1   **Transmission of human-associated microbiota along family and social networks**

2

3   **Authors:**
4   Ilana L. Brito[1]*, Thomas Gurry[2,3]*, Shijie Zhao[2,3], Katherine Huang[4], Sarah K. Young[4], Terrence P.
5   Shea[4], Waisea Naisilisili[5], Aaron P. Jenkins[6,7], Stacy D. Jupiter[5], Dirk Gevers[8], Eric J. Alm[2,3,4]

6

7   * denotes equal contribution.

8

9

10   **Affiliations**
11   [1] Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY
12   [2] Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA
13   [3] Center for Microbiome, Informatics and Therapeutics, Massachusetts Institute of Technology,
14   Cambridge, MA
15   [4] Broad Institute, Cambridge, MA
16   [5] Wildlife Conservation Society, Suva Office, Fiji Islands
17   [6] Edith Cowan University, Joondalup, Australia
18   [7] School of Public Health, University of Sydney, Sydney, Australia
19   [8] Janssen Human Microbiome Institute, Cambridge, MA
20   [9] Center for Microbiome, Informatics and Therapeutics, Massachusetts Institute of Technology,
21   Cambridge, MA

22

23   **Abstract**

24   The human microbiome, described as an accessory organ because of the crucial functions it
25   provides, is composed of species that are uniquely found in humans[1,2]. Yet, surprisingly little is known
26   about the impact of routine interpersonal contacts in shaping microbiome composition. In a relatively
27   'closed' cohort of 287 people from the Fiji Islands, where common barriers to bacterial transmission are
28   absent, we examine putative bacterial transmission in individuals' gut and oral microbiomes using strain-
29   level data from both core SNPs and flexible genomic regions. We find a weak signal of transmission,
30   defined by the inferred sharing of genotypes, across many organisms that, in aggregate, reveals strong
31   transmission patterns, most notably within households and between spouses. We find that women harbor
32   strains more closely related to those harbored by their familial and social contacts than men; and that
33   transmission patterns of oral- and gut-associated microbiota need not be the same. Using strain-level data
34   alone, we are able to confidently predict a subset of spouses, highlighting the role of shared
35   susceptibilities, behaviors or social interactions that distinguish specific links in the social network.

36   Host-specificity rather than generalist life histories dominate in the colonization of the gut[3]. Thus,
37 colonization likely depends on direct interpersonal interactions where individuals are exposed to other
38 humans' microbiota. Nevertheless, the extent to which regular, repeated bacterial exposures result in
39 transmission is unknown. Mother-to-child transmission can be detected early in life[4,5,6], but these patterns
40 fade, whereas other factors—environment[7], behaviors and genetics[8] impact the strain-level composition of
41 each adult's microbiome[9,10]. The human microbiome remains remarkably stable in composition over
42 days[11] and even years, at the level of strains[10,12], raising the question: do we exchange oral and gut
43 commensals with our closest family and friends?

44   Here, we take advantage of rich familial and social network data obtained as part of the Fiji
45 Community Microbiome Project (FijiCOMP) (Figure 1a, Supplementary Tables 1-2) to explore the role
46 of transmission in human populations with strain-level resolution. Our data consists of shotgun
47 metagenomic sequences from 287 people living in 5 agrarian villages in the Fiji Islands (Supplementary
48 Table 3, Supplementary Table 4). Paired gut and oral microbiome samples were deeply sequenced to
49 enable molecular epidemiological analyses. The presence of locally endemic bacterial disease suggests
50 that commensal bacteria may also spread widely within the community. Due to the relative isolation of
51 these villages and the reliance on local food and water, we hypothesized that with comprehensive
52 sampling of eligible individuals in each village, we could capture all human sources and sinks of human-
53 associated bacteria, enabling the tracking of strains within this comparatively 'closed' network.

54   The bacteria present in the FijiCOMP microbiomes are largely distinct from those in existing
55 databases[13] resulting in poor read alignments to reference genomes (Supplementary Figure 1). Therefore,
56 we binned reads derived from oral or gut microbiomes using Latent Strain Analysis[14], and *de novo*
57 assembled a set of draft genomes (Supplementary Table 5). There were little to no detectable differences
58 in species-level sharing than expected by chance across any relationship type in either the gut or oral
59 microbiome samples (Figure 1b,c, Supplementary Figure 2), a finding at odds with that of households in
60 Kenya[15], Israelis[16] or metropolitan Americans[9], yet one that may reflect the high contact rates between
61 individuals in this cohort.

