Signatures of global processes shaping the structure of microbial co-occurrence networks

Geut Galai [ID] [1], Dafna Arbel[1], Keren Klass[2], Ido Grinshpen [ID] [3,1], Itzik Mizrahi [ID] [3,1], and Shai Pilosof [ID] [4,1,*]

[1]*Department of Life Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel*
[2]*The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel*
[3]*National Institute of Biotechnology in the Negev, Ben-Gurion University of the Negev, Be'er Sheva, Israel*
[4]*The Goldman Sonnenfeldt School of Sustainability and Climate Change, Ben-Gurion University of the Negev, Be'er Sheva, Israel*
[*]*Corresponding author: pilos@bgu.ac.il*

Abstract

Co-occurrence networks offer insights into the complexity of microbial interactions, particularly in highly diverse environments where direct observation is challenging. However, identifying the scale at which local and non-local processes structure co-occurrence networks remains challenging because it requires simultaneously analyzing network structure within and between local networks. In this context, the rumen microbiome is an excellent model system because each cow contains a physically confined microbial community, which is imperative for the host's livelihood and productivity. Employing the rumen microbiome of 1,012 cows across seven European farms as our model system, we constructed and analyzed farm-level co-occurrence networks to reveal underlying microbial interaction patterns. Within each farm, microbe co-occurrence was significantly transitive with a group structure that reflected functional equivalence in co-occurrence. Knowing the group composition (obtained with stochastic block modeling) in one farm provided significantly more information on the grouping in another farm than expected. Moreover, microbes strongly conserved co-occurrence patterns across farms (also adjusted for phylogeny). We further developed a meta-co-occurrence multilayer approach, which links farm-level networks, to test scale signatures simultaneously at the farm and inter-farm levels. Consistent with the comparison between groups, the multilayer network was not partitioned into clusters. This result was consistent even when artificially disconnecting farm-level networks. Our results show a prominent signal of processes operating across farms to generate a non-random yet similar co-occurrence patterns. Comprehending the processes underlying rumen microbiome assembly can aid in developing strategies for its manipulation to mitigate greenhouse gas emissions. More broadly, our results provide new evidence for the scale at which forces shape microbe co-occurrence. Finally, the hypotheses-based approach and methods we developed provide can be adopted in other systems to detect scale signatures in species interactions.

Introduction

The composition and interaction structure (who interacts with whom) of microbial communities determine their stability and function [1–3]. For example, different gut microbiome compositions are associated with animals' digestion [4,5]. Therefore, a central goal of microbial ecology is to understand the mechanisms underlying microbiome composition and structure. However, it is highly challenging in microbial systems to observe interactions (e.g., competition, cross-feeding) when diversity is high under natural conditions. Hence, researchers use co-occurrence networks to describe the structure of microbial communities. While co-occurrence is not direct evidence for interactions, it is a prerequisite [6–9]. In a co-occurrence network, a link represents a statistically significant association across space (or time). Such associations emanate from three primary processes: dispersal limitation, environmental filtering, and biotic interactions [10–12]. Therefore, the structure of microbial co-occurrence networks is governed by a mixture of processes operating at different scales, as is also true for other network types such as food webs [13]. Identifying the signatures of local vs. non-local processes that determine the structure of co-occurrence networks remains a challenge and is the primary goal of this work.

At local scales, co-occurrence depends on direct and indirect interactions with other microbes. For instance, competition is a selective force that occurs at a local scale because it typically requires physical proximity [14]. Cross-feeding networks in which microbes depend on each other directly or indirectly via the production and consumption of metabolites is another key example [15]. At the non-local scale, environmental filtering affects local microbe composition [16,17]. For instance, physical-chemical conditions [18,19] and temperature [17]. Therefore, co-occurrence patterns can vary in space, for example, due to variability in environmental conditions or host characteristics [18,19]. Moreover, local communities exchange species with a 'regional' species pool and are interconnected via colonization-extinction dynamics due to microbe dispersal, generating meta-communities [20,21]. Ecological networks, including co-occurrence networks, contain signatures—that is, observable patterns indicative of the processes and factors that generated the network [22,23].

To investigate the scale at which signatures of co-occurrence are detected, we used the highly complex microbial ecosystem of the rumen microbiome [24]. The rumen ecosystem is a confined abiotic microbial environment. Therefore, it is particularly advantageous because it allows for a focused examination of interactions within this ecosystem, reducing biases resulting from a mismatch between spatial scales of analysis and interactions [9]. Moreover, the rumen ecosystem's considerable importance to food security drives comprehensive monitoring of on-farm parameters, including diet composition and occasionally even host genetics, which enables us to establish connections between ecological patterns and their potential determinants. We chose a rumen microbiome data set of 1,012 cows from seven European farms of different localities that were monitored for diet and host genetics [25]. We focused our analysis on the 'common core' microbiome—a subset of microbes that occur in a minimum proportion of hosts [26,27]. Using core microbes allowed us to focus on those more likely to influence other microbes and drive co-occurrence patterns [28]. Core microbes are also associated with cow phenotypes (e.g., methane emissions and dairy production) [25].
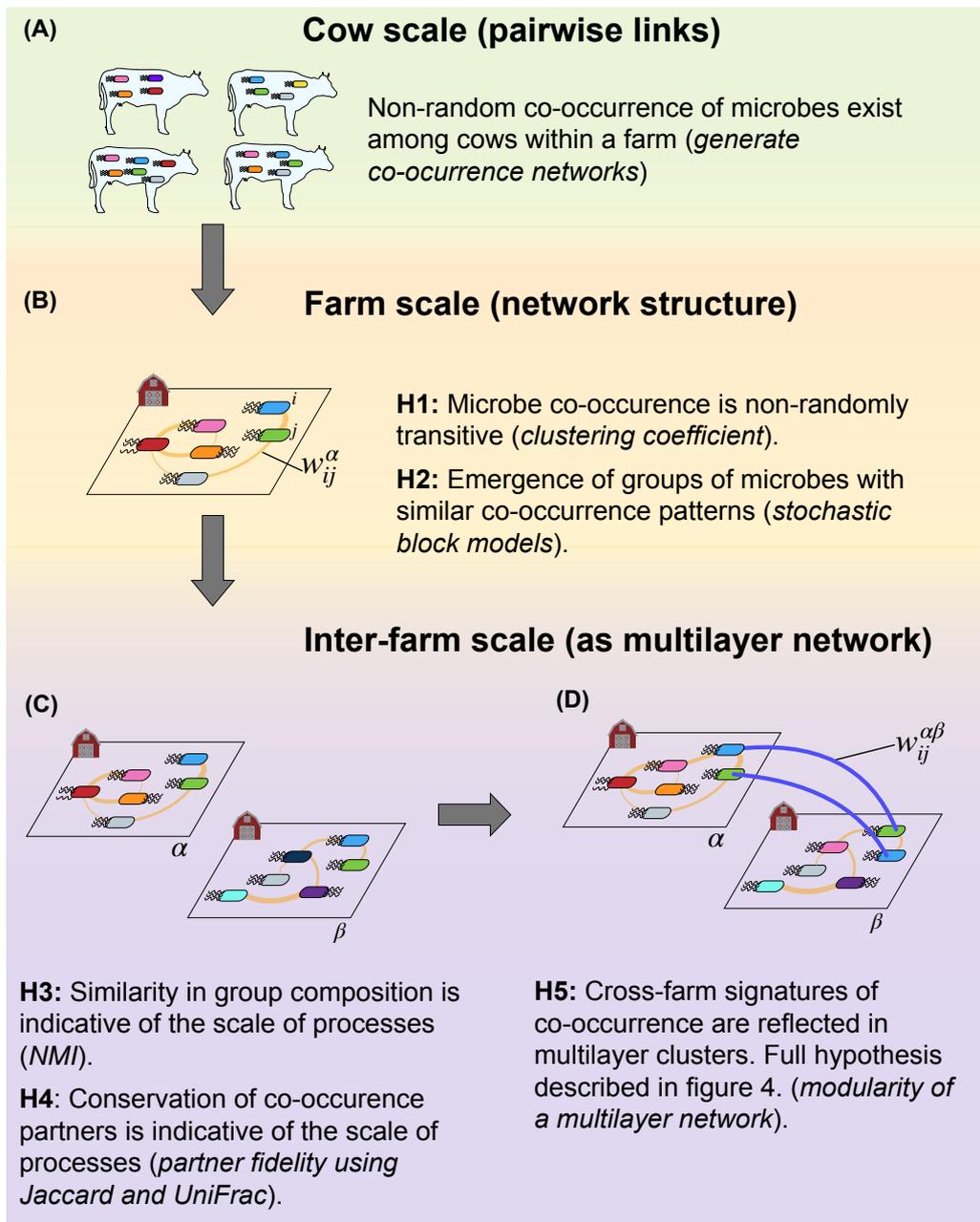
The advantage of using networks is the ability to go beyond pairwise co-occurrence because microbe occurrence and co-occurrence involve indirect effects. We developed a set of hypotheses to detect structural signatures of scale (Fig. 1). One approach is to analyze local co-occurrence networks separately

and compare them [19,29]. We hypothesized that within each farm, there would be a non-random transitive signature such that if microbe i co-occurred with j and j with k, then microbe i would also co-occur with k (hypothesis **H1**). Non-random transitivity would indicate that from a microbe perspective, co-occurrence structure is driven by either similar environmental conditions (i, j and k occupy the same niche) or interaction dependence [14]. For example, rock-paper-scissors competition dynamics can lead to the coexistence of three species [30].

Transitivity is a well-known measure of clustering in a node's neighborhood [31]. Such "neighborhood" clustering can lead to the emergence of groups of microbes that interact with microbes that themselves interact with similar microbes (**H2**). We tested for this network pattern using a stochastic block model (SBM). Stochastic block models aim to identify nodes with similar patterns of connections, suggesting they serve similar functions or roles within the network [32,33]. They can, therefore, be used to determine functional groups in ecological networks [32,34], including in microbial networks [35], although this has been rarely done. In our study system, microbes from the same group have a higher probability of co-occurring with each other than with microbes from other groups. By comparing the composition of groups across farms, we can gain insights into the scale at which co-occurrence is determined. If group composition is similar across farms, then similar processes likely operate in those farms (**H3**). Moreover, if similar processes operate at different farms, microbes will likely conserve co-occurrence partners across farms more than expected by chance. Therefore, co-occurrence links will be repeatable across farms (**H4**). We tested this hypothesis using partner fidelity [36], which measures how similar a microbe's partners are across farms.

Comparing farm-level networks treats these networks as separate, disconnected communities. Nevertheless, microbe co-occurrence is governed by various processes operating at different scales. Hence, it is necessary to simultaneously analyze network structure within and between farm-level networks. For that, we need a mathematical object that separates farm networks on the one hand and connects them on the other. Multilayer networks [37] are a powerful tool that can represent multiple interlinked co-occurrence networks, effectively creating a meta-co-occurrence network. This framework allowed us to investigate, for the first time, local and non-local network signatures of co-occurrence simultaneously (Fig. 1D, Methods). We analyzed this multilayer network using modularity to quantify at which scale—farm, country, or regional (across countries)—the partition of microbes to modules (clusters of nodes) was prominent.

Across scales, we detected a prominent signal of forces operating across farms to generate a non-random yet similar pattern of co-occurrence.

(A) Cow scale (pairwise links)

Non-random co-occurrence of microbes exist among cows within a farm (*generate co-ocurrence networks*)

(B) Farm scale (network structure)

w_{ij}^{α}

H1: Microbe co-occurence is non-randomly transitive (*clustering coefficient*).

H2: Emergence of groups of microbes with similar co-occurrence patterns (*stochastic block models*).

Inter-farm scale (as multilayer network)

(C)

α

β

(D)

$w_{ij}^{\alpha\beta}$

α

β

H3: Similarity in group composition is indicative of the scale of processes (*NMI*).

H4: Conservation of co-occurence partners is indicative of the scale of processes (*partner fidelity using Jaccard and UniFrac*).

H5: Cross-farm signatures of co-occurrence are reflected in multilayer clusters. Full hypothesis described in figure 4. (*modularity of a multilayer network*).

