Transmission of human-associated microbiota along family and social networks

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

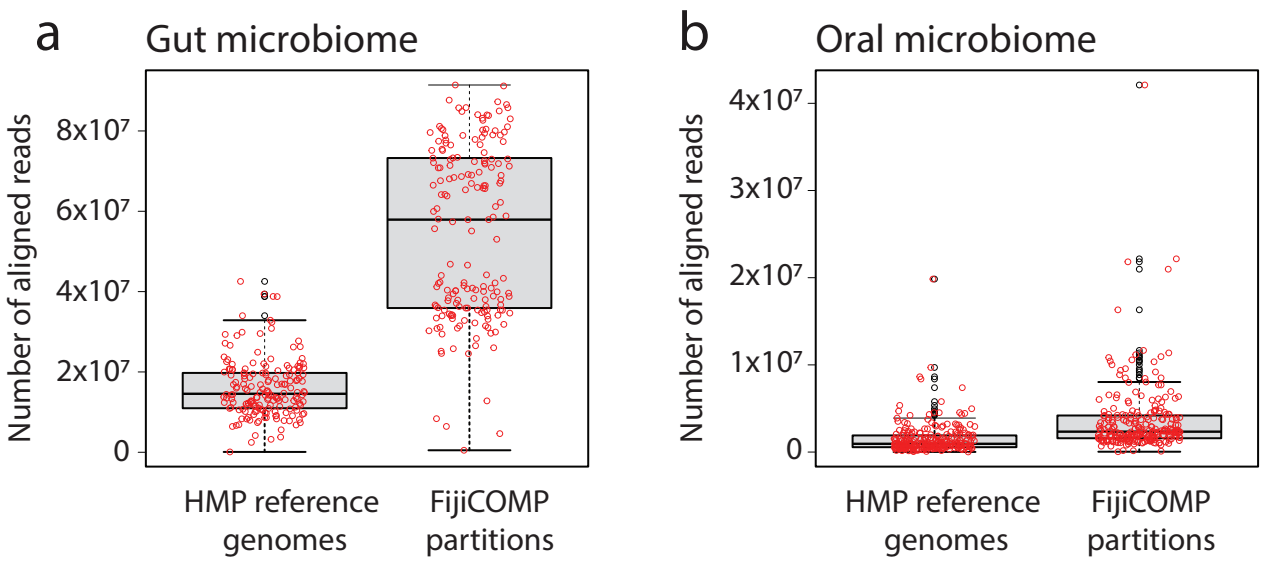
* denotes equal contribution.

Supplementary Information

Supplementary Figures 1-27

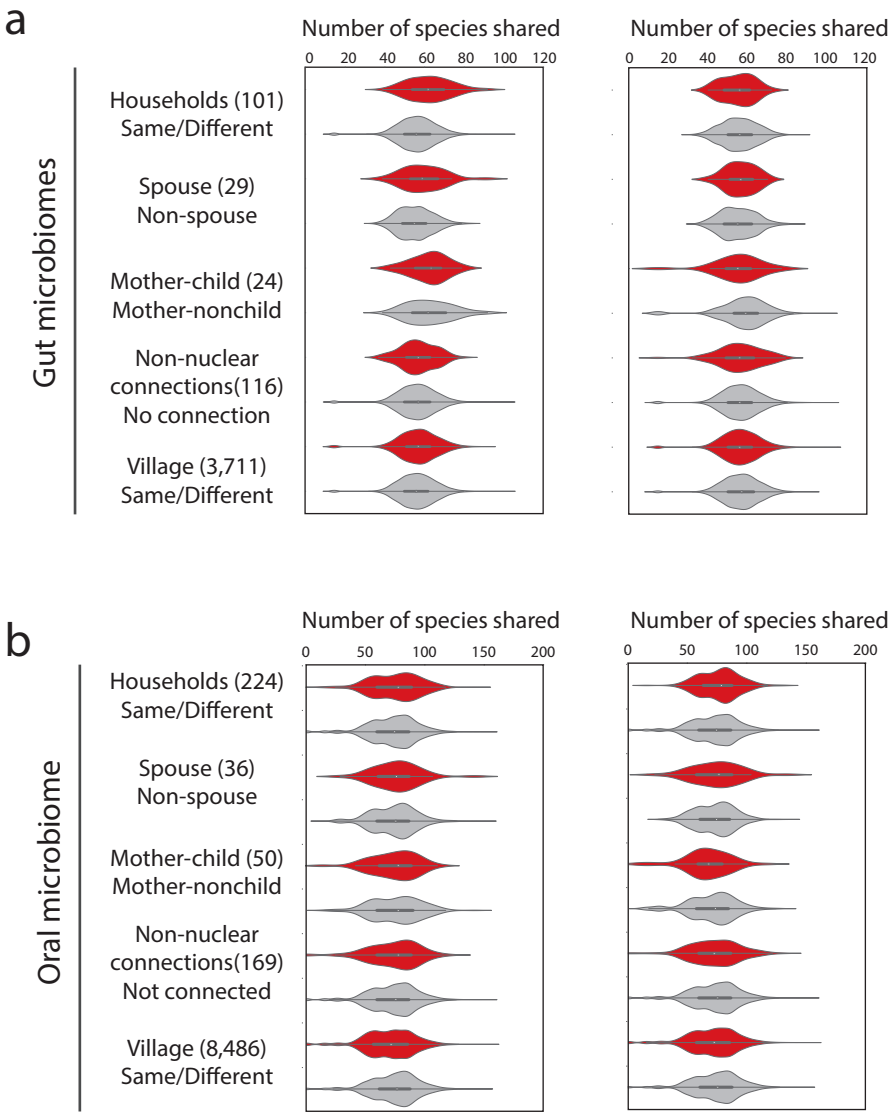
Supplementary Tables

Supplementary Figure 1



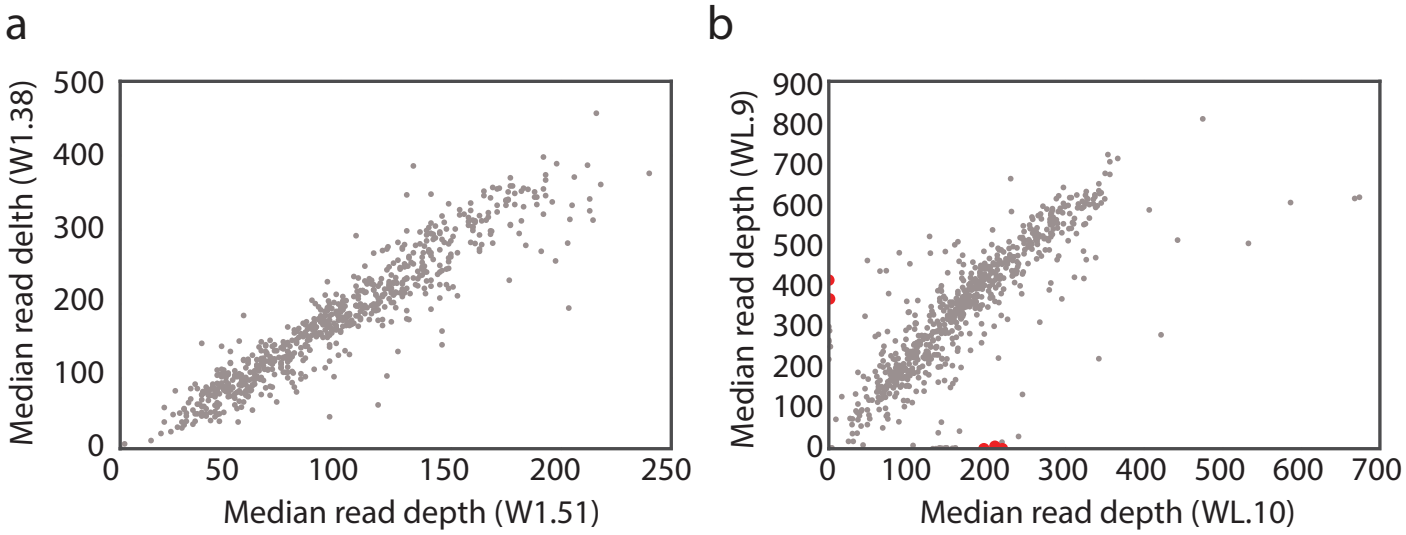
Supplementary Figure 1. Alignments of FijiCOMP reads to the HMP reference genomes and LSA-binned FijiCOMP assemblies. The total number of primary read alignments of (a) gut (*N*=176) and (b) oral FijiCOMP metagenomes (*N*=244) (each represented by a red circle) to either the 2,191 reference genomes that make up the Human Microbiome Project (HMP) reference genome collection (downloaded from https://www.hmpdacc.org on August 3, 2018) or the complete collection of LSA-binned FijiCOMP assemblies. Reads were filtered at 95% identity. Boxplots are drawn with lines at the median, whiskers that show quartile values, and black circles representing those samples that fall outside the quartile values.

Supplementary Figure 2



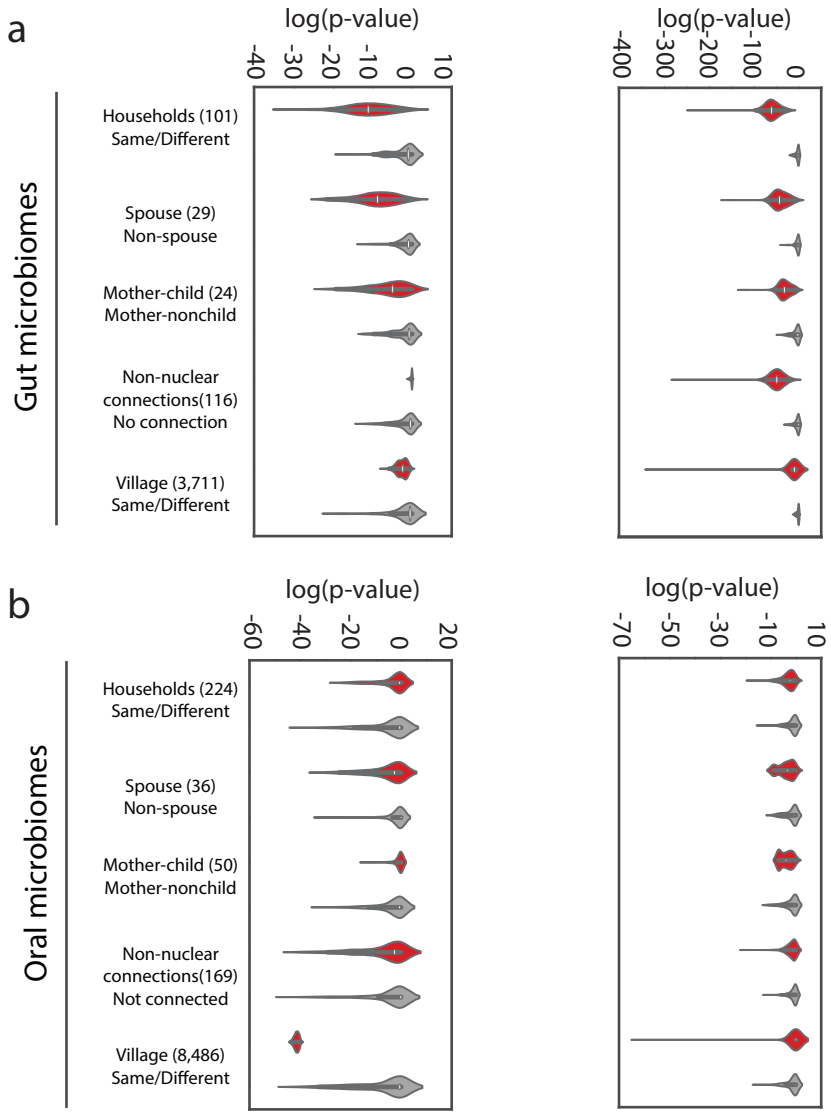
Supplementary Figure 2. Transmission of species using MetaPhlan2-annotated metagenomes. The number of MetaPhlan2-annotated species shared between related pairs and 100 bootstraps of unrelated pairs are shown for the (a) gut and (b) oral microbiomes. The violin plot distributions represent results from 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 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486). Red violin

