






# Longitudinal gut microbiota tracking reveals the dynamics of horizontal gene transfer

Received: 29 September 2025

Accepted: 10 November 2025

Published online: 22 November 2025

 Check for updates

Haoran Peng <sup>1</sup>, Sergio Andreu-Sanchez<sup>1,2</sup>, Angel J. Ruiz-Moreno<sup>1,2</sup>, Asier Fernández-Pato<sup>3</sup>, Jiafei Wu <sup>1,2</sup>, Ranko Gacesa <sup>1,3</sup>, Alexandra Zhernakova<sup>1</sup>, Daoming Wang <sup>1</sup> & Jingyuan Fu <sup>1,2</sup> ✉

Horizontal gene transfer (HGT) is a major driver of bacterial evolution, but its role in shaping the human gut microbiome over time remains poorly understood. Here, we present a longitudinal metagenomic analysis of 676 fecal samples from 338 individuals in the Lifelines-DEEP study collected ~4 years apart, using a newly developed workflow to detect recent HGT events from metagenome-assembled genomes. We identified 5,644 high-confidence HGT events occurring within the past ~10,000 years across 116 gut bacterial species. We find that species pairs with an HGT relationship were significantly more likely to maintain stable co-abundance relationships over the 4-year period, suggesting that gene exchange contributes to community stability. Notably, HGT and strain replacement act together to disseminate mobile genes in the population. Furthermore, our observation that an individual's mobile gene pool remains highly personalized and stable over time indicates that host lifestyles drive specific gene transfer. For example, proton pump inhibitor usage is linked to increased transfer of multidrug transporter genes. Our findings demonstrate, at the individual gut microbiome level, that HGT is both an integral and stabilizing force in the human gut ecosystem and an important mechanism for disseminating adaptive functions, underscoring HGT potential for tracking host lifestyle.

Horizontal gene transfer (HGT) is a microbial mechanism for the acquisition of genetic material from non-parental lineages that plays a significant role in microbial adaptive evolution and interactions<sup>1,2</sup>. Although it has been proposed that bacteria rarely work together<sup>3</sup>, highly abundant co-occurring bacteria tend to exchange more genes<sup>4,5</sup>, and HGT has been shown to help competitive bacteria co-exist<sup>6</sup> and to maintain the stability of the microbial community<sup>7,8</sup>.

The human gut microbiome is among the body's densest, most interactive microbial ecosystems; it has coevolved with humans<sup>9</sup> and exhibits pervasive HGT<sup>10</sup>. Although modern humans arose approximately 300,000 years ago, widespread global immigration, the rise of large civilizations, and industrialization have mainly happened in the

past 10,000 years. Prior work indicates that genes transferred within the last ~10,000 years often confer contemporary adaptive advantages, such as defense mechanisms and antimicrobial resistance, whereas more ancient transfers tend to involve core metabolic functions<sup>4</sup>. These recent transfers (≤10,000 years) have also been linked to host lifestyle factors, including geography, diet, and medication use<sup>11–14</sup>. Non-industrialized populations, for instance, exhibit higher rates of HGT involving carbohydrate-active enzymes compared to industrialized populations<sup>5</sup>. Moreover, HGT can also contribute to bacterial functionality at both the level of individual bacteria and the community level. Evidence has suggested that the metabolic potential of the infant gut microbiome can be influenced by mobile genetic

<sup>1</sup>University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands. <sup>2</sup>University of Groningen, University Medical Center Groningen, Department of Pediatrics, Groningen, The Netherlands. <sup>3</sup>Department of Gastroenterology and Hepatology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. ✉ e-mail: [j.fu@umcg.nl](mailto:j.fu@umcg.nl)

elements from the maternal microbiome<sup>15</sup>. Accordingly, there is increasing interest in understanding the role of HGT in shaping the dynamics of microbial ecosystems from global to individual scales<sup>12</sup>, particularly for recent human history. However, few studies have investigated HGT dynamics within individual gut microbiomes over time. One recent study that assessed HGT in 26 samples longitudinally over a maximum of 1 year concluded that HGT events are both unique within individuals and change over time<sup>14</sup>. To date, no comprehensive long-term study has addressed how ongoing HGT shapes gut community structure or how these transfers relate to host phenotypic traits<sup>16</sup>.

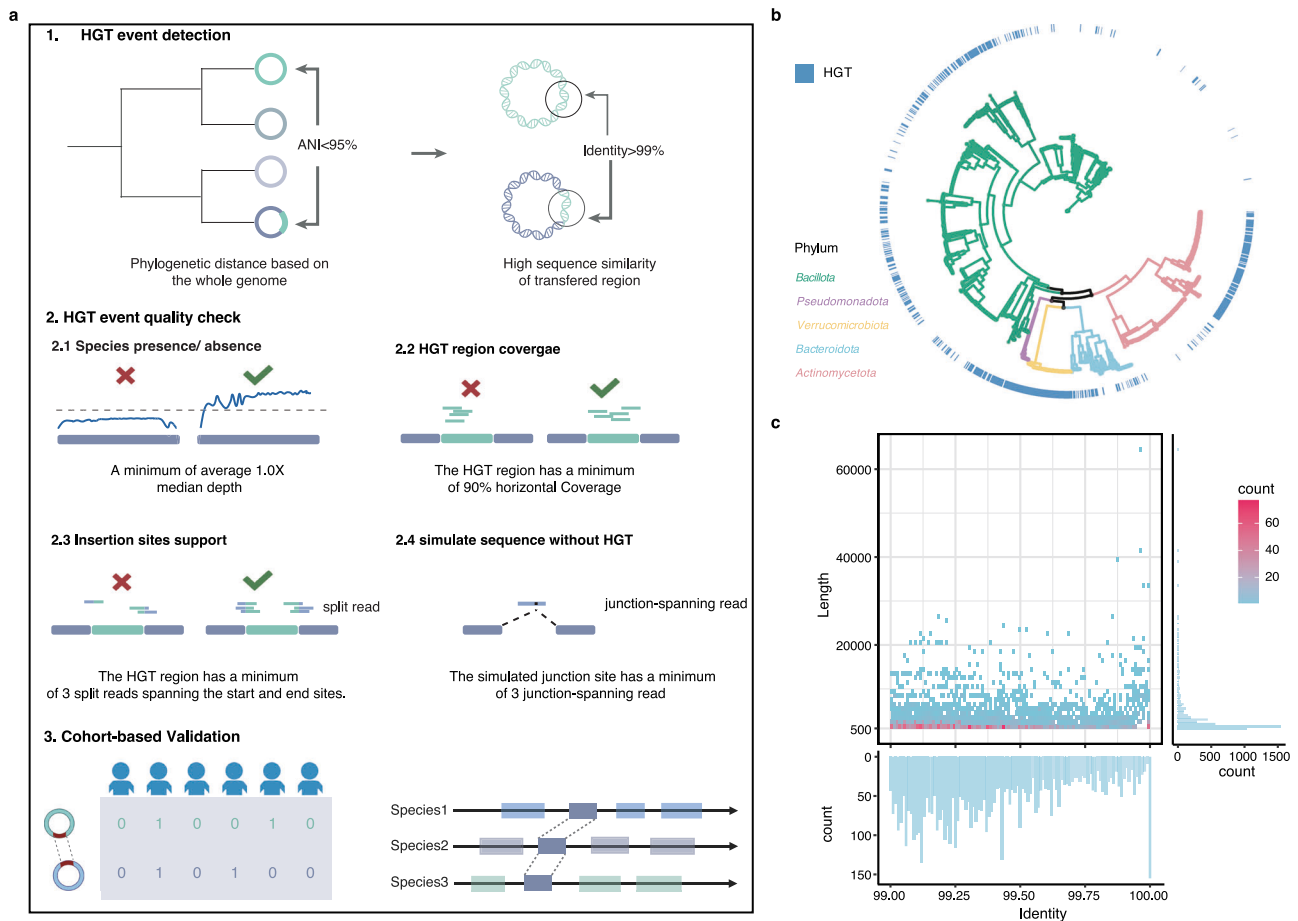
In this study, we profiled HGT events within a longitudinal microbiome cohort of 338 participants from the Lifelines-DEEP study. For these participants, fecal microbiome samples had been collected and sequenced at two time points 4-years apart<sup>17</sup>, accompanied by collection of detailed phenotype information. To capture recent HGT events (within ~10,000 years), we developed a dedicated HGT detection pipeline based on metagenome-assembled genomes (MAGs) called HGT Detection from MAGs at the Individual level (HDMI). Using this approach, we profiled thousands of HGT events and evaluated their persistence, spread, and impact on the gut microbial community. We show that HGT events contribute to stabilizing species interactions over time and can disseminate adaptive functions across hosts. We also link specific gene transfers to host factors such as medication use, providing new insight into how lifestyle pressures drive microbiome evolution.

## Results

### Cohort and the detection of HGTs in the gut microbiome

This study included 676 paired microbiome samples from 338 individuals in the Lifelines-DEEP cohort from the northern Netherlands, collected at two time points approximately 4 years apart. At baseline, participants had a mean age of 48.2 years (range = 18–80, SD = 11.7) and a mean BMI of 25.4 (range = 17.6–43.3, SD = 4.08). At follow-up, the mean age was 51.7 years (range = 22–84, SD = 11.7) and the mean BMI was 25.6 (range = 16.1–37.6, SD = 4.0). The phenotypic data assessed included anthropometric traits (e.g., age, sex, BMI, height) and the use of 14 medications (e.g., proton pump inhibitors [PPIs], oral contraceptives, beta-blockers, statins). From the baseline samples ( $n = 338$ ), we reconstructed 1473 high-quality MAGs (mean completeness =  $95.14 \pm 2.68\%$ , mean contamination =  $0.69 \pm 0.75\%$ ) (see Methods) that represent 192 distinct bacterial species (Supplementary Data 1).

To identify HGT events, we developed the HDMI pipeline, which reliably identifies recent HGT events between MAGs. We defined an HGT event as a DNA transfer detected between two distinct genomes that is characterized by the presence of a pair of highly similar DNA regions (>500 bp, >99% identity, and <95% average nucleotide identity (ANI) between genomes). Each transferred DNA region in the donor or recipient genome is referred to as an HGT segment. Thus, a single HGT event involves two corresponding segments, each located in a separate genome, which may differ in length, genomic context, or precise boundaries (Fig. 1a, see Methods).



**Fig. 1 | Recent horizontal gene transfers (HGTs) are commonly present in the gut microbiome.** **a** Overview of the HDMI workflow (Created in BioRender. Fu, J. (2025) <https://BioRender.com/k9mvzkl>). **b** Phylogenomic tree of the 1473 metagenome assembled genomes (MAGs) derived from the Lifelines-DEEP Baseline Cohort ( $n = 338$ ). Branches are colored according to the phylum classification. The

outer ring indicates whether the genome is involved in at least one HGT. **c** Distribution of HGT events. Two-dimensional histogram depicting distributions of the length (y-axis, right marginal histogram) against the identity (x-axis, bottom marginal histogram) of HGT events ( $n = 5644$ ).

In brief, HDMI works as follows. First, HDMI searches for genomic regions of at least 500 bp that show >99% identity between any pair of MAGs with ANI < 95%. Second, HDMI performs several quality checks. It first checks MAG contamination and identifies split reads that span the junction between the HGT regions and their flanking regions. We also require at least 90% coverage of the breadth and a median of 1x genome depth for HGT segments. Third, HDMI excludes false positive HGTs due to highly conserved genes with >99% identity. Lastly, we performed cohort-based validation by inferring, for each sample, the presence of each HGT event from read evidence, calculating the number of reads mapping to the split sites of sequences with HGT insertion and the junction site with no HGT insertion (Fig. 1a). Compared to other HGT detection tools, such as the recently developed reference-based WAAFL<sup>14</sup>, which does not distinguish recent from ancient transfers, HDMI offers several advantages: (1) it focuses on detecting recent HGT events (occurring within the past 0–10,000 years), (2) it enables reference-free HGT detection, and (3) it has high sensitivity for detecting HGT between bacteria with various phylogenetic distances, especially for intra-genus events (Fig. S1, Supplementary Note 1). Detailed descriptions of all methods, including metagenomic assembly, genome binning, HGT detection, and benchmarking, are provided in the Methods and Supplementary Note 1.

