Analysis of Rumen Microbial Community in Cattle through the Integration of Metagenomic and Network-based Approaches

Haiying Wang[1], Huiru Zheng[1*], Fiona Browne[1], Rainer Roehe[2], Richard J. Dewhurst[2], Felix Engel[3], Matthias Hemmje[3], 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
*h.zheng@ulster.ac.uk

Abstract— **A better understanding of the composition of rumen microbial communities and the association between host genetic and microbial activities has important applications and implication in bioscience. Being capable of revealing the full extent of microbial gene diversity, metagenomics-based approaches hold great promises in this endeavor. 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 of this study is to develop a random matrix theory-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 of the three main modules well separated. The involvement of KEGG pathways in each module was analysed. A close look at the abundance profiles highlights that Module B is strongly associated with methane emissions while Module C is highly enriched with microbial genes associated with feed conversion efficiency.**

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

I. 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. It harbors a highly diverse microbial community predominantly consisting of bacteria, archaea, protozoa and fungi. These rumen microbes living in a symbiotic manner break down ingested feed constituents to produce primarily volatile fatty acids and bacterial protein that are major nutrient sources for the host animal and used in its energy metabolism and protein synthesis [2], [3].

While being capable of harvesting energy from otherwise indigestible food components, 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 effect and global warming [1], [4], [5]. 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 [5], [6].

Early exploration of rumen microbiology was mainly dominated by culture-based approaches such as isolation, enumeration and nutritional characterization. Pioneering studies include the description of well characterized rumen bacteria based on the isolation of the functionally significant bacterial groups [8], [9]. While successfully identifying more than 200 microbial species including bacteria and protozoa from the rumen [1], [7], culture-dependent techniques requiring a careful design of protocol for growth of organism exhibit several significant limitations [10]. They are not only time consuming and cumbersome [7] 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 the setting of culture conditions [10], [11].

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 further shaping microbiology [12], [13]. Recent years have seen a growing trend toward using metagenomics based approaches for the analysis of the composition of rumen microbial community. Based on 742 samples from 32 animal species and 35 countries, Henderson et al. [6] 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. [4] brought new insights on the interactions between microbial populations and the association with the host. To characterize biomass-degrading genes and genomes, Hess et al. [14] analyzed 268 gigabases of metagenomic DNA from microbes adherent to plant fiber incubated in cow rumen. A total of 27,755 candidate genes with a significant match to at least one relevant catalytic domain or carbohydrate-binding module were identified, greatly expanding our knowledge of genes and genomes participating in the degradation of cellulosic biomass.

More recently, based on the relative abundance of 1570 microbial genes identified in a metagenomics analysis, Roehe and his colleagues [5] 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.

This study aims to study the rumen microbial community in cattle through the integration of metagenomic and network-based approaches. One of the main objectives is to develop an automatic computational technique to objectively determine the correlation threshold used to construct the co-abundance network. In this paper 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.

II. Methodology

A. Datasets under study

The abundance dataset used was released by the recent studies conducted at the Beef and Sheep Research Centre of Scotland's Rural College [3], in which 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)). Methane emissions of individual animals were measured in respiration chambers. 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). 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. 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%. The detailed description of data generation can be found in [3].

B. RMT-based appraoches

Random matrix theory (RMT)-based approach to an objective determination of the threshold used to construct the co-abundance network 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 [16], [18]. 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 described by the Wigner surmise [17].

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

- For a non-random matrix in which no correlation between nearest-neighbor eigenvalues is observed, the NNSD tends to follow the Poisson distribution given below, indicating the system represented by the matrix can be separated into several relatively independent clusters in which members exhibit similar behaviours and properties [18], [19].

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

It has been highlighted that the transition of NNSD between GOE and Poisson statistics can potentially serve as a reference point to automatically construct a condition-specific correlation network by removing random noise in an objective manner [16].

C. Construction of co-abundance networks

Based on the recent study 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, similar to the study [16], 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. 1.

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.

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

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.