Fig. 1: Overview of study scales, hypotheses, and analyses. The figures summarize the hypotheses and the methods used to test them (italic text in parentheses). **(A)** Microbes occur in cows. The four cows in this toy example belong to the same farm. We calculated the probability that two microbes will co-occur across the cows within each farm. We connected microbe pairs whenever their co-occurrence was statistically significantly greater than their co-occurrence based on their prevalence alone. The result is a microbial co-occurrence network (cows are not part of the network) for each farm as in panel **(B)**. **(B)** The farm-level networks allow testing patterns that transcend pairwise interactions within each farm, such as transitivity (H1) and grouping (H2). **(C)** At the inter-farm scale, we compared farm-level networks to gain insights into the scale at which processes determine the structure (H3 and H4). **(D)** We also created a meta-co-occurrence network using a multilayer network approach. In this network, layers (depicted as polygons) are farm-level networks. Interlayer links ($w_{ij}^{\alpha\beta}$; blue) connect between microbes i and j that have a co-occurrence link in both farms α and β. Here, the blue and green microbes co-occur in cows on both farms. In contrast, the blue and pink occur in both farms but do not co-occur in both and are therefore disconnected between farms. We used this network to detect scale signatures while simultaneously considering within and between farm processes (see Fig. 4 for a complete set of hypotheses).

Methods

Data

We used data from Wallace et al. [25]. Rumen microbiome samples were collected from 1012 cows. The cows were divided into two cow breeds located in four different countries in Europe: 816 Holstein dairy cows in the United Kingdom (farms UK1 and UK2) and Italy (farms IT1, IT2, and IT3), and 196 Nordic Red dairy cows in Finland (farm FI1) and Sweden (farm SE1). Cows were fed highly controlled diets. The Holsteins received a maize silage–based diet, while the Nordic Reds received a nutritionally equivalent diet based on grass silage as forage (Table S1). In the original study [25], the farmers did not reveal the farm identity or location. Hence, the regions the farms are located in are available, but not their exact coordinates. This does not affect our study because the signatures we obtained are not related to distance. Unlike wildlife, cows and the microbes within them do not disperse freely across space.

Each rumen sample was analyzed for paired amplicon reads [25]. The paired reads were then run through DADA2 using default parameters as pre-merged reads. The taxonomic assignments for the reads were performed using the SILVA database (silva_nr_v138). Because the data was for sequences of the V4 region, only Prokaryotes and Archaea are included. The complete data set included 12,109 microbe ASVs (amplicon sequence variants), clustered at 100% similarity. The microbiome samples were composed primarily of bacteria (99%), and the rest were archaea, or unknown.

We used the 'common core' microbes in each farm [26,27] to focus on the signal produced by the key microbes. Wallace et al. [25] defined the common core microbes as those occurring at $\geq 50\%$ of cows within a farm, remaining with 454 microbes. At this strict threshold, the core microbes strongly predicted the cow phenotype (e.g., methane emission, lactation). Therefore, this can be considered as the 'functional core' [38]. Nevertheless, a 50% threshold is stringent as it includes only 3.7% of ASVs, which omits variation in occurrence that can influence signatures in the meta-co-occurrence network. We opted not to deviate drastically from the functional core because it is biologically meaningful. Hence, we focused our analysis on microbes occurring in at least 30% of the cows, which removes biases due to rare species. We also tried other thresholds (except for SBM analysis, which is computationally impossible when there are too many nodes). Overall, the sensitivity analysis shows that the patterns we obtained in the 30% threshold provides a good balance between a biologically meaningful set of microbes and the ability to discern the scale of the processes that define it (see details in SI note 1).

Construction of farm-level co-occurrence networks

Within each farm, we constructed a co-occurrence network using a combinatorial model that provides the probability that two ASVs (nodes in the network) significantly positively or negatively co-occur [39] (implemented with R package cooccur v1.3 [40]) (Fig. 1A-B). The combinatorial model was solely used to determine co-occurrence links. That is, a link between ASV i and j in farm α could exist (value 1) or not (value 0): $w_{ij}^\alpha \in \{0, 1\}$. Statistical significance of co-occurrence was determined per ASV-pair by calculating the combinatorial probability that the two ASVs will randomly co-occur with a probability equal to, or more than, observed in the data. The calculation considers their prevalence in cows and the number of cows. If this probability is smaller than 5%, we considered an intralayer co-occurrence link ($w_{ij}^\alpha = 1$, Fig. 1B). Co-occurrence is calculated across cows, so these are not included

5

in the resulting network.

We note that there are several methods to detect co-occurrence in microbial communities, each with its advantages and disadvantages [7]. In reality, no method will detect the 'true' co-occurrence. We used Veech's method [39] for three reasons. First, it has a 'built-in' null hypothesis whereby p values are calculated using a distribution-free probabilistic model [39]. It, therefore, avoids potential biases due to randomization-based null models [41]. Second, our data were binary (presence/absence of a microbe in cows). Third, it has been widely used and found to be a robust method, including in studies of microbial communities [42,43].

We tested for non-random co-occurrence patterns by creating 500 shuffled networks for each farm. We randomly distributed the microbes between cows within each farm, conserving the number of microbes a cow hosts. This is a conservative algorithm because allowing the number of microbes per cow to vary would make it easier to detect statistically significant results (increasing false positive rates) [44]. This strengthens our results, ensuring that the non-random patterns revealed are statistically strong and biologically meaningful. After redistributing microbes, we recalculated the farm's co-occurrence network. This process was repeated 500 times per farm, resulting in 3,500 shuffled networks.

Clustering coefficient

We tested transitivity in co-occurrence by calculating the clustering coefficient. The clustering coefficient of a microbe i, CC_i, is defined as the number of closed triangles (three connected nodes, including i) out of all possible triangles: $CC_i = 2N_i/(k_i(k_i - 1))$, where N_i is the total number of links between i's neighbors and k_i is the degree (number of links) of i. If none of the co-occurring partners of a focal microbe co-occur with one another, CC_i will be 0. In contrast, if all the neighbors of a microbe co-occur among themselves, CC_i will be 1. Because CC_i can be biased when the degree is low (it is easier to close triangles), we only calculated CC for ASVs with ten or more partners [45]. An alternative analysis would be to calculate if the probability of three microbes to co-occur is larger than the joint probabilities of each to occur. Nevertheless, this analysis is conceptually different than transitivity because it does not test triangle closure. It is also computationally effectively impossible as there are too many microbe trios.

Stochastic block models

We tested for the emergence of groups of microbes with similar co-occurrence patterns (H2) using an SBM analysis on each farm. Stochastic block models identify nodes within a network with a similar connectivity profile. Accordingly, nodes are "stochastically equivalent" if they share the same connection patterns to other nodes [33,46]. In the standard SBM, nodes are divided into K groups with an affinity matrix, denoted as Ω, which encodes the probabilities of links within and between groups. Consequently, the probability that a link exists between two nodes, i and j, which belong to groups c_x and c_y respectively, is $P_{ij} = \Omega_{c_x c_y}$. This arrangement implies that nodes within the same group have identical probabilities of forming connections with nodes from any other group.

The primary objective of the SBM algorithm is to uncover these hidden groupings within the network. The algorithm determines the model parameters that most accurately represent the observed network structure. It seeks the grouping configuration most likely to have produced the observed data. The algorithm partitions the network to K groups and calculates the likelihood of such clustering, con-

sidering the membership of nodes in groups. The best partition minimizes the AIC score (calculated based on the likelihood and number of model parameters). We performed the SBM analysis with the sbm R package (v0.4.5) [47].

Normalized mutual information

We used a measure of Normalized Mutual Information (NMI) [48], based on information theory, to examine the differences in SBM group composition between each pair of farms. The index quantifies, for nodes that appear in both farms, the amount of information that group composition in one farm provides to deduce the group composition in the other. The NMI score is based on defining a confusion matrix $\mathbf{N}$, in which the rows correspond to the groups in one farm (farm A), and the columns correspond to those in the other farm (farm B). The element of $\mathbf{N}$, N_{xy}, is the number of nodes in the group x from farm A that appear in a group y from farm B. A measure of similarity between the partitions, based on information theory, is defined as:

$$I(A,B) = \frac{-2 \sum_{x=1}^{c_A} \sum_{y=1}^{c_B} N_{xy} \log(N_{xy} N / N_{x.} N_{.y})}{\sum_{x=1}^{c_A} N_{x.} \log(N_{x.}/N) + \sum_{y=1}^{c_B} N_{.y} \log(N_{.y}/N)} \tag{1}$$

Here, N is the total number of nodes compared (only those that occur in both farms are included), the number of groups in farm A is denoted c_A and the number of groups in farm B is denoted c_B, the sum over row x of the matrix $\mathbf{N}$ is denoted $N_{x.}$, and the sum over column y is denoted $N_{.y}$. If the two farm networks are partitioned into identical group compositions, the index takes its maximum value of 1; that is, knowing the assignment of nodes to groups in one farm provides perfect information on knowing the other. If the partitions in the two farms have no common assignments, or one of the farms is partitioned into a single group, then the value is 0.

To test if our NMI scores are different than what is expected by a random grouping, we rearranged the membership of the nodes within each farm without altering the number of groups and amount of microbes in each group. We then recalculated the NMI between each pair of rearranged farms and compared the observed NMI value to that obtained randomly. We performed NMI analysis using the NMI R package (v2.0) [49].

Partner fidelity

Partner fidelity captures the tendency of an ASV to co-occur with the same ASVs in different layers (farms) (Fig. S1). Mathematically, this translates to how similar the partners of a focal ASV are across layers [36]. For each focal ASV i, we created a table in which columns are the farms in which the ASV occurs, and rows are the focal ASVs co-occurring partners within each farm (see diagram in Fig. S1). Table entries were 1 if a partner occurred in a farm or 0 otherwise. We used the table to calculate the Jaccard similarity in partners between farms (performed with R package vegan v2.5-7; [50]). We used the mean of the resulting similarity matrix as a measure of partner fidelity that we termed P_i^J. This index ranges from 0 to 1 when the identities of co-occurring partners of a microbe i are completely discordant or exactly the same, respectively.

However, some microbes are highly similar in their sequence and can effectively be phylogenetically close strains. Therefore, microbe identity (the ASV label in the data set) may not reveal the complete picture. To account for phylogeny, we calculated partner fidelity as we did for P_j^U, using UniFrac [51]

instead of Jaccard. We used the R package GUniFrac version 1.4 [52]. We termed this index P_i^U. UniFrac is a similarity index that considers the degree of divergence between ASV sequences instead of ASV identity. If the lineages of a microbe's partners in different farms are distinct (no branches shared), P_i^U approximates 0; in contrast, if they share the maximum possible branch length, P_i^U is 1.

For UniFrac, we first constructed a phylogenetic tree. We used the genetic sequences to reconstruct an initial phylogenetic tree using a neighbor-joining method (using the R package ape v5.6-2; [53]). Specifically, the conversion from sequence to distance is calculated according to the Tamura and Nei model [54]. The likelihood of this initial tree was computed given the original data, and its parameters were optimized (using the R package phangorn v2.8-1; [55]). The significance of the fitting was tested using ANOVA and AIC (R package stats v4.1.2 [56]).

Calculating significance using z-scores

We tested for non-random patterns in transitivity by comparing CC_i of each ASV i to that obtained for the same ASV in the layer across 500 shuffled networks (see Methods section on constructing networks) using a z-score, following [36]. Z-scores are calculated as

$$z_i = \frac{CC_i^{empirical} - \overline{CC}_i^{shuffled}}{\sigma^{shuffled}}, \tag{2}$$

where $\overline{CC}_i^{shuffled}$ and $\sigma^{shuffled}$ are the mean and standard deviation of the transitivity obtained from the 500 shuffled networks. Hence, a positive z-score suggests that transitivity is higher than expected if microbes were randomly distributed across cows, while a negative one suggests it is lower. The significance of each empirical value was determined at the 0.05 level: a z-score > 1.96 or < -1.96 indicates that CC_i is greater or lower than the random expectation, respectively.

Network shuffling also alters the partners of each microbe in a farm and, therefore, serves as an adequate null model for testing partner fidelity. The calculation of z-scores and significance for P_i^J and P_i^U was done exactly the same as for CC_i.

Construction of a meta-co-occurrence multilayer network

For the modularity analysis, we connected the farm-level co-occurrence networks to form a meta-co-occurrence multilayer network (Fig. 1D). Interlayer links are the mathematical way to encode hypotheses related to inter-layer processes. This approach has been applied multiple times in ecological multilayer networks and is the major advantage of using multilayer networks [37,57–61]. We want the interlayer link definition to encode the extent to which processes that determine co-occurrence operate across layers. We therefore connected a microbe i in layer α with microbe j in layer β if i and j significantly co-occur in cows in both layers ($w_{ij}^{\alpha\beta} = 1$; Fig. 1D). This definition is closely related to partner fidelity because it is not enough that the microbes occur in two farms; they also need to maintain their co-occurrence in both to be linked.