Supplementary Figure 3



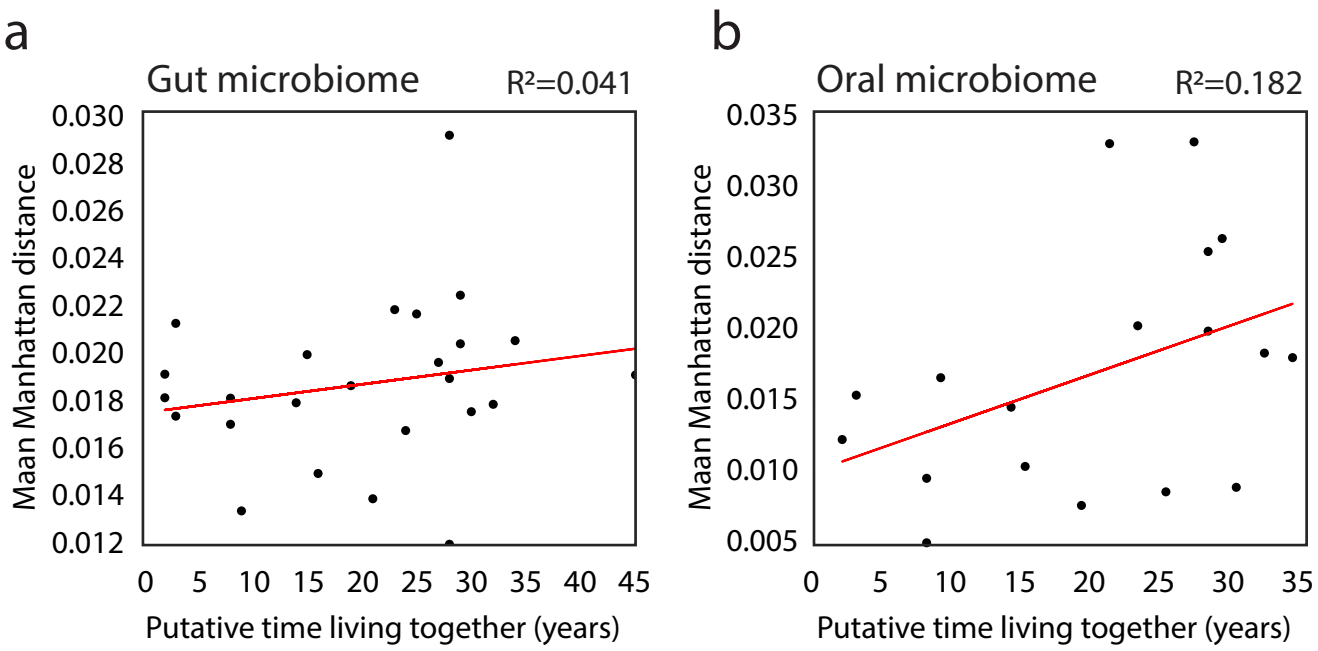
Supplementary Figure 3. Example of 1kb windows in a pair of individuals that share and do not share a strain. Median read depths, averaged across all sites within each 1kb window, are plotted for two individuals. The example shown here is gut microbiome partition 3558 for two individuals who share (left) the organisms versus two individuals who are not counted as sharing the organism (right). Outliers are shown in red. One partition is shown for two pairs of people, those that (a) share the same strain; and (b) those that do not. To consider that two individuals shared the same strain, they were not allowed to have any outlying 1kb windows.

Supplementary Figure 4



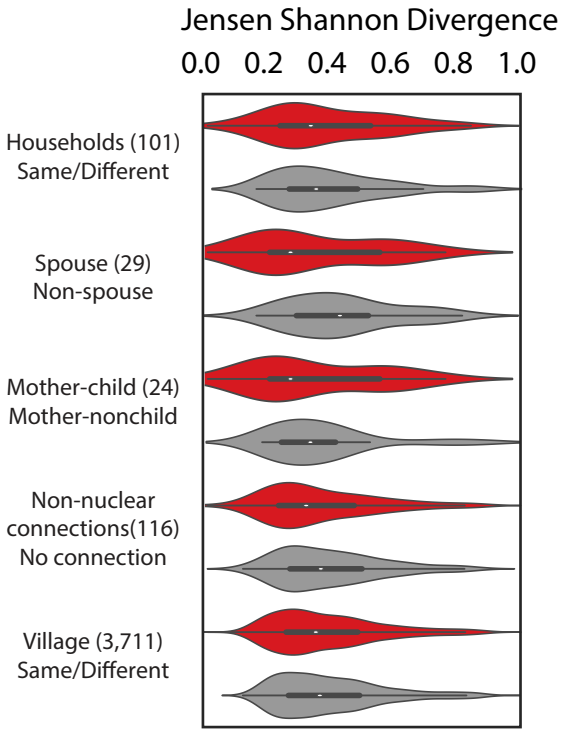
Supplementary Figure 4. Distributions of p-values for the down-sampled networks.
p-values for the core gene SNPs and 1kb windows shown in Figure 1 are the median p-values are for one-sided binomial tests between the linked members and subsampled unlinked members. Here, we show the full range of log(p-values), without multiple test correction, for the SNPs and 1kb windows in the (a) gut and (b) oral microbiomes for 100 subsampled networks each. 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. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Supplementary Figure 5



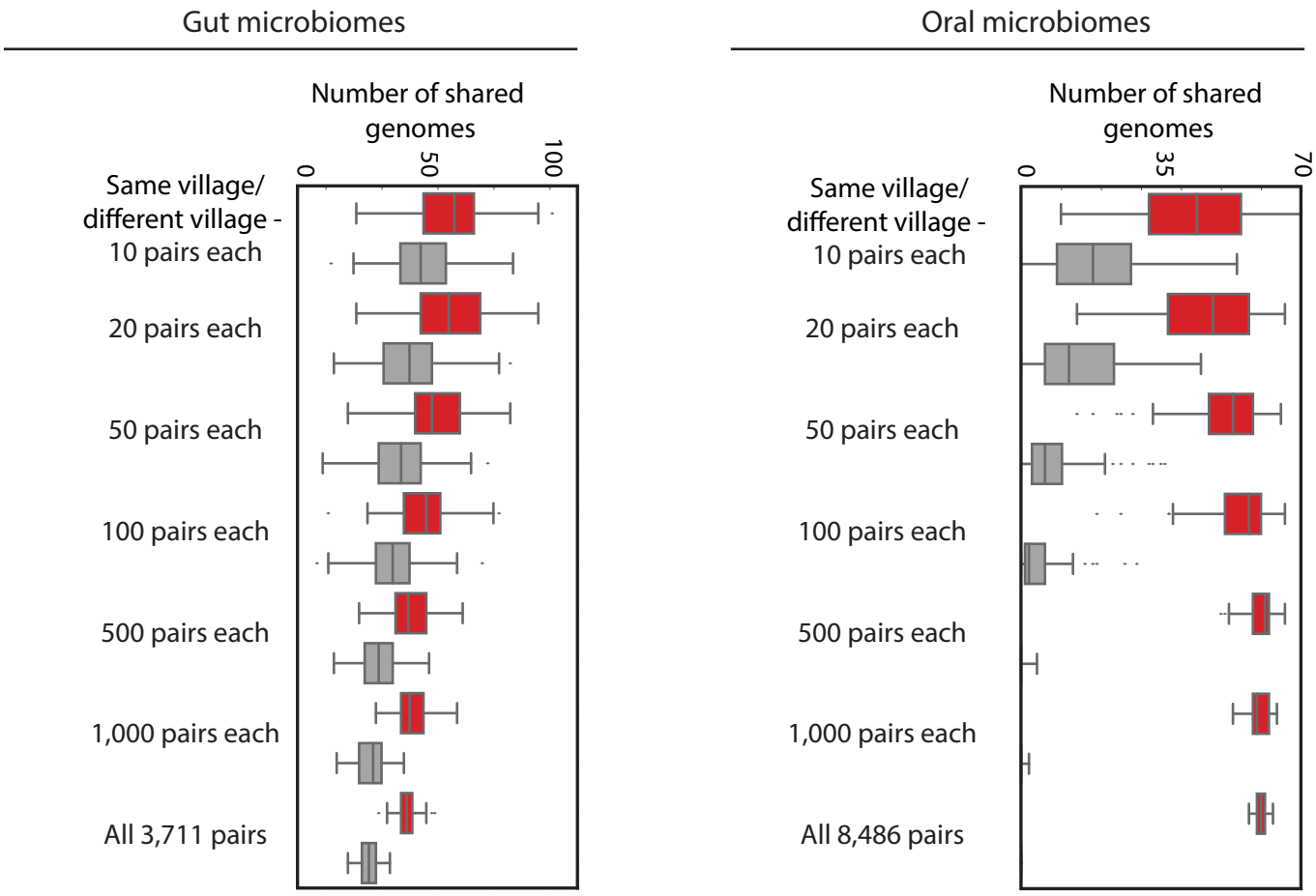
Supplementary Figure 5. Strain transmission in spouses according to the length of cohabitation. The length of cohabitation, inferred by the age of the couple's oldest child, is plotted for each spousal pair against the mean Manhattan distances across genomes for that pair in the (a) gut (*N*=25) and (b) oral (*N*=18) microbiomes. Linear regressions are shown in red.

Supplementary Figure 6



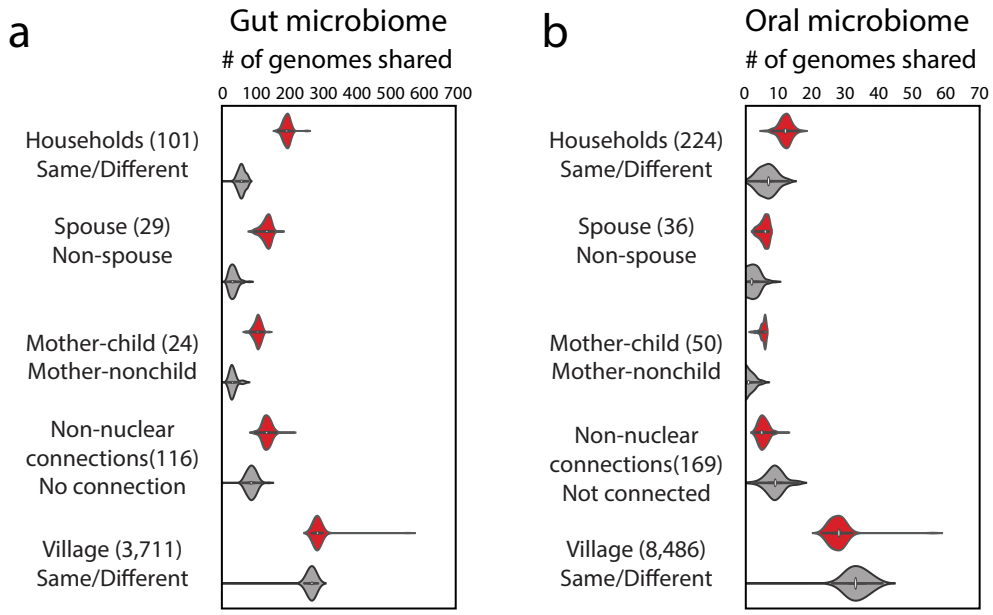
Supplementary Figure 6. Transmission of mobile genes is mildly detectable between spouses using a previously identified set of mobile genes. We calculated the Jensen-Shannon divergence for the vectors of abundances of mobile genes identified in Brito et al., (2016) in individuals' gut microbiomes across social networks. 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 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network are as follows: household 101; spouse 29; mother-child pairs 24; any connection 116; village 3,711.

