Microbial network inference for longitudinal microbiome studies with LUPINE

Saritha Kodikara[1], Kim-Anh Lê Cao[1]

1 Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, Royal Parade, 3052, Victoria, Australia

kimanh.lecao@unimelb.edu.au

Abstract

The microbiome is a complex ecosystem of interdependent taxa that has traditionally been studied through cross-sectional studies. However, longitudinal microbiome studies are becoming increasingly popular. These studies enable researchers to infer taxa associations towards the understanding of coexistence, competition, and collaboration between microbes across time. Traditional metrics for association analysis, such as correlation, are limited due to the data characteristics of microbiome data (sparse, compositional, multivariate). Several network inference methods have been proposed, but have been largely unexplored in a longitudinal setting.

We introduce LUPINE (LongitUdinal modelling with Partial least squares regression for NEtwork inference), a novel approach that leverages on conditional independence and low-dimensional data representation. This method is specifically designed to handle scenarios with small sample sizes and small number of time points. LUPINE is the first method of its kind to infer microbial networks across time, while considering information from all past time points. It . We validate LUPINE and its variant, LUPINE_single (for single time point anlaysis) in simulated data and three case studies, where we highlight LUPINE's ability to identify relevant taxa in each study context, across different experimental designs (mouse and human studies, with or without interventions, as short or long time courses). To detect changes in the networks across time, groups or in response to external disturbances, we used different metrics to compare the inferred networks.

LUPINE is a simple yet innovative network inference methodology that is suitable for, but not limited to, analysing longitudinal microbiome data. The R code and data are publicly available for readers interested in applying these new methods to their studies.

1 Introduction

The microbiome, a term derived from the combination of "microbe" and the suffix "-ome", refers to the entirety of all taxa that coexist within a specific environment (Ursell et al., 2012). This ecosystem is composed of billions of taxa, forming complex community structures that are challenging to decipher. Given the inherent dynamism of the microbiome, longitudinal studies are essential to gain insights into the impact, variations over time, and responses to external perturbations such as dietary changes or medications on the microbiome. The temporal dimension, while adding complexity, provides richer insights into the dynamic processes, patterns, and interactions within microbial communities, surpassing what cross-sectional studies can offer (Lyu et al., 2023). The advancements in sequencing technologies, coupled with reduced sequencing costs, have enabled researchers to conduct an increasing number of microbiome longitudinal studies.

One objective of microbiome studies is to accurately infer microbial networks from the datasets generated by sequencing technologies. These networks can be viewed as temporal or spatial snapshots of ecosystems (Röttjers and Faust, 2018). In these networks, interactions signify significant associations between taxa, which are typically non-directional in nature (Kurtz et al., 2015). The simplest way to detect associations is via correlation analysis, such as Pearson or Spearman correlation. However, these correlation methods are not suitable for microbiome data as they ignore the compositional structure of the data, which may lead to spurious results (Gloor et al., 2017). A more valid approach is to use the concept of partial correlation, which effectively measure pairwise associations between variables subject to a constant-sum constraint (i.e. compositional data) (Erb, 2020). Existing compositional-aware microbiome network methods have currently focused on single time point studies, rather than longitudinal. Widely used single time point methods are either based on correlation approaches, such as SparCC (Friedman and Alm, 2012) or precision based approaches, such as SpiecEasi (Kurtz et al., 2015). In contrast to correlation-based methods, precision or partial correlation-based methods concentrate on direct associations by removing indirect associations.

Longitudinal microbiome studies enables researchers to create temporal snapshots of microbial networks for a comprehensive view of the microbiome. However, network methods for such type of

studies are still in their infancy and exhibit diverse analytical objectives (Kodikara et al., 2022). Some methods aim to model relationships between taxa for each individual subject. Other methods aim to model collective microbial interactions across all time points. However, the assumption that microbial interaction stays constant through time, whether on an individual basis or across all subjects, proves limiting. This limitation is particularly evident when there are external interventions, such as alterations in diet or antibiotic usage, which disrupt the stability of these relationships. We have address this limitation with a sequential approach, that, to the best of our knowledge, is the first of its kind for longitudinal microbiome studies. We acknowledge the dynamic nature of microbial interactions, especially in the face of interventions, and introduce a flexible framework for the temporal evolution of these interactions.

LUPINE (LongitUdinal modelling with Partial least squares regression for NEtwork inference) combines one-dimensional approximation and partial correlation to measure the linear association between a pair of taxa, accounting for the effects of the other taxa. We developed two variants, one for single time point network inference, and another for longitudinal sequential network inference. The former focuses on inferring the network at a specific time point, in the vein of SpiecEasi (Kurtz et al., 2015) and SparCC (Friedman and Alm, 2012). The latter, however, incorporates information from all previous time points. The inferred network from our method is in the format of binary graphs $G = (V, E)$, where the vertex set $V = \{v_1, \ldots, v_p\}$ represents p taxa and the edge set $E = \{(v_i, v_j), (v_k, v_l), \ldots\}$ represents significant associations among the taxa.

In Section 2, we explain the rationale of our approaches along with downstream analyses that can be conducted to compare inferred networks. In the results Section 3, we first benchmark LUPINE and LUPINE_single with two state-of-the-art network inference methods (solely designed for single time point microbiome data) on simulated data. Next, through a case study, we demonstrate the robustness of LUPINE compared to LUPINE_single. We then showcase the applicability of LUPINE in real data with different experimental designs in three case studies. These experiments varied from human studies to mouse studies, from short to longer time courses, and ranged from cross-sectional to intervention studies. Finally, in the discussion Section 4, we delve into the implications of our findings, special aspects we had to consider when developing LUPINE, and suggest avenues for future research.

2 Materials and methods

2.1 Method

Our network inference methodology is tailored for the analysis of longitudinal microbiome data, where we assume that the dynamics of taxa interactions are temporal. In order to uncover these intricate relationships, we leverage both past and current information from the data in a sequential manner. Including past information in the model enables insights into inherent variations in microbial interactions across diverse temporal scenarios, including interventions. Our methodology includes three variants that are described in the following sections:

- Single time point modelling: given a pair of taxa, we estimate pairwise partial correlations while accounting for the influence of the other taxa using one-dimensional approximation with principal component analysis (PCA).
- Two time point modelling: given a pair of taxa, we estimate pairwise partial correlations while accounting for the influence of the other taxa with projection to latent structures (PLS) regression. PLS regression performs dimension reduction such that the covariance between the current and preceding time point data sets is maximised.
- Several time point modelling: given a pair of taxa, we estimate pairwise partial correlations while accounting for the influence of the other taxa using generalised PLS for multiple blocks of data ('blockPLS') regression. We use blockPLS to maximise the covariance between the current and any past time point datasets.

We assume that individuals within a specific group (e.g. control group) have a common network structure at a particular time point. Therefore, we consider each group separately for studies with multiple sample groups (as illustrated in case studies 1 and 3) to focus on each of their unique microbial characteristics.

Notations. In the following we denote $\boldsymbol{X}$ an $(n \times p)$ data matrix (where n is the number of samples and p is the number of taxa); $\boldsymbol{X}^i$ and $\boldsymbol{X}^j$ are columns in $\boldsymbol{X}$ for taxa i and j; $\boldsymbol{X}^{-(i,j)}$ is an $(n \times p-2)$ data matrix excluding taxa i and j columns in $\boldsymbol{X}$; $\boldsymbol{l}$ is vector of library sizes for n samples. In situations where time is of interest (i.e. in longitudinal modelling), all of the above matrices and vectors would include a subscript indicating the time point. For example, $\boldsymbol{X}_t$ represents an $(n \times p)$ data matrix at time point t.

Single time point modelling with PCA

We first introduce the single time point approach which we will generalise for multiple time points in the next sections. This type of analysis gives insights of microbial associations at a single time point. It can also be used when there is a large time gap between time points but only one particular time point is of interest.

To assess the strength of the association between two taxa, we use partial correlation – a statistical technique that measures association while controlling for other variables. For example, we estimate the partial correlation between a pair of taxa, denoted by i and j in Figure 1, to assess the strength of their association while controlling for other taxa. Controlling for these taxa requires regressing taxa i and j against $p-2$ taxa, which becomes unsolvable when the number of taxa p is larger than the number of samples n (e.g. in typical microbiome studies $p = 150 - 500$ and $n < 50$). Hence, the first step in our approach is to calculate a one-dimensional approximation of the control variables (e.g. all taxa except the pair i and j), through the first principal component.

We first start with a generic formulation of PCA (Jolliffe, 2002). The objective function to obtain the first principal component is

$$\underset{\|\boldsymbol{w}\|=1}{\operatorname{argmax}}\, var(\boldsymbol{X}\boldsymbol{w}) \tag{1}$$

where w is the p dimensional loading vector associated to the first principal component u, with $\boldsymbol{u} = \boldsymbol{X}\boldsymbol{w}$.

First, we modify Eq (1) to find the first principal component of the control taxa that excludes taxa i and j as

$$\underset{\|\boldsymbol{w}^{-(i,j)}\|=1}{\operatorname{argmax}}\; var(\boldsymbol{X}^{-(i,j)}\boldsymbol{w}^{-(i,j)}) \tag{2}$$

where $\boldsymbol{w}^{-(i,j)}$ is the $(p-2)$ dimensional loading vector associated to the first principal component $\boldsymbol{u}^{-(i,j)}$ with $\boldsymbol{u}^{-(i,j)} = \boldsymbol{X}^{-(i,j)}\boldsymbol{w}^{-(i,j)}$.

Next, we fit two log linear regressions on each of the taxa i and j. They are regressed against the first principal component, $\boldsymbol{u}^{-(i,j)}$, similar to principal component regression (PCR, Jolliffe (1982). PCR combines PCA and multiple linear regression, where the principal components of the explanatory variables serve as predictors rather than using the original variables directly in the regression. Here, we fit two log linear regressions to extract their residuals denoted by $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$:

