Association Between Oral Microbiota and Close-Range Proximity in a Primary School

Lorenzo Dall'Amico[1], Xiangning Bai[2,3], Sandra Marie Weltzien[4], Simon Rayner[5,6], Daniela Paolotti[1], Isabelle Sylvie Budin Ljøsne[7], Andreas Matussek[2,8], Anne-Sofie Furberg[4,9], Ciro Cattuto[1,*], Christopher Sivert Nielsen[10,11]

January 14, 2026

Abstract

The microbiota is the ensemble of microorganisms inhabiting the human body and it deeply influences human health and well-being. Recent studies showed its interplay with social behavior, suggesting that part of the microbiota might be socially transmissible. In this work, we investigate the association between close-range proximity and the oral microbiota composition in a group of children attending primary school. Unlike most related studies, our cohort comprises non-cohabiting individuals, and we use high-resolution proximity sensors to objectively measure social interactions. Our analysis shows that prolonged interactions between children pairs correlate with a higher similarity between their oral microbiota. These findings complement what is known for co-habiting individuals, suggesting that microbiota might be transmitted in social settings.

Introduction

The human microbiota is a complex ecosystem of microorganisms inhabiting the human body. It is predominantly composed of bacteria but also includes viruses, archaea, and fungi that engage in a continuous interplay of cooperation and competition [1]. It is known that the human microbiota significantly influences the host's phenotype, energy metabolism, immunity, and, ultimately, human health and well-being [2, 3, 4]. Recent studies also demonstrated its impact on psychological development and social behaviors in humans and animals [5, 6, 7, 8, 9]. As such, the human microbiome is considered the second human genome [10].

Social interactions and human health are deeply intertwined and are known to influence one another. Seminal studies have shown that obesity [11] and depression [12] are shaped by social interactions and can spread in a community. Likewise, a growing body of literature suggests that social interactions can influence the composition of human microbiota [13]. In humans, microbiota transmission can occur along a vertical direction of inheritance – primarily from the mother – and via a horizontal route through interactions with peers [14]. With time, the latter transmission mode dominates, and differences in dietary habits, environmental factors, and social contexts lead to differences in the microbiota between the child and the parents [15]. The association between microbiota composition and social structure was first documented in several non-human mammals, including mice [16], baboons [17, 18], chimpanzees [19], and lemurs [20], among others. To date, only a handful of studies have investigated the correlations between human microbiota and close-range social interactions. Ref. [21] showed that the skin microbiota of people moving to a new house rapidly aligns with that of the previous tenants. Similarly, Ref. [22] observed that family members and their dogs living

[1]ISI Foundation, Turin, Italy; [2]Department of Microbiology, Division of Laboratory Medicine, Oslo University Hospital, Oslo, Norway; [3]Division of Infectious Diseases, Department of Medicine Huddinge, Karolinska Institutet, Stockholm, Sweden; [4]Faculty of Health Sciences and Social Care, Molde University College, Molde, Norway; [5]Department of Medical Genetics, Oslo University Hospital and University of Oslo, Oslo, Norway; [6]Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [7]Department of Food Safety, Norwegian Institute of Public Health, Oslo, Norway; [8]Department of Microbiology, Division of Laboratory Medicine, Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [9]Department of Microbiology and Infection Control, University Hospital of North Norway, Tromsø, Norway; [10]Department of Chronic Diseases, Norwegian Institute of Public Health, Oslo, Norway; [11]Department of Pain Management and Research, Oslo University Hospital, Oslo, Norway. *Corresponding author: ciro.cattuto@isi.it

in the same household share skin microbiota [22]. Other studies showed that spouses can be identified from their skin microbiota [23], and that they have more similar gut microbiota and share more bacterial taxa than siblings [24]. These works suggest that, in adult life, prolonged interactions and similar lifestyles play a predominant role in shaping microbiota.

In the case of non-cohabiting individuals, Ref. [25] investigated the relation between the oral microbiota and the social network in a group of hunters and gatherers. They introduced and identified the *social microbiome* as a portion of the oral microbiome transmitted through social interactions. The authors also observed a higher diversity in the oral microbiota composition of the most central nodes in the social network, in agreement with Ref. [8], which obtained a similar result for the gut microbiota.

The cited works proved an association between the microbiota composition and the social structure, while only in a few cases transmission could be assessed [26, 27]. Ref. [26] showed that it is possible to infer cohabiting couples from their gut microbiota, while Ref. [27] found greater gut microbiota similarity among individuals having frequent interactions in a group of isolated villages.

Although the correlation between social interactions and the microbiota composition has been observed in many studies, research in this field is still in its early stages because of the many technical challenges it poses. One challenge is the accurate measurement of close-range proximity relations. In previous studies, contacts were typically defined in terms of known, long-standing relations or proxies of thereof. The most frequent contacts were usually identified by direct observation [18, 19] or using surveys, contact diaries, and self-reported information on interpersonal affinity between individuals [27]. These measurements are subjective and tend to be biased towards long-lasting relations [28, 29]. In many cases [15, 16, 20, 22, 26, 24], membership in a family/household/village was used as a proxy for interaction. While information on these associations is readily available and reliable, using such strong relationships does not allow the study of the microbiota's similarities (or lack thereof) among individuals who have social contacts reaching beyond their closest social circle. One exception is Ref. [25] in which proximity contacts were objectively measured with radio sensing technology.

In this work, we investigate the relationship between social interactions and oral microbiota composition in a group of children attending primary school in Molde, Norway. We focus on short-range proximity interactions happening at school, and we measure them using the high-resolution proximity sensors developed by the `SocioPatterns` collaboration [30, 31]. We analyze the oral rather than the gut microbiota because changes in its composition are more likely to be observed in relatively short time frames [15]. Moreover, based on long-term studies of individuals in a confined environment [32], it was suggested that social influence might impact the oral microbiome more than the gut microbiome. It is also worth mentioning that, although oral microbiota appears to be more easily transmitted, a strong relationship has been shown between oral and gut microbiota [33].

Results

Data description

Our data describe a group of 35 children from two third-grade classes at Nordbyen primary school, Molde, Mid-Norway. During class hours, the children sat in pairs in fixed positions, except for group work and practical teaching sessions. They engaged in outdoor and indoor free activities during breaks and in the after-school care program. This cohort comprises individuals from different families, allowing us to investigate the correlation between close-range social contacts and oral microbiota among people who share a common environment, a school class in our case. Additional details on the cohort composition and the enrollment procedure can be found in the Methods section.

We characterized participants' oral microbiota through saliva samples collected during class hours at eight time points between September and November 2022. Microbial composition was profiled using 16S rRNA gene sequencing, yielding relative abundance distributions for each taxon [35], recorded at the amplicon sequence variant (ASV) resolution. Across all samples, we identified 12,322 amplicon ASVs.

