Microbial network inference for longitudinal microbiome studies with LUPINE

Prof. Kim-Anh Lê Cao

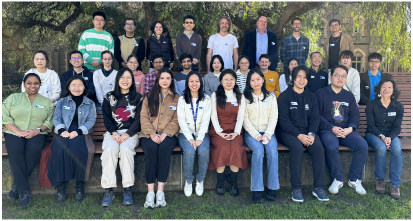
NHMRC investigator
Melbourne Integrative Genomics
School of Mathematics and Statistics

THE UNIVERSITY OF
MELBOURNE

@mixOmics_team | www.lecao-lab.science.unimelb.edu.au

Director of Melbourne Integrative Genomics

MIG is a cross-disciplinary initiative at the interface between biology, mathematics and statistics.



Schools of Maths & Stats and Biosciences; $\sim$ 45 members

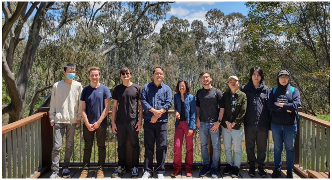
Two ARC Centres of Excellence

QUBIC The Australian Research Council Centre of Excellence in Quantum Biotechnology MACSYS

Computational workshops, monthly seminar series, reading groups

Lê Cao lab

- Expertise in statistics and computational biology
- Team of statisticians, bioinformaticians, data analysts and software developers
- Strengths: multi-disciplinary research, accessible software for the community, methods that are (often) technology agnostic
- mixOmics team finalist for the Australian Eureka prize 2023



We develop methods for microbiome and omics data integration.

Australian Government Australian Research Council

Australian Government National Health and Medical Research Council

QUBIC The Australian Research Council Centre of Excellence in Quantum Biotechnology

The microbiome is a complex organism

- Technological advances: culture-independent & NGS
- Characterize composition and interaction of microbes in an ecosystem
- Not only about cataloging organisms, but biological functions that affect host and participate in disease processes
- Require advanced bioinformatics and computational statistics

In the average adult are 100 trillion human cells and 1,500 trillion microbes.
At best you are little more than 10% you.
We're all just petri dishes with shoes.
-ANYLETTER
schuhlelewis.blogspot.com

Types of questions for data analysis

1. Identification of microbial features (taxa) whose relative abundance is associated to a phenotype of interest

2. Identification of microbial signatures as biomarkers of disease risk and prognostic.
   $\rightsquigarrow$ Variable selection

3. Understand the relationship between the host and their microbiome at different omics levels
   $\rightsquigarrow$ Data integration

Note: I haven't mentioned causal mechanisms here!

Challenging characteristics of microbiome data

1. Sparse (large number of zeroes)
2. Compositional
3. Multivariate

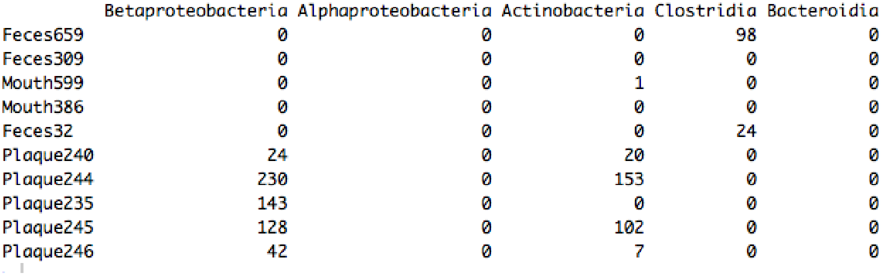
In addition:

- High variability, non Gaussian distribution
- Different types of sequencing technology
- Different levels of bacterial taxonomy for analysis
- Prone to batch effects

• Lê Cao et al (2016) mixMC: a multivariate statistical framework to gain insight into Microbial Communities. PLoS ONE.
• Wang, Lê Cao (2019) Managing Batch Effects in Microbiome Data. Briefings in Bioinformatics.

