2015	https://cran.r-project.org/web/packages/micropan/index.html
Diamond (version 0.8.22.84)	Buchfink et al., 2015	http://www.diamondsearch.org/index.php
Mafft (version 7.310)	Nakamura et al. 2018	https://mafft.cbrc.jp/alignment/software/
BMGE (version 1.12)	Criscuolo and Gribaldo, 2010	ftp://ftp.pasteur.fr/pub/gensoft/projects/BMGE/
Seaview (version 4.7)	Gouy et al. 2010	http://doua.prabi.fr/software/seaview
FastTree (version 2.1.10)	Price et al., 2010	http://www.microbesonline.org/fasttree/
Roary (version 3.11.2)	Page et al., 2015	https://github.com/sanger-pathogens/Roary
Gubbins (version 2.2.0)	Croucher et al., 2015	https://sanger-pathogens.github.io/gubbins/
SNP-sites (version 2.4.1)	Page et al., 2016	https://github.com/sanger-pathogens/snp-sites
Blastn (version 2.6.0)	Camacho et al., 2009	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
Trim Galore (version 0.5.0)		https://github.com/FelixKrueger/TrimGalore
Fastuniq (version 1.1)	Xu et al. 2012	https://sourceforge.net/projects/fastuniq/
BWA (version 0.7.13)	Li and Durbin 2009	https://github.com/lh3/bwa
Kraken2 (version 2.0.8-beta)	Wood et al. 2019	https://github.com/DerrickWood/kraken2/wiki

Bracken (version 2.5)	Lu et al. 2017	https://github.com/jenniferlu717/Bracken
vegan R package		https://cran.r-project.org/web/packages/vegan/index.html
Prodigal (version 2.6.3)	Hyatt et al., 2010	https://github.com/hyattpd/Prodigal
vsearch (version 2.3.4)	Rognes et al., 2016	https://github.com/torognes/vsearch
eggNOG-mapper	Huerta-Cepas et al., 2017	http://eggnogdb.embl.de/#/app/emapper
InterProScan (version 5.36-75.0)	Jones et al., 2014	https://www.ebi.ac.uk/interpro/search/sequence/
Hmmer3 (version 3.1b2)	Mistry et al., 2013	http://hmmer.org/
lme4 R package	Bates et al., 2015	https://cran.r-project.org/web/packages/lme4/index.html
lmtest R package		https://cran.r-project.org/web/packages/lmtest/index.html
Other		
NCBI Genome database		ftp://ftp.ncbi.nlm.nih.gov/genomes/
RiboDB database	Jauffrit et al., 2016	https://umr5558-bibiserv.univ-lyon1.fr/
Resfam database	Gibson et al., 2014	http://www.dantaslab.org/resfams
dbCAN database	Yin et al., 2012	http://bcb.unl.edu/dbCAN/

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Eric J Alm.

Materials Availability

Bacterial strains generated in this study are available upon request to the Lead Contact, Eric J Alm.

Data and Code Availability

Newly generated data (raw reads and genome assemblies for GMbC isolates and shotgun metagenomic data for GMbC individuals) will be made available online on the dbGaP server upon acceptance of this manuscript (Study ID: 38715; Accession: phs002235.v1.p1). Metagenomes and isolate genomes of USA individuals from the Boston area are available on the NCBI (BioProject PRJNA544527).

Scripts and command lines used to process the sequencing and genomic data are available at https://github.com/almlab/GMbC_HGTs

HGT data (genomic coordinates, species, individual host, length, functional annotations) are available on the Open Science Framework at https://osf.io/pr2fw/

Experimental Model and Subject Details

Study cohorts

Stool samples from 37 individuals recruited worldwide as part of the Global Microbiome Conservancy project (microbiomeconservancy.org) were obtained from Inuit individuals in Canadian Arctic, Sami and Finnish individuals in Finland, Beti and Baka individuals in Cameroon, Hadza and Datoga individuals in Tanzania, individuals from the North Plain Tribes in Montana (USA), Igbo and Yoruba individuals in Nigeria and Ashanti, Fante, Ga and Ahafo individuals in Ghana. Written informed consent was obtained from all participants. Research & ethics approvals were obtained from the MIT IRB (protocol #1612797956), but also in each sampled country prior to the start of sample collection, from the following local ethics committees: Chief Dull Knife College (Montana), protocol #FWA00020985; Comite National d'Ethique de la Recherche pour la Sante Humaine (Cameroon), protocol #2017/05/901/CE/CNERSH/SP; Nunavut Research Institute (Canada), protocol #0205217N-M; National Institute for Medical Research (Tanzania), protocol #NIMR/HQ/R.8a/Vol. IX/2657; Coordinating Ethics Committee of Helsinki and Uusimaa Hospital District (Finland), protocol #1527/2017; Cape Coast Teaching Hospital Ethical Review Committee (Ghana), protocol #CCTHERC/RS/EC/2016/3; Committee on Human Research, Publication and Ethics of the Komfo Anokye Teaching Hospital (Ghana), protocol #CHRPE/AP/398/18; National Health Research Ethics Committee of Nigeria (Nigeria), protocol #NHREC/01/01/2007-29/04/2018.

Sample collection

Participants produced a fecal sample in a sterile container that was immediately returned to researchers in the field. Raw stool was diluted 1:5 in 25% pre-reduced (anaerobic) glycerol solution containing acid-washed glass beads, and were immediately homogenized and aliquoted into cryogenic 2mL tubes. Stool samples aliquoted in cryoprotectant were immediately flash frozen in the field at -196C, using a cryoshipper tank. Samples were then shipped to MIT for processing, culturing and storage.

Isolate genome dataset

In this study, we sequenced the genome of 4,149 gut bacterial isolates that we cultured from the stool sample of 37 individuals. We completed our genome dataset with the 3,632 isolate genomes of the BIO-ML collection that we previously generated from 11 USA individuals

recruited in the Boston area (Poyet et al., 2019), providing a dataset of 48 individuals from 15 populations.

Supplementary Table 1 contains metadata information about each subject enrolled in this study; Supplementary Table 2 contains metadata for each of the 7,781 isolates, and Supplementary Table 9 provides information about the genomes that were used in the longitudinal analysis.

Method Details

DNA extraction, library construction and Illumina sequencing for shotgun metagenomics

We used the DNeasy PowerSoil Kit (Qiagen) with manufacturers' protocols to extract microbial genomic DNA from stool samples. Genomic DNA libraries were constructed from 1.2ng of cleaned DNA using the Nextera XT DNA Library Preparation kit (Illumina) according to the manufacturer's recommended protocol, with reaction volumes scaled accordingly. Prior to sequencing, libraries were pooled by collecting equal quantity of each library from batches of 94 samples. Insert sizes and concentrations of each pooled library were determined using an Agilent Bioanalyzer DNA 1000 kit (Agilent Technologies). Paired-end sequencing (2x150-bp reads) was performed using an Illumina NextSeq 500 instrument (Illumina Inc) at the Broad Institute.