To assess microbiota stability over time and across individuals, we compared samples using the UniFrac distance [36]. Figure 1A shows that each individual's microbiota composition remained remarkably stable throughout the study period, consistent with previous evidence on similar timescales [15, 37]. Each point in the figure represents a saliva sample, color-coded by individual and projected into two dimensions with

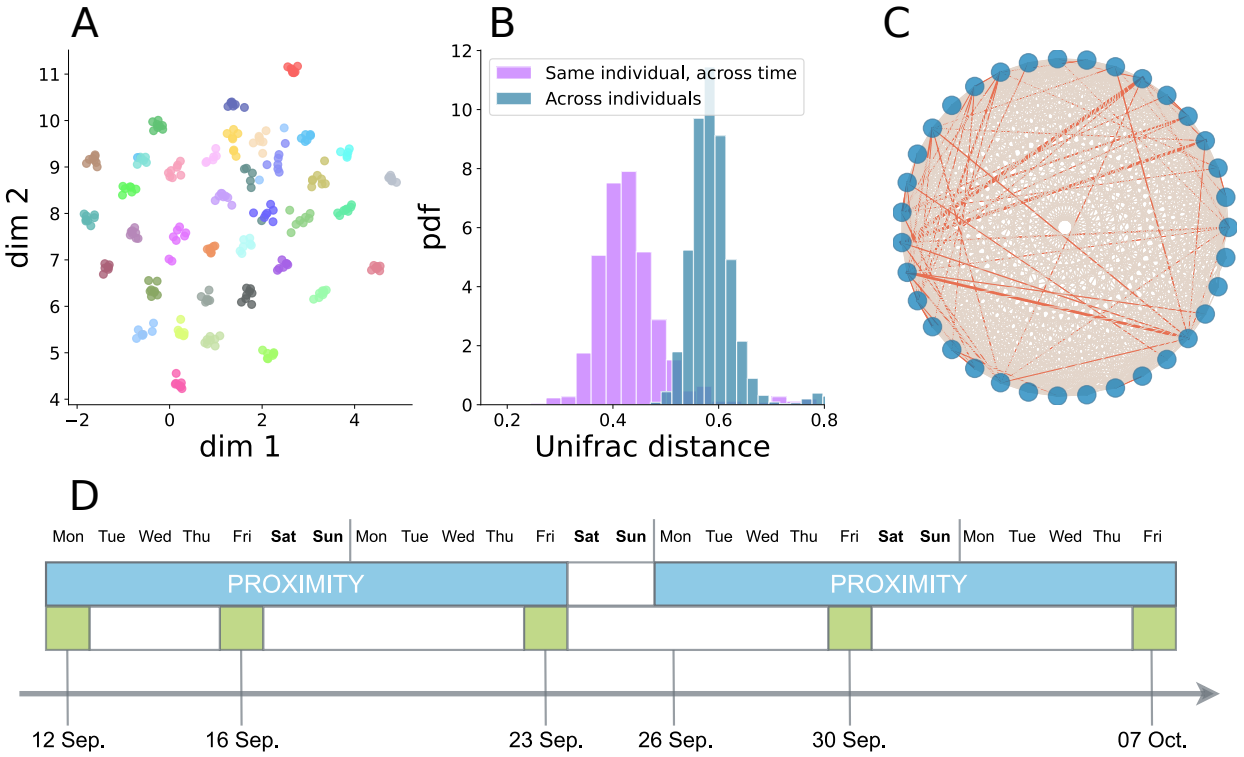

Figure 1: **Microbiota and proximity network measurements**: ***Panel A****: Empirical microbiota measurements.* Every point represents one microbiota measurement for an individual (color-coded). The two-dimensional representation is obtained with the `UMAP` algorithm [34], and the Euclidean distance between points correlates with the Unifrac distance between samples (Spearman $r = 0.42$ and p value below machine precision). ***Panel B****: Histogram of microbiota similarity across individuals and across time for the same individual.* The two histograms represent the Unifrac distance between all measurements involving the same person ("Across samples", purple) and between all measurements involving different individuals ("Across individuals", blue). ***Panel C****: Aggregated proximity network.* The blue nodes represent children, while the edges stand for interactions between them. Longer cumulative interaction times are represented with thicker lines. We highlighted in red those with a weight exceeding the 90th percentile of edge weights and in gray all others. ***Panel D****: Data collection timeline relevant for this study.* The arrow at the bottom indicates time and reports the dates in 2022 relevant for data collection. The light blue bars indicate the time intervals during which proximity sensors were used. Saliva samples were collected on the days highlighted in green, on the bottom row. The figure only shows the saliva samples and the proximity measurements that are considered for analysis.

UMAP [34]. Distances in this space correlate with the UniFrac distances between ASV profiles (Spearman $r = 0.42$, with p value below machine precision). Figure 1B confirms this stability: samples from the same individual show significantly lower UniFrac distances than those from different individuals (Mann–Whitney test, rank-biserial correlation 0.9 and p-value below machine precision). We reinforce these results with PERMANOVA analysis using the UniFrac distance matrix. Significant results are obtained when grouping the measurements by individual (p-value $= 0.001$, $R^2 = 0.51$), but not when grouping by sampling date (p-value $= 0.84$). In the Supplementary Information, we further demonstrate, using different alpha-diversity metrics, that there is no significant difference in microbial diversity across measurements.

In parallel with the microbiota sampling, we measured a short-range interaction network among children with close-range proximity sensors, developed by the `SocioPatterns` collaboration [30, 31]. These sensors are a well-established, unobtrusive, scalable technology that has been used to quantify human close-range proximity in a variety of real-world settings spanning schools [38, 39, 40], offices [41, 42], hospitals [43] and households in low-resource settings [44, 45, 46, 47], among others. Like the microbiota sampling, the proximity measurements were collected in 3 waves. Due to accumulating damage to proximity sensors (see Methods section), only the first two waves achieved satisfactory coverage of the participant's contacts. The total interaction times

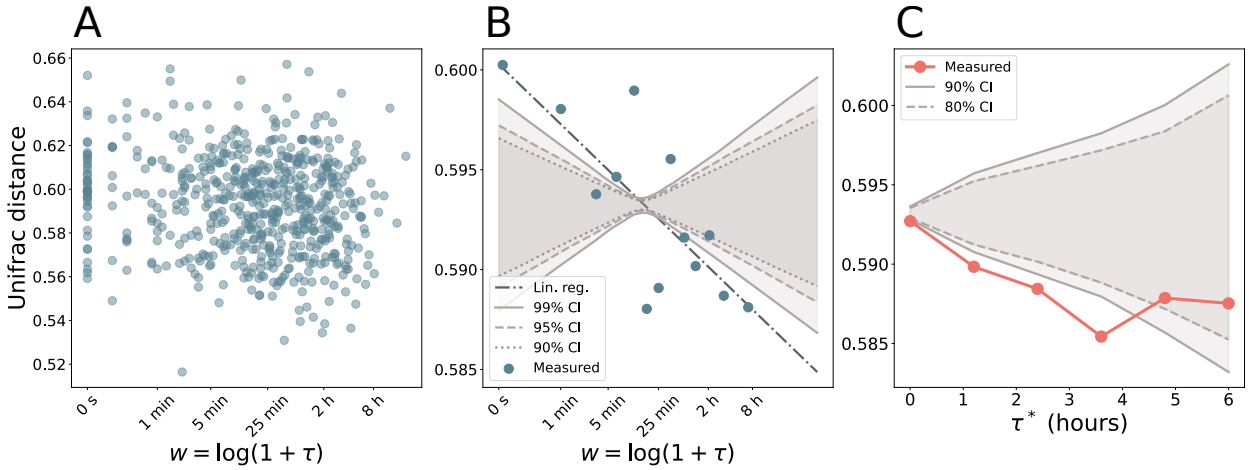

Figure 2: **Microbiota similarity as a function of time in proximity.** ***Panel A****: Scatter plot of edge weights in the physical proximity graph $\mathcal{G}_{\mathrm{p}}$ (x-axis) versus edge weights in the microbiota distance graph $\mathcal{G}_{\mathrm{m}}$ (y-axis).* Each dot corresponds to a pair of subjects; τ is the total interaction time expressed in seconds. Time in proximity increases from left to right, whereas microbiota similarity increases top to bottom. ***Panel B****. Scatter plot of the weights of $\mathcal{G}_{\mathrm{m}}$ (physical proximity) and $\mathcal{G}_{\mathrm{p}}$ (microbiota distance) averaged over edges with a similar contact duration τ.* The blue dots represent the empirical data, and the dash-dotted black line is the corresponding linear regression. The data show a strong negative correlation (Spearman $r = -0.69$, p value $= 0.01$). The style-coded gray lines delimit the 90%, 95%, and 99% confidence intervals obtained from 1500 realizations of a null model in which we randomize the association between the weights of $\mathcal{G}_{\mathrm{m}}$ and $\mathcal{G}_{\mathrm{p}}$, as detailed in the main text. Time in proximity increases from left to right, whereas microbiota similarity increases from top to bottom. ***Panel C.*** *Average microbiota distance between node pairs forming strong ties as a function of the definition of strong tie.* Strong ties have a duration τ larger than the threshold τ^*, spanning the x-axis. The y-axis shows the average Unifrac distance (weight of $\mathcal{G}_{\mathrm{m}}$) between nodes forming strong edges. The two style-coded gray lines are the 80% and 90% confidence intervals obtained over 1500 realizations of the null model randomizing the association between the weights of $\mathcal{G}_{\mathrm{p}}$ and $\mathcal{G}_{\mathrm{m}}$. Proximity strength increases from left to right, whereas microbiota similarity increases from top to bottom.