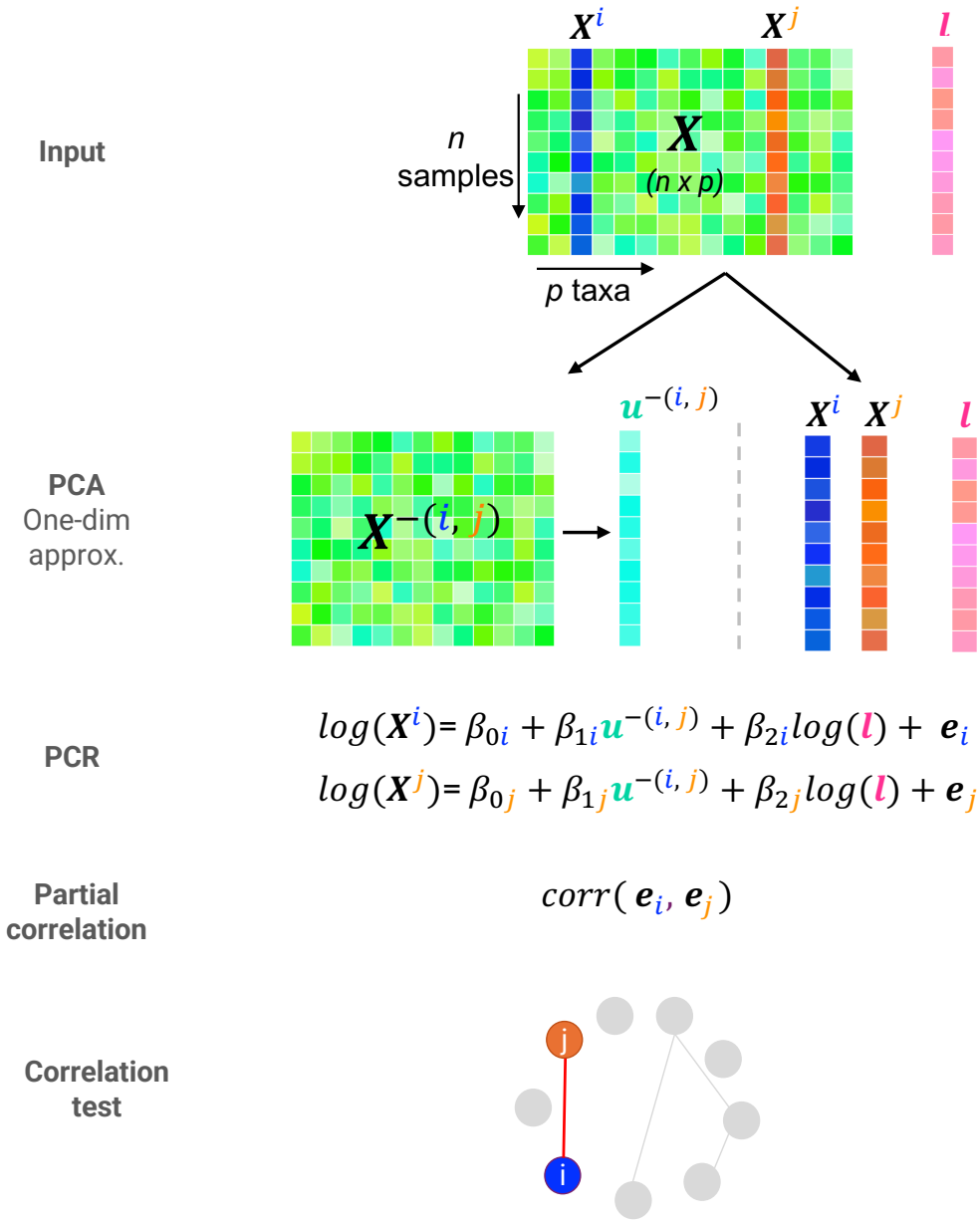


Fig 1. Single time point modelling overview. Consider the $(n \times p)$ input data matrix $\boldsymbol{X}$ and $\boldsymbol{l}$ the library size per sample. First, we use PCA to derive a one-dimensional approximation, $\boldsymbol{u}^{-(i,j)}$ to control for taxa other than i and j from the $\boldsymbol{X}^{-(i,j)}$ data matrix. We then apply two independent log linear regression models (a.k.a. Principal Component Regression) for $\boldsymbol{X}^i$ and $\boldsymbol{X}^j$, regressed on component $\boldsymbol{u}^{-(i,j)}$ and $\log(\boldsymbol{l})$ to obtain the residuals $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$. We estimate the partial correlation estimate for taxa i and j as the correlation between the two residuals. Finally, we test for the statistical significance of the true correlation between $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ to infer whether taxa i and j are connected in the network. We iterate this process for all pairs of taxa (i, j) to obtain a full binary network. Section 2.2 explains how to avoid repeated PCA approximations to improve computational time.

$$\boldsymbol{e}_i = log(\boldsymbol{X}^i) - E(log(\boldsymbol{X}^i)|\boldsymbol{u}^{-(i,j)}); \qquad \boldsymbol{e}_j = log(\boldsymbol{X}^j) - E(log(\boldsymbol{X}^j)|\boldsymbol{u}^{-(i,j)}) \tag{3}$$

where $E(.|\boldsymbol{u}^{-(i,j)})$ is the conditional expectation with respect to the first principal component $\boldsymbol{u}^{-(i,j)}$. We use a log linear regression as the response variables are counts. This is similar to the bias-corrected version of ANCOM (Lin and Peddada, 2020) that uses a linear regression framework based on log-transformed taxa counts.

In addition, to correct for library size (i.e. the variation in the total sequence reads), we modify Eq 3 and include an offset, similar to the approach of Zhang et al. (2018, 2020):

$$\boldsymbol{e}_i = log(\boldsymbol{X}^i) - E(log(\boldsymbol{X}^i)|\boldsymbol{u}^{-(i,j)}, log(\boldsymbol{l}); \qquad \boldsymbol{e}_j = log(\boldsymbol{X}^j) - E(log(\boldsymbol{X}^j)|\boldsymbol{u}^{-(i,j)}, log(\boldsymbol{l})) \tag{4}$$

where $E(.|\boldsymbol{u}^{-(i,j)}, log(\boldsymbol{l}))$ is the conditional expectation with respect to the first principal component $\boldsymbol{u}^{-(i,j)}$ and the log library size $\log(\boldsymbol{l})$.

We then quantify the strength of association between taxa i and j based on partial correlation using the residuals $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$. Partial correlations reveal direct (i.e. conditionally independent) associations between two variables while controlling for the other variables. We calculate the partial correlation between the pair of taxa i and j as

$$\pi_{\boldsymbol{e}_i,\boldsymbol{e}_j} = corr(\boldsymbol{e}_i, \boldsymbol{e}_j). \tag{5}$$

Finally, we test whether the true correlation $\rho_{\boldsymbol{e}_i,\boldsymbol{e}_j}$ between $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ (estimated with $\pi_{\boldsymbol{e}_i,\boldsymbol{e}_j}$) is significantly different from zero. We consider that a pair of taxa is connected in a network if the associated p-value from the correlation test is less than a chosen significance level (e.g. 0.05). We apply this approach to all possible pairs of taxa to obtain a binary network summarising the pairwise connections between all taxa.

In the pseudo-algorithm 1 and Figure 1, we outline the different steps in LUPINE_single. In Section 2.2, we explain how we improve computational time by avoiding repeated PCA approximations.

Algorithm 1 Single time point scheme

Input: $\boldsymbol{X}(n \times p)$: count matrix of n samples and p taxa

$\boldsymbol{l}$: vector of n library sizes

Output: $\pi_{\boldsymbol{e}_i,\boldsymbol{e}_j}$ (partial correlation) and $p_{\boldsymbol{e}_i,\boldsymbol{e}_j}$ (p-value) for all pairs of taxa (i, j)

for each pair of taxa (i, j) **do**
  Extract first principal component $\boldsymbol{u}^{-(i,j)}$ from $\boldsymbol{X}^{-(i,j)}$ using PCA
  Extract residuals, $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ by fitting log linear regressions f_i and f_j as:
    $f_i \rightarrow \log(\boldsymbol{X}^i) \sim \boldsymbol{u}^{-(i,j)} + log(\boldsymbol{l})$
    $f_j \rightarrow \log(\boldsymbol{X}^j) \sim \boldsymbol{u}^{-(i,j)} + log(\boldsymbol{l})$
  Compute partial correlation $\pi_{\boldsymbol{e}_i,\boldsymbol{e}_j} = corr(\boldsymbol{e}_i, \boldsymbol{e}_j)$
  Test significance of $\rho_{\boldsymbol{e}_i,\boldsymbol{e}_j}$ and extract p-value $p_{\boldsymbol{e}_i,\boldsymbol{e}_j}$
end for

Two time point modelling with PLS regression

Here we use PLS to obtain a one-dimensional approximation between two datasets from two time points (Figure 2). PLS is a projection-based method that maximises the covariance between the latent components associated to two data sets while managing correlated information (Wold et al., 2001). The PLS latent variables (or components), are linear combinations of variables from each data set. These latent variables aim to uncover subtle biological effects not evident otherwise if each data set is considered independently (Lê Cao and Welham, 2021).

The objective function for the first dimension of PLS is

$$\underset{\|\boldsymbol{w}_{t-1}\|=1,\|\boldsymbol{w}_t\|=1}{\operatorname{argmax}} \quad Cov(\boldsymbol{X}_{t-1}\boldsymbol{w}_{t-1}, \boldsymbol{X}_t\boldsymbol{w}_t) \tag{6}$$

where $\boldsymbol{w}_{t-1}$ and $\boldsymbol{w}_t$ are the two p dimensional loading vectors associated to the first latent components $\boldsymbol{u}_{t-1} = \boldsymbol{X}_{t-1}\boldsymbol{w}_{t-1}$ and $\boldsymbol{u}_t = \boldsymbol{X}_t\boldsymbol{w}_t$.

We modify Eq 6 to find the first latent components of the controlling taxa at time point t that has a maximum covariance with the preceding time point latent component as

$$\underset{\|\boldsymbol{w}_{t-1}^{-(i,j)}\|=1,\|\boldsymbol{w}_t^{-(i,j)}\|=1}{\operatorname{argmax}} \quad Cov(\boldsymbol{X}_{t-1}^{-(i,j)}\boldsymbol{w}_{t-1}^{-(i,j)}, \boldsymbol{X}_t^{-(i,j)}\boldsymbol{w}_t^{-(i,j)}) \tag{7}$$

where $\boldsymbol{w}_{t-1}^{-(i,j)}$ and $\boldsymbol{w}_{t}^{-(i,j)}$ are the two $(p-2)$ loading vectors associated to the first latent components $\boldsymbol{u}_{t-1}^{-(i,j)} = \boldsymbol{X}_{t-1}^{-(i,j)}\boldsymbol{w}_{t-1}$ and $\boldsymbol{u}_{t}^{-(i,j)} = \boldsymbol{X}_{t}^{-(i,j)}\boldsymbol{w}_{t}$.

To infer the network at time point t using the information from the preceding time point $(t-1)$, we fit two log linear regressions. These regressions are fitted on $\boldsymbol{X}_t^i$ and $\boldsymbol{X}_t^j$ that are regressed against the first latent component $\boldsymbol{u}_t^{-(i,j)}$, and the log library size $\log(\boldsymbol{l}_t)$. Note that $\boldsymbol{u}_t^{-(i,j)}$ differs from the single time point approach $\boldsymbol{u}^{-(i,j)}$ as we maximise the covariance with the preceding time point latent component $\boldsymbol{u}_{t-1}^{-(i,j)}$. The updated Equations (4) and (5) become

$$\boldsymbol{e}_{it} = log(\boldsymbol{X}_t^i) - E(log(\boldsymbol{X}_t^i)|\boldsymbol{u}_t^{-(i,j)}, log(\boldsymbol{l}_t)); \qquad \boldsymbol{e}_{jt} = log(\boldsymbol{X}_t^j) - E(log(\boldsymbol{X}_t^j)|\boldsymbol{u}_t^{-(i,j)}, log(\boldsymbol{l}_t)) \tag{8}$$

$$\pi_{\boldsymbol{e}_{it},\boldsymbol{e}_{jt}} = corr(\boldsymbol{e}_{it}, \boldsymbol{e}_{jt}) \tag{9}$$

where $E(.|\boldsymbol{u}_t^{-(i,j)}, log(\boldsymbol{l}_t))$ is the conditional expectation with respect to the first latent component $\boldsymbol{u}_t^{-(i,j)}$ and the log library size, $\log(\boldsymbol{l}_t)$ at time point t; and $\boldsymbol{e}_{it}$ and $\boldsymbol{e}_{jt}$ are the residuals from the two log linear models.