Modularity analysis

We used Infomap via the R package 'infomapecology' [59] to capture the modular structure of the meta-co-occurrence multilayer network. Infomap is a flow-based algorithm designed explicitly for multilayer networks. The advantage of using an algorithm dedicated to multilayer networks is that

it allows the detection of modules that cross layers [37,59] and, therefore, contain microbes from different farms. Hence, a module is a group of ASVs that co-occur more with each other than with other ASVs within and across farms. Briefly, Infomap detects the optimal network partition based on the movement of a random walker on the network (see [59,62,63] for details). For any given network partition, the random walker moves across nodes in proportion to the weight of the edges. The amount of information it costs to describe the walk is quantified using the objective function L called the map equation. The optimal network partition is the one that minimizes L [62]. In multilayer networks, nodes representing observable entities such as ASVs are called *physical nodes*, and nodes describing the occurrence of ASVs in the layers are called *state nodes*. The random walker moves from state node to state node within and across the layers on intra- and inter-layer links, respectively. Therefore, the multilayer network representation is not merely an extended network with unique nodes in all layers, and a module can encompass multiple farms. In addition, a physical node (ASV) can be assigned to different modules in different layers.

Infomap is particularly appropriate for testing our hypothesis because the next step in a random walk is determined probabilistically by the number of links a node has. Therefore, if a node has many intralayer (co-occurrence) links and only a few interlayer (partner fidelity) links, the next step will be inside the layer, generating local partitions, which are a signature of farm-level processes. In contrast, if many nodes have more or less equal numbers of intra- and interlayer links, the modules are more likely to encompass more than one farm, which is a signature of inter-farm level processes.

Code and data

All data and code are available in the GitHub repository https://github.com/Ecological-Complexity-Lab/rumen_microbiome_structure.

Results

Microbe diversity

Our threshold criteria resulted in 946 microbes (931 bacteria and 15 archaea) across the 1012 cows. The vast majority belonged to the Bacteroidetes (52.7%) and Firmicutes (27.8%) phyla. At the family level, microbes belonged mostly to Prevotellaceae (37.6%), followed by Lachnospiraceae (12.9%) (Fig. S2). The number of microbes ranged from 42 to 505 per cow across farms (mean±sd: 269.7±104; Figs. S3, S4). Farms varied in the number of microbes and cows they contained (Table 1). In addition, the occurrence of microbes in farms and cows was skewed, with most microbes occurring in a few farms and cows and a tail of microbes that occur across all the meta-community (Fig. 2). Hence, in terms of beta-diversity, farms shared microbes, but the maximum Jaccard similarity was 0.61 (Fig. S5)

Co-occurrence farm networks are non-randomly structured (H1 and H2)

The co-occurrence structure of each farm network differed from shuffled networks generated by randomizing microbe occurrence in cows (Fig. S6). Therefore, biological processes determine microbe co-occurrence structure at each farm, which we investigated further (H1 and H2; Fig. 1B). The mean CC_i ranged from 0.497-0.8 in the seven farms and was significantly larger than random for the vast majority of microbes in all farms (Fig. 3A). The farm co-occurrence networks had a block structure containing 6-20 groups (Table 1, Fig. S7).

Table 1: Summary statistics of farms. a: number of cows; b: number of microbial ASVs (richness); c: mean number [range] of ASVs per cow; d: number of co-occurrence links. Density is calculated as the number of links divided by the number of possible links: $2L/(S(S-1))$. K is the number of groups detected in the stochastic block models analysis.

Farm	Cows[a]	ASVs(S)[b]	ASVs per cow[c]	Links[d](L)	Density	K
FI1	100	329	185.8 [88-300]	23298	0.49	11
SE1	96	461	259.4 [42-389]	36783	0.35	15
IT1	185	307	179.8 [89-292]	20465	0.45	12
IT2	176	566	349.6 [129-505]	63310	0.42	20
IT3	48	217	125.8 [73-188]	2776	0.13	6
UK1	243	490	280.3 [100-442]	71101	0.61	20
UK2	164	630	369.1 [145-498]	51551	0.27	17

Microbes tend to conserve partners and co-occurrence links across farms (H3 and H4)

We compared the assignment of microbes into groups between pairs of farms (H3; Fig. 1C). The NMI ranged from 0.217 (between farms FI1 and IT3) to 0.579 (between farms IT2 and UK1) (Fig. 3B). Moreover, all the NMI pairwise comparisons were statistically significantly higher than a random group assignment of microbes (Fig. S8). Each NMI comparison only considers microbes that occur on both farms. Therefore, despite overall topological differences between the farm networks (Fig. S6), microbes assigned to the same group in one farm tend to be assigned to the same group in another.

Non-random transitivity and similar grouping patterns indicate that microbes maintain similar co-occurrence partners across farms. We tested this hypothesis (H4; Fig. 1C) by calculating partner fidelity for the 674 out of 946 microbes ($\approx 70\%$) that occurred in more than one farm (Fig 2B). We first used Jaccard similarity as a partner fidelity score (P_i^J). The mean P_i^J across the microbes that occurred in two or more farms was 0.19 ± 0.08 (mean±sd) with a maximum of ≈ 0.4 (Fig. 3C). In 87% of microbes, P_i^J was higher than expected from a random distribution of microbes across cows. Considering phylogeny (using P_i^U) increased partner fidelity. The mean P_i^U score was 0.48 ± 0.08. Comparison to shuffled networks showed that almost 50% of microbes (284 out of 649) tended to have smaller genetic distances between their co-occurring partners across farms than expected at random (Fig. 3C). The higher values and lower proportion of significant partner fidelity in P_i^U are expected because the probability of conserving the same ASV labels (used in P_i^J) is low due to the high number of ASVs. The qualitative pattern of partner fidelity was similar when using the taxonomic classification of ASVs; for example, microbes tended to conserve the genera with which they interact (Fig. S9).

Meta-co-occurrence network analysis shows signatures of global processes (H5)

We used modularity analysis to detect clusters (modules) of highly connected microbes within and between farms. Our alternative hypotheses are as follows. If local processes that determine co-occurrence dominate, then co-occurrence links within a farm will be dense compared to the conservation of partners in other farms. At the extreme, microbes will not conserve partners and, therefore, will not be linked to any microbe between farms. In that case, farms will be disconnected, and the modules will correspond to farms (H5a, Fig. 4A). At the other extreme, if similar processes operate in all the farms, then microbes will conserve their partners, creating similar patterns of interactions within and between farms. In that case, the network will have a single module encompassing all microbes in
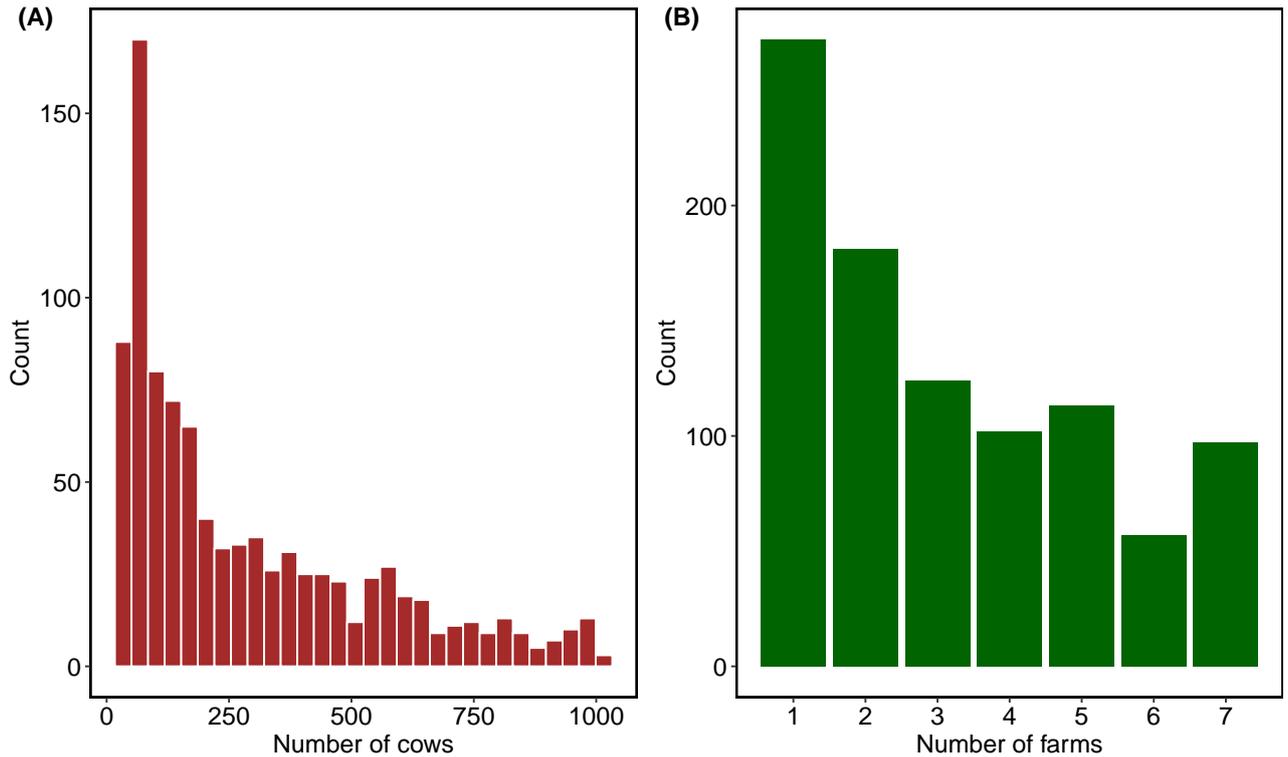
Fig. 2: Microbe distribution across cows and farms is skewed. The histograms depict the number of cows (A) and farms (B) in which microbes occur. Most microbes occur in a few cows, while a few (the tail of the distribution in (A)) occur in almost all cows. Over 70% occur in at least two farms.

all farms (H5c, Fig. 4C). This scenario is likely because we observed similar microbe assignment to groups across farms (H3) and high, nonrandom partner fidelity (H4). Between these two extremes, other partitions are possible. Specifically, we find a strong genetic population structure in which the two breeds are sorted into two genetic clusters, with all the Nordic Reds from Finland and Sweden forming one cluster and all Holsteins from the UK and Italy forming a second cluster (SI note 2). If host genetic background is a strong determinant of co-occurrence, then we should expect two modules (H5b, Fig. 4B), corresponding to the northern (SE, FI) and southern (UK, IT) farms. Testing these alternative hypotheses requires a mathematical object that separates network farms on the one hand and connects them on the other. We solved this problem using a multilayer network, representing a meta-co-occurrence [37,57] (Methods).

We found that the network was not partitioned and contained a single module, strongly supporting hypothesis H5a (Fig. 4A). This result is consistent with H3 and H4, indicating that similar processes occur across all farms, maintaining partner fidelity.