Culturing and isolation of bacterial isolates

To culture and isolate the 4,149 bacterial strains generated in this study, we used stool samples collected from 37 individuals across 14 human populations. To obtain an exhaustive representation of the diversity of human gut bacteria, human fecal samples were processed anaerobically at every step in a chamber, using gas monitors controlling physico-chemical conditions (5% Hydrogen, 20% Carbon dioxide, balanced with Nitrogen). Human fecal samples were diluted in pre-reduced PBS (with 0.1 % L-cysteine hydrochloride hydrate). Diluted samples were then plated onto pre-reduced agar plates and incubated anaerobically at 37°C for 7 to 14 days. Both general (nonselective) and selective media were used to culture diverse groups of organisms. We used different culturing media, combined with antibiotic, acid, and ethanol treatments to isolate 4,149 bacterial strains. See Supplementary Table 2 for culturing media used in this study and other metadata for each isolate. After incubation, bacteria were isolated by picking individual colonies with an inoculation loop. They were streaked onto a second pre-reduced agar plate to increase colony purity. After 2 days of incubation at 37°C, one colony was re-streaked again onto third agar plate for 2 additional days of incubation. One colony from each individual streak was then inoculated in liquid media in a 96-well culture plate. After 2 days of anaerobic incubation at 37°C, the taxonomy of the isolate was identified using 16S rRNA gene Sanger sequencing (starting at the V4 region). We first amplified the full 16S rRNA gene by PCR (27f 5'-AGAGTTTGATCMTGGCTCAG-3' - 1492r 5'-GGTTACCTTGTTACGACTT-3') and then generated a ~1kb long sequence by Sanger reaction (u515 5'-GTGCCAGCMGCCGCGGTAA-3'). All isolates are stored in -80°C freezers in a pre-reduced cryoprotectant glycerol buffer.

DNA extraction, library construction and Illumina sequencing of Whole Genomes

We used the DNeasy UltraClean96 MicrobioalKit (Qiagen) and the PureLinkPro96_gDNAkit (Invitrogen) kits to extract whole genome DNA from isolate colonies, following manufacturers' protocols. Genomic DNA libraries were constructed from 1.2ng of DNA using the Nextera XT DNA Library Preparation kit (Illumina), following the manufacturer's protocol, with reaction volumes scaled accordingly. Prior to sequencing, we pooled on average 250 samples with equal quantities of DNA. Insert size and concentration of each pooled library were determined using an Agilent Bioanalyzer DNA 1000 kit (Agilent Technologies). Paired-end (2x150bp) reads sequencing was performed using an Illumina NextSeq 500 instrument (Illumina Inc) at the Broad Institute.

Draft assembly and annotation of whole genome sequences

All parameters used to generate whole genome assemblies from 2x150bp paired-end data and used to perform downstream genomic analyses are embedded in the method descriptions below.

Briefly, reads were first demultiplexed using in-house scripts. We used cutadapt v1.12 (Martin, 2011) to remove barcodes and Illumina adapters (with parameters -a CTGTCTCTTAT -A CTGTCTCTTAT). We used Trimmomatic v0.36 (Bolger et al., 2014) for the quality filtering of data (with parameters PE -phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:5:20 MINLEN:50). Reads were assembled de novo into contigs using SPAdes v.3.9.1 (Bankevich et al., 2012) (with parameter --careful). To iteratively improve genome assemblies, we used SSPACE v3.0 (Boetzer et al., 2011) and GapFiller v1-10 (Nadalin et al., 2012) to scaffold contigs and to fill sequence gaps (with default parameters). Scaffolds smaller than 1kb were removed from genome assemblies. We aligned all reads back to the assembly to compute genome coverage using BBmap v37.68 (https://jgi.doe.gov/data-and-tools/bbtools/) and the covstats option (with default parameters). The final assemblies were annotated using Prokka v1.12 (Seemann, 2014) (with default parameters).

Assessing assembly quality

We measured genome assembly statistics using CheckM v1.0.7 (Parks et al., 2015) (with parameters lineage_wf --tab_table -x fna Prokka_annotations/). All summary and quality statistics can be found in Supplementary Table 3. The median assembly completeness of all 7,781 genomes is 99.33%, the median contamination is 0.3%, the median scaffold N50 is 144kb, and the median coverage is 120X.

Clustering genomes into species

We used whole genomic information to group genomes into species clusters. We used an open-reference approach and computed all-against-all genomic distances using Mash (Ondov et al., 2016) (with default parameters). A Mash distance lower than 0.05 is equivalent to using an Average Nucleotide Identity higher than 95 %, which is a standard threshold for delineating species (Konstantinidis and Tiedje, 2005). We used an unsupervised hierarchical clustering approach to group genomes that had Mash distances <= 0.05 into taxonomic units using the bClust function from the micropan R package (Snipen and Liland, 2015). We then measured the genetic distance between the representative genome of each species cluster (defined as the genome with the highest N50) and 79,226 non-contaminated complete and draft genomes downloaded from the NCBI FTP repository (ftp://ftp.ncbi.nlm.nih.gov/genomes/) on March 27th, 2017. Clusters with a Mash distance to NCBI genomes lower than 0.05 were assigned