across the two measurements are highly correlated (Spearman $r = 0.25$ with p value $7 \cdot 10^{-14}$). Therefore, we choose to work with static representation of contacts obtained by aggregating all the proximity information collected by the proximity sensors into a single, undirected, weighted social graph. Graph nodes represent children, edges are proximity relations, and edge weights are defined as the total time spent in close-range proximity. Figure 1C displays the resulting graph. Edges highlighted in red correspond to time in proximity above the 90-th percentile, which is approximately 3 hours. Our analysis focuses on the five microbiota sampling times that fall within the period covered by the first two waves of proximity sensor wearing. For each participant, we define a boolean vector whose entries indicate whether that individual harbored a given variant at any of the five time points we consider. The length of the vector is therefore the number of globally detected variants. Similarly to the proximity measurements, temporal aggregation within the chosen interval is justified by the stability of microbiota measurements over the considered period, as depicted in Figures 1A, 1B. Figure 1D provides an overview of the data collection schedule used in this study.

Microbiota similarity as a function of the social interaction strength

We want to study the relationship between microbiota composition and close-range proximity, testing the hypothesis that the microbiota composition is more similar for pairs of subjects who spend a longer time in proximity. To address this, we construct two social graphs encoding, respectively, physical proximity and microbiota similarity. Nodes represent the students, and both graphs are complete, *i.e.*, connections exist between all pairs of nodes. The two graphs differ in how edge weights are defined. In the microbiota similarity graph, denoted by $\mathcal{G}_{\mathrm{m}}$, the edge weights are defined as the Unifrac distance between the microbiota compositions of the incident nodes. In the proximity graph, denoted by $\mathcal{G}_{\mathrm{p}}$, edge weights increase with the total interaction time expressed in seconds, τ, between the incident nodes. It is well known that in human proximity networks, the probability distribution of the interaction times τ is fat-tailed [31, 38, 40, 41, 44, 46], and we consequently use a logarithmic weighting for the edges of $\mathcal{G}_{\mathrm{p}}$, expressed as $w = \log(1 + \tau)$. This expression leads to zero edge weights for subject pairs that did not

interact according to the proximity sensors ($\tau = 0$).

Figure 2A reports the correlation between edge weights of the proximity graph and the microbiota similarity graph. The correlation is weak (Spearman $r = -0.13$) but significant (p value $= 0.001$). The negative correlation indicates that Unifrac microbiota distances are *smaller* for subjects who have spent more time in proximity. Figure 2B reports the relation between time in proximity and microbiota similarity at a coarser level: it shows the average microbiota distance for groups of edges with similar weight in the proximity graph. More specifically, we bin edges according to their weight (duration) in $\mathcal{G}_{\mathrm{p}}$ and then, for each bin, we compute the average weight (microbiota distance) of those same edges in $\mathcal{G}_{\mathrm{m}}$. We observe a high negative correlation between average weights of the microbiota similarity graph $\mathcal{G}_{\mathrm{m}}$ and those of the proximity graph $\mathcal{G}_{\mathrm{p}}$ (Spearman $r = -0.69$, p value $= 0.01$). We assess the significance of this result by comparing it to 1500 realizations of the null model that randomizes the association between the weights of the proximity and microbiota similarity graphs. For each randomization r, we identify the linear regression line $f_r(x)$ that best fits the data. The shaded areas of Figure 2B show the 90%, 95%, and 99% percentiles of the values of $f_r(x)$ obtained from the null model.

Finally, to complement the results of Figures 2A and B, we carry out an analysis focusing on the strong links of the proximity graph, *i.e.*, on edges corresponding to the longest lasting proximity relations, as measured by the wearable sensors. We introduce a threshold proximity duration, τ^*, and define "strong" edges as those with duration $\tau > \tau^*$. We filter the proximity graph according to this threshold, retaining only strong edges, and study the properties of this graph as a function of the threshold τ^*. Figure 2C displays the average microbiota distance for strong links as a function of the proximity duration threshold τ^*. To assess the significance of our findings, we compare our results with a null model that randomizes the association between the weights of $\mathcal{G}_{\mathrm{p}}$ (related to contact duration) and $\mathcal{G}_{\mathrm{m}}$ (microbiota distance), as we did above. The light and dark shaded areas in Figure 2C correspond, respectively, to the 80% and 90% confidence intervals computed over 1500 realizations of the null model. We observe that the microbiota distance for interacting pairs is significantly smaller than the null model one, thus supporting our hypothesis that the microbiota composition is more similar for subjects pairs that spend a longer time in proximity. We remark that for large values of the duration threshold, $\tau^* \geq 5$ hours, the results are not significant. This result is expected, as for such large values of the threshold, only a very small portion of the graph's edges (approximately 3%) are retained for the analysis, and noise tends to dominate the results.

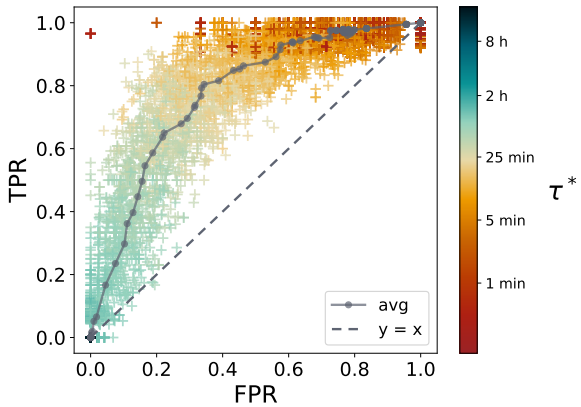

Figure 3: **ROC curve for the binary classifier used to infer the proximity links based on microbiota profiles.** We train a support vector machine to predict whether the weight (duration in proximity) of a given edge (ij) exceeds a threshold τ^*. We use, as feature vectors, the first 50 principal components of the one-hot encoding vectors indicating the ASVs shared by node pairs. The classifier is validated by training on a random selection of 90% of the links and tested on the remaining 10%. 100 random realizations of the training/test split are used, for each of the 250 logarithmically spaced values of τ^* we scan. Every cross in the plot corresponds to one realization of the training and test sets. The x coordinate is the false positive rate (FPR) and the y coordinate is the true positive rate (TPR). Markers are color-coded according to the value of the threshold τ^*. The solid gray line represents the average over the random splits and has an area under the curve auc $= 0.77$. The dashed line is the curve $y = x$ for a random-guess binary classifier.

