Improving the Inference of Co-occurrence Networks in the Bovine Rumen Microbiome

Huiru Zheng, *Senior Member*, *IEEE*, Haiying Wang*, Richard J. Dewhurst, and Rainer Roehe

Abstract— The importance of the composition and signature of rumen microbial communities has gained increasing attention. One of the key techniques was to infer co-abundance networks through correlation analysis based on relative abundances. While substantial insights and progress have been made, it has been found that due to the compositional nature of data, correlation analysis derived from relative abundance could produce misleading results and spurious associations. In this study, we proposed the use of a framework including a compendium of two correlation measures and three dissimilarity metrics in an attempt to mitigate the compositional effect in the inference of significant associations in the bovine rumen microbiome. We tested the framework on rumen microbiome data including both 16S rRNA and KEGG genes associated with methane production in cattle. Based on the identification of significant positive and negative associations supported by multiple metrics, two co-occurrence networks, e.g. co-presence and mutual-exclusion networks, were constructed. Significant modules associated with methane emissions were identified. In comparison to previous studies, our analysis demonstrates that deriving microbial associations based on the correlations between relative abundances may not only lead to missing information but also produce spurious associations. To bridge together different co-presence and mutual-exclusion relations, a multiplex network model has been proposed for integrative analysis of co-occurrence networks which has great potential to support the prediction of animal phytotypes and to provide additional insights into biological mechanisms of the microbiome associated with the traits.

Index Terms— Compositional data, co-occurrence networks, rumen microbiome, methane emission

1 Introduction

WHILE ruminant livestock play a major role in human food production and sustainable agricultural systems, methane production from ruminants contributes significantly to global anthropogenic greenhouse gas emissions [1-3]. Given the role of rumen microorganisms which predominantly consist of bacteria, archaea, protozoa and fungi [4] in the fermentation process, there is a growing effort to examine the composition of rumen microbial communities and their associations with phenotypical traits. It has been highlighted that rumen microorganisms play a vital role in their host's physiology and without a healthy microbial population in the rumen, ruminants cannot function properly [5]. For example, the rumen microbiota is important for fermentation of fibre to produce short-chain fatty acids, which are the main source of energy for ruminants. Without this fermentation, ruminants would not have their unique role of converting human inedible fibrous feeds, such as grass and forages, into high-quality protein foods, such as milk and beef. Hydrogen is also produced by fermentation and must be utilised, by processes such as methanogenesis, biohydrogenation of unsaturated fatty acids and reduction of nitrate or sulphate, to avoid it accumulating and so inhibiting fermentation. Various studies have demonstrated the influence of rumen microbial communities on animal phenotypes [6], [7]. More recently, Schären et al. [8] investigated the interrelations between the rumen microbiota and a range of production traits in dairy cows and concluded that in order to have a better understanding of the host-microbiome interaction and its dynamic, further investigation using more sophisticated methods to describe phenotypical traits of the host as well as the rumen microbiome is needed.

Due to the ability to reveal the full spectrum of microbial diversity, next-generation sequencing (NGS)-based metagenomics analysis has attracted great attention. Examples include the study conducted by Henderson et al. [9], which performed metagenomics analysis of 742 samples collected from 32 animal species and 35 countries and found that rumen microbial community composition varies with diet and host. However, similar bacteria and archaea are found to dominate in nearly all samples across a wide geographical range while protozoal communities were more variable. It has been suggested that differences in microbial community compositions were predominantly attributable to diet, with the host being less influential. Lengowski et al. [10] examined ruminal microbial community composition alterations during adaption and incubation in an *in vitro* rumen simulation system using different forages. It was shown that the ruminal microbial community can be influenced significantly by sampling time and forage source, but was a stable system after 48h. . Using young ruminants subjected to different microbial-modulating interventions, Morgavi et al. [7] reveals the affect of the gut microbiota on animal phenotype and its metabolites.. The recent study carried out by Roehe et al. provides a comprehensive insight into host-microbe interactions in the rumen

- *H. Zheng is with the School of Computing, Ulster University, Northern Ireland, United Kingdom, BT37 0QB. E-mail: h.zheng@ulster.ac.uk.*
- *HY.Wang* is with the School of Computing, Ulster University, Northern Ireland, United Kingdom, BT37 0QB. E-mail: hy.wang@ulster.ac.uk.*
- *R. J. Dewhurst is with Future Farming Systems, Scotland's Rural College, Edinburgh, United Kingdom, EH9 3JG. Email: Richard.Dewhurst@sruc.ac.uk.*
- *Rainer Roehe is with Future Farming Systems, Scotland's Rural College, Edinburgh, United Kingdom, EH9 3JG. Email: Rainer.Roehe@sruc.ac.uk.*

and highlighted that the host animal controls its own microbiota to a significant extent [6].

One of the key research areas in NGS-based metagenomics data analysis is to infer association and dependencies between members of microbial communities through correlation analysis [11]. For example, Williams et al. [12] introduced a framework to explore biological interactions occurring within microbial communities, in which the strength of correlation is derived from the calculation of the Spearman's correlation. The co-occurrence analysis can be performed at multiple scales ranging from the community level down to pairwise interactions between microbial taxa. Based on the relative abundance of 1570 KEGG genes across 8 samples, Roehe et al. [6] constructed a co-abundance network where nodes represent microbial genes and edges reflect the correlation in their abundance. They have successfully identified a close sub-network of the microbial genes associated with feed conversion efficiency and methane emissions respectively. Wang et al. [13] applied a random matrix theory-based approach for determination of the correlation threshold used to construct the co-abundance microbial network.

Despite encouraging results and substantial insights being obtained, the correlation-based approaches to the inference of associations between microbial genes exhibit some limitations [11], [14]. Due to the nature of data generation and the normalization process involved, the abundance derived from NGS is a relative measurement associated with each microbial gene. As such, abundances of genes estimated under certain condition are not completely independent of each other. It has been shown that simply applying correlation-based techniques to the analysis of such compositional data may produce misleading results [14].

Based on our previous investigation [15], this study aims to further explore the ways to enhance the inference of co-occurrence networks in rumen microbiome. The framework including a compendium of correlation and dissimilarity measures to mitigate the effect of compositionality has been further tested on a 16S rRNA data. One of the main objectives is to infer both co-presence and mutual exclusion networks associated with methane emissions. To bridge together different co-presence and mutual-exclusion relations, a multiplex network-based model has been proposed and utilized. The remainder of this paper is organized as follows. Section II briefly describes the framework and methodology used in this study, including datasets and an ensemble of correlation of dissimilarity metrics. Section III presents the results and discussion. The paper concludes with a summary of contributions and limitations of this study followed by the direction of future research.