881 the taxonomy of the closest reference genome (we manually curated Mash results to assign
882 a taxonomy to each cluster when NCBI taxonomies were incomplete or incorrect). All genome
883 taxonomies are compiled in Supplementary Table 2 & 3.
884
885 **Phylogenetic reconstructions**
886 To reconstruct the phylogenomic tree of all 7,781 genomes, we first built a concatenated
887 alignment of 47 nearly universal and single-copy ribosomal protein families. We used Diamond
888 v0.8.22.84 (Buchfink et al., 2015) (with parameters blastx —more-sensitive -e 0.000001 —id
889 35 —query-cover 80) to BLAST all 7,781 proteomes against the RiboDB database (v1.4.1) of
890 bacterial ribosomal protein genes (Jauffrit et al., 2016). We excluded proteins bL17, bS16,
891 bS21, uL22, uS3 and uS4, as they were not sufficiently distributed across all genomes. In
892 each RiboDB gene family, we excluded genomes that contained gene duplicates. Then, we
893 aligned all protein families individually with Mafft v7.310 (Nakamura et al., 2018) (with
894 parameter —auto). We filtered out misaligned sites using BMGE v1.12 (Criscuolo and
895 Gribaldo, 2010) (with parameters -t AA -g 0.95 -m BLOSUM30) and concatenated all individual
896 alignments using Seaview v4.7 (Gouy et al., 2010). We reconstructed the phylogenomic tree
897 using FastTree v2.1.10 (with parameters -lg –gamma) (Price et al., 2010). To reconstruct
898 phylogenetic trees of *B. vulgatus*, *B. ovatus*, *B. longum* and *A. muciniphila* (Figure 3 and Supp.
899 Fig 7), we reconstructed the alignment of core protein-coding genes with Roary v3.11.2 (Page
900 et al., 2015), removed recombining regions with Gubbins v2.2.0 (Croucher et al., 2015),
901 extracted SNPs with SNP-sites v2.4.1 (Page et al., 2016) and inferred the tree with FastTree.
902
903 **Detection of HGTs**
904 We looked for gene transfers that occurred between genomes of different bacterial species.
905 We used Blast (blastn, v2.6.0) (Camacho et al., 2009) to systematically detect blocks of DNA
906 that are shared by two genomes of different species. We retained blast hits with 100%
907 similarity and that are larger than 500bp. To further increase the likelihood of looking at transfer
908 events that occurred on timescales compatible with human lifetime, we focused many of our
909 analyses on transferred blocks that are larger than 10kb. To remove putative contaminants
910 from our set of blast hits when calculating HGT frequencies, we removed HGTs that involve
911 contigs with both k-mer assembly coverage lower than 3 (as provided by SPAdes) and a
912 relative read coverage lower than 0.2 compared to the average genome coverage in at least
913 one of the two compared genomes.
914
915 **Calculating HGT counts and frequencies**
916 To avoid inflating estimations of HGT counts and frequencies, we did not consider the absolute
917 number of distinct blast hits between two genomes, as poor assembly or genomic processes,
918 such as transposition, might result in splitting a single large HGT into many smaller apparent
919 HGT events. Instead, we used a conservative approach to quantify HGTs that was previously
920 published (Smillie et al., 2011), defining the HGT count as the number of between-species
921 genome pairs that share at least one HGT (either one 500bp+ or 10kb+ HGT). To measure
922 the frequency of HGT between two species, we then divided the HGT count by the total
923 number of between-species genome pairs.
924
925 **Simulation of HGT transmission across host generations**
926 To simulate the fraction of 100% similar HGTs seen in the present generation 0 (HGT_0s) that
927 would result from HGT events that occurred in past generation, we simulated a population of
928 constant size with N species in each individual. As the median number of species in the

microbiome of our sampled individuals is 187 based on Kraken2 metagenomic profiles, we fixed N=200. At each generation, each pair of species had an H% probability to engage in HGT. In our dataset, the average proportion of species pairs engaging in HGT is 0.885. We then chose to fix H to 9%. In a previous report (Ferretti et al., 2018), the intergenerational mother-to-infant rate of strain transmission was found to be 16%. In our simulation, we compared a 16% rate of species transmission to the next host generation to a more extreme rate of 50%. So each species had a probability to transmit into the next generation drawn from one of four possible distributions:

- U(0.16), a uniform distribution with a 16% chance of vertical transmission into offspring
- U(0.5), a uniform distribution with a 50% chance of vertical transmission into offspring
- $\mathbf{B}(2,11)$, each species probability to transmit into offspring was chosen from a Beta distribution with parameters $\alpha = 2$ and $\beta = 11$
- $\mathbf{B}(2,2)$, each species probability to transmit into offspring was chosen from a Beta distribution with parameters $\alpha = 2$ and $\beta = 2$

We chose a Beta distribution to allow for some species to have an increased probability to transfer into later generations, even though the overall average was fixed at ~16% for $\mathbf{B}(2,11)$ and 50% for $\mathbf{B}(2,2)$.

We then run the simulation across 5 generations, and recorded the generation of origin of each HGT. At the last generation (generation 0, corresponding to the generation at present time), we calculated the fraction of observed HGTs in the microbiome that occurred at each generation. We run 100 simulation replicates for each possible distribution of vertical transmission of strains into offspring. Simulations were run in Python.

Calculating gene gain and loss rates in the pangenome

We used Prokka gene annotations and Roary to reconstruct the core-genome alignment and the host individual-specific gene repertoires for *B. vulgatus*, *B. ovatus*, *B. longum* and *A. muciniphila* genomes that were longitudinally sampled in individual *am* (Poyet et al., 2019), and for *B. fragilis* genomes that were longitudinally sampled in individuals L01, L03, L04, L05, L06, L07 (Zhao et al., 2019). Note that individual *am* from (Poyet et al., 2019) and L01 from (Zhao et al., 2019) are the same individual. We used the following options with Roary: -e -n -z -i 90 -cd 95. We restricted our analysis to closely-related genomes that diversified within the host of origin upon colonization of the gut: genomes from individual am differed by 111, 42, 2,328 and 338 SNPs for *B. vulgatus, B. ovatus, B. longum and A. muciniphila*, respectively. When looking at genomes from all host individuals, isolate genomes differed by 68,746, 202,262, 51,064 and 33,793 SNPs, respectively. In addition, all *B. fragilis* genomes from the same individual differed by less than 100 SNPs, while those from different individuals differed by more than 10,000 SNPs (Zhao et al., 2019). This pattern suggests that we are only including closely-related genomes, limiting the potential impact of co-colonization of different major lineages or strain replacement on the analysis of the dynamics of gene gain and loss over time. We filtered genomes that had genome completeness as measured by CheckM below 99% out of the gene tables. For each species within each individual, we excluded genomes with low average coverage. With the final set of genomes, we checked whether the genome coverage was different across time points, as this could bias estimations of gene presence/absence profiles and gene gain/loss rates in pangenomes. We found that, for each species within each individual, genome coverage was homogenous across time points (Kruskal-Wallis tests, see Supp. Fig 7, panel j). Genome assemblies used to calculate gene gain and loss rates for each species within each individual is listed in Supplementary Table 9.

Because of assembly errors, genes truly 'present' in a genome may not have been detected in the assembly by Prokka, and were later called 'absent' by Roary. We confirmed the presence and absence of genes in a given genome by mapping reads of each genome onto each gene sequence inferred by Prokka. For genes initially called 'absent' in a given genome but 'present' in other genomes, we used a representative sequence of this gene for mapping. To call for the presence of a gene in a genome, genes must be covered by a minimum of 20 reads over 90% of their length, and have a minimum relative coverage of 0.2 compared to the average genome coverage. To call 'present' a gene that was initially called 'absent', the gene was also required to have less than 30% ambiguous mappings to be called 'present', in addition to the criteria listed above.