Supplementary Figure 7



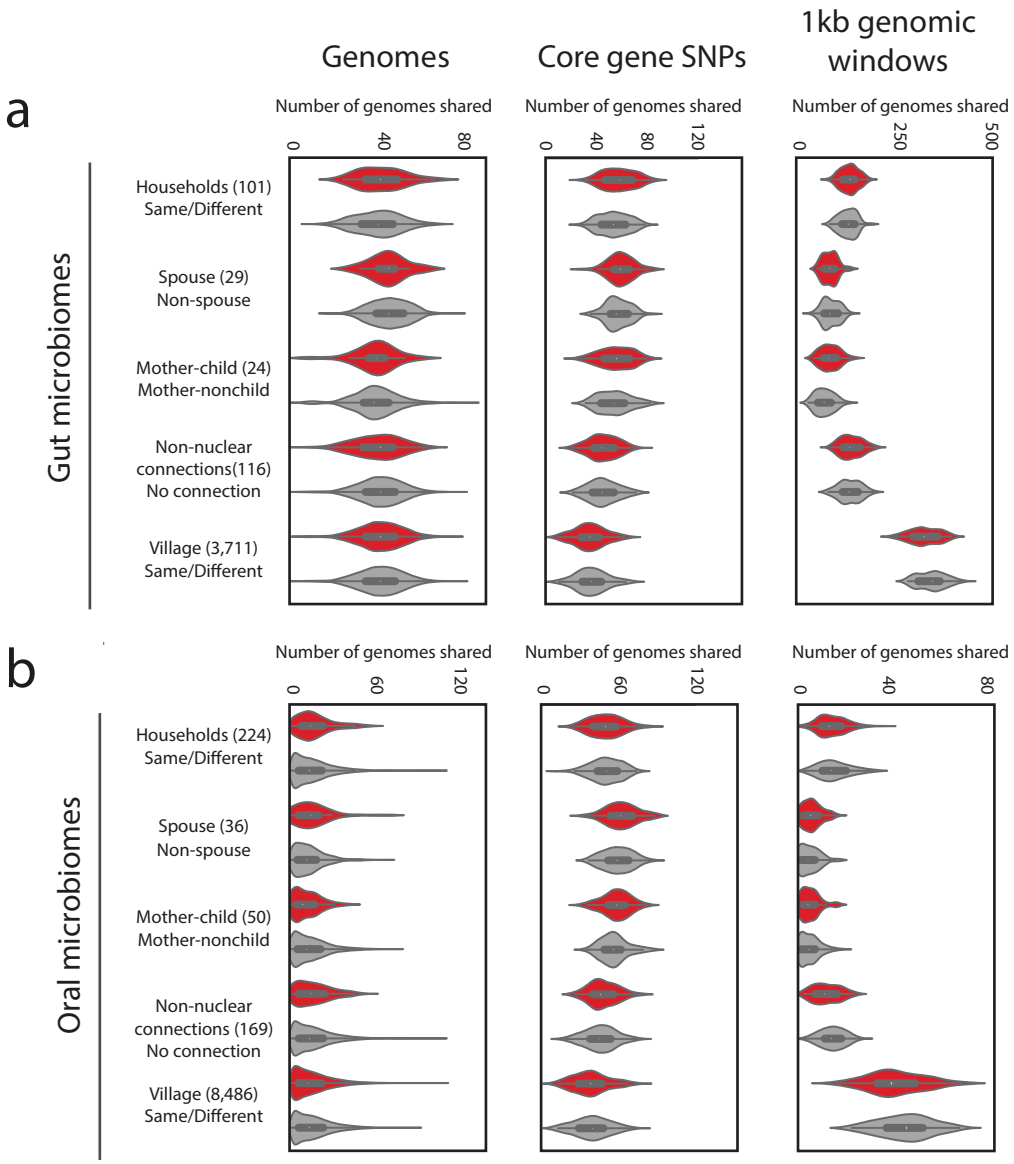
Supplementary Figure 7. Transmission is detected even at small numbers of pairs in our network. We rarefied the number of people considered in each iteration of our network analysis to determine the detection limit for the observed transmission patterns, to *N*=10, 20 50, 100, 500, or 1,000 pairs of individuals for the gut microbiome (left) and oral microbiome (right) samples. Whiskers inside boxplots extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The box 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.

Supplementary Figure 8



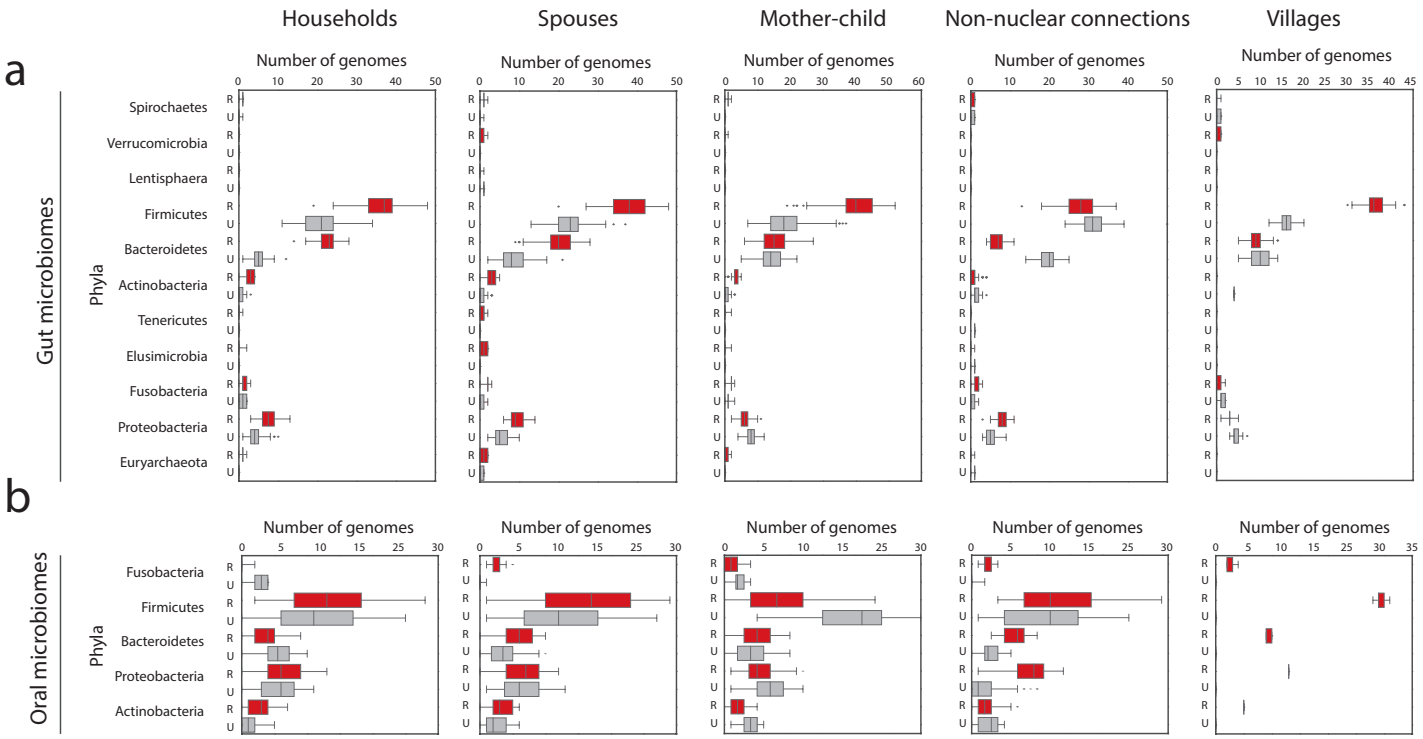
Supplementary Figure 8. Transmission signals are maintained with fewer, albeit higher quality, partitions. We reanalyzed 1kb segments from partitions with low amounts of putative contamination (less than 10% as determined by CheckM) for signals of transmission within the FijiCOMP cohort. 46 out of 207 gut microbiome partitions were removed, in addition to 34, for which CheckM was unable to run. 258 partitions out of 1,091 oral microbiome partitions were removed. The number of genomes shared according to 1kb segments for the (a) gut and (b) oral microbiomes are shown. 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. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486). Whiskers inside the violin extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median.

