Integrated Metagenomic Analyses of the Rumen Microbiome of Cattles Reveals Key Biological Mechanisms Associated with Methane Traits

Haiying Wang[1], Huiru Zheng[1*], Fiona Browne[1], Rainer Roehe[2], Richard J. Dewhurst[2], Felix Engel[3], Matthias Hemmje[3], Xiangwu Lu[4], Paul Walsh[4]

[1]School of Computing and Mathematics, Computer Science Research Institute
Ulster University, United Kingdom
[2]Future Farming Systems, Scotland's Rural College, Edinburgh, United Kingdom
[3]Research Institute for Telecommunication and Cooperation, Germany
[4]NSilico Life Science Ltd., Cork, Ireland
*Corresponding author: Huiru Zheng, h.zheng@ulster.ac.uk

Abstract

Methane is one of major contributors to global warming. The rumen microbiota is directly involved in methane production in cattle. The link between variations in rumen microbial communities and host genetics has important applications and implications in bioscience. Having the potential to reveal the full extent of microbial gene diversity and complex microbial interactions, integrated metagenomics and network analysis holds great promises in this endeavour. This study investigates the rumen microbial community in cattle through the integration of metagenomic and network-based approaches. Based on the relative abundance of 1570 microbial genes identified in a metagenomics analysis, the co-abundance network was constructed and functional modules of microbial genes were identified. One of the main contributions is to develop a random matrix theory-based approach to automatically determining the correlation threshold used to construct the co-abundance network. The resulting network, consisting of 549 microbial genes and 3349 connections, exhibits a clear modular structure with certain trait-specific genes highly over-represented in modules. More specifically, all the 20 genes previously identified to be associated with methane emissions are found in a module (hypergeometric test, $p < 10^{-11}$). One third of genes are involved in methane metabolism pathways. The further examination of abundance profiles across 8 samples of genes highlights that the revealed pattern of metagenomics abundance has a

strong association with methane emissions. Furthermore, the module is significantly enriched with microbial genes encoding enzymes that are directly involved in methanogenesis (hypergeometric test, $p < 10^{-9}$).

Keywords—rumen microbial community, metagenomics, network-based approaches, random matrix theory

Highlights:

- RMT-based approach can be used to automatically determine the correlation threshold used to construct the co-abundance network
- The constructed co-abundance network exhibits a clear modular structure
- Certain trait-specific genes including those associated with methane emissions are highly over-represented in modules
- Key biological mechanisms associated with methane emission including those directly involved in methanogenesis were revealed.

1 Introduction

As one of the most complicated anaerobic microbial ecosystems in nature [1], the rumen provides an environment with stable and favorable physiological conditions for microbial growth and fermentation. Microbes in the rumen are a complex ecosystem predominantly consisting of bacteria, archaea, protozoa and fungi. These microorganisms confer the ability to break down complex polysaccharides and harvest energy from otherwise indigestible food components [2], [3]. It has been shown that *Bos taurus* gut microbiota has a paramount role in cattle performance, productivity, health and immunity [4].

However, the rumen microbes are also responsible for the production of the highly potent greenhouse gas methane and nitrogen-rich wastes causing not only the loss of feed gross energy but also contributing to the greenhouse gas emissions and global warming [1], [5], [6]. Understanding the topological difference in gut microbial community composition is crucial to provide knowledge on the functions of each member of the microbiota to the physiological maintenance of the host. Thus a better understanding of the composition of rumen microbial communities and the association between host genetic and microbial activities has significant applications and implication in bioscience [6], [7].

Early exploration of rumen microbiology was mainly dominated by culture-based approaches. Examples include the description of well characterized rumen bacteria based on the isolation of the functionally significant bacterial groups [9], [10]. While successfully identifying more than 200 microbial species including bacteria and protozoa from the rumen [1], [8], culture-dependent techniques requiring a careful design of protocol for growth of organism exhibit several significant limitations [11]. They are not only time consuming and cumbersome [8] but more importantly, culture-based studies are usually unable to reveal the

full extent of microbial diversity due to the nature of protocol design and constraints due to culture conditions [11], [12].

Advances in next-generation sequencing (NGS) have opened up new avenues in microbial ecology studies. Metagenomics, defined as the direct genetic analysis of DNA from microbial communities sampled in their specific environment without prior need for culturing, is offering unparalleled coverage and depth in determining microbial gut dynamics as long as the analytic resources are available [13], [14].

A number of metagenomics studies have investigated rumen microbial populations. These include research by Henderson et al. [7], which investigated whether the microbial community composition was influenced by diet, host species, or geography. It has been found that the composition of rumen microbial community varies with diet and host, but similar bacteria and archaea dominated in nearly all samples. Based on the simultaneous exploration of rumen microbiota and the metabolic phenotype, the study carried out by Morgavi et al. [5] brought new insights on the interactions between microbial populations and the association with the host. By varying a host's diet, Faith et al. [15] predicted a human gut microbiota's response to diet in gnotobiotic mice, in which 60% of the variation in species abundance was predicted due to the differences in diet.

More recently, based on the relative abundance of 1570 microbial genes identified in a metagenomics analysis, Roehe and his colleagues [6] developed new selection criteria to be used for predicting methane emissions and other traits such as feed conversion efficiency. Using the partial least squares analysis, 20 and 49 microbial genes were found to be associated with methane emissions and feed conversion efficiency in cattle respectively. Furthermore, functional clusters of microbial genes were identified based on the analysis of the co-abundance network in which the correlation threshold was manually set to 0.9.

By extending our preliminary analysis [16], this study aims to further examine the rumen microbial community in cattle through the integration of metagenomic and network-based approaches. The main objectives include

- to develop an automatic computational technique to objectively determine the correlation threshold used to construct a condition specific co-abundance network.
- to adopt network systems biology approaches for the identification of key biological mechanisms associated the methane traits

The rest of the paper is organized as follows. Section II briefly describes the methodology and datasets under study. The detailed description of automatic determination and its implementation is provided. The results and discussion are presented in Section III. The conclusions, together with future research, are given in Section IV.

2 Methodologies

The framework for integrated metagenomic analyses adopted in this study is illustrated in Fig. 1. Based on the relative abundance of 1570 microbial genes identified in a metagenomics analysis, a random matrix theory (RMT)-based approach used to automatically determine the correlation threshold for the construction of the co-abundance network has been developed. By incorporating domain knowledge including KEGG pathways, trait specific genes and genes encoding enzymes involved in methanogensis, the co-abundance network was further analysed in terms of topological structure, functional enrichment and biological relevance.