Using our HGT detection pipeline, we identified 5644 high-confidence, recent HGT events (occurring within the past -10,000 years, Supplementary Data 2). Our data show that HGT is common in the gut microbiome. Out of 1473 MAGs, 901 MAGs (61.2%), representing 116 unique species, were involved in at least one HGT event, with a total of 7,581 HGT segments identified (Fig. 1b, Supplementary Data 3). Segment sizes ranged from 0.5 to 64 kb (median -1.65 kb), and the median nucleotide identity was 99.36%, consistent with transfers within the past -10,000 years (Fig. 1c). By reconciling phylogenetic trees of species constructed using RANGER-DTL<sup>18</sup> based on whole genomes or HGT segments, we identified the donor and recipient genomes. Notably, we also identified some exceptionally long segments, including a -64 kb element flanked by mobile elements at both ends that encodes multiple transporters and multidrug efflux systems (Supplementary Note 2, Fig. S2).

### Gene functions enriched in HGT

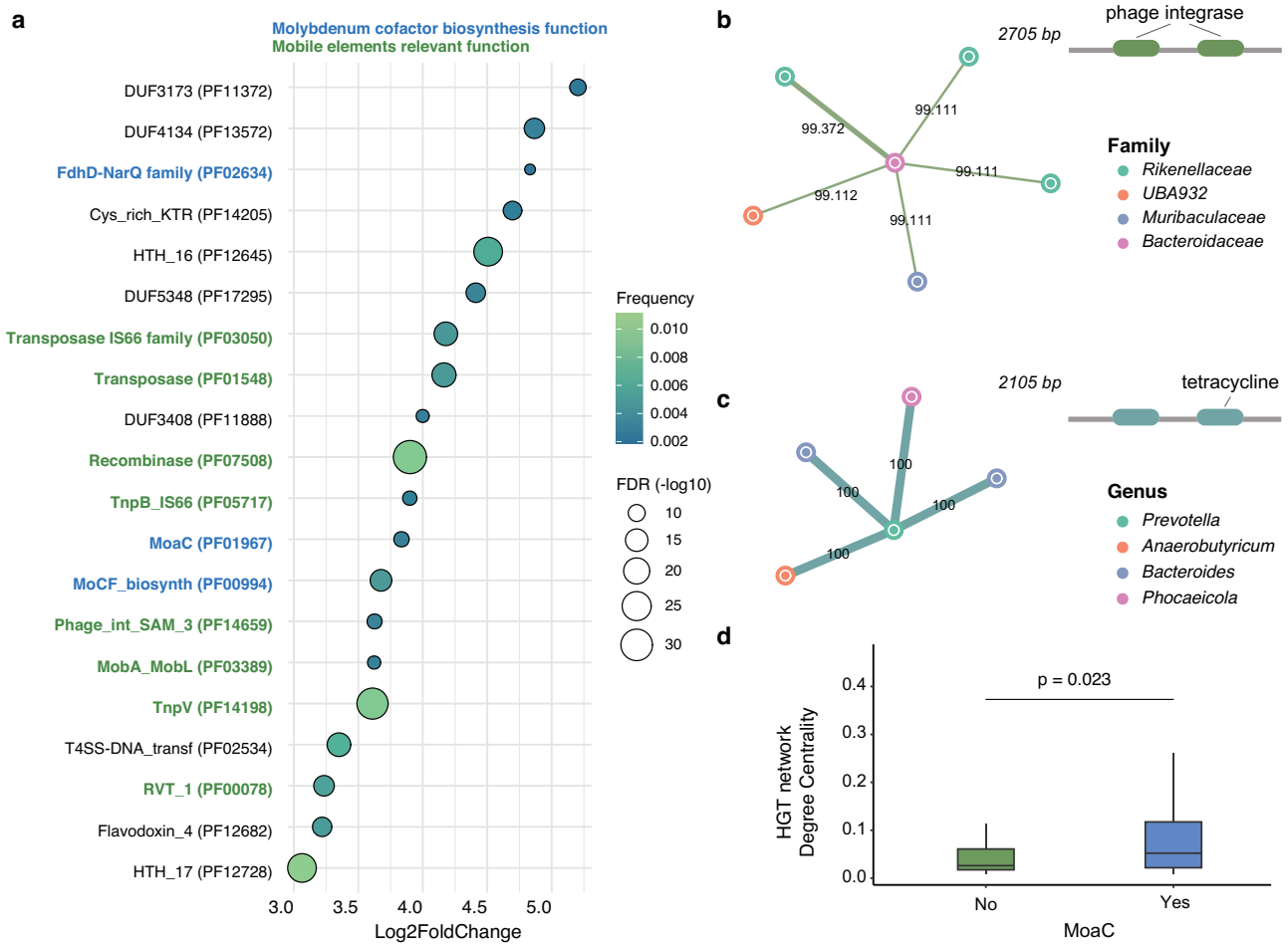
Next, we hypothesized that two sets of genes would likely be enriched in recent HGT events in the human gut: genes facilitating gene transfer (e.g., mobilome components such as prophages and transposons) and those involved in defense mechanisms, including antimicrobial resistance. To investigate this, we performed a functional enrichment analysis comparing the genes present in HGT segments to those in the species pangenome (see Methods, Supplementary Data 4). In agreement with previous studies<sup>19</sup>, functions associated with gene transfer were significantly enriched for transposase, recombinase, MobA/MobL, and phage integrase (Fig. 2a). These self-transmissible genes, which can move among various species and occasionally carry host genes, tend to have conserved transfer elements despite the variability in the host genes they transport. For example, we observed that a 5-kb HGT segment carrying a recombinase was shared among seven genera (Fig. S3). Additionally, we found two phage integrase homologs that were identified in a 2.7-kb HGT segment using *geNomad*<sup>20</sup> to be shared among four families (Figs. 2b, S3, Supplementary Note 3). Moreover, we estimated that 154 HGT events had occurred within the past 100 years, showing 100% nucleotide identity. This included four recent transfer events linked to the spread of tetracycline resistance genes that showed 100% identity across four different genera: *Prevotella*, *Anaerobutyricum*, *Bacteroides*, and *Phocaeicola* (Figs. 2c, S4, Supplementary Note 4). This observation aligns with the widespread use of tetracyclines starting in the 1950s<sup>21</sup>.

Furthermore, we observed significant enrichment for functions involved in the biosynthesis of the molybdenum cofactor (MoCo), including MoCF biosynthesis, MoaC, and FdhD-NarQ (Fig. 2a). Proteins in the FdhD-NarQ family, such as FdhD, facilitate the insertion of MoCo into target enzymes, while MoaC and MoCF\_biosynth catalyze the early and later steps of MoCo biosynthesis, respectively. Together, these components are critical for the MoCo pathway, and MoCo-dependent enzymes are essential for anaerobic respiration in microorganisms<sup>22,23</sup>. Additionally, motile bacteria (Likelihood-ratio test (LRT),  $p$ -value = 0.00228) and bacteria harboring MoaC (LRT,  $p$ -value = 0.0233) displayed significantly higher degree centralities within the HGT network, suggesting that these species act as hubs capable of engaging in gene transfer with a diverse array of other species (Figs. 2d, S5, Supplementary Note 5). Previous studies have shown that overexpression of MoCo synthesis genes promotes the growth of *Escherichia coli* in inflamed intestines<sup>23</sup> and that disruption of these genes leads to poor *E. coli* growth<sup>24</sup>. This suggests that MoCo plays a critical role in enabling microorganisms to thrive in the hypoxic environment of the intestine. Our findings extend this observation by demonstrating that HGT can facilitate the widespread prevalence of MoCo synthesis genes among intestinal microorganisms, not just in *E. coli*. However, further research is needed to fully elucidate the role of HGT in this process.

### Species involved in HGT tend to have stronger co-abundance correlation

HGT can disseminate genes, e.g., antibiotic resistance genes (ARGs), virulence factors (VFs), and antimicrobial-peptide resistance genes (AMPRs), that align species' responses under certain environmental pressures, leading to parallel abundance dynamics. We therefore hypothesized that species with HGT may show a stable co-abundance relationship. To test this, we calculated pairwise correlations among 192 species based on their abundance across individuals at each time point using Fastpar<sup>25</sup>. Based on 1,000 permutations, 6,994 species pairs (38.1% of all species pairs) showed a significant co-abundance relationship (permutation-based  $p$ -value < 0.01 and  $|\rho| > 0.2$ ) for at least one time point. Out of these, 1,968 species pairs exhibited stable positive co-abundance, while 1,702 pairs had negative correlations at both time points (Fig. 3a, see Methods). Notably, 73.7% (73/99) of species pairs with an HGT relationship at baseline (Fig. 3b) exhibited stable positive co-abundance relationships, showing a 2.24-fold enrichment ( $p$ -value =  $4.5 \times 10^{-21}$ ) compared to only 32.9% (1979/6001) for species pairs without an HGT relationship (Fig. 3b).

We then constructed a co-abundance network for species pairs with a stable and consistent co-abundance relationship at both time points (1968 pairs with a positive relationship and 1702 with a negative relationship), which could then be divided into seven communities using the Leiden clustering<sup>26</sup> (Fig. 3c, see Methods). The three largest sub-communities were Communities A ( $n = 47$ ), B ( $n = 44$ ), and C ( $n = 92$ ). Within each sub-community, nearly all species were positively correlated (Fig. 3d), whereas species across different communities were predominantly negatively correlated. Interestingly, the occurrence of HGT events was not evenly distributed among these communities. In Community C, about 10.5% (135/1280) of the co-abundance relationships also exhibited an HGT relationship, compared to 1.4% (5/353) in Community A (Fisher's exact test,  $p$ -value =  $5.88 \times 10^{-10}$ ) and 3.1% (11/354) in Community B (Fisher's exact test,  $p$ -value =  $2.65 \times 10^{-6}$ ) (Fig. 3d). Furthermore, Community C was significantly enriched in VFs (LRT,  $p$ -value =  $2.0 \times 10^{-5}$ ), ARGs (LRT,  $p$ -value =  $3.6 \times 10^{-2}$ ), and AMPRs (LRT,  $p$ -value =  $7.84 \times 10^{-5}$ ) (Fig. S6, Supplementary Data 5) compared to Communities A and B, while no significant differences in these gene categories were observed between Communities A and B (LRT,  $p$ -value > 0.1). In addition, temporal changes in the composition of Community C, as measured by



**Fig. 2 | Characterization of recently transferred functions in the gut microbiome.** **a** Functions enriched in transferred genes. Plot shows fold enrichments for Pfam domains among genes transferred within species pairs relative to genes in all species. Fisher's exact test was used to determine the significance of the annotation enrichment. Dot size and color gradient indicate  $-\log$  (base 10) of the FDR values. The two-tailed  $p$ -values were corrected using the Benjamini-Hochberg FDR method. Only  $-\log_{10}(\text{FDR}) \geq 8$  and  $\text{Log}_2\text{FoldChange} \geq 3$  are shown. Domains highlighted in blue are functionally related to molybdenum cofactor biosynthesis. Domains highlighted in green are functionally related to mobile elements. Gene transfer network of the same segment shared between multiple genomes encoding

(b) phage integrase and (c) antibiotic resistance genes. Each node represents a genome, with the corresponding taxonomy classification indicated by color. Each edge represents an HGT event observed between the genomes. The identity of the HGT event is labeled at the edge. The gene encoded in the segments is shown beside the network. **d** Association between species' degree centrality in the HGT network and the presence of the MoaC gene ( $n = 126$ ). Box plots show the median (center line), the 25th and 75th percentiles (box), and whiskers extending to points within  $1.5 \times \text{IQR}$ . Statistical significance was evaluated using a linear mixed-effects model to account for random effects.

Bray-Curtis distances, were significantly lower than those observed for Communities A and B ( $p$ -value =  $1.13 \times 10^{-20}$ , Fig. 3e, see Methods).