Inference of the proximity network from the microbiota

We further explore the relationship between oral microbiota similarity and close-range proximity by studying whether we can predict the proximity network using the oral microbiota information alone. Specifically, we train a binary classifier that predicts, given the microbiota profiles of two individuals i and j and a duration threshold τ^*, whether a link between i and j with weight (duration) exceeding τ^* is present in the measured proximity network. Missing links in the network are treated as zero weight edges. In other words, given a threshold duration τ^*, we consider the complete graph of relations between individuals, and assign a link (ij) to class "1" if nodes are i and j were observed in

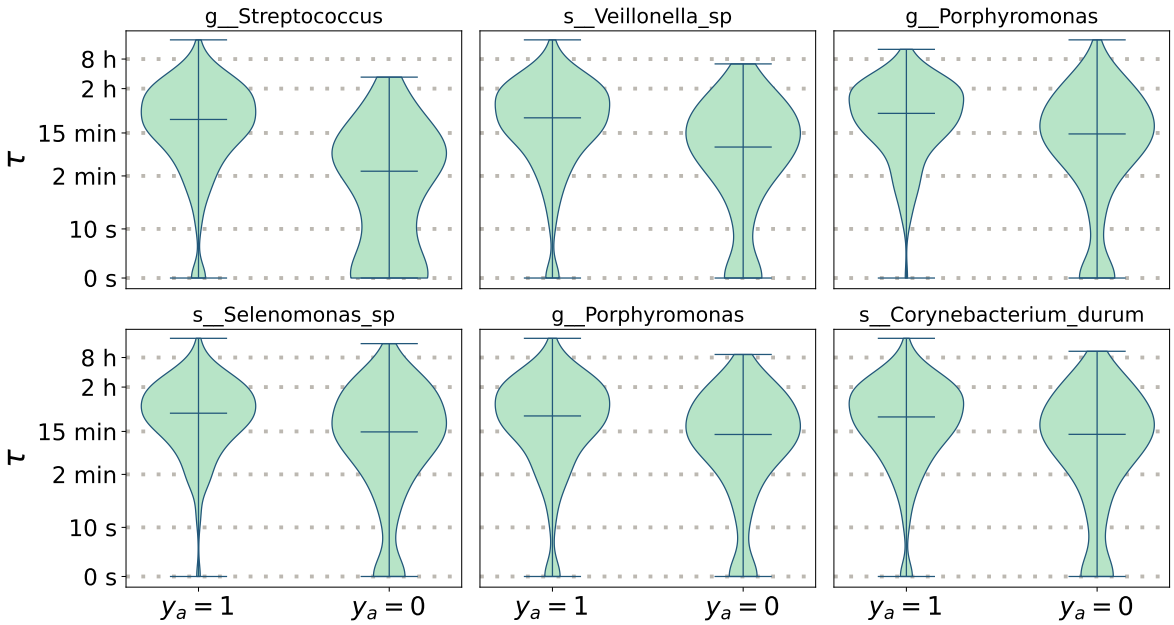

Figure 4: **Time spent in proximity by participants who share or do not share a given taxon**. Each panel corresponds to one ASV and shows in logarithmic scale the histogram of time spent in proximity by participants who share (left-hand violin plot, $y_a = 1$) or do not share (right-hand violin plot, $y_a = 0$) a given taxon. The title reports the last taxonomic level available with "s" being the *species*, and "g" the *genus*. Note that we imposed $\log(0) = 0$ to enhance readability. The plots show the two distributions for the 6 taxa for which the Mann-Whitney test is most significant, after the Bonferroni correction.

proximity for a total duration longer than τ^*, and to class "0" otherwise. As an input to the classifier, we use a binary feature vector $\boldsymbol{y}$ describing the taxa sharing between nodes i and j: entry y_a of $\boldsymbol{y}$ is 1 if taxon a is shared by i and j and 0 otherwise. To deal with the high dimensionality of ASV data, we adopt the standard machine learning approach of using a support vector machine (SVM) [48] trained on a reduced feature space comprising the 50 largest principal components of the matrix formed by the vectors $\{\boldsymbol{y}\}$. We validate the performance of the classifier by cross-validation, training it on a random sample of 90% of the links and testing it on the remaining 10%. Figure 3 shows the true positive rate (TPR) versus the false positive rate (FPR) for 100 random realizations of the training and test sets, for varying values of the threshold τ^*. The classifier achieves an average area under the curve (AUC) of 0.77. We remark that, while the accuracy of the inference task generally depends on the algorithm adopted, a single instance in which a classifier can be successfully trained (in our case, the SVM) is enough to assess the inference feasibility, hence to show the association between proximity and the microbiota composition. The Supplementary Information shows the ROC curve for a random forest classifier, which obtains a similar performance (AUC = 0.74).

Identification of taxa more strongly associated with network structure

In this section, we identify a set of taxa whose co-occurrence in pairs of individuals correlates with sustained close-range physical proximity. We independently consider each variant a and compare the interaction duration of edges in which both nodes hosted that variant (denoted with $y_a = 1$, following the notation of Section **Inference of the proximity network from the microbiota**) and the edges in which only one of the two nodes hosted that variant (denoted with $y_a = 0$). We test the hypothesis that the distribution of the microbiota variants is homophilic over the proximity network, *i.e.* that stronger interactions are associated with similar attributes. We thus identify the taxa a more strongly associated with social contacts as those in which the $y_a = 1$ edges have a significantly longer duration than those with $y_a = 0$. We attribute a duration equal to zero to the edges that were not measured by the sensors. Note that we exclude all edges in which neither of the two nodes ever hosted the considered variant, since no transmission can occur between these pairs.

We employ the Mann-Whitney signed test to assess the significance of the difference between the two distributions. This is a non-parametric signed

Taxon	p value	Corr.	δ_t
g_Streptococcus	7.81E-13	0.60	25
s_Veillonella sp	6.92E-08	0.35	22
g_Porphyromonas	4.46E-06	0.29	23
s_Selenomonas sp	4.53E-05	0.27	20
g_Porphyromonas	1.08E-03	0.24	18
s_Corynebacterium durum	5.67E-03	0.23	16
s_Capnocytophaga leadbetteri	8.93E-03	0.22	16
g_Streptococcus	1.19E-02	0.24	22
g_Porphyromonas	2.08E-02	0.24	21
s_Prevotella nanceiensis	2.70E-02	0.22	22
g_Stomatobaculum	3.59E-02	0.31	16
g_Parvimonas	3.88E-02	0.26	28

Table 1: **List of taxa more strongly associated with the proximity network**. The column "Taxon" is the variant identifier, which is preceded by a letter denoting the last available taxonomic level. The letter "g" stands for genus, and "s" for species. The second column reports the p value of the Mann-Whitney test after the Bonferroni correction, while "Corr." is the corresponding rank bi-serial correlation. The column "δ_t" reports the difference between the medians of the two distributions, expressed in minutes.

test allowing us to test if the edges in which the nodes share a taxon ($y_a = 1$) have *larger* weights than the others. We repeat the test on each detected taxon independently, and apply the Bonferroni correction to the statistical significance of the Mann-Whitney test. The test shows that edges with $y_a = 1$ (those in which both nodes hosted the variant a) have longer durations for several taxa. Figure 4 compares the weight distributions for the edges with $y_a = 1$ and $y_a = 0$ for the taxa in which the Mann-Whitney test p value is smallest. Table 1 reports the last available taxonomic level for all the taxa in which the p value is smaller than 0.05, after the Bonferroni correction, the corresponding rank bi-serial correlation, and the size effect express as difference between the medians of the two distributions. These results show the correlation between the microbiota and prolonged close-range physical proximity. We note that some of the genera in Table 1 were also identified as part of the "social microbiome" of [25], which are associated with periodontal diseases. This is particularly the case of *Corynebacterium*, *Capnocytophaga*, *Porphyromonas*, and *Prevotella*. We remark that we identified three variants (*Prevotella pallens*, *Prevotella melaninogenica*, *Clostridiales bacterium*) for which the Mann-Whitney is significant for the hypothesis that $y_a = 1$ edges have *shorter* durations.