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

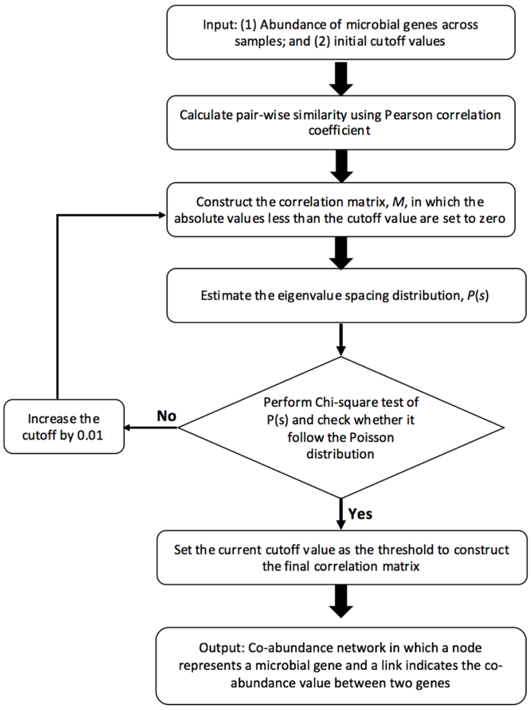
Input: (1) Abundance of microbial genes across samples; and (2) initial cutoff values
Calculate pair-wise similarity using Pearson correlation coefficient
Construct the correlation matrix, M, in which the absolute values less than the cutoff value are set to zero
Estimate the eigenvalue spacing distribution, P(s)
Perform Chi-square test of P(s) and check whether it follow the Poisson distribution
No
Increase the cutoff by 0.01
Yes
Set the current cutoff value as the threshold to construct the final correlation matrix
Output: Co-abundance network in which a node represents a microbial gene and a link indicates the co-abundance value between two genes

Fig. 1 A diagram to illustrate the key steps to construct the co-abundance network.

D. *Evaluation metrics and software packages used*

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 $\chi^2(df, \alpha)$.

The estimation of the distribution of unfolded eigenvalue spacing was implemented using the pipeline of Molecular Ecological Network Analysis [20]. The NNSD was plotted using the R package RMThreshold (https://cran.r-project.org/web/packages/RMThreshold/index.html).

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

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \qquad (5)$$

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.

The computation of topological parameters was with the NetworkAnalyzer [21] and CentiScaPe [22] plugins. 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 [23] respectively.

III. Results and Discussion

A. *The impact of the threshold*

As shown in Fig. 2, 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. 2(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 increases, the clear transition of the NNSD from GOE to Poisson was observed (Fig. 2(b) to Fig. 2(d)). As depicted in Fig. 2(c), the NNSD began to deviate from GOE at the threshold of 0.95. It appears to closely follows the Poisson distribution when the threshold set to 0.99 (Fig. 2(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 ($\chi^2 = 82.535, p = 0.028$). 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.

B. *Co-abundance network*

The network analysis of microbial gene abundance was illustrated in Fig. 3, in which each node stands for a microbial gene and the strength of each edge denotes the correlation in their abundance. Only the correlation across 8 samples greater than 0.99 was 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 I, each having a clustering coefficient significantly greater than a random graph constructed on the same number of nodes.