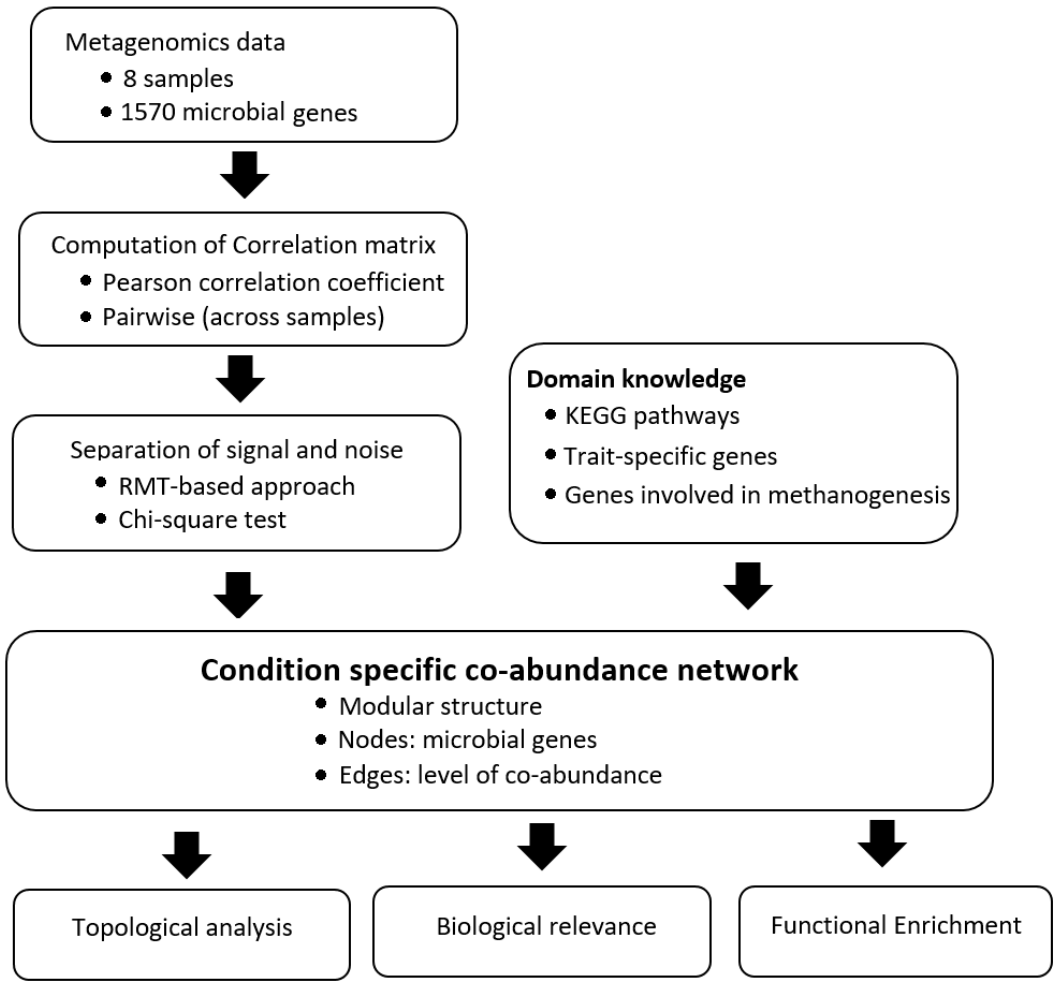


Fig. 1 The framework for integrated metagenomic analyses of the rumen micobiome

2.1 Metagenomic data

The metagenomics data applied in this research were released by the recent studies conducted at the Beef and Sheep Research Centre of Scotland's Rural College [3]. A brief overview of experiment design and DNA sequencing is given below. The reader is referred to [3] and [6] for a detailed description of data generation.

2.1.1 Experiment design and methane measurements

A 2 × 2 factorial design experiment was performed using two breed types (Aberdeen Angus (AA) and Limousin (LIM) rotational crosses) and two diets (defined as concentrate (CON) and forage (FOR)) using 72 steers from a two-breed rotational cross between AA and LIM. All

animals were raised on the Research Farm. Methane emissions of individual animals were measured in respiration chambers [15].

2.1.2 DNA sequencing and KEGG analysis

A total of 8 extreme animals were identified for deep sequencing analysis (4 high and 4 low) based on methane emissions balanced for breed type (Aberdeen-Angus or Limousin cross) and diet (CON or FOR) as depicted in Table I. DNA was extracted from rumen samples and subject to qPCR for the 16S rRNA genes to determine the abundance [6]. Sequence data between 8.6 and 14.6 GB per sample (between 43.4 and 72.7 million paired reads) were assembled de novo. To identify the microbial genes, the genomic reads were aligned to the KEGG genes database allowing for up to a 10% mismatch. The read and best hits belonging to a single KEGG orthologue group (KO) were retained. In total 3970 KEGG gene orthologues were identified in rumen contents samples, of which 1570 genes showed a relative abundance of more than 0.001%.

TABLE I THE CHARACTERISTICS OF 8 SAMPLES USED IN THE SRUC STUDIES. AA: ABERDEEN ANGUS; LIM: LIMOUSIN CROSS; CON: CONCENTRATE BASED DIET; FOR: FORAGE BASED DIET; DMI: DRY MATTER INTAKE; AND FCR: FEED CONVERSION RATIO

Animal code	Breed	Diet	Methane emission group	Methane (kg/DMI)	FCR (kg intake/kg gain)
2019N0001	AA	CON	LOW	7.635	6.102
2019N0002	AA	CON	HIGH	18.137	6.096
2019N0003	LIM	CON	LOW	9.290	9.327
2019N0004	LIM	CON	HIGH	20.130	8.039
2019N0005	AA	FOR	LOW	17.412	10.381
2019N0006	AA	FOR	HIGH	32.415	6.719
2019N0007	LIM	FOR	LOW	19.373	8.065
2019N0008	LIM	FOR	HIGH	30.372	8.118

2.2 RMT-based approaches

Since the influential work carried out by Wigner in 1950's [18], RMT has found a wide range of applications across a number of areas including physics, finance and bioinformatics [17] - [20]. One of applications areas is to determine a threshold to objectively separate signal from noise which is based on the following two universal predictions associated with statistical properties of the nearest neighbor spacing distribution (NNSD) of unfolded eigenvalues, i.e. $P(s)$.

- The NNSD of any random matrix representing systems largely composed of noise closely follows Gaussian orthogonal ensemble (GOE) statistics [17], [20]. Let *N* represent the order of the matrix, e_i be the unfolded eigenvalue and $s_i = e_{i+1} - e_i$ $(i = 1, 2,3, \cdots, N-1)$ denote the spacing between consecutive eigenvalues after unfolding. It has been shown that the distribution can be well represented by the Wigner surmise [18] as described by Eq. (1).

$$P(s) = \frac{\pi}{2} \times s \times e^{(-\pi s^2/4)} \quad (1)$$

- For a non-random matrix in which no correlation between nearest-neighbor eigenvalues is observed, the NNSD tends to follow the Poisson distribution as shown in Eq. (2), indicating the system represented by the matrix can be separated into several relatively independent clusters in which members exhibit similar behaviours and properties [20], [21].

$$P(s) = e^{-s} \quad (2)$$

It has been highlighted that the transition of NNSD between GOE and Poisson statistics as illustrated in Fig. 2 can potentially serve as a reference point to automatically construct a condition-specific correlation network by removing random noise in an objective manner [17].

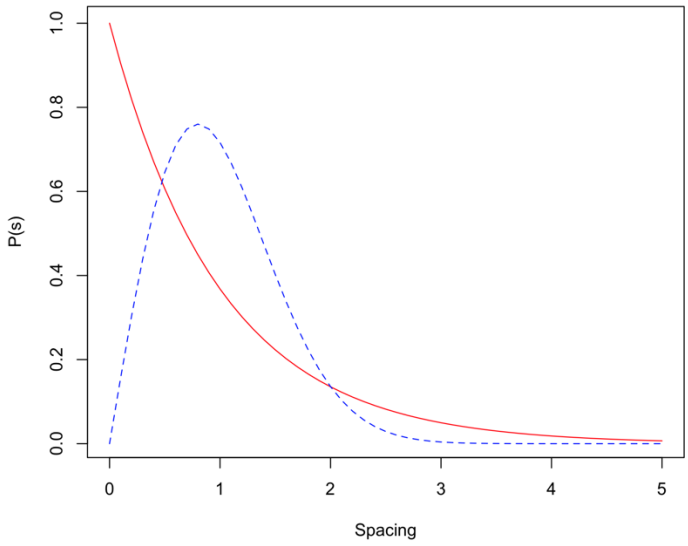


Fig. 2 Transition between GOE and Poisson distributions in RMT. The blue dotted line depicts the GOE distribution while the solid red curve represents a Poisson distribution. The transition of NNSD between GOE and Poisson statistics can potentially serve as a reference point to construct a condition-specific correlation network by separating high and weak correlation.

2.3 Construction of co-abundance networks

Based on the recent study [6] which demonstrates that the abundance of a suite of microbial genes was highly informative for predicting certain traits and the co-abundance network exhibits a modular structure, we hypothesized that the correlation matrix derived from the abundance of microbial genes under different conditions can be broken into two parts: the high correlation part encoding the correlation of microbial genes specified to the changes in conditions and the weak correlation part associated with non condition specific correlation between gene abundances. In order to construct a network specified to the conditions under

study, we gradually remove pairs with absolute correlation values below the selected cutoff values as illustrated in Fig. 3.