2 METHODOLOGY

We followed the approach introduced in [14] by using a compendium of similarity/dissimilarity measures for the analysis of rumen metagenomics data which include the relative abundance of 1570 KEGG genes and 76 taxonomic units at genus level [6]. Without loss of generality, a network including 1000 top-ranking and 1000 bottom-ranking

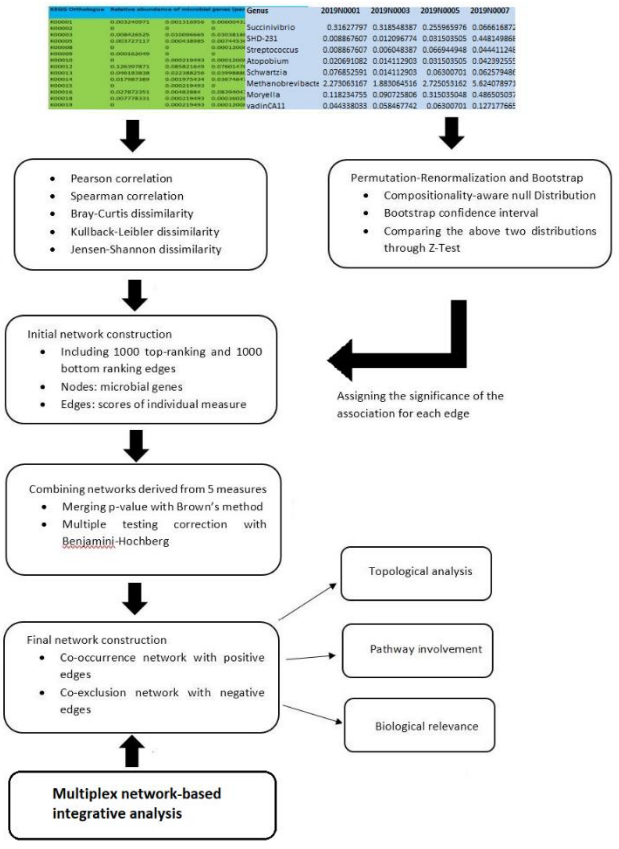

Fig. 1 A diagram to illustrate the key steps involved in the study.

edges was constructed for each measure. To assess the significance of scores associated with each edge, we applied the Permutation-Renormalization and Bootstrap (ReBoot) method [14], which can construct a null distribution that reflects the compositional nature of the data. After merging *p*-values and multiple testing corrections, a final network consisting of significant co-presence (positive interaction) and mutual-exclusion (negative interaction) patterns was extracted. The resulting network was further examined in terms of topological analysis, biological relevance and pathway involvement. To investigate the potential crosstalk between co-occurrence networks, a multiplex network-based approach was proposed. The key steps involved in the study are illustrated in Fig. 1

2.1 Rumen Metagenomics Data

The data applied in this research was released by Roehe and his colleagues in a study [6] in which a 2 × 2 factorial design experiment of breed types and diets was performed using 72 steers from a two-breed rotational cross between Aberdeen-Angus (AA) or Limousin cattle (LIM). For each of the breed/diet combination, the lowest and highest methane emitters were identified. DNA were extracted from the rumen content taken from these 8 extreme animals and subjected to qPCR of 16S rRNA genes and to deep sequencing using the Illumina HiSeq platform. For 16SrRNA gene analysis, the genomic reads were aligned to the GREENGENES database. The number of reads that were assigned to taxonomic groups at kingdom, phylum and genus levels were counted and normalized. For functional

analysis, the geneomic reads were aligned to the KEGG genes database and a total of 3970 KEGG gene orthologues were identified. The detailed description of data generation can be found in [2] and [6].

In this study, the abundances of 76 genera and 1570 KEGG genes showing a relative abundance of more than 0.001% were used. The charactersitics of 8 extreme animals are depicted in Table I.

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	Methane emissions	Breed/Diet	Archaea:Bacteria ratio
2019N0001	LOW	AA/CON	1.16:98.84
2019N0002	HIGH	AA/CON	2.28:97.72
2019N0003	LOW	LIM/CON	0.76:99.24
2019N0004	HIGH	LIM/CON	4.92:95.08
2019N0005	LOW	AA/FOR	1.18:98.82
2019N0006	HIGH	AA/FOR	3.40:96.60
2019N0007	LOW	LIM/CON	2.94:97.07
2019N0008	HIGH	LIM/CON	4.40:95.60

2.2 An Emsemble of Similarity and Disimilarity Measures

In order to mitigate the effect of compositionality on the analysis of rumen microbiome data, a compendium of two correlation measures, i.e. Spearman and Pearson correlations, and three dissimilarity metrics that are intrinsically robust to compositionality [14], i.e. Bray-Curtis dissimilarity (BC), Kullback-Leibler dissimilarity (KL), and Jensen-Shannon dissimilarity (JS) were utilized.

Let x and y be two vectors containing relative abundances across samples for two microbial genes. The three dissimilarities are defined as follows.

$$BC(x,y) = 1 - \frac{2\sum_k |x_k - y_k|}{\sum_k x_k + \sum_k y_k} \quad (1)$$

$$KL(x,y) = \sum_k (x_k \times \log\frac{x_k}{y_k} + y_k \times \log\frac{y_k}{x_k}) \quad (2)$$

$$JS(x,y) = \sum_k (x_k \log\frac{2x_k}{(x_k + y_k)} + y_k \log\frac{2y_k}{(x_k + y_k)}) \quad (3)$$

2.3 Statistical Significance of Emsemble Scores

To evaluate the significance of the association accounting for compositionality, we applied a nonparametric test based on the ReBoot method introduced in [14]. Unlike a standard procedure based on permutation test that essentially removes compositional effects and thus fails to identify spurious compositional correlations, the ReBoot method introduces sample-wise renormalization after permuting the abundance across samples. Such an approach leads to the construction of compositionality-aware null distribution. Comparing this null distribution to a standard bootstrap confidence interval, an appropriate significance level of the observed correlation can be established.

In this study, both permutation and bootstrap score distributions were computed with 100 iterations. Any edge with a score that falls outside of the bootstrapped confidence interval was removed.

2.4 Network Merging

After constructing a measure-specific network in which a node stands for a microbial gene and a score associated with each edge represents the strength of the association between two genes, we combined all the networks using Brown's method [16] which is an extension of Fisher's method for combining tests of significance when all the variables are not jointly independent. The merged p-values on each final edge were adjusted using the Benjamini-Hochberg false discovery rate (FDR) correction and the final network was thresholded at a q-value less than 0.05.

2.4 Multiplex networks

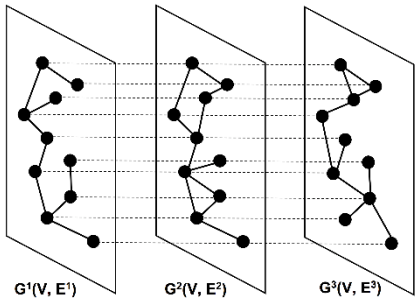

Fig. 2 An illustration of a multilayer network in which a set of 10 nodes, V, are present in three complementary layers: G1(V, E1), G2(V, E2), and G3(V, E3). Dot lines represent inter-layer connec-tions and black solid lines link nodes within a layer. Superscript 1, 2, and 3 denote a layer.