To measure rates of gene gain and loss in the pangenome of each species between two timepoints, we identified the set of gene families that were absent in all genomes at initial sampling and present in at least 1 genome at the later time point. We repeated this procedure for all pairs of time points, and we normalized the rates of gene gain and loss to a number of events per year. We employed the same strategy for calculating rates of gene loss. When measuring differences in pangenome gene repertoires between two timepoints, we downsampled genomes at each timepoint to perform comparisons with the same number of genomes.

Analysis of metagenomic data

Metagenomic data were quality-filtered with Trim Galore v0.5.0 and Trimmomatic (same options as with isolate genomic sequencing data), dereplicated with FastUniq v1.1 (Xu et al., 2012) (default parameters) and mapped against the hg38 human reference genome with BWA v0.7.13 (Li and Durbin, 2009) (default options) to remove human reads. We used Kraken2 v2.0.8-beta (Wood et al., 2019) with default options and the Kraken2 database to call for taxonomies. We then used Bracken v2.5 (Lu et al., 2017) to refine Kraken2 taxonomic profiles at the species level, with the following options: -t 20 -k 35 -l 150. We rarefied the OTU (species) table, by downsampling reads to the minimum number of reads among all samples. We measured beta-diversities with the Bray-Curtis dissimilarity metric using the 'vegdist' function from the 'vegan' R package. Metagenomic data were not used to reconstruct metagenome-assembled genomes, as only genome assemblies generated from isolate bacteria were analyzed in this study.

Measuring the abundance of isolate genomes

We measured average species abundances of isolates within each individual host. For species with more than five isolate genomes per individual, we randomly selected 5 genomes to compute the average abundance. For species with less than five isolate per individual, we used all isolates to calculate the average abundance. We mapped metagenomic data generated from the same individual host against each isolate genome, and used the per base coverage K, the average read length L, the size of each genome S and the total number of reads T in the shotgun data to calculate the relative abundance A of each genome in the metagenome with A = (K*S/L) / T. We used a threshold of 1% to define lowly and highly abundant bacteria.

Assigning Gram stain to bacterial species

We used Gram staining data from reference microbiology databases (ATCC (http://www.lgcstandards-atcc.org/en.aspx), DSMZ (https://www.dsmz.de/) & the Microbe Directory database (https://microbe.directory)) and from publications characterizing the

phenotype of bacterial isolates to assign a consensus Gram stain to each of our bacterial species. Species with contradictory Gram staining information or with unknown taxonomy were excluded from the analysis of the correlation between HGT frequency and cell wall architecture. Our data recapitulate what we know from the literature (Garrity, 2005; Krieg et al., 2011): Bacteroidetes are Gram-; Bifidobacterium are Gram+; Firmicutes are Gram+, to the exception of Negativicutes species, which are known diderm bacteria, and of a few other species; Fusobacterium are Gram-; Akkermansia are Gram-; Proteobacteria are Gram-.

Annotating transferred genes

Functional annotation followed the basic approach described previously (Brito et al., 2016). Briefly, CDS were assigned to all 500bp+ HGTs using Prodigal v2.6.3 (Brito et al., 2016; Hyatt et al., 2010) in metagenome mode to capture gene fragments. The resultant CDS were dereplicated and clustered at 90% nucleotide identity using vsearch v2.3.4 (Rognes et al., 2016). These gene centroids were used for subsequent functional annotation steps. Both eggNOG-mapper (Huerta-Cepas et al., 2017) and InterProScan v5.36-75.0 (Jones et al., 2014) were used to assign putative function predictions to gene centroids. For additional classification of antibiotic resistance genes and carbohydrate active enzymes, hmmer3 v3.1b2 (Mistry et al., 2013) was used with the Resfam (Gibson et al., 2014) and dbCAN (Yin et al., 2012) hmm databases with a cutoff e-value of $1e^{-5}$ and score of 22. Text mining with a set of regular functional annotations that we previously used (Brito et al., 2016) was employed to determine the assignment of genes into the following categories: phage, plasmid, transposons, and antibiotic resistance.

Quantification and Statistical Analyses

When the R output of a *p-value* calculation equalled to 0, we used the smallest double-precision machine number, which is $2.2*10^{-308}$. Such p-values are shown with an asterisk in figures.

Comparing HGT frequencies and counts

Statistical analyses were performed in R. **When comparing HGTs between two categories, e.g. within-person vs. between-people or Urban industrialized vs. Rural non-industrialized, the numbers of genome and individual pairs for any pair of bacterial species that were sampled are different between the two categories. This difference in sampling could interfere with comparisons of HGT frequencies.** To correct for differences in sampling, we employed the following approach. Consider the comparison of within-person to between-people HGTs: we calculated, for each species pair, the observed within-person HGT count (corresponding to the number of within-person genome comparisons with at least 1 HGT) and the expected within-person HGT count based on the between-people HGT frequency of the same species pair. We then summed observed and expected HGT counts across all species pairs and compared the observed total HGT count within individual people to its expected value based on the amount of transfer seen between individuals, and calculated a p-value using the Poisson distribution (ppois R function). The same approach was used to compare HGT counts of the same species pairs found in different cohorts that have different lifestyles (Figure 4), for instance to compare counts of HGT in the Industrialized & Urban cohort to the Non-industrialized & Rural cohort. This approach allows us to control for differences in the number of genome, species and individual pairs sampled between two compared cohorts (within-person vs. between-people or Industrialized & Urban vs. Non-

industrialized & Rural). Note that when measuring the effect of lifestyle on HGT, observed HGTs, expected HGTs and p-values were calculated for each pair of cohorts (4 lifestyle categories, 6 cohort pairs in total). Also, as this analysis is a-symmetrical, we also performed all our tests in the other direction, *i.e.* testing whether the observed between-people HGT count is lower than the expected between-people HGT count based on within-person HGT frequencies (118,210 vs. 671,160, *p-value* < $2.2*10^{-308}$); and whether the observed *Rural & Industrialized* HGT count is *lower* than the expected count based on Urban & Industrialized HGT frequencies (42,254 vs. 66,276, *p-value* < $2.2*10^{-308}$).

We also controlled for the effect of including multiple genome pairs of the same species pairs sampled in individuals when comparing total observed and expected HGT counts. We downsampled our dataset by randomly drawing a single genome pair per species pair and per individual pair. We run this control for the comparison of within-person to between-people HGTs, and for the comparison of Urban & Industrialized (UI) to Rural & Non-industrialized (RN) HGTs. For each comparison, we run 100 random replicates. For each replicate, we calculated the total observed and expected HGT counts for the within-person category or the UI group, as described above. We then compared the distributions of observed and expected HGTs with a Welsh t-test.

