An Integrative Framework for Functional Analysis of Cattle Rumen Microbiomes

Jyotsna Talreja Wassan[1], Huiru Zheng[1], Fiona Browne [1], Jenna Bowen[3], Paul Walsh[2], Rainer Roehe [3], Richard Dewhurst[3], Cintia Palu[2], Brian Kelly[2], and , Haiying Wang[1]

[1] School of Computing, Ulster University, Co. Antrim, Northern Ireland, UK
hy.wang@ulster.ac.uk
[2] NSilico Life Science Ltd., Dublin, Ireland

[3] Future Farming Systems, Scotland's Rural College, Edinburgh, Scotland, UK

Abstract— **Metagenomics is the study of environmental microbial communities and has various applications and implications in biological research. This paper aims to study the role of microbial communities in cattle rumen and their relation to probiotic diet supplement usage as part of the EU H2020 MetaPlat project[1]. In this research, we proposed and evaluated a computational framework to classify 16S rRNA samples from *Bos taurus* (cattle) rumen microbiome into a diet phenotype. We performed analysis by benchmarking various phylogeny-driven methods based on integration of biological domain knowledge of relationships and non-phylogenetic methods based on the raw abundances. The integrative approach incorporating phylogenetic tree structure into machine learning (ML) modelling achieved a high predictive performance with *Accuracy* of 0.925 and *Kappa* of 0.900 for classifying cattle microbiomes into diets supplemented with oil, nitrate, a combination and controls.**

Keywords— Metagenomics, Phylogeny, Machine Learning, Classification, Ordination, Cattle Microbiomes

I. INTRODUCTION (*HEADING 1*)

Metagenomics [1] involves the study of genome sequences of microorganisms existing in an ecological niche. Categorization of the microbial genomes into their functional roles (i.e. phenotypes) forms a significant machine learning (ML) problem of supervised classification in metagenomics [2]. In this current study, we performed the classification of cattle rumen microbiomes into different diet supplements. The combinations of different dietary supplement strategies are expected to reduce the methane emissions (CH_4) in livestock systems to further improve the cattle productivity. Previous work [3, 4, 5], has highlighted the potential of controlled feeding of cattle with nitrate or oil treated diet and its effects on cattle rumen metabolism. Our aim is to perform data analysis to identify a predictive model for differentiating cattle microbiomes into four categories of diets: - oil-based, nitrate-based, combined diet (Oil-Nitrate) and controls. The objective was addressed through a ML-based experimental framework.

In this metagenomics use case, we exploited the integration of phylogenetic tree structure connecting the microbial group of microbial taxas/species (also known as Operational Taxonomic Units (OTUs) [6] or analogous Amplicon Sequence Variants (ASVs)) [6], and their neighbourhood, as naturally defined by common ancestral history in addition to their abundance count values. Recent advances in improving methods of grouping species [6], have devised a novel taxa picking method to create "amplicon sequence variants" or ASVs analogous to OTUs [6]. ASVs are advantageous outcome of metagenomic pipeline as they are obtained at finer resolution level of single-nucleotide differences and independently from a reference database, unlike OTUs which are sequenced over a gene region at a similarity threshold [6]. We implemented a computational framework to assess the predictive power of ASVs and their ancestors from phylogeny for the classification of cattle microbial samples. This resulted in an integrative ML pipeline to investigate the presence of related species (i.e. OTUs/taxas/ASVs) and their role in determining phenotype unlike studies which considered the microbial species as independent features [7-10]. We found that the incorporation of the phylogenetic tree structure into analysis has potential to increase the prediction power. More accurate predictions were obtained with the phylogeny-aware pipeline as proposed in [11], when benchmarked with other non-phylogenetic [7-9] and phylogenetic-based approaches [12-14]. Some high-level visualizations and their applicability to our data are also presented. The findings of the paper would aid readers in analysing the structure and function of metagenomes in cattle rumen effectively. The paper is organized as follows. Section 2 highlights the related work. Materials and methods are discussed in Section 3. Experimental results and discussions are enlisted in Section 4. Section 5 provides the conclusion and future research directions.

II. RELATED WORK

The studies by Knights et al. [7] and Statnikov et al. [8], highlighted the commonly used supervised classification methods such as Random Forest (RF), Logistic Regression (LR) and Support Vector machines (SVMs) for determining functions in metagenomic studies. RF has been established as a gold-standard method for classifying the metagenomes [7, 8]. Wassan et al. [9], suggested regularized-LR (reg-LR) and extreme gradient Boosting (XgBoost) ensemble with RF as a potential method in metagenomic data. Pasolli et al. [15] suggested feature engineering by RF is effective for functional meta-analysis of OTU abundances. Here, we also review and summarize the ML-based approaches based on phylogenetic measures from literature [12-14,16]. Phylogeny is typically represented as the evolutionary distances embarked on a phylogenetic tree consisting of nodes representing the common hierarchal ancestral [17]. Phylogeny proved potentially useful for functional analysis in our current study, since phylogenetically close microorganisms would likely to share the metabolism. Magee et al. [18], highlighted the increasing availability of