The multilayer network had 269K intralayer and 287K interlayer links. This comparable proportion of links results from strong partner fidelity and likely promotes a lack of farm-specific partitions. To further quantify the relative effect of inter-farm level processes, we repeated the modularity analysis while gradually removing an increasing proportion of the interlayer links (uniformly at random). Removing interlayer links should facilitate partitioning between farms because the random walk is less likely to move between farms than within. The network consistently had a single module containing all farms until 80% of the interlayer links were removed (Fig. 4D). Therefore, there is strong evidence that the processes determining co-occurrence operate similarly in all farms.
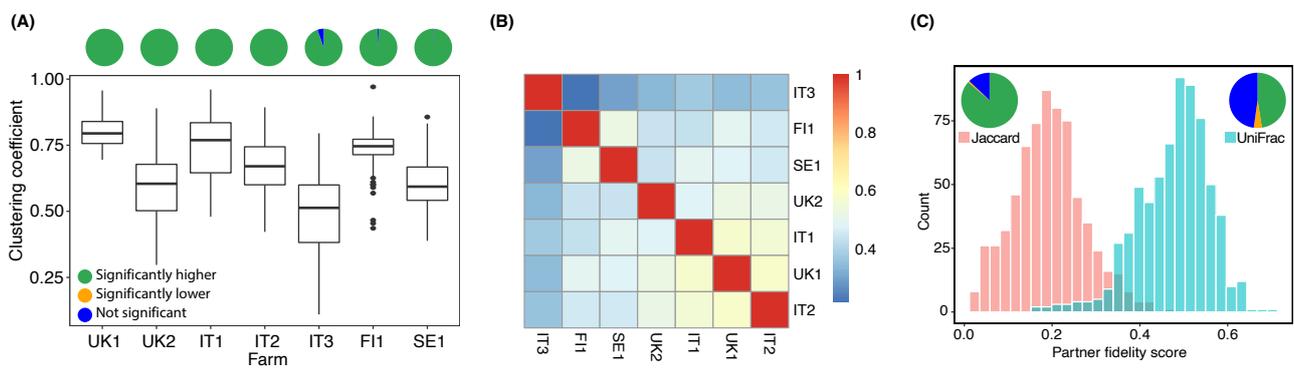
Fig. 3: Non-random scale signatures are observed at the farm and inter-farm levels. (A) The box plots depict the values of the clustering coefficient CC_i across all microbes within a farm. Microbes are predominantly involved in transitive co-occurrence within each farm. **(B)** Comparison of SBM group structure between pairs of farms using normalized mutual information (NMI). NMI quantifies the extent to which knowing the group composition in one farm provides information on the group composition in the second. The diagonal is the control because knowing the group composition in a given farm provides perfect information (NMI=1) on the group structure of that same farm. **(C)** Distributions of 649 microbes' partner fidelity calculated with Jaccard (P_i^J) and UniFrac (P_i^U). The pie charts in panels A and C depict the proportion of microbes whose CC_i, P_i^J, and P_i^U are significantly greater (green), lower (orange), or non-significant (blue). Significance was calculated with a z-score test comparing the observed index of each microbe to that obtained from 500 shuffled networks of each farm (see Methods for the description of the null model).
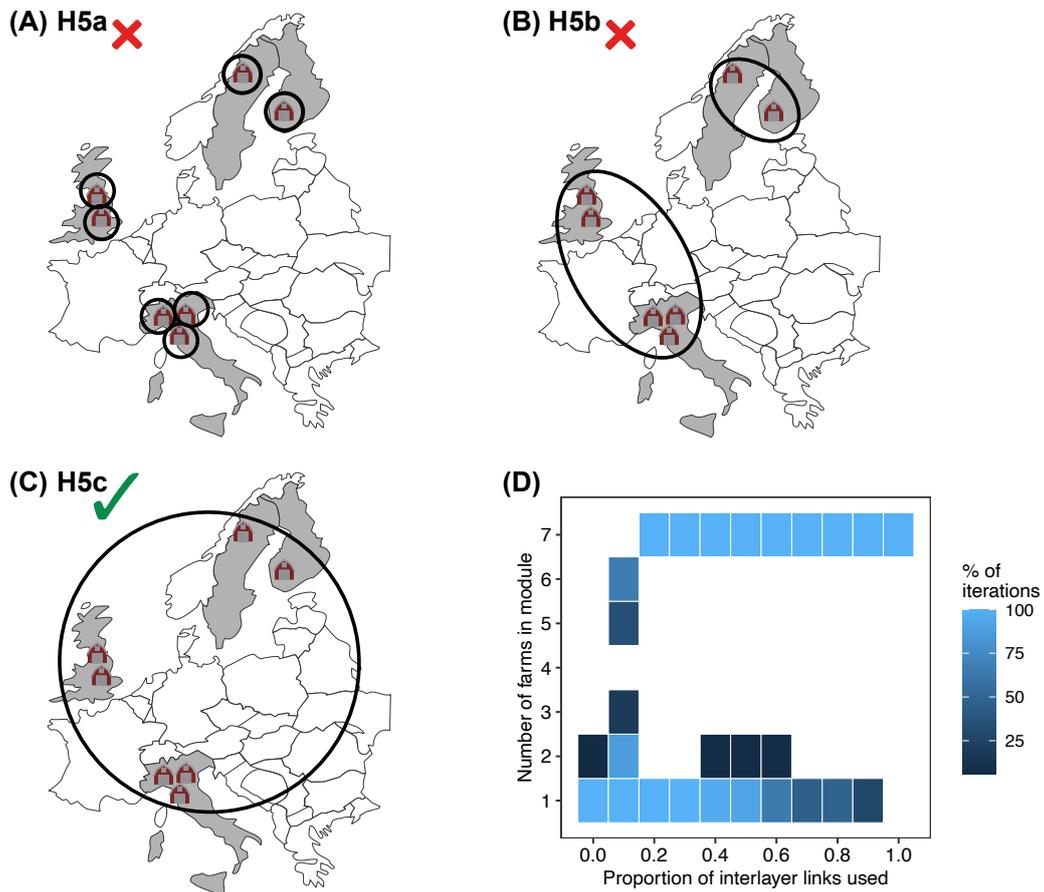
Fig. 4: Hypotheses for modularity analysis. Farms are depicted with icons and modules with colored circles. **(A) H5a:** Local processes dominate. In this hypothesis, each farm (layer in the multilayer network) will be assigned to its own module. **(B) H5b:** Host genetic background determines modularity. The two modules correspond to two cow breeds (Methods, Table S1). **(C) H5c:** Global processes determine co-occurrence. In this scenario, all the farms are assigned to a single module encompassing the entire network. The green tick and red x denote hypotheses that were supported or not, respectively. **(D)** We repeated the modularity analysis while randomly removing a certain proportion of the interlayer links (x-axis). Each cell shows the proportion of iterations in which at least one module contains a given amount of farms (y-axis). Note that columns do not necessarily sum to 1 because, in a given run, several modules could contain any number of farms.

Discussion

Highly diverse microbial communities are often studied using co-occurrence networks because direct physical interactions are impossible to observe. Nevertheless, the relative contributions of local and non-local processes to the structure of co-occurrence networks have been little studied. We addressed this challenge using a combination of hypothesis-driven network analyses at increasing scales. We found non-random signatures of co-occurrence within farms, which scaled up to the inter-farm level. At this level, there were clear signatures in the microbial co-occurrence structures, which indicated that similar processes across farms drive farm-level co-occurrence.

At the farm scale, co-occurrence was non-randomly transitive (triangle closure) in virtually all microbes regardless of the farm. Although several studies have quantified transitivity, its interpretation has been vague [7,17,64]. The coexistence of three microbes can result from shared habitat preferences (environmental filtering). This is likely because, within a farm, cow hosts were genetically similar (same breed) and fed the same diet. Another possible mechanism is dependency due to cross-feeding. However, theory predicts that when two microbes depend on metabolites produced by a third one, they will compete with each other [7,65]. Therefore, faster-growing bacteria should prevail without stabilizing mechanisms that maintain coexistence. Our results suggest that stabilizing mechanisms, such as fluctuating resources and niche separation, allow microbe coexistence. Identifying these mechanisms and their interplay with environmental filtering is currently an open question. From a statistical perspective, a transitivity value is only informative when proper comparisons are made with a null model. We solved this issue for the first time by comparing the transitivity of each microbe to random expectations.

A previous study showed mixed effects of climate and local interactions on microbial communities of the pitcher plant *Sarracenia purpurea* [66]. In our system, host genetics can serve as an environmental filter for microbes, determining occurrence [20]. However, such local (farm-level) signatures were not detected. Instead, at the inter-farm scale, microbes maintained co-occurrence links (with a significant phylogenetic signal) across farms. This can be explained by environmental filtering because similar environments would allow similar microbes to co-occur. However, the farm networks were partitioned into groups of microbes that were highly repeatable across farms: the group composition in one farm provides sufficient information to know the group composition in another (non-significant and high NMI). Such a functional division, which is repeatable across farms, can be driven by biotic dependence because microbes must co-occur with the same microbes to maintain cross-feeding or other complementary functions [67].

The analytical methods we employed (e.g., SBM) are rarely used in microbial and community ecology and provide a novel aspect to this work. In particular, the meta-co-occurrence multilayer network we developed in this study was used here for the first time. Future studies can adopt these methods. Furthermore, our approach is not limited to co-occurrence networks. Future studies can use it to explore local and non-local signatures in species interaction (meta)networks or metacommunities. Identifying such signatures remains little studied [13].

Along with that, our study has several limitations. First, we used presence-absence data. We opted not to use reads as abundance estimates due to known limitations [68]. Nevertheless, microbes with low prevalence but high occurrence could influence co-occurrence structure. Second, we focused on the community composition of the microbiome without considering its functional aspect. Our SBM

analysis indicates that functional aspects such as metabolic pathways should be included in future studies (we did not have such data). It is also worth noting that this study was conducted on gut microbiome inhabitants. Studying samples of free-living microbes could reveal a different structure and is a relevant direction for further research.

In conclusion, our results provide strong evidence that co-occurrence patterns are driven by similar forces across locations in the rumen ecosystem. Comprehending the processes underlying rumen microbiome assembly can aid in developing strategies for its manipulation to mitigate greenhouse gas emissions [67]. More broadly, understanding the mechanisms underlying the scale of microbe distribution and interactions is an open question in microbial ecology. We provided a set of hypotheses and accompanying analytical methods to detect scale-depend signatures in network structure. These hypotheses and tools can be adopted in other microbial and non-microbial systems.

References

1. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350,** 663–666. doi:10.1126/science.aad2602 (2015).

2. Gonze, D., Lahti, L., Raes, J. & Faust, K. Multi-stability and the origin of microbial community types. *ISME J.* **11,** 2159–2166. doi:10.1038/ismej.2017.60 (2017).

3. Liu, J., Meng, Z., Liu, X. & Zhang, X.-H. Microbial assembly, interaction, functioning, activity and diversification: a review derived from community compositional data. *Marine Life Science & Technology* **1,** 112–128. doi:10.1007/s42995-019-00004-3 (2019).

4. Shabat, S. K. B. *et al.* Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. *ISME J.* **10,** 2958–2972. doi:10.1038/ismej.2016.62 (2016).

5. Bergamaschi, M. *et al.* Gut microbiome composition differences among breeds impact feed efficiency in swine. *Microbiome* **8,** 110. doi:10.1186/s40168-020-00888-9 (2020).

6. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.* **5,** 219. doi:10.3389/fmicb.2014.00219 (2014).

7. Röttjers, L. & Faust, K. From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiol. Rev.* **42,** 761–780. doi:10.1093/femsre/fuy030 (2018).

8. Blanchet, F. G., Cazelles, K. & Gravel, D. Co-occurrence is not evidence of ecological interactions. *Ecol. Lett.* **23,** 1050–1063. doi:10.1111/ele.13525 (2020).

9. Goberna, M. & Verdú, M. Cautionary notes on the use of co-occurrence networks in soil ecology. *Soil Biol. Biochem.* **166,** 108534. doi:10.1016/j.soilbio.2021.108534 (2022).

10. Román, S., Ortiz-Álvarez, R., Romano, C., Casamayor, E. O. & Martin, D. Microbial community structure and functionality in the deep sea floor: Evaluating the causes of spatial heterogeneity in a submarine canyon system (NW Mediterranean, Spain). *Front. Mar. Sci.* **6.** doi:10.3389/fmars.2019.00108 (2019).

11. Ortiz-Álvarez, R., Cáliz, J., Camarero, L. & Casamayor, E. O. Regional community assembly drivers and microbial environmental sources shaping bacterioplankton in an alpine lacustrine district (Pyrenees, Spain). *Environ. Microbiol.* **22,** 297–309. doi:10.1111/1462-2920.14848 (2020).

12. Ontiveros Vicente J. *et al.* Biological Microbial Interactions from Cooccurrence Networks in a High Mountain Lacustrine District. *mSphere* **0,** e00918–21. doi:`10.1128/msphere.00918-21` (2022).

13. Saravia, L. A., Marina, T. I., Kristensen, N. P., De Troch, M. & Momo, F. R. Ecological network assembly: How the regional metaweb influences local food webs. *J. Anim. Ecol.* **91,** 630–642. doi:`10.1111/1365-2656.13652` (2022).

14. Araújo, M. B. & Rozenfeld, A. The geographic scaling of biotic interactions. *Ecography* **37,** 406–415. doi:`10.1111/j.1600-0587.2013.00643.x` (2014).

15. Goldford, J. E. *et al.* Emergent simplicity in microbial community assembly. *Science* **361,** 469–474. doi:`10.1126/science.aat1168` (2018).

16. Ladau, J. & Eloe-Fadrosh, E. A. Spatial, Temporal, and Phylogenetic Scales of Microbial Ecology. *Trends Microbiol.* **27,** 662–669. doi:`10.1016/j.tim.2019.03.003` (2019).

17. Yuan, M. M. *et al.* Climate warming enhances microbial network complexity and stability. *Nat. Clim. Chang.* **11,** 343–348. doi:`10.1038/s41558-021-00989-9` (2021).

18. Ma, B. *et al.* Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. *ISME J.* **10,** 1891–1901. doi:`10.1038/ismej.2015.261` (2016).

19. Mandakovic, D. *et al.* Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Sci. Rep.* **8,** 5875. doi:`10.1038/s41598-018-23931-0` (2018).

20. Mittelbach, G. G. & Schemske, D. W. Ecological and evolutionary perspectives on community assembly. *Trends Ecol. Evol.* **30,** 241–247. doi:`10.1016/j.tree.2015.02.008` (2015).

21. Miller, E. T., Svanbäck, R. & Bohannan, B. J. M. Microbiomes as Metacommunities: Understanding Host-Associated Microbes through Metacommunity Ecology. *Trends Ecol. Evol.* **33,** 926–935. doi:`10.1016/j.tree.2018.09.002` (2018).