Supplementary Figure 9



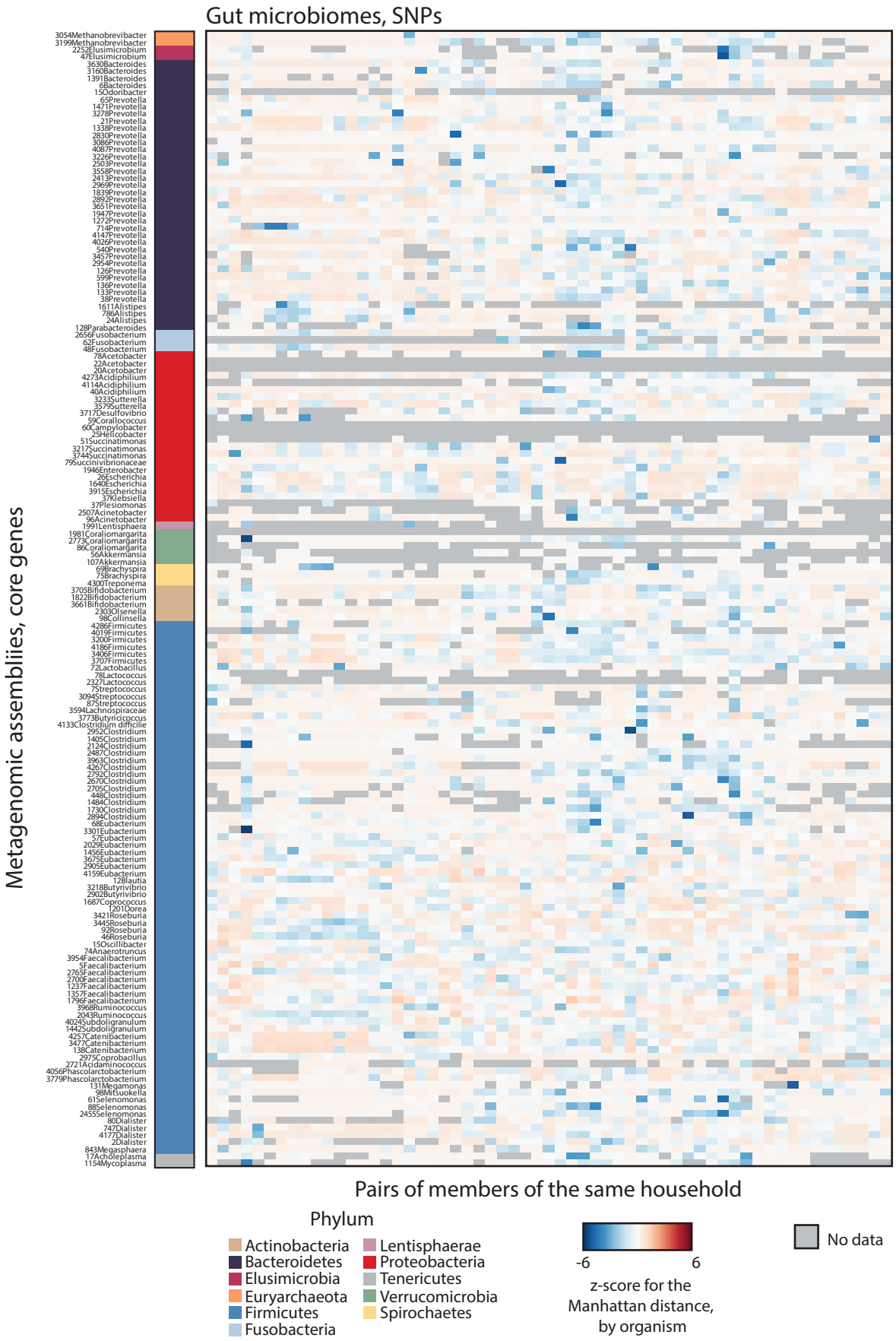
Supplementary Figure 9. Transmission signals are ablated with the shuffled networks. First the network was shuffled, maintaining the overall network structure. In the (a) gut and (b) oral microbiomes, the number of shared genomes (left); the number of genomes with shared core gene SNP profiles, determined by Manhattan distances (middle); or the number of genomes sharing flexible genomic regions, determined by 1kb genomic windows (right), whose data are represented in Figure 1; are shown for linked and subsampled unlinked members of the social network as indicated for the shuffled networks. 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. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486). Whiskers inside the violin extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median.

Supplementary Figure 10



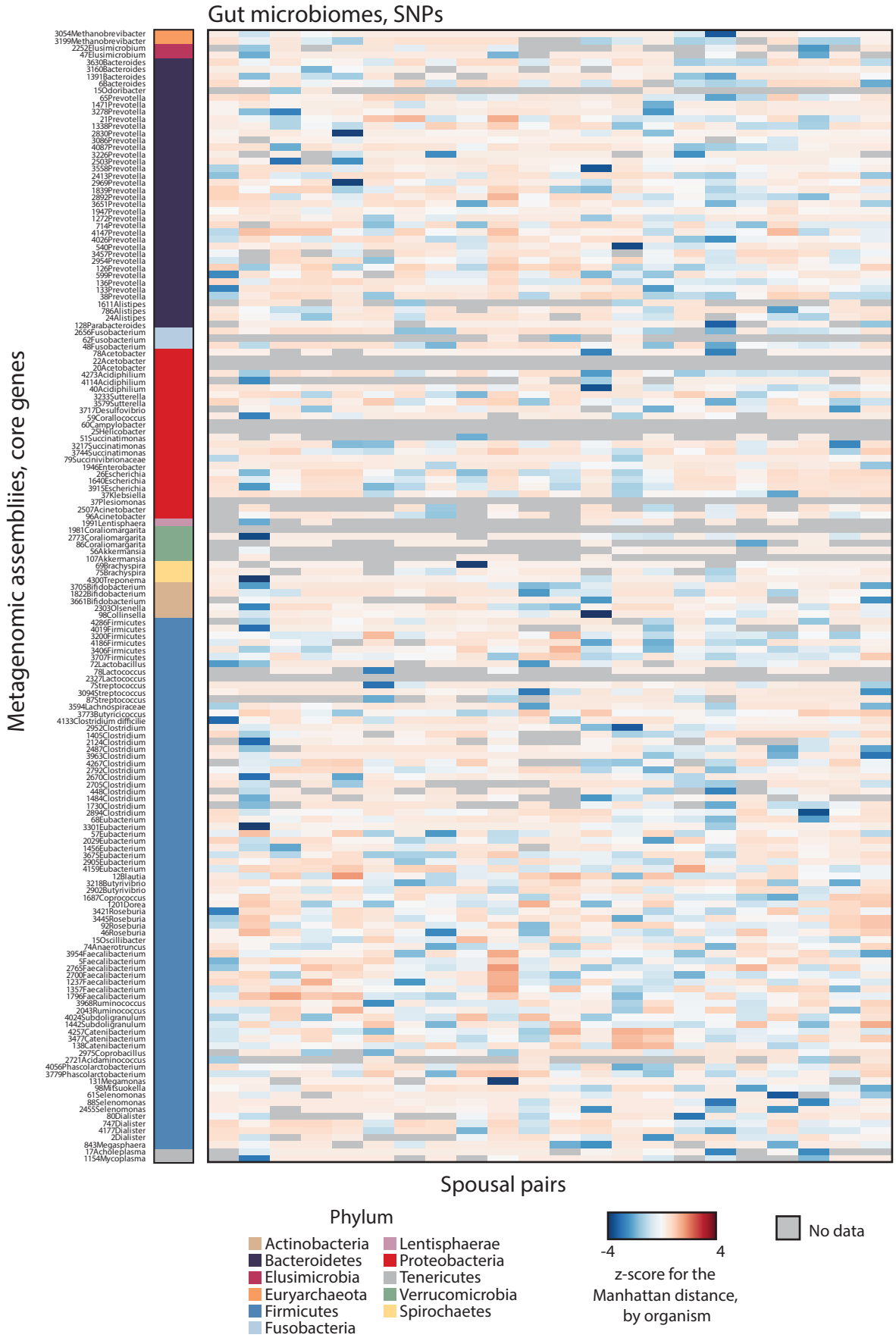
Supplementary Figure 10. Phylogeny does not correlate with transmission patterns. In the (a) gut and (b) oral microbiome samples, the number of genomes with shared core gene SNP profiles, determined by Manhattan distances, were determined for organisms according to their phylum. For each social group, we compared 'related' versus 'unrelated', shorthand for directly linked and unlinked in the network. There were no consistent phyla associated with greater sharing across social contexts in either gut or oral microbiomes. The box 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 boxplots extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Supplementary Figure 11



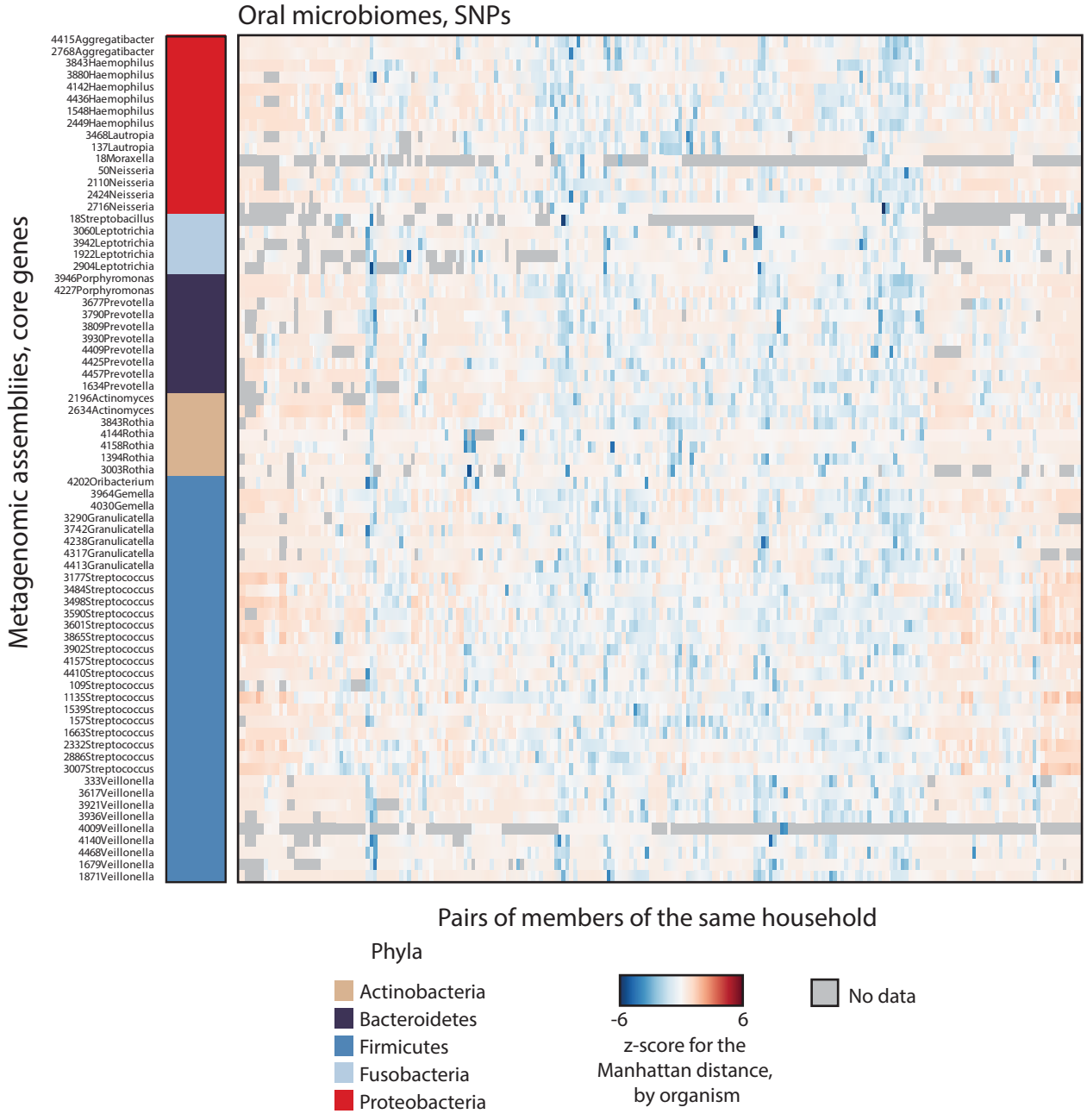
Supplementary Figure 11. Heatmaps show a dispersed transmission signal across organisms and pairs of individuals. For the core gene SNPs, a z-score of the Manhattan distance was calculated for each organism across the gut microbiomes of pairs of individuals living in the same household.