A multilayer network as illustrated in Fig. 2, in which each layer represents one type of interactions between nodes in the network, has emerged as a new paradigm in network science to study complex systems arising from diverse disciplines including biology, physics and social science. In this study, we used a multiplex network model for integrative analysis of co-occurrence networks. The model has been successfully applied to combine heterogeneous omics data for the identification of cancer subtypes [20].

2.4 Software packages used

The co-occurrence networks were constructed using the CoNet app [17] which offers a variety of approaches for inference of biological meaning. The network visualization was implemented using Cytoscape [18], an open source software platform for visualizing complex networks. The computation of topological parameters was with the NetworkAnalyzer plugins [19]. Analysis and visualization of multiplex networks was implemented within the

framework provided by MuxViz [22].

3 Results

3.1 Co-occurrence networks derived from relative abundances of KEGG microbial genes

Only the interactions with an FDR corrected p value (q-value) less than 0.05 were kept in the final co-ocurrence networks derived fom the relative abundance of 1570 microbial genes. The co-presence network (Fig. 3) consists of 790 nodes (microbial genes) and 2106 edges with positive scores, 537 of which are supported by at least two metrics with a q value below 0.05. The mutual-exclusion network (Fig. 4) is composed of 763 negative significant associations interactions between 473 microbial genes, in which 382 edges are supported by more than one metric. A close look highlights that all the pairs supported by the Spearson's correlation in the co-presence network exhibit a perfect monotonic relationship. None of links in the mutual-exclusion network is supported by all the 5 metrics used while in the co-presence network a total of 10 pairs are significantly supported by all 5 metrics as listed in TABLE II

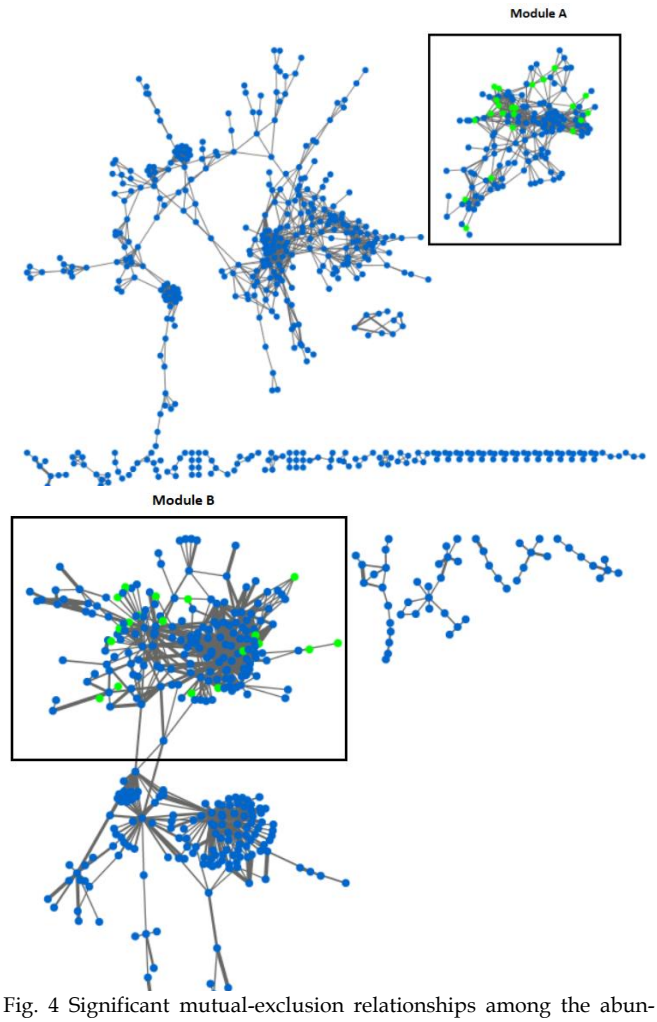

Fig. 4 Significant mutual-exclusion relationships among the abundances of KEGG microbial genes in the rumen microbiome. The width of edges is proportional to the level of significance of supporting evidence. Green nodes represent genes encoding enzymes that are directly involved in the methane production pathway.

TABLE II INTERACTIONS SUPPORTED BY BOTH CORRELATION METRICS AND THREE DISSIMILARITY MEASURES

Interaction type	Interactor A	Interactor B	Corrected p value
co-presence	K13812	K00577	0.000
co-presence	K13812	K00400	0.000
co-presence	K02908	K14105	0.000
co-presence	K00577	K00400	0.000
co-presence	K00320	K14127	0.000
co-presence	K07388	K01623	0.000
co-presence	K00123	K14128	0.000
co-presence	K00123	K03388	0.000
co-presence	K09154	K03042	0.000
co-presence	K03832	K03303	0.000

A. Biological relevance

Given that the extreme animals selected in the data collection carried out by SRUC [2], [6] were based on methane emissions, we first checked the distribution of methane specific-microbial genes in both networks. The level of the enrichment of trait-specific genes can be quantitatively expressed by the hypergeometric distribution probability calculated as follows.

$$p = 1 - \sum_{i=0}^{m-1} \binom{m}{i} \binom{N-m}{n-i} / \binom{N}{n} \quad (4)$$

where m is the number of microbial genes found in a module, i is the number of genes in the module associated with certain trait, N is the total number of microbial genes contained in the network and n is the number of trait-specific genes associated found in the network.

We found that, out of 31 genes that are directly involved in the methane production pathway studied in Wallace et al. [2], twenty-two and nineteen were found in the co-presence and mutual-exclusion networks respectively and all of them are grouped in Module A and B respectively ($p < 10^{-15}$). Furthermore, nineteen out of 20 methane emission specific genes identified by Roehe et al. [6] are contained in the co-presence network and grouped together in Module A ($p < 10^{-11}$). Based on these figures, one may confidently assume that Modules A and B are co-occurrence networks significantly associated with methane production.

We then turned to the topological analysis of Modules A and B. Both modules have a low average path length of less than four in comparison to 6 found in random networks on average [14]. Surprisingly, the clustering coefficient of Module B is equal to 0, indicating that none of the neighbours of nodes in Module B are connected. Moreover, Module B is more heterogeneous than Module A as indicated by the metric of network heterogeneity which reflects the tendency of a network containing hub nodes.

There are two hub nodes in Module B having a degree more than 50. The top node (K06013, STE24 endopeptidase

[EC:3.4.24.84]) exhibits significant mutual exclusion patterns over samples with 59 microbial genes supported by all three dissimilarity measures (BC, KL and JS) with a q-value less than 0.05. Similarly, K03780 is linked to 57 microbial genes in the form of strong mutual exclusions ($q < 0.05$).