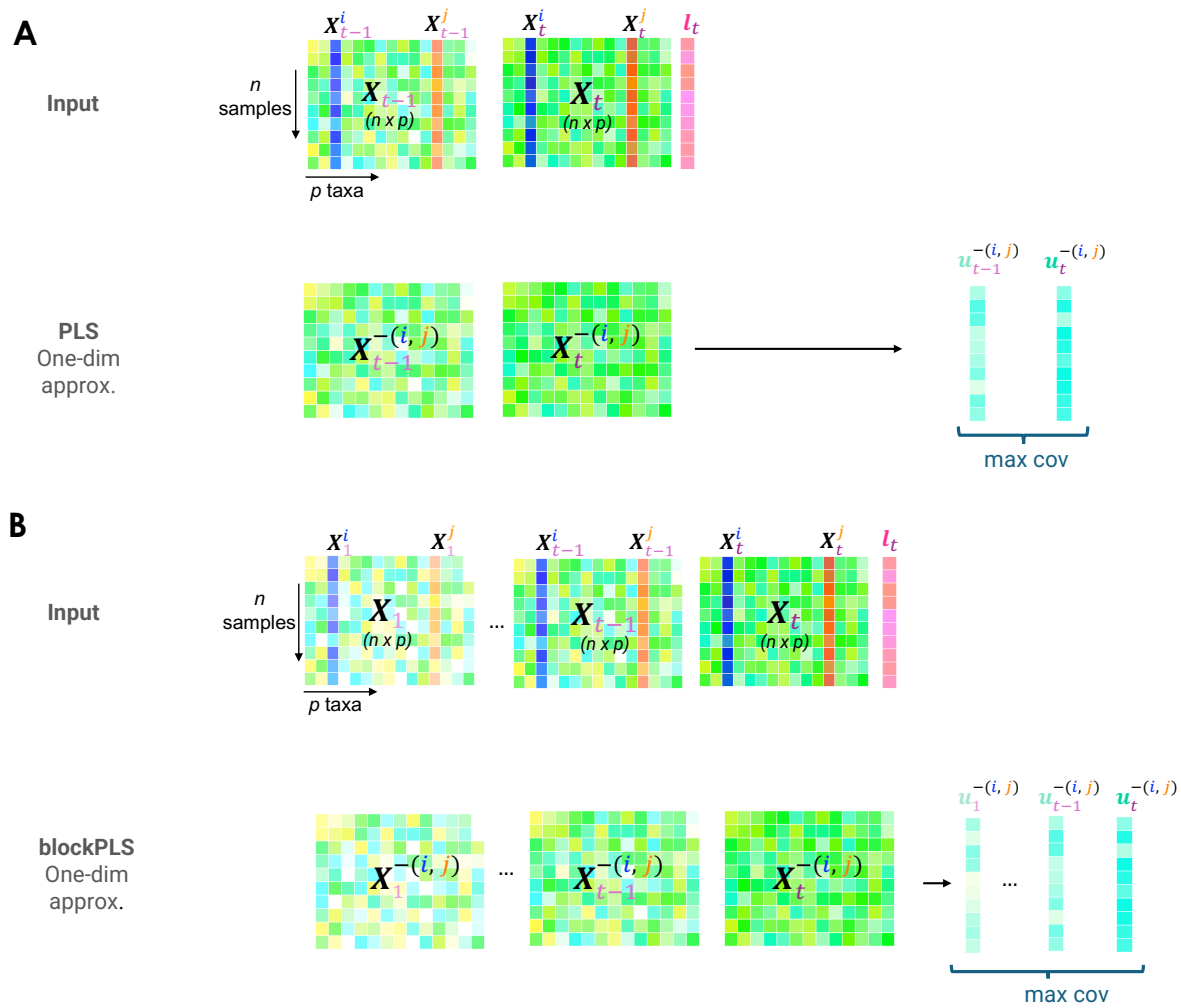
Similar to the single time point scheme described above, we test for the strength of association between taxa i and j with a correlation test, and obtain a binary network summarising the pairwise connections between all taxa after controlling for all other taxa from current time point and previous time point.

Multiple time point modelling with blockPLS regression

We use blockPLS, an extension of PLS, to calculate a one-dimensional approximation that maximises the covariance between latent components associated to the current time point, and all past time points. BlockPLS is based on generalised PLS (Tenenhaus and Tenenhaus, 2011, Lê Cao and Welham, 2021) that involves regressing the response matrix (e.g. $\boldsymbol{X}_t$) on multiple datasets $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{t-1})$. The objective function for the first dimension of blockPLS is

$$\underset{\|\boldsymbol{w}_1\|=1,\ldots,\|\boldsymbol{w}_t\|=1}{\text{argmax}} \sum_{q,k=1,q\neq k}^{t} Cov(\boldsymbol{X}_q w_q, \boldsymbol{X}_k \boldsymbol{w}_k) \tag{10}$$

where $\boldsymbol{w}_q$ is a p dimensional loading vector associated to the first principal component $\boldsymbol{u}_q$, where



C

Regression

$$log(\boldsymbol{X}_t^i) = \beta_{0it} + \beta_{1it}\boldsymbol{u}_t^{-(i,j)} + \beta_{2it}log(\boldsymbol{l}_t) + \boldsymbol{e}_{it}$$

$$log(\boldsymbol{X}_t^j) = \beta_{0jt} + \beta_{1jt}\boldsymbol{u}_t^{-(i,j)} + \beta_{2jt}log(\boldsymbol{l}_t) + \boldsymbol{e}_{jt}$$

Fig 2. Two or more time points modelling overview. Consider as input the $(n \times p)$ data matrices for **A** two time points or **B** more than two time points (denoted $\boldsymbol{X}_1 \ldots, \mathbf{X}_t$) and their respective library size $\boldsymbol{l}_t$ at time t. Similar to single time point modelling presented in Figure 1, first, we derive a one-dimensional approximation, $\boldsymbol{u}_t^{-(i,j)}$ to control for taxa other than i and j. Instead of using PCA as previously for a single time point, we use either **A** PLS maximising the covariance with the control taxa at the preceding time $t-1$, or **B** blockPLS maximising the covariance with the control taxa at all previous times $1, \ldots, t-1$. **C** We then fit two independent log linear models for $\boldsymbol{X}_t^i$ and $\boldsymbol{X}_t^j$, regressed on $\boldsymbol{u}_t^{-(i,j)}$ and $\log(\boldsymbol{l}_t)$ to compute the residuals $\boldsymbol{e}_{it}$ and $\boldsymbol{e}_{jt}$ following either **A** or **B**. The remainder of the modeling process, including the partial correlation calculation and the correlation test, is identical to that in Figure 1.

$\boldsymbol{u}_q = \boldsymbol{X}_q \boldsymbol{w}_q$.

We modify Eq 11 to find the first latent components of all controlling taxa at time point t that has a maximum covariance with previous time points latent components as

$$\underset{\|\boldsymbol{w}_1^{-(i,j)}\|=1,\ldots,\|\boldsymbol{w}_t^{-(i,j)}\|=1}{\operatorname{argmax}} \sum_{q,k=1,q\neq k}^{t} Cov(\boldsymbol{X}_q^{-(i,j)}\boldsymbol{w}_q^{-(i,j)}, \boldsymbol{X}_k^{-(i,j)}\boldsymbol{w}_k^{-(i,j)}) \tag{11}$$

where $\boldsymbol{w}_q^{-(i,j)}$ is a $p-2$ dimensional loading vector associated to the first latent component $\boldsymbol{u}_q^{-(i,j)}$ at time point q, with $\boldsymbol{u}_q^{-(i,j)} = \boldsymbol{X}_q^{-(i,j)}\boldsymbol{w}_q^{-(i,j)}$.

To infer a network at any time point t using all time points until time point t, we fit two log linear regressions on the counts of $\boldsymbol{X}_t^i$ and $\boldsymbol{X}_t^j$ against the first latent component at t, $\boldsymbol{u}_t^{-(i,j)}$, and log library size at t, $\log(\boldsymbol{l}_t)$. However, we need to keep in mind that the latent component $\boldsymbol{u}_t^{-(i,j)}$ maximises the covariance with all previous time points latent variables, $[\boldsymbol{u}_1^{-(i,j)}, \ldots, \boldsymbol{u}_{t-1}^{-(i,j)}]$, thus effectively taking into account all past time points. We then use Equations (8) and (9) to quantify and test the statistical significance of the partial correlation between taxa i and j.

Pseudo-algorithm 2 outlines the different steps in either two time points and multiple time points modelling.

Algorithm 2 Multi time point scheme

Input: $\boldsymbol{X}_t(n \times p)$: count matrix of n samples and p taxa at time t

$\boldsymbol{l}_t$: vector of n library sizes at time t

Output: $\pi_{\boldsymbol{e}_{it},\boldsymbol{e}_{jt}}$ (partial correlations) and $p_{\boldsymbol{e}_{it},\boldsymbol{e}_{jt}}$ (p-value) for all pairs of taxa (i, j) at time t

for $t = 2, \ldots, t$ **do**
  for each pair of taxa (i, j) **do**
    if two time modelling **then**
      Extract first latent component $\boldsymbol{u}_t^{-(i,j)}$ using PLS regression between $\boldsymbol{X}_t^{-(i,j)}$ and $\boldsymbol{X}_{t-1}^{-(i,j)}$
    else
      Extract first latent component $\boldsymbol{u}_t^{-(i,j)}$ using blockPLS regression between $\boldsymbol{X}_t^{-(i,j)}$ and $\boldsymbol{X}_{t-1}^{-(i,j)}, \ldots, \boldsymbol{X}_1^{-(i,j)}$
    end if
    Extract residuals, $\boldsymbol{e}_{it}$ and $\boldsymbol{e}_{jt}$ by fitting log linear regressions f_{it} and f_{jt} as:
      $f_{it} \rightarrow \log(\boldsymbol{X}_t^i) \sim \boldsymbol{u}_t^{-(i,j)} + log(\boldsymbol{l}_t)$
      $f_{jt} \rightarrow \log(\boldsymbol{X}_t^j) \sim \boldsymbol{u}_t^{-(i,j)} + log(\boldsymbol{l}_t)$
    Compute partial correlation $\pi_{\boldsymbol{e}_{it},\boldsymbol{e}_{jt}} = corr(\boldsymbol{e}_{it}, \boldsymbol{e}_{jt})$
    Test significance of $\rho_{\boldsymbol{e}_{it},\boldsymbol{e}_{jt}}$ and extract p-value $p_{\boldsymbol{e}_{it},\boldsymbol{e}_{jt}}$
  end for
end for

2.2 Improvement in computational time

Our approach requires to compute one-dimensional approximation for each pair of taxa (i, j), i.e. $p * (p - 1)/2$ times. Using matrix decomposition principles, we propose instead to approximate the full data matrix $\boldsymbol{X}$ with the first one-dimension component $\boldsymbol{u}$ that summarises most of the variation in $\boldsymbol{X}$, and its associated loading vector that are obtained from PCA, PLS, or blockPLS:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p} \approx \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1} \times \begin{bmatrix} w_1 & w_2 & \cdots & w_p \end{bmatrix}_{1\times p} \tag{12}$$

Using the approximation from Eq 12, we then can calculate the $\boldsymbol{u}^{-(i,j)}$ vector by shrinking the loading coefficients $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ of the pair of taxa (i, j) to zero:

$$\begin{bmatrix} u_1^{-(i,j)} \\ u_2^{-(i,j)} \\ \vdots \\ u_n^{-(i,j)} \end{bmatrix}_{n\times 1} \approx \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p} \times \begin{bmatrix} w_1 & w_2 & \cdots & w_i = 0 & \cdots & w_j = 0 & \cdots & w_p \end{bmatrix}^{\intercal}_{1\times p} \tag{13}$$

This allows us to obtain the vectors $\boldsymbol{u}$ and $\boldsymbol{w}$ only once. We show the efficacy of this approximation in Supplementary section S1.

2.3 Comparisons to other network inference methods

Existing longitudinal network methods either infer networks for a single individual or derive a single network from all data (Kodikara et al., 2022), which can be problematic when we expect changes over time. LUPINE infers networks sequentially to include the information learned across time. We compared our results with two state-of-the-art single time point network inference methods, SpiecEasi (Matchado et al., 2021) and SparCC (Matchado et al., 2021), using simulations where the true network is known. SpiecEasi and SparCC represent two distinct branches of network analysis. Specifically, SparCC employs a correlation-based approach, while SpiecEasi uses a conditional dependence approach. The SpiecEasi method further includes two methods based on different graphical model estimators, representing different types of network analysis: (i) neighborhood selection and (ii) sparse inverse covariance selection, which we will refer to as SpiecEasi_glasso' and SpiecEasi_mb', respectively.

2.4 Measures for network comparisons

We used two distinct measures to identify similarities and dissimilarities among the inferred networks: a network distance measure for quantitatively evaluating the network topology (see Figure 3B1 - B2), and

a node-wise measure for assessing the influential nodes in the network (see Figure 3C1 - C2).

Pairwise distance between networks

We use Graph diffusion distance (GDD) (Hammond et al., 2013) to measure pairwise differences in network typologies. GDD assesses the average similarity between two networks by exploring the information flow and connectivity within their structures, based on the heat diffusion process on graphs. The GDD score is calculated by searching for a diffusion time that maximises the Frobenius norm of the difference between the diffusion kernels. By calculating the GDD for all pairs, we obtain a pairwise distance matrix representing the network distances between pairs of graphs. We then transform these pairwise distances into a lower-dimensional space using classical Multidimensional Scaling (MDS) (Torgerson, 1958).