Example of microbial count data

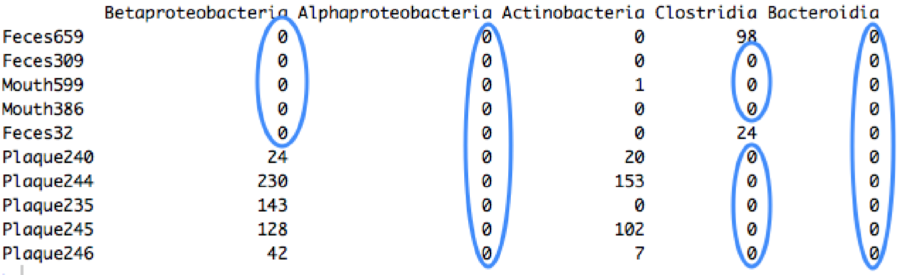
16S rRNA sequencing data after taxonomic classification:



```
          Betaproteobacteria Alphaproteobacteria Actinobacteria Clostridia Bacteroidia
Feces659                   0                   0              0         98           0
Feces309                   0                   0              0          0           0
Mouth599                   0                   0              1          0           0
Mouth386                   0                   0              0          0           0
Feces32                    0                   0              0         24           0
Plaque240                 24                   0             20          0           0
Plaque244                230                   0            153          0           0
Plaque235                143                   0              0          0           0
Plaque245                128                   0            102          0           0
Plaque246                 42                   0              7          0           0
```

Note: we often consider the Operational Taxonomy Unit level (OTU), but may report our microbial OTU signatures at higher taxonomic ranks

Problem 1: Sparse

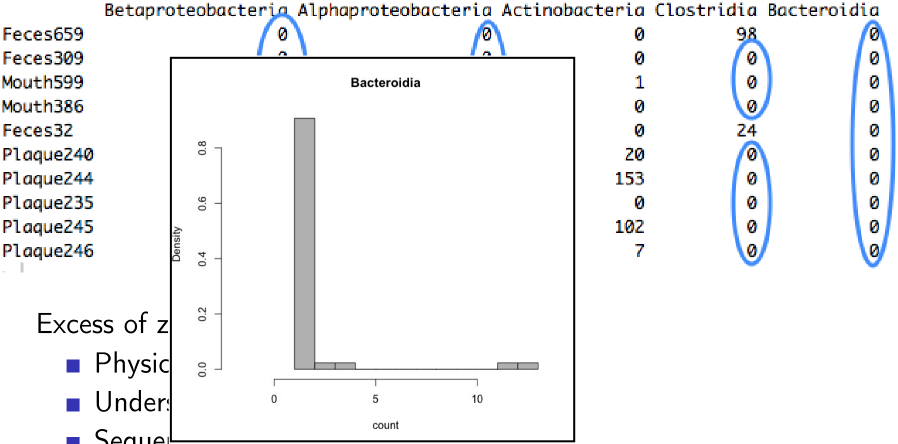


```
           Betaproteobacteria Alphaproteobacteria Actinobacteria Clostridia Bacteroidia
Feces659                    0                   0              0         98           0
Feces309                    0                   0              0          0           0
Mouth599                    0                   0              1          0           0
Mouth386                    0                   0              0          0           0
Feces32                     0                   0              0         24           0
Plaque240                  24                   0             20          0           0
Plaque244                 230                   0            153          0           0
Plaque235                 143                   0              0          0           0
Plaque245                 128                   0            102          0           0
Plaque246                  42                   0              7          0           0
```

Excess of zeroes:

- Physical absence?
- Undersampling?
- Sequencing error?

Problem 1: Sparse



```
            Betaproteobacteria Alphaproteobacteria Actinobacteria Clostridia Bacteroidia
Feces659                     0                   0              0         98           0
Feces309                                                        0          0           0
Mouth599                                                        1          0           0
Mouth386                                                        0          0           0
Feces32                                                         0         24           0
Plaque240                                                      20          0           0
Plaque244                                                     153          0           0
Plaque235                                                       0          0           0
Plaque245                                                     102          0           0
Plaque246                                                       7          0           0
```

Excess of z

- Physic
- Unders
- Sequencing error?