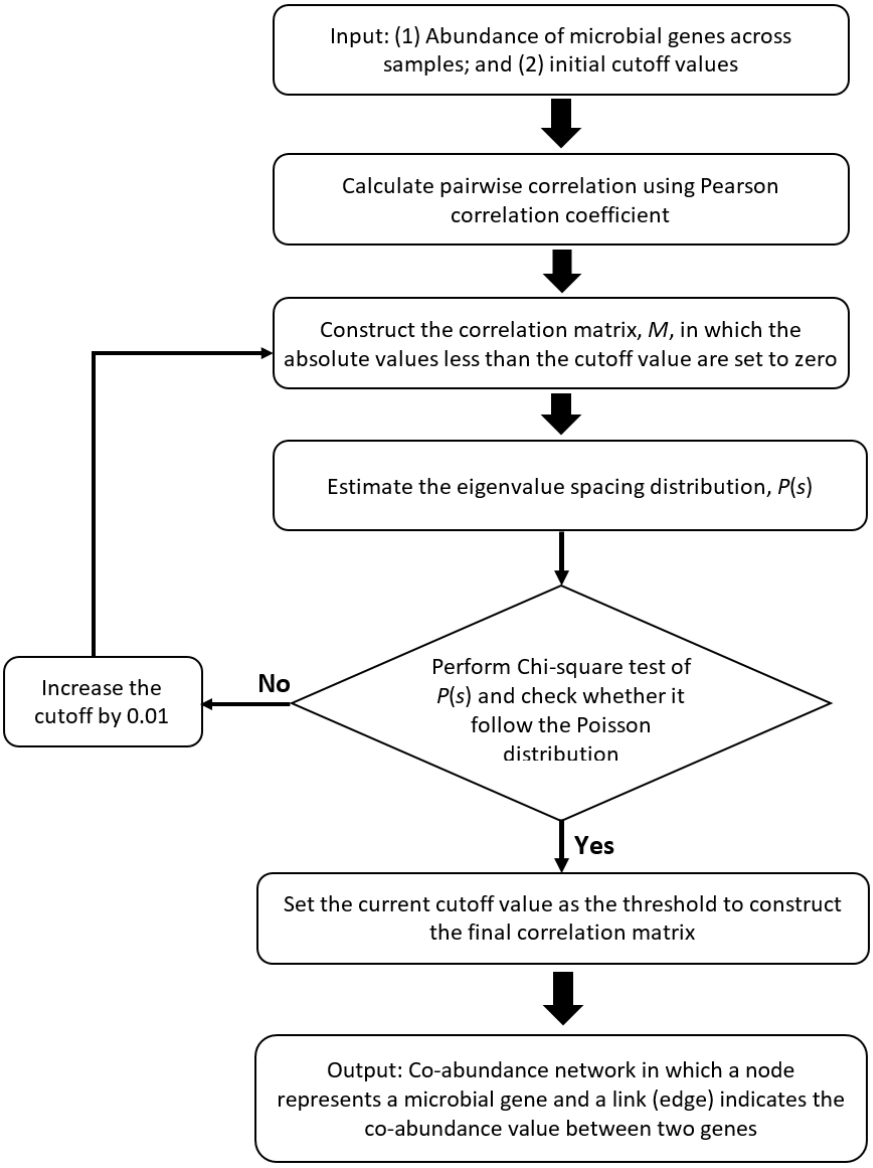


Fig. 3 A diagram to illustrate the key steps to construct the co-abundance network.

Let g_{ik} denote the abundance of microbial gene *i* in sample *k*. The pair-wise similarity between two microbial genes was estimated using Pearson correlation coefficient, $c(g_i, g_j)$ as defined below where $\overline{g_i}$ is the average abundance of gene *i* over the samples.

$$c(g_i, g_j) = \frac{\sum_{k=1}^{n}(g_{ik} - \overline{g_i})(g_{jk} - \overline{g_j})}{\sqrt{\sum_{k=1}^{n}(g_{ik} - \overline{g_i})^2}\sqrt{\sum_{k=1}^{n}(g_{jk} - \overline{g_j})^2}} \quad (3)$$

The eigenvalues were calculated based on the Eq. (4) where *M* is an *n* by *n* correlation matrix, λ is an eigenvalue, *v* is the corresponding eigenvector and *I* is the *n* by *n* identity matrix.

$$(M - \lambda I)v = 0 \quad (4)$$

2.4 Centrality metrics

The constructed network was further analysed using a number of topological metrics including degree, betweenness, eigenvector, bridging, closeness, PageRank, and power centralities, which have been previously applied to identified key players in biological processes. A brief definition of eigenvector, pagerank and power centralities is presented below. A detailed description of the rest of metrics can be found in [23] and [24].

Let $A \in \mathbb{R}^{N\times N}$ be an adjacency matrix associated with a graph $G = (V, E)$ representing a network where $N = |V|$ representing the number of nodes in the network and E is the set of links between nodes. Each entry A_{ij} indicates the strength of association between nodes v_i and v_j. Eigenvector, pagerank and power centralities of nodes v_i denoted by EC_i, PR_i, and PC_i respectively can be calculated using Eqs. (5), (6) and (7) respectively as defined below:

$$EC_i = \frac{1}{\lambda}\sum\nolimits_{j\in N(v_i)} A_{ij}\times EC_j \quad (5)$$

$$PR_i = \frac{1-d}{N} + d\times\sum\nolimits_{j\in N(v_i)} A_{ij}\times\frac{PR_j}{N(j)} \quad (6)$$

$$PC_i = \sum\nolimits_{j\in N(v_i)} (\alpha - \beta\times PC_j)A_{ij} \quad (7)$$

where λ is the eigenvalue, α is used to normalize the measure, *β* sets the dependence of each nodes centrality to the adjacent nodes and $N(v_i)$ is the set of neighbours of nodes v_i.

2.5 Evaluation metrics and software packages used

2.5.1 Chi-square goodness-of-fit test

To check whether the distribution of nearest neighbor eigenvalues spacing follows the Poisson statistic as defined by Eq. (2), the Chi-square (χ^2) goodness-of-fit test was applied with the null and alternative hypotheses being as follows:

H_0: $P(s)$ follows the Poisson distribution.

H_1: $P(s)$ does not follow the Poisson distribution.

Let $\chi^2(df, \alpha)$ be the critical value of Chi-square with *df* degrees of freedom at a significant level of α (α is set to 0.01 in this study). The H_0 will be rejected if the calculated χ^2 is greater than $\chi^2(df, \alpha)$.

2.5.2 Enrichment analysis

The level of the enrichment of certain trait specific genes was quantitatively expressed by the *hypergeometric distribution* probability calculated as follows.

$$p = 1 - \sum_{i=0}^{K-1} \binom{K}{i} \binom{N-K}{n-i} \Big/ \binom{N}{n} \quad (8)$$

where *K* is the number of genes that fall into a module, *k* is the number of trait-specific genes in the module, *N* is the total number of genes included in the network and *n* is the number of genes associated with a trait found in the network.

2.5.3 Software packages used

The estimation of the distribution of unfolded eigenvalue spacing was implemented using the pipeline of Molecular Ecological Network Analysis [22]. The NNSD was plotted using the R package RMThreshold (https://cran.rproject.org/web/packages/RMThreshold/index.html). The computation of topological parameters was with the NetworkAnalyzer [23] and CentiScaPe [24] plugins and the iGraph library in R available in: http://igraph.org/r/doc/. The construction of co-abundance network and interaction visualization of networks were

achieved using ExpressionCorrelation plugin available at http://www.baderlab.org/Software/ExpressionCorrelation and Cytoscape 3.3 [25] respectively.

3 Results and discussion

3.1 The conformity of the co-abundance network

To apply the RMT-based algorithm to determine a signal-noise threshold for a matrix, the matrix must meet certain criteria. For example, the matrix must be large, real-valued, symmetric and should not have a rank much smaller than its dimension. The eigenvalues must be unfolded.

To assert if the co-abundance network is well-conditioned for the proposed algorithm, we applied the validation function provided by the RMThreshold package. The matrix associated with the co-abundance network is not sparse and has a rank of 7. Two unfolding methods have been tested. One is based on the estimation of the Gaussian kernel density of the eigenvalue spectrum; another is based on fitting the cumulative eigenvalue distribution function to a cubic spline. As depicted in Fig. 4, the scatter plot of the derived eigenvalue spacing has a linear trend line with a slope of zero and an intersect of one (dotted line in Fig.4(a)), suggesting the average eigenvalue spacing is kept to one over the whole spectrum and thus confirming the eigenvalues have been correctly unfolded. As expected, when no threshold is applied, the NNSD is close to the GOE distribution with small eigenvalue spacings approaching zero (Fig.4 (b)) highlighting the co-abundance network is dominated by noise.