MDS maintains pairwise dissimilarities between objects. As a result, in the MDS plot, points that represent each network are deemed similar if they are positioned closely, and dissimilar if they are spaced further apart. This visual technique allows us to compare multiple networks at once.

Node influence

To assess influential nodes in the network, we use the Integrated Value of Influence (IVI) algorithm that integrates local (degree centrality and ClusterRank), semi-local (neighbourhood connectivity and local H-index), and global (betweenness centrality and collective influence) measures (Salavaty et al., 2020). IVI combines these measures to capture both local prominence as well as the broader impact of the nodes. After calculating the IVI values for all nodes, we perform a PCA on the resulting matrix, where the rows represent different networks, and the columns represent different nodes (here taxa). While MDS aims to preserve pairwise distances or dissimilarities as much as possible, PCA prioritise explaining the maximum variation in the data and highlighting strong patterns.

2.5 Simulation and case studies

In this section, we outline our simulation strategy and detail the three case studies analysed with LUPINE.

2.5.1 Simulation study

In our simulation, we create realistic network structures based on our high-fat high-sugar (HFHS) case study described in Section 2.5.2. We assume that the individual mice undergo a normal diet until day 5,

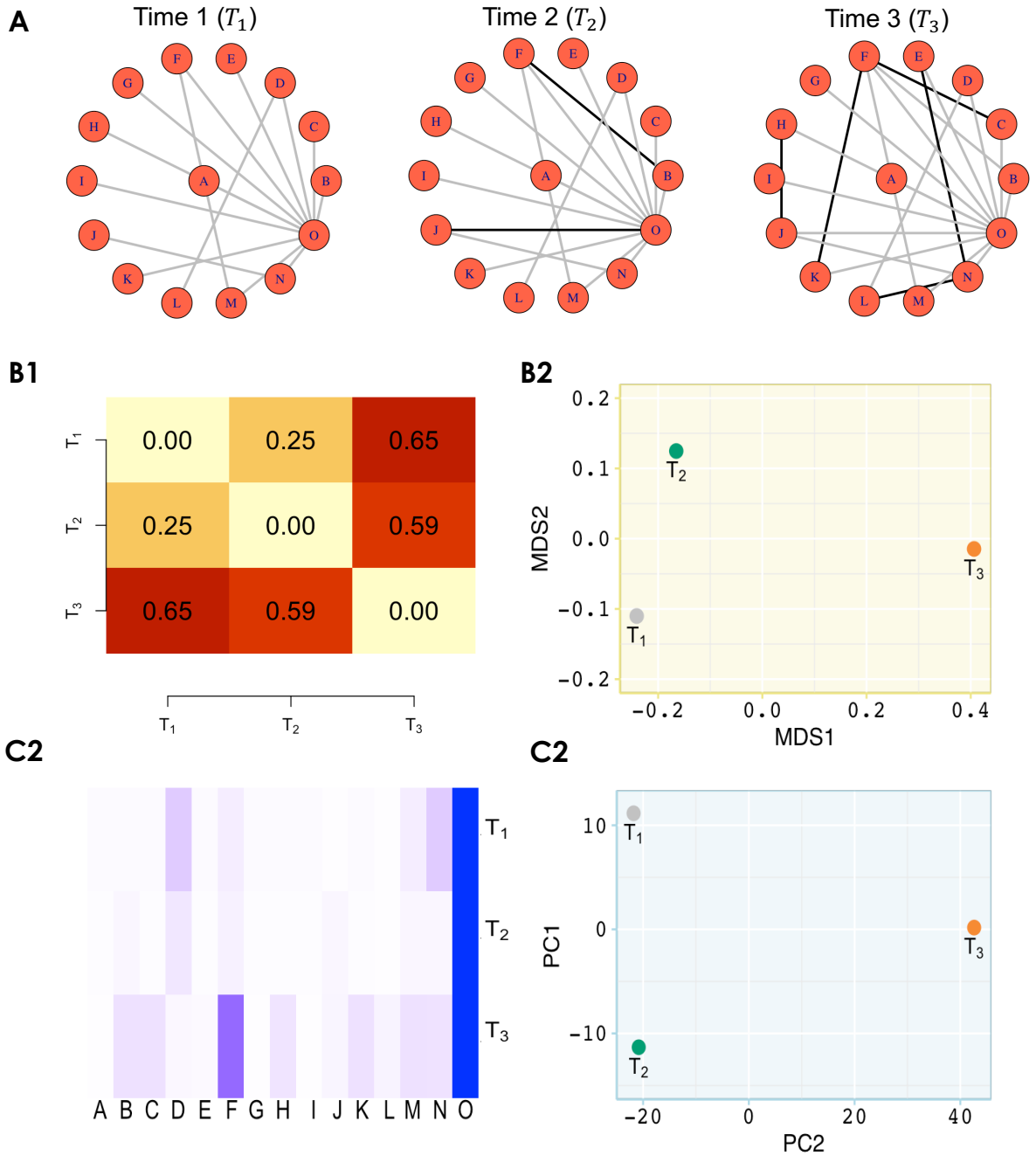


Fig 3. Toy example for network comparison. A Consider three networks across three time points, where nodes represent taxa and edges represent significant association between taxa. Black edges denote new connections compared to the preceding time point. **B** Network dissimilarity through graph diffusion distance (GDD). **B1** Pairwise GDD distances reveal that the network at T_3 is dissimilar to the networks at T_1 and T_2. **B2** Transformation of pairwise GDD into a lower-dimensional space using classical Multidimensional Scaling (MDS) represents the separation of time 3 network from the other two networks. **C** Network discrimination through influential nodes (i.e taxa). Node influence is computed by integrating local, semi-local, and global measures. **C1** Heatmap showing the node influence (more influential = darker blue shades). For example, node 'O' is influential in all time points, but node 'F' is more influential at T_3 rather than T_1 and T_2. **C2** Principal Component Analysis (PCA) plot to represent the networks based on their respective node influence. Similar to what we observed with the GDD distance, the largest source of variation is explained between the network at T_3 and the networks at earlier time points.

and then transition to a HFHS diet on day 6. This dietary shift is reflected in our simulation through two different networks: one from days 1 to 5 and another from days 6 to 10, indicating a network change on day 6. Both networks are inferred using the SpiecEasi method (Kurtz et al., 2015), applied to concatenated data from each respective diet. Due to computational constraints, we restricted our simulation to the sub-network of the *Bacteroidales* order, resulting in 54 nodes. The networks used across the two time periods are represented in Figure 4.

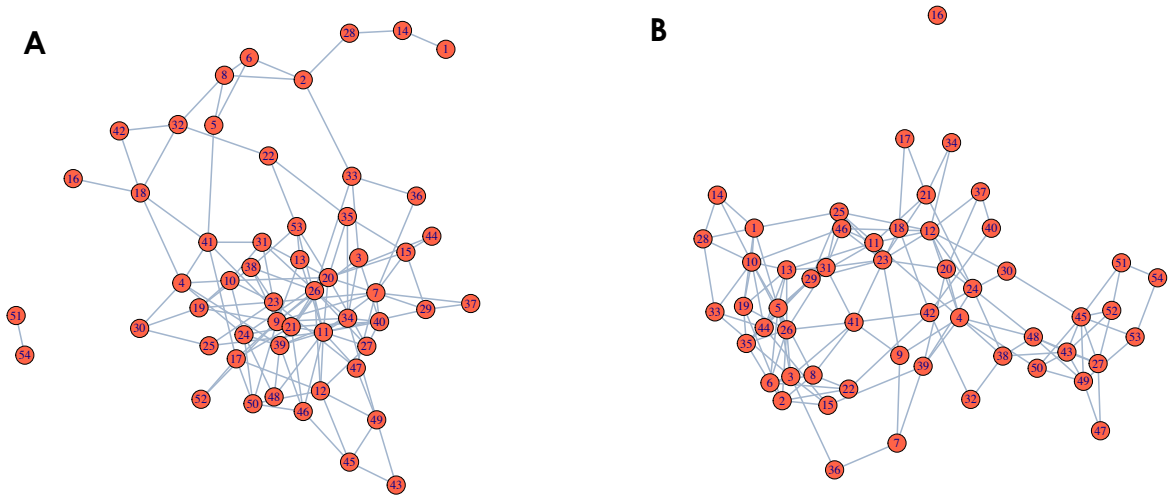


Fig 4. Networks used in simulation. Inferred network from SpiecEasi for **A** normal diet and **B** HFHS diet among *Bacteroidales*. These networks form the basis of our simulation for the two time periods (i.e. from day 1 to 5 and from day 6 to 10).

These networks are then used to generate data using a multivariate Poisson distribution. More specifically, for each new time point, we simulate new births with a multivariate Poisson process and update the taxa counts. To simulate 'death', we use the First-order Integer-valued Autoregressive (INAR(1)) mechanism, which models the decrease in taxa population between two time points. More details on the simulation strategy are provided in the Supplemental Section S2.

To explore the impact of sample size, we included three settings by varying the sample size from 23 to 50, and then to 120. For each sample size, 50 data sets were simulated, each with 52 taxa and 10 time points. In each simulation, we generated a new correlation matrix from the two networks given in Figure 4 to be used in the multivariate Poisson distribution and in INAR(1).

2.5.2 Case studies

We analysed three 16S rRNA amplicon datasets, the experimental designs of which are outlined in Table 1. In each study, we grouped data by 'group' and 'time point', and calculated the relative abundance of

each taxa within these specific groups and time points. Any taxa with a relative abundance less than 0.1% across all groups and all time points were excluded. This filtering process, while significantly reducing the number of taxa, aligns with the goal of conducting a more focused analysis (Sankaran and Jeganathan, 2023). However, we only calculated the partial correlations for taxa that had a mean relative abundance exceeding 0.1% at a given time point and group. As a result, there were no edges in the network from the taxa with low abundance at that time point and group. When we fit a longitudinal model on a given time point, we only consider the taxa filtered on that specific time point and group, and across the past time points.

Table 1. Summary of Experimental Frameworks: Key Parameters for Each Case Study. Number of taxa are indicated after the filtering step described in Section 2.5.2.

Name	Sample size	Number of taxa	Time points	Collection days	Interventions
High-fat high-Sugar (HFHS)	HFHS = 23 Normal = 24	212	4	(0, 1, 4, 7)	NA
Vancomycin-resistant Enterococcus faecium (VREfm) colonisation	9	126	11	(0-2, 5-9, 12-14)	Ceftriaxone antibiotic treatment (days 6-7) VREfm colonisation (day 9)
Diet study	Plant diet = 10 Animal diet = 10	311	15	(-4, -3, -2, -1 0, 1, 2, 3, 4 5, 6, 7, 8, 9, 10)	Diet intervention (days 0-4)

Case Study 1: HFHS (a four time point case-control mouse study)

We investigated the effect of a high-fat high-sugar (HFHS) diet on the mouse microbiome, as outlined in Susin et al. (2020). To assess the diet effect on the gut microbiome, 47 C57/B6 female black mice were fed with either an HFHS or a normal diet. Fecal samples were collected at Days 0, 1, 4 and 7.

The raw data contained 1,172 taxa across all four time points. We removed one outlier mouse from the normal diet group with a very large library size. After filtering, we modelled 102,107,105, and 91 taxa respectively for the normal diet for each day, and 99, 147, 85, and 92 taxa for the HFHS diet.

Case Study 2: VREfm (an eleven time point mouse study with two interventions)

Mu et al. (2020) investigated the functional roles of the gut microbiome during Vancomycin-resistant Enterococcus faecium (VREfm) colonisation. To assess the bacterial community composition in the murine gut, 9 C57BL/6 mice (co-housed wild-type males) were monitored. Fecal samples were collected over 14 days from the same three mice. Over the course of 14 days, the mice underwent two

interventions: (1) administration of ceftriaxone treatment at a concentration of 0.5g/liter in the drinking water over 2 days on days 6 and 7, and (2) colonisation with a dosage of 1×106 VREfm ST796 on day 9. The days before the antibiotic administration were considered as the naive phase of the experiment. The two days when the antibiotic was given were considered the antibiotic phase, and all other days after that were considered the VRE phase.