Calculating the frequency of transferred genes within bacterial populations

The population frequency of a given mobile gene carried by a 10kb+ HGT detected in a given species and in a given individual was calculated by counting the number of genomes carrying this mobile gene, divided by the total number of genomes of this species in this individual. Only species with a minimum of 10 genomes in each individual were included.

Controlling for the effect of phylogeny on within-person vs. between-people HGT

To measure the difference between within-person and between-people HGT across phylogenetic distance bins (Fig 3B), we compared for each separate bin the total observed within-person HGT count across all species pairs to its expected count value based on the between-people HGT frequencies of the same species pairs in that bin, with a Poisson distribution. P-values were then combined into a single p-value with Fisher's method ('sumlog' function from the 'metap' R package).

Controlling for the effect of *in vitro* culturing

To control for the effect of *in vitro* culturing on the estimation of within-person HGTs and its comparison with between-people HGTs, we used our set of 10kb+ HGTs to test (i) whether within-person HGTs are more frequent when genome pairs are sampled from the same vs. different culturing plates and (ii) for genome pairs isolated from the same plate, whether HGTs are more frequent when genome pairs are sampled from a media containing antibiotics. These tests control for (i) HGTs that may occur during the culturing on the plate and (ii) HGTs that may be triggered by antibiotics present in the media. We compared HGTs for all bacterial species pairs from each individual host that were sampled in both categories of each of the aforementioned variables being tested. As we are comparing HGTs for genome pairs from the same species pairs sampled from the same individual, we do not need to control for differences in bacterial phylogenetic distances or abundances. We compared the total observed HGT counts for genome pairs cultured within the same plate to its total expected value based on the HGT frequency of genome pairs of the same species being cultured from different plates, using a Poisson distribution. We used the same approach for genome pairs

being grown on antibiotic-containing media vs. without antibiotics. We also correlated observed to expected HGT counts for each species pair using a Pearson correlation. Finally, we also compared within-plate to between-plate HGT frequencies and with-antibiotics vs. without-antibiotics HGT frequencies using paired Wilcoxon tests. All results are shown in Supplementary Table 8.

Permutation test to compare HGTs from populations with different lifestyles

We also used a permutation test to compare HGTs in two cohorts of different lifestyles. We defined the statistic S = (HGTcounts_observed - HGTcounts_expected) / HGTcounts_expected. For the more industrialized and urban cohort, HGTcounts_observed > HGTcounts_expected. We tested if the difference between observed and expected counts is higher with real data than under a null hypothesis. We computed the null distribution of S by rearranging the lifestyle labels of either each individual participant, or each pair of participant before calculating average HGT frequencies. The value of S obtained with real data was then compared to the null distribution to calculate the p-value. Null distributions of S for these tests are shown in Supplementary Figure 8.

Measuring the effect of bacterial phylogeny, abundance and cell-wall architecture on HGT

The effect of phylogeny on HGT frequency was measured using Generalized Linear Mixed Effects (GLME) models, assuming an intercept that is different for each pair of species. We also accounted for the effects of bacterial abundance and cell-wall architecture (Gram-negative vs. Gram-positive) in the models. We used the lme4 R package (Bates et al., 2015) (glmer function) to fit the GLME models, and used Likelihood Ratio Tests (with the lmtest package and the lrtest function) to calculate the *p-value* for phylogeny. Confident intervals for odds ratios were calculated with the Wald method.

We defined the following variables:

- phylogeny: Continuous variable. Phylogenetic distance between two species derived from the phylogenomic tree shown in Fig. 2A.
- abundance: Discrete variable. Abundance category for each pair of species in each sampled host individual, derived from the abundance category of each individual species. We used a threshold of 1% relative abundance to classify species as highly or lowly abundant in each individual.
- gram_staining: Discrete variable. Gram staining category for each pair of species derived from the individual Gram staining of each individual species.
- hgt_freqs: Continuous variable. Average within-person HGT frequency for each individual species pair. Average within-person HGT frequencies were calculated for each population separately, to account for population-level differences.
- species_pairs: Discrete variable. Names of species pairs. Because we calculated within-person HGT frequencies on a per-population basis, a given species pair can be represented multiple times in the model. We accounted for this by considering the variable species_pairs as a random effect term in the GLME models.

We fitted the following models, with HGT frequencies either derived from the dataset of 10kb+ HGTs or from the dataset of 500bp+ HGTs:

model1=glmer(hgt_freqs ~ phylogeny + abundance + gram_staining + (1|species_pairs), family="binomial")
model2=glmer(hgt_freqs ~ abundance * gram_staining + (1|species_pairs), family="binomial")

To assess whether phylogeny is significantly contributing to HGTs, we performed the following LRT:

Phylogeny: LRT_phylogeny = lrtest(model1, model2)

To measure the effect of lifestyle on HGT with the dataset of 500bp+ HGTs, while controlling for the effects of phylogeny, abundance and cell-wall architecture, we defined the discrete variable 'lifestyle' as the level of host industrialization associated with the sampled species pair (i.e. 'industrialized' or 'non-industrialized'), and run the following GLME models:

model3=glmer(hgt_freqs ~ phylogeny + abundance + gram_staining + lifestyle + (1|species_pairs), family="binomial")
model4=glmer(hgt_freqs ~ abundance + gram_staining + lifestyle + (1|species_pairs), family="binomial")
model5=glmer(hgt_freqs ~ phylogeny + gram_staining + lifestyle + (1|species_pairs), family="binomial")
model6=glmer(hgt_freqs ~ phylogeny + abundance + lifestyle + (1|species_pairs), family="binomial")
model7=glmer(hgt_freqs ~ phylogeny + abundance + gram_staining + (1|species_pairs), family="binomial")

We run the following LRTs to evaluate the contribution of each factor to HGT:
Phylogeny: LRT_phylogeny = lrtest(model3, model4)
Abundance: LRT_abundance = lrtest(model3, model5)
Cell-wall architecture: LRT_cell-wall = lrtest(model3, model6)
Lifestyle: LRT_lifestyle = lrtest(model3, model7)

Comparing functional profiles of HGTs

Profiles of COG functional categories were compared using a chi-square Goodness-of-fit test (chisq.test function). HGT frequencies of phage, plasmid, transposon, ARG, CAZyme and Virulence genes were compared between host populations of different lifestyles (Figure 6) using two-proportions Z-tests (prop.test function), and a Bonferroni correction for multiple tests (p.adjust function).