62   To achieve strain-level resolution within individuals' microbiomes, we employed two orthogonal
63 approaches, focusing on either polymorphisms in core proteins or the presence/absence of flexible
64 genomic regions. The former involved aligning sequencing reads to sets of core genes from each of the
65 assemblies (Supplementary Table 5), similar to several established methods[6,7,17], adjusted for use within
66 the context of a social network. Specifically, we calculated the Manhattan distances between pairs of
67 individual's putative genotypes, inferred by the dominant SNP at each polymorphic position in the
68 alignment. For individuals in the same village, household members or non-nuclear connections, we
69 compared the distances for each genome of all connected pairs and a balanced random subset of
70 unconnected pairs; whereas we simply shuffled the associations of spouses and mother-child pairs. We
71 performed 100 bootstraps of the unconnected pairs or shuffles, each time tallying the number of genomes
72 for which the median Manhattan distance was lower in connected individuals versus unconnected (Figure
73 1b,c). We next implemented an alternate strategy, largely based on the previous observation that flexible
74 genomic regions may be highly personalized[18]. Coverage of one kilobase windows of contigs over 10kb
75 were compared across pairs of individuals. Shared genotypes were defined by the complete lack of
76 outlying 1kb regions present in one individual and absent in the other (Supplementary Figure 3). We
77 tallied the number of assembled genomes more frequently shared in each relationship type in over 100
78 shuffles or bootstraps, again controlling for class imbalances, resulting in the distributions in Figure 1.

79   Transmission, loosely defined by shared inferred genotypes, has been observed for strains within
80 the gut microbiomes of mother-child pairs[19], albeit most notably in the first year of life[4,5,6], in cases where
81 fecal material was used for transplantation[17,20] , and between twin-pairs[10]. Within the village setting, we
82 are unable to determine whether strain transfer is direct or indirect, or from a common source, nor can we
83 infer its directionality. However, we refer to the presence of shared genotypes as 'transmission' as the

84 putative explanation for the observed patterns. Here, consistent patterns of transmission were revealed
85 across individuals' social networks in both gut and oral microbiomes, independent of the metric used
86 (Figure 1b,c, distributions of p-values in Supplementary Figure 4). Household members showed high
87 levels of strain similarities in their gut microbiomes, across mother-child pairs and, most notably, among
88 spouses, who share no genetic relatedness. The length of cohabitation was positively correlated, albeit
89 weakly, with the measure of strain dissimilarities (Supplementary Figure 5), which may reflect long-term
90 changes in intimacy or lifestyle.

91   The signal varies across our two metrics, potentially highlighting interactions in which organisms
92 versus mobile genetic elements are transmitted between individuals. Using a set of gut microbiome
93 mobile genes previously identified in the FijiCOMP cohort[13], we find mobile genes weakly shared
94 between spouses (Supplementary Figure 6). Using strain-level metrics, the transmission signals are
95 robust. Transmission within villages in both gut and oral microbiomes was detectable in core gene SNPs
96 even when we rarefied the number of village pairs from over one-thousand down to 10 pairs each of
97 connected and unconnected individuals (Supplementary Figure 7). Furthermore, our results were
98 consistent even when we reduced the number of genomes considered using only those LSA-informed
99 assemblies with low putative contamination (Supplementary Figure 8). In all cases, shuffling network
100 relations, while retaining network architecture, ablated observable transmission patterns (Supplementary
101 Figure 9).

102   We next examined the contributions of specific organisms, as familial transmission has been
103 previously observed for certain gut and oral commensals[8,21,22,23]. There was no consistent signal of
104 transmission across any single phyla (Supplementary Figure 10). Instead, each pair of connected
105 individuals had a unique signature of shared organisms (Figure 2a,b, Supplementary Figures 11-18) ,
106 suggesting that transmission may be largely driven by chance events and indirect transfer. Interestingly,
107 the fidelity of our LSA-informed assemblies did not strongly impact our results, as transmission may still
108 be observed even if core genes are shuffled between assemblies (Supplementary Figure 19), supporting
109 the notion that signatures of transmission are distributed broadly over many strains. Overall microbiome
110 functional profiles also failed to capture transmission signals (Supplementary Figure 20), though this does
111 not negate the potential contributions of individual virulence- or transmission-associated genes
112 contributing to transmissibility. We hypothesized that perhaps the abundance of each organism would be
113 indicative of its overall transmissibility, favoring a mass-action model of transmission, yet this was not
114 the case (Figure 2c,d).

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

130   Borrowing from the framework of disease ecology, we sought to test the impact of specific
131 individuals within the social network on overall network-level transmission. 'Superspreading' is a

132   phenomenon observed for the transmission of diseases such as severe acute respiratory syndrome (SARS)
133   and human immune-deficiency virus (HIV), where the majority of the transmission observed is
134   attributable to a relatively small number of people[24]. Across our cohort, there were detectable differences
135   in transmission per individual of both stool and saliva (Figure 3a-c,e, Supplementary Figure 22). Since we
136   cannot determine the direction of transmission, we refer to this phenomenon as 'supersharing' in this
137   cohort. Supersharing was largely agnostic to the individuals' read depth, once a threshold is achieved for
138   obtaining accuracy in Manhattan distances (Supplementary Figure 23). Interestingly, individuals who
139   were strong supersharers of gut microbiota were not the same as those of oral microbiota (Figure 3g),
140   revealing differences between the transmission routes of commensals. There was also no specific
141   association with individuals' overall sharing and their network positions, either in terms of the number of
142   connections ('degree') or the centrality (measured by 'betweenness') (Supplementary Figure 24).