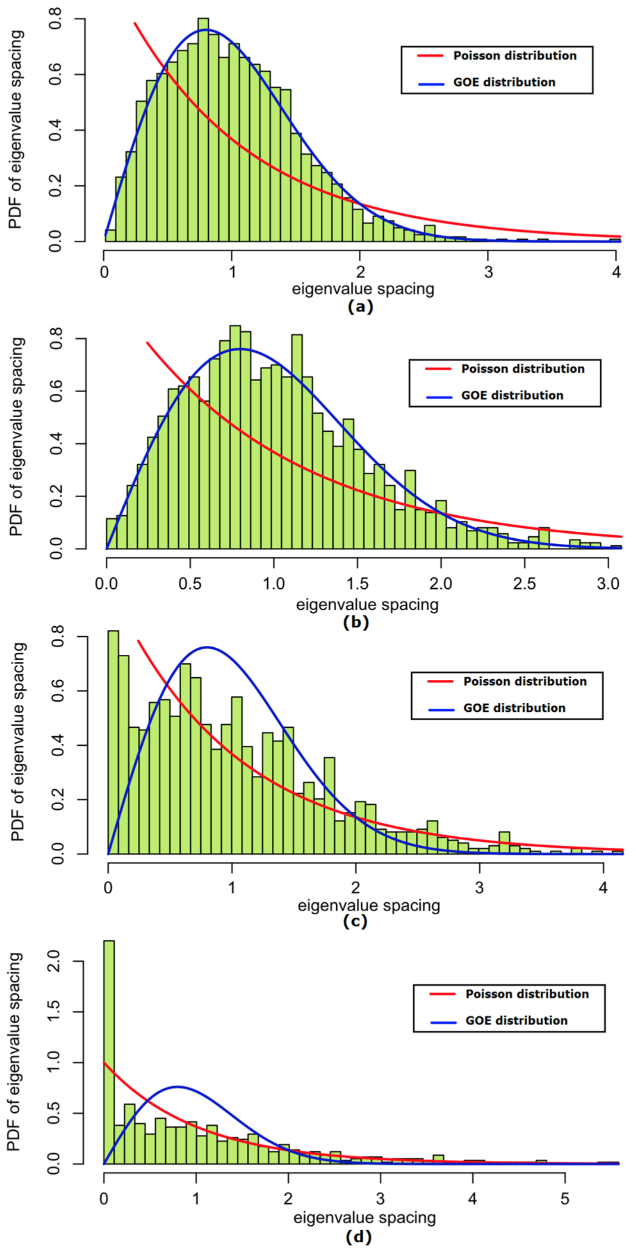

Fig. 2 The NNSD of the correlation matrix constructed from the abundance of 1570 microbial genes across 8 samples with different thresholds: (a) threshold = 0.0; (b) threshold = 0.90; (c) threshold = 0.95; and (d) threshold = 0.99.

Table I 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

C. *Biological relevance*

We first checked 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). A close look at the abundance profile of genes across 8 samples as depicted in Fig. 3 suggests that Module B be heavily linked to methane emissions. The high level of abundance was observed in the samples in the high methane emission group (2019N002, 2019N004, 2019N006, and 2019N008). Thirty out of 91 genes in Module B are involved in methane metabolism pathway.

An analysis with regard to the distribution of microbial genes strongly associated with traits such as methane emissions and feed conversion efficiency indicates that certain trait-specific genes are highly over-represented in modules. For example, all the 20 genes identified to be associated with methane emissions by Roehe et al. [3] represented by red triangle nodes in Fig. 3 are found in Module B (hypergeometric test, $p < 10^{-11}$). A total of 15 genes linked to feed conversion efficiency (green diamond nodes in Fig.3) are found in the network, 9 of which are assigned to Module C (hypergeometric test, $p < 10^{-6}$). It is worth noting that the extreme animals selected in the data collection were based on methane emission, therefore the power to detect microbial genes associated with methane are substantially higher than those associated with feed efficiency.