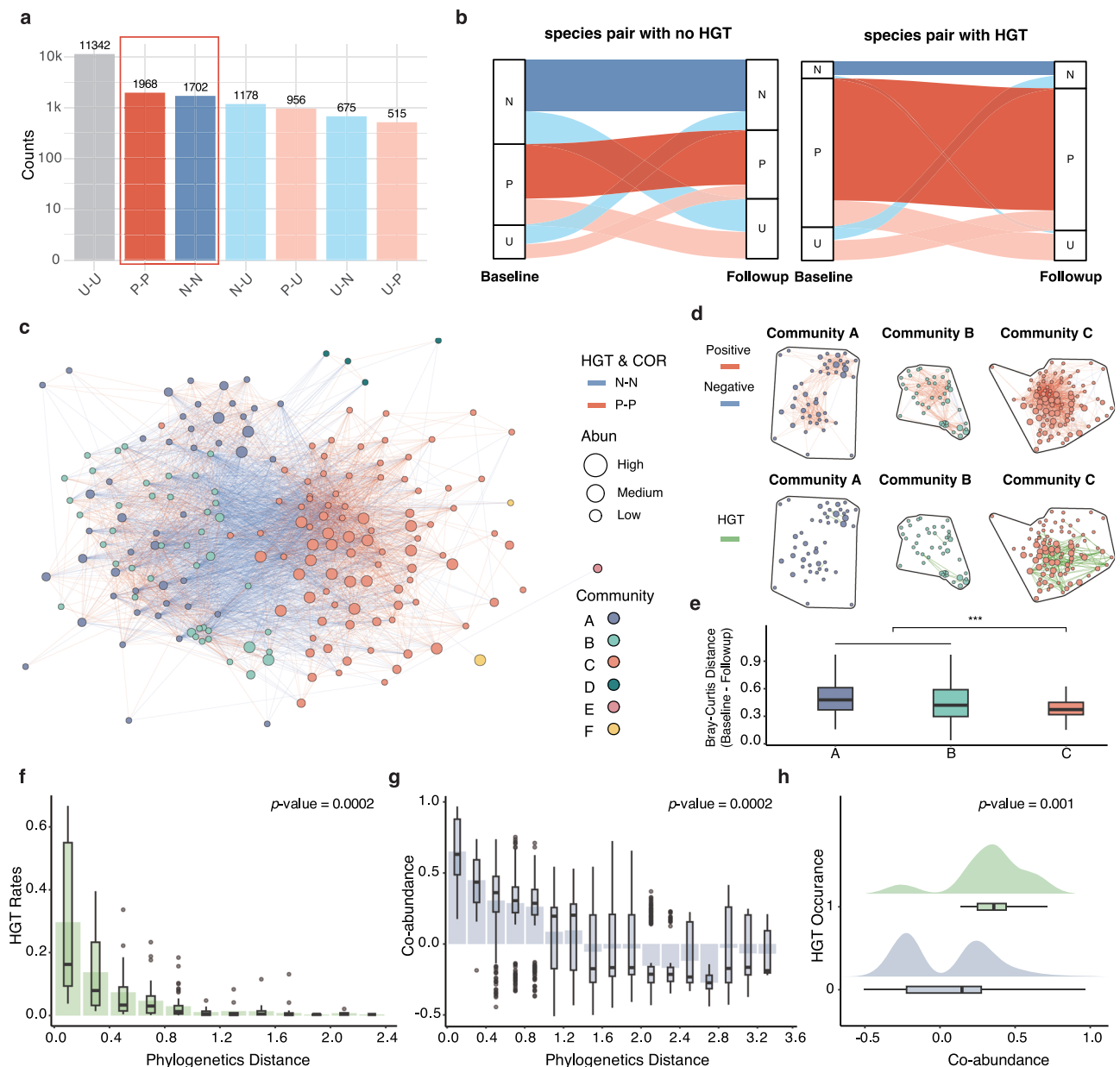
Consistent with previous reports that HGT is more likely to occur between phylogenetically close species<sup>5,14</sup>, our data show that phylogenetically close bacteria exhibited a higher HGT rate (9999 permutation partial Mantel tests,  $p$ -value = 0.0002, Fig. 3f, Supplementary Data 6). We also observed a negative relationship between phylogenetic distance and co-abundance strength (partial Mantel tests,  $p$ -value = 0.002, Fig. 3g). Importantly, even after controlling for phylogenetic distance, species pairs with an HGT relationship still demonstrated stronger positive co-abundance relationships (9999 permutation partial Mantel tests,  $p$ -value = 0.001, Fig. 3h). Together, these findings suggest that species pairs involved in HGT tend to accumulate more resistance genes and show stronger correlation in their abundances.

### Propagation of HGT segments in the gut community is mediated by both HGT and strain replacement

As detected HGTs represent selection-filtered variants, we expected them to be maintained and detected in an increasing proportion of

hosts over time; accordingly, we hypothesized that cross-individual prevalence would increase from baseline to follow-up. To quantify the spread of HGT segments over time, we compared the prevalence of each HGT segment across individuals at baseline versus follow-up. Out of 338 individuals, an HGT segment was present in 25.6% of individuals on average at baseline, and this prevalence increased to 27.1% at follow-up ( $p$ -value =  $3.57 \times 10^{-200}$ , Fig. 4a). Overall, 4696 of the 7581 segments (62.0%) were detected in a greater fraction of individuals after 4 years compared to baseline, while 2530 segments (33.4%) became less prevalent (Fig. 4b). This indicates a net propagation of transferable segments within the population over the study period. The top increased segment was a 2.7-kb region annotated as bacteriophage in *Agathobacter rectalis*, with a prevalence that increased from 7.0% to 30.2% after 4 years (Fig. 4c). Another notable increase in *Anaerobutyricum hallii* was linked to a 5.2-kb segment encoding  $\beta$ -lactamase, with its prevalence rising from 40.5% to 52.1%. As  $\beta$ -lactamase is involved in bacterial resistance to  $\beta$ -lactam antibiotics, this observation suggests a potential risk of widespread dissemination of ARGs within the population.

Note that the observed increase in HGT segment prevalence may be explained by two scenarios. In the first, a segment is directly

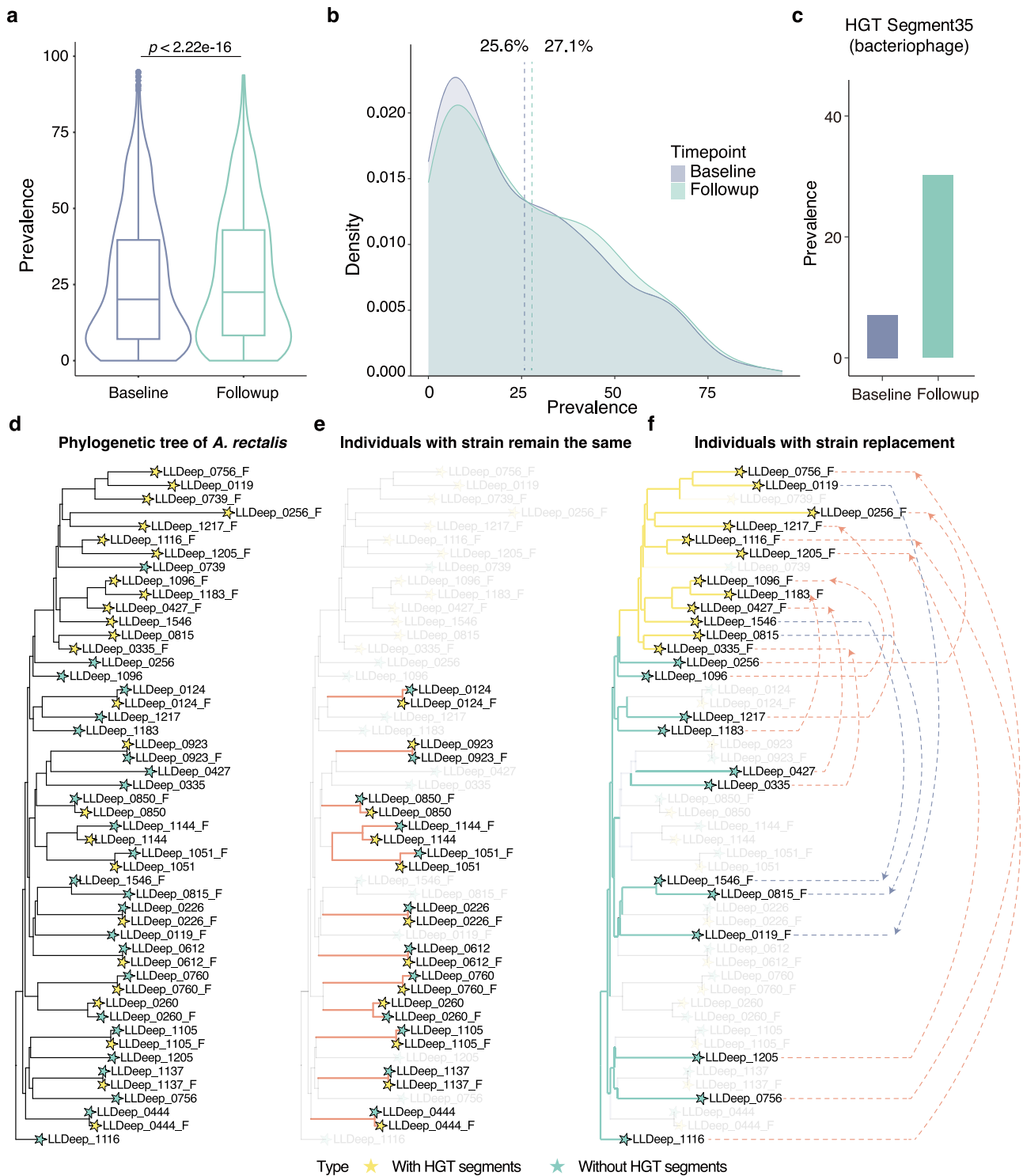


**Fig. 3 | Species involved in HGT tend to have stronger co-abundance correlation.** **a** Distribution of various types of correlation among species pairs. Each two-letter code represents the correlation observed at baseline and follow-up (N for negative, P for positive, U for uncorrelated). Correlations that are stable over time (N-N and P-P) are highlighted in the red box. **b** Plots showing the change in correlation in species pairs at baseline and follow-up. Temporal changes in species pairs' correlation with no HGT are shown at left and in those with HGT at right. **c** Microbial Co-Abundance Network. In this network, edge lengths reflect the strength of species correlations, with shorter edges indicating stronger associations. Nodes are color-coded based on Leiden clustering, and edges are colored according to the type of relationship (blue for negative correlations, red for positive correlations, green for HGT transfers). **d** The three largest sub-communities identified are Community A ( $n = 47$ ), Community B ( $n = 44$ ), and Community C ( $n = 92$ ). Upper panel shows the co-abundance relationship, where edge colors indicate positive correlations in red and negative correlations in blue. The lower panel shows the HGT relationships, with green edges representing HGT events. **e** Temporal Stability of Sub-Communities. The y-axis represents the Bray-Curtis

distance of species abundance within each sub-community between baseline and follow-up ( $n = 338$ ). Box plots show the median (center line), the 25th and 75th percentiles (box), and whiskers extending to points within  $1.5 \times \text{IQR}$ . The LMM with phylogenetic distance matrices was fitted to examine the association. **f** Association between HGT rates and phylogenetic distance of species pairs ( $n = 237$ ). Bars show the average HGT rates across all species pairs in each bin. The association was evaluated using a linear regression model. **g** Association between co-abundance and phylogenetic distance of species pairs ( $n = 3896$ ). Bars show the average correlation value across all species pairs in each bin. The association was evaluated using a linear regression model (see Methods). **h** Association between HGT occurrence and co-abundance of species pairs ( $n = 237$ ). The half violin presents the co-abundance distribution when HGT occurred (blue) and did not occur (green) between species pairs. The association between co-abundance and HGT occurrence (0/1) while adjusting for phylogenetic distance was assessed using a partial Mantel test with 9999 permutations. Boxplots show median (center line), IQR (box), and  $1.5 \times \text{IQR}$  whiskers.  $p$ -values were FDR-corrected.

transferred from one species to another. In the second, a strain lacking the HGT segment is replaced by one that carries it. We hypothesized that phylogenetic analysis can distinguish between these scenarios. If the strain acquired the segments during the 4 years, the samples at

both time points should cluster together because their genetic background is the same. If a baseline strain without the segment is replaced by another strain with the segment, samples from the two timepoints would cluster in different clades due to their diverse genetic