⇝ 'zero-inflated' distribution

Problem 2: co-dependency

from https://www.nps.gov

from http://english.samajalive.in

Ecology: abundance of ladybugs does not affect number of tigers.

Microbiome: most microorganisms are co-dependent + data are not in absolute but relative abundance!

Problem 3: compositional

Definition

Compositional data are naturally described as proportions: they contain information about the relationships between the proportions.

Origins of compositional data:

- 'Naturally' (e.g. ladybugs and tigers)
- Technical artefacts (sequencing a finite amount of reads)
- Data transformations (rarefaction, proportions)

$\rightsquigarrow$ most statistical methods assume unbounded data whereas proportions are bounded

$\rightsquigarrow$ spurious correlations (as noted by Pearson in 1897!)

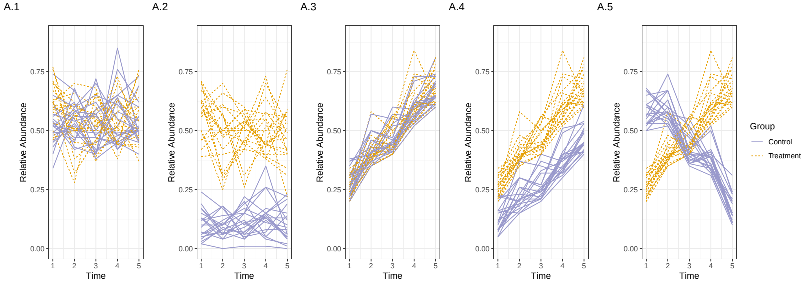
• Gloor GB, et al (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. Front. Microbiol. 8:2224

Some solutions

1. Ratio transformation of the data (e.g. log-ratios):
   - Data become non bounded so that classical statistical methods can now be used
     → Centered log-ratio (CLR), Additive log-ratio (ALR)
     Isometric log-ratio (ILR)
     → Special care should be given to interpretation!
2. Other approaches that use ratios:
   - Proportionality distance between pairs of variables (Lovell et al., 2015)
   - Compositional balances (log-contrasts) of taxa (Rivera-Pinto 2018)

• Susin, Wang, Lê Cao, Calle L. Variable selection in microbiome compositional data analysis, NAR Genomics and Bioinformatics

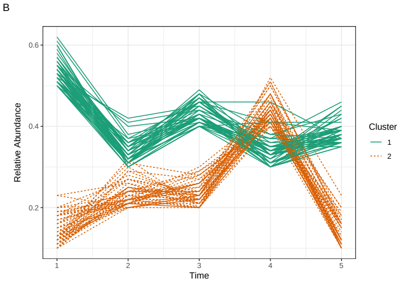
Obj 1: Differential abundance over time and between sample groups



Each line = a given taxon for each individual

Models: Zero-inflated beta regression, Negative binomial mixed model, SplinectomeR, Zero-inflated Gaussian mixed models, Linear mixed model splines

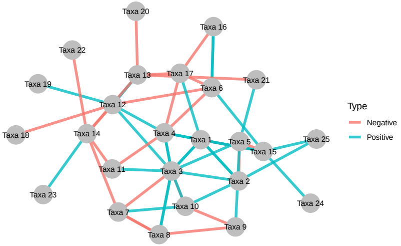
Obj 2: Clustering profiles of microorganisms and omics measurements correlated across time



Several taxa are shown, each averaged across several individuals

Methods: Dynamic time warping, Partitioning around medoids and agglomerative clustering, PCA / PLS on linear mixed model splines

Obj 3: Network modelling to identify temporal relationships between microorganisms



One edge = two taxa nodes are 'associated' (correlated)

Methods to infer networks: Dynamic Bayesian Network, Granger causality based interaction networks, LUPINE