Supplementary Figure 12



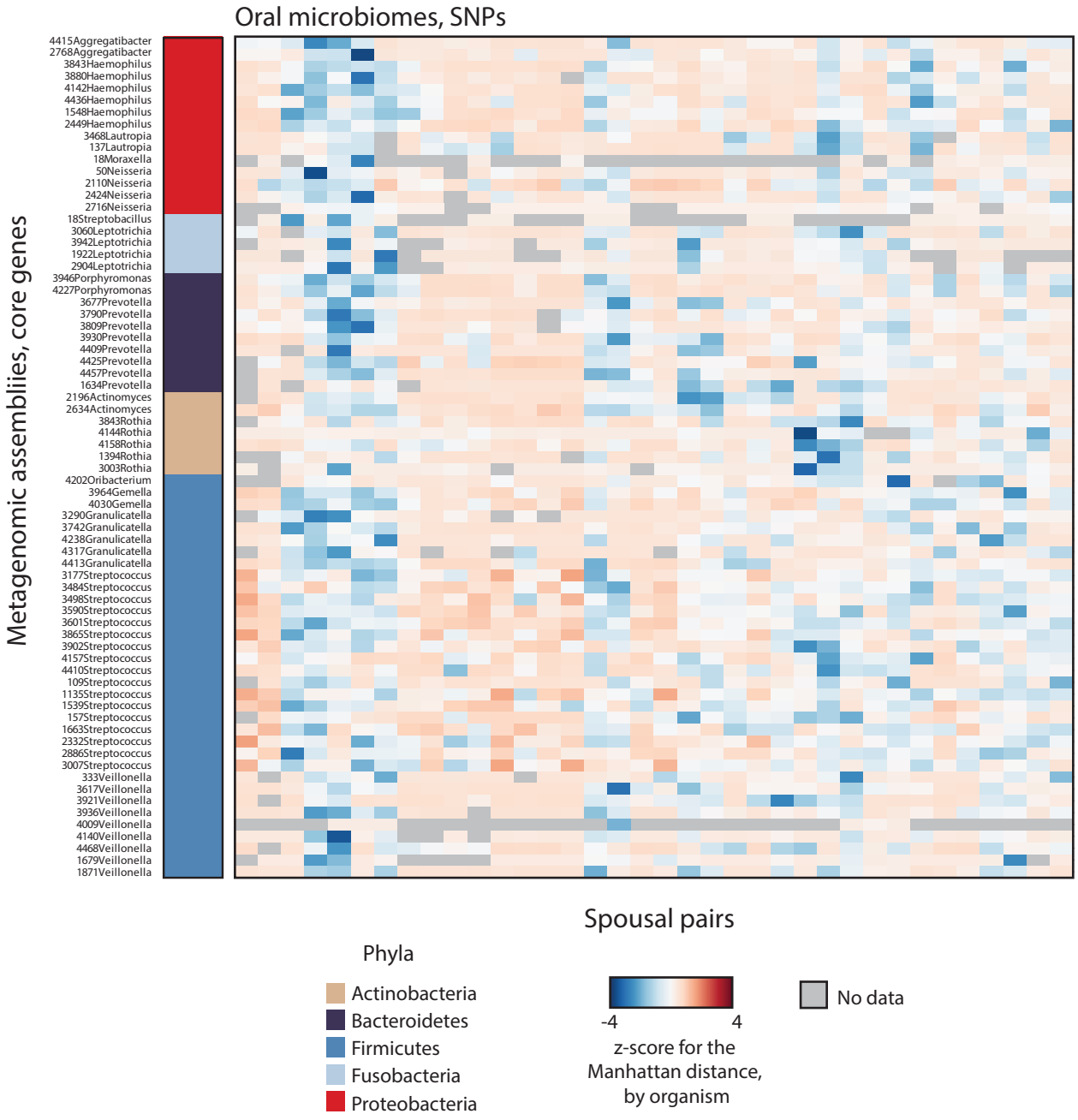
Supplementary Figure 12. Heatmaps show a dispersed transmission signal across organisms and pairs of individuals. For the core gene SNPs, a z-score of the Manhattan distance was calculated for each organism across the gut microbiomes of spouses.

Supplementary Figure 13



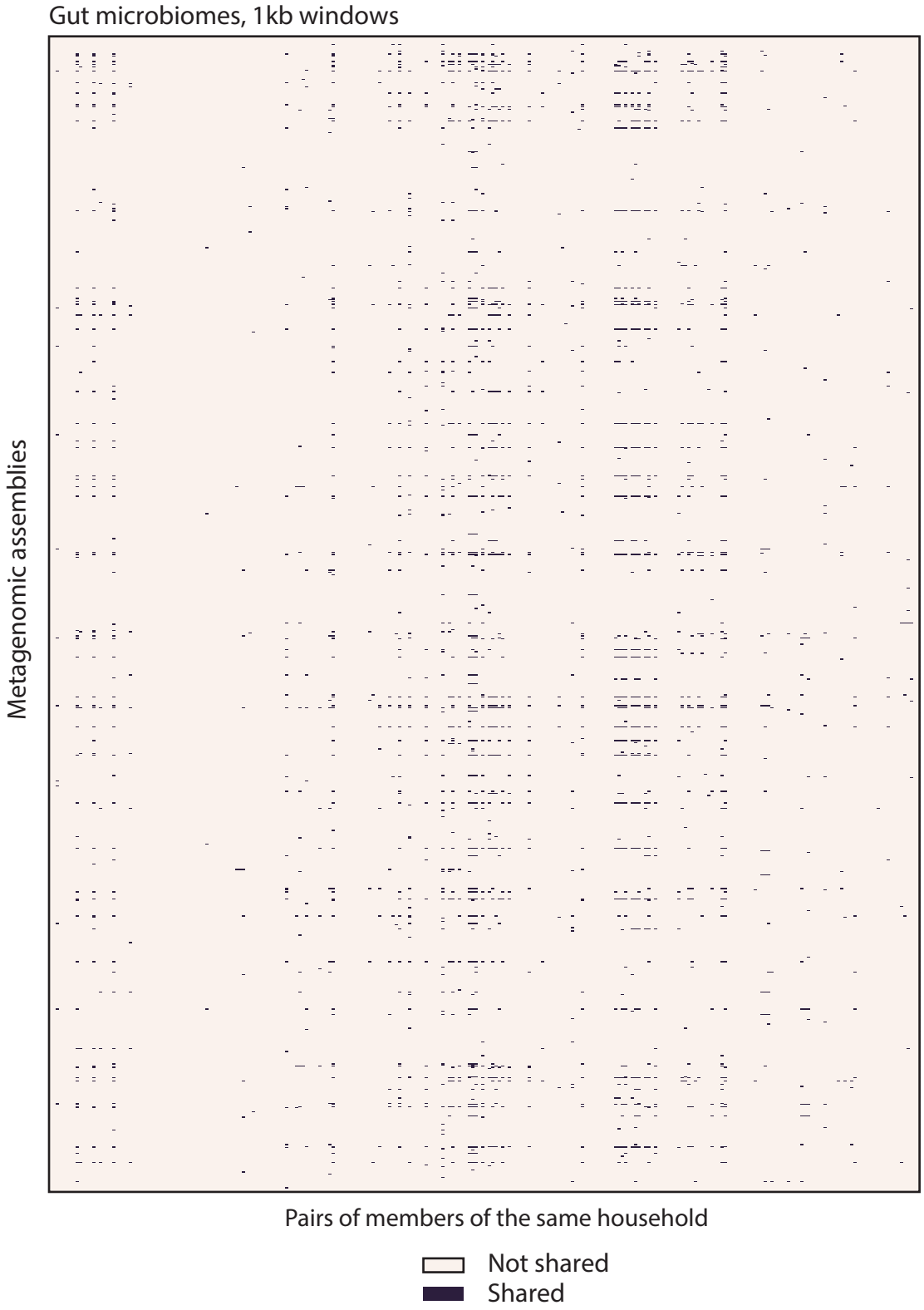
Supplementary Figure 13. Heatmaps show a dispersed transmission signal across organisms and pairs of individuals. For the core gene SNPs, a z-score of the Manhattan distance was calculated for each organism across the oral microbiomes of pairs of individuals living in the same household.

Supplementary Figure 14



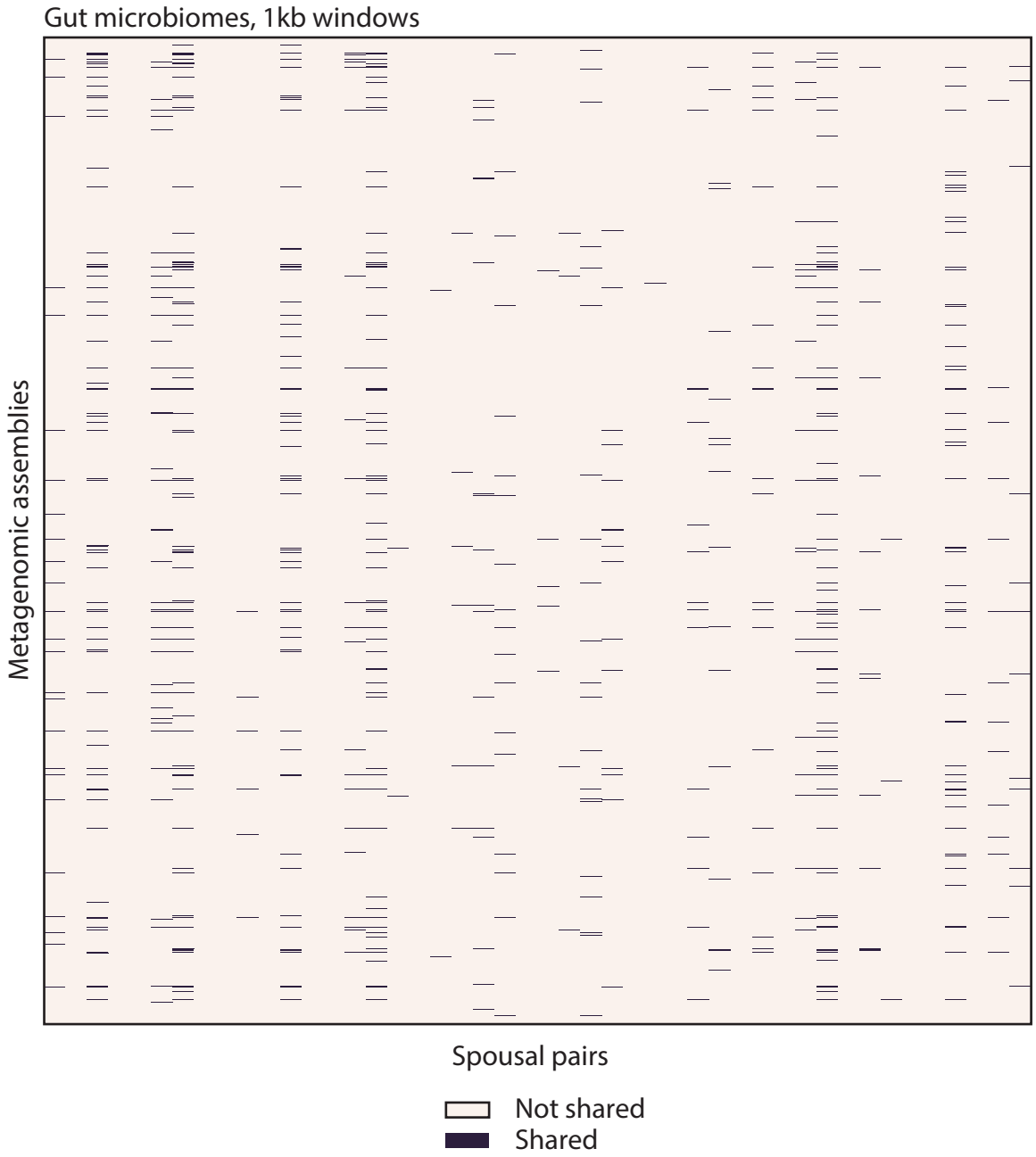
Supplementary Figure 14. Heatmaps show a dispersed transmission signal across organisms and pairs of individuals. For the core gene SNPs, a z-score of the Manhattan distance was calculated for each organism across the oral microbiomes of spouses.

