

Transmission of human-associated microbiota along family and social networks

Ilana L. Brito^{1,9*}, Thomas Gurry^{2,3,9}, Shijie Zhao^{2,3}, Katherine Huang⁴, Sarah K. Young⁴, Terrence P. Shea⁴, Waisea Naisilisili⁵, Aaron P. Jenkins^{6,7}, Stacy D. Jupiter⁵, Dirk Gevers⁸ and Eric J. Alm^{2,3,4*}

The human microbiome, described as an accessory organ because of the crucial functions it provides, is composed of species that are uniquely found in humans^{1,2}. Yet, surprisingly little is known about the impact of routine interpersonal contacts in shaping microbiome composition. In a relatively 'closed' cohort of 287 people from the Fiji Islands, where common barriers to bacterial transmission are absent, we examine putative bacterial transmission in individuals' gut and oral microbiomes using strain-level data from both core single-nucleotide polymorphisms and flexible genomic regions. We find a weak signal of transmission, defined by the inferred sharing of genotypes, across many organisms that, in aggregate, reveals strong transmission patterns, most notably within households and between spouses. We were unable to determine the directionality of transmission nor whether it was direct. We further find that women harbour strains more closely related to those harboured by their familial and social contacts than men, and that transmission patterns of oral-associated and gut-associated microbiota need not be the same. Using strain-level data alone, we are able to confidently predict a subset of spouses, highlighting the role of shared susceptibilities, behaviours or social interactions that distinguish specific links in the social network.

Host specificity rather than generalist life histories dominate in the colonization of the gut³. Thus, colonization probably depends on direct interpersonal interactions where individuals are exposed to other humans' microbiota. Nevertheless, the extent to which regular, repeated bacterial exposures result in transmission is unknown. Mother-to-child transmission can be detected early in life^{4–6}, but these patterns fade, whereas other factors—environment⁷, behaviours and genetics⁸—impact the strain-level composition of each adult's microbiome^{9,10}. The human microbiome remains remarkably stable in composition over days¹¹ and even years, at the level of strains^{10,12}, raising the question: do we exchange oral and gut commensals with our closest family and friends?

Here, we take advantage of rich familial and social network data obtained as part of the Fiji Community Microbiome Project (FijiCOMP) (Fig. 1a and Supplementary Tables 1 and 2) to explore the role of transmission in human populations with strain-level resolution. Our data consist of shotgun metagenomic sequences from 287 people living in 5 agrarian villages in the Fiji Islands (Supplementary Tables 3 and 4). Paired gut and oral microbiome samples were deeply sequenced to enable molecular

epidemiological analyses. The presence of locally endemic bacterial disease suggests that commensal bacteria may also spread widely within the community. Owing to the relative isolation of these villages and the reliance on local food and water, we hypothesized that, with comprehensive sampling of eligible individuals in each village, we could capture all human sources and sinks of human-associated bacteria, enabling the tracking of strains within this comparatively 'closed' network.

The bacteria present in the FijiCOMP microbiomes are largely distinct from those in existing databases¹³, resulting in poor read alignments to reference genomes (Supplementary Fig. 1). Thus, we binned reads derived from oral or gut microbiomes using latent strain analysis (LSA)¹⁴, and de novo assembled a set of draft genomes (Supplementary Table 5). There were little-to-no detectable differences in species-level sharing than expected by chance across any relationship type in either the gut or oral microbiome samples (Fig. 1b,c and Supplementary Fig. 2), a finding at odds with that of households in Kenya¹⁵, Israelis⁷ or metropolitan Americans⁹, yet one that may reflect the high contact rates between individuals in this cohort.

To achieve strain-level resolution within individuals' microbiomes, we employed two orthogonal approaches, focusing on either polymorphisms in core proteins, or the presence or absence of flexible genomic regions. The former involved aligning sequencing reads to sets of core genes from each of the assemblies (Supplementary Table 5), similar to several established methods^{6,7,16}, adjusted for use within the context of a social network. Specifically, we calculated the Manhattan distances between pairs of individual's putative genotypes, inferred by the dominant single-nucleotide polymorphism (SNP) at each polymorphic position in the alignment. For individuals in the same village, household members or non-nuclear connections, we compared the distances for each genome of all connected pairs and a balanced random subset of unconnected pairs, whereas we simply shuffled the associations of spouses and mother-child pairs. We performed 100 bootstraps of the unconnected pairs or shuffles, each time tallying the number of genomes for which the median Manhattan distance was lower in connected individuals versus unconnected individuals (Fig. 1b,c). We next implemented an alternate strategy, largely based on the previous observation that flexible genomic regions may be highly personalized¹⁷. Coverage of 1-kb windows of contigs over 10 kb were compared across pairs of individuals. Shared genotypes were defined by the complete lack of outlying 1-kb regions present in one individual and absent in the

¹Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA. ²Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. ³Center for Microbiome, Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁴Broad Institute, Cambridge, MA, USA. ⁵Wildlife Conservation Society, Suva, Fiji. ⁶Edith Cowan University, Joondalup, Western Australia, Australia. ⁷School of Public Health, University of Sydney, Sydney, New South Wales, Australia. ⁸Janssen Human Microbiome Institute, Cambridge, MA, USA. ⁹These author contributed equally: Ilana L. Brito, Thomas Gurry. *e-mail: ibrto@cornell.edu; ejalm@mit.edu