In Module A, the most connected node is an uncharacterized protein (K07161), which shows significant co-presence patterns with 30 microbial genes across samples with a corrected p value less than 0.00001. In particular, it exhibits a similar abundance pattern (Fig. 5) across 8 samples with five genes (K00581, K00125, K00202, K00402 and K00401) encoding enzymes involved in the methane production pathway and four microbial genes associated with methane emissions (K00581, K00125, K01499 and K00169). As shown in Fig. 5, K07161 has a relative high level of abundance in the samples in the high methane emission group, suggesting this uncharacterized protein might be involved in the methane production pathway.

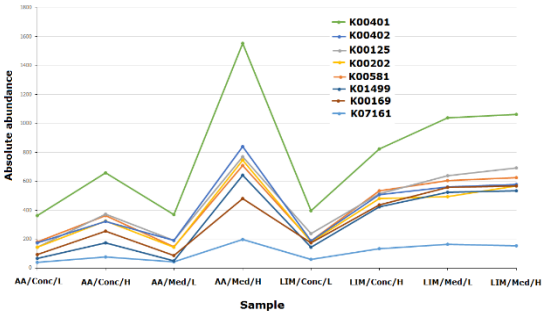

Fig. 5 The absolute abundance profile of the co-presence pattern observed in K07161 and seven microbial genes relevant to methane emissions. AA: Aberdeen Angus, LIM: Limousin, Conc: concentrate diet, Med: medium concentrate diet, L: Low methane emissions, H: High methane emissions.

B. Pathway analysis

The interaction partners in the co-presence network for genes encoding enzymes involved in methanogenesis are depicted in Fig. 6. As expected, no mutual-exclusion patterns were observed among KEGG orthologues representing enzymes involved in methane production while there are a number of strong positive interactions among methane specific microbial genes. The significant co-presence patterns were also observed among genes encoding interacting enzymes. Examples include significant positive associations between K00125 encoding formate dehydrogenase (EC:1.2.1.2) and K00201 encoding formylmethanofaran dehydrogenase (EC:1.2.99.5). Similar observation is made between genes K00443 and K03388 encoding interacting enzymes, heterodisulfide reductase (EC:1.8.98.1) and coenzyme F420 hydrogenase (EC:1.12.98.1) respectively. However, no mutual exclusion patterns have been found among genes either associated with methane emissions or involved in the methane production pathway.

C. Comparisons with previous studies

In comparison to our previous studies [6], [13] in which a co-abundance network was constructed using Pearson correlation coefficient to measure the similarity between two

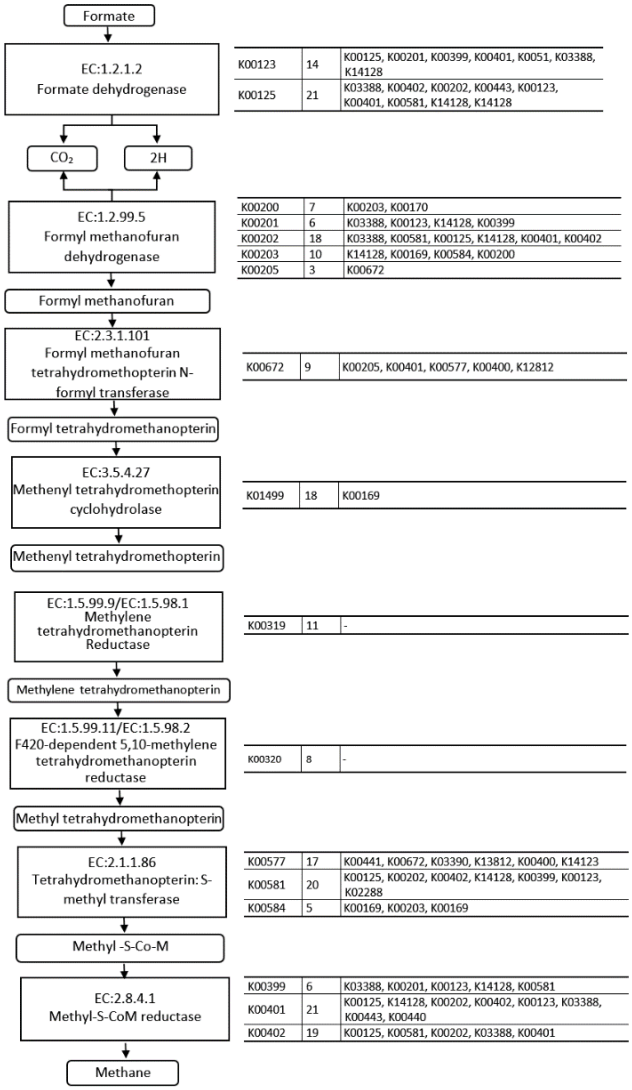

Fig. 6. Interaction partners in the co-presence network for key units involved in methane production pathway [2]. The 3 columns in the table represent the enzyme encoding gene, degree and Methane specific interaction partners respectively.

genes based on their relative abundances, the current study introduces two major improvements: (1) the system is able to construct a network containing either co-presence or mutual-exclusion patterns; and (2) the compositional effect in the analysis of rumen microbial communities based on relative abundance data is mitigated through an ensemble approach [14] containing two correlation measurements and 3 dissimilarity metrics.

It has been shown that assessing relationships between relative abundance profiles purely based on correlation-based metrics may lead to spurious correlation. For example, the actual counts and relative abundances which sum to one of K02986 and K00790 are shown in Fig. 7. Two microbial genes only have a weak negative correlation with Pearson correlation coefficient equal to -0.291 in Fig. 7(a) while they exhibit a strong negative correlation based on their relative abundance (-0.995) in Fig. 7(b). Another example is the correlation between K07636 and K03742

which show a strong positive correlation well above the threshold (0.99) used in our previous study [13] to construct the co-abundance network. However, if we look at their actual abundance profile, they have a correlation of 0.948 which is below the threshold identified (0.99).