22. Pascual, M. & Dunne, J. A. *Ecological Networks: Linking Structure to Dynamics in Food Webs* (Oxford University Press, 2006).

23. Bascompte, J. Networks in ecology. *Basic Appl. Ecol.* **8,** 485–490. doi:`10.1016/j.baae.2007.06.003` (2007).

24. Mizrahi, I., Wallace, R. J. & Moraïs, S. The rumen microbiome: balancing food security and environmental impacts. *Nat. Rev. Microbiol.* **19,** 553–566. doi:`10.1038/s41579-021-00543-6` (2021).

25. Wallace, R. J. *et al.* A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. *Sci Adv* **5,** eaav8391. doi:`10.1126/sciadv.aav8391` (2019).

26. Shade, A. & Handelsman, J. Beyond the Venn diagram: the hunt for a core microbiome. *Environ. Microbiol.* **14,** 4–12. doi:`10.1111/j.1462-2920.2011.02585.x` (2012).

27. Lemanceau, P., Blouin, M., Muller, D. & Moënne-Loccoz, Y. Let the Core Microbiota Be Functional. *Trends Plant Sci.* **22,** 583–595. doi:`10.1016/j.tplants.2017.04.008` (2017).

28. Kokou, F. *et al.* Core gut microbial communities are maintained by beneficial interactions and strain variability in fish. *Nat Microbiol* **4,** 2456–2465. doi:`10.1038/s41564-019-0560-0` (2019).

29. Li, J. *et al.* Distinct mechanisms shape soil bacterial and fungal co-occurrence networks in a mountain ecosystem. *FEMS Microbiol. Ecol.* **96.** doi:`10.1093/femsec/fiaa030` (2020).

30. Hibbing, M. E., Fuqua, C., Parsek, M. R. & Peterson, S. B. Bacterial competition: surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.* **8,** 15–25. doi:`10.1038/nrmicro2259` (2010).

31. Newman, M. E. J. *Networks* Second edition.. (Oxford University Press, Oxford, United Kingdom, 2018).

32. Allesina, S. & Pascual, M. Food web models: a plea for groups. *Ecol. Lett.* **12,** 652–662. doi:`10.1111/j.1461-0248.2009.01321.x` (2009).

33. Peixoto, T. P. in *Elements in the Structure and Dynamics of Complex Networks* (Cambridge University Press, 2023). doi:`10.1017/9781009118897`.

34. Baskerville, E. B. *et al.* Spatial guilds in the Serengeti food web revealed by a Bayesian group model. *PLoS Comput. Biol.* **7,** e1002321. doi:`10.1371/journal.pcbi.1002321` (2011).

35. Cobo-López, S., Gupta, V. K., Sung, J., Guimerà, R. & Sales-Pardo, M. Stochastic block models reveal a robust nested pattern in healthy human gut microbiomes. *PNAS Nexus* **1,** gac055. doi:`10.1093/pnasnexus/pgac055` (2022).

36. Trøjelsgaard, K., Jordano, P., Carstensen, D. W. & Olesen, J. M. Geographical variation in mutualistic networks: similarity, turnover and partner fidelity. *Proc. Biol. Sci.* **282,** 20142925. doi:`10.1098/rspb.2014.2925` (2015).

37. Pilosof, S., Porter, M. A., Pascual, M. & Kéfi, S. The multilayer nature of ecological networks. *Nat Ecol Evol* **1,** 0101. doi:`10.1038/s41559-017-0101` (2017).

38. Perlman, D. *et al.* Concepts and Consequences of a Core Gut Microbiota for Animal Growth and Development. *Annu Rev Anim Biosci* **10,** 177–201. doi:`10.1146/annurev-animal-013020-020412` (2022).

39. Veech, J. A. A probabilistic model for analysing species co-occurrence. *Glob. Ecol. Biogeogr.* **22,** 252–260. doi:`10.1111/j.1466-8238.2012.00789.x` (2013).

40. Griffith, D. M., Veech, J. A. & Marsh, C. J. cooccur: Probabilistic Species Co-Occurrence Analysis in R. *J. Stat. Softw.* **69,** 1–17. doi:`10.18637/jss.v069.c02` (2016).

41. Mainali, K. P., Slud, E., Singer, M. C. & Fagan, W. F. A better index for analysis of co-occurrence and similarity. *Sci Adv* **8,** eabj9204. doi:`10.1126/sciadv.abj9204` (2022).

42. Mainali, K. P. *et al.* Statistical analysis of co-occurrence patterns in microbial presence-absence datasets. *PLoS One* **12,** e0187132. doi:`10.1371/journal.pone.0187132` (2017).

43. Adair, K. L., Wilson, M., Bost, A. & Douglas, A. E. Microbial community assembly in wild populations of the fruit fly Drosophila melanogaster. *ISME J.* **12,** 959–972. doi:`10.1038/s41396-017-0020-x` (2018).

44. Gotelli, N. J. Null Model Analysis of Species Co-Occurrence Patterns. *Ecology* **81,** 2606–2621. doi:`10.2307/177478` (2000).
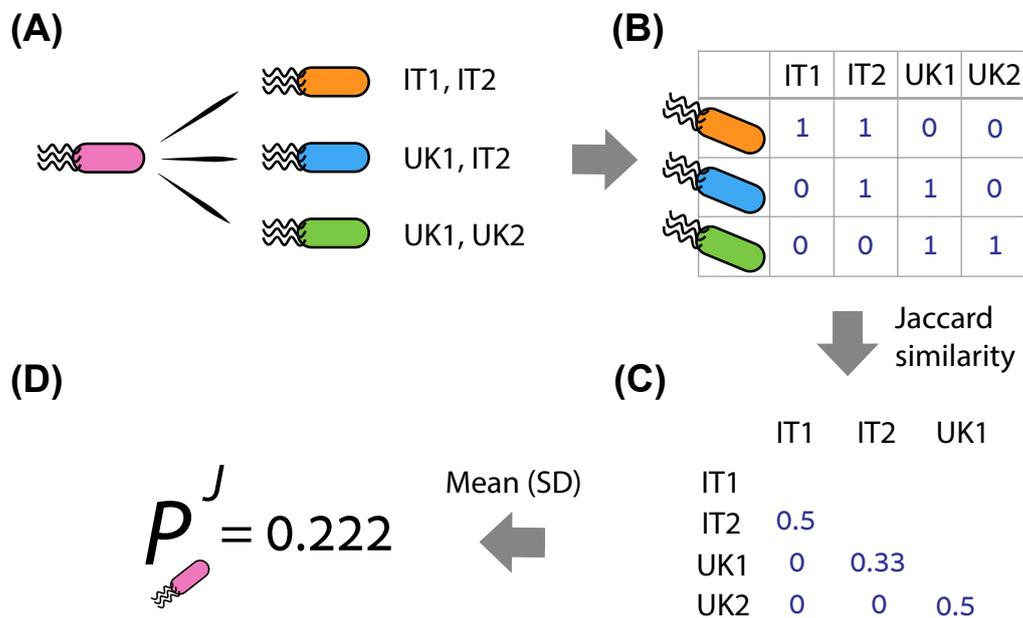
45. Delmas, E. *et al.* Analysing ecological networks of species interactions: Analyzing ecological networks. *Biol Rev* **94,** 16–36. doi:`10.1111/brv.12433` (2019).

46. Rosvall, M., Delvenne, J.-C., Schaub, M. T. & Lambiotte, R. in *Advances in network clustering and blockmodeling* (eds Doreian, P., Batagelj, V. & Ferligoj, A.) 71–87 (Wiley, 2018).

47. Chiquet, J., Donnet, S. & Barbillon, P. *sbm: Stochastic Blockmodels* 2024.

48. Danon, L., Díaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *J. Stat. Mech.* **2005,** P09008. doi:`10.1088/1742-5468/2005/09/P09008` (2005).

49. Wu, T. *NMI: Normalized Mutual Information of Community Structure in Network* 2016.

50. Oksanen, J. *et al.* vegan: Community Ecology Package (version 2.5-7). *https://CRAN.R-project.org/package=* (2020).

51. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235. doi:`10.1128/AEM.71.12.8228-8235.2005` (2005).

52. Chen, J. *et al.* Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28,** 2106–2113. doi:`10.1093/bioinformatics/bts342` (2012).

53. Paradis & Schliep. *ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R* 2019.

54. Tamura, K & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10,** 512–526. doi:`10.1093/oxfordjournals.molbev.a040023` (1993).

55. Schliep, K. P. *phangorn: phylogenetic analysis in R* 2011. doi:`10.1093/bioinformatics/btq706`.

56. R Core Team. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2021.

57. Hutchinson, M. C. *et al.* Seeing the forest for the trees: Putting multilayer networks to work for community ecology. *Funct. Ecol.* **33** (ed Godoy, O.) 206–217. doi:`10.1111/1365-2435.13237` (2019).

58. Hervías-Parejo, S *et al.* Species functional traits and abundance as drivers of multiplex ecological networks: first empirical quantification of inter-layer edge weights. *Proc. Biol. Sci.* **287,** 20202127. doi:`10.1098/rspb.2020.2127` (2020).

59. Farage, C., Edler, D., Eklöf, A., Rosvall, M. & Pilosof, S. Identifying flow modules in ecological networks using Infomap. *Methods Ecol. Evol.* **12,** 778–786. doi:`10.1111/2041-210x.13569` (2021).

60. Vitali, A. *et al.* Invasive species modulate the structure and stability of a multilayer mutualistic network. *Proc. Biol. Sci.* **290,** 20230132. doi:`10.1098/rspb.2023.0132` (2023).

61. Shapiro, J. T. *et al.* Multilayer networks of plasmid genetic similarity reveal potential pathways of gene transmission. *ISME J.,* 1–11. doi:`10.1038/s41396-023-01373-5` (2023).

62. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 1118–1123. doi:`10.1073/pnas.0706851105` (2008).

63. De Domenico, M., Lancichinetti, A., Arenas, A. & Rosvall, M. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Phys. Rev. X* **5,** 011027. doi:10.1103/PhysRevX.5.011027 (2015).

64. Faust, K. *et al.* Cross-biome comparison of microbial association networks. *Front. Microbiol.* **6,** 1200. doi:10.3389/fmicb.2015.01200 (2015).

65. Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31,** 343–366 (2000).

66. Freedman, Z. B. *et al.* Environment-host-microbial interactions shape the Sarracenia purpurea microbiome at the continental scale. *Ecology* **102,** e03308. doi:10.1002/ecy.3308 (2021).

67. Moraïs, S. & Mizrahi, I. The Road Not Taken: The Rumen Microbiome, Functional Groups, and Community States. *Trends Microbiol.* **27,** 538–549. doi:10.1016/j.tim.2018.12.011 (2019).

68. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8,** 2224. doi:10.3389/fmicb.2017.02224 (2017).

69. Greenbaum, G., Templeton, A. R. & Bar-David, S. Inference and Analysis of Population Structure Using Genetic Data and Network Theory. *Genetics* **202,** 1299–1312. doi:10.1534/genetics.115.182626 (2016).

70. Greenbaum, G., Rubin, A., Templeton, A. R. & Rosenberg, N. A. Network-based hierarchical population structure analysis for large genomic data sets. *Genome Res.* **29,** 2020–2033. doi:10.1101/gr.250092.119 (2019).