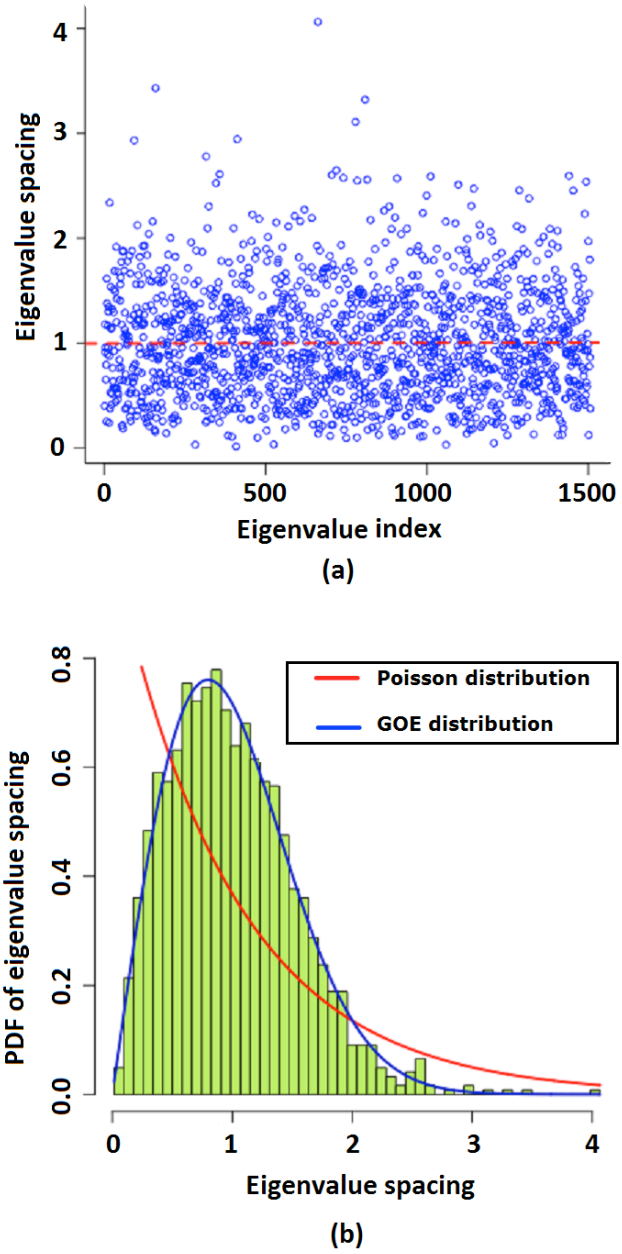


Fig. 4 Diagnostic results after the validation: (a) a scatterplot of the eigenvalue spacing with linear fit (red dotted line); and (b) the NNSD distribution.

3.2 The impact of the threshold

As shown in Fig. 5, the selection of the cutoff value has significant impact on the NNSD derived from the co-abundance matrix. As expected, the NNSD clearly follows the GOE distribution when no threshold was applied (Fig. 5(a)), suggesting that the correlation matrix directly derived from the abundance data failed to distinguish condition specific relationship

embedded in the correlation matrix from random noise. As the threshold is increased gradually, the clear transition of the NNSD from GOE to Poisson was observed (Fig. 5 (b) to Fig. 5(d)). This was further confirmed when we examined small eigenvalue spacings (<0.003) and the log likelihood of the empirical NNSD as depicted in Figs 6 and 7 respectively. For example, as shown in Fig. 6, the percentage of small spacings approaches zero for threshold less than 0.9 which suggests that eigenvalues somehow repel each other. This implies the data are still largely covered by noise. When the threshold has increased to a sufficiently high level, the log likelihood of the NNSD belonging to Poisson distribution increased sharply (blue curve with triangle markers), indicating the patterns hidden by noise start to prevail.

As depicted in Fig. 5(c), the NNSD began to deviate from GOE at the threshold of 0.95. It appears to closely follow the Poisson distribution when the threshold set to 0.99 (Fig. 5(d)). This was indeed the case when we applied the Chi-square goodness of fit test, in which the null hypothesis that the data are governed by a Poisson statistic was accepted ($\chi^2 = 84.85, p = 0.019$) as shown in Table II.

Thus, the clear transition from GOE to Poisson statistics at the threshold of 0.99 was used as a reference point to construct the co-abundance network in which condition specific relationships encoded in the correlation matrix can be better represented.

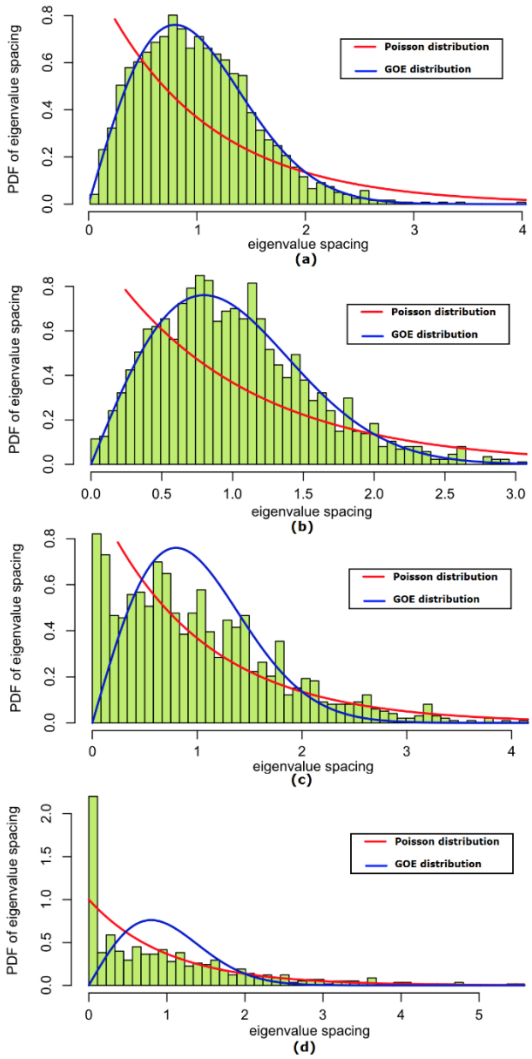


Fig. 5 The NNSD of the correlation matrix constructed from the abundance of 1570 microbial genes across 8 samples with different thresholds: (a) threshold = 0.0; (b) threshold = 0.90; (c) threshold = 0.95; and (d) threshold = 0.99.

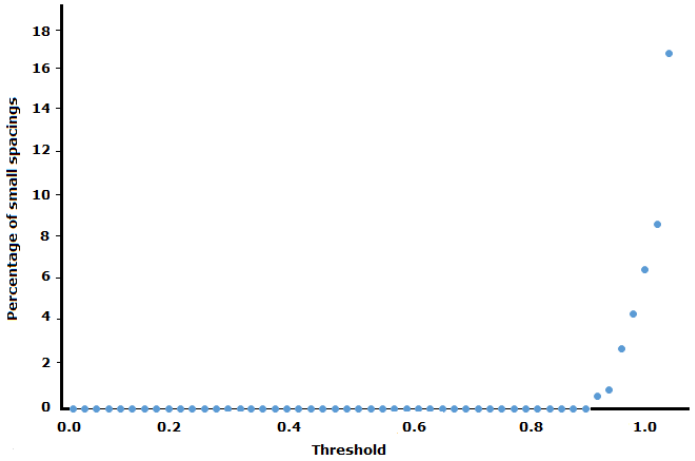


Fig. 6 The percentage of small eigenvalue spacings (less than 0.003) derived at different threshold

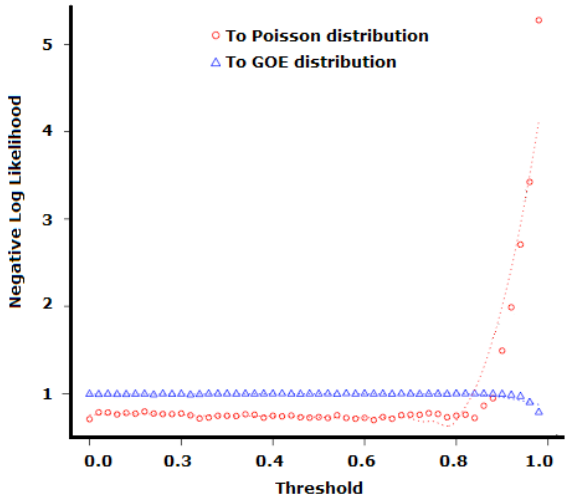


Fig. 7 The distance of the empirical NNSD to the GOE and Poisson distributions at the different thresholds.