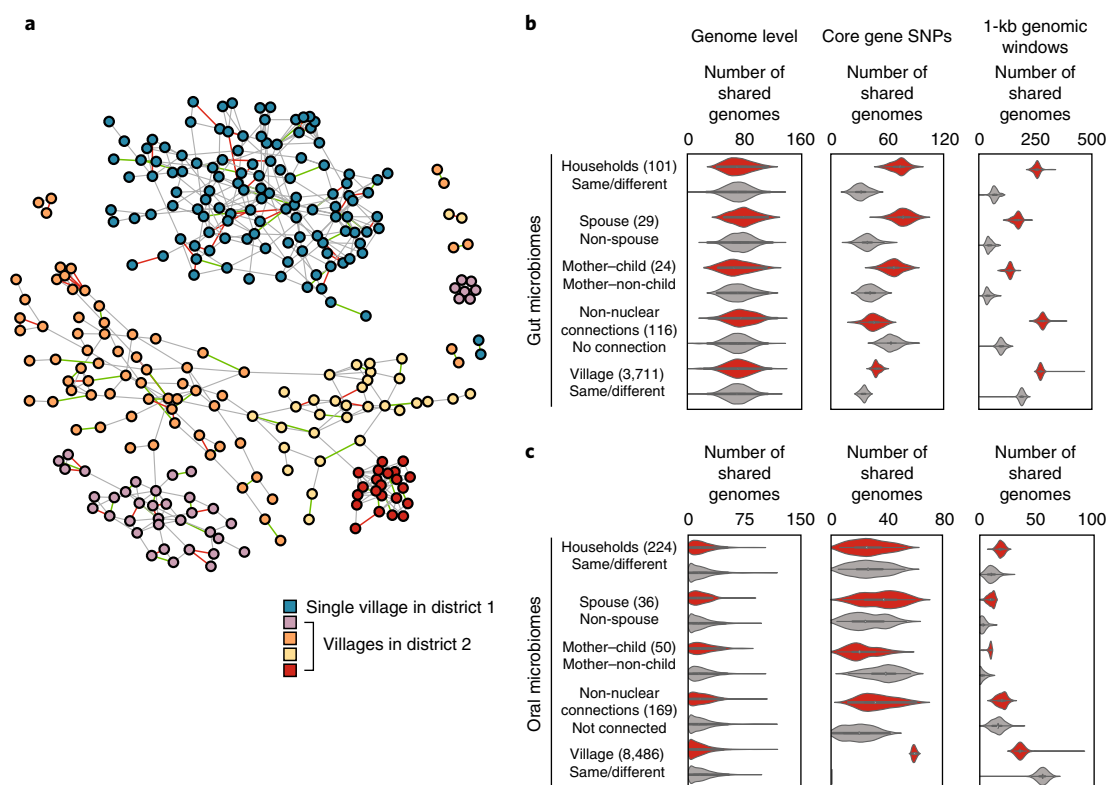


Fig. 1 | Household membership results in shared bacterial lineages. a, The family and social network of the FijiCOMP cohort, coloured by village membership. Four villages are in the same district, whereas the fifth village is in a different district. Spousal relationships are designated by edges coloured red, whereas mother-child relationships are designated by green edges. Grey edges represent all other familial or social network relations. **b,c**, In the gut (**b**) and oral (**c**) microbiome samples, the number of shared genomes, the number of genomes with shared core gene SNP profiles, determined by Manhattan distances, and the number of genomes sharing flexible genomic regions, determined by 1-kb genomic windows, were significantly associated with pairs of linked, rather than unlinked, members of the social network. A ‘genome’ refers to each assembled set of core proteins for each species (left and middle), or to each assembled LSA partition (right). Any connection refers to friendship or distant familial connections in the network, excluding nuclear family and household connections. Full *P* value distributions for the distributions shown are in Supplementary Fig. 4. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (grey) with $n=100$ independent sets of the unlinked pairs obtained by bootstrapping. Whiskers inside the violin extend to points within 1.5 interquartile ranges of the lower and upper quartiles for a distribution, and the centre points represent its median. The numbers of linked pairs for each network (stool or saliva) are as follows: household (101 out of 224); spouse (29 out of 36); mother-child pairs (24 out of 50); any connection (116 out of 169); and village (3,711 out of 8,486).

other (Supplementary Fig. 3). We tallied the number of assembled genomes more frequently shared in each relationship type in over 100 shuffles or bootstraps, again controlling for class imbalances, resulting in the distributions in Fig. 1.

Transmission, loosely defined by shared inferred genotypes, has been observed for strains within the gut microbiomes of mother-child pairs¹⁸, albeit most notably in the first year of life^{4–6}, in cases in which faecal material was used for transplantation^{16,19}, and between twin pairs¹⁰. Within the village setting, we are unable to determine whether strain transfer is direct or indirect, or from a common source, nor can we infer its directionality. However, we refer to the presence of shared genotypes as ‘transmission’ as the putative explanation for the observed patterns. Here, consistent patterns of transmission were revealed across individuals’ social networks in both gut and oral microbiomes, independent of the metric used (Fig. 1b,c and distributions of *P* values in Supplementary Fig. 4). Household members showed high levels of strain similarities in their gut microbiomes, across mother-child pairs and, most notably, among spouses, who share no genetic relatedness. The length of cohabitation was positively correlated, albeit weakly, with the measure of strain dissimilarities (Supplementary Fig. 5), which may reflect long-term changes in intimacy or lifestyle.

The signal varies across our two metrics, potentially highlighting interactions in which organisms versus mobile genetic elements are transmitted between individuals. Using a set of gut microbiome mobile genes previously identified in the FijiCOMP cohort¹³, we find mobile genes weakly shared between spouses (Supplementary Fig. 6). Using strain-level metrics, the transmission signals are robust. Transmission within villages in both gut and oral microbiomes was detectable in core gene SNPs, even when we rarefied the number of village pairs from over 1,000 down to 10 pairs each of connected and unconnected individuals (Supplementary Fig. 7). Furthermore, our results were consistent even when we reduced the number of genomes considered using only those genomes from LSA-informed assemblies with low putative contamination (Supplementary Fig. 8). In all cases, shuffling network relations, while retaining network architecture, ablated observable transmission patterns (Supplementary Fig. 9).