The data contained 3,574 taxa across eleven time points. Day 8 was excluded from the analysis due to insufficient taxonomic counts for two mice. After filtering, we considered the following number of taxa in each phase: naive (53, 63, 58, 72), antibiotic (34, 22) and VRE (10, 15, 15, 15).

Case Study 3: Diet (a fifteen time point case-control human study with one intervention)

David et al. (2014) investigated the human gut microbiome responses to a short-term diet intervention. To evaluate the dynamics of the diet intervention, they assigned 10 participants to each diet intervention, namely the 'plant-based' and 'animal-based' diets for five consecutive days. The study participants were observed for four days before the diet intervention, a period referred to as the 'baseline period' to study their regular eating habits. After intervention, a six-day 'washout period' was also observed to assess microbial recovery. Data were obtained from the MITRE repository (Bogart et al., 2019). The data contained 17,310 taxa across all time points. As taxonomic details were lacking, we used the sequence information along with dada2 reference database "silva_nr99_v138.1_wSpecies_train_set.fa.gz" for taxonomic classification (Callahan et al., 2016).

After filtering, we retained the following number of taxa in the plant-based diet group for each period: baseline (105, 100, 111, 103), intervention (118, 116, 116, 114, 114), and washout (121, 120, 90, 96, 115, 109); and in the animal-based diet group: baseline (120, 109, 128, 115), intervention (113, 113, 108, 105, 110), and washout (112, 104, 106, 105, 107, 123). Across the duration of the experiment, only 8-15 samples were obtained from each participant because of unsuccessful sampling at certain time points. Thus, we imputed the missing values using a cubic splines on the centered log ratio (clr) transformed data. This was motivated by the linear interpolation and cubic splines interpolation adopted in Sankaran and Jeganathan (2023), Ruiz-Perez et al. (2021). Since, clr transformation accounts for library size differences, we used Eq 3 instead of Eq 4 in our modelling.

3 Results

We first evaluated our two approaches against two widely recognised network inference methods (described in Section 2.3) for cross sectional microbiome data. Using simulated data, our objective was to demonstrate the effectiveness and computational efficiency of our proposed approaches compared to commonly used single time point methods. We then highlighted the advantages of LUPINE over its single time point counterpart LUPINE_single in the HFHS study.

After this benchmark step, we applied LUPINE in three case studies. In the longitudinal case-control mouse study, we evaluated whether the difference in diets could be detected in our network models. In the mouse study with two interventions, we evaluated the effectiveness of LUPINE in detecting abrupt changes. Finally, in the human case-control study, we assessed the ability of LUPINE to handle inherent high variability typical of human studies across fifteen time points.

Benchmarking analysis

Simulation study: LUPINE methods outperformed existing single time point methods.

We first compared our proposed longitudinal method (LUPINE) and its single time point version (LUPINE_single), with two SpiecEasi models and SparCC. Note that all methods except LUPINE were designed for single time point modeling.

We assessed model performance across 50 simulated datasets, focusing on scenarios where the number of samples was smaller than the number of variables. We used two metrics: the receiver operating characteristic (AUC-ROC) curve and the precision-recall characteristic (AUC-PRC) curve. Higher values in both metrics signified better model performance, with AUC-ROC evaluating overall discriminatory power and AUC-PRC evaluating precision and recall characteristics.

Our proposed methods and SparCC outperformed SpiecEasi with higher AUC-ROC values. Additionally, both our approaches were computationally more efficient than SparCC (Figure 5B). However, all methods had similar AUC-PRC values, suggesting a similar level of precision-recall trade-off (Figure S4). When the sample size increased, LUPINE and LUPINE_single led to superior model performance based on AUC-ROC and AUC-PRC values (Figures S5 and S6).

Next, we compared the inferred networks across time points, as explained in Section 2.4. For SparCC and our methods, taxa connections were determined via a correlation test (this was not required for SpiecEasi methods, as they inherently output a binary network).

To compare networks based on structure, we computed the pairwise GDDs for all inferred networks

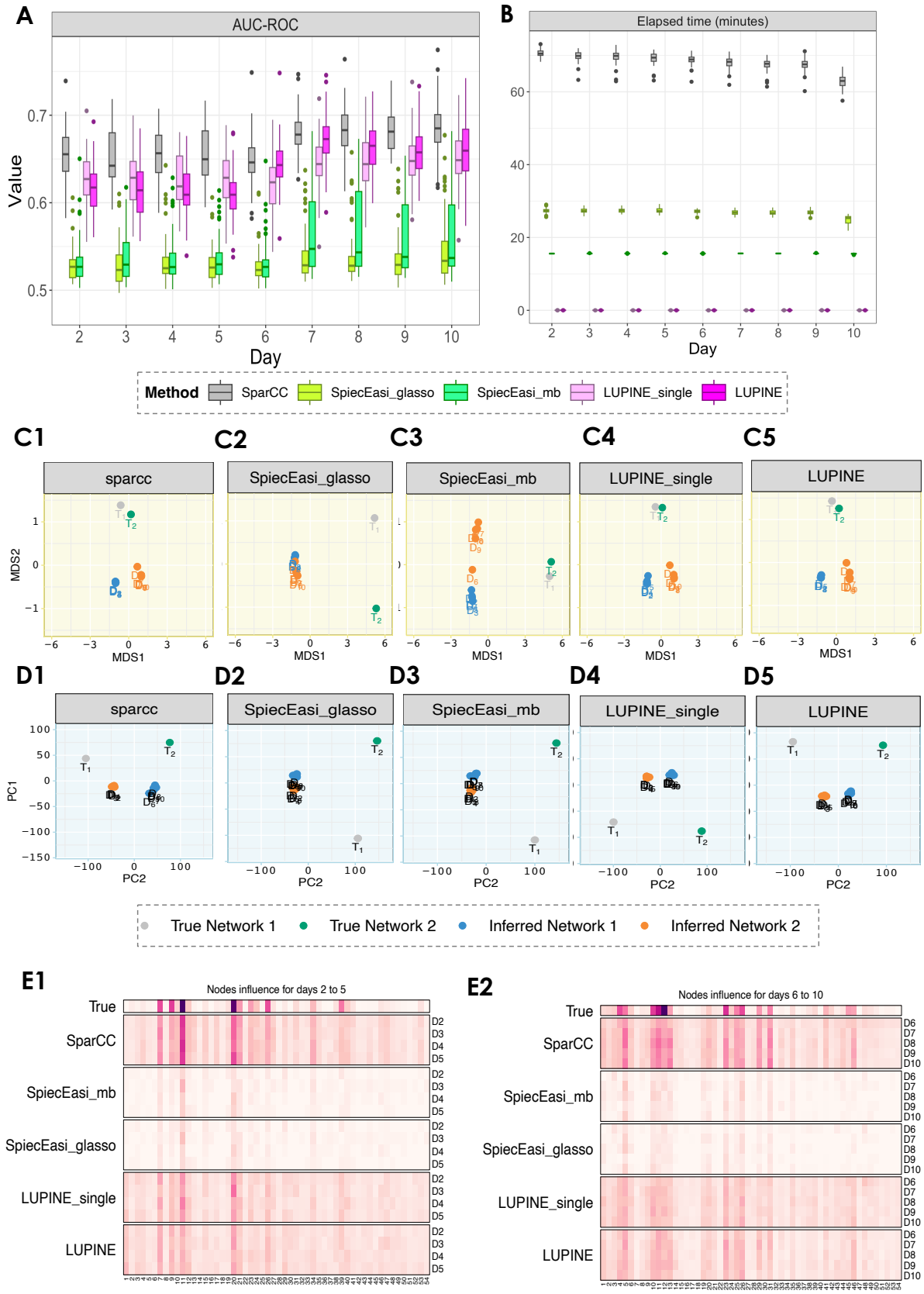


Fig 5. (Caption next page.)

Fig 5. (Previous page.) **Simulation results (n = 23)**. **A** Box plots of the Area Under the Receiver Operating Characteristic (AUC-ROC) over time for different methods. Both LUPINE models and SparCC outperformed the two SpiecEasi methods based on AUC-ROC values. **B** Elapsed time for each method, showing the superiority of the two LUPINE methods. **C1-C5** MDS plots of the pairwise GDD for the inferred networks from the different methods. All but SpiecEasi methods distinguish networks based on the two periods (days 2-5 and 6-10) in the first coordinate, regardless of the network being true or inferred. **D1-D5** PCA plots of the IVI scores for the inferred networks for each method. The largest node influence variation occurs across days, with a clear contrast between the two periods (days 2-5 and 6-10) for both LUPINE models and SparCC. Heatmaps of mean IVI scores for each node **E1** from day 2 to 5 and **E2** from day 6 to 10. The first row displays IVI scores for true networks from 4. All methods identified the most influential nodes.

across time points, then averaged these distances for each time point across all simulations. In our simulation (described in Section 2.5.1), we generated two stable networks that were distinct, across two time periods (days 2-5 and 6-10). Thus, we expect the inferred networks to be similar within each period, and dissimilar between the two periods. However, SpiecEasi displayed the largest variation between the inferred and the actual networks in the MDS plot of the pairwise GDDs (Figures 5C2, 5C3). In contrast, our two proposed methods (Figures 5C4, 5C5) and SparCC (Figure5C1) showed the largest between-day variation, regardless of whether the networks were true or inferred.

We applied the same procedure used for GDDs to conduct node influence based comparisons based on IVI values. Similar to the MDS plots, SparCC (Figure 5D1) and LUPINE methods (Figures 5D4 and 5D5) explained the greatest between-day variation in the PCA plots. While all methods identified the most influential nodes (Figures 5E1 and 5E2), the two SpiecEasi methods showed less pronounced influential nodes due to their sparse network inference, compared to the other methods.

HFHS study: LUPINE highlighted more robust longitudinal network patterns than LUPINE_single.

We compared LUPINE and LUPINE_single on the HFHS study. This study includes a small number of time points (Fig 7A), where mice were subjected to either a HFHS or a normal diet.

Both LUPINE_single and LUPINE differentiated the diet groups in the first axis of the MDS and PCA plots (Figure 6). For LUPINE_single, microbial networks from the normal diet group and the HFHS diet at day 0 were clustered, indicating a similar network structure at the onset of the two different diets (Figure 6A2). However, we did not observe a close cluster of microbial networks within the normal diet networks in the PCA plot of the IVI scores (Figure 6A2), suggesting that the node influence within the normal diet group is changing over time.

In contrast, for LUPINE, the groupings in the PCA plot (Figure 6B1) closely matched to those observed in the MDS plot (Figure 6B2), indicating consistent network results based on GDD and IVI

scores. Thus, these results suggest that LUPINE is able to model robust longitudinal patterns. Note that the initial day 0 microbial network is absent from the longitudinal scheme as we require at least one prior time point for inference.

In this benchmarking section, we first demonstrated the superior model performance of both LUPINE and LUPINE_single using AUC-ROC and AUC-PRC values. We then highlighted the advantages of LUPINE in comparison to our own LUPINE_single approach. We showed that LUPINE led to robust longitudinal network patterns. The following sections focus on the biological interpretation of the microbial communities identified by LUPINE in three case studies.