Supplementary Information

Table S1: Details of Wallace et al.'s 2019 study design.

| Farm	Breed	Diet	Housing
UK1 UK2 IT1 IT2 IT3	Holstein dairy cows	Maize silage, grass silage or grass hay, and concentrates	Group-housed
SE1 FI1	Nordic Red dairy cows	Grass silage and concentrates	Group-housed Individual standings

(A)

IT1, IT2

UK1, IT2

UK1, UK2

(B)

	IT1	IT2	UK1	UK2
	1	1	0	0
	0	1	1	0
	0	0	1	1

Jaccard similarity

(C)

	IT1	IT2	UK1
IT1			
IT2	0.5		
UK1	0	0.33	
UK2	0	0	0.5

(D)

Mean (SD)

$$P_i^{J} = 0.222$$

Fig. S1: Partner fidelity calculation. Partner fidelity calculation is done for each microbe separately by comparing the set of co-occurring partners it has in one layer (farm) to the set of partners it has in another layer using a beta-diversity index. (A) In this diagram partner fidelity (P_i^J) was calculated for the pink microbe i by first (B) identifying the partners is has in each layer, then (C) calculating for each pair of layers how similar are the partners in one layer compared to the other, using Jaccard. (D) Lastly, the mean value of all pairwise similarities across layers was calculated and defined as the partner fidelity of the pink microbe. We also repeated the same calculation using UniFrac (P_i^U) instead of Jaccard.
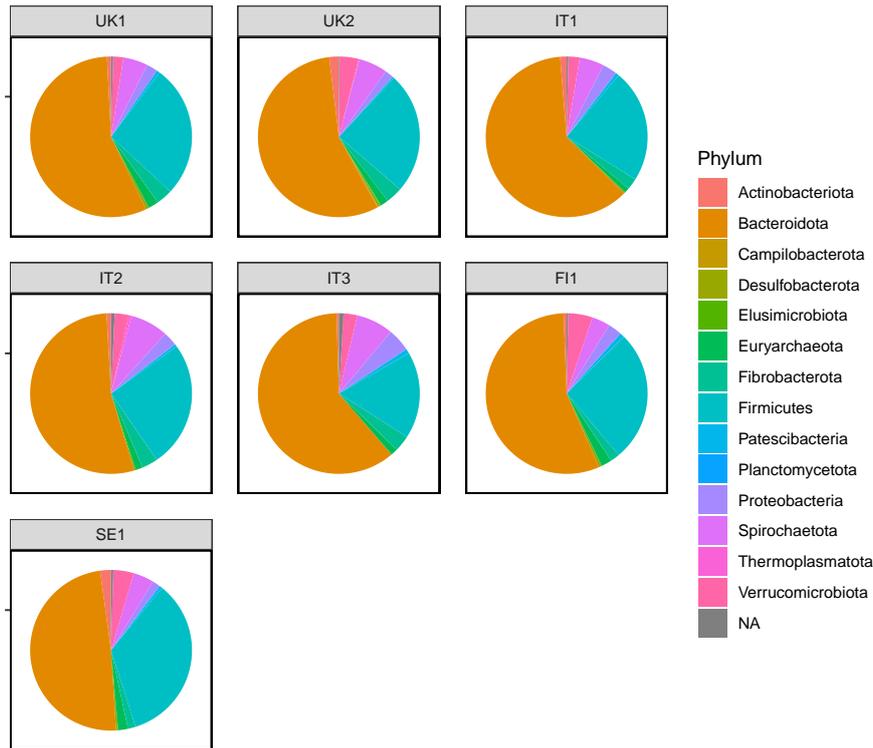
Fig. S2: ASV taxonomy dominated by Bacteroidetes and Firmicutes. The relative richness—number of ASVs that belong to a certain Phylum out of all ASVs—in each farm. The two largest phyla are Bacteroidetes and Firmicutes.
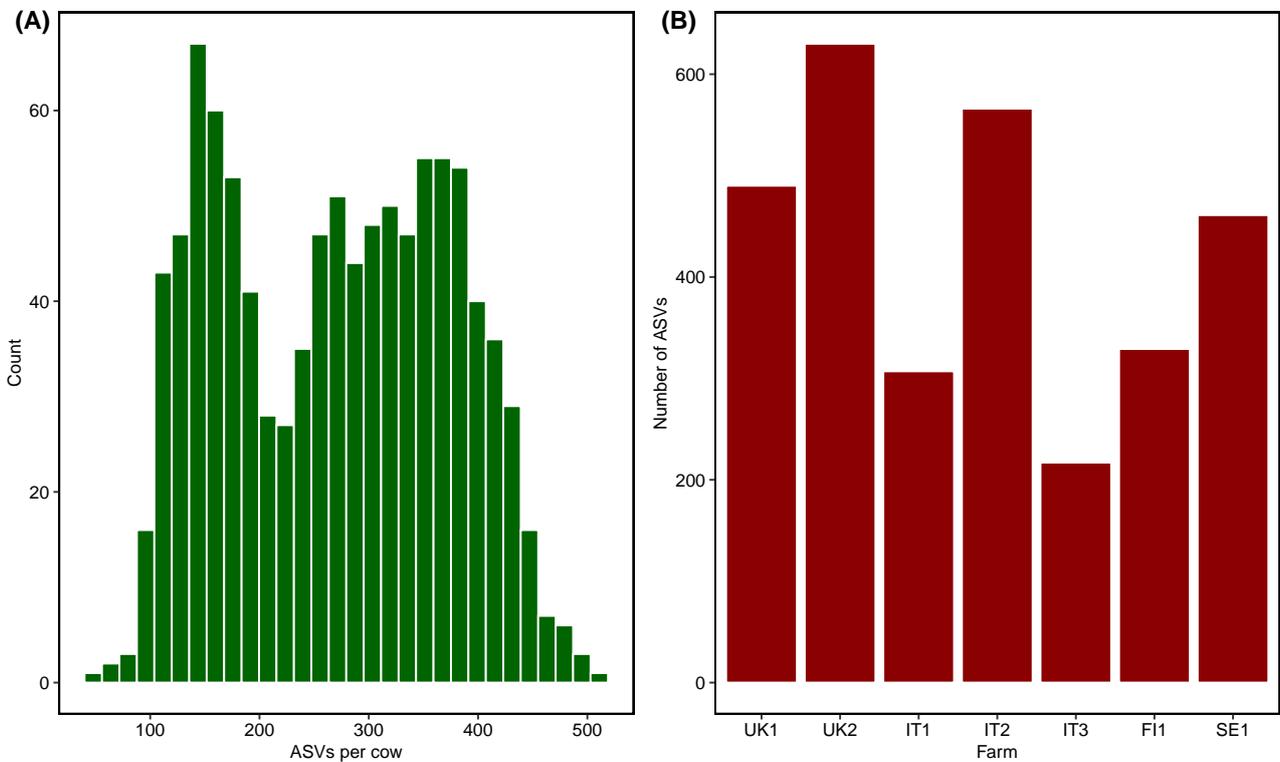

Fig. S3: ASV distribution. (A) A histogram showing the number of ASVs per cow (across farms). **(B)** Bar plots of the number of ASVs per farm (across cows).
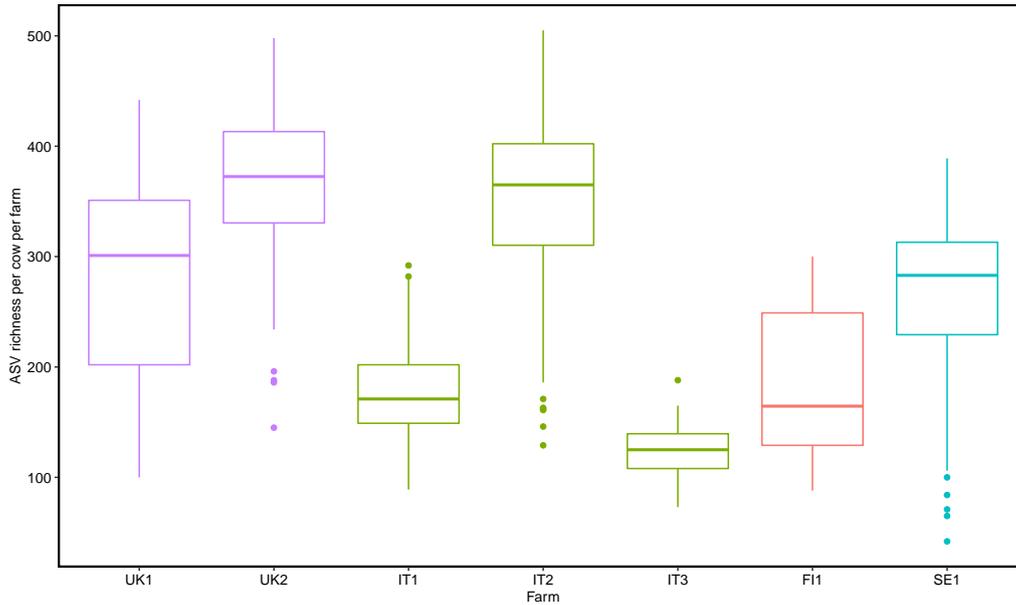
Fig. S4: ASV richness. Box plots of the distribution of number of ASVs in cows, within each farm. Colors depict countries.
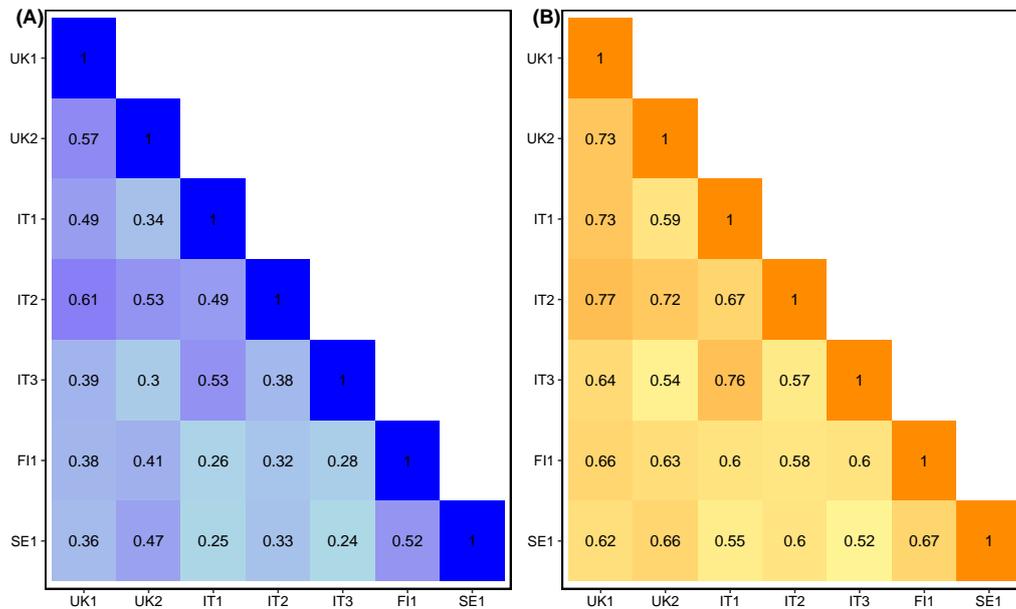