Supplementary Figure 15



Supplementary Figure 15. Heatmaps show a dispersed transmission signal across organisms and pairs of individuals. For the 1kb genomic regions, we simply plot the presence and absence of a shared partition for gut microbiomes in pairs of individuals living in the same households.

Supplementary Figure 16

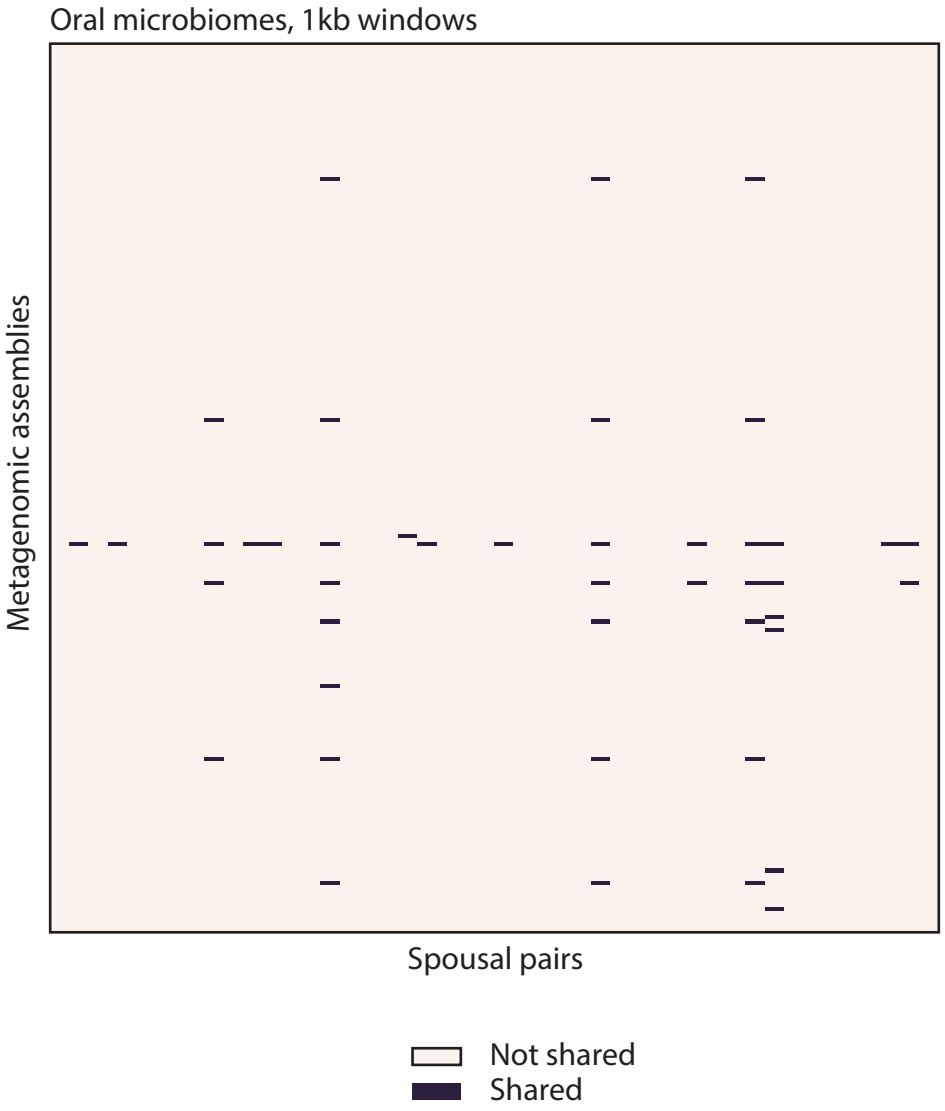


Supplementary Figure 16. Heatmaps show a dispersed transmission signal across organisms and pairs of individuals. For the 1kb genomic regions, we simply plot the presence and absence of a shared partition for gut microbiomes in spouses.

Supplementary Figure 17

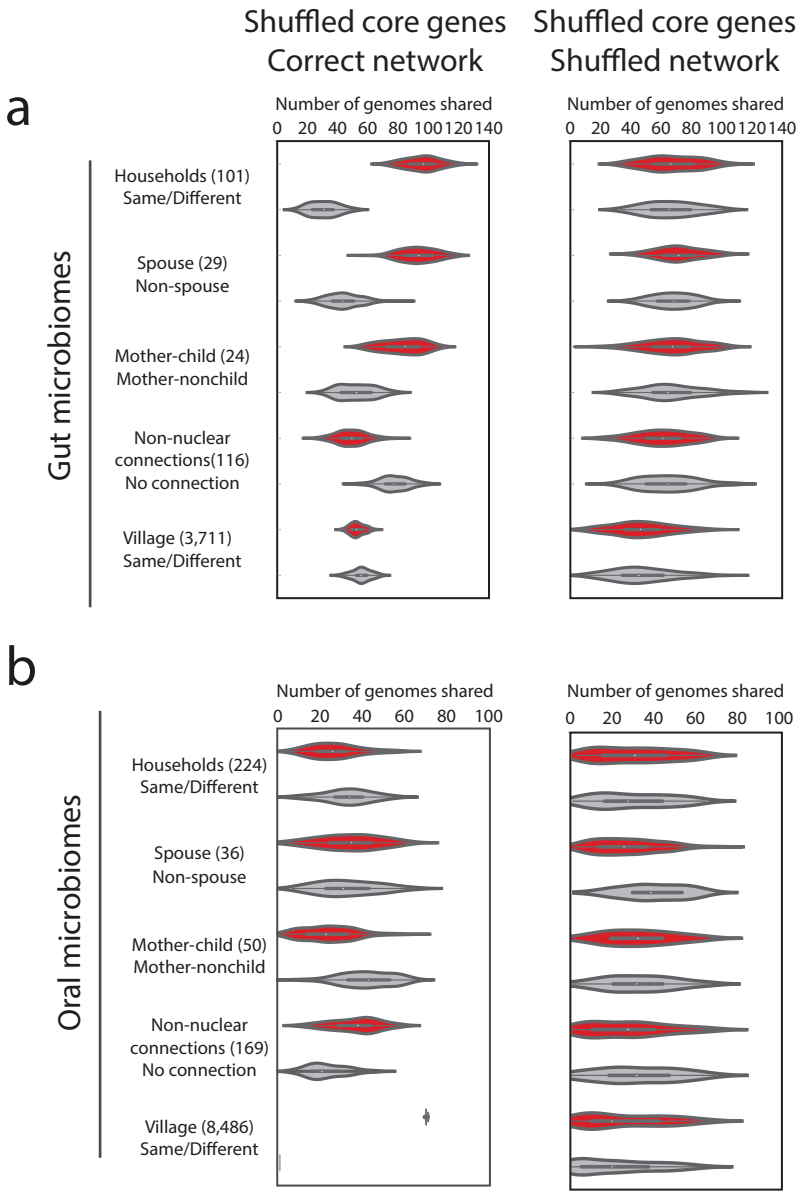
Supplementary Figure 17. Heatmaps show a dispersed transmission signal across organisms and pairs of individuals. For the 1kb genomic regions, we simply plot the presence and absence of a shared partition for oral microbiomes in pairs of individuals living in the same households.

Supplementary Figure 18



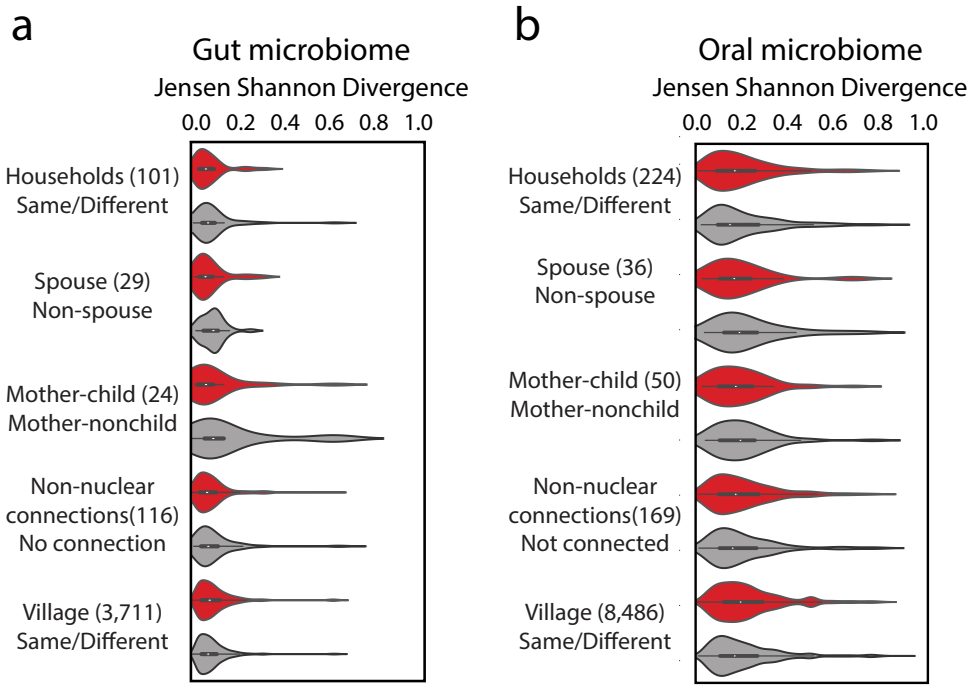
Supplementary Figure 18. Heatmaps show a dispersed transmission signal across organisms and pairs of individuals. For the 1kb genomic regions, we simply plot the presence and absence of a shared partition for oral microbiomes in spouses.