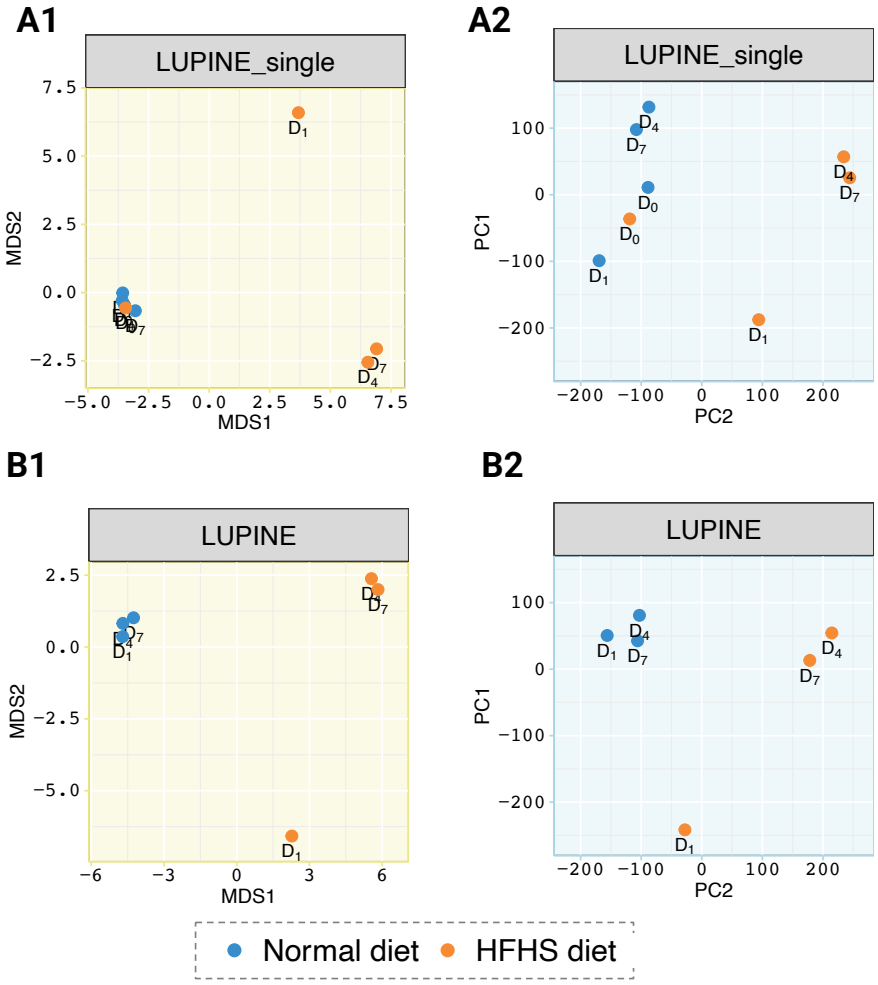


Fig 6. Network comparison results for LUPINE_single and LUPINE. A1-B1 MDS plot of pairwise GDD and **A2-B2** PCA plot of the IVI scores illustrating network similarities. **A** from LUPINE_single scheme and **B** from LUPINE longitudinal scheme. Compared to LUPINE_single, LUPINE exhibits consistent grouping patterns for both types of measures.

HFHS study: LUPINE identified taxonomic orders that differentiate microbial networks between different diet groups in mice across four time points

We identified patterns in the microbial network plots that distinguished the two diet groups (Figure 7B). In the normal diet group, we observed denser connections among nodes within the *Bacteroidales* order, particularly on days 4 and 7, compared to the HFHS diet group. Additionally, nodes within the *Lactobacillales* order exhibited a higher number of connections in the HFHS diet group, specifically with nodes within the *Erysipelotrichales* and *Clostridales*. Similar inferences can also be made from the IVI scores in Figures 7C1 and 7C2: In the normal diet group, we observed decreased IVI scores across days for nodes within the *Clostridales* order, with the majority having a zero IVI value. Compared to the normal diet group, the HFHS diet group networks had a higher node influence in *Lactobacillales* order (Figure 7C2). While these findings do not imply causation, they are consistent with previous studies that reported decreased relative abundances of *Bacteroidales* and enrichment of *Lactobacillales* in mice that had been fed with a high-fat diet (Guo et al., 2017, Yin et al., 2018), suggesting a potential influence of diet on these taxonomic orders. Daniel et al. (2014) also found that the proportions of *Lactobacillales* and *Erysipelotrichales* were higher in mice fed with high-fat diet.

To summarise, LUPINE analysis highlighted two different microbial community networks between normal and HFHS diets, with an increase in connections in the taxa nodes belonging to *Lactobacillales* in mice that were fed with a HFHS diet.

VREfm study: LUPINE highlighted changes in the network structure across two abrupt interventions in a mouse study with 11 time points

In the VREfm (Figure 8A), mice underwent two interventions: antibiotic administration and VRE colonisation. Similar to the first case study, this is a much more controlled experiment compared to human studies. However, in this study the time periods are divided into three phases: naive, antibiotic, and VRE, reflecting different stages of the experiment. In contrast to the first case study, where the objective was to identify the group differences the objective here is to explore network differences between the phases.

During the antibiotic phase, the number of edges in the inferred networks decreased (Figure 8B). However, during the VRE phase, the number of edges increased, but was not observed across all taxonomic orders. Specifically, we observed an increased number of edges in the nodes belonging to

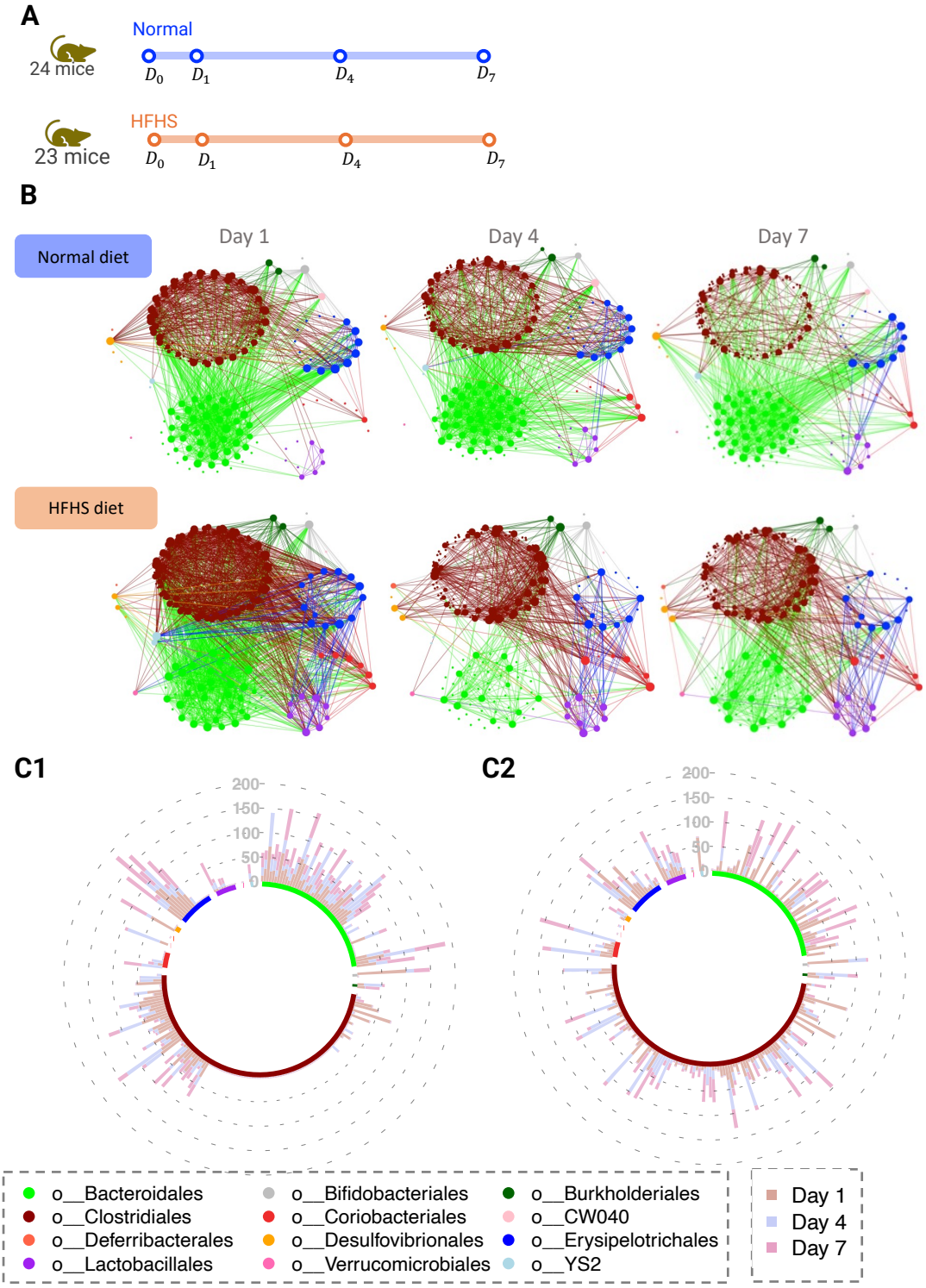


Fig 7. (Caption next page.)

Fig 7. (Previous page.) **HFHS case study A** Graphical representation of the study time line. **B** Inferred microbial networks over time for both normal group and HFHS diet groups. In the HFHS diet group, we observe a decrease in connections among nodes belonging to the *Bacteroidales* order over time. In the HFHS diet group, nodes belonging to the *Erysipelotrichales* order are more connected to those from the *Bacteroidales* order, but in the normal diet group, but we observe that nodes belonging to the *Erysipelotrichales* order are more connected to those from the *Lactobacillales* order. **C** Circular stacked bar plot of the IVI scores for **C1** normal diet and **C2** HFHS diet groups show a reduction in IVI scores for nodes associated with the *Clostridales* order in the normal diet group compared to the HFHS diet group.

Bacteroidales. In contrast, and even at the end of the experiment, the nodes belonging to *Clostridiales* had very few connections compared those detecting in the naive phase. From the antibiotic phase to day 9, two nodes belonging to taxonomic order *Streptophyta* and *Rickettsiales* exhibited higher connections. This finding is in agreement with Lankelma et al. (2016), who studied the disruption of the gut microbiota after the administration of broad-spectrum antibiotics and found a strong relative increase in bacteria of the genus *Streptococcus* across all subjects. Therefore, the increased connections of *Streptophyta* and *Rickettsiales* could potentially be attributed to the impact of antibiotics on the gut microbiota.

The MDS plot of the pairwise GDD revealed a clear separation between the naive phase, antibiotic phase, and VRE phase networks (Figure 8C). Within the VRE phase, we observed a temporal progression, indicating a unique network structure on day 9, which marked the initiation of VREfm colonisation. The PCA plot of the IVI scores (Figure 8D) showed a similar grouping structure to the MDS plot. The similarity between the two plots indicated a consensus in the network grouping, regardless of the two metrics employed (as we showed previously in the HFHS study). The naive phase networks IVI scores were well separated from the other phases, accounting for 35% of total variation of the scores. After the naive phase, we observed a decrease in the influential nodes for the majority of taxa. However, taxa belonging to the *Enterobacterales* and *Lactobacillales* orders became more influential following the antibiotic phase (Figure 8E). In Kodikara et al. (2022), we also found that these taxa belonging to families *Enterobacteriaceae* and *Enterococcaceae* showed a group difference between the naive and the VRE phases. We also noted a decrease in the IVI scores for *Clostridiales* nodes following the naive phase, a decrease that persisted until the end of the experiment. This taxonomic order was a key taxa group that has been studied for its ability to restrict gut colonisation by *Enterobacteriaceae* pathogens in mice, as shown by Byndloss et al. 2017, Kim et al. 2017, Djukovic et al. 2022.