TABLE II CHI-SQUARE (χ^2) GOODNESS-OF-FIT TESTS ASSOCIATED WITH EACH THRESHOLD

Threshold	χ^2	*p*-value
0.90	619.87	0.000
0.91	624.34	0.000
0.92	573.66	0.000
0.93	544.35	0.000
0.94	435.24	0.000
0.95	411.12	0.000
0.96	262.98	0.000
0.97	215.84	0.000
0.98	108.40	0.000
0.99	84.85	0.019

3.3 Co-abundance network

The network analysis of microbial gene abundance was illustrated in Fig. 8, in which each node stands for a microbial gene and the strength of each edge denotes the correlation in their abundance. Only the correlations between microbial gene abundances across 8 samples greater than 0.99 were kept. The network including 549 genes and 3349 links shows a clear modular structure with the largest component (Module A) having 237 nodes and 2860 edges. The topological parameters of the top 3 largest components, i.e. Modules A, B, and C, are

shown in Table III, each having a clustering coefficient significantly greater than a random graph constructed on the same number of nodes.

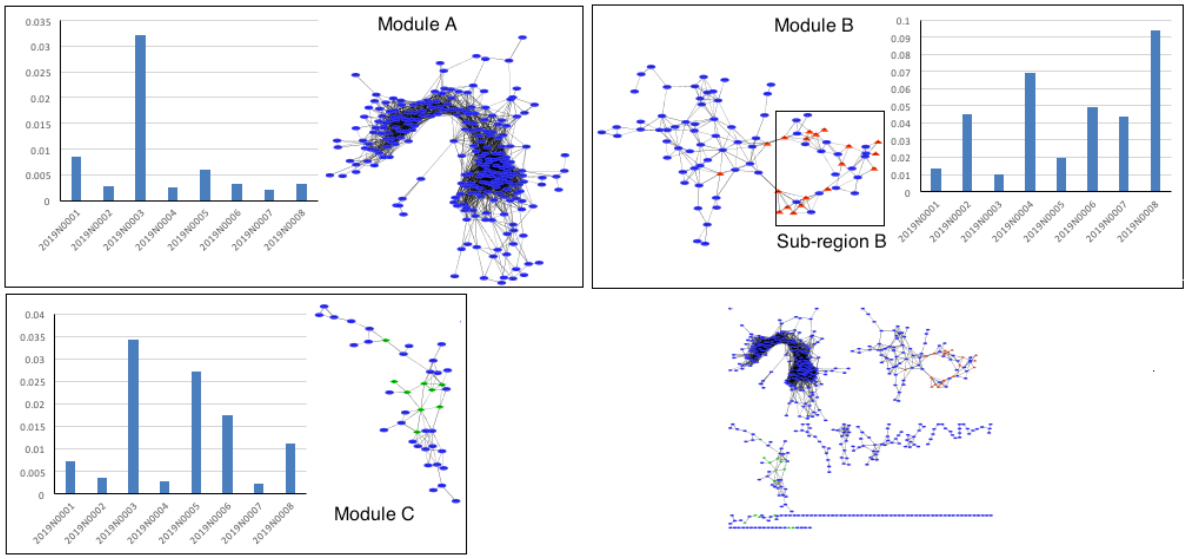


Fig. 8 Network-based approach to the correlation analysis of microbial gene abundance. The threshold used to construct the co-abundance network was set to 0.99. The network, in which each node represents a microbial gene and each edge indicates the correlation in their abundance, exhibits a clear modular structure. The average abundance of genes in top 3 largest modules, i.e. Modules A, B, and C, across 8 samples were shown. The whole network constructed is shown at the bottom right. The red triangle nodes denote genes associated with methane emissions while green diamond nodes are microbial genes linked to feed conversion efficiency.

TABLE III THE TOPOLOGICAL FEATURES OF TOP 3 LARGEST MODULES, I.E. MODULES, A, B, AND C. CPL: CHARACTERISTICS PATH LENGTH

Parameters	Module A	Module B	Module C
Number of nodes	237	91	41
Number of edges	2860	219	77
Network diameter	11	14	13
Network radius	6	7	7
Network density	0.102	0.053	0.094
Clustering coefficient	0.621	0.469	0.392
CPL	3.671	4.888	4.449
Network centralization	0.158	0.082	0.138
Network heterogeneity	0.736	0.531	0.163

3.4 Topological analysis

In an attempt to assess the topological relevance of each node, which may be linked to critical roles in certain biological mechanisms, we computed centrality indexes for each node

including degree, betweenness, eigenvector, bridging, closeness, PageRank, and power centralities, each representing a process by which a node might influence the flow of information through a network [26]. For example, it has been suggested that a node with high betweeness often referred to bottlenecks plays an important role in maintaining network integrity and paths of information flow [27]. Pangrank centrality can be used to identify important nodes of low degree [28]. Peng and Schork [29] explored the application of network centrality-based analysis to the identification of potential therapeutic targets for a tumor. They highlighted that eigenvector centrality has the potential to reveal genes that could serve as alterative therapeutic targets as nodes captured by eigenvector centrality are often connected to otherwise critical nodes.

Table IV shows the top 5 ranked microbial genes according to 7 centralities. Interestingly, the top 5 genes in Module B ranked by power centrality include two genes, i.e. K00584 and K00201, that are directly involved in methanogenesis [3].

TABLE IV THE TOP 5 GENES RANKED BY 7 CENTRAKITY METRICS IN MODULES A, B AND C

Centrality indexes	Top 5 ranked microbial genes in Module A
Degree	K02315, K00878, K00111, K04070, K00805
Betweeness	K02566, K05297, K07078, K01682, K11358
EigenVector	K06923, K00878, K00111, K00805, K04070
Bridging	K01626, K00016, K01615, K01878, K03152
Closeness	K05297, K01682, K11358, K02315, K06967
PageRank	K01295, K00763, K02315, K00878, K09687
Power Centrality	K07078, K05810, K06399, K01992, K07464
Centrality indexes	Top 5 ranked microbial genes in Module B
Degree	K01959, K07161, K03432, K09482, K04483
Betweeness	K03679, K00400, K07161, K01959, K04483
EigenVector	K01959, K07161, K03432, K04483, K07574
Bridging	K03390, K09726, K00440, K00125, K03679
Closeness	K07161, K04483, K03679, K03420, K00125

PageRank	K07161, K01959, K03432, K09482, K03044
Power Centrality	K09726, K06863, K00584, K00201, K14105
Centrality indexes	Top 5 ranked microbial genes in Module C
Degree	K13542, K03500, K06969, K07090, K03458
Betweeness	K13542, K07090, K09117, K00956, K03500
EigenVector	K02203, K09816, K02048, K03529, K06023
Bridging	K06023, K00956, K09117, K06179, K00974
Closeness	K13542, K00375, K00394, K00974, K07090
PageRank	K13542, K07090, K06969, K03500, K03458
Power Centrality	K11189, K00882, K04758, K03500, K02654

Further analysis was performed to discern if top ranked genes using topological analysis metrics on the co-abundance similarity network overlaped with KEGG pathways. We hypothesized that application of these metrics are important as previous studies have uncovered key players from biological networks using metrics such as degree (hubs), whereby network hubs are often essential [30]. For each centrality, a ranked list including the top 20% genes were selected and analysed. This percentage was selected as it has been previously applied as a cut-off threshold in the study [31].

A total of 8 KEGG pathways are shown representing the top pathways enriched with genes obtained from the various topological analyses (Table V). As shown in Table V, all the metrics have high overlap with the metabolic pathways. Interestingly, we can see that the ranked list derived from Module B using Degree, Bridging and PageRank centralities have the highest overlap with the methane metabolism pathway and microbial metabolism in diverse environments (which is of interest as it is related to methane production). Furthermore, these overlaps are statistically significant (Fisher Exact Test, $p<0.05$).

TABLE V THE OVERLAPPED BETWEEN THE TOP RANKED GEGES IDENTIFIED BY 7 CENTRALITIES WITH KEGG PATHWAYS.