State-of-the-art

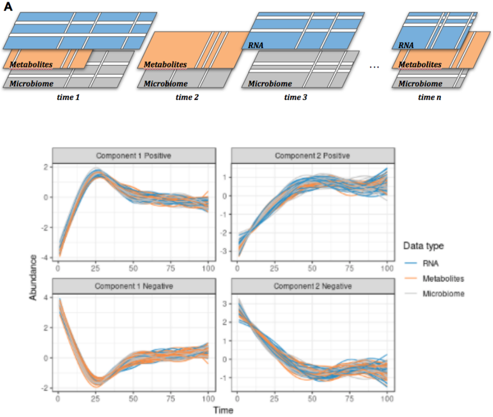
- Differential analysis (univariate):
  several methods, either based on counts or continuous (normalised) data

- Clustering analysis (multivariate):
  require small number of time points (5 - 10), expect regular or similar time trends

- Network modelling to identify temporal associations between variables (multivariate):
  not many approaches, either across all individuals, or per individual

Simulations and real data analysis:
• Kodikara S, Ellul S, Lê Cao K-A (2022), Statistical challenges in longitudinal microbiome data analysis, *Briefings in Bioinformatics*

Identify correlated omics profiles across time



- Data are in 3D
  → Dimension reduction to 2D with smoothing splines
  → Estimate missing data with splines

- PLS or PCA to identify correlated spline profiles (genes, taxa, metabolites) across time and omics

- Bodein A, Chapleur O, ...and Lê Cao K-A (2019). A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. *Frontiers in Genetics*
- Bodein A, ..., Lê Cao K-A and Droit A (2021). `timeOmics`: an R package for longitudinal multi-omics data integration. *Bioinformatics*

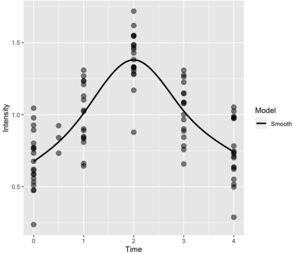
Smoothing spline modelling

Advantages

- Reduce noise
- Interpolates time points when irregular sampling times

Disadvantages

- Requires large number of time points
- Smoothing parameter to choose

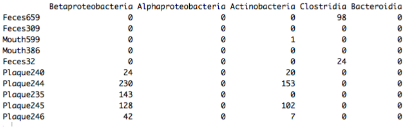


Linear mixed model splines

No smoothing parameter; Flexible models; Interpolation; Hypothesis testing; Downstream clustering analysis

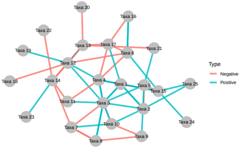
- Straube, Gorse, Huang and Lê Cao (2015). A linear mixed model spline framework for analyzing time course 'omics' data. PLOS ONE, 10(8).

Modelling microbial networks



```
          Betaproteobacteria Alphaproteobacteria Actinobacteria Clostridia Bacteroidia
Feces659                   0                   0              0         98           0
Feces309                   0                   0              0          0           0
Mouth599                   0                   0              1          0           0
Mouth386                   0                   0              0          0           0
Feces32                    0                   0              0         24           0
Plaque240                 24                   0             20          0           0
Plaque244                230                   0            153          0           0
Plaque235                143                   0              0          0           0
Plaque245                128                   0            102          0           0
Plaque246                 42                   0              7          0           0
```

⟹

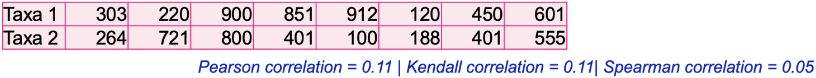


Extracting co-occurrence of taxa to:

- Study changes in associations between microorganisms
- Identify important members of a microbial community
- Characterise change in associations between microorganisms resulting from intervention

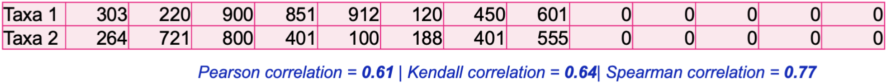
Common metrics used for association inference are based on correlation (Pearson, Spearman, Kendall)