Association between network centrality and microbiota diversity

Previous studies showed that both in humans [8, 25] and animals [18], individuals with a larger social network have higher gut and oral microbiota diversity. We investigate this relation with our data, following Ref. [25]. Here, the authors introduce the *bacterial sharing network* by letting the weight of the edge (ij) be the number of shared taxa between i and j among the ones associated with the proximity network structure (those listed in Table 1, in our case). The authors then investigate the correlation between the eigenvector centralities in the proximity network and in the bacterial co-sharing network. Figure 5 shows the results of this analysis on our data and displays a strong positive correlation (Spearman $r = 0.52$, with p value $= 1.5 \cdot 10^{-3}$) between the centrality values obtained from the two networks. In the Supplementary Information we explore other centrality measures and show that the node strength in the proximity network $\mathcal{G}_{\mathrm{p}}$ also correlates with the number of taxa displayed in Table 1 hosted by each node.

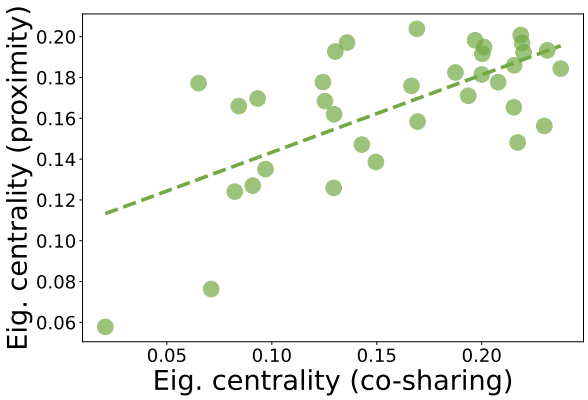

Figure 5: **Association between node eigenvector centralities in the proximity and in the bacterial co-sharing networks**. Each dot refers to an individual. The y-axis represents the node's eigenvector centrality in the proximity network $\mathcal{G}_{\mathrm{p}}$. The x-axis represents the node's eigenvector centrality in the bacterial co-sharing network, whose edge weights are defined as the number of taxa from Table 1 shared by the edge nodes. The dashed green line is the result of the linear regression between the two centrality metrics. The Spearman correlation between the two variables is $r = 0.52$ with p value $= 1.5 \cdot 10^{-3}$.

Discussion

Our study shows a statistically significant association between the composition of oral microbiota and close-range proximity of non-cohabiting children in a primary school. Our findings are qualitatively consistent with previous studies involving cohabiting individuals [22] and with recent, larger-scale studies focusing on gut or oral microbiome [27, 15]

in isolated social settings, including studies using wearable proximity sensors in isolated communities [25].

The distinctive feature of our study is the choice of a highly non-isolated social environment – a primary school in a developed country – and the adoption of a comparatively low-resource approach. We use 16S rRNA gene sequencing and low-cost proximity sensors deployed for just a few weeks in the chosen social setting, without observing any of the contacts children have off-school with household members and their communities.

We remark that children were instructed to wear the proximity sensors also before and after school hours, including the weekend, but off-school interactions contribute negligibly to our analysis. Of course, this might be due to lower compliance with sensor wearing outside of school hours, but it is important to remark that the interaction strengths (durations) we observe are almost exclusively due to close-range proximity in the school.

We also show that information about the oral microbiota predicts sustained close-range proximity relations, with a prediction performance similar to that reported by other studies [26, 27] in isolated settings. This suggests that the oral microbiota structure is significantly shaped by social interactions, at least for some taxa. We also identify a set of ASVs whose presence in the oral microbiota correlates with prolonged social interaction. Possibly, these variants are transmitted via social interactions at school, and we remark that several of the taxa of Table 1 have been associated with social interactions by other studies [25]. Moreover, we observe that the most central nodes in the proximity network have a higher number of these variants: this is consistent with previous studies showing that individuals with a larger number of social interactions have a higher microbiota diversity [8, 25].

On the one hand, our results are consistent with the hypothesis that, at least for some taxa, the oral microbiota can be socially transmitted via sustained close-range proximity at school. On the other hand, it is important to remark that time proximity at school, measured during our observation period, is probably a proxy for the time in proximity spent by the same individuals during the whole school year, as well as for other types of unobserved kinships among children. Since social influence and latent homophily on unobserved attributes – socio-economic attributes, dietary habits, etc. – generally cannot be teased apart, care needs to be taken in interpreting our results as evidence for the social transmission of oral microbiota, even assuming transmission is at play. In fact, it is important to stress that the interactions times we observe should not be regarded as causally relevant for transmission. For instance, the interaction times in Figure 2 should not be interpreted as typical times required for microbiota transmission, as they only provide a definition of a social tie sufficiently strong to be associated, in our study, with the composition of the oral microbiota.

The time scales relevant for our study also need to be discussed. We observe that the microbiota of all individuals is essentially stationary over the 2 months spanning the days on which saliva samples were collected. This is compatible with two opposed scenarios: one in which the time scale of the presumed transmission is much longer than the duration of our study, and another in which this duration is much shorter. In the former case (long time scale needed for transmission), the duration of our study is insufficient for observing any significant microbiota dynamics determined by close-range proximity. In the latter case (short time scale needed for transmission), the microbiota of all individuals may have already achieved a stationary state, because our data collection started weeks after the beginning of the school year. In both cases, this study only observed a snapshot of a complex process that likely involves latent homophily on several unobserved attributes.

To overcome some of the limitations discussed above, future studies might envision a longer data collection period, with the goal of exposing the temporal dynamics of oral microbiota and identifying the time scales relevant for changes in microbiota determined by close-range proximity. Increasing the sampling frequency of oral microbiota might also be necessary to expose microbiota dynamics, and a more fine-grained approach to characterizing the microbiota composition, going beyond the limitations of 16S rRNA gene sequencing, might be essential. Similarly, to deal with confounders and tease apart oral *microbiota association with social interactions*, such as the one observed in this study, and *microbiota transmission determined by social interactions*, future studies might enroll a larger cohort, possibly including the students' households, and collect richer information about study participants, such as socio-economic attributes, dietary habits, and more.

Methods

Ethics

The Regional Committee for Medical and Health Research Ethics (REK East Norway) reviewed the research protocol and concluded it did not fall under the scope of "The Act on Medical and Health Research", as the aim was not to directly generate knowledge about health and disease. Consequently, no REK approval was required for this study. The project received approval from the Data Protection Officer at Molde University College (September 2022). All re-identifiable data were stored at Molde University College and only de-identified data files were shared with research group members.

Enrollment

Our study was conducted in a Norwegian primary school with 292 children, subdivided into 7 grades. Before recruitment commenced, the details of the study were discussed with the faculty at the primary school to ensure they had a full understanding of what the study entailed. Parents/guardians were invited to a meeting where the study was presented, and questions could be asked. Written information about the study was provided to the parents/guardians in two versions: one for the adults and another for the children, along with a consent form. Children were informed about the study orally in class. Inclusion required consent from both parents/guardians and assent from the child.

We invited 41 children aged 7-8 years (22 boys and 19 girls), 37 participated (21 boys and 16 girls). Due to sensor damage, however, the proximity data were collected for 35 children that constitute the cohort analyzed in this study. The experiment was conducted from September to November 2022. The children had a five-day school week with four to five class hours daily. During October 2022, the school had an autumn break with the duration of one week. After the data collection was completed, all children in the two classes were invited to an afternoon trip to the local cinema and a class trip to an aquarium in a neighboring city, both financed by the study. No distinction was made between children participating in the study and non-participating children.