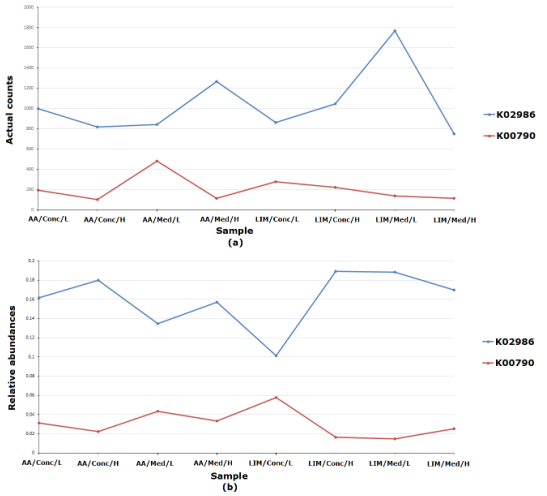

Fig. 7 The abundance profiles of two microbial genes, i.e. K02986 and K00790 across 6 samples: (a) actual counts; and (b) relative abundance

Our results also provide the evidence that analysis of relative abundance profiles purely based on correlation-based metrics may lead to loss of information. For example, out of 31 genes encoding enzyme directly involved in the methane production pathway, 22 were found in Module A which is strongly associated with methane emissions while only 18 were included in the module found in our previous study [13]. Out of 2106 positive interactions included in the co-presence network and 763 negative associations in the mutual-exclusion network, only 775 are found to have an absolute value of the Pearson correlation greater than the threshold identified in [13]. In particular, there is only one pair of microbial genes, i.e. K00790 and K02986, having a negative correlation less than -0.99. On the other hand, the interactions supported by the Pearson correlation measure found in the co-presence network have a positive value higher than 0.995, suggesting that inferring a microbial association network solely based on a correlation measure may not only lead to missing information but also cause artefactual associations. A close examination of the interaction partners of K00123 (formate dehydrogenase major unit [EC:1.2.1.2]) in the co-expression network confirms our analysis. K00123 is found to be associated with methane emissions [6] and involved in the methane production pathway [2]. It has 14 significant positive interactions with a corrected p value less than 0.05. However, more than half of interactions have a Pearson correlation coefficient less than 0.99 and thus were not included in our previous study including the interactions with another subunit of formate dehydrogenase (K00125) and K00201(formylmethanofuran dehydrogenase subunit B [EC:1.2.99.5]).

3.2 Co-occurrence networks derived from relative abundances of 16S rRNA KEGG genes

Similar to the analysis of KEGG genes, the co-presence network constructed based on 16S rRNA genes exhibits a clear modular structure as shown in Fig. 8. It has 56 nodes and 153 significant associations (q value < 0.05) consisting of two main modules, e.g. Modules C and D, both having a clustering coefficient much greater than in case of random networks with the same number of nodes and edges (TABLE III) and a low average path length of less than four in comparison to 6 found in random networks on average [14].

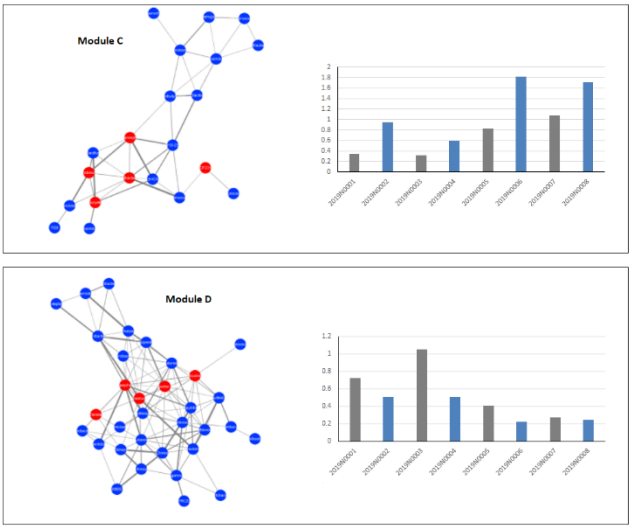

Fig. 8 Significant co-presence relationships among the abundances of 16S rRNA genes in the rumen microbiome. The network, in which node represents genera and edges indicate strong positive associations, consists of two main modules: Modules A and B. The width of edges is proportional to the level of significance of supporting evidence (q value). Red nodes represent genera whose abundances determined by qPCR differed between low and high emitting animals with p < 0.05 listed in [2]. The average abundance of genera across 8 samples were shown in a chart

TABLE IV THE TOPOLOGICAL FEATURES OF MODULES A AND B FOUND IN THE CO-PRESENCE NETWORK DERIVED FROM 16S RRNA DATA. CPL: CHARACTERISTICS PATH LENGTH

Parameters	Module C	Module D
Number of nodes	21	33
Number of edges	40	112
Network diameter	6	5
Network radius	3	3
Network density	0.190	0.212
Clustering coefficient	0.425	0.502
CPL	2.924	2.280
Network centralization	0.232	0.273
Network heterogeneity	0.510	0.578

All the associations in the copresence network are supported by at least 3 metrics with a q value less than 0.05 and by both Pearson and Spearman correlations. Interestingly the thresholds derived are much lower than those determined by KEGG microbial genes as depicted in TABLE V.

TABLE V DISTRIBUTION OF THE NUMBER OF INTERACTIONS SUPPORTED BY 5 METRICS

Metrics	Co-presence network		Mutual-exclusion network	
	Threshold	Number of pairs	Threshold	Number of pairs
Pearson correlation	0.203	153	-0.368	41
Spearman correlation	0.307	153	-0.355	41
Bray-Curtis	0.072	147	0.669	41
Kullback-Leibler	0.050	151	10.252	41
Jensen-Shannon	0.0062	17	0.317	41

In terms of biological relevance, we found that on average genera in Module A were nearly twice abundant in animals with high methane emissions (2019N002, 2019N004, 2019N006, and 2019N008) than in the low emitters (2019N001, 2019N003, 2019N005, and 2019N007) as illustrated in the chart shown in Fig. 8. Examples include archaeal genus *Methanobrevibacter* whose abundance differed significantly between low and high embittering animals with more than 2.54 times abundant in high emitters (p < 0.05) [2] and bacterial genus *Mogibacterium* from the family of Eubacteriaceae, which was 2.17 times abundant in high emitters compared to low emitters (p < 0.05). *Methanobrevibacter* is known to be one such major intestinal genus of the Methanobacteriaceae family that produces methane through the reduction of CO_2 with H_2 [23]. Another example is the genus *Ruminococcus* being more than twice as abundant in samples with high methane emissions. It has been found that species belonging to *Ruminococcus* have higher abundance in the high methane emitter due to excess hydrogen production [24].

On the contrary, a significantly high level of abundance was observed in animals with low methane emissions in Module B (t-test, p < 0.005). Out of 33 genera in Module B, 30 were found to be less abundant in the high emitters. For example, the bacterial genus *Dialister* was more than 4-fold abundant in low emitters (p < 0.05). Among 12 genera differing between low and high emitters with an unadjusted p value less than 0.05 identified in the study conducted by Wallace et al. [2], the genera *Megasphaera*, *Pseudoramibacter_Eubacterium*, *Mitsuokella*, *Roseburia*, and *Dialister* were less abundant in high emitters and were all included Module B.

Turning to the mutual-exclusion network which consists of 26 genera and 41 links (Fig. 9), we found no significant difference was observed in terms of the average abundance between low and high emitters (p = 0.26). Out of 40 genera, only 12 were more abundant in samples belonging to the high methane emission group with a ratio (H/L) ranging from 1.5 to 3.7. it is worth noting that all the associations included in the network were supported by all the 5 metrics with q < 0.05 (TABLE V).