143   Surprisingly, sharing of both gut and oral microbiota was more associated with females in the
144   network (p<0.005 for gut microbiomes, p<0.05 for oral microbiomes, Pearson correlation, Figure 3d,f,
145   Supplementary Figure 25), yet had no relationship with age (Supplementary Figure 26). Although gender-
146   related differences in pathogenic bacterial transmission are well known, as are the myriad factors that
147   affect exposure and susceptibility[25], these are less well understood for commensal microbiota with no
148   clear mechanisms of transmission. Nevertheless, exposure risks may be associated with occupations and
149   behaviors that are highly gendered within this cohort (housekeeping, $p<10^{-15}$; farming and fishing, $p<10^{-15}$,
150   caring for ill family members, p<0.05; and soap usage p<0.05, chi-squared test). It remains to be
151   determined how the transmission observed in this low-income, agrarian population would compare to a
152   population living in an industrialized nation, where interventions such as the use of antiseptics,
153   disinfectants and antibiotics, sanitation infrastructure and food safety restrictions, may influence the
154   transmission of commensal bacteria.

155   We next asked whether strain-level information alone could be used to predict specific social
156   relationships. We implemented a machine learning approach that utilized organism abundances, core SNP
157   profiles, flexible region similarity or combinations thereof, without considering demographics. Our
158   household predictions were moderately accurate (AUC = 0.64±0.02 and 0.61±0.01, for gut and oral
159   microbiomes, respectively), whereas our model to predict spousal relationships performed better (AUC =
160   0.70±0.03 and 0.72±0.02, for gut and oral microbiomes, respectively) (Figure 4, Supplementary Figure
161   27). Despite the poorer overall performance of our household models, the predictions appeared dependent
162   on the network structure, as all of the relationships within some households were accurately predicted, in
163   both gut and oral samples. Remarkably, our model reveals that close to 25% of spouses are exceedingly
164   easy to predict with high confidence (Figure 4c,d). Within the household network, some of these spousal
165   pairs were obscured, highlighting the subtle nature of these transmission signals. Why certain couples are
166   easier to predict than others is unknown, but may reflect shared susceptibilities, specific behaviors, or the
167   relative importance of extra-marital relationships. Interestingly, spouses have been found to share immune
168   repertoires[26] and households display family-specific signatures[27], providing evidence for shared
169   exposures.

170   Although it is well-established that shared environments significantly affect the gut microbiome
171   composition and phenotype of isogenic mice[28,29] and that social interactions shape wild primate
172   microbiomes[30], this work opens the door to understanding the process of transmission and its implications
173   in human society. Within this small community of individuals with relatively homogeneous living
174   environments, diets and microbiomes, bacterial DNA alone can be used to accurately predict certain
175   intimately linked pairs of individuals. Our research begins to tease apart relevant transmission patterns
176   evident in a social network and a role for gender in commensal transmission, revealing that long-term
177   intimate interactions that occur later in life, such as through marriage or co-habitation, can result in
178   stochastic transmission events in both the gut and oral microbiomes. Given the wide array of
179   microbiome-associated health conditions, this study further hints at the possibility that diffuse

180   transmission patterns of pathogenic or protective commensals may contribute to individuals' overall
181   health status.

182   **Methods**

183

184   **Social network construction**

185   The Fiji Community Microbiome Project (FijiCOMP) consisted of interviewing and sampling the gut and
186   oral microbiomes of almost 300 individuals living in 5 village communities in two districts approximately
187   50 miles away from one another on Vanua Levu in the Fiji Islands. The sampling all took place within a
188   4-week period, each village taking approximately 1-2 weeks. IRB approval was received from
189   Institutional Review Boards at Columbia University, the Massachusetts Institute of Technology and the
190   Broad Institute and ethics approvals were received from the Research Ethics Review Committees at the
191   Fiji National University and the Ministry of Health in the Fiji Islands. Informed consent was obtained
192   from all study participants.

193

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

217

218   Additional information was obtained from all participants including having individuals name their
219   occupation (of which domestic duties, farmer, and fisherman were all possible answers), whether the
220   individual had cared for a sick family member in the past year, and whether they used soap (with possible
221   answers: always, sometimes and never).

222

223   **Alignments and identification of single nucleotide polymorphisms**

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

230   More precisely, quality-filtered, dereplicated metagenomic datasets (on average, over 52 million and 10
231   million reads for our gut and oral microbiomes, respectively), devoid of human genetic material (filtered

232 as described in Brito et al., 2016[13]), were partitioned prior to assembly using Latent Strain Analysis[14]
233 according to covarying kmer content across samples. Read partitions were then assembled using Velvet[31].
234 Sets of core genes were identified using AMPHORA2[32]. Core genes were assigned taxonomies using
235 genera-level best hits using BLAST+ against the NCBI NR database. Partitions with complete (31 single-
236 copy genes for bacteria) or near-complete gene sets of AMPHORA genes deriving from the same genera
237 were retained for analysis (Supplementary Tables). If a core gene set contained more than 2 of the same
238 assembled gene, we removed both copies of that gene.