[1] MetaPlat, *http://www.metaplat.eu*

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

phylogenetic data to inform the research in biological domains. Pertaining to metagenomics, phylogeny has contributed to the following tools. MetaPhyl [12], is based on a hierarchical grouping of coefficients of LR model by regularizing it with tree penalty function, as was formulated by Kim et al. [13]. Phylogenetic Isometric Log-Ratio Transform (PhILR) [14] is primarily focused on the compositional nature of microbiomes. This tool used reference weights derived from a phylogenetic tree to transform the microbial feature space to an unconstrained Euclidean space, overcoming the challenges associated with the compositional nature of abundance count data. The approaches [12, 14], were applied to the data obtained from the Human Microbiome Project (HMP) [19]. Chen et al. [16], integrated relative taxon abundance with phylogenetic distances obtained from a phylogenetic tree to attain a score vector (i.e. taxon-proportion) for microbial analysis and applied adaptive sum of powered tests (SPU) on the obtained microbiome vector (i.e. aMiSPU) for microbiome studies [16]. The related permutation scheme is based on LR to calculate the *p*-values for the analysis [16].

Recently, we proposed an approach involving integration of phylogenetic weights from the tree and abundance counts of leaves via formulation of phylogeny and abundance aware matrix (i.e. PAAM) [11]. The matrix consists of each node of the tree (i.e. leaves as well as ancestral nodes) as the microbial features and was used as a pre-processed input to ML models. The entries in the matrix were computed by combining phylogenetic distances (PD) and abundances of constituting nodes (OTUs/ASVs). The abundance of each leaf node was weighted by the phylogenetic weights on the branches (i.e. PD), to span them to the ancestral nodes at each level of the tree forming a hierarchal topology (Eq. (1)) [11].

Assuming *m* leaf nodes and that A_{ms} is the abundance count of m^{th} leaf node in a sample *s*; there exists $n = m\text{-}1$ ancestral nodes. Considering PD_{nm} as the phylogenetic distance of m^{th} leaf node from the n^{th} ancestral node (derived from tree) and Y_{nm} as a binary variable to represent whether m^{th} leaf is descendent of n^{th} ancestral node (i.e. $Y_{nm}= 1$ if m^{th} leaf is descendent of n^{th} ancestral node and $Y_{nm} = 0$ if m^{th} leaf is not a descendent of the n^{th} ancestral node); weighted abundance of each ancestral node *n* in a sample *s* (WA_{ns}), was calculated by Eq. (1), where *i* ranges from 1 to the number of leaf nodes [11].

$$WA_{ns} = \sum_{i=1}^{m} \frac{A_{is}\ Y_{ni}}{PD_{ni}} \tag{1}$$

In the current study, the comprehensive evaluation of different ML methods of RF, SVMs, reg-LR, PhILR, MetaPhyl, and PAAM-based ML was conducted for classifying cattle rumen ASVs into functional phenotypes of diet. The findings indicate that the method based on phylogeny-aware matrix and feature engineering with RF importance and reg-LR provides a drive for a comparative very good classification performance.

III. Materials And methods

In this section, we provide a brief description of the dataset used in the current study and proposed methodology to effectively analyse the related metagenomes.

A. Dataset under Study

The current study involves analysis over the use case dataset of 16S rRNA sequences obtained form *Bos taurus* rumen samples playing an important role in cattle productivity, health, and immunity. The data was collected by *Future Farming Systems, Scotland's Rural College, Edinburgh*,U.K.(*https://www.sruc.ac.uk/info/120060/future_farming_systems*), Edinburgh, U.K. to investigate *Bos taurus* rumen microbiota in the context of an environmental trait of supplemented diet. The community composition was determined in 80 case samples provided by the MetaPlat project[1]. The dataset consisted of two breeds of Aberdeen Angus or Limousin sired steers; and four dietary treatments of Control (443 g concentrate and 25 g lipid / kg diet DM); Nitrate (18 g nitrate / kg DM); Oil/Lipid (maize distiller's dark grains, 37 g lipid / kg diet DM) and Combined (18 g nitrate and 37 g lipid / kg dietary DM)]. The data consisted of 20 samples from each of the dietary treatments. ASV table (analogues to OTU table), consisted of raw abundance count of species. The related phylogenetic tree, was obtained by *NSilico Life Science Ltd.* (*http://* www.nsilico.com/), using the QIIME2 pipeline (*https://qiime2.org/*) [20]. The samples were associated with meta-data describing their relationship with environmental factor of diet. The data consisted of 727 microbial organisms at species level of study.