References

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2020). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol.

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

1215 Microbiome during the First Year of Life. Cell Host Microbe *17*, 852.

1216 Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M.,
1217 Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly
1218 algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

1219 Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects
1220 Models Usinglme4. Journal of Statistical Software *67*.

1221 Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding
1222 pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.

1223 Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina
1224 sequence data. Bioinformatics *30*, 2114–2120.

1225 Brewster, R., Tamburini, F.B., Asiimwe, E., Oduaran, O., Hazelhurst, S., and Bhatt, A.S.
1226 (2019). Surveying Gut Microbiome Research in Africans: Toward Improved Diversity and
1227 Representation. Trends Microbiol. *27*, 824–835.

1228 Brito, I.L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S.D., Jenkins, A.P., Naisilisili, W.,
1229 Tamminen, M., Smillie, C.S., Wortman, J.R., et al. (2016). Mobile genes in the human
1230 microbiome are structured from global to individual scales. Nature *535*, 435–439.

1231 Browne, H.P., Forster, S.C., Anonye, B.O., Kumar, N., Neville, B.A., Stares, M.D., Goulding,
1232 D., and Lawley, T.D. (2016). Culturing of “unculturable” human microbiota reveals novel taxa
1233 and extensive sporulation. Nature *533*, 543–546.

1234 Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using
1235 DIAMOND. Nat. Methods *12*, 59–60.

1236 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and
1237 Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

1238 Caudell, M.A., Quinlan, M.B., Subbiah, M., Call, D.R., Roulette, C.J., Roulette, J.W., Roth,
1239 A., Matthews, L., and Quinlan, R.J. (2017). Antimicrobial Use and Veterinary Care among
1240 Agro-Pastoralists in Northern Tanzania. PLoS One *12*, e0170328.

1241 Coyne, M.J., Zitomersky, N.L., McGuire, A.M., Earl, A.M., and Comstock, L.E. (2014).
1242 Evidence of extensive DNA transfer between bacteroidales species within the human gut.
1243 MBio *5*, e01305–e01314.

1244 Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy):
1245 a new software for selection of phylogenetic informative regions from multiple sequence
1246 alignments. BMC Evol. Biol. *10*, 210.

1247 Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill,
1248 J., and Harris, S.R. (2015). Rapid phylogenetic analysis of large samples of recombinant
1249 bacterial whole genome sequences using Gubbins. Nucleic Acids Res. *43*, e15.

1250 Didelot, X., Sarah Walker, A., Peto, T.E., Crook, D.W., and Wilson, D.J. (2016). Within-host
1251 evolution of bacterial pathogens. Nature Reviews Microbiology *14*, 150–162.

1252 Drake, J.W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. Proc.
1253 Natl. Acad. Sci. U. S. A. *88*, 7160–7164.

1254 Duchêne, S., Holt, K.E., Weill, F.-X., Le Hello, S., Hawkey, J., Edwards, D.J., Fourment, M.,
1255 and Holmes, E.C. (2016). Genome-scale rates of evolutionary change in bacteria. Microb
1256 Genom *2*, e000094.

Faith, J.J., Guruge, J.L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A.L., Clemente, J.C., Knight, R., Heath, A.C., Leibel, R.L., et al. (2013). The Long-Term Stability of the Human Gut Microbiota. Science *341*, 1237439–1237439.

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.

Forsberg, K.J., Reyes, A., Wang, B., Selleck, E.M., Sommer, M.O.A., and Dantas, G. (2012). The shared antibiotic resistome of soil bacteria and human pathogens. Science *337*, 1107–1111.

Forster, S.C., Kumar, N., Anonye, B.O., Almeida, A., Viciani, E., Stares, M.D., Dunn, M., Mkandawire, T.T., Zhu, A., Shao, Y., et al. (2019). A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat. Biotechnol. *37*, 186–192.

Garrity, G. (2005). Bergey's Manual of Systematic Bacteriology: Volume 2 : The Proteobacteria (Springer).

Garud, N.R., Good, B.H., Hallatschek, O., and Pollard, K.S. (2019). Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLoS Biol. *17*, e3000102.

Gibbons, S.M., Kearney, S.M., Smillie, C.S., and Alm, E.J. (2017). Two dynamic regimes in the human gut microbiome. PLoS Comput. Biol. *13*, e1005364.

Gibson, M.K., Forsberg, K.J., and Dantas, G. (2014). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. *9*, 207.

Goodman, A.L., Kallstrom, G., Faith, J.J., Reyes, A., Moore, A., Dantas, G., and Gordon, J.I. (2011). Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. Proc. Natl. Acad. Sci. U. S. A. *108*, 6252–6257.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. *27*, 221–224.

Hansen, M.E.B., Rubel, M.A., Bailey, A.G., Ranciaro, A., Thompson, S.R., Campbell, M.C., Beggs, W., Dave, J.R., Mokone, G.G., Mpoloka, S.W., et al. (2019). Population structure of human gut bacteria in a diverse cohort from rural Tanzania and Botswana. Genome Biology *20*.

Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature *464*, 908–912.

Hendriksen, R.S., Munk, P., Njage, P., van Bunnik, B., McNally, L., Lukjancenko, O., Röder, T., Nieuwenhuijse, D., Pedersen, S.K., Kjeldgaard, J., et al. (2019). Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat. Commun. *10*, 1124.

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.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC

1300 Bioinformatics *11*, 119.

1301 Jauffrit, F., Penel, S., Delmotte, S., Rey, C., de Vienne, D.M., Gouy, M., Charrier, J.-P.,
1302 Flandrois, J.-P., and Brochier-Armanet, C. (2016). RiboDB Database: A Comprehensive
1303 Resource for Prokaryotic Systematics. Mol. Biol. Evol. *33*, 2170–2172.

1304 Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen,
1305 J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function
1306 classification. Bioinformatics *30*, 1236–1240.

1307 Koenig, J.E., Spor, A., Scalfone, N., Fricker, A.D., Stombaugh, J., Knight, R., Angenent, L.T.,
1308 and Ley, R.E. (2011). Succession of microbial consortia in the developing infant gut
1309 microbiome. Proc. Natl. Acad. Sci. U. S. A. *108 Suppl 1*, 4578–4585.

1310 Konstantinidis, K.T., and Tiedje, J.M. (2005). Genomic insights that advance the species
1311 definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A. *102*, 2567–2572.

1312 Krieg, N.R., Ludwig, W., Whitman, W.B., Hedlund, B.P., Paster, B.J., Staley, J.T., Ward, N.,
1313 and Brown, D. (2011). Bergey's Manual of Systematic Bacteriology: Volume 4: The
1314 Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres,
1315 Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia,
1316 Chlamydiae, and Planctomycetes (Springer Science & Business Media).