240 Each individuals' samples were then aligned to the sets of core genes using BWA-MEM[33]. Reads were
241 subjected to more stringent trimming using TRIMMOMATIC[34] (in addition to trailing low quality bp
242 (quality < 4), we also implemented a sliding window, trimming when the quality < 15). Reads were then
243 aligned to regions that included 1 read-length (100bp) downstream and upstream of each core gene to
244 avoid edge effects within the alignments. 100bp from each end of the alignment, regardless of whether the
245 gene was positioned at the end of the contig, was then trimmed from the final pile-up. Reads were filtered
246 to retain those with greater than 40% of the length aligning at 90, 95, 97 or 99 percent identity. A lower
247 cut-off was chosen to capture a wide variety of strains for each alignment. Setting a lower threshold
248 would be more inclusive of strains more distantly related to the reference, which would only obfuscate a
249 signal for a given species should it include too distant strains. Previous work[20] estimated the species
250 boundary at approximately 85-90% identity in core genes (analogous to ~97% in the 16S rRNA gene).
251 90% identity also resulted in the most consistent coverage across core genes, and it was therefore chosen
252 for all subsequent SNP-level calculations. Reads with soft- or hard-clipping were removed. To further
253 validate our gene sets, we filtered out genes with abnormal coverage relative to the rest of the gene set.
254 We expected the depth of each gene to be uniform across a genome, and sequencing depth to be Poisson
255 distributed at each locus. To avoid including genes within a species' genomes that recruiting abnormal
256 numbers of reads compared to the remainder of the genome (and thus more likely to be recruiting reads
257 from other species), we computed a chi-squared goodness of fit test for each gene between the empirical
258 coverage distribution and the equivalent Poisson distribution of the same mean. Genes with median p-
259 value lower than 0.05 across subjects were discarded from any subsequent analysis. Results were mostly
260 bimodal, where most genes fit the equivalent Poisson distribution very well, giving us confidence that
261 reads were being recruited uniformly across the full length of the considered genes.

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

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

275 As a comparison, we also ran the quality-filtered forward metagenomic reads through the MetaPhlan2[36]
276 pipeline.

Abundance comparisons of 1kb windows in assembled genomes
278 Contigs under 10kb were removed from LSA-assembled draft genomes. Reads were aligned to contigs
279 with 95% identity. Reads with hard and soft clipping were removed, as were Supplementary alignments.

280 We only considered pairs where both individuals had a median coverage of 10 or more across the
281 genome. 1kB regions were considered present in an individual and absent in another if its coverage was
282 greater than the median in the first individual; and lower than one thousandth of the median in the other.
283 Pairs of individuals were considered to share the same strain if there were no such 1kB regions across the
284 entire genome (i.e. all regions were either present or absent in both individuals) and that it was present
285 with a median coverage of 10 or more in both individuals.

Mobile genetic element analysis
287 For Supplementary Figure 6, we used the abundances of mobile genes identified in Brito *et al.* 2016 to
288 determine whether there was a transmission signal. We calculated the Jensen-Shannon divergence
289 between all pairs and compared the number of pairs within each group with a balanced, subsampled
290 group.

Functional contribution to transmission
292 Genes in the LSA-assembled genomes were first clustered at 90% identity using CD-HIT[37].
293 Representative genes were then annotated using DIAMOND[38] against the Kyoto Encyclopedia of Genes
294 and Genomes (KEGG) database (release 73.0). Abundances for each gene were then calculated as the
295 median read depth across genes with over 85% coverage. Abundances were summed for each functional
296 gene family (represented by a single KO number). For each pair, the Jensen Shannon divergence was
297 calculated.

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

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

Social network predictions
321 For each pair of individuals, we created feature vectors containing the mean pair-wise SNP distance for
322 each genome, the relative abundance of that genome in each individual, the number of shared genomes
323 using 1kB outlier regions, and True/False values for whether a given genome was considered to be the
324 same strain in both individuals using the 1kB outlier regions. These features were then used to train
325 Random Forest Classifiers (RFCs) to predict spousal and household connections, where class-balanced
326 datasets were constructed by downsampling the number of unrelated pairs to equal the number of related

327   pairs (spouse/non-spouse; same household/different household). In order to train the RFCs on different
328   data than those used in the predictions, we performed a 3-fold split of the related pairs and trained on two
329   thirds of the data while predicting on the remaining third. Predictions from the three separate test sets
330   were combined. ROC curves were constructed from the average of ten sets of 3-fold cross-validation, and
331   p-values computed for the resulting AUCs using a Mann-Whitney U statistic on the confusion matrices.

332   **Data availability**
333   Additional information on the samples can be obtained from www.fijicomp.org. All samples may be
334   downloaded from NCBI Short Read Archive under Bioproject PRJNA217052. Note that the name for
335   sample SRS475548 in SRA was incorrectly entered. It should be the oral microbiome sample from
336   M2.33, not W2.33. All accession numbers are listed in Supplementary Table 1. Sample collection was
337   voluntary, therefore not all of the individuals have oral and gut microbiome samples associated with the
338   surveys.
339
340   Code for the analyses in this paper start with an alignment table in the form of a Python dictionary
341   containing individual core genes as its highest level keys, where for each core gene there is a M x N x 4
342   numpy array, for M subjects, N loci, and four different alleles (A, G, C and T). Code for filtering these
343   alignment tables into SNP tables, Manhattan distance calculations, and scripts for identifying non-shared
344   mobile genetic elements from 1kb regions are posted on Github at
345   https://github.com/thomasgurry/fijiComp_transmission.