B. The Proposed Framework

1. The Schematic Workflow

The schematic workflow of our computational framework for studying and classifying cattle metagenomes is described below.

a. Inputs. ASV table and the phylogenetic tree served as two inputs to the proposed approach. The ASV table consisted of rows representing the microbial samples and columns representing ASVs; and the phylogenetic tree structure represented hierarchal relationships of ASVs.

b. ML Modelling. To maximize the performance of our experimental design, the integrated workflow focussed majorly on two aspects:- (1) analysis based on phylogenetic tree and abundance count data [11,12,14] (2) analysis based on only raw abundances [21-24] by evaluating its performance using the standard measures. The ML methods were trained with leave-one-out cross validation (LOOCV) [25]. The construction of the proposed framework can be divided into following distinct experimental pathways (as also shown in Fig.1.)

i. RF, reg-LR, SVMs, XgBoost models were applied over the raw-abundances of ASVs obtained by MetaPlat[1] into functional phenotype of diet [21-24].

ii. The feature engineering technique of ranking taxonomic features using RF importance (RFI) was applied over the ASVs [27-28]. The top 10, 30, 50 and 70 % features were derived with RFI modelling.

iii. RF and reg-LR were applied over the PhILR [14] transformed data. The input to the PhILR was filtered based on taxa not seen with more than 3 counts in at least 20% of samples [14]. Subsequently, those with a coefficient of variation ≤ 3 were filtered [14]. Thereafter, data were normalized by adding a pseudo count of 1 to the remaining taxas to avoid calculating log-ratios involving zeros [14].

iv. The phylogenetic method of MetaPhyl was applied along the experimental pathway [12].

v. The phylogeny-aware classification was also conducted on the pre-processed input of phylogeny and abundance aware matrix (i.e. PAAM) [11], derived by the tree structure connecting the ASVs (based on phylogenetic weights) and their abundance count values according to Eq.1. We extended the matrix-based approach by applying ML with RFI [27], reg-LR and XgBoost [24].

vi. The basic statistical association of cattle micro biome with phenotypes was operatively implemented with the aMiSPU [16].

vii. Diversity within the microbial community between different phenotypic functions was visualized by comparing microbial composition in one environment to another .

The two predominant measures for performing diversity analysis are : - i) alpha diversity which measures the number of different microbial species present in a microbial sample, ii) beta diversity which measures species composition and abundance between different samples [29].

An abundance-based estimator of chao1 estimator for species richness and shannon diversity for each sample were used for visualizing the alpha diversity [29]. UniFrac distances represent the fraction of branch length shared between two taxas placed on a phylogenetic tree. Principal Coordinate Analysis over UniFrac distances was used as an informative tool for representing the beta diversity [30,31].

c. Performance Evaluation. The Accuracy [32] and Kappa [33] performance assessment metrics were used for evaluating the classification models in our study. The accuracy is defined as the fraction of correctly classified samples (Eq. (2)).

$$Accuracy = \frac{Number\ of\ Corrrect\ Classifications}{\text{Total Number of Classifications}} \quad (2)$$

Kappa is used to evaluate the agreement between two classifications on ordinal or nominal scales (Eq. (3)). [33].

$$Kappa = \frac{\mathrm{P(a) - P(e)}}{\mathrm{1 - P(e)}} \quad (3)$$

, where $P(a)$ represents the actual observed agreement between the input/s and the outcome of interest, and $P(e)$ represents the chance agreement.

2. Description of ML Methods Used in the Workflow

The following ML methods facilitated the predictive modelling over cattle metagenomes. The aim of this study is to identify ML model which provides good predictive performance.

a. RF. The RF model employs a collection of decision trees to classify metagenomic samples [21]. The model outputs the predictive class majority voting amongst the constituent individual trees [21]. Several studies have applied RF for the prediction of metagenomic functions [7-9].

b. Reg-LR. The LR method tries to fit a generalized logistic model (Eq. (4)) for classifying metagenomes [23].

$$m = \text{natural log}(p/(1\text{-}p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n \quad (4)$$

,where natural log of the probability an event occurring is predicted as a linear function of the regression coefficients β_i's and the input features X_j's. The regularization applied to LR penalizes highly weighted coefficients in Eq. (4) to optimize the cost of the model and preventing model to pick the noisy values.

c. SVMs. This method aims to calculate the optimal hyper-plane by using kernel functions to separate the metagenomic classes. The separating hyper-plane tends to maximize the margin between separable classes [24].

d. XgBoost. XgBoost is a scalable ensemble method of decision tree boosting, supporting the fine tuning and regularization to learn the decisions iteratively [22]. The method continuously tries to improve its prediction in subsequent tree iterations to improve the classification performance.

e. RFI. RF consists of a number of decision trees to create a forest. The role each feature plays in decreasing the weighted impurity in a tree is noted in the approach [27-28]. For the whole forest, the impurity decrease over each feature is averaged to rank the features globally [27-28].

f. The PhILR [14], transform was used to map high-dimensional compositional metagenomic data to a phylogeny-driven Euclidean space.