Correlation may not detect true association but partial correlation can



Taxa 1	303	220	900	851	912	120	450	601
Taxa 2	264	721	800	401	100	188	401	555

Pearson correlation = 0.11 | Kendall correlation = 0.11| Spearman correlation = 0.05



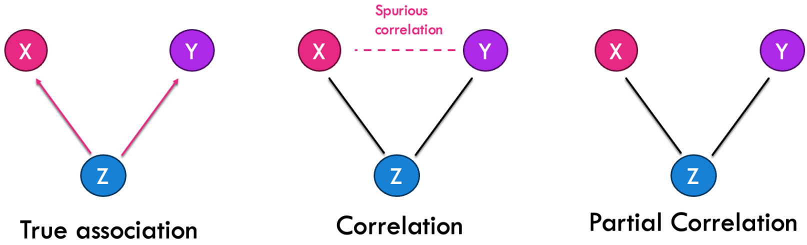
Taxa 1	303	220	900	851	912	120	450	601	0	0	0	0	0
Taxa 2	264	721	800	401	100	188	401	555	0	0	0	0	0

*Pearson correlation = **0.61** | Kendall correlation = **0.64**| Spearman correlation = **0.77***

Matching zeros inflate correlation coefficients

⇝ Pearson and Spearman correlation coefficients can be severely distorted

Correlation may not detect true association but partial correlation can



LUPINE (LongitUdinal modelling with PLS regression for NEtwork inference)

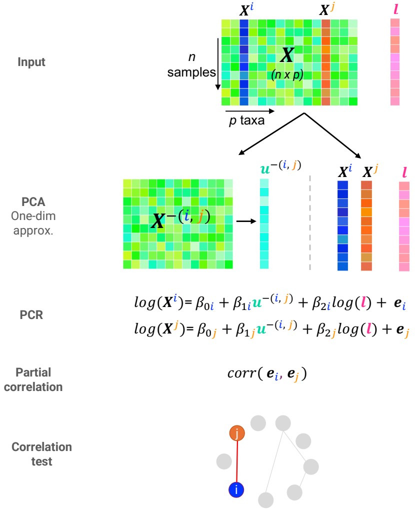
LUPINE learns from past information for network inference of microbiome longitudinal data.

We estimate partial correlation between taxa based on low-dimensional data representation.

Dr Saritha Kodikara, MIG

- Kodikara S, Lê Cao K-A. Microbial network inference for longitudinal microbiome studies with LUPINE. *bioRxiv* 2024.05.08.593086

Based on PCA for a single time point



Input

PCA
One-dim
approx.

PCR

$$log(X^i) = \beta_{0i} + \beta_{1i} u^{-(i,j)} + \beta_{2i} log(l) + e_i$$

$$log(X^j) = \beta_{0j} + \beta_{1j} u^{-(i,j)} + \beta_{2j} log(l) + e_j$$

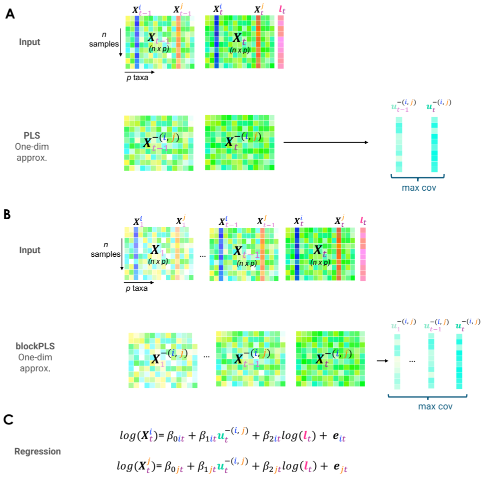
Partial correlation

$$corr(e_i, e_j)$$

Correlation test

- Aim: estimate partial correlation between taxa at a single time point
- We use low-dimensional data representation (PC) for conditional independence
- Component $u^{-(i,j)}$ controls for all taxa except taxa i and j and is fitted into a regression model to extract residuals (we also add library size)
- Estimate partial correlation between taxa i and j based on residuals