346   **Correspondences**
347   All correspondences should be addressed to Eric Alm (ejalm@mit.edu).

355   **Author contributions**
356   This study was conceived of by I.L.B. and E.J.A. The study was designed by I.L.B., A.P.J., S.P.J., and
357   E.J.A. Raw data was collected by I.L.B. and W.N. Metagenomic assemblies and metrics were developed
358   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
359   I.L.B. and T.G. The paper was written by I.L.B., T.G., and E.J.A.
360
361   The authors have no conflicts of interest to report.

References

1.   Moeller, A. H. *et al.* Rapid changes in the gut microbiome during human evolution. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 16431–16435 (2014).
2.   Davenport, E. R. *et al.* The human microbiome in evolution. *BMC Biol.* **15,** 127 (2017).
3.   Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551,** 457–463 (2017).
4.   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).

5.   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)

6.   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).

7.   Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555,** 210–215 (2018).

8.   Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159,** 789–799 (2014).

9.   Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).

10.   Xie, H. *et al.* Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell Syst* **3,** 572-584.e3 (2016).

11.   David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15,** R89 (2014).

12.   Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science* **341,** 1237439 (2013).

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.   Cleary, B. *et al*. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat Biotechnol*. **33**, 1053-60. (2015).

15.   Mosites, E. Microbiome sharing between children, livestock and household surfaces in western Kenya. *PLoS One*. **12**(2):e0171017. (2017).

16.   Rothschild, D., *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature.* **555**, 210-215 (2018).

17.   Li, S. S. *et al.* Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352,** 586–589 (2016).

18.   Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci. U.S.A.* **112,** E2930-2938 (2015).

19.   Korpela, K. *et al*. Selective maternal seeding and environment shape the human gut microbiome. *Genome Res*. **28**, 561-568 (2018).

20.   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.e5 (2018).

21.   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).

22.   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).

23.   Van Winkelhoff, A. J. & Boutaga, K. Transmission of periodontal bacteria and models of infection. *J. Clin. Periodontol.* **32 Suppl 6,** 16–27 (2005).

24.   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).

25.   World Health Organization. Taking sex and gender into account in emerging infectious disease programmes: an analytical framework. (Manila : WHO Regional Office for the Western Pacific, Geneva Switzerland, 2011).

26.   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).

27.   Lax, S. *et al*., Longitudinal analysis of microbial interaction between humans and the indoor environment . *Science*. **345**, 1048-1052 (2014).

28.   Sivan, A. *et al.* Commensal Bifidobacterium promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science* **350,** 1084–1089 (2015).

29.   Rosshart, S. P. *et al.* Wild Mouse Gut Microbiota Promotes Host Fitness and Improves Disease Resistance. *Cell* **171,** 1015-1028.e13 (2017).

30.   Moeller, A.H. *et al.*, Social behavior shapes the chimpanzee pan-microbiome. *Science Advances.* 2:e1500997.

Figures

Figure 1. Household membership results in shared bacterial lineages.

(a) The family and social network of the FijiCOMP cohort, colored 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 colored red, whereas mother-child relationships are designated by green edges. Gray 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 (left), the number of genomes with shared core gene SNP profiles, determined by Manhattan distances (middle), and the number of genomes sharing flexible genomic regions, determined by 1kb genomic windows (right) significantly associated with pairs of linked rather than unlinked members of the social network. A 'genome' refers to each assembled sets 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 Figure 4. The violin plot distributions represent results from comparing the linked pairs in a given social network (red) or the shuffled network (gray) with $N=100$ independent sets of the unlinked pairs obtained by bootstrapping. Whiskers inside the violin extend to points within 1.5 interquartile ranges (IQRs) of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (stool/saliva) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Figure 2. Organisms vary in their transmissibility across the social network.

(a,b) The mean Manhattan distance, prevalence (number of individuals who harbor that organism), $\log_{10}$(mean abundance) and phyla are plotted for organisms in the (a) gut ($N=1,988$) and (b) oral microbiomes ($N=1,111$) 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 (c) gut and (d) oral microbiomes. Linear regressions are plotted in red.

Figure 3. Some individuals are 'supersharers'.

(a,b) For each person in the network, the average distance, defined as the median of mean Manhattan distances across all genomes to all directly connected individuals, is plotted for organisms within the (a) gut and (b) oral microbiomes. Arrows point out examples in which individuals' sharing patterns 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) The distribution of average distances (median of mean Manhattan distances) for each individual to all of their directly connected individuals is plotted for female and male individuals' (c) gut (*N*=173) and (e) oral microbiomes (*N*=243).

(d,f) A histogram of the average distances (median of mean Manhattan distances) for each individual to all of their directly connected individuals is plotted for individuals' (d) gut (*N*=173) and (f) (*N*=243) oral microbiomes. Boxes indicate the upper and lower quartiles, whiskers extend to highest and lowest values excluding outliers, and center lines indicate medians. P-values were obtained from one-tailed Wilcoxon rank sum tests.