3. Software Packages and Tools

The various software packages related to ML models used in the current study are enlisted below.

a. ML models of RF, reg-LR, SVM, XgBoost were implemented with the help of caret package in R [26]. The source code for ML models is available at https://github.com/topepo/caret/tree/master/models/ files.

b. PhILR [14], is available as an R package available at https://bioconductor.org/packages/release/bioc/html/philr.html.
c. RFI [27], was implemented with the help of *importance()* function available in R package of randomForest.
d. MetaPhyl [12], was compiled and run as a C++ library available at http://alumni.cs.ucr.edu/~tanaseio/metaphyl.htm.
e. Cattle microbiomes' distribution was studied by alpha diversity, beta diversity, canonical and redundancy analysis with the help of *plot_ordination()* function available in phyloseq and ggplot2 packages in R[29].
f. aMiSPU [16] was implemented as a R package.

IV. Experiments And Results

Predictive modelling provides a holistic understanding for functional metagenomics analysis. The objective of this study is to identify efficient ML models for classifying cattle metagenomes.

The statistical pipeline of aMiSPU [16], was applied over the cattle rumen microbiomes, assuming the null hypothesis of "there exists no correlation between ASVs and diet as phenotype". The advantages of using aMiSPU are: - i) it is independent of any parametric assumptions on the distribution of functional microbiome data as aMiSPU is based on adaptive sum of powered score (aSPU) tests with phylogenetic measure [16], ii) does not suffer from this curse of the compositionality problem in OTUs or ASVs.

aMiSPU reported significant relationship of cattle microbiomes (p < 0.01) with diet by rejecting the null hypothesis. Therefore, we performed ML modelling over the cattle rumen microbiomes to associate it and classify into *diet* phenotype.

We further evaluated the proposed experimental framework (shown in Fig.1.) by following a systematic comparison of ML models benchmarked on phylogenetic and non-phylogenetic approaches over the cattle ASVs. The results of the conducted experiments are summarized in Table 1 and 3.

Table 1 reports the ML models providing performance improvement over the state-of-the-art [7, 8]. The application of reg-LR (Accuracy:0.875) and XgBoost (Accuracy:0.850) models, substantially depicted higher predictive performance in comparison to other state-of-art conventional ML classifiers of RF (Accuracy:0.787) [7] and SVMs (Accuracy:0.675) [8], when applied over the raw abundances. It was also observed that feature engineering with RFI further improved the performance of RF over the raw abundance count of ASVs.

Feature engineering over PAAM [11], was conducted to deal with the high-dimensional nature of taxonomic features obtained from cattle microbiome. The following strategies with: i) RF applied over the features selected (top 10,30,50,70 %) by RFI from the PAAM ii) XgBoost over PAAM and iii) reg-LR over PAAM; provided significantly better performance with highest Accuracy of 0.887 over the other state-of-the-art phylogenetic methods of MetaPhyl (Accuracy: 0.662) [12] and PhILR (Accuracy: 0.40) [14]. RF over the top 10 % of features selected from PAAM provided the best predictive performance (Accuracy: 0.887, Kappa: 0.850) for classifying cattle microbiomes (Table 1). It was also noted that the ancestral nodes which attain weights by combining the abundance of leaves and phylogenetic distances annotated on the respective branches, played important role in top-ranked features by RFI in cattle microbiomes. Some of the important ancestors of ASVs (from phylum to genus level) in cattle microbial classification are noted in Table 2.

In order to further improve the performance of classification with RF over the RFI selected features from the PAAM, we experimented with different ensembles of ML methods. The ensemble ML was constructed by applying XgBoost and reg-LR over the RFI selected features from PAAM. The ensemble ML methods as enlisted in Table 3, provided comparatively better performance over PAAM in comparison to the other ML models listed in Table 1. The results of ensemble ML methods are useful for further comparative analysis (Table 3). The ensemble of top 30 % of RFI selected features with reg-LR (over the PAAM), provided highest performance with Accuracy: 0.925 and Kappa: 0.900 (Table 3).