1317 Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
1318 transform. Bioinformatics *25*, 1754–1760.

1319 Li, S.S., Zhu, A., Benes, V., Costea, P.I., Hercog, R., Hildebrand, F., Huerta-Cepas, J.,
1320 Nieuwdorp, M., Salojärvi, J., Voigt, A.Y., et al. (2016). Durable coexistence of donor and
1321 recipient strains after fecal microbiota transplantation. Science *352*, 586–589.

1322 Lopatkin, A.J., Meredith, H.R., Srimani, J.K., Pfeiffer, C., Durrett, R., and You, L. (2017).
1323 Persistence and reversal of plasmid-mediated antibiotic resistance. Nat. Commun. *8*, 1689.

1324 Lu, J., Breitwieser, F.P., Thielen, P., and Salzberg, S.L. (2017). Bracken: estimating species
1325 abundance in metagenomics data.

1326 Makki, K., Deehan, E.C., Walter, J., and Bäckhed, F. (2018). The Impact of Dietary Fiber on
1327 Gut Microbiota in Host Health and Disease. Cell Host Microbe *23*, 705–715.

1328 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing
1329 reads. EMBnet.journal *17*, 10.

1330 McDonald, D., Hyde, E., Debelius, J.W., Morton, J.T., Gonzalez, A., Ackermann, G.,
1331 Aksenov, A.A., Behsaz, B., Brennan, C., Chen, Y., et al. (2018). American Gut: an Open
1332 Platform for Citizen Science Microbiome Research. mSystems *3*.

1333 Mehta, R.S., Abu-Ali, G.S., Drew, D.A., Lloyd-Price, J., Subramanian, A., Lochhead, P.,
1334 Joshi, A.D., Ivey, K.L., Khalili, H., Brown, G.T., et al. (2018). Stability of the human faecal
1335 microbiome in a cohort of adult men. Nature Microbiology *3*, 347–355.

1336 Mira, A., Ochman, H., and Moran, N.A. (2001). Deletional bias and the evolution of bacterial
1337 genomes. Trends Genet. *17*, 589–596.

1338 Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in
1339 homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids
1340 Res. *41*, e121–e121.

1341 Modi, S.R., Lee, H.H., Spina, C.S., and Collins, J.J. (2013). Antibiotic treatment expands the

resistance reservoir and ecological network of the phage metagenome. Nature *499*, 219–222.

Munck, C., Sheth, R.U., Freedberg, D.E., and Wang, H.H. (2020). Recording mobile DNA in the gut microbiota using an Escherichia coli CRISPR-Cas spacer acquisition platform. Nat. Commun. *11*, 95.

Nadalin, F., Vezzi, F., and Policriti, A. (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics *13*.

Nakamura, T., Yamada, K.D., Tomii, K., and Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics *34*, 2490–2492.

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.

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

Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A., and Harris, S.R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics *2*.

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.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell *176*, 649–662.e20.

Pehrsson, E.C., Tsukayama, P., Patel, S., Mejía-Bautista, M., Sosa-Soto, G., Navarrete, K.M., Calderon, M., Cabrera, L., Hoyos-Arango, W., Bertoli, M.T., et al. (2016). Interconnected microbiomes and resistomes in low-income human habitats. Nature *533*, 212–216.

Poyet, M., Groussin, M., Gibbons, S.M., Avila-Pacheco, J., Jiang, X., Kearney, S.M., Perrotta, A.R., Berdy, B., Zhao, S., Lieberman, T.D., et al. (2019). A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. Nat. Med. *25*, 1442–1452.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE *5*, e9490.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ *4*, e2584.

Schnorr, S.L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., et al. (2014). Gut microbiome of the Hadza hunter-gatherers. Nat. Commun. *5*, 3654.

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

1384 Sieff, D.F. (1999). The effects of wealth on livestock dynamics among the Datoga
1385 pastoralists of Tanzania. Agric. Syst. *59*, 1–25.

1386 Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.A., and Alm, E.J. (2011).
1387 Ecology drives a global network of gene exchange connecting the human microbiome.
1388 Nature *480*, 241–244.

1389 Smillie, C.S., Sauk, J., Gevers, D., Friedman, J., Sung, J., Youngster, I., Hohmann, E.L.,
1390 Staley, C., Khoruts, A., Sadowsky, M.J., et al. (2018). Strain Tracking Reveals the
1391 Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota
1392 Transplantation. Cell Host Microbe *23*, 229–240.e5.

1393 Smits, S.A., Leach, J., Sonnenburg, E.D., Gonzalez, C.G., Lichtman, J.S., Reid, G., Knight,
1394 R., Manjurano, A., Changalucha, J., Elias, J.E., et al. (2017). Seasonal cycling in the gut
1395 microbiome of the Hadza hunter-gatherers of Tanzania. Science *357*, 802–806.

1396 Snipen, L., and Liland, K.H. (2015). micropan: an R-package for microbial pan-genomics.
1397 BMC Bioinformatics *16*, 79.

1398 Sonnenburg, E.D., and Sonnenburg, J.L. (2019a). The ancestral and industrialized gut
1399 microbiota and implications for human health. Nat. Rev. Microbiol. *17*, 383–390.

1400 Sonnenburg, J.L., and Sonnenburg, E.D. (2019b). Vulnerability of the industrialized
1401 microbiota. Science *366*.

1402 Stecher, B., Denzler, R., Maier, L., Bernet, F., Sanders, M.J., Pickard, D.J., Barthel, M.,
1403 Westendorf, A.M., Krogfelt, K.A., Walker, A.W., et al. (2012). Gut inflammation can boost
1404 horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. Proc.
1405 Natl. Acad. Sci. U. S. A. *109*, 1269–1274.

1406 Stewart, C.J., Ajami, N.J., O'Brien, J.L., Hutchinson, D.S., Smith, D.P., Wong, M.C., Ross,
1407 M.C., Lloyd, R.E., Doddapaneni, H., Metcalf, G.A., et al. (2018). Temporal development of
1408 the gut microbiome in early childhood from the TEDDY study. Nature *562*, 583–588.

1409 Thomas, C.M., and Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal gene
1410 transfer between bacteria. Nat. Rev. Microbiol. *3*, 711–721.

1411 Van Boeckel, T.P., Pires, J., Silvester, R., Zhao, C., Song, J., Criscuolo, N.G., Gilbert, M.,
1412 Bonhoeffer, S., and Laxminarayan, R. (2019). Global trends in antimicrobial resistance in
1413 animals in low- and middle-income countries. Science *365*.