Supplementary Figure 19



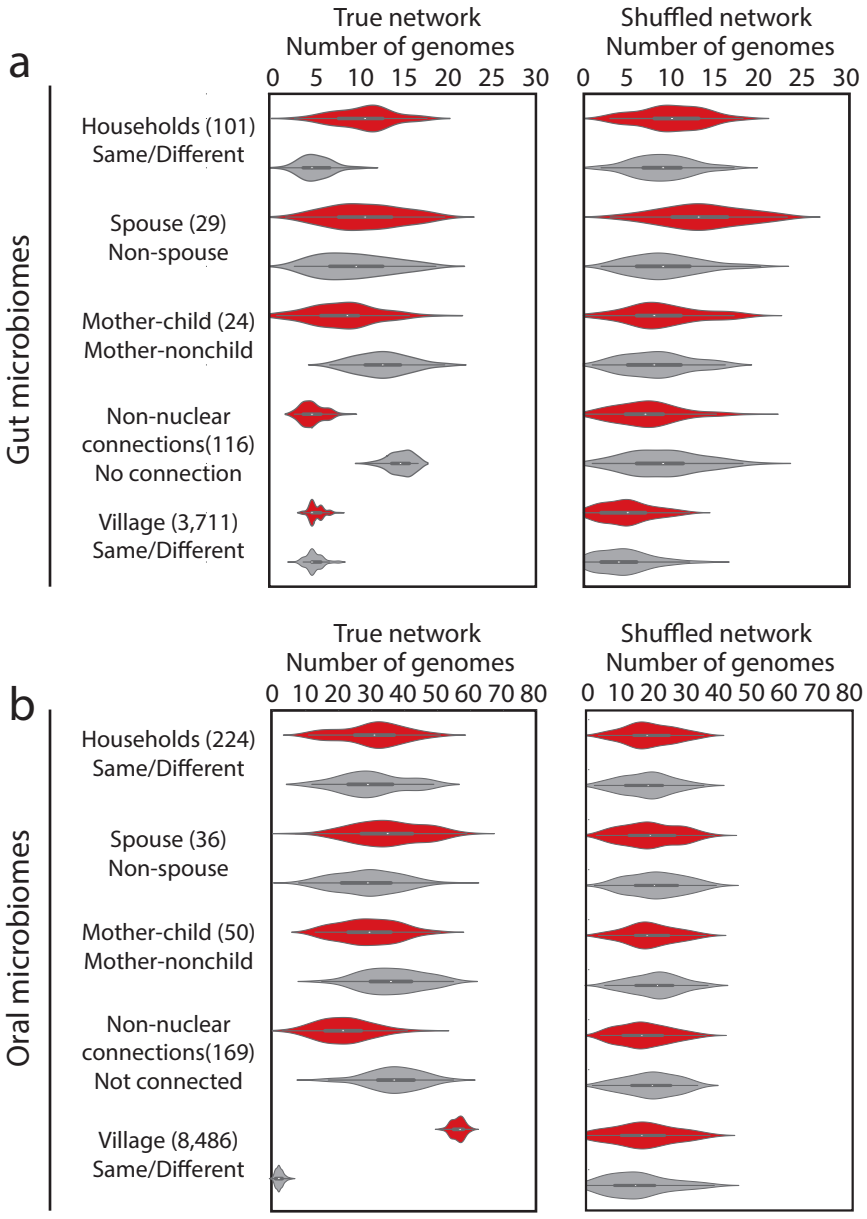
Supplementary Figure 19. Shuffling core genes does not ablate signals of transmission. We shuffled the core genes that contributed to each of the genomes in the (a) gut and (b) oral microbiomes. Briefly, an equal number of new, synthetic partitions were created from the original LSA partitions, and core genes from the originals were distributed randomly to the new partitions, with the constraint that a partition could only have one copy of a given core gene. 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 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Supplementary Figure 20



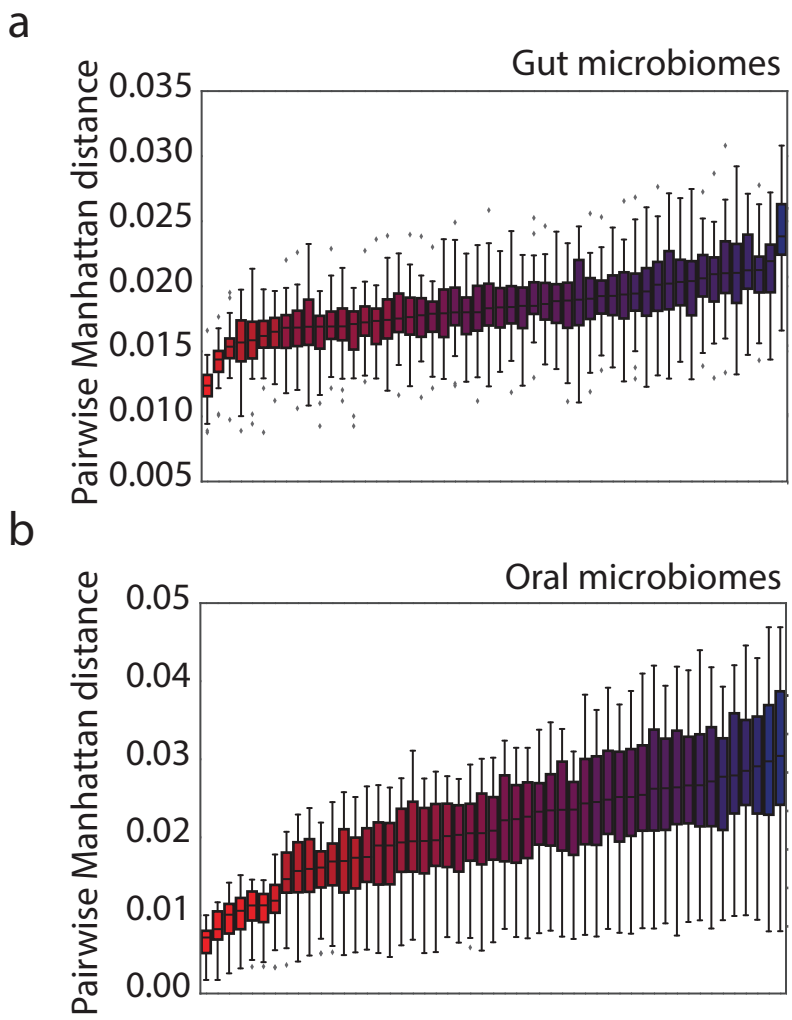
Supplementary Figure 20. Functional profiles are not indicative of transmission routes. We calculated the Jensen-Shannon divergence for abundances of genes in the (a) gut and (b) oral microbiomes, aggregated by KEGG pathway, between pairs according to each relationship type. Classes were balanced for comparison. 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 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Supplementary Figure 21



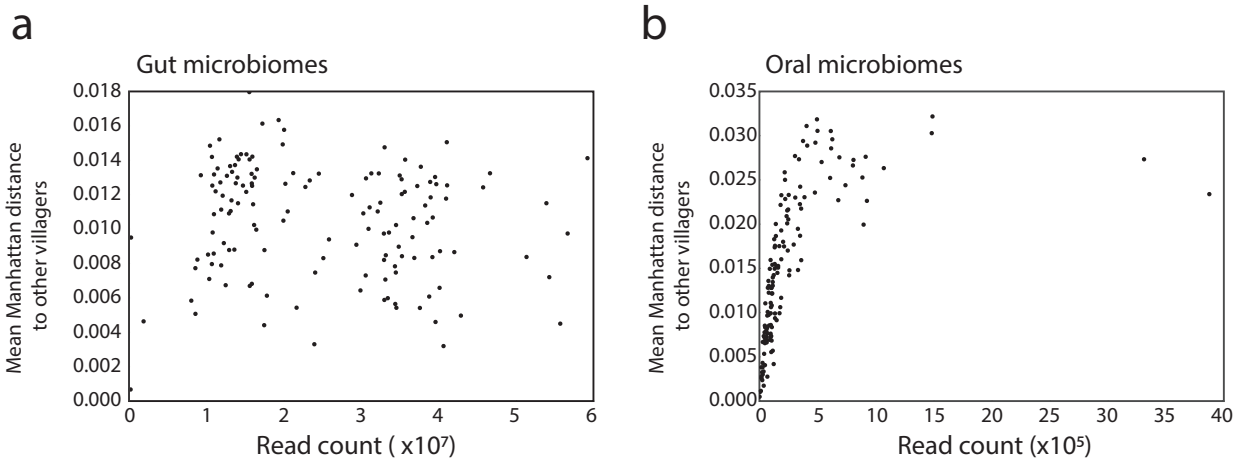
Supplementary Figure 21. Transmission signals are not reliable detected after rarefying read counts. We rarefied individuals' metagenomic libraries to 5 million quality-filtered reads and re-computed the number of shared genomes in the (a) gut and (b) oral microbiome according to their shared core gene SNP profiles, determined by Manhattan distances, and compared these with the shuffled social network computed using this many reads. 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 IQRs of the lower and upper quartile for a distribution, and center points represent its median. The numbers of linked pairs for each network (gut/oral) are as follows: household (101/224); spouse (29/36); mother-child pairs (24/50); any connection (116/169); village (3,711/8,486).

Supplementary Figure 22



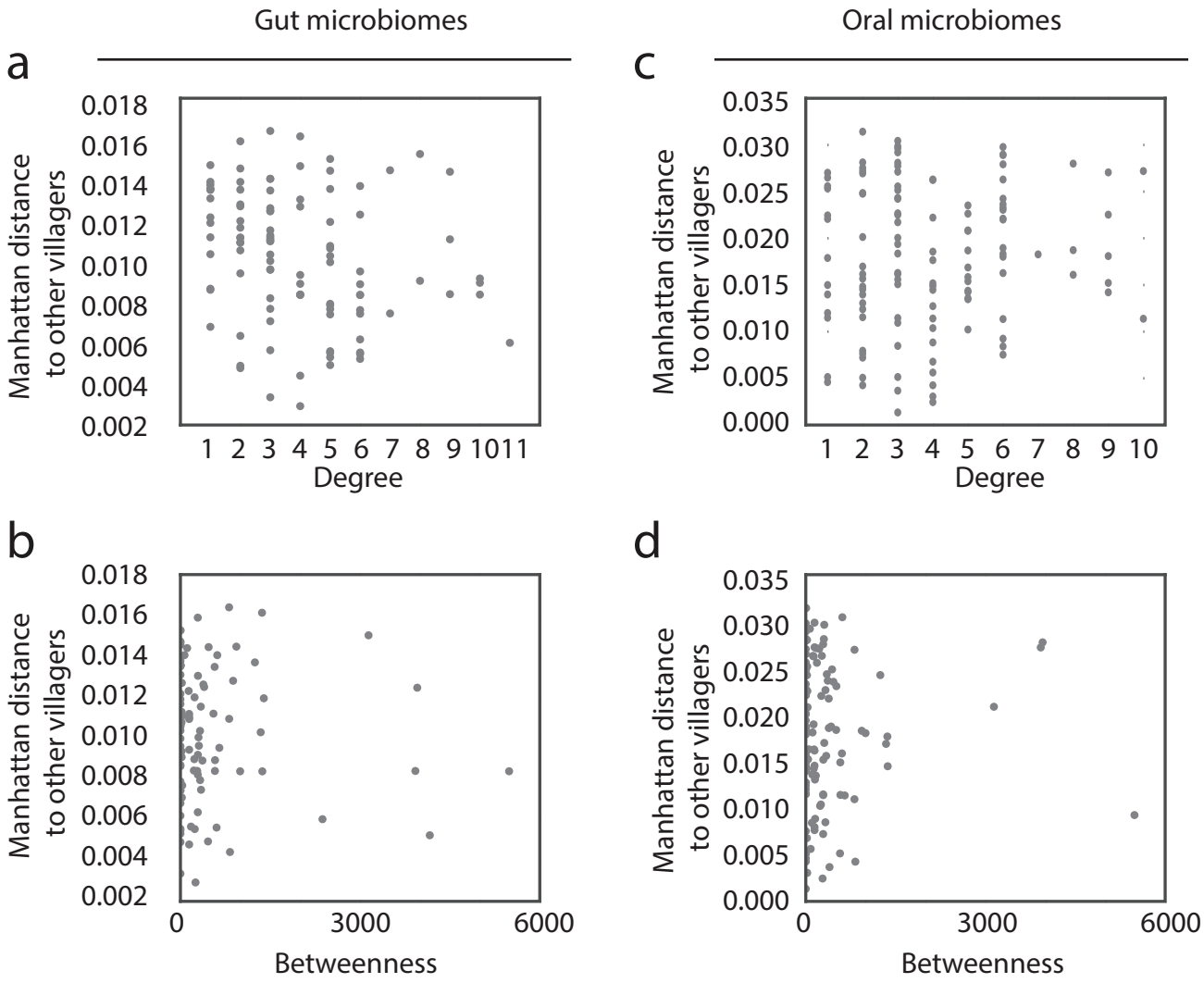
Supplementary Figure 22. Distributions of Manhattan distances between all direct linkages for each individual. For simplicity's sake, distributions of Manhattan distances for (a) gut and (b) oral microbiomes for individuals living in one of the villages are shown (*N*=51). Whiskers inside boxplots extend to points within 1.5 IQRs of the lower and upper quartile for a distribution, and center points represent its median.