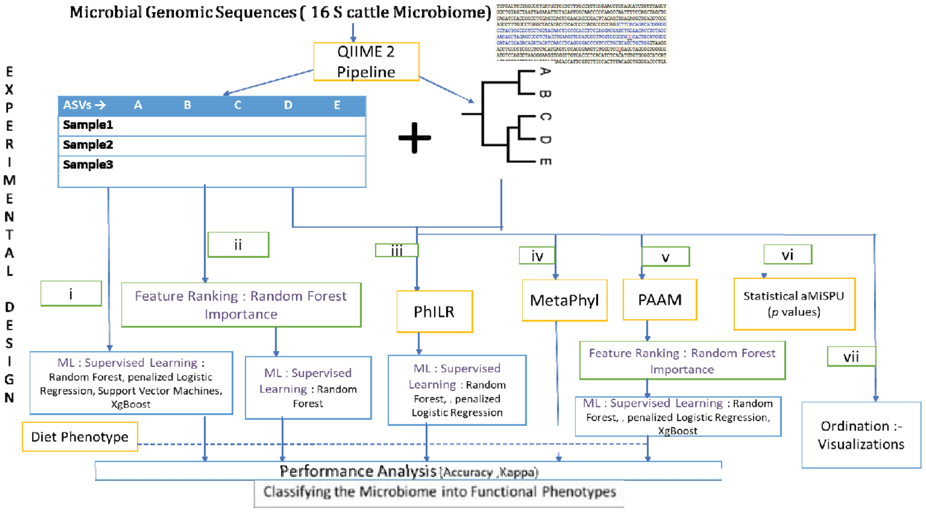
Microbial Genomic Sequences (16 S cattle Microbiome)
QIIME 2 Pipeline
EXPERIMENTAL DESIGN
ASVs → A B C D E
Sample1
Sample2
Sample3
+
A
B
C
D
E
i
ii
iii
iv
v
vi
vii
Feature Ranking : Random Forest Importance
PhILR
MetaPhyl
PAAM
Statistical aMiSPU (p values)
ML : Supervised Learning : Random Forest, penalized Logistic Regression, Support Vector Machines, XgBoost
ML : Supervised Learning : Random Forest
ML : Supervised Learning : Random Forest, , penalized Logistic Regression
Feature Ranking : Random Forest Importance
ML : Supervised Learning : Random Forest, , penalized Logistic Regression, XgBoost
Ordination :- Visualizations
Diet Phenotype
Performance Analysis (Accuracy ,Kappa)
Classifying the Microbiome into Functional Phenotypes

Fig. 1. Proposed Experimental Workflow for classifying cattle metagenomes

TABLE 1. RESULTS OF ML METHODS BASED ON PHYLOGENY AND RAW ABUNDANCES OVER THE CATTLE MICROBIOME

Approach	**ML Model with (LOOCV)**	**Accuracy**	**Kappa**
Phylogenetic (Integrating phylogenetic weights with raw abundances of ASVs)	**RF over Top 10 % derived from PAAM as ranked by RFI**	**0.887**	**0.850**
	RF over Top 30 % derived from PAAM as ranked by RFI	0.875	0.833
	RF over Top 50 % derived from PAAM as ranked by RFI	0.863	0.817
	RF over Top 70 % derived from PAAM as ranked by RFI	0.850	0.800
	reg-LR over PAAM	0.750	0.660
	XgBoost over PAAM	0.850	0.800
	RF over PhILR transformed	0.400	0.200
	reg-LR over PhILR transformed data	0.387	0.183
	MetaPhyl	0.662	0.559
Non-Phylogenetic (over Raw Abundances of ASVs)	RF over Original Abundance Counts	0.787	0.716
	reg-LR over Original Abundance Counts	**0.875**	**0.833**
	SVMs over Original Abundance Counts	0.675	0.566
	XgBoost over Original Abundance Counts	0.825	0.766
	RF over Top 10 % of ASVs as ranked by RFI	**0.850**	**0.799**
	RF over Top 30 % of ASVs as ranked by RFI	0.813	0.750
	RF over Top 50 % of ASVs as ranked by RFI	0.825	0.766
	RF over Top 70 % of ASVs as ranked by RFI	0.813	0.750

TABLE 2. TAXONOMY OF ASVS (TILL GENUS) THAT PLAYED IMPORTANT ROLE FOR CATTLE MICROBIAL CLASSIFICATION USING RFI OVER PAAM

Taxonomy from Kingdom to Genus
k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Corynebacteriaceae; g__Corynebacteriumftable
k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium;
k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Alcaligenaceae; g__Sutterella;
k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhodospirillales; f__Acetobacteraceae; g__Rhodovarius
k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhodospirillales; f__Rhodospirillaceae; g__Azospirillum;
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__Veillonella;

TABLE 3. RESULTS OF ENSEMBLE ML METHODS OVER PAAM

Ensemble ML Models (with LOOCV)	**Accuracy**	**Kappa**
reg-LR over Top 10 % derived from PAAM as ranked by RFI	0.887	0.850
XgBoost over Top 10 % derived from PAAM as ranked by RFI	0.912	0.883
reg-LR over Top 30 % derived from PAAM as ranked by RFI	**0.925**	**0.900**
XgBoost over Top 30 % derived from PAAM as ranked by RFI	0.900	0.867
reg-LR over Top 50 % derived from PAAM as ranked by RFI	0.900	0.866
XgBoost over Top 50 % derived from PAAM as ranked by RFI	0.912	0.883
reg-LR over Top 70 % derived from PAAM as ranked by RFI	0.837	0.783
XgBoost over Top 70 % derived from PAAM as ranked by RFI	0.862	0.816

Few high-level ordinations were applied to visualize the distribution of cattle microbiomes within (alpha diversity) and across samples (beta diversity) w.r.t diet phenotype.

The *plot_richness ()* in phyloseq R package [29] was used to estimate the alpha diversity with variously defined indices [29]. The resultant ordination is shown in Fig.2. For the beta diversity analysis, we used Principal Coordinate Analysis (PCoA) over weighted UniFrac matrix [30,31]. UniFrac is derived from the species' distances obtained from the phylogenetic tree. The calculation of UniFrac is based on the fraction of branch length that is shared between two samples or unique to one or the other sample [30]. *plot_ordination* () from phyloseq [29], was used to visualize the beta diversity (Fig.3.). PCoA revealed the separation of the cattle samples from 4 different diets along the first axis, suggesting the observed diversity across the two different breeds of cattle.