Module A

Centrality	KO00680	KO01100	KO01110	KO01130	KO02010	KO01120	KO01230	KO01200
Degree	0	10	7	3	2	2	5	1
Betweeness	0	14	7	6	3	2	4	2
Eigenvzector	0	9	6	2	2	1	4	0
Bridging	0	18	7	6	3	3	4	0
Closeness	0	9	9	5	2	2	5	0
PageRank	0	12	7	4	3	1	4	0
Power Centrality	0	14	8	7	0	2	4	1

Module B

Centrality	KO00680	KO01100	KO01110	KO01130	KO02010	KO01120	KO01230	KO01200
Degree	7	11	2	2	0	8	2	7
Betweeness	5	9	0	0	0	6	1	5
EigenVector	5	8	1	1	0	6	2	5
Bridging	7	8	0	0	0	7	0	3
Closeness	4	6	0	1	0	5	1	3
PageRank	3	9	2	2	0	4	2	3
Power Centrality	5	7	1	1	0	5	0	4

Module C

Centrality	KO00680	KO01100	KO01110	KO01130	KO02010	KO01120	KO01230	KO01200
Degree	0	1	1	0	0	0	0	0
Betweeness	0	2	1	1	0	1	0	0
EigenVector	0	1	1	0	1	0	0	0
Bridging	0	1	0	1	0	1	0	0
Closeness	0	2	1	1	1	1		
PageRank	0	2	1	0	0	1	0	0
Power Centrality	0	4	2	1	0	5	0	0

*KO00680: Methane metabolism; KO01100: Metabolic pathways; KO01110: Biosynthesis of secondary metabolites; KO01130: Biosynthesis of antibiotics; KO02010: ABC transporters; KO01120: Microbial metabolism in diverse environments; KO01230: Biosynthesis of amino acids; KO01200: Carbon metabolism

3.5 Biological relevance

We first checked the abundance profile of genes in each module across 8 samples as depicted in Fig. 8. Interestingly, genes in both Modules A and C have a higher level of abundance in the low methane emission group (2019N001, 2019N003, 2019N005, and 2019N007) than in the other 4 samples with high methane emission, i.e. 2019N002, 2019N004, 2019N006, and 2019N008 (t-test, p<0.00001). On the contrary, a significantly high level of abundance was observed in the samples in the high methane emission group for 91 genes found in Module B (t-test, p = 4.2E-18) in which more than two third of genes in Module B have abundances differing between the high and low methane emission groups (T-test, p<0.05). This suggests that Module B be heavily linked to methane emissions. The examination of abundance profiles of 35 genes grouped in the sub-region in Fig. 8 further confirms the observation as shown in Fig. 9 in which all the genes have a low level of abundance in the samples assigned to the low methane emission group especially in the 2019N001, 2019N003, and 2019N005 samples. The top 10 ranked genes based on 7 centralities are shown in Table VI. Unexpectedly, among these 35 genes, K00400 involved in Methane metabolism pathway (ko00680) is ranked at the top in terms of 5 centrality metrics used (degree: 5; closeness: 0.0029; betweenness: 2005.81; eigenVector: 0.050; and bridging centrality: 160.36).

TABLE VI THE TOP 10 GENES IN THE SUB-REGION IN FIG. 4 RANKED BY 7 CENTRALITY INDEXES

Ranking	Degree	Betweenness	Bridging	Closeness	Eigenvector	pagerank	Power
1	K00400	K00400	K00400	K00400	K00400	K03044	K00584
2	K00577	K00581	K00581	K00581	K00581	K07041	K00201
3	K02007	K03045	K06174	K02930	K02007	K14128	K02322
4	K14128	K02930	K04076	K00577	K00577	K00400	K00399
5	K03044	K02122	K02930	K14128	K13812	K02007	K04076
6	K00581	K00577	K00203	K03045	K02930	K00577	K06932
7	K03045	K04076	K03045	K02007	K00672	K03045	K00441
8	K00672	K06174	K02122	K00672	K14128	K02322	K11600
9	K07041	K02007	K00577	K13812	K00441	K13525	K14123
10	K13812	K00203	K00672	K00441	K03388	K00581	K00205

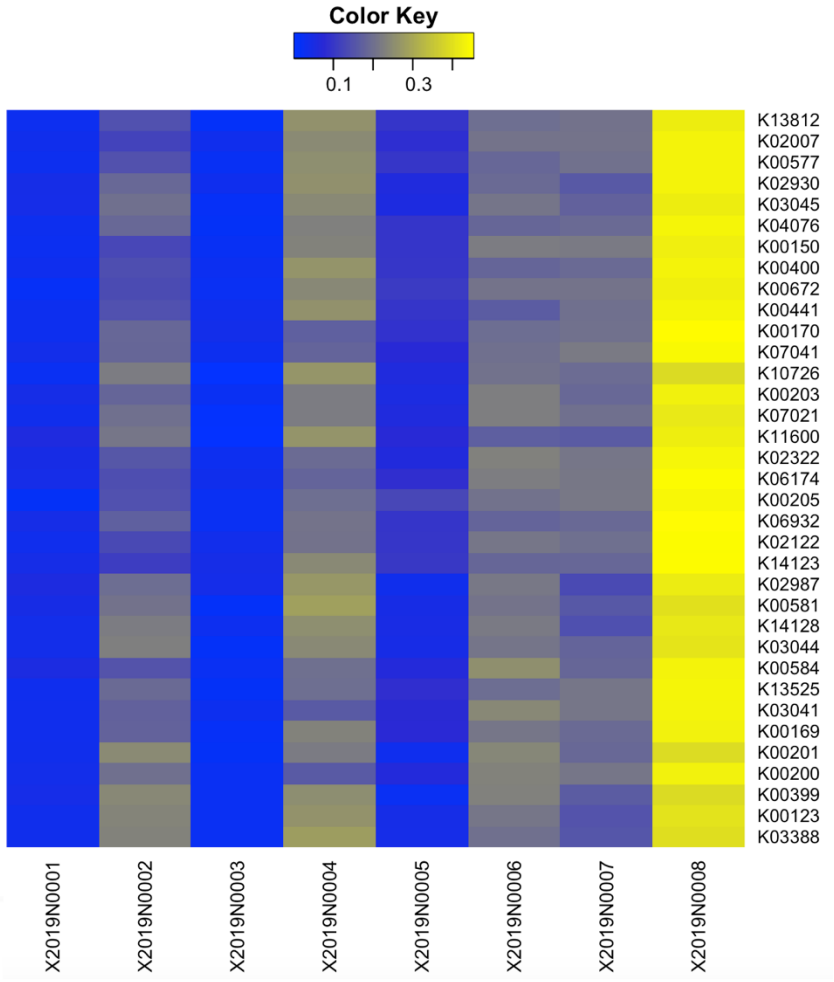


Fig. 9 The heatmap of the relative abundance of microbial genes grouped in the sub-region in Fig. 8