**Fig. 4 | Propagation of HGT segments in the gut community is mediated by both HGT and strain replacement.** **a** Comparison of HGT prevalence at baseline (blue) and follow-up (green) (two-sided Wilcoxon Signed Rank Tests,  $n = 338$ ). Box plots show the median (center line), the 25th and 75th percentiles (box), and whiskers extending to points within  $1.5 \times$  IQR. Violin plots indicate data distribution. **b** Distribution of HGT segment prevalence at baseline (blue, 338 individuals,  $n = 7581$ ) and follow-up (green, 338 individuals,  $n = 7581$ ). **c** Prevalence of HGT Segment 35 at baseline and follow-up. **d** Phylogenetic tree of *Agathobacter rectalis* (112) (SGB4933) for individuals in selected clusters at baseline and follow-up (see

Figure S6). **e** As in (d) but with individuals hosting the same *A. rectalis* strain at baseline and follow-up highlighted in red. **f** As in (d) but highlighting individuals with different *A. rectalis* strains at baseline and follow-up. Samples with *A. rectalis* with the HGT segments are indicated in yellow. Samples without the HGT segments are indicated in green. Red dashed lines indicate the strain replacement direction of individuals who gained HGT segments at follow-up. Blue dashed lines indicate individual loss of HGT segments at follow-up. In (d–e), the tag <sub>B</sub> indicates a baseline sample and <sub>F</sub> indicates a follow-up sample.

backgrounds. To assess this, we first visualized changes in HGT segments within individuals over the 4-year period (Fig. S7a), which identified an interesting cluster of HGT segments encoding acetylxyloan esterase genes, with markedly higher coverage than adjacent regions, that were flanked by genes characteristic of mobile genomic islands (e.g., Tad/Flp pilus components<sup>27</sup>, type IV secretion system<sup>28</sup>, and toxin-antitoxin system<sup>29</sup>; Fig. S7b, c). For 24 individuals, the presence or absence status of this *A. rectalis* segment had changed in 4 years, and we constructed the phylogenetic tree of *A. rectalis* based on its marker genes for those individuals at both time points (Fig. 4d, see Methods). Interestingly, for 12 individuals, the *A. rectalis* strains at the two timepoints clustered together (phylogenetic distance <0.1) (Fig. 4e), indicating a true HGT event resulting in the gain (or loss) of the segment within the 4-year interval (Fig. 4e). For the other 12 individuals, the *A. rectalis* strains with or without the segment from the two timepoints clustered into different clades (Fig. 4f), demonstrating that the change in the presence of HGT segments was due to strain replacement.

### An individual's HGT remains highly personalized and stable over time, indicating that HGT may serve as a lasting record of host lifestyle

Although the overall prevalence of HGT segments increased over time, the prevalence of 5454 segments (71.9% of all HGT segments) changed minimally over the 4-year period, varying by no more than 5% from baseline (Fig. 5a). We therefore investigated whether overall HGT profiles are individual-specific and persistent over time. Using Jaccard distance to assess the similarity of HGT profiles between the two time points, we found that these profiles were highly individual-specific, with temporal changes smaller than inter-individual differences (9999 permutation Wilcoxon-Mann-Whitney test,  $p$ -value <  $1 \times 10^{-4}$ , Fig. 5b). Based on HGT profiles, we could correctly match 194 out of 338 (57%) paired samples from the same donor, a significantly better performance than could be achieved using taxonomic composition alone (42/338, Fig. 5c) and microbial pathway alone (16/338). To validate this, we further tested whether the HGT profiles we identified could also be used to classify longitudinal samples from the American Human Microbiome Project (HMP) cohort, where 40 donors had microbiome samples collected one-year apart. Our HGT profiles correctly matched samples for 34 out of the 40 donors (Fig. S8), resulting in an 85% accuracy, which is higher than the previously reported 30% prediction accuracy based on microbial species abundance and the 80% prediction accuracy reported for microbial gene abundance<sup>30</sup>. These findings suggest that microbial individual-specificity<sup>17</sup> is reflected not only in microbial composition and genetic makeup but also in HGT profile. We do note that successful matching of longitudinal HMP samples using HGT had to meet two conditions: 1) that the HGT could be found in the samples, and 2) that its individual-specificity could also be validated in the cohort. Overall, the performance of HGT profiles for identifying HMP pairs (one-year apart) is even better than for Lifelines-DEEP pairs (four-years apart), suggesting that HGT profiles are more dynamic over longer time spans.

Since mobile genes are stable and highly individual-specific, we hypothesized that specific lifestyle factors might drive specific gene transfer events. Given that antibiotic use is already well known to promote the transfer of ARGs<sup>31</sup>, we instead focused on the often underestimated influence of non-antibiotic drug use on HGT in the gut microbiome. To estimate the effects of non-antibiotic drugs on HGT changes, we calculated the abundance of mobile genes in HGT segments (see Methods) and applied linear mixed models (LMMs) that accounted for age, sex, and read count as fixed effects and treated individual ID and timepoints as random effects. This analysis identified 48 positive and 14 negative associations (false discovery rate (FDR) < 0.05, Supplementary Data 7).

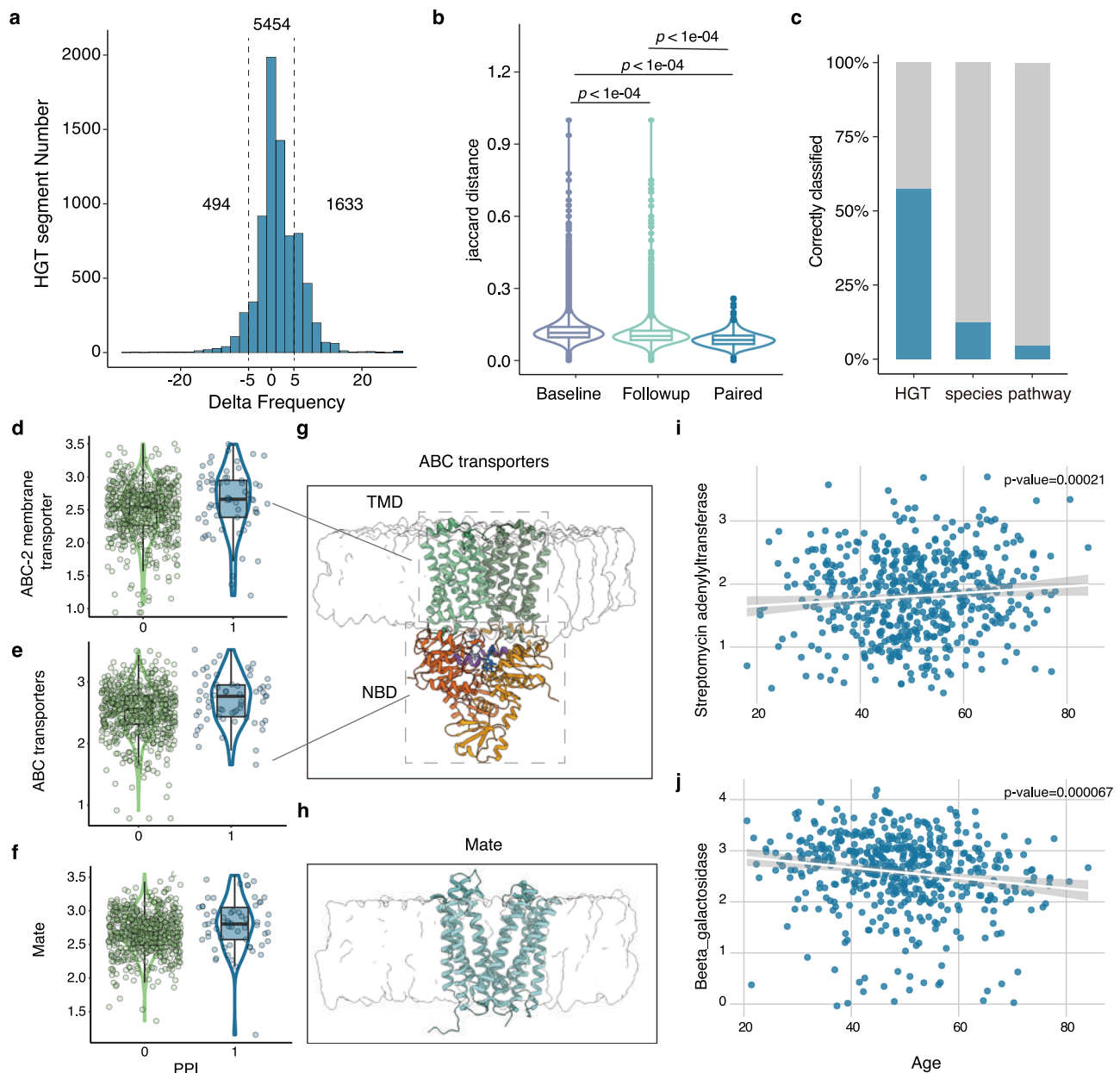
Among the associations, PPI usage was significantly linked to an increased abundance of mobile genes encoding the ABC-2 membrane

transporter (Beta = 0.51,  $p$ -value = 0.0011), the ABC transporter (Beta = 0.54,  $p$ -value = 0.00034), and the multi-antimicrobial extrusion protein (Mate) (Beta = 0.43,  $p$ -value = 0.00052) (Fig. 5d–f). Notably, the genes encoding the ABC transporter and the ABC-2 membrane transporter, which were shared between *Blautia\_A wexlerae* and *Anaerobutyricum hallii*, were located on the same HGT segment. Protein modeling of these genes confirmed that they code for a complete ABC transporter structure comprising two nucleotide-binding domains and two transmembrane domains (Fig. 5g). Similarly, modeling verified that Mate, which was also shared between *Blautia\_A wexlerae* and *Anaerobutyricum hallii*, encodes a complete membrane transporter (Fig. 5h). Both the ABC transporter and Mate can translocate a variety of toxic compounds across membranes, suggesting a potential mechanism by which bacteria gain beneficial genes from other bacteria to mitigate PPI-induced toxicity<sup>32</sup>. Additionally, we observed that the abundance of streptomycin adenylyltransferase within mobile regions was significantly positively associated with age (Beta = 0.030,  $p$ -value = 0.00021, Fig. 5i), corresponding to an approximate 2.0-fold increase per decade. This association may reflect increased antibiotic exposure in elderly individuals, but it could also indicate that resistance genes accumulate and disseminate more readily within the microbial communities of older hosts. In contrast, the abundance of beta-galactosidase within mobile regions showed a significant negative association with age (Beta = -0.042,  $p$ -value = 0.000067, Fig. 5j), corresponding to an approximate 0.38-fold decrease per decade. This observation might be explained by higher dairy product consumption among younger individuals, which stimulates microbial beta-galactosidase production for lactose degradation, while reduced dairy intake in older adults could lead to decreased enzyme abundance. Overall, our results suggest that the mobile gene pool within the human microbiome can serve as a reflection of host lifestyle.

## Discussion

This study presents a large-scale, longitudinal metagenomic investigation into the dynamics of recent HGT in the human gut microbiome. Our findings collectively demonstrate that HGT segments can disseminate among hosts, either directly through gene transfer events or indirectly via strain transmission, thereby enhancing their prevalence within populations. Additionally, we observe that species engaged in HGT frequently establish stable co-abundance relationships and tend to accumulate resistance genes. Our results also highlight the individual-specific nature of HGT events and their responsiveness to host lifestyle factors such as aging and non-antibiotic drug usage. These insights expand current knowledge regarding microbial interactions mediated by gene transfer within the human gut and underscore the potential of HGT as a valuable record of recent human lifestyle.