Supplementary Figure 23



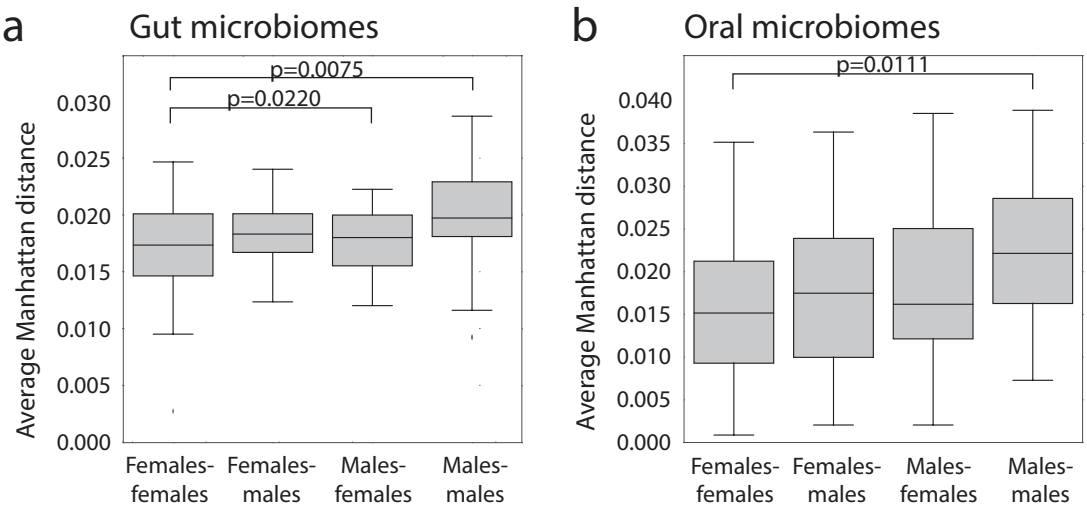
Supplementary Figure 23. Supersharing is agnostic to read depth. The number of quality-filtered reads per individual's (a) gut or (b) oral microbiomes are plotted against their median of mean Manhattan distances across all genomes to all of their directly connected individuals (N=141).

Supplementary Figure 24



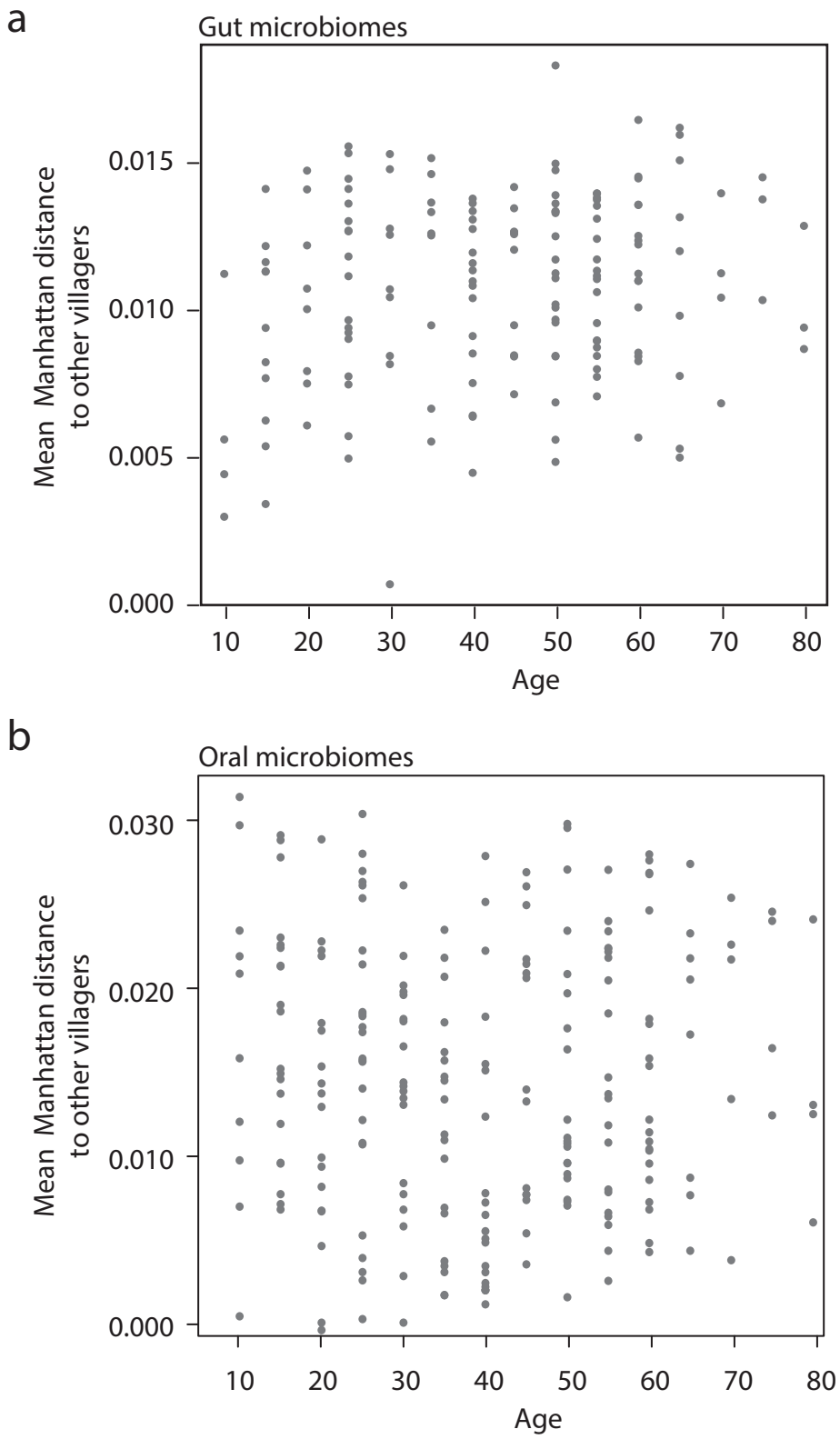
Supplementary Figure 24. Network statistics do not capture 'supersharing' metrics. Node degree and 'betweenness' are plotted for each person in the (a,b) gut and (c,d) oral microbiomes.

Supplementary Figure 25



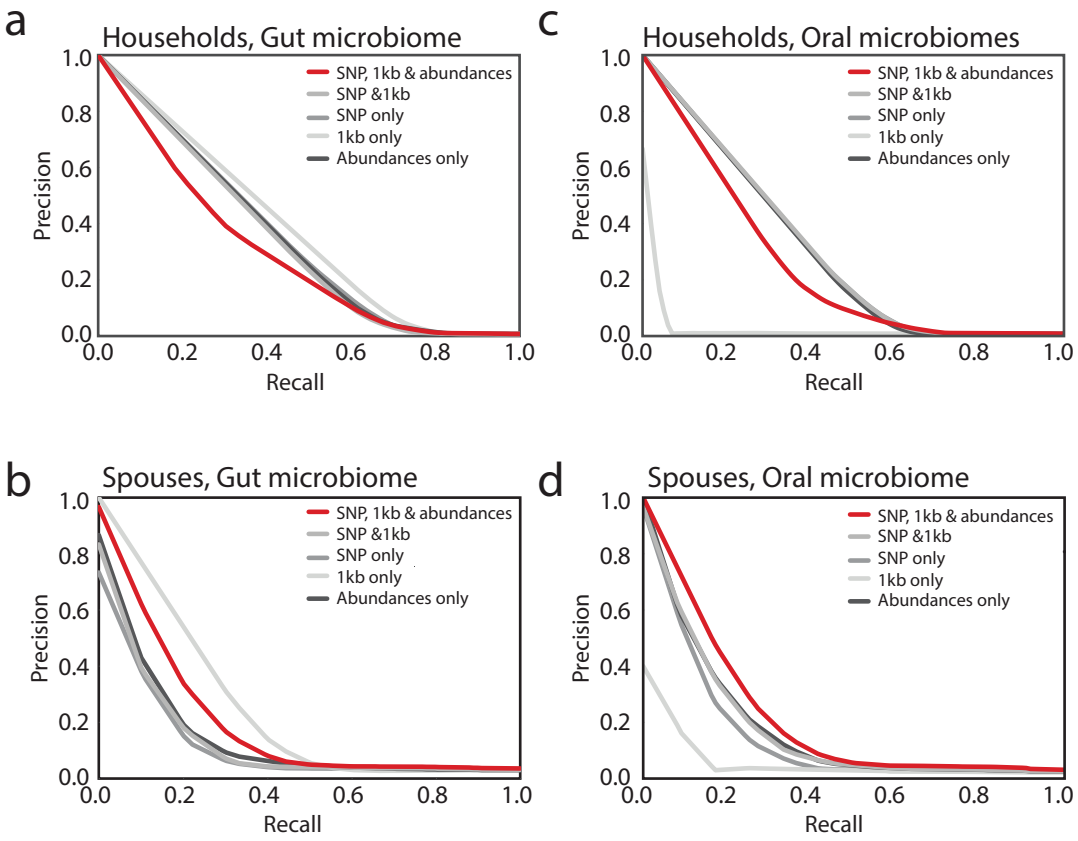
Supplementary Figure 25. Females average strains are closer to their female links. The distribution of mean Manhattan distances for each individual to all of their directly connected female or male linkages is plotted for female (*N*=125) and male (*N*=118) individuals' (a) gut and (b) oral microbiomes. Boxes indicate the upper and lower quartiles, whiskers extend to highest and lowest values excluding outliers, and center lines indicate medians. P-values reflect a two-tailed Wilcoxon Rank-sum test.

Supplementary Figure 26



Supplementary Figure 26. Age is unrelated to 'supersharing'. Individuals are plotted according to their age and their mean Manhattan distance to all individuals they are connected to in the (a) gut and (b) oral microbiomes.

Supplementary Figure 27



Supplementary Figure 27. Precision-recall curves for predictive family and social interaction models. Precision-recall curves for the models in Figure 4, namely (a) gut microbiomes of household members; (b) gut microbiomes of spouses; (c) oral microbiomes of household members; and (d) oral microbiomes of spouses.

Supplementary Tables

Supplementary Table 1. Information on the FijiCOMP study participants. We report the identification number for each person, whether gut and/or oral microbiomes were available for that individual, their sex, village and household association.

Supplementary Table 2. Social network relationships. Pairs of individuals who are spouses or have familial or social connections are listed. Note that shared household connections are not listed here, but can be found in Supplemental Table 1.

Supplementary Tables 3 and 4. Information about the gut and saliva microbiome metagenomic assemblies, respectively. For each assembled genome used to identify distinct 1kb regions across individuals, we report the number of contigs over 10kb, the size of the partition (only including contigs over 10kb, taxonomic classification based on 31 core genes (where available), and the number of people that harbor each assembly.

Supplementary Table 5. Core genes used for SNP analyses. For each assembly used to identify dominant SNP genotypes, we report the taxonomic classification, the number of core genes retained after filtering (31 genes maximum for bacteria), the number of people harboring that organisms, the length of the alignment and the number of SNPs. The taxonomic classification was determined by taking the best hit after BLASTing the core genes to NCBI's NR database and finding consensus across the core genes.