Experiment description

One dedicated researcher from our team served as the contact point for the children, their guardians, and the teachers. This researcher organized the collection of materials at the school throughout the whole data collection period, with visits to both classes every one to two weeks. Saliva samples were collected at 8 time points, with a one-week interval between the first five samples and a two-week interval between the last three. The samples were taken from September 12th to November 25th, 2022. Daily use of sensors overlapped this entire period. The proximity sensors had to be worn outside the clothing, making them visible to other children. They had a unique identifier to simplify their identification by each child. To avoid the possible stigmatization, children who did not participate were given the option to take saliva samples without storing them for analysis and to wear a sensor with the battery removed, i.e., without data recording. Four of the 41 children invited used this opportunity.

There were no compliance issues with either procedure. The children were highly motivated to give saliva samples and to use the sensors throughout the whole study period. Feedback from parents and teachers indicated that the children were self-motivated and independent. They also looked out for each other to ensure they remembered to wear the sensors, e.g., after gym when the sensors had to be taken off. The study did not impact the children's routine, except for $10-15$ minutes set aside for collecting saliva samples on specific days. Our third-grade cohort participated in an after-school care program in the mornings ($07:15-08:20$), and joined fourth-grade children in the program during the afternoons ($13:00-16:45$). Most of the children in our cohort also live near each other and frequently spend time together outside of school hours.

Saliva sampling

The children performed saliva sampling during class hours using the GeneFiX™ Saliva Microbiome DNA collection kits (Isohelix, Cell Projects Ltd, Harrietsham, England)[1] containing sample tubes with 1 ml stabilizing solution and funnel. They were required to not eat or drink during the 30 minutes before the procedure. All the steps in the saliva collection were performed simultaneously by all the children in each class, under the instruction of the dedicated researcher from our team. First, an envelope containing the collection kit and labeled with each child's name was handed out. Next, the children followed the manufacturer's instructions. Fi-

[1] www.isolhelix.com

nally, the children placed the collection tube back into the envelope and delivered it to the researcher. The collection tube was premarked with a GeneFix™ seven-digit serial number and a barcode, which were linked to each child's study ID and the date of sampling. The saliva samples were stored at room temperature before DNA extraction. All samples were transported overnight from Molde University College to the laboratory at Oslo University Hospital HF in autumn 2022 and received in two batches on October 4th and November 30th.

16S rRNA gene sequencing and microbiota profiling

DNA from each saliva sample was extracted using the NAxtra™ nucleic acid extraction kit (Lybe Scientific) following the manufacturer's instructions, and then sent to the Norwegian Sequencing Center (Oslo, Norway) for 16S rRNA gene amplification, library preparation, and sequencing. In brief, the V3-V4 region of the 16S rRNA gene was amplified based on the two-step PCR procedure as described in the Illumina application note using primer sequences derived from Ref. [49]. Libraries were pooled and sequenced on an Illumina MiSeq platform (Illumina, San Diego, CA, USA) with 300 bp paired-end reads (v3 reagents). 20% PhiX control library was added to the libraries, and cluster density was reduced to 80% of regular levels. Base calling and production of demultiplexed FASTQ files were performed by running RTA v1.18.54.4 and bcl2fastq v2.20.0.422.

The demultiplexed FASTQ files were analyzed using QIIME2 version 2022.8 run via a SLURM script. The analysis comprised the following steps:

1. Reads were demultiplexed and adapters trimmed using cutadapt with forward and reverse sequences CCTACGGGN-GGCWGCAG and GACTACHVGGGTATCTAATCC. To remove low-quality calls at the ends of the reads, adapter-trimmed reads were truncated at 250 bp for the forward read and 190 bp for the reverse read.

2. Denoising and amplicon sequence variant (ASV) inference was carried out using DADA2.

3. Taxonomic classification of the ASVs was carried out using the feature-classifier classify-sklearn command with SILVA release 138 as the reference database.

4. A phylogenetic tree was estimated using the align-to-tree-mafft-fasttree command, which generates a MAFFT alignment from reads, and fasttree was used to estimate a phylogenetic tree from these alignments.

5. To remove possible contaminations (taxa present in the negative control water sample), a vector was created with all ASVs except the ones present in the negative control. This vector was then applied to the 'prune_taxa()' function from the R phyloseq package on the 'phyloseq' object.

The QIIME2 analysis SLURM script is available on GitHub at github.com/CBGOUS/.

Proximity network construction

The proximity network was measured using the `SocioPatterns` wireless proximity sensors [30, 31]. These are wearable, unobtrusive devices embedded in a badge worn by the participants using a lanyard. Each sensor emits low-power radio packets at known power on a radio channel within the ISM radio band at 2.4 GHz, at a rate of approximately 40 packets per second. When two sensors exchange at least one information packet, the receiving sensor records the unique identifier of the sender, the interaction time-stamp, and the signal attenuation value. Proximity interactions within approximately 1.5 meters are obtained by filtering the measurements according to their attenuation. Moreover, each sensor periodically records some status properties that are used during data cleaning to determine the correct functioning of the sensor. Among these, an accelerometer detects whether the sensor is moving or not. This signal can be used to filter out contact data collected while the sensor is not moving which is likely because it was not worn. This technology has been extensively tested and used for quantifying human contacts in a variety of contexts, including schools [38, 40], hospitals [43] and African rural villages [47, 46], among others.

Based on a preliminary analysis, we observed a rapid deterioration in the data quality collected by the proximity sensors over successive waves, primarily due to damage. Consequently, we opted to only utilize the initial two measurements as reliable inputs for our analysis. These two measurements show a high correlation in the contact patterns: the Spearman correlation between the weights is $r = 0.25$ with p value $7 \cdot 10^{-14}$. These correlations, paired with the high compliance observed in these measurements, make them reliable data sources.

Acknowledgments We thank the children for their participation in the study, the parents for their invaluable help ensuring that the sensors were worn every day, the rector and teachers for their support and practical assistance in executing the study.

Funding This project was funded by the Norwegian Institute of Public Health and the Molde University College. LD, CC and DP acknowledge the support from the Lagrange project of Fondazione CRT. LD and CC acknowledge the support from Fondation Botnar (EPFL COVID-19 Real Time Epidemiology I-DAIR Pathfinder).

Author contributions CSN and CC designed the study. SMW and ASF supervised the data collection in the field. XB, SR, and AM carried out the microbiota profiling. CC, LD, and DP performed the data analysis. LD, CC, and XB wrote the first version of the manuscript. All authors interpreted the results, contributed to the final version of the manuscript, and approved the final version of the manuscript.

References

[1] Katharine Z Coyte, Jonas Schluter, and Kevin R Foster. The ecology of the microbiome: networks, competition, and stability. *Science*, 350(6261):663–666, 2015.

[2] Barbara A. Methé et al. A framework for human microbiome research. *Nature*, 486:215 – 221, 2012.

[3] Richard J. Lamont, Hyun Koo, and George Hajishengallis. The oral microbiota: dynamic communities and host interactions. *Nature Reviews Microbiology*, 16:745–759, 2018.

[4] Amar Sarkar, Cameron JA McInroy, Siobhán Harty, Aura Raulo, Neil GO Ibata, Mireia Valles-Colomer, Katerina V-A Johnson, Ilana L Brito, Joseph Henrich, Elizabeth A Archie, et al. Microbial transmission in the social microbiome and host health and disease. *Cell*, 187(1):17–43, 2024.

[5] Elizabeth A Archie and Jenny Tung. Social behavior and the microbiome. *Current opinion in behavioral sciences*, 6:28–34, 2015.

[6] Amar Sarkar, Siobhán Harty, Katerina V-A Johnson, Andrew H Moeller, Elizabeth A Archie, Laura D Schell, Rachel N Carmody, Timothy H Clutton-Brock, Robin IM Dunbar, and Philip WJ Burnet. Microbial transmission in animal social networks and the social microbiome. *Nature ecology & evolution*, 4(8):1020–1035, 2020.