HDMI represents a novel workflow capable of detecting recent HGT events (0–10,000 years) at the individual level. Although recent HGT in the human gut microbiome has been shown to be closely related to human lifestyle and metabolic capabilities<sup>5,13,15</sup>, most HGT detection tools cannot distinguish between recent and long-term events<sup>14,33,34</sup>. MetaCHIP is one of the few methods capable of distinguishing between recent and long-term HGT events through phylogenetic tree comparisons, but it cannot be applied at the individual level<sup>35</sup>. In contrast, the recently published tool WAAFL<sup>14</sup> can infer HGT events from assembled contigs at the individual level and has demonstrated superior performance to MetaCHIP<sup>14</sup>. Our benchmark results show that HDMI outperforms WAAFL<sup>14</sup>, especially in detecting intra-genus HGT events. As both previous studies<sup>5,14</sup> and our findings indicate, phylogenetically close bacteria are more likely to exchange genes, making intra-genus HGT events a major component of recent gene transfer, further highlighting the advantage of our tool. Building on this approach, we identified previously unrecognized gene transfer events in longitudinal samples from 338 individuals, significantly



**Fig. 5 | Association between mobile genes and human lifestyle.** **a** The delta prevalence of HGT segments from baseline to follow-up (338 individuals,  $n = 7581$ ). **b** Comparison of the inter-individual Jaccard distance of the HGT profile at baseline (blue,  $n = 56,953$ ) and follow-up (green,  $n = 56,953$ ), as well as the Jaccard distance within paired samples at the two time points (cyan,  $n = 338$ ). Group differences were assessed using the Wilcoxon rank-sum test. The empiric  $p$ -value was calculated by permuting samples 9999 times. **c** Accuracy in classifying longitudinal samples of 338 individuals using their HGT profile, species abundance, and microbial pathway abundance. **d-f** Association between proton pump inhibitor (PPI) intake and mobile gene abundance ( $n = 676$ ). Gene abundance was transformed by  $\log_{10}$ . Violin plots show the kernel density of the data. For the box plots,

the center line denotes the median; the box bounds indicate the 25th and 75th percentiles (IQR); whiskers extend to the smallest and largest observations within  $1.5 \times \text{IQR}$  from the lower and upper quartiles, respectively. **g** Structure of the ABC transporter. Gray surface: membrane. Green: two transmembrane domains (TMD). Orange: two nucleotide-binding domains (NBD). Blue sticks: ATP molecules. Purple: ABC signature sequence. **h** Structure of Mate. Gray surface: membrane. Green: TMD. **i, j**, Association between age and mobile gene abundance ( $n = 676$ ). Points represent individual samples. The solid line denotes the fitted linear regression, and the shaded band shows the 95% confidence interval. See Methods for model details.

broadening our understanding of the mobile gene pool in the human gut microbiome. This methodological advance enables the analysis of microbial gene flow at the individual level and lays the groundwork for future investigations into the ecological role of HGT.

This study also demonstrates that HGT promotes positive co-abundance relationships and enhances the stability of microbial community structure. While previous cross-sectional studies have indicated that co-occurring microorganisms are more likely to participate in HGT<sup>36</sup>, no research has examined the role of HGT in human

gut microbiota temporal stability. Theoretical models and experimental studies support the notion that HGT fosters cooperative coexistence among microorganisms<sup>6,37</sup>. For instance, community stability analyses based on the generalized Lotka–Volterra model have shown that introducing highly transferable mobile genes, such as resistance genes, can improve the community's resilience to disturbances<sup>38</sup>. By enabling the widespread dissemination of beneficial functions, HGT allows more species to share critical survival advantages, thereby promoting multi-species coexistence and enhancing



overall system homeostasis. Our longitudinal observations revealed that species involved in HGT tend to accumulate more resistance genes and exhibit more stable co-abundance relationships, strongly confirming the predictions of both experimental and modeling studies<sup>37–39</sup>. Our findings provide new evidence for the long-term impact of HGT on human gut microbial communities and suggest that, in addition to species diversity and relative abundance, the extent of interspecies genetic exchange may be a key factor in maintaining intestinal microecological homeostasis.

In previous studies, HGT has been reported to facilitate the spread of mobile genes through two primary mechanisms. First, mobile genes can be directly transferred between individuals via mobile elements<sup>15</sup>. For example, in a mother-to-child cohort, phage-associated gene fragments in maternal strains were transferred to distinct strains in infants without direct strain transfer, enabling the infant's intestinal microbes to acquire additional functions<sup>15</sup>. Second, mobile genes may be acquired from new strains through HGT. In fecal microbiota transplantation studies, newly introduced strains were observed to transfer genes to other species in the recipient microbiota<sup>40</sup>, illustrating a scenario involving both strain and gene transfer. Our results add another dimension by showing that strain replacement, where strains carrying mobile genes supplant the original strains, leads to the acquisition of mobile genes by the species as a whole. Moreover, we demonstrate that, in addition to exceptional cases like mother-to-child transmission and fecal microbiota transplantation, the natural spread of mobile genes via HGT is a common phenomenon in the population. Together, HGT and strain replacement ensure the persistence and dissemination of key genes within individuals and across populations<sup>11</sup>.

It is noteworthy that numerous previous studies have demonstrated that the mobile gene pool in the microbiome is largely shaped by environmental factors and lifestyle<sup>5,13</sup>. These findings suggest that each individual's intestinal microbiome may accumulate a unique series of HGT events that reflect their lifestyle history. Our data strongly support this view: we observed that HGTs are highly individual-specific and remain relatively stable over multiple years of follow-up, transcending the effects of short-term fluctuations in microbial communities. These long-term preserved mobile genes act as a “history book” of the microbiome, recording the host's environmental exposures and selection pressures. For instance, we detected the recent transfer of tetracycline resistance genes and the enrichment of specific resistance genes associated with PPI usage. Although our current detection accuracy has not yet achieved full personalization, we believe that an individual's microbial mobile gene pool holds promise for future applications, such as inferring disease exposure history and evaluating the efficacy of personalized medical interventions, thereby expanding the role of the microbiome in precision medicine.

### Limitations of the study

We acknowledge several limitations of our study. The Lifelines-DEEP cohort comprises participants of Dutch ancestry from the northern region of the Netherlands, so the results might be biased toward a region-specific microbial background and local environmental exposures. In addition, the Lifelines-DEEP cohort includes mainly healthy individuals, which limits its power to detect associations between HGT, diseases, and drug usage. We also acknowledge that our workflow is sensitive to MAG quality and number, and therefore captures only a part of the HGT landscape within the gut ecosystem. Since MAGs are constructed by assembling short sequence reads and binning genomic contigs, reads from HGT regions may map to multiple genomes. This can lead to inaccurate binning and the formation of incomplete or erroneous MAGs, with such contamination (i.e., heterologous sequences) potentially being misidentified as HGT events<sup>41</sup>. Multi-sample metagenomic binning can significantly improve both the number and quality of MAGs<sup>42</sup>, and subsequent detection<sup>41</sup> and removal of chimerism and contamination can help reduce these issues.

However, contigs containing heterogeneous regions may not be binned with the rest of the genome<sup>9</sup>, causing many HGT events to be lost. Moreover, in our approach, we identified HGT in the baseline samples and then tracked its spread over time. This approach will not detect HGTs that occurred only in the follow-up samples.

Our current HDMI-based workflow, by default, does not infer HGT direction and does not exploit all available contigs, so we had to predict HGT direction by comparing gene trees to species trees created using RANGER-DTL. In future work, combining MAGs from both time points, integrating un-binned contigs with those assigned to MAGs, and comparing MAGs with reference databases (and gene–species tree reconciliation where appropriate), may enhance HGT discovery and directionality inference. These considerations suggest that complementary methods such as WAAFL will continue to play an important role. Additionally, although HGT can integrate into a recipient genome through both recombination and insertion, our individual-level detection only captures HGT events via insertion, as our method does not detect gene replacement through recombination. This limitation, combined with the low coverage depth of low-abundance species, results in numerous NA values in the temporal analysis of HGT segments between baseline and follow-up.

## Methods

### Ethical Approval

The Lifelines study was approved by the Ethics Committee of the University Medical Center Groningen (METc 2007/152). All participants signed an informed consent form prior to enrollment. The LLD study was approved by the Institutional Ethics Review Board of the University Medical Center Groningen (UMCG) (ref. M12.113965), the Netherlands.

### Experimental model and subject details

**Human subjects.** The Lifelines-DEEP study, part of the Lifelines biobank with over 167,729 participants, focuses on a select group of 1,539 individuals to explore various factors affecting health outcomes in the northern Netherlands. A follow-up study was recently conducted on 338 cohort participants who had been analyzed in 2013<sup>43</sup>. As described previously<sup>17</sup>, follow-up stool samples were collected for these 338 individuals (55.6% female and 44.4% male) at the second time point. The duration between the two time points ranged from 3.33 to 3.92 years (mean = 3.53, SD = 0.12). At baseline, the mean age of participants was 48.2 years (range = 18–80, SD = 11.7) and their mean BMI was 25.4 (range = 17.6–43.3, SD = 4.08). At follow-up, the mean age was 51.7 years (range = 22–84, SD = 11.7), and the mean BMI was 25.6 (range = 16.1–37.6, SD = 4.0). Phenotypic data assessed in this study included anthropometric traits (e.g., age, sex, BMI, height) and usage of 14 medications (e.g., PPIs, oral contraceptives, beta-blockers, statins).

### Method Details

**Metagenomic data generation and preprocessing.** Stool sample collection and processing followed the same protocol at both time points. All participants were asked to collect fecal samples at home and to place them in their home freezer (–20 °C) within 15 minutes after production. Subsequently, a nurse visited the participant to pick up the fecal samples on dry ice and transfer them to the laboratory. Aliquots were then made and stored at –80 °C until further processing. The same protocol for fecal DNA isolation and metagenomics sequencing was used at both time points. Fecal DNA isolation was performed using the AllPrep DNA/RNA Mini Kit (QIAGEN cat. 80204). After DNA extraction, fecal DNA was sent to the Broad Institute of Harvard and MIT in Cambridge, Massachusetts, USA, where library preparation and whole-genome shotgun sequencing were performed on the Illumina HiSeq platform. From the raw metagenomics sequencing data, low-quality reads were discarded by the sequencing facility, and reads belonging to the human genome were removed by mapping the data to the human

reference genome (version NCBI37) with KneadData (version 0.4.6.1) and Bowtie2 (version 2.1.0)<sup>44</sup> (Table 1). After filtering, the average read depth was 12.3 million for both baseline and follow-up samples. The read depths of all samples at both time points were not significantly different (paired Wilcoxon test,  $p$ -value = 0.89).

**De novo assembly and quality control.** MetaSPAdes<sup>45</sup> was used to perform de novo assembly for each sample in Lifelines-DEEP Baseline ( $n = 338$ ). The assembled contigs were further binned and refined using MetaWRAP<sup>46</sup>. The quality of the MAGs was assessed by CheckM<sup>47</sup>. Genomes with >90% completeness and <5% contamination were retained. dRep<sup>48</sup> was used to dereplicate MAGs with the option -sa 0.998 -pa 0.95 to ensure non-identical genomes were included.

## MAG clustering and taxonomy classification

dRep compare was used to compare and cluster MAGs. MASH<sup>49</sup> was used to form primary clusters with a threshold of 0.95. fastANI<sup>50</sup> was used to create secondary clusters with a threshold of 0.99. The primary and secondary clusters were used for grouping genomes to species- and strain-level. Taxonomic classification of the genomes was performed using GTDB-Tk<sup>51</sup> with default parameters. All genome taxonomies and groups are compiled in Supplementary Data 1. To robustly infer vertical phylogenetic distances, we used UBCGtree v2.0<sup>52</sup> to construct MAG phylogenomic trees based on 81 universal single-copy core genes identified by UBCG2 using profile-specific HMMER trusted cutoffs derived from a reference set of >3,500 species spanning 43 phyla and present in ≥95% of genomes.