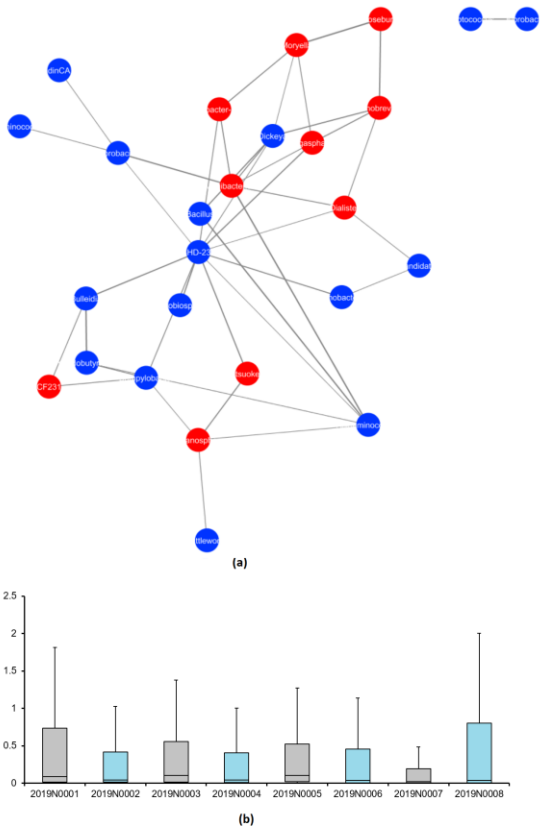

Fig. 9 Significant mutual-exclusion relationships among the abundances of 16S rRNA genes in the rumen microbiome. (a) The network, in which node represents genera and edges indicate strong negative associations, consists of 26 genera and 41 negative associations. The width of edges is proportional to the level of significance of supporting evidence (q value). Red nodes represent genera whose abundances determined by qPCR differed between low and high emitting animals with p < 0.05 listed in [2], and (b) The box and whisker chart represents abundance of genera across 8 samples.

In terms of topological analysis, the major observation different to the analysis of the co-presence network shown in Fig. 8 is all the nodes in the mutual-exclusion network have a clustering coefficient of zero suggesting there is no clustering at all in the network.

There are a total of 10 genera included in both the co-presence and mutual-exclusion networks whose abundance were found to differ significantly between low and high emitters [2]. To assess the topological relevance of these nodes, we computed two well-studied centrality indexes for each node, e.g. degree and betweenness, as listed in TABLE VI. The bacterial genera *Roseburia*, *Megasphaera*, and *Pseudoramibacter-Eubacterium* are ranked as top 3 most connected nodes which have a degree of 15, 14 and 12 respectively. In the mutual-exclusion network, the genus *Mogibacterium* is ranked as the second most connected genus which connects to 6 other genera.

TABLE VI THE DEGREE AND BETWEENNESS FOR 10 GENERA INCLUDED IN BOTH THE CO-PRESENCE AND MUTUALEXCLUSION NETWORKS WHOSE ABUNDANCE DIFFERED SIGNIFICIANTLY BETWEEN LOW AND HIGH EMITTERS. CP: CO-PRESENCE NETWORK; ME: MUTUAL-EXCLUSION NETWORK

Genus	Degree		Betweenness	
	CP	ME	CP	ME
Methanobrevibacter	5	4	0.0526	0.0537
Dialister	5	4	0.0016	0.0990
Mitsuokella	9	2	0.0703	0.0304
CF231	2	2	0.1	0.0026
Methanosphaera	6	4	0.193	0.100
Roseburia	15	2	0.180	0.0025
Moryella	5	4	0.104	0.0456
Mogibacterium	8	6	0.244	0.122
Pseudoramibacter_Eubacterium	12	3	0.061	0.035
Megasphaera	14	4	0.141	0.0666

Betweenness centrality was estimated based on communication flow and it has been suggested that a node with high betweenness may play an important role in controlling the flow of information through a network and maintaining the integrity of a network [25]. As a genus having the highest betweenness in the copresence network and ranked as the top connected node in the mutual-exclusion network, we hypothesized that the bacterial genus *SHD-231* could play an important role in methane production. Indeed, we found that the genus *SHD*-231 from the phylum *Chloroflexi* was about 3.69 times more abundant in samples with high methane emissions (Fig. 10), though not significantly so (t-test, $p = 0.19$). It has also been found that *SHD-231* is one of the most abundant bacterial genera in the study recently published by Cunha et al. [26] which evaluated how the gut microbiota affects both methane emissions and animal production.

Comparing to previous studies in which the co-abundance network was constructed using a random matrix theory(RMT)-based approach for determining correlation threshold, we found that the number of significant associations identified in this study has been dramatically reduced. For example, the correlation network constructed using the RMT approach consists of more than 1600 links with the threshold was estimated to be 0.31 while there are only 41 and 153 associations identified in the copresence and mutual-exclusion networks respectively with a q value less than 0.05. Having a close look at the networks constructed, we found that due to the compositional nature of the data, construction of a network solely based on the relative abundance may lead to misleading results. For example, the absolute abundance profile of *Methanobrevibacter* across 8 samples is different to the one based on the relative abundance as illustrated in. The count of *Methanobrevibacter* in high emitter sample 2019N007is greater than the one observed in 2019N006. However it is not the case when analyzing based on relative abundances. Furthermore, it has been observed that, in some cases, the correlation derived from relative values are substantially different to those estimated based on absolute counts as illustrated in TABLE *VII*. For instance, the bacterial genera *YRC22* and *Prevotella* was negatively correlated based on their relative abundances across 8 samples. However, they exhibit a positive correlation when using absolute counts.

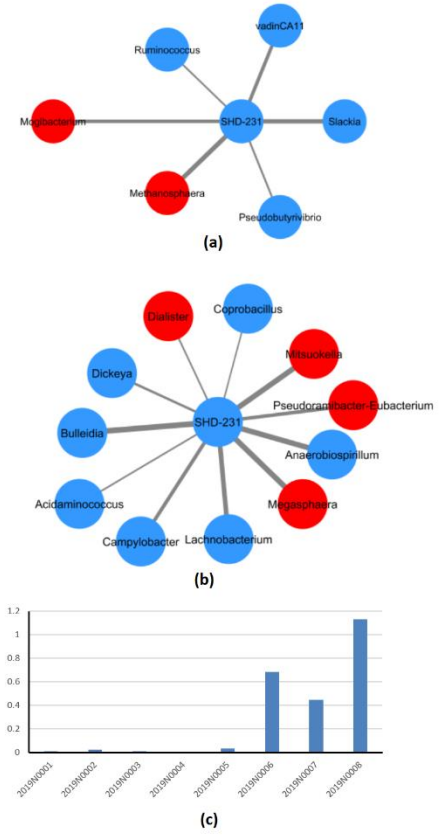

Fig. 10 Strong associations identified for the bacterial genus SHD-231 in (a) the co-presence network; and (b) the mutual-exclusion network. (c) its abundance across 8 samples. The width of edges is proportional to the level of significance of supporting evidence (q value). Red nodes represent genera whose abundances determined by qPCR differed between low and high emitting animals with $p < 0.05$ listed in [2].