We next examined the contributions of specific organisms, as familial transmission has been previously observed for certain gut and oral commensals^{8,20–22}. There was no consistent signal of transmission across any single phyla (Supplementary Fig. 10). Instead, each pair of connected individuals had a unique signature of shared organisms (Fig. 2a,b and Supplementary Figs. 11–18),

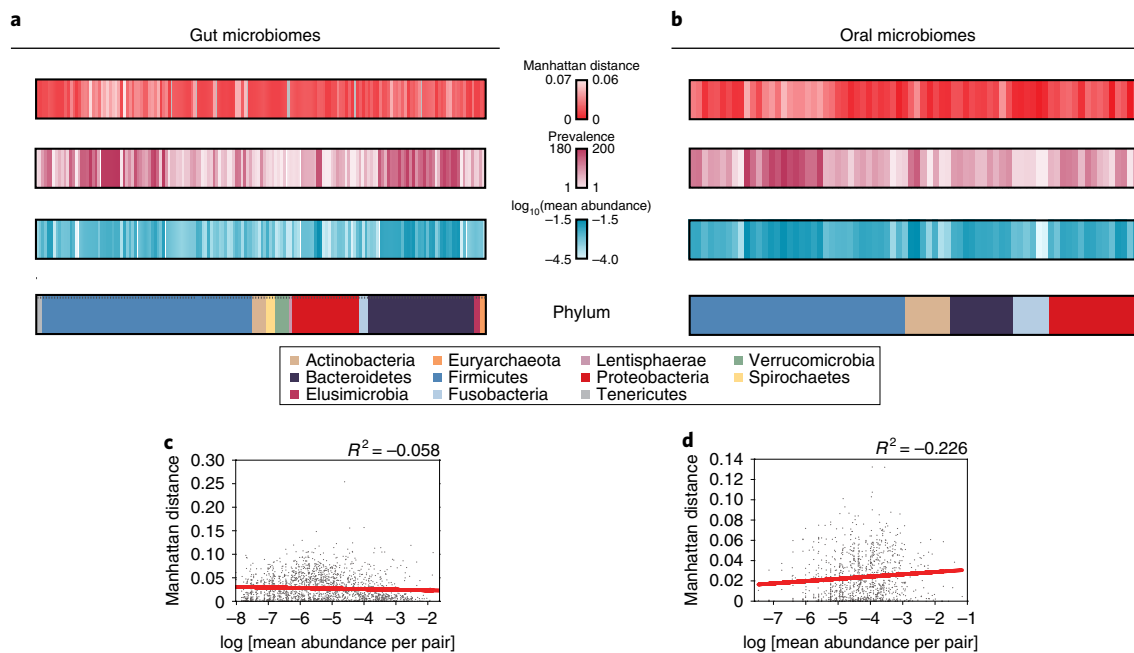


Fig. 2 | Organisms vary in their transmissibility across the social network. a,b, The mean Manhattan distance, prevalence (the number of individuals who harbour that organism), \log_{10} [mean abundance] and phyla are plotted for organisms in the gut ($n=29$) (**a**) and oral ($n=36$) (**b**) microbiomes of spouses. **c,d**, The mean abundance of each organism across each pair of individuals is plotted against the Manhattan distance of that organism for that pair of individuals in the gut ($n=1,988$) (**c**) and oral ($n=1,111$) (**d**) microbiomes. Linear regressions are plotted in red.

suggesting that transmission may be largely driven by chance events and indirect transfer. Interestingly, the fidelity of our LSA-informed assemblies did not strongly affect our results, as transmission may still be observed even if core genes are shuffled between assemblies (Supplementary Fig. 19), supporting the notion that signatures of transmission are distributed broadly over many strains. Microbiome functional profiles also failed to capture transmission signals (Supplementary Fig. 20), although this does not negate the potential contributions of individual virulence-associated or transmission-associated genes contributing to transmissibility. We hypothesized that perhaps the abundance of each organism would be indicative of its overall transmissibility, favouring a mass-action model of transmission, yet this was not the case (Fig. 2c,d).

These findings lead to an apparent paradox: if most bacteria are transmitted directly between members of the community, then why don't we observe clearer patterns of transmission? We believe there are several factors that contribute to the 'diffuse' signal for transmission observed across this population. First, despite this relatively 'closed' network of individuals, there are inherent difficulties in capturing the full range of individuals' contacts and exposures. Our best approximations of direct transfer may be far from actual events, where indirect transfer between individuals outside the network or transmission from unknown and unsampled environmental reservoirs may play a consequential role. Second, we focus on a snapshot in time, not knowing a priori what types of interpersonal interactions result in transfer nor whether transmission occurs during particularly volatile points in an individual's microbiome history. Third, despite our achieved sequencing depth, perhaps longer-read sequencing or a massive increase in sequencing depth is required to achieve greater strain resolution. We reached the limit of detecting transmission when we rarefied samples to 5 million reads (Supplementary Fig. 21). Last, this community may actually be more prone to transmission between a wide range of community members, even when compared to other non-industrialized populations.

This is best illustrated by regular gatherings to drink kava, in which a communal vessel and cup are shared.

Borrowing from the framework of disease ecology, we sought to test the effect of specific individuals within the social network on overall network-level transmission. 'Superspreading' is a phenomenon observed for the transmission of diseases, such as severe acute respiratory syndrome and human immunodeficiency virus, in which the majority of the transmission observed is attributable to a relatively small number of people²³. Across our cohort, there were detectable differences in transmission per individual of both stool and saliva (Fig. 3a–c,e and Supplementary Fig. 22). As we cannot determine the direction of transmission, we refer to this phenomenon as 'supersharing' in this cohort. Supersharing was largely agnostic to the individual's read depth, once a threshold is achieved for obtaining accuracy in Manhattan distances (Supplementary Fig. 23). Interestingly, individuals who were strong supersharers of gut microbiota were not the same as those of oral microbiota (Fig. 3g), revealing differences between the transmission routes of commensals. There was also no specific association with individual's overall sharing and their network positions, either in terms of the number of connections ('degree') or the centrality (measured by 'betweenness') (Supplementary Fig. 24).