Fig. S5: ASV similarity across farms. The similarity between farms in ASV composition, calculated with Jaccard **(A)** and Unifrac **(B)**.
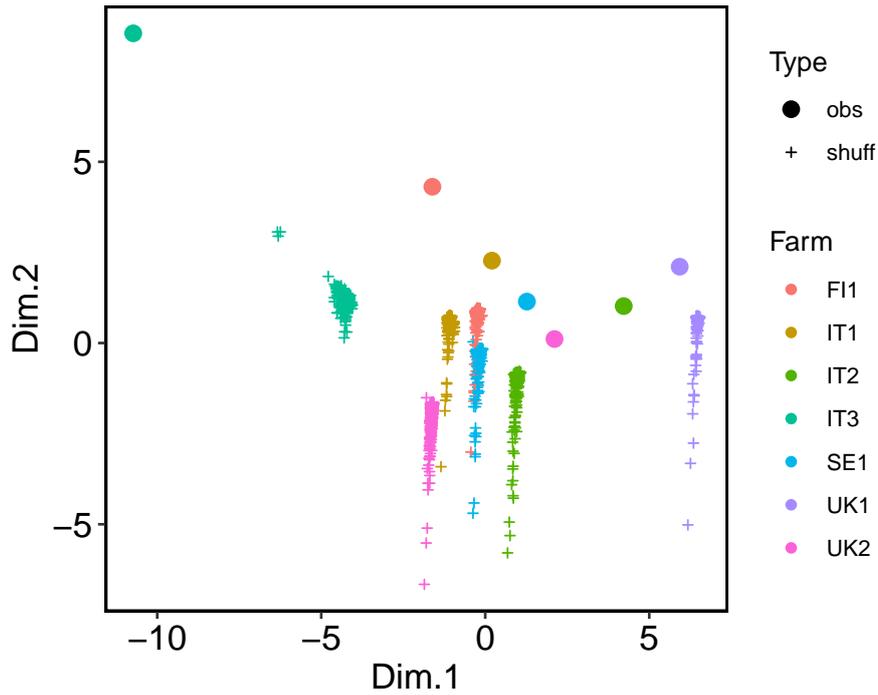
Fig. S6: Farm-level co-occurrence network structure is significantly different than random. Each of the seven co-occurrence networks (circles) was regenerated 500 times after redistributing microbes in cows (plus sign; see Methods). We calculated 14 network properties for each observed and shuffled networks. The PCA analysis was performed on all networks using the rescaled properties. It is clear that for each farm (colors), the observed co-occurrence structure is non-random (circles do not fall within the plus sign ranges)
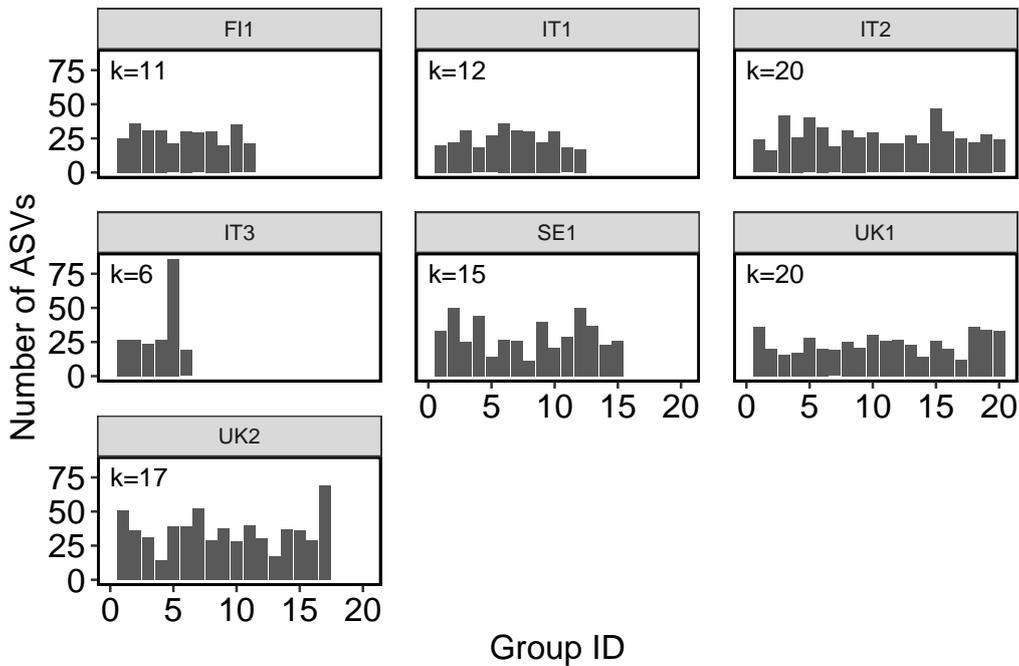
Fig. S7: Stochastic block model groups. The plots show the number of groups in each farm (number of columns), and the number of microbes assigned to each group (column height).
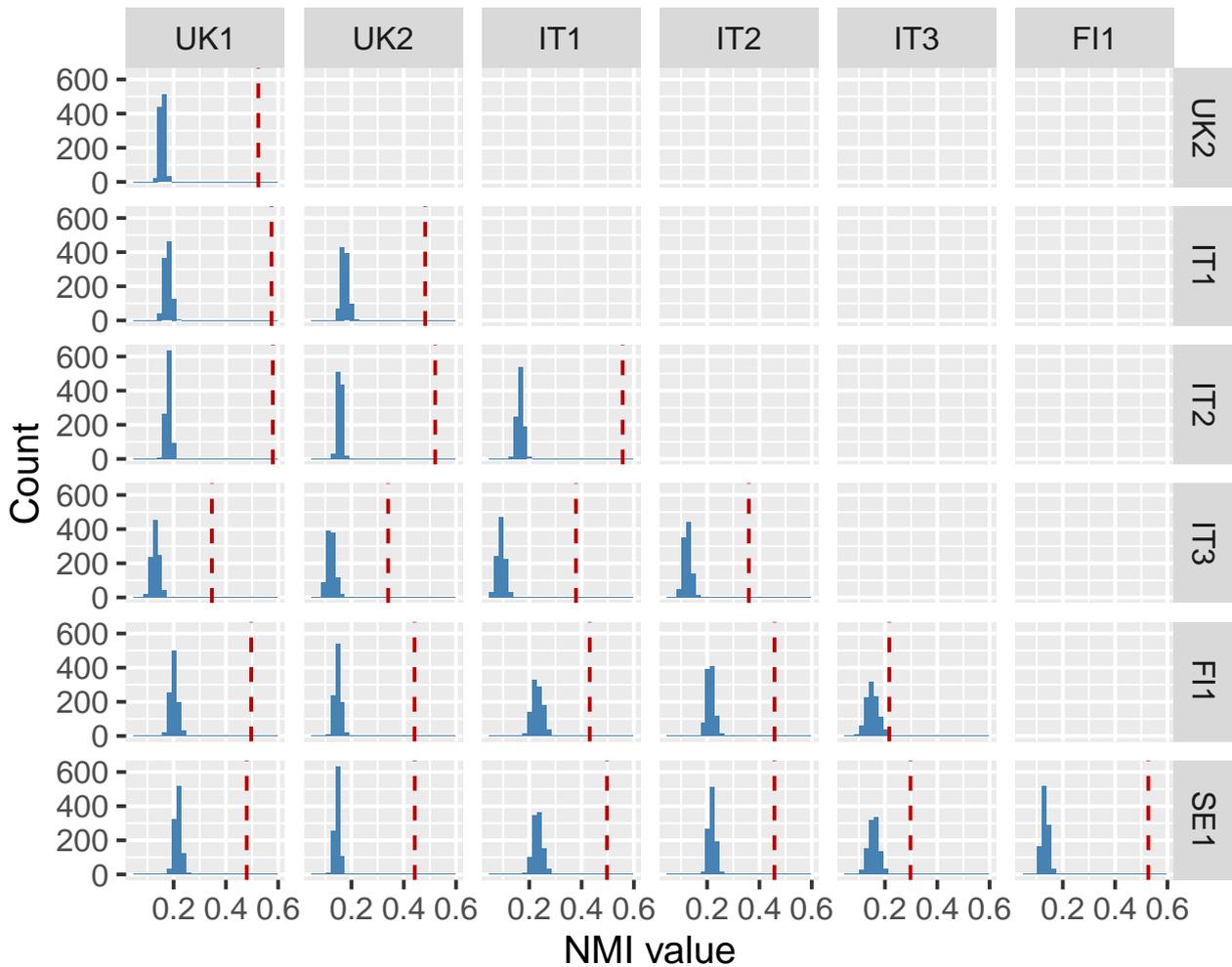
Fig. S8: Comparison of stochastic block model results to random. To test if the observed NMI in groups between farms is non random, we redistributed the ASVs across groups within each farm while keeping the size of each group. We then recalculated the NMI between each pair of farms using the shuffled node memberships. We repeated this process 500 times. The distributions of the NMI values of the shuffled networks are presented in the histograms, with the NMI value from the observed node membership depicted as a red dashed line.
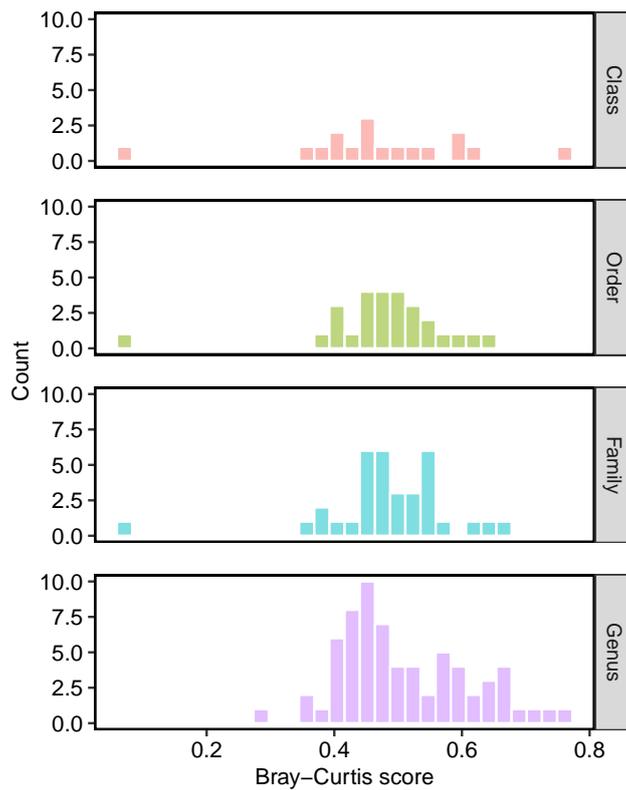
Fig. S9: Taxa-level partner fidelity. We repeated the partner fidelity analysis considering taxonomic levels instead of phylogenetic distances. We quantified how many ASVs connect each taxa pair (e.g., genus 1 with genus 2) within a farm. Then, we compared the identity and quantity of partners of each focal taxa between farms using Bray–Curtis similarity (weighted version of Jaccard similarity). The overall level of partner fidelity in all taxa was similar to that obtained using phylogenetic distances (UniFrac).

SI note 1: Sensitivity to thresholds of core microbes

To focus on the signal produced by the key microbes in the community, we used the common core microbes in each farm [25,38]. We did not work with abundance or relative abundance due to known issues with using reads as abundance estimates [68]. Therefore, core microbes were defined as ASVs occurring in a certain proportion of cows within each farm (the 'common core' definition of [38]). A general problem in deciding on a relevant threshold for setting core microbes is the trade off between environmental adaptability to biological variation [38]. On the one hand, including too many ASVs can skew the signal because typically most microbes are rare or occur in too few samples (here, cows). In that case, network structure may not be representative of the relevant gut biota because low-prevalence microbes may have little effect on community function. Moreover, computing co-occurrence networks is extremely extensive, even when run on High Performance Computing systems. Including too many ASVs may restrict the amount of analyses we can practically perform. On the other hand, including too few ASVs will cause loss of ecologically meaningful variation. There is obviously no one true answer to this problem. Our approach was therefore to start with prior knowledge of this data set. In the original paper, Wallace et al. [25] defined core microbes as those occurring at $\geq 50\%$ of cows within a farm, remaining with 454 ASVs out of $\approx 12,000$ (3.7%). They showed that even at this strict threshold, the remaining microbes can strongly predict the cow phenotype (e.g., methane emission, lactation). Therefore, this can be considered as the 'functional core' [38].

High-occupancy typically suggests that microbes are well adapted to their environment and carry functions that lead to high prevalence [38]. We therefore opted not to deviate drastically from the functional core because it is biologically meaningful. Nevertheless, a 50% threshold is extremely strict as it includes only 3.7% of ASVs. Additionally, the nature of the question we asked requires more variation because the functional core, being so restricted, will be highly repetitive across cows and farms. Therefore, we conducted a sensitivity analysis measuring the percentage of ASVs present in cows in each farm as a function of different threshold levels (Fig. S10). While the sharpest decline was at 5%, it was a large deviation from the functional core. In addition, this threshold had too many bacteria for practical computational analysis, especially because we had shuffled networks and SBM. Given that after this drop, the curve did not change drastically, we opted to work with a threshold of 30%, remaining with 946 ASVS. This allowed us to relax Wallace et al.'s [25] definition while doubling the number of microbes in the analysis. In addition, we repeated the analyses (without the shuffled networks) for the 5%, 10% and 20% threshold levels as described hereafter.

Sensitivity analysis showed that, qualitatively, the general pattern of CC_i was maintained across farms. Lower threshold include more microbes, lowering the probability that microbes will close triangles. Therefore, within each farm, CC_i increased with threshold (Fig. S11). The qualitative differences between P_i^J and P_i^U was maintained as with the 30% threshold. That is, UniFrac had generally higher similarity than Jaccard. (Fig. S12). The multilayer modularity resulted in a single module that encompassed all farms at all thresholds. Therefore, over-
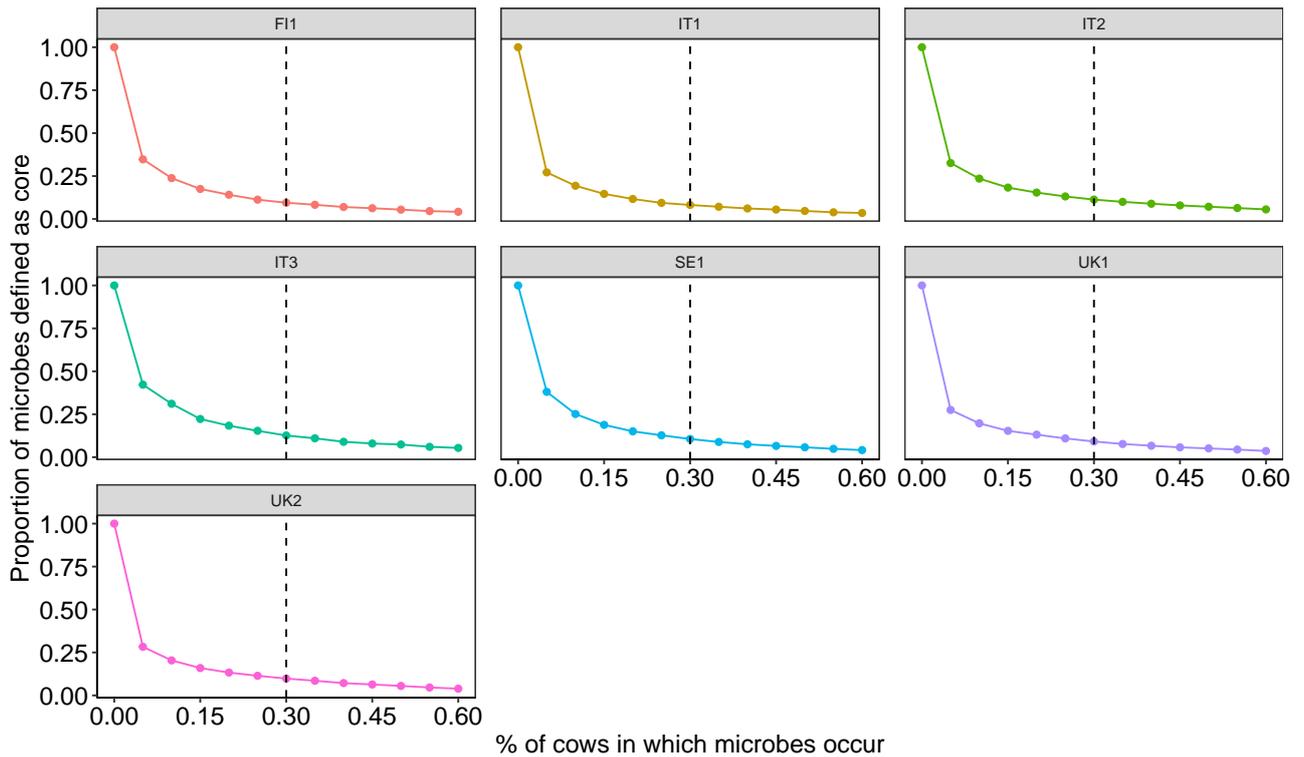
Fig. S10: Sensitivity analysis for core microbes. Within each farm we included ASVs that occurred in a given proportion of the cows (x-axis), and quantified the proportion of ASVs that were defined as core microbes. The vertical dashed line marks the threshold with which we work in the main text.