TABLE VII TEN EXAMPLES OF GENUS PAIRS WHOSE CORRELATIONS DERIVED FROM RELATIVE ABUNDANCES ARE SUBSTANTIALLY DIFFERENT TO THOSE DERIVED FROM ABSOLUTE COUNTS

Genus		Pearson correlation	
A	B	Absolute counts	Relative abundances
Anaerobiospirillum	TG5	-0.317	-0.718
Lachnospira	Selenomonas	-0.353	-0.757
CF231	Succiniclasticum	-0.400	-0.804
TG5	Coprococcus	-0.319	-0.787
YRC22	Prevotella	0.389	-0.347
Methanobrevibacter	SMB53	-0.843	-0.353
Bacillus	SMB53	-0.710	-0.334
Methanobrevibacter	Desulfovibrio	0.457	0.800
SMB53	Mogibacterium	-0.790	-0.449
Sphaerochaeta	Clostridium	0.536	0.871

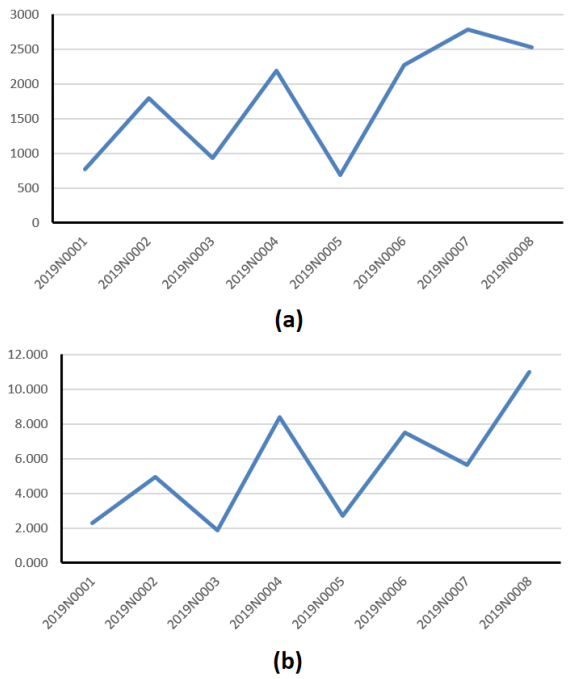

Fig. 11 The abundance profile of *Methanobrevibacter* across 8 samples: (a) absolute values; and (b) relative values.

3.3 Integrative analysis with a multiplex network appraoch

In this section, we introduced a multiplex network-based approach for integrative analysis of co-occurrence networks. Instead of treating co-presence and mutual-exclusion networks as two independent networks, we proposed to use a multiplex network model to bridge together different co-presence and mutual-exclusion relations (Fig. 12).

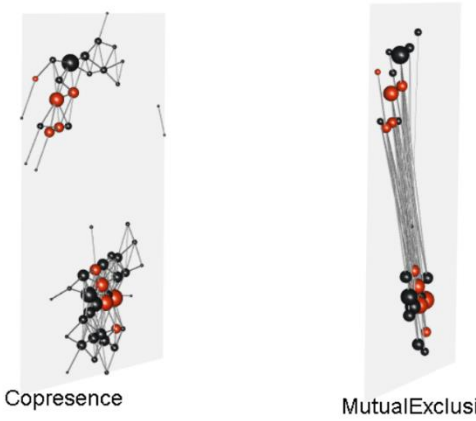

Fig. 12 Visualisation of the copresence and mutual-exclusion networks simultaneously using the MuxViz platform [22]. The size of nodes is proportional to degree. Red nodes represent genera whose abundances determined by qPCR differed between low and high emitting animals with $p < 0.05$ listed in [2]. Nodes were arranged based on the layout of the copresene network.

As expected, there are no links shared by both networks. The overlap between the copresence and mutual-exclusion networks is nearly half (46%). Surprisingly the similarity between two networks is zero in terms of the shortest path distance between the all pairs of genera in both networks.

As a proof of concept, we applied the PageRank and eigenvector centralities developed for interconnected multilayer networks [20], [22] which were based on a random walk on a multilayer network to rank the nodes. Let $T_{j\beta}^{i\alpha}$ denote the tensor of transition probabilities for jumping from node v_i in layer α to node v_j in layer β, and let $p_{i\alpha}(t)$ be the time-dependent tensor representing the probability to find a walker at node v_i in layer α. Thus, the equation governing the discrete-time evolution of the probability $p_{i\alpha}(t)$ can be denoted as

$$p_{j\beta}(t+1) = T_{j\beta}^{i\alpha} * p_{i\alpha}(t) \quad (5)$$

The full description of the calculation of PageRank and eigenvector centralities using tensor formalism can be found in [20]. Without loss of generality, we applied the framework to combine the copresence and mutual-exclusion networks derived from 16S rRNA genes. The top 3 are all bacterial genera as depicted in TABLE VIII. Interestingly, Roseburia, *Pseudoramibacter_Eubacterium*, and *Megasphaera* included in Modue D in the mutul-exclusion network were all lower in high emitters compared to low emitters. The bacterial genus SHD-231 which were over 3.5 times more abundant in animals with high methane emissions was ranked as the top one in terms of PageRank, which might again suggest its role in the methane production pathway.

TABLE VIII THE TOP 3 GENERA RAKED USING IN A MULTIPLEX NETWORK-BASED PAGERANK AND EIGENVECTOR CENTRALITIES

Rank	PageRank in multiplex network	Eigenvector in multiplex network
1	SHD-231	Roseburia
2	Megasphaera	Pseudoramibacter_Eubacterium
3	Roseburia	Megasphaera

4 CONCLUSION

Advances in NGS-based approaches have opened up new avenues in rumen microbial ecology studies. One of the key research areas is to infer association and dependencies between members of rumen microbial communities through correlation analysis. However, it has been found that due to the nature of data generation and the normalization process involved, traditional correlation-based analysis exhibits some significant limitations [10], [13]. Using a compendium of 2 correlation and 3 dissimilarity measures, this paper applied a new framework for the analysis of rumen metagenomics data which include the relative abundance of 1570 KEGG genes and 76 genera. Robust co-presence and mutual exclusion networks were constructed which contains 1000 top-ranking and 1000 bottom-ranking edges with an FDR corrected value less than

0.05. Biological relevance of derived co-occurrence networks were assessed in terms of both pathway analysis and the level of enrichment of trait specific genes. It has been found that Modules A and B in the co-presence and mutual exclusion networks constructed from KEGG genes and Module C in the co-presence network derived from 16S rRNA are strongly related to methane emissions. Based on the assessment of level of enrichment of trait-specific microbial genes, co-presence and mutual-exclusion modules associated with methane production, i.e. Modules A and B, were identified. While there exist strong positive correlations between methane specific genes, no mutual-exclusion patterns were observed among genes associated with methane emissions and encoding enzymes included in the methane production pathway. The results demonstrate that deriving microbial associations based on the correlations between relative abundances may not only lead to loss of information but also produce spurious associations.