Surprisingly, sharing of both gut and oral microbiota was more associated with females in the network ($P < 0.005$ for gut microbiomes; $P < 0.05$ for oral microbiomes, Pearson correlation; Fig. 3d–f and Supplementary Fig. 25), yet had no relationship with age (Supplementary Fig. 26). Although gender-related differences in pathogenic bacterial transmission are well known, as are the myriad factors that affect exposure and susceptibility²⁴, these are less well understood for commensal microbiota with no clear mechanisms of transmission. Nevertheless, exposure risks may be associated with occupations and behaviours that are highly gendered within this cohort (housekeeping: $P < 10^{-15}$; farming and fishing: $P < 10^{-15}$; caring for ill family members: $P < 0.05$; and soap usage: $P < 0.05$, chi-squared tests). It remains to be determined how the transmission

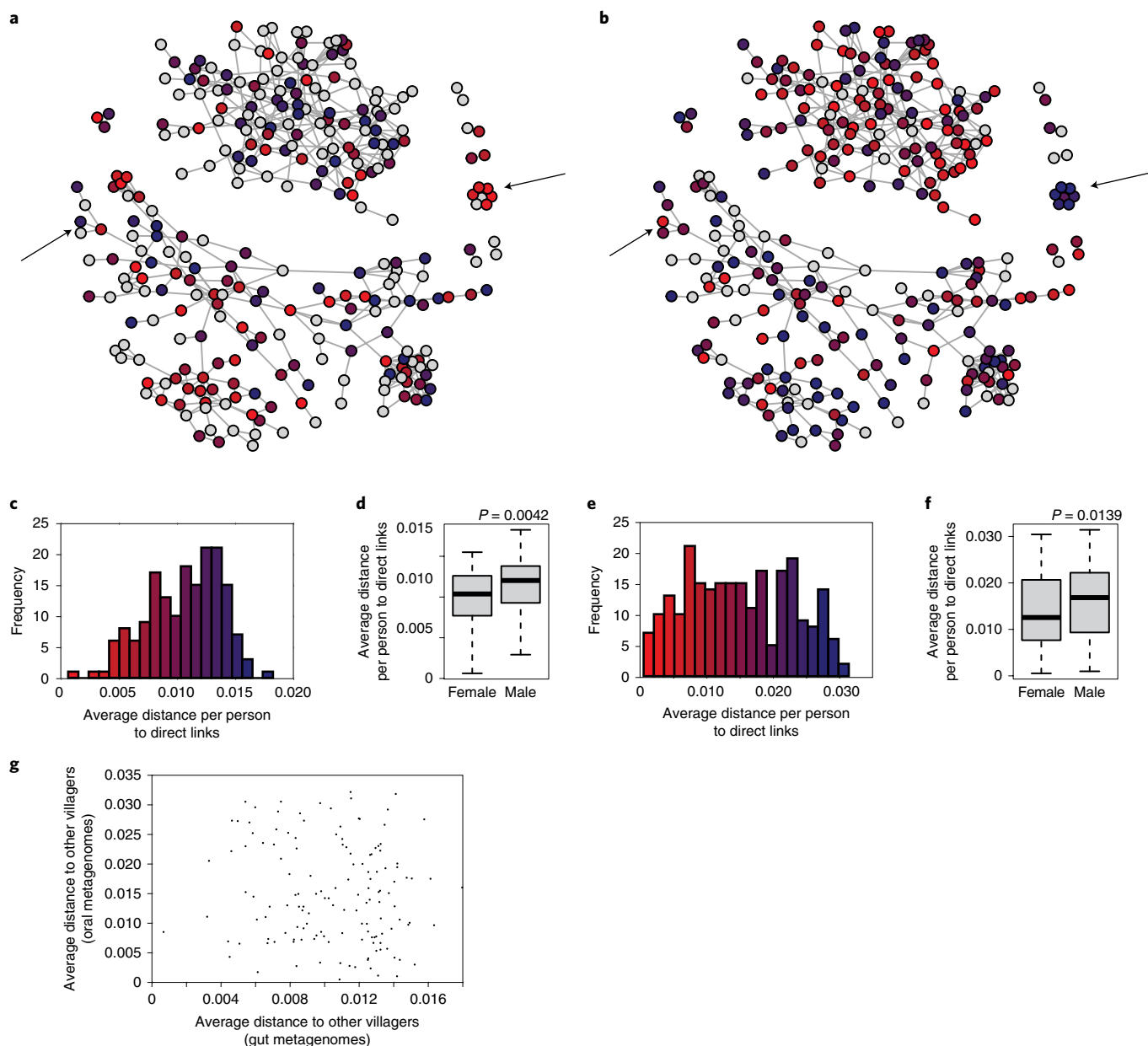


Fig. 3 | Some individuals are 'supersharers'. a,b, For each person in the network, the average distance, defined as the median of the mean Manhattan distances across all genomes to all directly connected individuals, is plotted for organisms within the gut (**a**) and oral (**b**) microbiomes. The arrows point to examples in which the sharing patterns of individuals are different for gut and oral microbiota. The red and blue in plots **a** and **b** match the values plotted in parts **c** and **e**, respectively. **c,e**, Histograms of the average distances for each individual to all of their directly connected individuals is plotted for individuals' gut ($n=173$) (**c**) and oral ($n=243$) (**e**) microbiomes. **d,f**, The distribution of the average distances for each individual to all of their directly connected individuals is plotted for female and male individuals' gut ($n=173$) (**d**) and oral ($n=243$) (**f**) microbiomes. Boxes indicate the upper and lower quartiles, the whiskers extend to the highest and lowest values excluding outliers, and the centre lines indicate the medians. P values were obtained from one-tailed Wilcoxon rank-sum tests. **g**, Each individual's median of the mean Manhattan distances to all individuals within the same village is plotted for their gut and oral microbiomes ($n=142$).

observed in this low-income, agrarian population would compare to a population living in an industrialized nation, where interventions such as the use of antiseptics, disinfectants and antibiotics, sanitation infrastructure and food safety restrictions may influence the transmission of commensal bacteria.