all, the sensitivity analysis shows that the patterns we obtained in the 30% threshold, which is close to the 'functional core' found in Wallace et al. [25] provides a good balance between a biologically meaningful set of microbes and ability to discern the scale of the processes that define it.
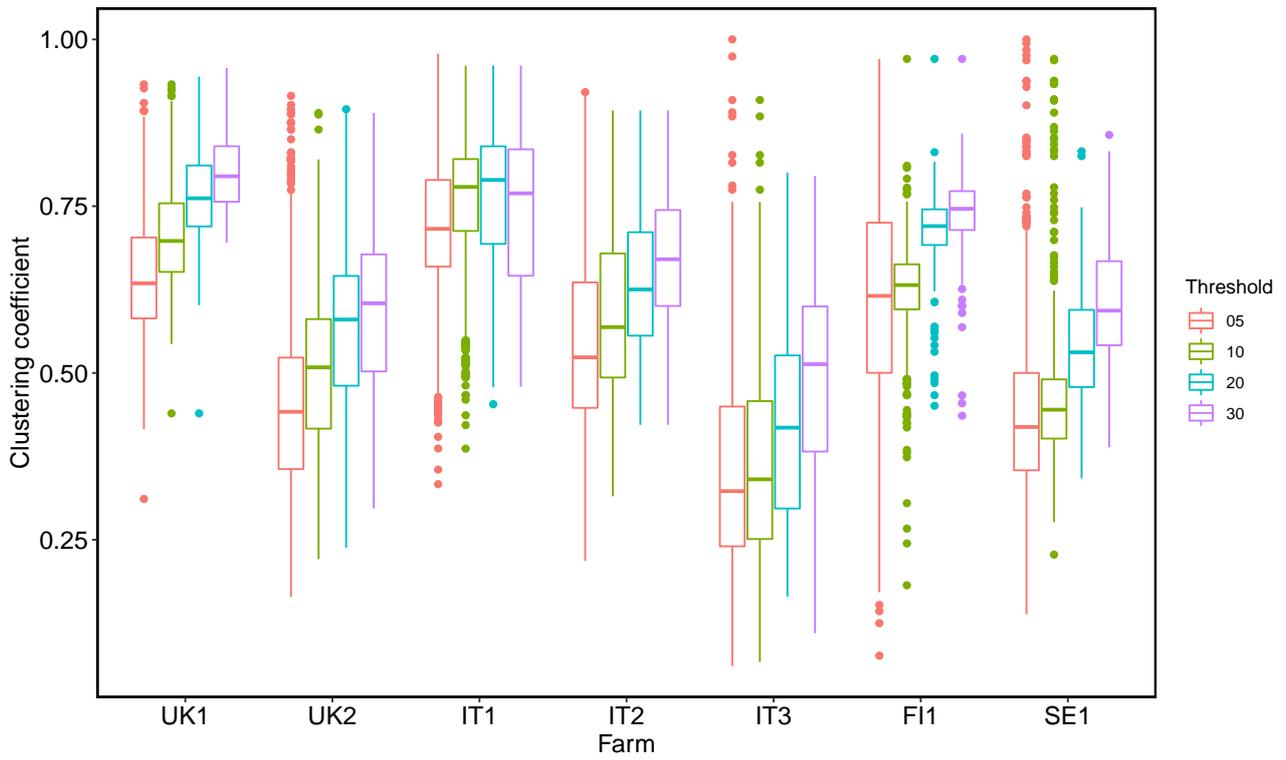
Fig. S11: Sensitivity analysis of transitivity. Box plots depict the values of the clustering coefficient CC_i across all microbes within a farm for different thresholds of the common core (depicted in colors). As expected, the lower the threshold, the lower the CC_i because lower thresholds include more microbes.
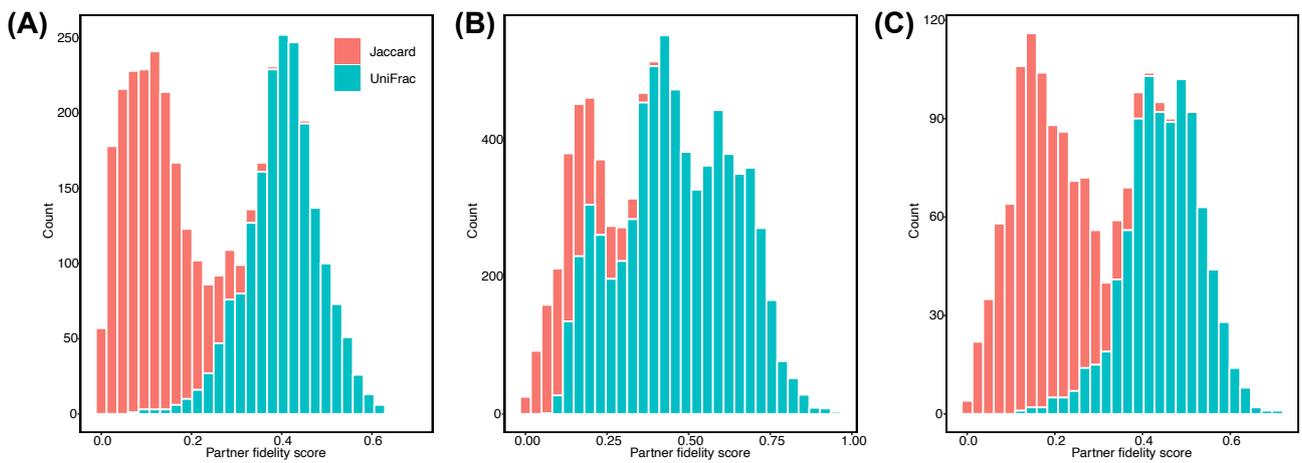

Fig. S12: Sensitivity analysis of partner fidelity. Distributions of P_i^J (red) and P_i^U (blue) for the 5% (A), 10% (B) and 20% (C) thresholds.

SI note 2: Analysis of cow population genetics

We analyzed population genetic structure for cows for which we had genetic and microbiome data (935 out of 1012 cows: 741 Holsteins and 194 Nordic Reds). The data set had 67785 shared SNPs. First, we computed a genetic similarity matrix. We used a frequency-weighted allele-sharing similarity (FWASS) approach, in which a greater weight is assigned to shared rare alleles than to shared common alleles when computing genetic similarity between a given pair of individuals [69,70]. We then used the genetic similarity matrix to compute hierarchical population genetic structure by computing hierarchical clustering using the Ward method. The hierarchical population genetic structure was as follows (Fig. S13). *At level 1*, the two breeds were clearly sorted into two clusters, with all the Nordic Reds from Finland and Sweden forming one cluster, and all Holsteins from both the UK and Italy forming a second cluster. This level corresponds to hypothesis H5b, stating a partitioning between the northern (SE, FI) and southern (UK, IT) farms.

There are some dedicated packages for analyzing genetic structure related to strains, such as anapan, which includes automated filtering of metagenomic functional profiles. However, the goal of our population genetic analysis was merely to test for simple population genetic structure to formulate hypothesis H5b (Fig. 4 in the main text).
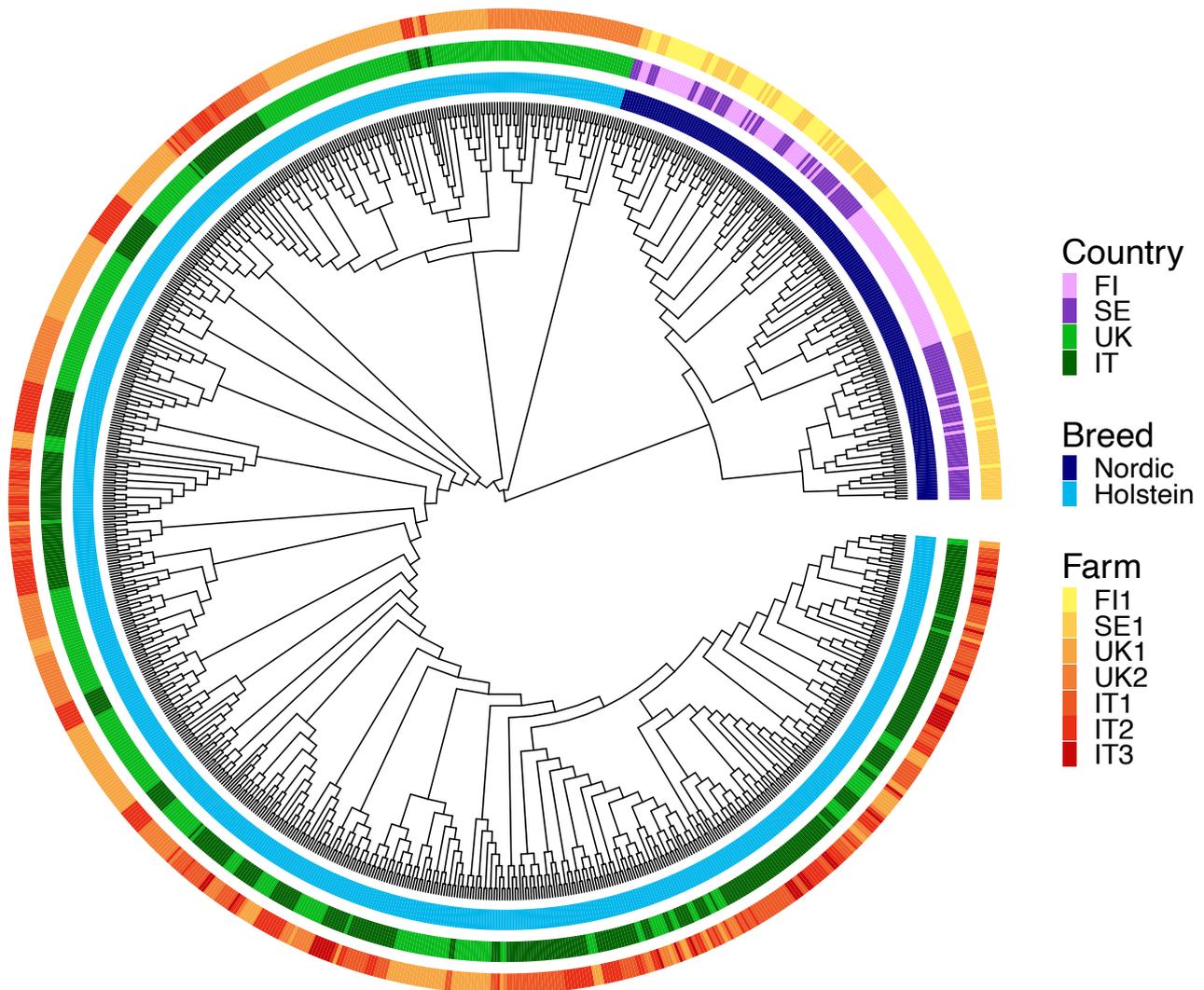
Fig. S13: Cow genetic population structure. The genetic similarity was calculated using SNPs data (see Methods), followed by clustering using Ward hierarchical clustering according to the genetic distance. Clustering was performed on 935 cows for which both SNP and microbiome data was available. Each leaf in the tree is a cow and colors indicate the cows' breed, country and farm.