Based on PLS for multiple time points



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

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

- Aim: estimate partial correlation between taxa based on previous time points
- We use **A:** PLS for two time points, and **B:** block PLS for multiple time points
- Estimate component $u_t^{-(i,j)}$ to control for all taxa except taxa i and j that is maximally correlated with **A:** previous time point $u_1^{-(i,j)}$ or **B:** previous time points $u_{t-1}^{-(i,j)}, \ldots, u_1^{-(i,j)}$

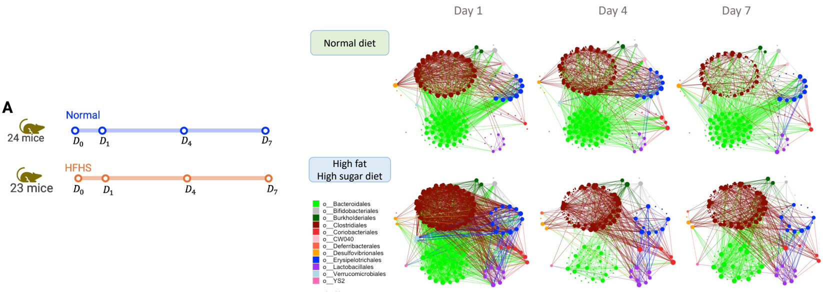
Measures to compare networks

- Pairwise distance between network topologies: graph diffusion distance (**GDD**, Hammond et al, 2013) visualised with MDS
- Node influence: Integrated Value of Influence (**IVI**, Salavaty et al, 2020) combines local prominence and broader impact of the nodes visualised with PCA
- Correlation test between two networks: Hamming distance and Mantel test to assess similarity / differences

In simulation studies we show that:

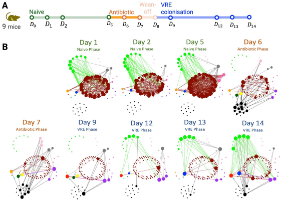
- LUPINE has better performance than sparCC and spiecEasi (incl computational)
- LUPINE highlighted more robust longitudinal network patterns than LUPINE_single

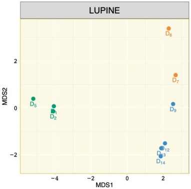
1. HFHS diet study in mice

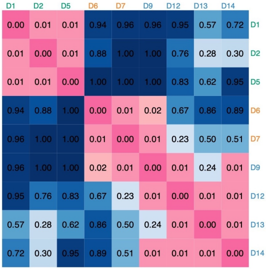


LUPINE highlighted two different microbial community networks between between diets, larger connections in *Lactobacillales* in HFHS diet mice

2. Vancomycin-resistant Enterococcus faecium in mice







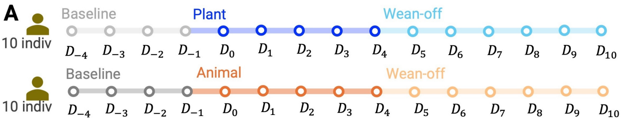
	D1	D2	D5	D6	D7	D9	D12	D13	D14
D1	0.00	0.01	0.01	0.94	0.96	0.96	0.95	0.57	0.72
D2	0.01	0.00	0.01	0.88	1.00	1.00	0.76	0.28	0.30
D5	0.01	0.01	0.00	1.00	1.00	1.00	0.83	0.62	0.95
D6	0.94	0.88	1.00	0.00	0.01	0.02	0.67	0.86	0.89
D7	0.96	1.00	1.00	0.01	0.00	0.01	0.23	0.50	0.51
D9	0.96	1.00	1.00	0.02	0.01	0.00	0.01	0.24	0.01
D12	0.95	0.76	0.83	0.67	0.23	0.01	0.00	0.01	0.01
D13	0.57	0.28	0.62	0.86	0.50	0.24	0.01	0.00	0.01
D14	0.72	0.30	0.95	0.89	0.51	0.01	0.01	0.01	0.00