**Table 1 | Key resources table**

| REAGENT or RESOURCE              | SOURCE                                | IDENTIFIER  |
|----------------------------------|---------------------------------------|---|
| <b>Biological samples</b>        |                                       |   |
| Fecal samples                    | This study                            | <a href="https://www.Lifelines.nl">https://www.Lifelines.nl</a>   |
| <b>Deposited data</b>            |                                       |   |
| LLD baseline metagenomics        | European Genome-Phenome Archive (EGA) | EGAD00001001991   |
| LLD follow-up metagenomics       | EGA                                   | EGAD00001006959   |
| <b>Software and algorithms</b>   |                                       |   |
| R (version 4.0.3)                | R Foundation                          | <a href="https://www.r-project.org/">https://www.r-project.org/</a>   |
| Python (version 3.7.4)           | Python                                | <a href="https://www.python.org">https://www.python.org</a>   |
| KneadData (version 0.4.6.1)      | 44                                    | <a href="https://huttenhower.sph.harvard.edu/kneaddata/">https://huttenhower.sph.harvard.edu/kneaddata/</a>                               |
| Megahit                          | 86                                    | <a href="https://github.com/voutcn/megahit">https://github.com/voutcn/megahit</a>   |
| dRep (version 3.4.3)             | 48                                    | <a href="https://github.com/MrOlm/drep">https://github.com/MrOlm/drep</a>   |
| Metawrap (version 1.3.2)         | 46                                    | <a href="https://github.com/bxlab/metaWRAP">https://github.com/bxlab/metaWRAP</a>   |
| CheckM (version 1.2.2)           | 47                                    | <a href="https://github.com/ECogenomics/CheckM">https://github.com/ECogenomics/CheckM</a>   |
| GTDBTk (version 2.1.1)           | 51                                    | <a href="https://github.com/ECogenomics/GTDBTk">https://github.com/ECogenomics/GTDBTk</a>   |
| MASH (version 2.3)               | 49                                    | <a href="https://github.com/marbl/Mash">https://github.com/marbl/Mash</a>   |
| FastANI (version 1.33)           | 50                                    | <a href="https://github.com/ParBLISS/FastANI">https://github.com/ParBLISS/FastANI</a>   |
| Bowtie 2 (version 2.5.1)         | 53                                    | <a href="https://github.com/BenLangmead/bowtie2">https://github.com/BenLangmead/bowtie2</a>   |
| CD-HIT (version 4.8.1)           | 63                                    | <a href="https://sites.google.com/view/cd-hit">https://sites.google.com/view/cd-hit</a>   |
| Prodigal (version 2.6.3)         | 64                                    | <a href="https://github.com/hyattpd/Prodigal">https://github.com/hyattpd/Prodigal</a>   |
| Barrnap (version 0.9)            | 87                                    | <a href="https://github.com/tseemann/barrnap">https://github.com/tseemann/barrnap</a>   |
| Prokka (version 1.14.6)          | 71                                    | <a href="https://github.com/tseemann/prokka">https://github.com/tseemann/prokka</a>   |
| Blastn (version 2.6.0)           | 58                                    | <a href="https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/">https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/</a> |
| eggNOG-mapper (version 2.1.11)   | 65                                    | <a href="http://eggnog5.embl.de">http://eggnog5.embl.de</a>   |
| Resistance Gene Identifier (RGI) | 67                                    | <a href="https://github.com/arpcard/rgi">https://github.com/arpcard/rgi</a>   |
| Traitair (version 1.1.1)         | 72                                    | <a href="https://github.com/hzi-bifo/traitar">https://github.com/hzi-bifo/traitar</a>   |
| Fastspar (version 1.0.0)         | 25                                    | <a href="https://github.com/scwatts/fastspar">https://github.com/scwatts/fastspar</a>   |
| UBCGtree (version 2.0)           | 52                                    | <a href="http://leb.snu.ac.kr/ubcg2/about">http://leb.snu.ac.kr/ubcg2/about</a>   |
| vegan R package                  | 81                                    | <a href="https://cran.r-project.org/web/packages/vegan/index.html">https://cran.r-project.org/web/packages/vegan/index.html</a>           |
| lme4 R package                   | 80                                    | <a href="https://cran.r-project.org/web/packages/lme4/index.html">https://cran.r-project.org/web/packages/lme4/index.html</a>             |
| lmtest R package                 | 84                                    | <a href="https://cran.r-project.org/web/packages/lmtest/index.html">https://cran.r-project.org/web/packages/lmtest/index.html</a>         |
| ggplot2                          | 88                                    | <a href="https://cran.r-project.org/web/packages/ggplot2/index.html">https://cran.r-project.org/web/packages/ggplot2/index.html</a>       |
| igraph v1.2.6                    | 76                                    | <a href="https://cran.r-project.org/web/packages/igraph/index.html">https://cran.r-project.org/web/packages/igraph/index.html</a>         |
| ggraph v2.0.5                    | 77                                    | <a href="https://cran.r-project.org/web/packages/ggraph/index.html">https://cran.r-project.org/web/packages/ggraph/index.html</a>         |
| ncf v 1.3-2                      |                                       | <a href="https://github.com/objornstad/ncf">https://github.com/objornstad/ncf</a>   |
| lme4qtl v0.2.2                   | 78                                    | <a href="https://rdr.io/github/variani/lme4qtl/man/relmatLmer.html">https://rdr.io/github/variani/lme4qtl/man/relmatLmer.html</a>         |
| ComplexHeatmap v2.18.0           | 89                                    | <a href="https://github.com/jokergoo/ComplexHeatmap">https://github.com/jokergoo/ComplexHeatmap</a>                                       |
| <b>Database</b>                  |                                       |   |
| Pfam database                    | 66                                    | <a href="https://www.ebi.ac.uk/interpro/entry/pfam">https://www.ebi.ac.uk/interpro/entry/pfam</a>   |
| dbCAN                            | 90                                    | <a href="https://bcb.unl.edu/dbCAN/">https://bcb.unl.edu/dbCAN/</a>   |
| CARD database                    | 67                                    | <a href="https://card.mcmaster.ca/analyze/rgi">https://card.mcmaster.ca/analyze/rgi</a>   |

### Species abundance calculation

Species relative abundance was calculated via the sum of the genome abundance within each species in the samples. For species with more than five genomes, we randomly selected five genomes to calculate the relative and median abundance of species. We mapped reads against all selected genomes using Bowtie2<sup>53</sup> and calculated the depth of coverage in all contigs in each genome as:

$$D_{genome_i} = (Coverage_{11}, Coverage_{12}, \dots, Coverage_{1N}, Coverage_{21}, \dots, Coverage_{NL_N})$$

Where N represents the number of the contigs and  $L_N$  represents the position in the contig. The median abundance of each genome in the metagenome was calculated with Median (D). The per-base depth of coverage K, the average read length L, the size of each genome S, and the total read number T in the shotgun data are used to calculate the relative abundance A of each genome in the metagenome according to  $A = (K*S/L) / T$ .

### HGT detecting workflow

In this work, we introduce a workflow to identify recent HGT at the individual level. In brief, HDMI works as follows. First, HDMI searches for genomic regions of at least 500 bp that show >99% identity between any pair of MAGs with ANI < 95%. These regions are likely to be transfer events that occurred within the last 10,000 years (hereafter referred to as “recent” HGTs), assuming a molecular clock of 1 SNP/genome/year for a genome size of  $10^6$  bp<sup>54–56</sup>. Second, HDMI performs several quality checks. It first checks MAG contamination and identifies split reads that span the junction between the HGT regions and their flanking regions. In addition, we required at least 90% coverage of the breadth and a median 1x genome depth for HGT segments. Third, HDMI excludes false positive HGTs due to highly conserved genes with >99% identity. For this, we extracted and calculated the identity of 81 nearly universal single-copy core genes (including 42 ribosomal proteins)<sup>57</sup> between all cross-species genome pairs. Lastly, inspired by the detection of differential alternative splicing events in transcriptomes, we inferred whether each HGT in the sample is inserted and determined the ratio of insertion to non-insertion by calculating the number of reads mapping to the split sites of sequences with HGT insertion and the junction site with no HGT insertion (Fig. S1). Each of these steps is described in more detail below.

### HGT candidate detection

In this pipeline, we only focused on transfers occurring between bacterial species (ANI < 0.95), ignoring within-species (ANI ≥ 0.95) gene recombination events. We screened all genomes and used Blastn v2.6.0<sup>58</sup> to identify genomic segments of at least 500 bp that were shared between any pair of genomes from different species with an identity >99%

To exclude segments carrying potentially conserved genes, which are evolving slowly within species and more likely to have a high identity, we referred to the UBCG2 resource<sup>57</sup>. This resource defines 81 nearly universal single-copy core genes, including 42 ribosomal proteins, that are thought to be vertically transmitted genes. We then calculated the identity of each vertically transmitted gene in all species pairs and excluded between-species genome pairs containing any vertically transmitted gene with >99% identity. For the remaining genome pairs, the best explanation for these high homologous segments is HGT rather than vertical inheritance because the expected identity between highly conserved and vertically inherited genes of different species exceeds the 99% identity threshold used in our approach to retain HGT candidates.

In addition, the Assembly algorithm (metaSPAdes) based on DeBruijn graphs can produce contamination for regions with sequencing errors but high similarity<sup>59</sup>. This process produces ‘bubbles’ that can result in the generation of two contigs with overlapping

sequences at their ends. If not handled properly, such duplications may lead to erroneous conclusions in HGT analysis. To mitigate this risk, we disregarded HGT candidates located within 100 bp of the end of the contig. We also excluded any putative HGT candidates found on contigs that matched >90% of their full-length with a longer contig, as these are likely to be artificial duplicates created during the assembly process<sup>35</sup>.

### Cohort-based HGT event validation

If the median abundance of all MAGs was 0 among the five randomly selected MAGs, the abundance of the species was considered NA. To make sure the transferred region exists in the genome, the HGT region should have a minimum of 90% breadth of coverage. Subsequently, we counted those reads where one part mapped in the putative HGT and the other part mapped to the flanking region at either end of the HGT to suggest true HGT events.

Bowtie2 with option -a --very-sensitive --no-unal was used for read-mapping. For each transferred sequence, we required it to have at least three reads mapping its start and end sites, with at least 10 bps of overhang on either side. If any end lacked sufficient read support, the median abundance of all selected genomes of the species was 0, or the HGT region had <90% breadth coverage, the presence of this putative HGT in that genome was considered NA. We then conducted detection for both genomes involved in each HGT event. If we observed the HGT in only one of the two genomes of an HGT event, across the entire cohort, the HGT event was considered a false positive.

### HGT profiling

Each HGT event is considered to be two HGTs in two species due to their unique insertion sites (start, end) and the potential to be acquired or lost. First, we extracted sequences containing an HGT and concatenated the flanking regions of the HGT to get sequences without the HGT. We then counted the reads spanning the start and end of the split site with HGT insertion ( $HI_1$ ,  $HI_2$ ) and the reads spanning the site with no HGT insertion ( $nHI$ ) in each sample using the same strategy we used in the second step above. Finally, we calculated the HGT’s presence/absence by:

$$HI = \min(HI_1, HI_2) \text{ if } \min(HI_1, HI_2) \geq 3 \text{ else } 0$$

$$nHI = nHI \text{ if } nHI \geq 3 \text{ else } 0$$

$$HGT \text{ presence} = \frac{HI}{HI + nHI} > 0$$

$$HGT \text{ absence} = \frac{HI}{HI + nHI} = 0$$

Here, if the HI and nHI are both 0, it is NA.

### Benchmark HDMI

Selected representative genomes were downloaded from NCBI with GCA ID (see Supplementary Data 8). Seqkit was used to randomly select genes from the donor genome<sup>60</sup>. HgtSIM<sup>61</sup> was used to simulate the insertion of genes with genetic divergence varying between 99% and 100% (-f genome -r 1-0-1-1 -x fna -mixed 0-1 -keep\_cds -a genebank). ART (Version 2.5.8)<sup>62</sup> was used to simulate metagenomic sequencing data at 6X, 9X, and 12X coverage (art\_illumina -ss HS25 -l 150 -f 9 -p -m 500 -s 10).

In our comparison of WAAFL and HDMI, we assessed the recovery rates of both tools at various taxonomic levels by simulating gene insertions in genome pairs spanning different evolutionary distances (i.e., intra-genus, intra-order, and intra-phylum, Supplementary Data 8). For each genome pair, we generated three metagenomic datasets with distinct ratios (9:12, 12:6, and 6:10). Simulated reads were

further assembled by metaSPAdes<sup>45</sup>, and the resulting contigs were binned using MetaWRAP<sup>46</sup> to obtain MAGs. In HDMI, a correct detection is defined as the identified insertion gene matching the simulated insertion gene, with a maximum length discrepancy of no more than 1% of the full gene length. Similarly, in WAAFLE, a correct detection event is defined as the transferred gene (i.e., directional HGT event) identified by WAAFLE as being identical to the simulated insertion gene, allowing for a length difference of no more than 1% of the complete gene length. Since real genomes exhibit both recent and ancient HGT events beyond the simulated gene insertions, our comparison focused exclusively on the recovery rates of the simulated gene insertions.