[7] Eoin Sherwin, Seth R. Bordenstein, John L. Quinn, Timothy G. Dinan, and John F. Cryan. Microbiota and the social brain. *Science*, 366, 2019.

[8] Katerina V-A Johnson. Gut microbiome composition and diversity are related to human personality traits. *Human Microbiome Journal*, 15:100069, 2020.

[9] Katerina V-A Johnson, Karli K Watson, Robin IM Dunbar, and Philip WJ Burnet. Sociability in a non-captive macaque population is associated with beneficial gut bacteria. *Frontiers in microbiology*, 13:1032495, 2022.

[10] Elizabeth A Grice and Julia A Segre. The human microbiome: our second genome. *Annual review of genomics and human genetics*, 13(1):151–170, 2012.

[11] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[12] J Niels Rosenquist, James H Fowler, and Nicholas A Christakis. Social network determinants of depression. *Molecular psychiatry*, 16(3):273–281, 2011.

[13] Brett B Finlay and CIFAR Humans. Are noncommunicable diseases communicable? *Science*, 367(6475):250–251, 2020.

[14] Catherine D Robinson, Brendan JM Bohannan, and Robert A Britton. Scales of persistence: transmission and the microbiome. *Current Opinion in Microbiology*, 50:42–49, 2019.

[15] Mireia Valles-Colomer, Aitor Blanco-Míguez, Paolo Manghi, Francesco Asnicar, Leonard Dubois, Davide Golzato, Federica Armanini, Fabio Cumbo, Kun D Huang, Serena Manara, et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature*, 614(7946):125–135, 2023.

[16] Aura Raulo, Bryony Allen, Tanya Troitsky, Arild Husby, Josh A. Firth, Tim Coulson, and Sarah C. L. Knowles. Social networks strongly predict the gut microbiota of wild mice. *The ISME Journal*, 15:2601 – 2613, 2020.

[17] Jenny Tung, Luis B Barreiro, Michael B Burns, Jean-Christophe Grenier, Josh Lynch, Laura E Grieneisen, Jeanne Altmann, Susan C Alberts, Ran Blekhman, and Elizabeth A Archie. Social networks predict gut microbiome composition in wild baboons. *elife*, 4:e05224, 2015.

[18] Laura E. Grieneisen, Joshua Livermore, Susan C. Alberts, Jenny Tung, and Elizabeth A. Archie.

Group living and male dispersal predict the core gut microbiome in wild baboons. *Integrative and comparative biology*, 57 4:770–785, 2017.

[19] Andrew H Moeller, Steffen Foerster, Michael L Wilson, Anne E Pusey, Beatrice H Hahn, and Howard Ochman. Social behavior shapes the chimpanzee pan-microbiome. *Science advances*, 2(1):e1500997, 2016.

[20] Amanda C Perofsky, Lauren Ancel Meyers, Laura A Abondano, Anthony Di Fiore, and Rebecca J Lewis. Social groups constrain the spatiotemporal dynamics of wild sifaka gut microbiomes. *Molecular Ecology*, 30(24):6759–6775, 2021.

[21] Simon Lax, Daniel P Smith, Jarrad Hampton-Marcell, Sarah M Owens, Kim M Handley, Nicole M Scott, Sean M Gibbons, Peter Larsen, Benjamin D Shogan, Sophie Weiss, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*, 345(6200):1048–1052, 2014.

[22] Se Jin Song, Christian Lauber, Elizabeth K Costello, Catherine A Lozupone, Gregory Humphrey, Donna Berg-Lyons, J Gregory Caporaso, Dan Knights, Jose C Clemente, Sara Nakielny, et al. Cohabiting family members share microbiota with one another and with their dogs. *elife*, 2:e00458, 2013.

[23] Ashley A. Ross, Andrew Charles Doxey, and Josh D. Neufeld. The skin microbiome of cohabiting couples. *mSystems*, 2, 2017.

[24] Kimberly A Dill-McFarland, Zheng-Zheng Tang, Julia H Kemis, Robert L Kerby, Guanhua Chen, Alberto Palloni, Thomas Sorenson, Federico E Rey, and Pamela Herd. Close social relationships correlate with human gut microbiota composition. *Scientific reports*, 9(1):703, 2019.

[25] Federico Musciotto, Begoña Dobon, Michael Greenacre, Alex Mira, Nikhil Chaudhary, Gul Deniz Salali, Pascale Gerbault, Rodolph Schlaepfer, Leonora H Astete, Marilyn Ngales, et al. Agta hunter–gatherer oral microbiomes are shaped by contact network structure. *Evolutionary Human Sciences*, 5:e9, 2023.

[26] Ilana Lauren Brito, Thomas Gurry, Shijie Zhao, Katherine H. Huang, Sarah K. Young, Terrance Shea, Waisea Naisilisili, Aaron P. Jenkins, Stacy D. Jupiter, Dirk Gevers, and Eric J. Alm. Transmission of human-associated microbiota along family and social networks. *Nature Microbiology*, 4:964 – 971, 2019.

[27] Francesco Beghini, Jackson Pullman, Marcus Alexander, Shivkumar Vishnempet Shridhar, Drew Prinster, Adarsh Singh, Rigoberto Matute Juárez, Edoardo M Airoldi, Ilana L Brito, and Nicholas A Christakis. Gut microbiome strain-sharing within isolated village social networks. *Nature*, pages 1–9, 2024.

[28] Timo Smieszek, E U Burri, Ryan Scherzinger, and Roland W. Scholz. Collecting close-contact social mixing data with contact diaries: reporting errors and biases. *Epidemiology and Infection*, 140:744 – 752, 2011.

[29] Rossana Mastrandrea and Alain Barrat. How to estimate epidemic risk from incomplete contact diaries data? *PLoS Computational Biology*, 12, 2016.

[30] sociopatterns.org. Official site of the `SocioPatterns` organization.

[31] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLoS ONE*, 5, 2008.

[32] Anders Esberg, Linda Eriksson, and Ingegerd Johansson. Site-and time-dependent compositional shifts in oral microbiota communities. *Frontiers in oral health*, 3:826996, 2022.

[33] Thomas SB Schmidt, Matthew R Hayward, Luis P Coelho, Simone S Li, Paul I Costea, Anita Y Voigt, Jakob Wirbel, Oleksandr M Maistrenko, Renato JC Alves, Emma Bergsten, et al. Extensive transmission of microbes along the gastrointestinal tract. *Elife*, 8:e42693, 2019.

[34] Tim Sainburg, Leland McInnes, and Timothy Q Gentner. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907, 2021.

[35] Huang Lin and Shyamal Das Peddada. Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ biofilms and microbiomes*, 6(1):60, 2020.

[36] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235, 2005.

[37] Steven N Steinway, Matthew B Biggs, Thomas P Loughran Jr, Jason A Papin, and Reka Albert. Inference of network dynamics and metabolic interactions in the gut microbiome. *PLoS computational biology*, 11(6):e1004338, 2015.

[38] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter van den Broeck, Corinne Régis, Bruno Lina, and Philippe Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6, 2011.

[39] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS one*, 10(9):e0136497, 2015.

[40] Julie Fournet and Alain Barrat. Contact patterns among high school students. *PLoS ONE*, 9, 2014.

[41] Mathieu Génois, Christian L Vestergaard, Julie Fournet, André Panisson, Isabelle Bonmarin, and Alain Barrat. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science*, 3(3):326–347, 2015.

[42] Mathieu Génois and Alain Barrat. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*, 7(1):1–18, 2018.

[43] Philippe Vanhems, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul a Kim, Brigitte Comte, and Nicolas Voirin. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE*, 8, 2013.