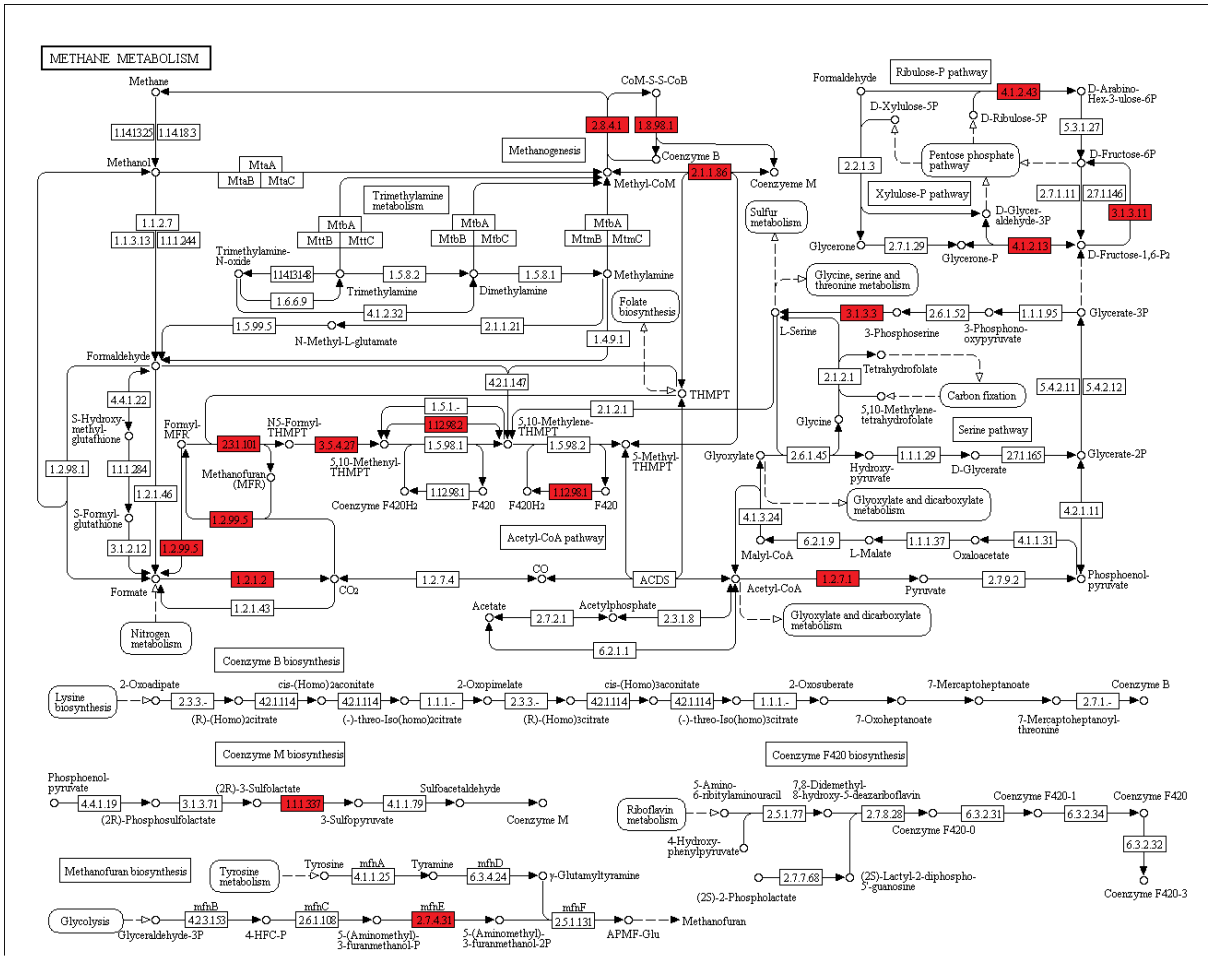
An analysis with regard to the distribution of microbial genes strongly associated with traits indicates that methane emission-specific genes are highly over-represented in Module B. For example, all the 20 genes identified to be associated with methane emissions by Roehe et al. [6] represented by red triangle nodes in Fig. 8 were found in Module B (hypergeometric test, $p < 10^{-11}$). Out of 25 genes encoding enzymes that are directly involved in the methane production pathway studied in Wallace et al. [3], 18 were found in the network, 15 of which were assigned to Module B (hypergeometric test, $p < 10^{-9}$) as dispicted in Table VII.

TABLE VII ABUNDANCE PROFILE OF GENES ENCODING ENZYMES INVOLVED IN METHANOGENESIS

KEGG genes	Abundance in low emission group	Abundance in high emission group	*p* value (t test)	Encoding enzymes involved in methanogenesis
K00123	0.092	0.250	0.002	EC:1.2.1.2 formate dehydrogenase
K00200	0.058	0.154	0.002	EC:1.2.99.5 formylmethanofuran dehydrogenase
K00201	0.066	0.181	1.1E-06	
K00672	0.017	0.048	0.014	EC:2.3.1.101 Formyl methanofuran--tetrahydromethanopterin N-formyl transferase
K01499	0.024	0.080	0.027	EC:3.5.4.27 Methenyl tetrahydromethanopterin cyclohydrolase
K00577	0.022	0.063	0.022	EC:2.1.1.86 tetrahydromethanopterin S-methyltransferase
K00581	0.035	0.102	0.006	
K00584	0.035	0.106	0.011	
K00399	0.102	0.275	7.6E-05	EC:2.8.4.1 methyl-coenzyme M reductase
K00401	0.069	0.185	0.025	
K00402	0.035	0.101	0.017	
K00440	0.032	0.083	0.042	EC:1.12.98.1 coenzyme F420 hydrogenase
K00441	0.016	0.047	0.029	
K03388	0.117	0.334	0.002	EC:1.8.98.1 heterodisulfide reductase
K03390	0.015	0.047	0.036	

We then turned to the analysis of the involvement of KEGG pathways in each module. A total of 86, 45, and 23 pathways were found to be involved by microbial genes in Modules A, B, C respectively. As expected, the largest portion of genes in each module are involved in KEGG metabolic pathway (KO01100). However, a close look reinforces our observation that genes in Module B have a strong association with methane emission. Nearly one third microbial genes grouped in Module B are involved in methane metabolism pathway. There

are a total of 36 KEGG genes in the co-abundance network that are involved in methane metabolism pathway, 30 of which are found in Module B (hypergeometric test, $p < 10^{-10}$) as illustrated in Fig. 10.



Genes and its encoding enzymes:

K00440, K00441: EC1.2.98.1; K00123: EC1.2.1.2; K00200, K00201: EC 1.2.99.5; K00672: EC 2.3.1.101; K03388, K03390: EC1.8.98.1; K01499: EC 3.5.4.27
K00399, K00401, K00402: EC 2.8.4.1; K0057, K00581, K00584: EC 2.1.1.86; K07144: EC 2.7.4.31; K13942: EC 1.12.98.2; K13812: EC 4.1.2.43
K00125: EC1.2.1.2; K00202, K00203: EC 1.2.99.5; K01622: EC 3.1.3.114.1.2.13; K01623: EC 4.1.2.13; K05884: EC 1.1.1.337; K01079: EC 3.1.3.3; K00169, K00170: EC 1.2.7.1

Fig. 10 KEGG methane metabolism pathway with EC gene numbers found in Module B (Highlighted in red). The genes and their coding enzymes are listed at the bottom.

4 Conclusions

Recent years have seen a growing use of metagenomics-based approaches to study the full extent of microbial diversity, as well as the association between host genetic and microbial activities. This study investigated the rumen microbial community in cattle through the integration of metagenomics and network-based approaches. Based on the relative

abundance of 1570 microbial genes identified in a metagenomics analysis, the co-abundance network was constructed and functional modules of microbial genes were identified. One of the main contributions is to develop a RMT-based approach to automatically determine the correlation threshold used to construct the co-abundance network. It has been shown that the network exhibits a highly modular structure with each module well separated. The involvement of KEGG pathways in each module was analysed and compared. A close look at the abundance profiles highlights that two modules i.e. Modules B and C are strongly associated with methane emissions and feed conversion efficiency respectively (hypergeometric test, $p < 10^{-6}$).

This study contributes to the development of automated computational methods to supporting the identification of functional modules of microbial genes through integration of metagenomics and network-based approaches. Given that the association between microbial genes can be realized via different mechanisms, we are now working toward a multiplex network-based approach to the analysis of the composition of rumen microbial community [32], [33]. In addition we are building user friendly interfaces to this metagenomics analysis on the *Simplicity* bioinformatics cloud computing platform to provide access to this analysis to researchers working on metagenomics projects in a reproducible manner [34].

This research has been undertaken as the European Commission (EC) funded MetaPlat project (www.metaplat.eu). The EC increasingly requests that funded projects follow specific data management regulations, to optimize sharing of research results and its later validation through proper reproducibility. In essence, sharing and later validation is enforced from the EC, because research undertakings are expensive and the return on investment needs to be secured by research purchasers through proper management of the knowledge that is required for long term research reuse. As such, we are working on OAIS (Reference Model for

an Open Archival Information Systems, (cf. [35], ISO 14721), that builds a framework of terms and concepts to specify an archival information system. Within OAIS so called Information Packages (IP) is used to describe the relation of applied research data, beside the knowledge required to enable its later comprehensive reuse. In terms of OAIS this is classified as Content Information and Preservation Description Information.

Our hypothesis is that enabling extensive reproducibility for long term reusability is fundamentally dependent on the substantial and consistent representation of all information that came into existence along the phases of the introduced information lifecycle. We argue that the OAIS Information Model, could act here as an abstract specification of the structure and the constituting components of a metagenomics research, that could be refined by means of further introduced community specific standards. Hence, we will, in the course of the project runtime, elaborate on the comprehensive representation, integration and validation of introduced standards into the OAIS information Model by means of technologies in the context of the Semantic Web. Furthermore, we will undertake research in the unambiguous documentation of involved resources and their interrelation (e.g. SRUC data set, technologies like *NetworkAnalyzer* or *CentiScaPe* and applied methods) and to clearly specify all these resources in compliance to OAIS.

In the current study, the co-abundance network was constructed by computing pairwise correlation between two microbial genes, which includes both direct and indirect associations. An important part of our future research is to investigate direct associations between variables by calculating partial correlation and its impact on the network modular structure [36], [37]. Another potential direction of the future research is to explore the potential to detect critical states of key players associated with methane emission [38].