**HGT clustering, gene annotation, and the direction in the cluster**  
HGT sequences were clustered using cd-hit<sup>63</sup> (-aS 0.9 -aL 0.9 -c 0.9), representing HGTs with similar gene content and function. The coding sequences (CDS) were assigned to all HGTs using Prodigal V2.6.3<sup>64</sup> in metagenome mode to capture gene segments. eggNOG-mapper<sup>65</sup> was used to assign putative function predictions to genes, and the queries were realigned to the Pfam<sup>66</sup> domains to get the Pfam function domain annotation (-evalue 0.001 --score 60 --pident 40 --query\_cover 20 --pfam\_realign realign). The RGI and CARD database<sup>67</sup> was used to predict ARGs with default parameters.

In each cluster, sequences were aligned using the auto option in mafft<sup>68</sup>, and the gene tree was constructed with IQ-TREE<sup>69</sup> (-m MFP). We then subsampled the species tree from the comprehensive MAGs tree using the ETE Toolkit. The subtree was used to infer the root of the gene tree using the OptRoot module from RANGER-DTL v.2.0. We then ran RANGER-DTL with default settings to reconcile the gene tree and the genomic tree a total of 500 times. Reconciliations from each optimal root were aggregated using the AggregateRanger module from RANGER-DTL v.2.0. The transfer direction was then extracted for annotation information in each cluster.

### Pangenome construction and trait prediction for each species

To represent the function of each species, we constructed the pangenome<sup>70</sup>. Based on the MAG's taxonomic classification, in each species, we performed gene calling using Prodigal<sup>71</sup> in all MAGs belonging to the species and clustered genes using cd-hit (-c 0.9). The set of non-redundant genes comprising all MAGs from the same species is the meta-pangenome of that species. Traitair<sup>72</sup> was used to predict different phenotype traits for each species, including energy resources for growth, enzymatic activities, and morphology. The Phypat and PGL algorithms were used to predict traits. To avoid over-interpretation of false positive traits, we only considered the 22 traits with >90% predictive accuracy using the Phypat + PGL method, following the accuracy evaluation in the original paper<sup>72</sup>.

ARGs were predicted using strict mode in SRID<sup>67</sup>. VF identification was done using the core set of Virulence Factors of Pathogenic Bacteria Database (VFDB 2022<sup>73</sup>) with the BLASTP option of the Diamond software with strict parameters (e-value < 10<sup>-5</sup>, >50% identity at the protein level, and 70% query sequence coverage). AMPRs were identified by performing a BLASTP sequence similarity search against the manually curated list of AMPRs<sup>74</sup> with the same parameters (e-value < 10<sup>-5</sup>, >50% identity at the protein level, and 70% query sequence coverage). Genes encoding CAZymes were identified using dbCAN (CAZyDB.08062022.dmnd). The proportion of CAZyme genes for a particular substrate was calculated as the number of the CAZyme genes involved in its utilization divided by the total number of the CAZyme genes. CAZyme classification was described in a previous study<sup>75</sup>.

### HGT network construction and degree calculation

The HGT network was constructed based on the transfer between species. The network's edges are unweighted. The R package igraph v1.2.6<sup>76</sup> was used to construct the network and calculate the

betweenness and degree centrality, with ggraph v2.0.5<sup>77</sup> used to visualize the network. An LMM with phylogenetic distance matrices as random effect was fitted to examine the association of a bacteria trait and its centrality in the HGT network, adjusting the effect of abundance and phylogenetic distance with the function relmatLmer in R package lme4qtl (version 0.2.2)<sup>78</sup>.

Model1: relmatLmer (Centrality - Abundance + (1 | Phylogenetics Distance), REML = FALSE)

Model2: relmatLmer (Centrality - Trait + Abundance + (1 | Phylogenetic Distance), REML = FALSE)

Model3: relmatLmer (Centrality - MoaC + Abundance + (1 | Phylogenetic Distance), REML = FALSE)

LRT was used to measure the effect of traits on HGT network centrality:

LRT\_Trait = lrtest(model1, model2)

### The effect of correlation and phylogenetic distance on HGT rates

Based on the species abundance calculated above, we used FastSpar<sup>25</sup>, a C++ implementation of the SparCC algorithm<sup>79</sup>, to calculate the correlations between the genomes, with 1000 permutations at each time point based on the abundances of the genomes across the individuals. Correlations with  $p \leq 0.01$  and  $|\rho| \geq 0.2$  were then retained for further analysis.

For MAG-based HGT rate estimation, we implemented a previously published conservative approach in isolates<sup>10</sup>, defining the HGT rates for a species pair as the proportion of between-species genome pairs that share at least one HGT among all between-species genome pairs. Species with <3 MAGs were not considered in calculating HGT rates.

Based on the phylogenetic tree built using 81 nearly universal single-copy genes, we further calculated the phylogenetic distance across all species pairs using the Python package etc3. The phylogenetic distances between species were calculated by averaging the distances of all genome pairs in each species pair.

We used the generalized linear and logistic regression models to measure the effect of microbial co-abundance on HGT rates and occurrence. The partial.mantel.test function from the ncf R package (version 1.3-2)<sup>80</sup> was used with pairwise deletion.

HGT\_Occur: At least one between-species genome pair shares at least one HGT.

Model1: partial.mantel.test (Co-abundant correlation matrix, HGT\_Rates matrix, Phylogenetics Distance matrix, resamp = 9999)

Model2: partial.mantel.test (Co-abundant correlation matrix, HGT\_Occur matrix, Phylogenetics Distance matrix, resamp = 9999)

The LMM with phylogenetic distance matrices was fitted to examine the association of the number of function genes in the species pangenome and their communities.

Model3: relmatLmer (Gene number - Abundance + (1 | Phylogenetics Distance), REML = FALSE)

Model4: relmatLmer (Gene number - Community + Abundance + (1 | Phylogenetics Distance), REML = FALSE)

Model5: relmatLmer (Bray-Curtis distance - Community + (1 | Phylogenetics Distance), REML = FALSE)

Model6: relmatLmer (Bray-Curtis distance - (1 | Phylogenetics Distance), REML = FALSE)

To assess whether the Community is significantly contributing to genes enrichment, we performed the following LRT: Trait = lrtest(-model3, model4). To assess whether the Community is significantly contributing to temporal abundance changes, we performed the following LRT: Trait = lrtest(model5, model6).

### Microbial co-abundance and HGT network

The microbial co-abundance and HGT network was constructed by correlation, using cluster\_leiden function with objective\_function

Constant Potts Model in *igraph*<sup>76</sup> to define communities. The layout of the network was visualized by *ggraph*<sup>77</sup>, and the length of the edges in the network was represented by the transformed correlation:  $1 - ((\text{correlation} + 1)/2)$ . Within Community C (Fig. S6), three *Enterobacteriaceae* bacteria were seen to contain a greater than 10-fold higher abundance of ARGs, AMPRs, and VF genes compared to other species, which might explain the enrichment in Community C. To avoid introducing bias to the following analysis, these three *Enterobacteriaceae* species were excluded from our analysis.

### Transmission profiling

Each HGT event ( $n = 5644$ ) refers to two HGTs ( $n = 11,288$ ) with identical sequences shared between two genomes, resulting in 7581 non-redundant HGT segments distinguished by genomes and insertion locations. First, we extracted the genomic sequences containing the detected HGTs from the MAGs. We then removed the HGT region from the genome, concatenating the flanking regions as if the HGT were not present. Finally, we searched for reads that span this region and align with the flanking regions with no read split, further confirming the absence of HGTs within the samples, as described in the 3<sup>rd</sup> step of the workflow. We then calculated the prevalence of each HGT across all individuals at baseline and follow-up, respectively, and determined the delta prevalence of each HGT. Based on the matrix of HGT presence and absence, we applied the *vegdist()* function from the R package *vegan* (version 2.5.5)<sup>81</sup> to calculate the Jaccard distance dissimilarity matrix. To compare the Jaccard distance dissimilarity between and within individuals, we used the R package *coin*<sup>78</sup> to calculate the empiric *p*-value by permuting samples of the HGT matrix 9999 times.

### Strain-level profiling of samples

Here we first applied *PhyloPhlan3* to assign a *MetaPhlan 4 v Jan21* species genome bin label to MAGs. We only looked at the species *Agathobacter rectalis*, which was assigned to SGB4933. We then ran *StrainPhlan4* in `t_SGB4933_group` with option `--marker_in_n_samples 50 --sample_with_n_markers 50 --secondary_sample_with_n_markers 50 --sample_with_n_markers_after_filt 33`. The multiple sequence alignment was built on 187 available markers, and 551/676 samples were used to build the phylogenetic tree. Only those individuals with two timepoints in the resulting tree were included. Samples in which strains' phylogenetic distance was  $< 0.1$  were considered to have the same strain. The tree was visualized with R package *ggtree*<sup>82</sup>.

### Function enrichment

To avoid overestimating the frequency of HGTs, we established a non-redundant HGT gene database for each species. First, we tagged the HGTs with information about their host species. We then separately predicted the genes of the HGTs present in each species and used *cd-hit* to de-replicate these genes, using parameters identical to those used in constructing the species pangenome. We assigned the highest-scoring Pfam functional domain annotations to each species' non-redundant genes and each species' non-redundant mobile genes using *eggNOG-mapper*<sup>65</sup> (`--pfam_realign realign --evaluate 0.001 --score 60 --pidnt 40 --query_cover 20 --subject_cover 20`). Subsequently, we merged the non-redundant genes and non-redundant mobile genes of each species and calculated the number and frequency of each Pfam functional domain across all genes and all mobile genes. *Foldchange* was calculated as: the frequency of Pfam function domain in HGTs / the frequency of Pfam function domain in all genes in each species. Fisher's exact test was used to determine the significance of the annotation enrichment. Two-tailed *p*-values were corrected using the Benjamini-Hochberg FDR method.

### HGT profile as a fingerprint

To test how well the HGT profiles distinguish samples from the same individual, we used the abundance matrix described previously and

generated the Bray-Curtis distances between all samples at two time-points. If two samples (and only these two samples) from the same individual had the closest distance, we considered them correctly linked. To assess generalizability, we obtained HMP metagenomes (Human Microbiome Project Data Analysis and Coordination Center, [hmpdacc.org](http://hmpdacc.org)) and for each sample estimated the abundance of the HGT fragments identified in our cohort, following the same procedures described above.

### Phenotype association

The mobile genes were predicted from HGTs using *Prodigal*, using *cd-hit* to dereplicate with the same parameters, as described above. *CoverM*<sup>83</sup> was used to calculate the abundance of those mobile genes with the parameters `--min-read-aligned-percent 50 --min-read-percent-identity 99 --min-covered-fraction 4`. To measure associations between mobile genes and drug usage, smoking, and alcohol intake frequency, we used the *glmer* function in the *lme4* R package<sup>80</sup> to fit the generalized linear mixed effects models. The *lmerTest* R package<sup>84</sup> was used to estimate the *p*-value.

Model: Gene abundance Association (joint Association)

Genes abundance ~ Phenotype + clean\_reads + sex + age + (1 | IndividualID) + (1 | TimePoint)

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

All relevant data supporting the key findings of this study are available within the article and its supplementary information files. The raw metagenomic sequencing data and basic phenotypes (i.e., age and sex) of the Lifelines-DEEP participants at both time points are available from the European Genome-Phenome Archive (<https://ega-archive.org>) via accession numbers EGAD00001001991 and EGAD00001006959, respectively. Due to informed consent regulations, detailed phenotypic data for the Lifelines-DEEP cohort can be requested from Lifelines (<https://www.lifelines.nl/researcher>) by submitting an intention letter to the Lifelines Data Access Committee responsible for the Lifelines-DEEP data (contact: Jackie Dekens, email: [j.a.m.dekens@umcg.nl](mailto:j.a.m.dekens@umcg.nl)). The availability of datasets is subject to a data transfer agreement, and specific rules and guidelines regulate data usage. The fecal samples of Lifelines participants can be requested via Lifelines biobank (<https://www.lifelines.nl/researcher>). All metagenome-assembled genomes generated in this study have been deposited in the European Nucleotide Archive (ENA) under BioProject PRJEB102570 (ERS27207502–ERS27208974).