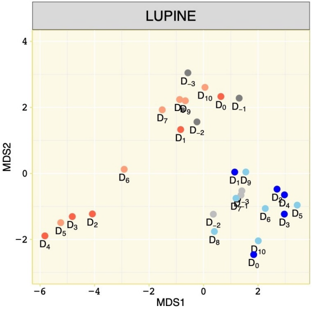
After antibiotic treatment: reduction of *Clostridales* order until day 14 while *Bacteroidales* order appears to recover after day 12.

Pairwise GDD/MDS shows different network structures for each phase of the experiment.

Mantel p-values indicates strong correlations within each phase.

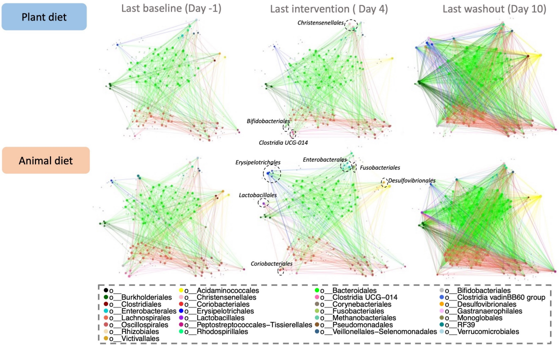
3. Case control diet study in humans


A
10 indiv
Baseline
Plant
Wean-off
D_{-4} D_{-3} D_{-2} D_{-1} D_0 D_1 D_2 D_3 D_4 D_5 D_6 D_7 D_8 D_9 D_{10}
10 indiv
Baseline
Animal
Wean-off
D_{-4} D_{-3} D_{-2} D_{-1} D_0 D_1 D_2 D_3 D_4 D_5 D_6 D_7 D_8 D_9 D_{10}



Pairwise GDD / MDS shows distinct patterns according to diet groups

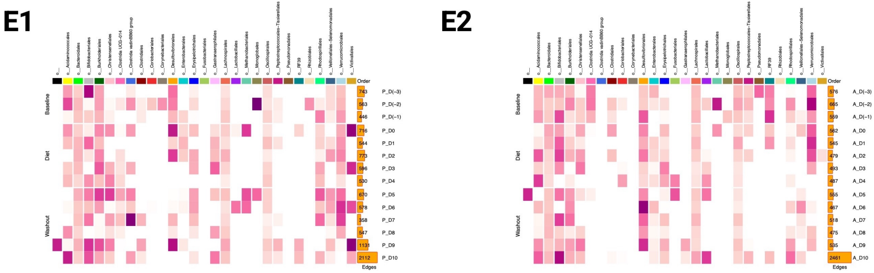
3. Case control diet study in humans



Inferred networks.

Day 4: plant-based network shows increased connections in *Christensenellales, Clostridia UCG 014* whereas animal-based network shows increased connections in *Erysipelotrichales, Lactobacillales, Coriobacteriales, Enterobacterales, Fusobacteriales, Desulfovibrionales*.

3. Case control diet study in humans



Average IVI score for each taxonomic order. **E1:** plant based and **E2:** animal based diet groups (pink = high value). *Bacteroidetes, Lachnospirales, Oscilospirales* consistently exhibit a non-zero IVI score, indicating their stable influence, unaffected by diet or daily variations.

LUPINE

- First sequential microbial network inference approach for longitudinal experiments
- Detects stability of taxa associations over time and as response of external disturbances
- Low-dimensional PLS approximation to calculate partial correlation and infer networks across time points
- Use of metrics to identify any abrupt network changes across time, groups, and key taxa nodes, and correlation tests to measure differences between networks
- Suitable for small sample sizes ($>$ 10 per group)

R package: https://github.com/SarithaKodikara/LUPINEmanuscript (soon in mixOmics)

Next online mixOmics workshop starts Oct 21 for 6 weeks, see www.mixOmics.org