In addition, a multiplex network model was proposed for integrative analysis of co-occurrence networks in an attempt to bridge together different co-presence and mutual-exclusion relations. By facilitating the crosstalk and interactions between co-occurrence networks, the proposed framework has great potential to support the prediction of complex animal phenotypes, such as methane production and feed efficiency, which are dependent on the rumen fermentation. This will also provide additional insights into biological mechanisms of microbiome associated with the traits.

In this study, we adopted the parameters used [13] and the network construction was based on the analysis of the 2000 edges with extreme scores, i.e. 1000 top-scores representing strong positive interactions and 1000 bottom scoring associated with negative association. A potential direction for our future research is to develop an advanced approach for the automatic determination of the optimal number of edges to be included for the inference of microbial association networks.

The current study was based on the analysis of rumen samples from 8 extreme animals balanced for breed type and diet. Applying the framework to the analysis of a large cohort of samples would be an important part of our future work.

ACKNOWLEDGMENT

This work was supported in part by the MetaPlat project, (www.metaplat.eu) funded by H2020-MSCA-RISE-2015. HY. Wang (hy.wang@ulster.ac.uk) is the corresponding author.

REFERENCES

[1] J. Broucek, "Production of Methane Emissions from Ruminant Husbandry: A Review," Journal of Environmental Protection, 2014, 5, pp. 1482-1493.

[2] 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

[3] GreenGas House online, 2017. [Online]. Avaialble: http://www.ghgonline.org/methaneruminants.htm, [accessed: 04-07-2017].

[4] A. Kumar, P. Rameshwar, S. Devki, and N Kamra, Rumen Microbiology: From Evolution to Revolution, Springer, July 11, 2015

[5] E. Jami, B.A. White, and I. Mizrahi I, "Potential Role of the Bovine Rumen Microbiome in Modulating Milk Composition and Feed Efficiency," PLoS ONE 9(1): e85423. https://doi.org/10.1371/journal.pone.0085423

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

[8] M. Schären, J. Frahm, S. Kersten, U. Meyer, J. Hummel, G. Breves, S. Dänicke, "Interrelations between the rumen microbiota and production, behavioral, rumen-fermentation, metabolic, and immunological attributes of dairy cows," Journal of Dairy Science, 2018, in press.

[9] 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).

[10] M.B. Lengowski, K.H.R. Zuber, M. Witzig, J. Möhring, J. Boguhn, and 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. https://doi.org/10.1371/journal.pone.0150115

[11] J. Friedman J and E. J.Alm EJ, "Inferring Correlation Networks from Genomic Survey Data," PLoS Comput Biol, 2012, 8(9): e1002687.

[12] R.J. Williams, A. Howe, and K. S. Hofmockel, "Demonstrating microbial co-occurrence pattern analyses within and between ecosystems," Front. Microbiol. 2014, 5:358. doi: 10.3389/fmicb.2014.00358

[13] Wang et al, "Integrated metagenomic analysis of the rumen microbiome of cattle reveals key biological mechanisms associated with methane traits," Methods, 2017. https://doi.org/10.1016/j.ymeth.2017.05.029.

[14] K. Faust, J.F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, et al. "Microbial Co-occurrence Relationships in the Human Microbiome," PLoS Comput Biol., 2012, 8(7): e1002606.

[15] HY.Wang, H. Zheng, R. J. Dewhurst, and R. Roehe, "Microbial Co-presence and Mutual-exclusion Networks in the Bovine Rumen Microbiome," in the Proc. Of 2017 IEEE International Conferene on Bioinformatics and Biomedicine (BIBM), pp.114-119, 2017

[16] M.B. Brown, "A Method for Combining Non-Independent, One-Sided Tests of Significance", Biometrics, 1975, 31, pp. 987-992

[17] K. Faust and J.Raes, "CoNet app: inference of biological association networks using Cytoscape," F1000Research, 2016, 5:1519 (doi: 10.12688/f1000research.9050.2)

[18] 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), pp.2498-504

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

[20] 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.

[21] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas, "Ranking in interconnected multilayer networks reveals versatile nodes," Nat. Commun., 2015, 6:6868 doi: 10.1038/ncomms7868

[22] M. De Domenico, M. Porter and A. Arenas, "MuxViz: a tool for multilayer analysis and visualization of networks," Journal of Complex Networks, 2015, 3(2), pp.159-17

[23] A.S. Dighe, K. Jangid, J. M. González, V.J. Pidiyar, M. S. Patole, D. R Ranade and Y.S. Shouche, "Comparison of 16S rRNA gene sequences of genus Methanobrevibacter", BMC Microbiol. 20044:20.

[24] S. Kittelmann, C. S. Pinares-Patiño, H. Seedorf, M. R. Kirk, S. Ganesh, J. C. McEwan, P. H. Janssen, “Two Different Bacterial Community Types Are Linked with the Low-Methane Emission Trait in Sheep,” PLoS ONE 9(7): e103171

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

[26] C.S. Cunha et al. “Assessing the impact of rumen microbial communities on methane emissions and production traits in Holstein cows in a tropical climate,” Syst Appl Microbiol. 2017, 40(8), pp.492-499.

Huiru Zheng (M’03, SM’16) received the Ph.D. degree on data mining and bioinformatics from Ulster University, UK, in 2003. Her research area lies on the broad area of data mining and artificial intelligence and their applications on systems biology, and healthcare. She has published over 230 research papers in peer reviewed international journals and conferences. Prof. Zheng is currently a Professor of Computer Sciene with the School of Computing at Ulster University.

Haiying Wang received the Ph.D. degree on artificial intelligence in biomedicine in 2004 and he is currently a Reader in the School of Computing at Ulster University, UK. His research area includes artificial intelligence, complex network analysis, computational biology and bioinformat-ics. He has a particular research interest and expertise in network-based approaches to the field of systems biology and meta-genomics. Since 2004, he has published more than 130 peer-reviewed research papers in international journals and conference proceedings.

Richard Dewhurst obtained a Ph.D. degree in ruminant nutrition in 1989. He has headed leading ruminant research units in the UK, New Zealand and Ireland and produced over 110 refereed papers. He is currently Professor of Ruminant Nutrition & Production Systems and Head of Future Farming Systems Group at SRUC.

Rainer Roehe obtained a Ph.D. degree in animal breeding and genomics in 1990. He is currently Professor of Animal Genetics and Microbiome at SRUC, UK. He has published over 120 referred papers, including the publication receiving the PLOS Genetics Research Prize 2017. His main research interests are in the understanding of the functional and genomic architecture of the gut microbiome and its interactions with the host genome.