### Code availability

The code of the HDMI<sup>85</sup> workflow is available via: <https://github.com/HaoranPeng21/HDMI>. Analysis code is available via: <https://github.com/HaoranPeng21/HGT-Project>.

### References

- Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
- Arnold, B. J., Huang, I. T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.* **20**, 206–218 (2022).
- Palmer, J. D. & Foster, K. R. Bacterial species rarely work together. *Science* **376**, 581–582 (2022).
- Dmitrijeva, M. et al. A global survey of prokaryotic genomes reveals the eco-evolutionary pressures driving horizontal gene transfer. *Nat. Ecol. Evol.* **8**, 986–998 (2024).

5. Groussin, M. et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**, 2053–2067.e2018 (2021).
6. Lee, I. P. A., Eldakar, O. T., Gogarten, J. P. & Andam, C. P. Bacterial cooperation through horizontal gene transfer. *Trends Ecol. Evol.* **37**, 223–232 (2022).
7. Fan, Y., Xiao, Y., Momeni, B. & Liu, Y.-Y. Horizontal gene transfer can help maintain the equilibrium of microbial communities. *J. Theor. Biol.* **454**, 53–59 (2018).
8. Wang, T. et al. Horizontal gene transfer enables programmable gene stability in synthetic microbiota. *Nat. Chem. Biol.* **18**, 1245–1252 (2022).
9. Suzuki, T. A. et al. Codiversification of gut microbiota with humans. *Science* **377**, 1328–1332 (2022).
10. Smillie, C. S. et al. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).
11. Hehemann, J. H. et al. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010).
12. Lester, C. H., Frimodt-Møller, N., Sørensen, T. L., Monnet, D. L. & Hammerum, A. M. In vivo transfer of the vanA resistance gene from an *Enterococcus faecium* isolate of animal origin to an *E. faecium* isolate of human origin in the intestines of human volunteers. *Antimicrob. Agents Chemother.* **50**, 596–599 (2006).
13. Brito, I. L. et al. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435–439 (2016).
14. Hsu, TY et al. Profiling lateral gene transfer events in the human microbiome using WAAFL. *Nat. Microbiol.* <https://doi.org/10.1038/s41564-024-01881-w> (2025).
15. Vatanen, T. et al. Mobile genetic elements from the maternal microbiome shape infant gut microbial assembly and metabolism. *Cell* **185**, 4921–4936.e4915 (2022).
16. Peng, H & Fu, J Unveiling horizontal gene transfer in the gut microbiome: bioinformatic strategies and challenges in metagenomics analysis. *Natl. Sci. Rev.* nwaf128 (2025).
17. Chen, L. et al. The long-term genetic stability and individual specificity of the human gut microbiome. *Cell* **184**, 2302–2315.e2312 (2021).
18. Bansal, M. S., Kellis, M., Kordi, M. & Kundu, S. RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics* **34**, 3214–3216 (2018).
19. Hsu, TY et al. Profiling novel lateral gene transfer events in the human microbiome. *bioRxiv* <https://doi.org/10.1101/2023.08.08.552500> (2023).
20. Camargo, AP et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-023-01953-y> (2023).
21. Hutchings, M. I., Truman, A. W. & Wilkinson, B. Antibiotics: past, present and future. *Curr. Opin. Microbiol.* **51**, 72–80 (2019).
22. Levillain, F. et al. Horizontal acquisition of a hypoxia-responsive molybdenum cofactor biosynthesis pathway contributed to *Mycobacterium tuberculosis* pathoadaptation. *PLOS Pathog.* **13**, e1006752 (2017).
23. Hughes, E. R. et al. Microbial respiration and formate oxidation as metabolic signatures of inflammation-associated dysbiosis. *Cell Host Microbe* **21**, 208–219 (2017).
24. Zhu, W. et al. Precision editing of the gut microbiota ameliorates colitis. *Nature* **553**, 208–211 (2018).
25. Watts, S. C., Ritchie, S. C., Inouye, M. & Holt, K. E. FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics* **35**, 1064–1066 (2019).
26. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
27. Tomich, M., Planet, P. J. & Figurski, D. H. The tad locus: postcards from the widespread colonization island. *Nat. Rev. Microbiol.* **5**, 363–375 (2007).
28. Costa, T. R. D. et al. Type IV secretion systems: Advances in structure, function, and activation. *Mol. Microbiol.* **115**, 436–452 (2021).
29. Goeders, N. & Van Melderen, L. Toxin-antitoxin systems as multi-level interaction systems. *Toxins* **6**, 304–324 (2014).
30. Franzosa, E. A. et al. Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci.* **112**, E2930–E2938 (2015).
31. Brito, I. L. Examining horizontal gene transfer in microbial communities. *Nat. Rev. Microbiol.* **19**, 442–453 (2021).
32. Akhtar, A. A. & Turner, D. P. J. The role of bacterial ATP-binding cassette (ABC) transporters in pathogenesis and virulence: Therapeutic and vaccine potential. *Microb. Pathogenesis* **171**, 105734 (2022).
33. Trappe, K., Marschall, T. & Renard, B. Y. Detecting horizontal gene transfer by mapping sequencing reads across species boundaries. *Bioinformatics* **32**, i595–i604 (2016).
34. Li, C., Jiang, Y. & Li, S. LEMON: a method to construct the local strains at horizontal gene transfer sites in gut metagenomics. *BMC Bioinforma.* **20**, 702 (2019).
35. Song, W., Wemheuer, B., Zhang, S., Steensen, K. & Thomas, T. MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome* **7** <https://doi.org/10.1186/s40168-019-0649-y>(2019).
36. Dmitrijeva, M. et al. A global survey of prokaryotic genomes reveals the eco-evolutionary pressures driving horizontal gene transfer. *Nat. Ecol. Evol.* <https://doi.org/10.1038/s41559-024-02357-0> (2024).
37. Stecher, B. et al. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. *Proc. Natl. Acad. Sci.* **109**, 1269–1274 (2012).
38. Coyte, K. Z. et al. Horizontal gene transfer and ecological interactions jointly control microbiome stability. *PLoS Biol.* **20**, e3001847 (2022).
39. Granato, E.T. et al. Horizontal gene transfer can reshape bacterial warfare. *bioRxiv*, 2024 <https://doi.org/10.1101/2024.08.28.610076> (2024).
40. Behling, A. H. et al. Horizontal gene transfer after faecal microbiota transplantation in adolescents with obesity. *Microbiome* **12**, 26 (2024).
41. Orakov, A. et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).
42. Shaw, J. & Yu, Y. unW. Fairy: fast approximate coverage for multi-sample metagenomic binning. *Microbiome* **12**, 151 (2024).
43. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
44. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421–432 (2019).
45. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
46. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
47. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
48. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved

- genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
49. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
  50. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
  51. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
  52. Na, S. I. et al. UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* **56**, 280–285 (2018).
  53. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
  54. Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W. & Wilson, D. J. Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* **14**, 150–162 (2016).
  55. Duchêne, S. et al. Genome-scale rates of evolutionary change in bacteria. *Micro Genom.* **2**, e000094 (2016).
  56. Zhao, S. et al. Adaptive evolution within gut microbiomes of healthy people. *Cell Host Microbe* **25**, 656–667.e658 (2019).
  57. Kim, J., Na, S. I., Kim, D. & Chun, J. UBCG2: Up-to-date bacterial core genes and pipeline for phylogenomic analysis. *J. Microbiol.* **59**, 609–615 (2021).
  58. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma.* **10**, 421 (2009).
  59. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
  60. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS one* **11**, e0163962 (2016).
  61. Song, W., Steensen, K. & Thomas, T. HgtSIM: a simulator for horizontal gene transfer (HGT) in microbial communities. *PeerJ* **5**, e4015 (2017).
  62. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
  63. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
  64. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
  65. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
  66. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
  67. Alcock, B. P. et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, D690–d699 (2023).
  68. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
  69. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
  70. Ma, B., France, M. & Ravel, J. Meta-pangenome: at the crossroad of pangenomics and metagenomics. *Pangenome*, 205 (2020).
  71. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
  72. Weimann, A. et al. From genomes to phenotypes: TraitR, the microbial trait analyzer. *MSystems* **1**, e00101–e00116 (2016).
  73. Liu, B., Zheng, D., Zhou, S., Chen, L. & Yang, J. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* **50**, D912–D917 (2022).
  74. Kintsjes, B. et al. Phylogenetic barriers to horizontal transfer of antimicrobial peptide resistance genes in the human gut microbiota. *Nat. Microbiol.* **4**, 447–458 (2019).
  75. Wu, G. et al. Two competing guilds as a core microbiome signature for health recovery. *bioRxiv*, 2022.2005.2002.490290 (2022).
  76. Csardi, M. G. Package ‘igraph’. Last accessed 3, 2013 (2013).
  77. Pedersen, T.L., Pedersen, M., LazyData, T., Rcpp, I. & RCPP, L. Package ‘ggraph’. Retrieved January 1, 2018 (2017).
  78. Hothorn, T., Hornik, K., van de Wiel, M. A. & Zeileis, A. Implementing a Class of Permutation Tests: The coin Package. *J. Stat. Softw.* **28**, 1–23 (2008).
  79. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
  80. Bates, D.M. (Springer New York, 2010).
  81. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
  82. Xu, S. et al. Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta* **1**, e56 (2022).
  83. Aroney, S. T. N. et al. CoverM: read alignment statistics for metagenomics. *Bioinformatics* **41**, btaf147 (2025).
  84. Hothorn, T. et al. Package ‘lmtest’. Testing linear regression models. <https://cran.r-project.org/web/packages/lmtest/lmtest.pdf>. Accessed 6 (2015).
  85. Peng, H. Longitudinal gut microbiota tracking reveals the dynamics of horizontal gene transfer. *GitHub Repository* <https://doi.org/10.5281/zenodo.17522213> (2025).
  86. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
  87. Seemann, T. barrnap 0.9: rapid ribosomal RNA prediction. *Google Scholar* (2013).
  88. Wickham, H., Chang, W. & Wickham, M. H. Package ‘ggplot2’. *Create elegant data visualisations using the grammar of graphics. Version 2*, 1–189 (2016).
  89. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
  90. Yin, Y. et al. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, W445–W451 (2012).

## Acknowledgements

We thank all the volunteers in the Lifelines cohort (<https://www.lifelines.nl/>) for their participation and the project staff for their help and management. We thank Kate Mc Intyre for English editing. We also thank Hongyu Jin, Johannes Björk, Yue Zhang, and Jiqui Wu for their suggestions that inspired this study. This study is supported by the Netherlands Organization for Scientific Research (NWO) VICI grant VI.C.202.022 (J.F.), NWO VIDI grant 016.178.056 (A.Z.), NWO VENI grant 222.016 (D.W.), European Research Council (ERC) Consolidator grant 101001678 (J.F.), ERC Starting Grant 715772 (A.Z.), and Dutch Heart Foundation grant IN-CONTROL (CVON2018-27 to J.F. and A.Z.). In addition, H.P. is supported by a joint fellowship from the University Medical Center Groningen and the China Scholarship Council with grant number CSC202208060107. J.F. is supported by a 2023 AMMODO Science Award for Biomedical Sciences from Stichting Ammodo and the Netherlands Organ-on-Chip Initiative, an NWO Gravitation project (024.003.001) funded by the Ministry of Education, Culture and Science of the government of the Netherlands. A.Z. is further supported by the NWO Gravitation grant Exposome-NL (024.004.017) and the EU Horizon Europe Program grant INITIALISE (101094099).

## Author contributions

J.F. coordinated and supervised the study. J.F., H.P., S.A.-S. and D.W. conceptualized the study. H.P. performed data analysis. A.J.R.-M. and H.P. conducted the protein-structure-based analysis. A.F.-P. performed plasmid and virus annotation. J.W. helped with the statistical analysis. R.G. performed metagenomic data assembly. J.F. and A.Z. set up the Lifelines-DEEP cohort. H.P. and J.F. drafted the manuscript. All authors reviewed and edited the manuscript.

## Competing interests

A.Z. received a speaker fee from Nestlé. All other authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-66612-z>.

**Correspondence** and requests for materials should be addressed to Jingyuan Fu.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025