[44] Moses C Kiti, Michele Tizzoni, Timothy M Kinyanjui, Dorothy C Koech, Patrick K Munywoki, Milosch Meriac, Luca Cappa, André Panisson, Alain Barrat, Ciro Cattuto, et al. Quantifying social contacts in a household setting of rural kenya using wearable proximity sensors. *EPJ data science*, 5:1–21, 2016.

[45] Laura Ozella, Daniela Paolotti, Guilherme Lichand, Jorge P Rodríguez, Simon Haenni, John Phuka, Onicio B Leal-Neto, and Ciro Cattuto. Using wearable proximity sensors to characterize social contact patterns in a village of rural malawi. *EPJ Data Science*, 10(1):46, 2021.

[46] Lorenzo Dall'Amico, Jackie Kleynhans, Laetitia Gauvin, Michele Tizzoni, Laura Ozella, Mvuyo Makhasi, Nicole Wolter, Brigitte Language, Ryan G. Wagner, Cheryl Cohen, Stefano Tempia, and Ciro Cattuto. Estimating household contact matrices structure from easily collectable metadata. *PLOS ONE*, 19(3):1–13, 03 2024.

[47] Jackie Kleynhans, SA-S-HTS Group, Lorenzo Dall'Amico, Laetitia Gauvin, M. Tizzoni, Lucia Maloma, S. Walaza, Neil Alexander Martinson, Anne von Gottberg, Nicole L. Wolter, Mvuyo Makhasi, Cheryl Cohen, Carlo Cattuto, and Stefano Tempia. Association of close-range contact patterns with sars-cov-2: a household transmission study. *eLife*, 12, 2023.

[48] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[49] Anna Klindworth, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. Evaluation of general 16s ribosomal rna gene pcr primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research*, 41(1):e1–e1, 2013.

Supplementary information

S.1 Analysis of microbial alpha diversity

Figure S.1 shows the microbial alpha diversity in the five measurements analyzed in the main text and shown in Figure 1D. Using different metrics (number of observed ASVs, Shannon, Fisher, and Simpson diversity indices), we show that the alpha microbial diversity remains stable over the five measurements. No pair of measurements has highly significant (Mann-Whitney $p < 0.05$ with Bonferroni correction) differences in the distributions, according to any diversity metric.

S.2 Inference of the proximity network from the microbiota with a random forest classifier

We follow the procedure described in **Inference of the proximity network from the microbiota** and use a random forest classifier instead of a support vector machine to test the robustness of our result. The test is conducted with the same parameters, and it only differs in the use of the classifier. Figure S.2 shows the ROC curve obtained for 100 random splits of the train and test sets. The average AUC equals 0.74.

S.3 Correlation between proximity and bacterial sharing network

We investigate the correlation between the social network $\mathcal{G}_{\mathrm{p}}$, measured using the proximity sensors, and the bacterial sharing network. We recall that,

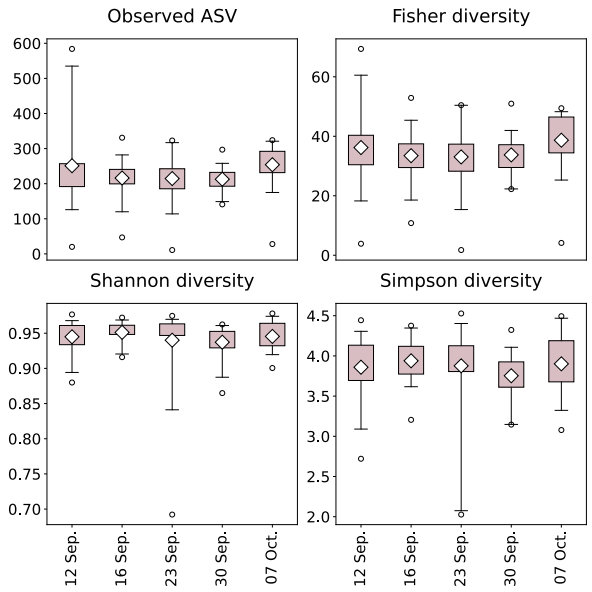

Figure S.1: **Alpha microbial diversity.** Boxplot of the alpha diversity according to four different diversity measures indicated in the frames' titles. The x-axis spans the five measurements considered in the main text. The whiskers indicate the $1-99$ confidence interval, while the white diamond is the sample average.

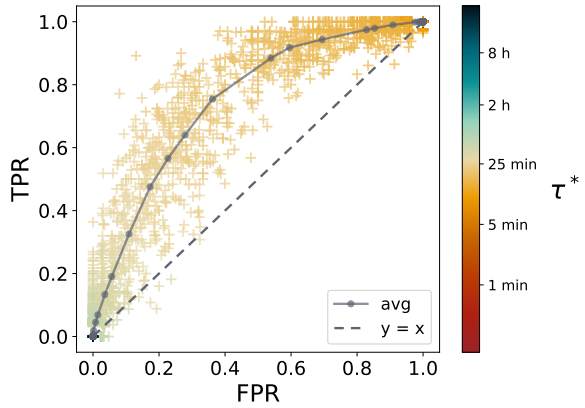

Figure S.2: **ROC curve for the random forest binary classifier used to infer the proximity links based on microbiota profiles.** We follow the procedure detailed for Figure 3 using a random forest classifier to predict whether the weight (duration in proximity) of a given edge (ij) exceeds a threshold τ^*. The results are obtained for 100 random split of the training/test sets, for each of the 250 logarithmically spaced values of τ^* we scan. Every cross in the plot corresponds to one realization of the training and test sets. The x coordinate is the false positive rate (FPR) and the y coordinate is the true positive rate (TPR). Markers are color-coded according to the value of the threshold τ^*. The solid gray line represents the average over the random splits and has an area under the curve auc $= 0.74$. The dashed line is the curve $y = x$ for a random-guess binary classifier.

for this network, the weights of an edge (ij) are

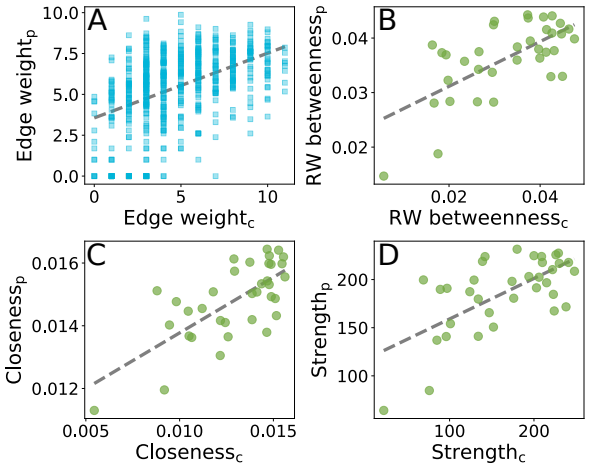

Figure S.3: **Correlation between the proximity network and the bacterial sharing networks.** *Panel A*: correlation between the edge weights in the two networks. *Panels B, C, D*: correlation of different node centrality measures (random walk betweenness, closeness, and strength) across the two networks. The subscript "c" stands for *co-sharing*, while the subscript "p" for *proximity*.

defined as the number of taxa appearing in Table 1 shared by the nodes i and j. Figure S.3A shows that the edge weights of the two networks are highly correlated (Spearman $r = 0.44$ and p value below machine precision). The subsequent panels complement Figure 5 by considering alternative node centrality measures for which we obtain strong across the two networks: random walk betweenness (Spearman $r = 0.51$, p value $= 1.4 \cdot 10^{-3}$), closeness (Spearman $r = 0.51$, p value $= 1.7 \cdot 10^{-3}$), and strength (Spearman $r = 0.51$, p value $= 1.7 \cdot 10^{-3}$).