We next asked whether strain-level information alone could be used to predict specific social relationships. We implemented a machine learning approach that utilized organism abundances, core SNP profiles, flexible region similarity or combinations thereof,

without considering demographics. Our household predictions were moderately accurate (area under the curve (AUC) = 0.64 ± 0.02 and 0.61 ± 0.01 for gut and oral microbiomes, respectively), whereas our model to predict spousal relationships performed better (AUC = 0.70 ± 0.03 and 0.72 ± 0.02 for gut and oral microbiomes, respectively) (Fig. 4 and Supplementary Fig. 27). Despite the poorer overall performance of our household models, the predictions seemed to be dependent on the network structure, as all of the relationships within some households were accurately predicted, in

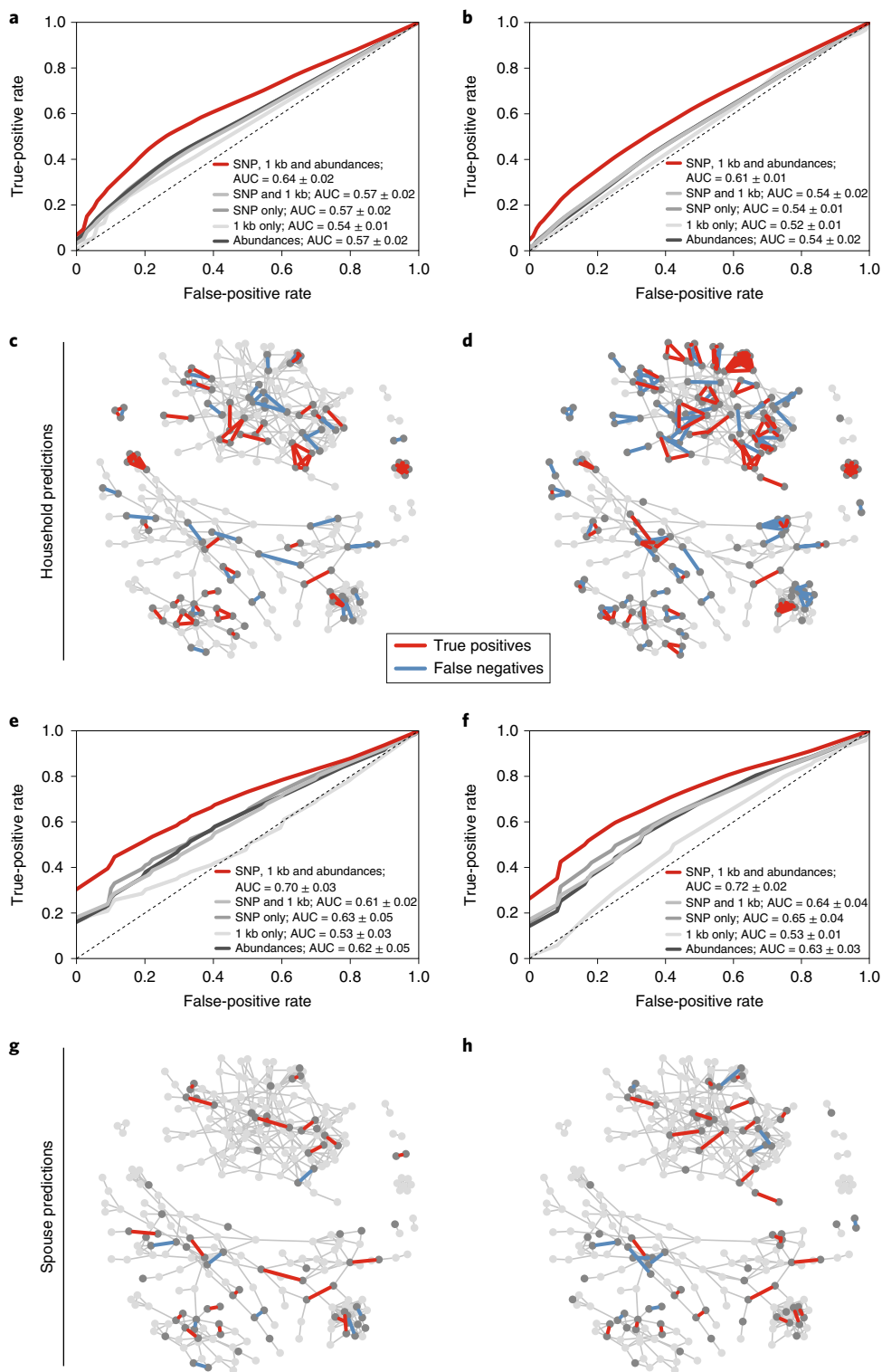


Fig. 4 | Machine learning predicts a subset of spouses with high confidence. **a, b**, ROC curves for a random forest model predicting household membership based on shared gut (**a**) or oral (**b**) microbiome strain-level data are plotted for models using SNP profiles, shared flexible regions, both, or both with organismal abundances. Random forest models were constructed from 1,000 decision trees and without constraint on maximum tree depth. The dotted line shows an ROC in which false positives equal false negatives. The legend reports the means and standard deviations for each classifier's AUC. **c, d**, The social network plotted with predicted true-positive household pairs and false-negative household pairs using gut (**c**) or oral (**d**) microbiome data. **e, f**, ROC curves for a random forest model predicting spousal relationships based on shared gut (**e**) or oral (**f**) microbiome strain-level data are plotted for models using SNP profiles, shared flexible regions, both, or both with organismal abundances. Random forest models were constructed from 1,000 decision trees and without constraint on maximum tree depth. The dotted line shows an ROC in which false positives equal false negatives. The legend reports the means and standard deviations for each classifier's AUC. **g, h**, The social network plotted with predicted true-positive spousal pairs and false-negative spousal pairs using gut (**g**) or oral (**h**) microbiome data.

both gut and oral samples. Remarkably, our model reveals that close to 25% of spouses are exceedingly easy to predict with high confidence (Fig. 4c,d). Within the household network, some of these spousal pairs were obscured, highlighting the subtle nature of these transmission signals. Why certain couples are easier to predict than others is unknown, but may reflect shared susceptibilities, specific behaviours or the relative importance of extra-marital relationships. Interestingly, spouses have been found to share immune repertoires²⁵ and households display family-specific signatures²⁶, providing evidence for shared exposures.