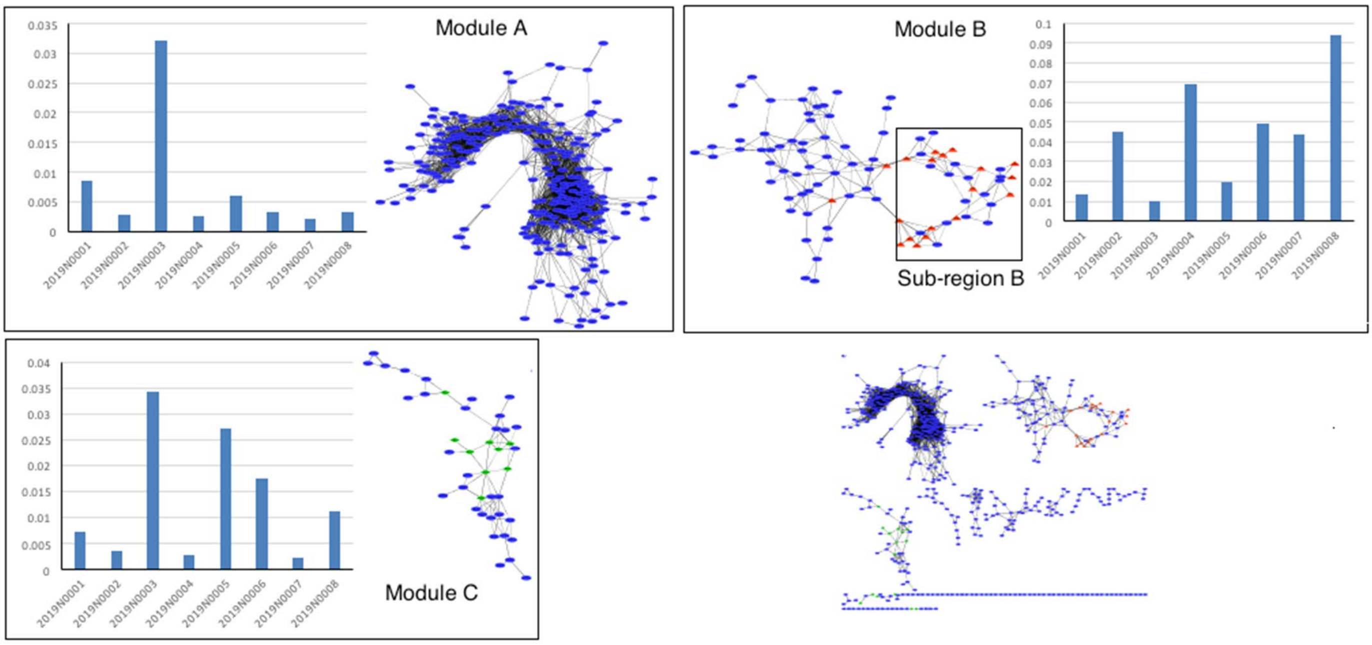

Fig. 3 Network-based approach to the correlation analysis of microbial gene abundance. The threshold used to construct the co-abundance network was set to 0.99. The network, in which each node represents a microbial gene and each edge indicates the correlation in their abundance, exhibits a clear modular 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.

The further examination of abundance profiles of 35 genes grouped in the sub-region in Fig. 3 support the above observation, which illustrates the relative. It has been shown that all the genes have a low level of abundance in the samples assigned to the low methane emission group. Unexpectedly, among these 35 genes, K00400 involved in Methane metabolism pathway (ko00680) is ranked at the top in terms of all 5 centrality metrics used (degree: 5; closeness: 0.0029; betweenness: 2005.81; eigenVector: 0.050; and bridging centrality: 160.36).

IV. Conclusions

Being capable of revealing the full extent of microbial diversity, recent years have seen a growing trend toward metagenomics-based approaches to study the composition of rumen microbial community and the association between host genetic and microbial activities. This study investigated 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 contribution in this study 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 [26], [27].

This research has been undertaken as part of 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 now working on OAIS (Reference Model for an Open Archival Information Systems, ([28], ISO 14721) to build a framework of terms and concepts to specify an archival information system, aiming to build a platform-neutral instrument to preserve and share digital objects in the long term. 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.

Acknowledgment

This work was supported in part by the MetaPlat project, (www.metaplat.eu) funded by H2020-MSCA-RISE-2015 and COST Action Open Multiscale Systems Medicine (OpenMultiMed) supported by COST (European Cooperation in Science and Technology).

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), pp.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.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.

[5] 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.

[6] 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).

[7] 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), pp. 341-351

[8] 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

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

[10] 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

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

[12] 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

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

[14] M. Hess, et al., “Metagenomic discovery of biomass-degrading genes and genomes from cow rumen”, *Science*, vol. 331, 2011, pp. 463-467.

[15] M.L. Mehta, Random Matrices, 2nd edition. Academic Press: 1990.

[16] 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.

[17] E. Wigner, “On the distribution of roots of certain symmetric matrices”, Ann. Math., 1958, vol.67, No. 2, pp.325-327.

[18] 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.

[19] 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), pp.660–668.

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

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

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

[23] 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

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

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

[27] 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, in press.

[28] 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.