Correspondence analysis (CCA) [29] was conducted to graphically represent the relationship between diet and microbial species (Fig.4a); and unconstrained redundancy data analysis (RDA) [29] (equivalent to principal component analysis) was conducted to visualize the variation in a functional phenotype of supplemented diet that can be explained by a set of abundance of microbial species (Fig.4b) [29]. CDA and RDA identified environmental gradients along the two main ordination axes. The impact of supplement usage on the cattle diet was mainly revealed by the dense proportion of nitrate-treated samples.

V. CONCLUSION

In this paper, we classified cattle microbiomes into functional role of supplemented diet as part of the *EU H2020 MetaPlat project*[1]. The project aims to analysis the cattle related metagenomic sequences to provide useful insights on the probiotic supplement usage, methane production, and feed conversion efficiency. It was shown that dietary nutrients supplements are significantly associated with cattle microbiome composition. We studied the effect of integrating phylogeny of microorganisms present in a microbial community into their abundance counts. Prioritizing nodes of a phylogenetic tree based on the integration of structural and abundance information, supported better metagenomic classification when compared to state-of-the-art [7, 8, 12, 14].

ASVs reportedly provide better biological resolution and relevance than OTU methods [6]. In current study, we proposed an experimental workflow which applied ML models over the: - i) features as independent ASVs in classification of metagenomic sequences and ii) features as related ASVs by phylogeny. We uniformly evaluated metagenomic cattle microbiome data using leave-one-out cross validation to train and predict the performance of ML models for determining phenotypes. We recommended some of the best models for functional metagenomic analysis of cattle microbiomes. We proposed that embedded ML methods of XgBoost or reg-LR are most effective in dealing with high-dimensional metagenomic raw data. RFI also played an important role over the inputs integrating phylogeny and abundance values with PAAM [11].

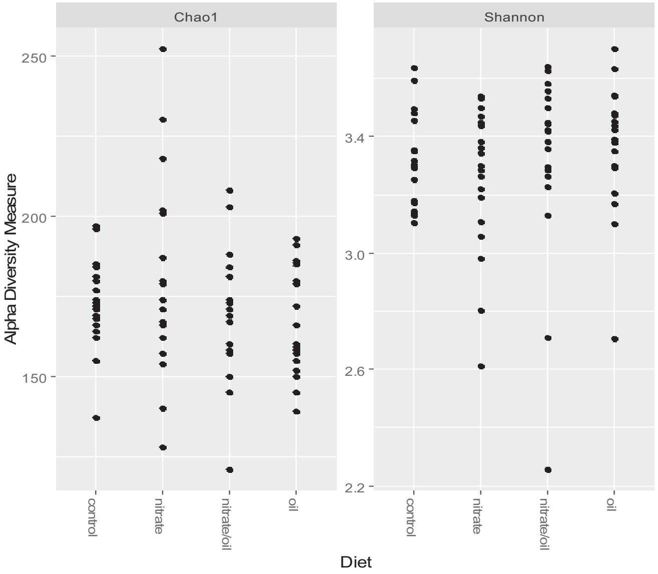

Fig.2. Alpha Diversity in Cattle Rumen Samples

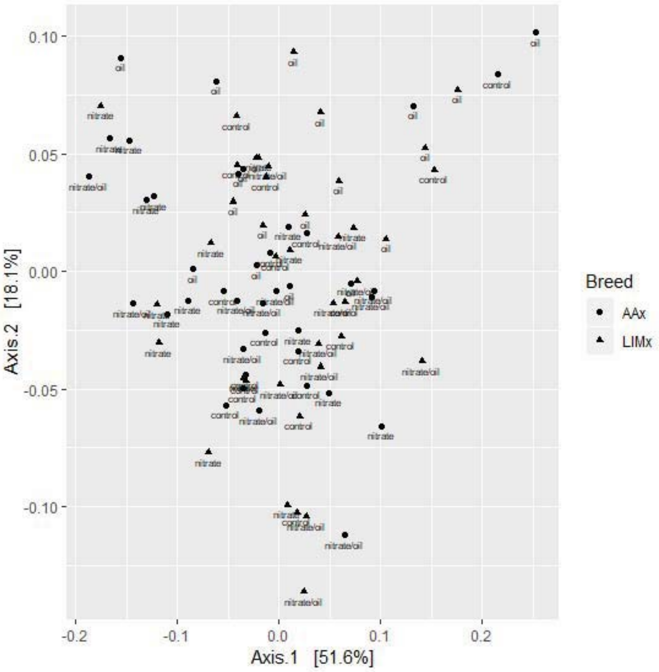

Fig.3. PCoA over weighted UniFrac in Cattle Microbiome w.r.t Diet in the two cattle breeds of Aberdeen Angus Cross (AAx) and Limonsin sired (LIMx)

The ensemble of XgBoost or reg-LR methods with RFI ranking of features over PAAM, further improved the classification performance over the high-dimensional metagenomes. Additionally, the analysis was benchmarked along the integrated workflow using phylogeny-driven methods of PhILR [14] and MetaPhyl [12]. Overall, the ensemble ML method combining reg-LR with top 30 % features obtained by RFI over PAAM, attained the best performance (over state-of-the-art phylogenetic and non-phylogenetic methods) in our use case. PAAM-based approach attained best performance over the sourced data.