Although it is well established that shared environments significantly affect the gut microbiome composition and phenotype of isogenic mice^{27,28} and that social interactions shape wild primate microbiomes²⁹, this work opens the door to understanding the process of transmission and its implications in human society. Within this small community of individuals with relatively homogeneous living environments, diets and microbiomes, bacterial DNA alone can be used to accurately predict certain intimately linked pairs of individuals. Our research begins to tease apart relevant transmission patterns evident in a social network and a role for gender in commensal transmission, revealing that long-term intimate interactions that occur later in life, such as through marriage or cohabitation, can result in stochastic transmission events in both the gut and the oral microbiomes. Given the wide array of microbiome-associated health conditions, this study further hints at the possibility that diffuse transmission patterns of pathogenic or protective commensals may contribute to the overall health status of individuals.

Methods

Social network construction. The FijiCOMP consisted of interviewing and sampling the gut and oral microbiomes of almost 300 individuals living in 5 village communities in 2 districts approximately 50 miles away from one another on Vanua Levu in the Fiji Islands. The sampling all took place within a 4-week period, each village taking approximately 1–2 weeks. Institutional Review Board approval was received from Institutional Review Boards at Columbia University (New York City, NY, USA), the Massachusetts Institute of Technology (Cambridge, MA, USA) and the Broad Institute (Cambridge, MA, USA), and ethics approvals were received from the Research Ethics Review Committees at the Fiji National University and the Ministry of Health in the Fiji Islands. Informed consent was obtained from all study participants.

As part of the survey, each head of the household was asked to draw their family trees, including all members of their household, even if they are not related. Individuals were specifically asked to name their spouse, if married, and the number and ages of their children. We inferred the number of years a married couple lived together by the age of their oldest child. We excluded 6 of the 63 couples from our analysis of the time they lived together because either they did not have any children or their children's ages were inconsistent (for example, if children came from a previous marriage). As houses commonly have names rather than specific addresses in these villages, individuals were asked the name of the house in which they live. Responses by individuals were cross-referenced for consistency and ambiguous links were removed from our analysis. Minor discrepancies, such as slight differences between spouses in the reporting of their children's ages that differed by 1 year were permitted. Individuals were further asked to provide the names of five individuals with whom they spent the most amount of time. Although the individuals mentioned the type of relationship (for example, mother–child, cousin, sister-in-law, friend, classmate and churchmate, among others), these relationship types were not solely relied on to define a particular relationship type. In a small number of examples, individuals cited social interaction with a third party whose identity could not be verified and were therefore not included in our analysis. In addition, some individuals mentioned siblings or parent–child relationships that could not be verified, so these were also counted as merely social interactions. This resulted in 489 unique social or familial interactions, in addition to household-level interactions. For the purposes of anonymity, the ages of individuals were rounded to the nearest 5-year increment and the number of children per person was not reported. Not all children of each family were surveyed, either because the children did not meet the inclusion criteria (they needed to be at least 8 years of age) or because they were inaccessible during the time when we were sampling. The social network was plotted using R package igraph (v.1.0.1). Network metrics (that is, betweenness and degree) were calculated using igraph standard functions.

Additional information was obtained from all participants, including having individuals name their occupation (of which domestic duties, farmer and fisherman were all possible answers), whether the individual had cared for a sick

family member in the past year and whether they used soap (with possible answers: always, sometimes and never).

Alignments and identification of SNPs. We calculated the Manhattan distances between the dominant SNPs within pairs of individuals' core gene alignments. This involved aligning each individual's reads to core genes in the assembled LSA partitions, extracting polymorphic loci and determining the dominant allele at each locus. For each pair of individuals, we computed the Manhattan distance at each locus, averaged these distances across loci and computed this quantity for every assembled genome. These distances were then used for the network comparisons described in the 'Network comparisons' section.

More precisely, quality-filtered, dereplicated metagenomic data sets (on average, over 52 million and 10 million reads for our gut and oral microbiomes, respectively), devoid of human genetic material (filtered as described in Brito et al.¹³), were partitioned before assembly using LSA¹⁴ according to covarying k-mer content across samples. Read partitions were then assembled using Velvet³⁰. Sets of core genes were identified using AMPHORA2 (ref. ³¹). Core genes were assigned taxonomies using genera-level best hits using BLAST+ against the NCBI nr database. Partitions with complete (31 single-copy genes for bacteria) or near-complete gene sets of AMPHORA genes deriving from the same genera were retained for analysis (Supplementary Tables 3 and 4). If a core gene set contained more than two of the same assembled gene, we removed both copies of that gene.

Each individual's samples were then aligned to the sets of core genes using BWA-MEM³². Reads were subjected to more stringent trimming using TRIMMOMATIC³³ (in addition to trailing low-quality base pairs (quality < 4), we also implemented a sliding window, trimming when the quality was < 15). Reads were then aligned to regions that included one read length (100 bp) downstream and upstream of each core gene to avoid edge effects within the alignments. One-hundred base pairs from each end of the alignment, regardless of whether the gene was positioned at the end of the contig, were then trimmed from the final pile-up. Reads were filtered to retain those with >40% of the length aligning at 90, 95, 97 or 99% identity. A lower cut-off was chosen to capture a wide variety of strains for each alignment. Setting a lower threshold would be more inclusive of strains more distantly related to the reference, which would only obfuscate a signal for a given species should it include too distant strains. Previous work¹⁹ estimated the species boundary at approximately 85–90% identity in core genes (analogous to ~97% in the 16S rRNA gene). Ninety percent identity also resulted in the most consistent coverage across core genes, and it was therefore chosen for all subsequent SNP-level calculations. Reads with soft or hard clipping were removed. To further validate our gene sets, we filtered out genes with abnormal coverage relative to the rest of the gene set. We expected the depth of each gene to be uniform across a genome, and sequencing depth to be Poisson distributed at each locus. To avoid including genes within species' genomes recruiting abnormal numbers of reads compared to the remainder of the genome (and thus more likely to be recruiting reads from other species), we computed a chi-squared goodness-of-fit test for each gene between the empirical coverage distribution and the equivalent Poisson distribution of the same mean. Genes with a median $P < 0.05$ across subjects were discarded from any subsequent analysis. Results were mostly bimodal, where most genes fit the equivalent Poisson distribution very well, giving us confidence that reads were being recruited uniformly across the full length of the considered genes.