(g) Each individual's median of mean Manhattan distances to all individuals within the same village is plotted for their gut and oral microbiomes (*N*=142).

Figure 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 (a) gut or (b) oral 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 where false positives equal false negatives. The legend reports means and standard deviations for each classifier's Area Under the Curve (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. Arrows point to examples of either families in which everyone in a household can be confidently predicted.

(e,f) ROC curves for a random forest model predicting household membership based on shared (e) gut or (f) oral 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 legend reports means and standard deviations for each classifier's AUC.

(g,h) The social network plotted with predicted true positive household pairs and false negative household pairs using gut (g) or oral (h) microbiome data. Arrows point to examples of either families in which everyone in a household can be confidently predicted.
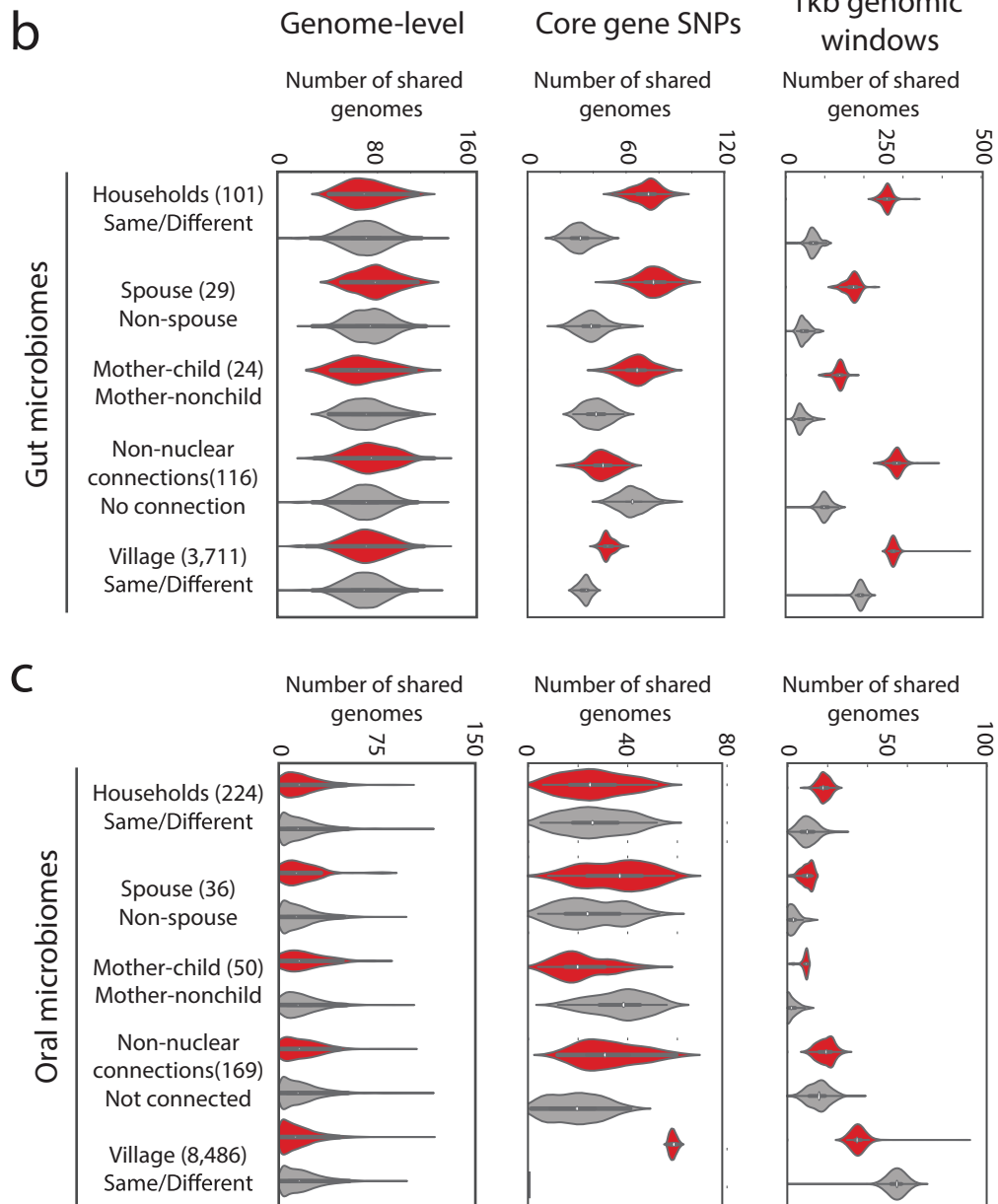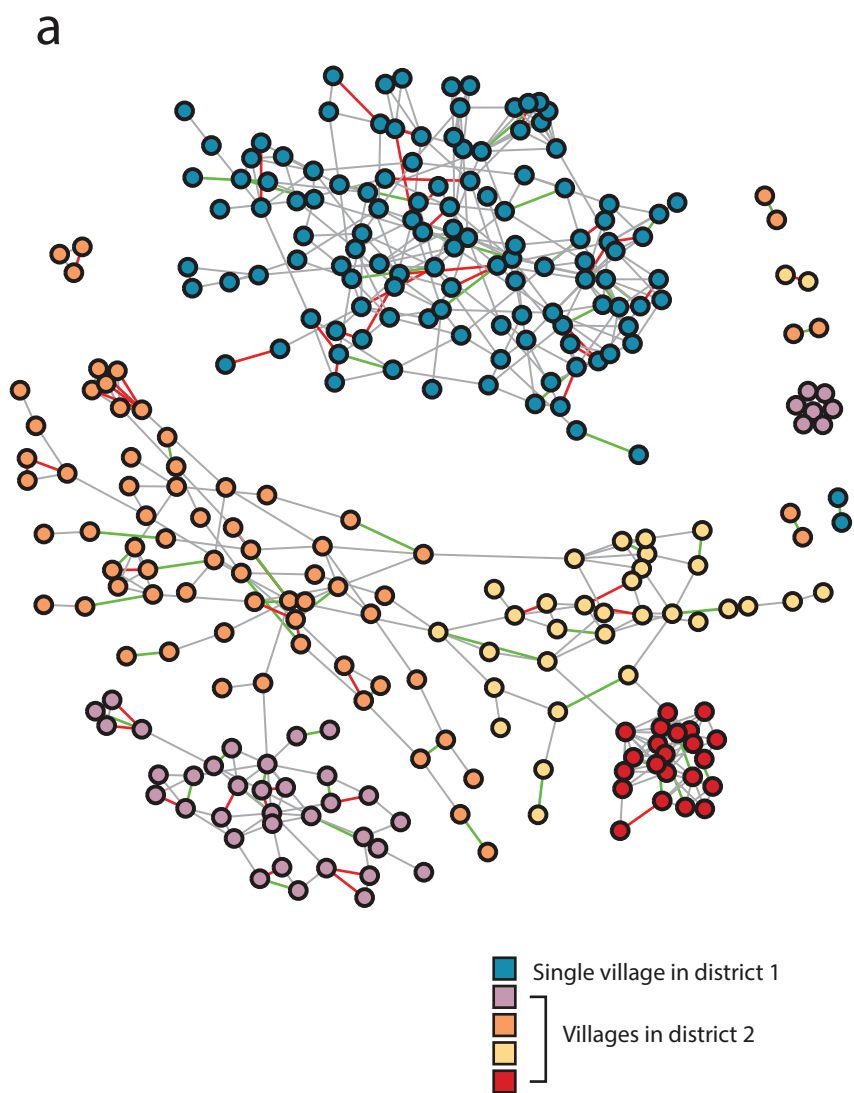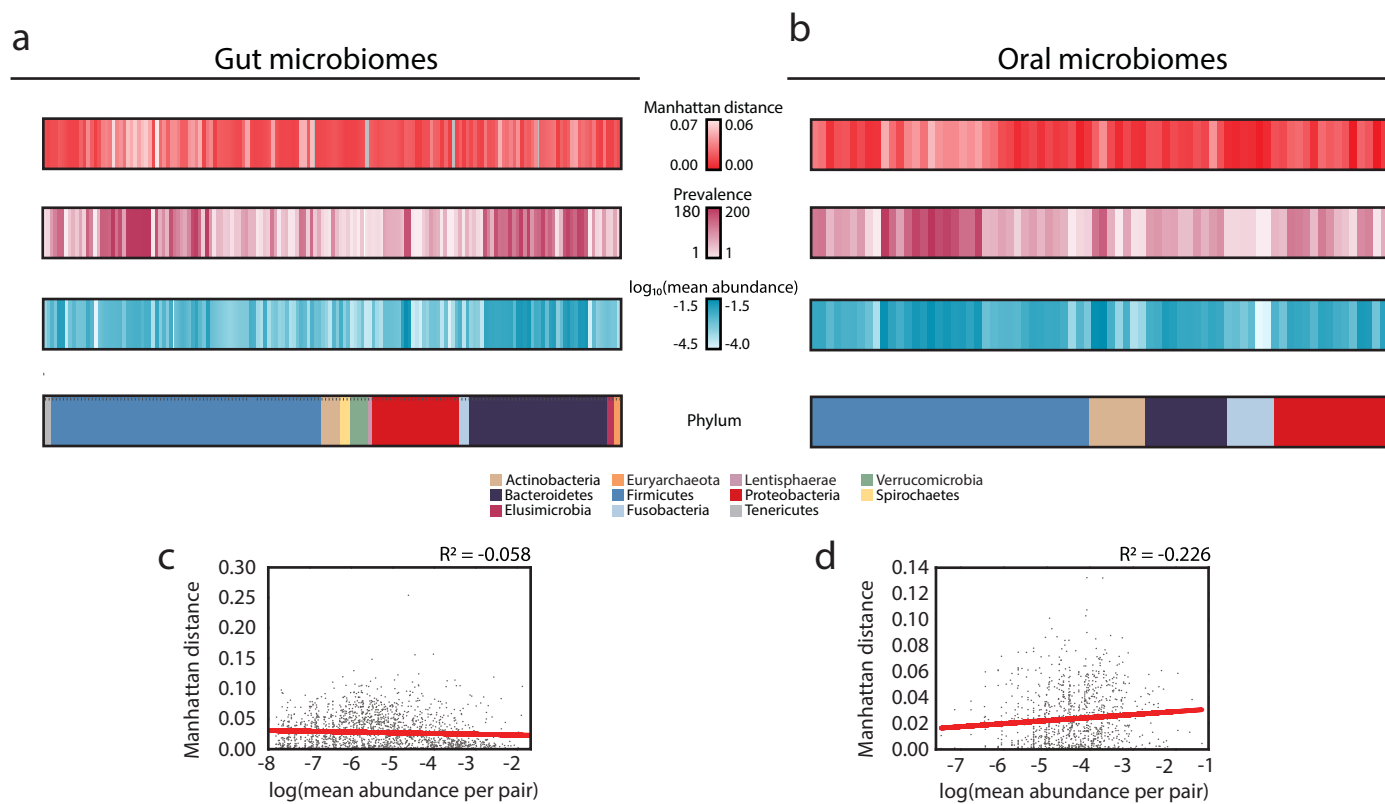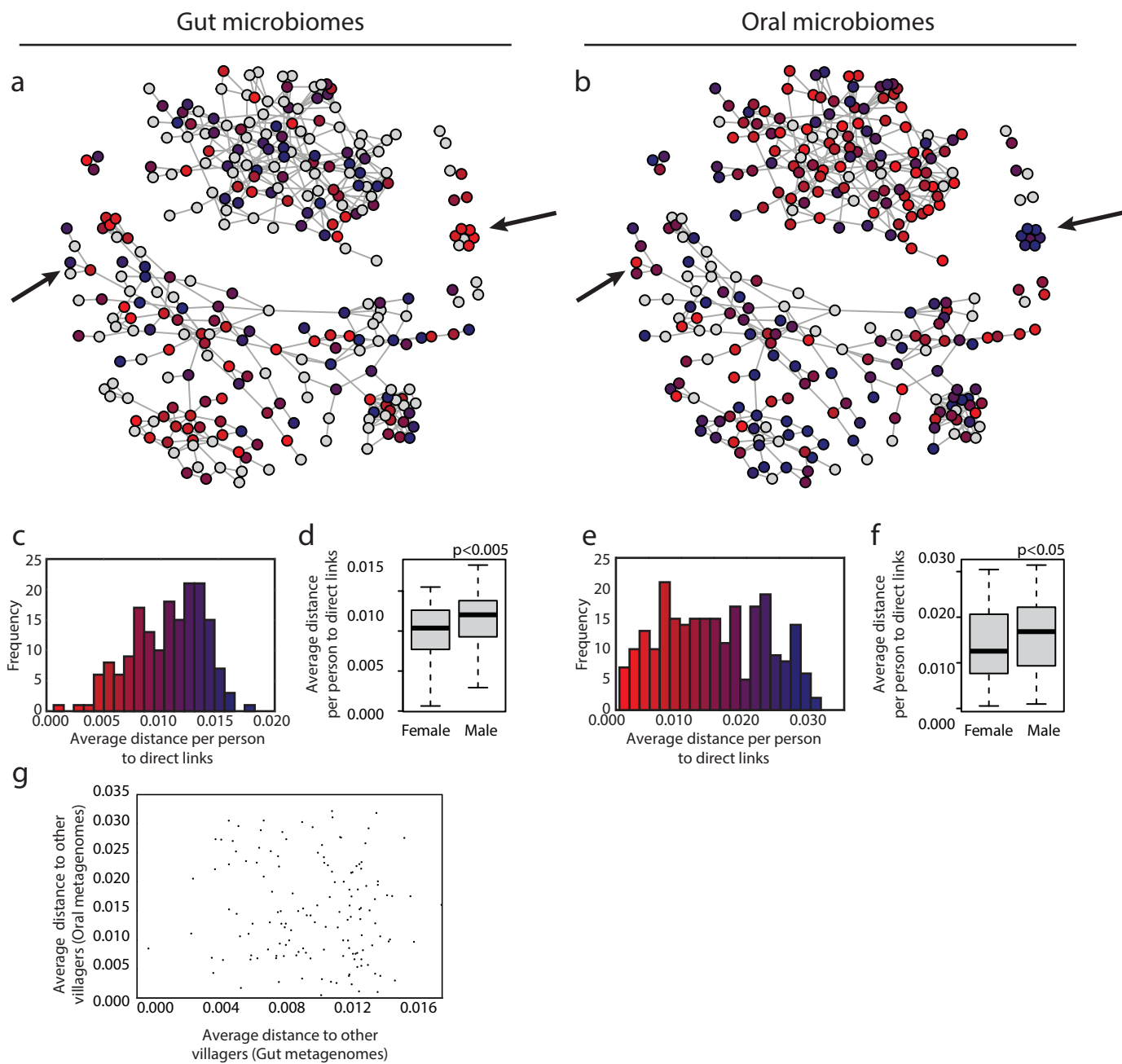
Figure 1

Figure 2

Figure 3

Figure 4