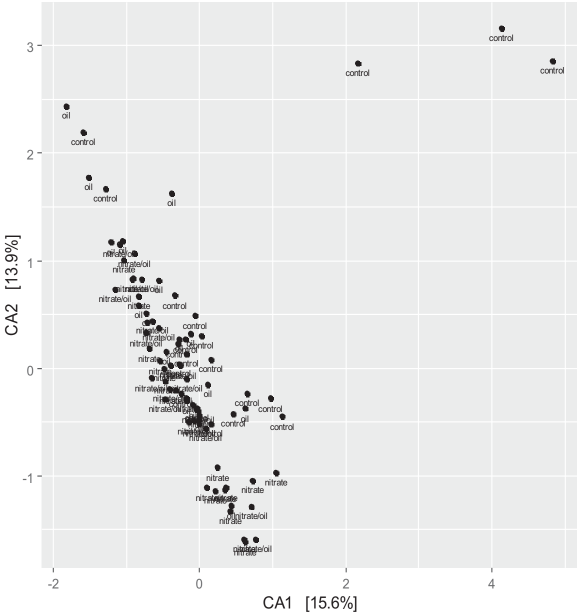

Fig.4a. Correspondence analysis over Cattle Microbial Samples

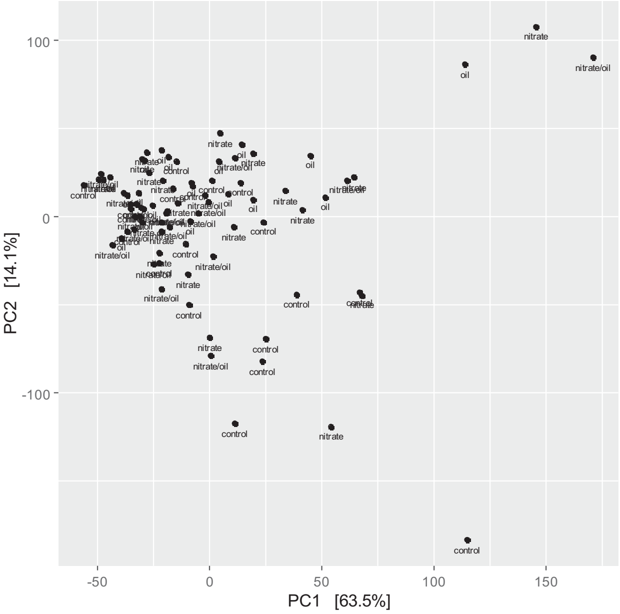

Fig.4b. Redundancy data analysis over cattle microbial samples

This indicates ancestral relationships between different microbial taxas (derived from their phylogeny) are important and drive a good classification performance. We also highlighted a few high-level visualizations of cattle microbiome composition in relation to the functional phenotype of diet.

In future, we would like to explore other advances in ML such as deep learning with 2D neural nets, networks analysis (depicting co-occurrence, interrelations), etc. for increasing the reliability of microbiome analysis using phylogeny along the current workflow. We will further explore other techniques to obtain phylogenetic rankings for all species and determine their effect on the classification process.

Acknowledgment

This research is supported by the Research Strategy Fund of Ulster University, UK and the MetaPlat project, funded by H2020-MSCA-RISE-2015-2019.

References

[1] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, “Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.,” Chem. Biol., vol. 5, no. 10, 1998.

[2] H. Soueidan and M. Nikolski, “Machine learning for metagenomics: methods and tools," pp. 1–23, 2015.

[3] R. Roehe, R. Dewhurst, C. Duthie, J. Rooke, N. McKain, D. Ross, J. Hyslop, A. Waterhouse, T. Freeman, M. Watson and R. Wallace, "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 Genetics, vol. 12, no. 2, p. e1005846, 2016.

[4] A. Patra, "The effect of dietary fats on methane emissions, and its other effects on digestibility, rumen fermentation and lactation performance in cattle: A meta-analysis", Livestock Science, vol. 155, no. 2-3, pp. 244-254, 2013.

[5] P. Walsh, C. Palu, B.Kelly, B.Lawor, J.Wassan, H.Zheng, and H.Wang. "A metagenomics analysis of rumen microbiome." In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2077-2082. IEEE, 2017.

[6] Callahan, P. McMurdie and S. Holmes, "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis", The ISME Journal, vol. 11, no. 12, pp. 2639-2643, 2017.