To summarise, LUPINE highlighted a reduction in the number of edges after antibiotic treatment, affecting nodes belonging to the *Clostridiales* order. Interestingly, the reduction in *Clostridiales* node connections persisted even after the antibiotic phase, suggesting that the VRE colonisation had an

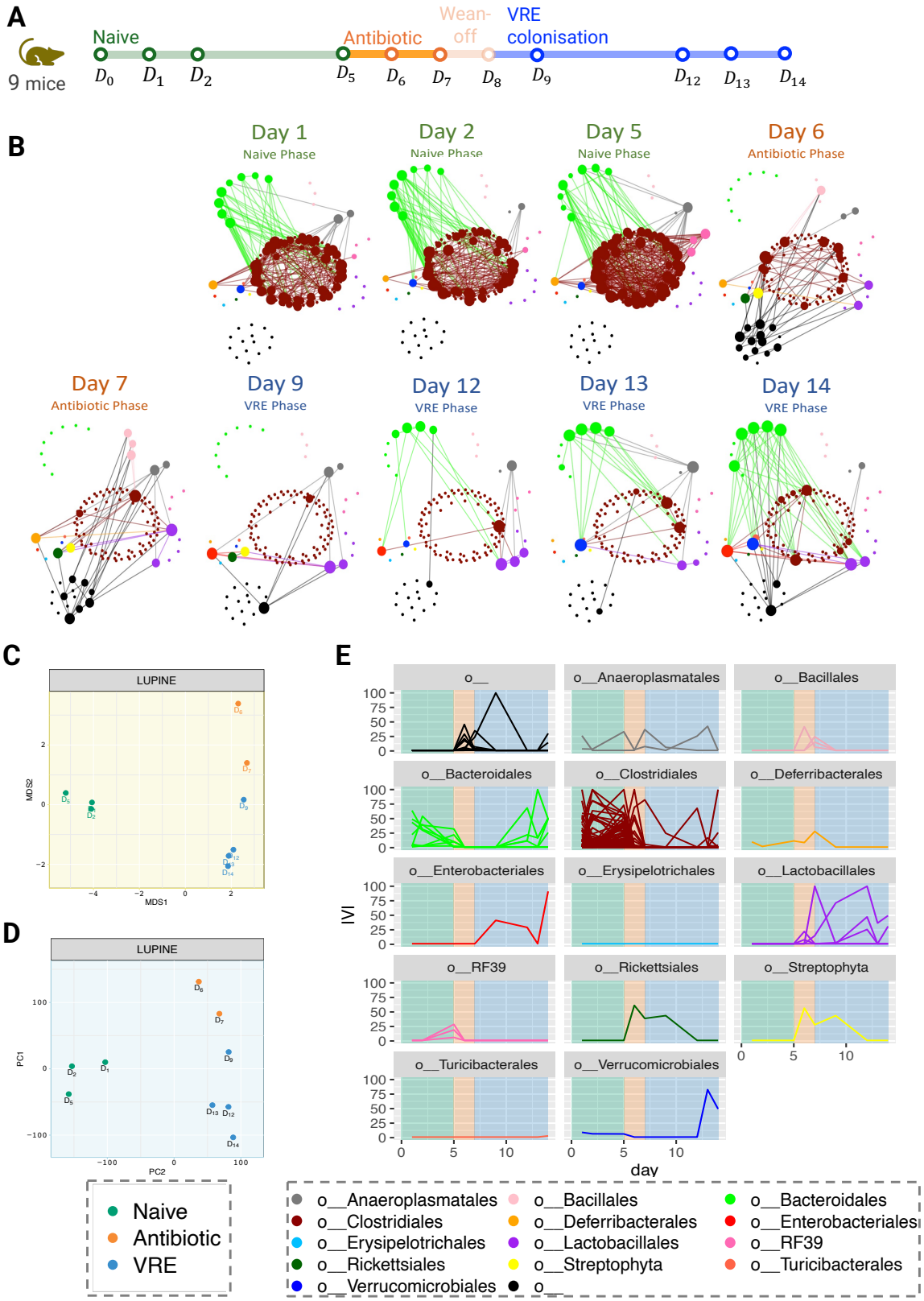


Fig 8. (Caption next page.)

Fig 8. (Previous page.) **VREfm case study. A** Graphical representation of the study time line. **B** Inferred networks across time. After antibiotic treatment on day 6, we observe a reduction in the associations in the nodes belonging to the *Clostridales* order, which are not recovered until day 14. However, the connections among nodes of the *Bacteroidales* order appear to recover from day 12. **C** MDS plot of the pairwise GDD shows that the network structures are significantly different for each phase of the experiment. **D** PCA plot of the IVI scores exhibits a grouping pattern consistent with the MDS plot. **E** Line plots of the IVI scores for each taxa across time, grouped by their taxonomic order shows a reduction of influential nodes after antibiotics treatment, indicating a less dense network structure

impact on the recovery of these connections. However, further studies are needed to assess whether these connections would be recovered without VRE intervention.

Diet study: LUPINE detected diet-specific and diet-stable taxonomic groups in a case-control human study spanning accross 15 time points

The Diet case-control study differs from the previous ones as it involves humans and numerous time points (Fig 9A) and a small number of participants. This results in expected higher variability than mouse studies.

The plant-based diet group differed from the animal-based diet group, with the former showing a tighter network grouping, indicating more significant changes in the animal diet intervention (Figure 9B). In the animal diet group, two clusters were observed. The first included networks from days 2 to 4, and the second from days -3 to 1 and 7 to 10, with the network on day 6 distinct from both clusters. This suggests that the network structure returned to its pre-intervention form two days into the washout period. A high variability was also noted across the animal diet group networks in the PCA plot (Figure 9C). However, the PCA plot revealed a less distinct separation between the two network groupings within the animal diet group compared to the MDS plot. This indicates a larger difference in the inferred networks, due to information flow or structure, compared to the IVI score variation. Furthermore, we noticed a close alignment in the baseline inferred networks across both groups. This implies a similar degree of node influence during the baseline phase for both groups.

We examined in more detail the six inferred networks from the two diet groups at three different days, each reflecting the final day of each phase: baseline, intervention, and washout (Figure 9D). We highlighted few taxa nodes in the networks at day 4 that differed between the two diet groups. On day 4, we observed the nodes belonging to *Clostridia UCG 014* had a comparatively higher degree in the plant based diet group than in the animal based diet group. *Clostridia UCG 014* is a bacterial taxon belonging to the class *Clostridia* in the phylum *Firmicutes*. While Pessoa et al. (2023) found that

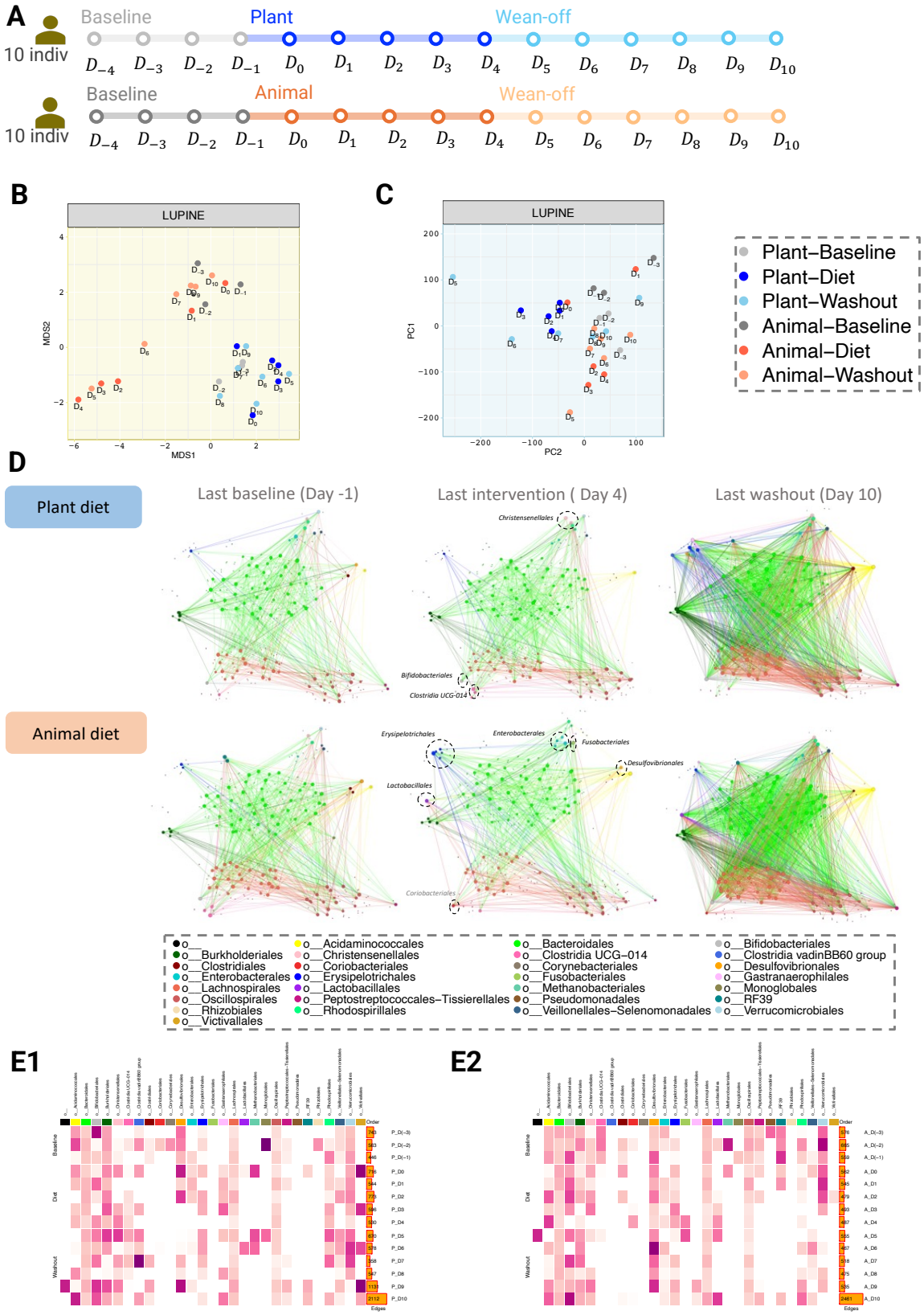


Fig 9. (Caption next page.)

Fig 9. (Previous page.) **Diet case study. A** Graphical representation of the study time line. **B** MDS plot of pairwise GDD shows that distinct patterns emerge based on diet groups (tighter cluster for plant-based networks compared to animal-based networks). **C** PCA plot of the IVI scores show that inferred networks at baseline are similar regardless of diet, indicating similar node wise importance.**D** Inferred networks across two diet groups and three time points (last day of baseline, intervention and washout phase). On day 4, the plant-based network shows increased connections in nodes associated with *Christensenellales*, *Christensenellales*, and *Clostridia UCG 014* compared to the animal-based network, which exhibits increased connections in nodes related to *Erysipelotrichales*, *Lactobacillales*, *Coriobacteriales*, *Enterobacterales*, *Fusobacteriales* and *Desulfovibrionales*.**E** Heatmap of the average IVI score for each taxonomic order for **E1** plant based and **E2** animal based diet groups, with darker pink representing high average IVI value. Nodes from *Bacteroidetes*, *Lachnospirales* and *Oscilospirales* consistently exhibit a non-zero IVI score, indicating their stable influence, unaffected by diet or daily variations.

Clostridia UCG 014 was reduced in the high-fat diet, Li et al. (2021) associated the decrease in *Clostridia UCG 014* with obesity. Additionally, He et al. (2022) found a positive correlation between *Clostridia UCG 014* and blood glucose levels, and Jiang et al. (2023) inferred *Clostridia UCG 014* as one of the taxa potentially linked to cholesterol regulation. Nodes belonging to *Lactobacillales* had comparatively higher degree in the animal based diet group compared to the plant based diet group. The original study also found that the reads associated with *Lactococcus lactis* and *Pediococcus acidilactici*, which belong to the taxonomy order *Lactobacillales*, significantly increased in the animal based diet. This is in agreement with the review from Lee et al. (2023) reporting an increased level of *Lactobacillus*, a member of the *Lactobacillales* order, resulting from the consumption of white meat protein (such as chicken and fish) and dairy products (like milk, cheese, yogurt, and kefir). By comparing these taxa with the average IVI scores across the intervention period, we observed some agreement between a high average IVI score (Figures 9E1 and 9E2) and a high number of connections in the networks in Figure 9D. For instance, *Clostridia UCG 014* had a non zero IVI average in the plant based diet group from day 2, whereas the IVI score in the animal based diet group was zero specifically during the intervention period (day 0 to day 4). We also observed that the nodes belonging to taxa order *Lactobacillales* had a non zero IVI average in the animal based diet group from day 2, whereas the IVI score in the plant based diet group was zero specifically during days 1 to 5. In the heat-maps, the bar plots revealed a high number of edges on the last day (i.e., day 10) compared to all other days in both groups. No clear reasoning can be given for this observation as there was no change at this particular time point. However, this network did not stand out in our network comparison, indicating that even though the number of connections increased, the network structure and the influential nodes on day 10 were similar to those on day 9. A densely connected network at the last time point was also visible in the network plots in Figure 9D.