To calculate genome-wide statistics (Fig. 1, left), we built a table of the median coverage across the SNP tables within the core genes, across different assemblies. Then, for each pair of people, we counted the number of these genomes that they shared and compared that between related and a balanced set of unlinked pairs.

Polymorphic loci were then identified from the alignment, resulting in a counts matrix for each genome containing read counts for each allele at each locus in each individual. We retained the dominant allele for each individual (the allele with the highest number of read counts) at each site, then computed the Manhattan distance between each individual's dominant allele at each site and averaged these distances across each genome to obtain an average Manhattan distance per SNP for each genome in a given pair of individuals. For each pair of individuals in a given social network (for example, the same household), this average Manhattan distance per SNP was computed for every genome, and the median distance for a given genome compared to the median distance observed in unrelated pairs of individuals. This calculation is described in more detail in the 'Network comparisons' section.

As a comparison, we also ran the quality-filtered forward metagenomic reads through the MetaPhlan2 (ref. ³⁴) pipeline.

Abundance comparisons of 1-kb windows in assembled genomes. Contigs under 10 kb were removed from LSA-assembled draft genomes. Reads were aligned to contigs with 95% identity. Reads with hard and soft clipping were removed, as were supplementary alignments. We only considered pairs in which both individuals had a median coverage of 10 or more across the genome. One-kilobase regions were considered present in an individual and absent in another if its coverage was greater than the median in the first individual, and lower than one-thousandth of the median in the other. Pairs of individuals were considered to share the same strain if there were no such 1-kb regions across the entire genome (that is, all regions were either present or absent in both individuals) and that it was present with a median coverage of 10 or more in both individuals.

Mobile genetic element analysis. For Supplementary Fig. 6, we used the abundances of mobile genes identified in Brito et al.¹³ to determine whether there was a transmission signal. We calculated the Jensen–Shannon divergence between all pairs and compared the number of pairs within each group with a balanced, subsampled group.

Functional contribution to transmission. Genes in the LSA-assembled genomes were first clustered at 90% identity using CD-HIT³⁵. Representative genes were then annotated using DIAMOND³⁶ against the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (release 73.0). Abundances for each gene were then calculated as the median read depth across genes with over 85% coverage. Abundances were summed for each functional gene family (represented by a single KO number). For each pair, the Jensen–Shannon divergence was calculated.

Network comparisons. Network comparisons on the mean pairwise SNP distance were performed by comparing the median value of the mean pairwise distance per SNP in related pairs with those in unrelated pairs for each genome. If a genome's median pairwise distance was lower in related pairs than in unrelated pairs, it was counted as a positive hit for related, and vice versa. The total number of genomes that fell in favour of related and unrelated was then compared. Similar analyses were performed comparing sharing of 1-kb windows in assembled genomes. A genome was assigned a positive hit for related if the number of related pairs sharing the same strain of that genome exceeded the number of unrelated pairs sharing the same strain, and vice versa. To avoid artefacts arising from the fact that the number of unrelated pairs often vastly exceeds the number of related pairs, we downsampled each of the sets of unrelated pairs 100 times, resulting in the *P* value distributions observed in Supplementary Fig. 4.

Networks considered were spousal relationships (spouses versus non-spouse), household relationships (same versus different household), mother–child relationships (mother–child versus non-mother–child), any social network connection (any connection versus no connection) and village (same versus different village). To ensure fair comparisons in the case of spousal relationships, a set of non-spousal pairs was constructed by considering all pairs possible between males of one marriage with females of a different marriage. Similarly, in the case of household relationships, a set of different household pairs was constructed by considering all pairs possible between members of one household and members of another. In addition, comparisons were also made between randomized networks of related and unrelated pairs, in which the identity of the network's nodes was shuffled but the connections preserved, thus preserving the structure of the network.

Social network predictions. For each pair of individuals, we created feature vectors containing the mean pairwise SNP distance for each genome, the relative abundance of that genome in each individual, the number of shared genomes using 1-kb outlier regions and true or false values for whether a given genome was considered to be the same strain in both individuals using the 1-kb outlier regions. These features were then used to train random forest classifiers to predict spousal and household connections, in which class-balanced data sets were constructed by downsampling the number of unrelated pairs to equal the number of related pairs (spouse/non-spouse; same household/different household). To train the random forest classifiers on different data than those used in the predictions, we performed a threefold split of the related pairs and trained on two-thirds of the data while predicting on the remaining one-third. Predictions from the three separate test sets were combined. Receiver operating characteristic (ROC) curves were constructed from the average of ten sets of threefold cross-validation, and *P* values were computed for the resulting AUCs using a Mann–Whitney *U*-statistic on the confusion matrices.

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

Code availability

The code for the analyses in this paper start with an alignment table in the form of a Python dictionary containing individual core genes as its highest-level keys, where for each core gene there is a $M \times N \times 4$ numpy array, for *M* subjects, *N* loci and four different alleles (A, G, C and T). The code for filtering these alignment tables into SNP tables and Manhattan distance calculations, and scripts for identifying non-shared mobile genetic elements from 1-kb regions are posted on GitHub at https://github.com/thomasgurry/fijiComp_transmission.

Data availability

Additional information on the samples can be obtained from www.fijicomp.org. All samples may be downloaded from the NCBI Short Read Archive under Bioproject [PRJNA217052](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA217052). Note that the name for sample SRS475548 in the Short Read Archive was incorrectly entered; it should be the oral microbiome sample from M2.33, not W2.33. All accession numbers are listed in Supplementary Table 1. Sample collection was voluntary; thus, not all of the individuals have oral and gut microbiome samples associated with the surveys.