[7] D. Knights, E. Costello and R. Knight, "Supervised classification of human microbiota", FEMS Microbiology Reviews, vol. 35, no. 2, pp. 343-359, 2011.

[8] A. Statnikov, M. Henaff, V. Narendra, K. Konganti, Z. Li, L. Yang, Z. Pei, M. Blaser, C. Aliferis and A. Alekseyenko, "A comprehensive evaluation of multicategory classification methods for microbiomic data", Microbiome, vol. 1, no. 1, p. 11, 2013.

[9] J. Wassan, H. Wang, F. Browne and H. Zheng, "A Comprehensive Study on Predicting Functional Role of Metagenomes Using Machine Learning Methods", IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1-1, 2018.

[10] H. Wang, H. Zheng, F. Browne, R. Roehe, R. Dewhurst, F. Engel, M. Hemmje, X. Lu and P. Walsh, "Integrated metagenomic analysis of the rumen microbiome of cattle reveals key biological mechanisms associated with methane traits", Methods, vol. 124, pp. 108-119, 2017.

[11] J. Wassan, H. Wang, F. Browne and H. Zheng, " PAAM-ML : A novel Phylogeny and Abundance aware Machine Learning Modelling for Microbiome Classification ", in the Proc of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2018),in press.

[12] O. Tanaseichuk, J. Borneman, and T. Jiang, “Phylogeny-based classification of microbial communities,” Bioinformatics, vol. 30, 2014.

[13] S. Kim and E. Xing, "Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping", The Annals of Applied Statistics, vol. 6, no. 3, pp. 1095-1117, 2012.

[14] J. D. Silverman, A. D. Washburne, S. Mukherjee, and L. A. David, “A phylogenetic transform enhances analysis of compositional microbiota data,” Elife, vol. 6, pp. 1–20, 2017.

[15] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, “Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights,” PLoS Comput. Biol., vol. 12, 2016.

[16] Wu, J. Chen, J. Kim, and W. Pan, “An adaptive association test for microbiome data,” Genome Med., vol. 8, no. 1, pp. 1–12, 2016.

[17] S. Whelan, P. Lio, and N. Goldman, “Molecular phylogenetics: state-of- ` the-art methods for looking into the past,” TRENDS in Genetics, vol. 17, no. 5, pp. 262–272, 2001.

[18] F., Magee, M.R. May & B.R. Moore, B. R. “The dawn of open access to phylogenetic data”. PLoS One, 9(10), e110268, 2014.

[19] "NIH Human Microbiome Project", Microbe Magazine, vol. 4, no. 9, pp. 393-393, 2009.

[20] J. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. Bushman, E. Costello, N. Fierer, A. Peña, J. Goodrich, J. Gordon, G. Huttley, S. Kelley, D. Knights, J. Koenig, R. Ley, C. Lozupone, D. McDonald, B. Muegge, M. Pirrung, J. Reeder, J. Sevinsky, P. Turnbaugh, W. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld and R. Knight, "QIIME allows analysis of high-throughput community sequencing data", Nature Methods, vol. 7, no. 5, pp. 335-336, 2010.

[21] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 958 2001.

[22] T. Chen and C. Guestrin, “XGBoost : A Scalable Tree Boosting System,” pp. 785–794, 2016.

[23] J.Hosmer, D.W., S.Lemeshow, R.Sturdivant, "Applied logistic regression",vol. 398,. John Wiley & Sons,2013.

[24] M. Hearst, S. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines", IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, 1998..

[25] T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", Pattern Recognition, vol. 48, no. 9, pp. 2839-2846, 2015.

[26] M. Kuhn, “Building predictive models in r using the caret package,” Journal of Statistical Software, Articles, vol. 28, no. 5, pp. 1–26, 2008. [Online]. Available: https://www.jstatsoft.org/v028/i05

[27] K.Archer and R. Kimes, "Empirical characterization of random forest variable importance measures", Computational Statistics & Data Analysis, vol. 52, no. 4, pp. 2249-2260, 2008.

[28] C. Strobl and Z. Achim Zeileis. "Why and how to use random forest variable importance measures (and how you shouldn’t)." In The R User Conference. 2008.

[29] P. McMurdie and S. Holmes, "phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data", PLoS ONE, vol. 8, no. 4, p. e61217, 2013.

[30] R.Wong, J. Wu and G. Gloor, "Expanding the UniFrac Toolbox", 2018.

[31] J. Jovel, J. Patterson, W. Wang, N. Hotte, S. O'Keefe, T. Mitchel, T. Perry, D. Kao, A. Mason, K. Madsen and G. Wong, "Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics", Frontiers in Microbiology, vol. 7, 2016.

[32] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview", Bioinformatics, vol. 16, no. 5, pp. 412-424, 2000.

[33] T. Nichols, P. Wisner, G. Cripe and L. Gulabchand, "Putting the Kappa Statistic to Use", The Quality Assurance Journal, vol. 13, no. 3-4, pp. 57-61, 2010.