To summarise, LUPINE applied to a case-control human study with intervention identified several

taxonomic associations that were specific to plant based diet group or animal based diet group. We also identified few taxonomic orders such as *Bacteroidetes*, *Lachnospirales* and *Oscilospirales* that were influential regardless of time, or diet intervention. This suggests that these taxa may exhibit stability in the face of dietary changes.

4 Discussion

We developed LUPINE to detect the stability of taxa associations within a microbial community over time, and the responses of these microbial communities to external disturbances such as dietary changes and medication. To the best of our knowledge, LUPINE is the first sequential microbial network inference approach for a longitudinal setting.

In LUPINE, we combined the concepts of low-dimensional approximation with partial correlation to infer networks across time points. We then used GDD and IVI metrics to identify any abrupt network changes across time, groups, and key taxa nodes in each of the networks. Note that a requirement for LUPINE is that samples match across time points, as LUPINE assumes correlation structure across time and individuals. In the case where different individuals are sampled across time, a better approach for network inference is our variant LUPINE_single applied to each time point.

In our simulation study, we demonstrated that LUPINE and LUPINE_single improved model performance and were computationally more efficient compared to existing network modelling approaches SparCC and SpiecEasi that were designed for single time point analyses. We also demonstrated the applicability of LUPINE in case studies with different study designs. The two controlled mouse studies were either case-control or intervention studies while the third study was a more complex human study incorporating elements of both a case-control and an intervention design midway through the experiment, with a large number of time points (15). In all case studies, our LUPINE analyses extracted meaningful biological insights, including clustering patterns in the inferred networks that were coherent within each study design.

While developing LUPINE, we had to consider several aspects related to the characteristics of microbiome data. First, taxa filtering is an important data processing step in this kind of analysis. We only retained taxa with a mean relative abundance exceeding 0.1% for any group at any time point to obtain a consistent set of taxa being examined across various time points and groups. Further, to avoid identifying connections between taxa with low abundance at certain time points in specific groups, we then only examined taxa with a mean relative abundance greater than 0.1% within each group and time

point in the inferred networks. Second, microbiome studies are typically low sample size, which are likely to result in false positive associations. We used a correlation test that is appropriate for small sample size when calculating test statistics. We chose to identify significant associations based on p-values with an arbitrary cutoff of 0.05, rather than a resampling based model selection approach (used in SpiecEasi (Kurtz et al., 2015)) or a bootstrap-based p-value calculation approach (used in SparCC (Friedman and Alm, 2012)) that are computationally expensive. Adjustments for false discovery on p-values could also be considered to obtain sparser networks. However, our experience has shown that these adjustments may lead to empty networks in some situations.

We have identified several potential extensions of LUPINE that could enrich longitudinal analysis of microbiome data. Firstly, when considering studies with several groups (e.g. HFHS and Diet studies), LUPINE must be applied to each group separately if we assume that the true networks within these groups differ from one another. A potential extension of LUPINE could include all groups together to infer a common network across all groups. Secondly, LUPINE main focus is to identify associations between taxa at specific time points, while taking into account information from previous time points. Therefore, we do not model time gaps between points. Future developments could include approaches that give more weight to closer time points than to distant time points. Finally, we have focused our analyses on amplicon sequencing for illustrative purposes, but similar analyses can be performed on metagenomic sequencing data, which not only provide taxonomy profiles represented by species abundance information but also functional profiles represented by gene abundance information. An interesting extension of LUPINE to fully harness these complex data would be to establish connections between these two layers, akin to multilayer networks at each time point. This advancement could significantly broaden our understanding of microbiome dynamics, opening up new possibilities for research and discovery.

Acknowledgments

KALC was supported in part by the National Health and Medical Research Council (NHMRC) Career Development fellowship (GNT1159458).

Availability of data and materials

All analyses were conducted using the R software and are fully reproducible using the codes available on Github https://github.com/SarithaKodikara/LUPINE.

Author Contributions

Conceived and designed the study: SK KALC. Performed the analysis: SK. Wrote the paper: SK KALC.

Ethics declarations

Competing interests

The authors declare no competing interests.

References

E. Bogart, R. Creswell, and G. K. Gerber. Mitre: inferring features from microbiota time-series data linked to host status. *Genome biology*, 20:1–15, 2019.

M. X. Byndloss, E. E. Olsan, F. Rivera-Chávez, C. R. Tiffany, S. A. Cevallos, K. L. Lokken, T. P. Torres, A. J. Byndloss, F. Faber, Y. Gao, et al. Microbiota-activated ppar-γ signaling inhibits dysbiotic enterobacteriaceae expansion. *Science*, 357(6351):570–575, 2017.

B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7):581–583, 2016.

H. Daniel, A. M. Gholami, D. Berry, C. Desmarchelier, H. Hahne, G. Loh, S. Mondot, P. Lepage, M. Rothballer, A. Walker, et al. High-fat diet alters gut microbiota physiology in mice. *The ISME journal*, 8(2):295–308, 2014.

L. A. David, C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563, 2014.

A. Djukovic, M. J. Garzón, C. Canlet, V. Cabral, R. Lalaoui, M. García-Garcerá, J. Rechenberger, M. Tremblay-Franco, I. Peñaranda, L. Puchades-Carrasco, et al. Lactobacillus supports clostridiales to

restrict gut colonization by multidrug-resistant enterobacteriaceae. *Nature Communications*, 13(1): 5617, 2022.

I. Erb. Partial correlations in compositional data analysis. *Applied Computing and Geosciences*, 6: 100026, 2020.

J. Friedman and E. J. Alm. Inferring correlation networks from genomic survey data. *PLoS computational biology*, 8(9):e1002687, 2012.

G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8:2224, 2017.

X. Guo, J. Li, R. Tang, G. Zhang, H. Zeng, R. J. Wood, Z. Liu, et al. High fat diet alters gut microbiota and the expression of paneth cell-antimicrobial peptides preceding changes of circulating inflammatory cytokines. *Mediators of inflammation*, 2017, 2017.

D. K. Hammond, Y. Gur, and C. R. Johnson. Graph diffusion distance: A difference measure for weighted graphs based on the graph laplacian exponential kernel. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 419–422, 2013. doi: 10.1109/GlobalSIP.2013.6736904.

G. He, T. Chen, and L. Huang. Tremella fuciformis polysaccharide reduces obesity in high-fat diet-fed mice by modulation of gut microbiota. *Frontiers in Microbiology*, 13:1073350, 2022.

X. Jiang, B. Zhang, F. Lan, C. Zhong, J. Jin, X. Li, Q. Zhou, J. Li, N. Yang, C. Wen, et al. Host genetics and gut microbiota jointly regulate blood biochemical indicators in chickens. *Applied Microbiology and Biotechnology*, 107(24):7601–7620, 2023.

I. T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 31(3):300–303, 1982.

I. T. Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

Y.-G. Kim, K. Sakamoto, S.-U. Seo, J. M. Pickard, M. G. Gillilland III, N. A. Pudlo, M. Hoostal, X. Li, T. D. Wang, T. Feehley, et al. Neonatal acquisition of clostridia species protects against colonization by bacterial pathogens. *Science*, 356(6335):315–319, 2017.

S. Kodikara, S. Ellul, and K.-A. Lê Cao. Statistical challenges in longitudinal microbiome data analysis. *Briefings in Bioinformatics*, 23(4):bbac273, 2022.

Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology*, 11(5): e1004226, 2015.

J. M. Lankelma, C. Belzer, A. J. Hoogendijk, A. F. De Vos, W. M. De Vos, T. Van Der Poll, and W. J. Wiersinga. Antibiotic-induced gut microbiota disruption decreases tnf-α release by mononuclear cells in healthy adults. *Clinical and translational gastroenterology*, 7(8):e186, 2016.

K.-A. Lê Cao and Z. M. Welham. *Multivariate data integration using R: methods and applications with the mixOmics package*. CRC Press, 2021.

C. Lee, J. Lee, J. Y. Eor, M.-J. Kwak, C. S. Huh, and Y. Kim. Effect of consumption of animal products on the gut microbiome composition and gut health. *Food Science of Animal Resources*, 43(5):723, 2023.

K. Li, M. W. Epperly, G. A. Barreto, J. S. Greenberger, and B. A. Methé. Longitudinal fecal microbiome study of total body irradiated mice treated with radiation mitigators identifies bacterial associations with survival. *Frontiers in Cellular and Infection Microbiology*, 11:715396, 2021.

H. Lin and S. D. Peddada. Analysis of compositions of microbiomes with bias correction. *Nature communications*, 11(1):3514, 2020.

R. Lyu, Y. Qu, K. Divaris, and D. Wu. Methodological considerations in longitudinal analyses of microbiome data: A comprehensive review. *Genes*, 15(1):51, 2023.

M. S. Matchado, M. Lauber, S. Reitmeier, T. Kacprowski, J. Baumbach, D. Haller, and M. List. Network analysis methods for studying microbial communities: A mini review. *Computational and structural biotechnology journal*, 19:2687–2698, 2021.

A. Mu, G. P. Carter, L. Li, N. S. Isles, A. F. Vrbanac, J. T. Morton, A. K. Jarmusch, D. P. De Souza, V. K. Narayana, K. Kanojia, et al. Microbe-metabolite associations linked to the rebounding murine gut microbiome postcolonization with vancomycin-resistant enterococcus faecium. *Msystems*, 5(4): 10–1128, 2020.

J. Pessoa, G. D. Belew, C. Barroso, C. Egas, and J. G. Jones. The gut microbiome responds progressively to fat and/or sugar-rich diets and is differentially modified by dietary fat and sugar. *Nutrients*, 15(9):2097, 2023.

L. Röttjers and K. Faust. From hairballs to hypotheses–biological insights from microbial networks. *FEMS microbiology reviews*, 42(6):761–780, 2018.

D. Ruiz-Perez, J. Lugo-Martinez, N. Bourguignon, K. Mathee, B. Lerner, Z. Bar-Joseph, and G. Narasimhan. Dynamic bayesian networks for integrating multi-omics time series microbiome data. *Msystems*, 6(2):10–1128, 2021.

A. Salavaty, M. Ramialison, and P. D. Currie. Integrated value of influence: an integrative method for the identification of the most influential nodes within networks. *Patterns*, 1(5), 2020.

K. Sankaran and P. Jeganathan. Microbiome intervention analysis with transfer functions and mirror statistics. *arXiv preprint arXiv:2306.06364*, 2023.

A. Susin, Y. Wang, K.-A. Lê Cao, and M. L. Calle. Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, 2(2):lqaa029, 2020.

A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76:257–284, 2011.

W. S. Torgerson. Theory and methods of scaling. 1958.

L. K. Ursell, J. L. Metcalf, L. W. Parfrey, and R. Knight. Defining the human microbiome. *Nutrition reviews*, 70(suppl_1):S38–S44, 2012.

S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

J. Yin, Y. Li, H. Han, S. Chen, J. Gao, G. Liu, X. Wu, J. Deng, Q. Yu, X. Huang, et al. Melatonin reprogramming of gut microbiota improves lipid dysmetabolism in high-fat diet-fed mice. *Journal of Pineal Research*, 65(4):e12524, 2018.

X. Zhang, Y.-F. Pei, L. Zhang, B. Guo, A. H. Pendegraft, W. Zhuang, and N. Yi. Negative binomial mixed models for analyzing longitudinal microbiome data. *Frontiers in Microbiology*, 9:1683, 2018.

X. Zhang, B. Guo, and N. Yi. Zero-inflated gaussian mixed models for analyzing longitudinal microbiome data. *Plos one*, 15(11):e0242073, 2020.