Received: 4 July 2018; Accepted: 14 February 2019;
Published online: 25 March 2019

References

- Moeller, A. H. et al. Rapid changes in the gut microbiome during human evolution. *Proc. Natl Acad. Sci. USA* **111**, 16431–16435 (2014).
- Davenport, E. R. et al. The human microbiome in evolution. *BMC Biol.* **15**, 127 (2017).
- Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
- Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
- Ferretti, P. et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe* **24**, 133–145 (2018).
- Yassour, M. et al. Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell Host Microbe* **24**, 146–154 (2018).
- Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
- Goodrich, J. K. et al. Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
- Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
- Xie, H. et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* **3**, 572–584 (2016).
- David, L. A. et al. Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15**, R89 (2014).
- Faith, J. J. et al. The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013).
- Brito, I. L. et al. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435–439 (2016).
- Cleary, B. et al. Detection of low-abundance bacterial strains in metagenomic datasets by eigengene partitioning. *Nat. Biotechnol.* **33**, 1053–1060 (2015).
- Mosites, E. et al. Microbiome sharing between children, livestock and household surfaces in western Kenya. *PLoS ONE* **12**, e0171017 (2017).
- Li, S. S. et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352**, 586–589 (2016).
- Franzosa, E. A. et al. Identifying personal microbiomes using metagenomic codes. *Proc. Natl Acad. Sci. USA* **112**, E2930–E2938 (2015).
- Korpela, K. et al. Selective maternal seeding and environment shape the human gut microbiome. *Genome Res.* **28**, 561–568 (2018).
- Smillie, C. S. et al. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* **23**, 229–240 (2018).
- Caugant, D. A., Levin, B. R. & Selander, R. K. Distribution of multilocus genotypes of *Escherichia coli* within and between host families. *J. Hyg. (Lond.)* **92**, 377–384 (1984).
- Preus, H. R., Zambon, J. J., Dunford, R. G. & Genco, R. J. The distribution and transmission of *Actinobacillus actinomycetemcomitans* in families with established adult periodontitis. *J. Periodontol.* **65**, 2–7 (1994).
- Van Winkelhoff, A. J. & Boutaga, K. Transmission of periodontal bacteria and models of infection. *J. Clin. Periodontol.* **32**, 16–27 (2005).
- Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
- Taking Sex and Gender into Account in Emerging Infectious Disease Programmes: an Analytical Framework* (WHO, 2011).
- Carr, E. J. et al. The cellular composition of the human immune system is shaped by age and cohabitation. *Nat. Immunol.* **17**, 461–468 (2016).
- Lax, S. et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* **345**, 1048–1052 (2014).
- Sivan, A. et al. Commensal *Bifidobacterium* promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science* **350**, 1084–1089 (2015).
- Rosshart, S. P. et al. Wild mouse gut microbiota promotes host fitness and improves disease resistance. *Cell* **171**, 1015–1028 (2017).
- Moeller, A. H. et al. Social behavior shapes the chimpanzee pan-microbiome. *Sci. Adv.* **2**, e1500997 (2016).
- Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
- Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).
- Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- Truong, D. T. et al. MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).

35. 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).
36. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

Acknowledgements

This work was supported by funding from the Center for Microbiome Informatics and Therapeutics at the Massachusetts Institute of Technology (MIT). This work was supported by grants from the National Human Genome Research Institute (U54HG003067) to the Broad Institute, the Center for Environmental Health Sciences at MIT, the Center for Microbiome Informatics and Therapeutics at MIT, and the Fijian Ministry of Health. I.L.B. is a Sloan Foundation Research Fellow, a Packard Fellowship in Science and Engineering, and a Pew Foundation Biomedical Scholar.

Author contributions

This study was conceived by I.L.B. and E.J.A. The study was designed by I.L.B., A.P.J., S.P.J. and E.J.A. Raw data were collected by I.L.B. and W.N. Metagenomic assemblies and

metrics were developed and assessed by I.L.B., T.G., S.Z., K.H., S.K.Y., T.P.S., D.G. and E.J.A. Final analyses were performed by I.L.B. and T.G. The paper was written by I.L.B., T.G. and E.J.A.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41564-019-0409-6>.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to I.L.B. or E.J.A.

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

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

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

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

Details on data collection are provided in a previous publication in Brito et al., 2016 Nature.

Data analysis

Versions are provided for all programs used. Code is available at https://github.com/thomasgurry/fijiComp_transmission, which is mentioned in the text.

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

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

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

The Fiji Community Microbiome Project (FijiCOMP) consisted of interviewing and sampling the gut and oral microbiomes of 300 individuals living in 5 village communities in two districts approximately 50 miles away from one another on Vanua Levu in the Fiji Islands. Additional information on the samples can be

obtained from www.fijicomp.org. All samples may be downloaded from NCBI Short Read Archive under Bioproject PRJNA217052. Note that sample SRS475548 in SRA was incorrectly entered. It should be the oral microbiome sample from M2.33, not W2.33. Sample collection was voluntary so not all of the individuals have oral and gut microbiome samples associated with the surveys.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

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

Sample size	We sampled all of the people that were feasible, based on cost and time. The strength of transmission signals was unclear when we embarked on the project. We comprehensively sampled every village (5 villages) going door-to-door to achieve near comprehension of these households. The dataset was on the size order of the NIH-funded Human Microbiome Project.
Data exclusions	In determining our social network, we asked a series of questions. We only included linkages between individuals that could be verified and that had strong support from the metadata. The presence of missing links would only reduce our efforts to see an effect, rather than the opposite. This is described in the methods.
Replication	We repeated our analyses using bootstrapping methods and we report the full range of p-values for these bootstraps. We also use two different metrics to examine transmission.
Randomization	We shuffled our social networks 100-times and compare our data to the shuffled networks. In our random forest models, we perform 10-fold cross-validations and the samples we are predicting on are never included in that specific model.
Blinding	NA

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	NA
Wild animals	NA
Field-collected samples	Samples were frozen shortly after obtaining the sample in liquid nitrogen dry dewers. They were transported in these dewers and transferred to a -80 for longer term storage.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Participants were surveyed and asked about their age, sex, family and friendship interactions, household identification, occupation and involvement in household/agricultural work, usage of soap, and caring for sick individuals.
----------------------------	---

Recruitment

Participants had to be older than 5 and able to fully consent or obtain parental consent.

Individuals were recruited by going house-to-house in each of the communities. The goal was to be as comprehensive as possible so survey administrators went to all households. Soap was offered to participants.