1414 Vangay, P., Johnson, A.J., Ward, T.L., Al-Ghalith, G.A., Shields-Cutler, R.R., Hillmann, B.M.,
1415 Lucas, S.K., Beura, L.K., Thompson, E.A., Till, L.M., et al. (2018). US Immigration
1416 Westernizes the Human Gut Microbiome. Cell *175*, 962–972.e10.

1417 Vries, J. de, and de Vries, J. (1994). The Industrial Revolution and the Industrious
1418 Revolution. The Journal of Economic History *54*, 249–270.

1419 Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken
1420 2. Genome Biol. *20*, 257.

1421 Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., Chen, J., and Chen, S. (2012).
1422 FastUniq: a fast de novo duplicates removal tool for paired short reads. PLoS One *7*,
1423 e52249.

1424 Yaffe, E., and Relman, D.A. (2019). Tracking microbial evolution in the human gut using Hi-
1425 C reveals extensive horizontal gene transfer, persistence and adaptation. Nat Microbiol.

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.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. *40*, W445–W451.

Zeng, M.Y., Inohara, N., and Nuñez, G. (2017). Mechanisms of inflammation-driven bacterial dysbiosis in the gut. Mucosal Immunol. *10*, 18–26.

Zhao, S., Lieberman, T.D., Poyet, M., Kauffman, K.M., Gibbons, S.M., Groussin, M., Xavier, R.J., and Alm, E.J. (2019). Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host Microbe *25*, 656–667.e8.

Zlitni, S., Bishara, A., Moss, E.L., Tkachenko, E., Kang, J.B., Culver, R.N., Andermann, T.M., Weng, Z., Wood, C., Handy, C., et al. (2020). Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. Genome Med. *12*, 50.

Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., et al. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. *37*, 179–185.

SEDAC Population Estimation Service, 2015 (https://sedac.ciesin.columbia.edu/mapping/popest/pes-v3/)

United Nations Development Program, 2020 (http://www.hdr.undp.org/en/data)

Acknowledgements

We are grateful to our field collaborators in Montana (US), Canada, Finland, Cameroon, Tanzania, Nigeria and Ghana. We thank all human participants that agreed to provide samples to the Global Microbiome Conservancy project. This work was supported by grants from the Center for Microbiome Informatics and Therapeutics at MIT and the Rasmussen Family Foundation, and by a BroadNext10 award from the Broad Institute. Additional support was provided by a Marie Skłodowska-Curie fellowship (A.S. - H2020-MSCA-IF-2016-780860) and an ANR grant (L.S. - MICROREGAL, ANR-15-CE02-0003). We thank Tamara Mason and the team at the Walkup Sequencing platform at the Broad Institute for support on sequencing efforts.

Author Contributions

M.G., M.P. and E.J.A. designed this study. M.G., M.P., A.S., K.M., R.E.S., R.J.X. and E.J.A founded the Global Microbiome Conservancy project under which field collections occurred. M.G. and M.P. managed field administrative work and performed the collection of data and samples. A.S., M.N., J.H., S.M.G., L.S., A.F., R.S.M., A.F., V.A.J., S.L., F.E.T., C.G., L.T.T.N., D.I., B.J.S., J.M.S.L., L.R., P.P.K., T.V., S.S., A.M., M.D-R, Y.A.N, A.A-N, A.D., Y.A.A, K.A.V.,

S.O.A., M.Y.A., L.R., A.P. and C.A.O. provided local support, and contributed to the field administrative work and the collection of data and samples. M.P. performed bacteria culturing, DNA extraction from isolates and library preparation for whole genome sequencing. M.P performed DNA extraction from stool samples and library preparation for metagenomics sequencing. M.G. performed computational work and data analyses. M.G and S.M.K. performed functional annotations on transferred genes. M.G., M.P. and E.J.A. analyzed the results. M.G., M.P. and E.J.A. wrote the manuscript, which was improved by K.M. and all other authors.

Declaration of Interests

Eric Alm is a co-founder and shareholder of Finch Therapeutics, a company that specializes in microbiome-targeted therapeutics.

Supplemental Information

Supplementary Figure 1: Sampled locations, populations and gut microbiomes.
Supplementary Figure 2: Detection and relative coverage of the 5,267,297 HGTs detected across all pairs of isolate genomes from different bacterial species.
Supplementary Figure 3: Functional profile of horizontally transferred coding sequences.
Supplementary Figure 4: Frequency in bacterial populations of transferred genes being present in within-person 10kb+ HGTs.
Supplementary Figure 5: Extensive within-person gene transfers in the gut microbiome is found when looking at different resolutions of geography, human population and size of HGTs.
Supplementary Figure 6: The vast majority of 100% similar HGTs detected in today's microbiome occurred in the generation of sampled individuals.
Supplementary Figure 7: Within-person gene gains in bacterial pangenomes.
Supplementary Figure 8: The signal for elevated HGT in urban industrialized populations is robust to heterogeneities in HGT across individuals.
Supplementary Figure 9: Elevated HGT frequency in the gut microbiome of individuals living in industrialized & urban populations as compared to several non-industrialized lifestyles.
Supplementary Figure 10: Pairwise comparisons of human populations with different lifestyles show elevated HGT frequencies in industrialized populations.
Supplementary Figure 11: Homogeneity of frequencies for HGT functions across pairs of bacterial species in the gut microbiome.
Supplementary Figure 12: Transferred genes involved in plasmid, transposon and antibiotic resistance functions viewed across host lifestyles.
Supplementary Table 1: Metadata for all individuals sampled in this study.
Supplementary Table 2: Metadata for all 7,781 isolate genomes analyzed in this study.
Supplementary Table 3: Genome assembly summary statistics for all 7,781 genomes analyzed in this study.

Supplementary Table 4: Counts of sampled species and isolate genomes per individual.
Supplementary Table 5: HGT counts and frequencies for all sampled species pairs across all individual pairs, for both sets of mobile elements (10kb+ & 500bp+ elements).
Supplementary Table 6: Counts of species and genomes pairs for all cohort comparisons.
Supplementary Table 7: Control for the effect of including multiple genome pairs of a given species pairs in sampled individuals.
Supplementary Table 8: Control for the effect of culturing from the same plate or from antibiotic-containing media on within-person 10kb+ HGT counts and frequencies
Supplementary Table 9: Species and isolate genomes sampled longitudinally within people to measure rates of gene gain and loss in pangenomes over time
Supplementary Table 10: Comparison of rates of gene gains between Bacteroides species and B. longum and A. muciniphila
Supplementary Table 11: Species pairs sampled across all pairwise comparisons of population groups with diverse lifestyles.