Acknowledgment

This work was supported in part by the MetaPlat project (www.metaplat.eu) funded by H2020-MSCA-RISE-2015.

References

[1] H.J. Lee, J.Y. Jung, Y.K. Oh, S.-S.Lee, E.L. Madsen, and C.O. Jeon, Comparative Survey of Rumen Microbial Communities and Metabolites across One Caprine and Three Bovine Groups, Using Bar-Coded Pyrosequencing and 1H Nuclear Magnetic Resonance Spectroscopy, Applied and Environmental Microbiology, 2012, 78(17), 5983–5993.

[2] M.B. Lengowski, KHR Zuber, M. Witzig, J. Möhring, J. Boguhn, M. Rodehutscord, Changes in Rumen Microbial Community Composition during Adaption to an *In Vitro* System and the Impact of Different Forages, PLoS ONE, 2016 11(2): e0150115.

[3] R. J. Wallace, J.A. Rooke, N. McKain, C-A. Duthie, J. J. Hyslop, D. W. Ross, et al. The rumen microbial metagenome associated with high methane production in cattle, BMC Genomics. 2015;16: 839. doi: 10.1186/s12864-015-2032-0. pmid:26494241

[4] D.E. Beever, The impact of controlled nutrition during the dry period on dairy cow health, fertility and performance, Animal Reproduction Science, 2006, 96(3–4), 212–226.

[5] D.P. Morgavi, E. Rathahao-Paris, M. Popova, J. Boccard, K. F. Nielsen, H. Boudra, Rumen microbial communities influence metabolic phenotypes in lambs, Frontiers in Microbiology, 2015;6:1060. doi:10.3389/fmicb.2015.01060.

[6] R. Roehe, R.J. Dewhurst, C-A. Duthie, J.A. Rooke, N. McKain, et al., Bovine host genetic variation influences rumen microbial methane production with best selection criterion for low methane emitting and efficiently feed converting hosts based on metagenomic gene abundance, PLoS Genet., 2016, 12: e1005846. doi:10.1371/journal.pgen.1005846.

[7] G. Henderson, F. Cox, S. Ganesh, A. Jonker, Y. Wayne, et al. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range, Scientific Reports. 2015, 5, 14567 (http://dx.doi: 10.1038/srep14567).

[8] C. McSweeney, S. Kang, E. Gagen, C. Davis, M. Morrison, and S. Denman, Recent developments in nucleic acid based techniques for use in rumen manipulation, Revista Brasileira de Zootecnia, 2009, 38(spe), 341-351

[9] C. J. Creevey, W. J. Kelly, G.Henderson, and S. C. Leahy, Determining the culturability of the rumen bacterial microbiome, Microbial Biotechnology, 2014, 7(5), 467–479. http://doi.org/10.1111/1751-7915.12141

[10] M.P. Bryant, Bacterial species of the rumen, Bacteriol Rev., 1959, 23(3), 125-153.

[11] H. Andreas, G. Joachim, R. Udo, and G. André, Analyses of intestinal microbiota: culture versus sequencing, ILAR J, 2015, 56 (2), pp.228-240 doi:10.1093/ilar/ilv017

[12] P. H. Janssen and M. Kirs, Structure of the Archaeal Community of the Rumen, Applied and Environmental Microbiology, 2008, 74(12),3619–3625. http://doi.org/10.1128/AEM.02812-07

[13] A. Oulas, C. Pavloudi, P. olymenakou, G.A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, et al. Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies, Bioinformatics and Biology Insights, 2015, 9, pp.75–88. http://doi.org/10.4137/BBI.S12462

[14] J. Handelsman, Metagenomics: Application of Genomics to Uncultured Microorganisms, Microbiology and Molecular Biology Reviews, 2004, 68(4), 669–685.

[15] J. J. Faith, N. P. McNulty, F. E. Rey, et al., Predicting a human gut microbiota's response to diet in gnotobiotic mice., Science, vol. 333, no. 6038, 101–104, Jul. 2011.

[16] HY. Wang, H. Zheng, F. Browne, R. Roehe, R.J. Dewhurst, F. Engel, M. Hemmje, and P. Walsh, Analysis of rumen microbial community in cattle through the integration of metagenomic and network-based approaches, In the Proc. Of 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), , pp. 198-203, 2016, doi:10.1109/BIBM.2016.7822518.

[17] F. Luo, Y. Yang, J. Zhong, H. Gao, L. Khan, D. Thompson, *et al.* Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory, BMC Bioinformatics, 2007 8:299.

[18] E. Wigner, On the distribution of roots of certain symmetric matrices, Ann. Math., 1958, 67(2), 325-327.

[19] J. Bouchaud and M.Potters, Financial applications of random matrix theory: a short review, in: The Oxford Handbook of Random Matrix Theory, 2009, pp.824-850.

[20] F.Luo, P.Srimani, and J. Zhou, Application of random matrix theory to analyze biological data, in: Handbook of data intensive computing, B.Furht and A. Escalante, Ed. Springer Science+Business Media, 2011, pp.711-732.

[21] Y. Malevergne and D. Sornette, Collective origin of the coexistence of apparent random matrix theory noise and of factors in large sample correlation matrices, Physica A: Statistical Mechanics and its Applications, 2003. 331(3–4), 660–668.

[22] Y. Deng, Y-H. Jiang, Y. Yang, Z. He, F. Luo, J. Zhou, Molecular ecological network analyses, BMC Bioinformatics, 2012,**13**, 113

[23] Y. Assenov, F. Ramírez, S.E. Schelhorn, T. Lengauer, M.Albrecht, Computing topological parameters of biological networks, Bioinformatics, 2008, **24**(2), 282-284.

[24] G. Scardoni, M. Petterlini, C. Laudanna, Analyzing biological network parameters with CentiScaPe, Bioinfomatics, 2009, 25 (21), 2857-2859

[25] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Research 2003 Nov; 13(11), 2498-504

[26] T.W. Valente, K. Coronges, C. Lakon, E. Costenbader, How Correlated Are Network Centrality Measures? Connections (Toronto, Ont). 2008;28(1):16-26.

[27] A. L. Barabási, Z. N. Oltvai, Network biology: understanding the cell's functional organization, Nat Rev Genet. 2004 Feb;5(2):101-13.

[28] D. Bánky, G. Iván, and V. Grolmusz, Equal Opportunity for Low-Degree Network Nodes: A PageRank-Based Method for Protein Target Identification in Metabolic Graphs, PLoS ONE, 2013. 8(1): e54204. doi:10.1371/journal.pone.0054204

[29] Q.Peng and N. Schork, Utility of network integrity methods in therapeutic target identification, Frontiers in Genetics, 2014, DOI=10.3389/fgene.2014.00012

[30] H. Jeong H, S.P. Mason, A.L. Barabasi, Z.N. Oltvai ZN, Lethality and centrality in protein networks, Nature, 411(2001), 41–42.

[31] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, Theimportance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics, PLoS Comput. Biol. vol. 3, no. 4, pp. 713–720, 2007.

[32] P. Mucha, T. Richardson, K. Macon, M. Porter, J. Onnela, Community structure in time-dependent, multiscale, and multiplex networks, Science, 208(2010), 876-878, 2010.

[33] H.Y. Wang H. Zheng, J. Wang, C. Wang and F.X. Wu, Integrating omic data with a multiplex network-based approach for the identification of cancer subtypes, IEEE Transactions on NanoBioscience, 2016, 15(4), 335-342.

[34] P. Walsh, J. Carroll, and R. D. Sleator, "Accelerating in silico research with workflows: a lesson in simplicity, Computers in biology and medicine, 2013, 43(12), 2028-2035

[35] CCSDS: Reference model for an open archival in-formation sys-tem (oais). Pink Book 1, Consulta-tive Committee for Space Data Systems (2012).Rec-ommendation for Space Data Sys-tems Standards, adopted as ISO 14721:2012.

[36] R. Opgen-Rhein anf K. Strimmer, From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data, BMC Systems Biology20071:37, DOI: 10.1186/1752-0509-1-37

[37] A. de la Fuente, N. Bing, I. Hoeschele and P. Mendes, Discovery of meaningful associations in genomic data using partial correlation coefficients, Bioinformatics, 20(18), 2004, pp. 3536-3574.

[38] L. Chen, R. Liu, Z. Liu, M. Li and K. Aihara, Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers, Scientific Reports 2: 342 (2012) doi:10.1038/srep00342