A Metagenomics Analysis of Rumen Microbiome

Paul Walsh[1], *Cintia Palu*[1,2], *Brian Kelly*[1], *Brendan Lawor,*[1]*Jyotsna Talreja Wassan*[3], *Huiru Zheng*[3], *Haiying Wang*[3]

[1] NSilico Life Science, Nova Centre, University College Dublin, Ireland
[2] School of Biochemistry & Cell Biology, University College Cork, Ireland
[3] School of Computing, Ulster University, United Kingdom
Paul.Walsh@nsilico.com

Abstract— **Climate change and food security are significant global challenges facing society. The dairy industry is inextricably linked to these challenges as it is concerned with the economies of food production, while acknowledging that it is a major contributor to greenhouse gas production. Action by microbial communities in the rumen is responsible for efficient breakdown of plant matter for food conversion, but a by-product of this action is substantial methane production. Insight into food conversion and methane production in rumen microbiota is possible through metagenomics analysis, which is the analysis of microbial communities and their interactions with the environment. However, metagenomic analysis is hampered by the sheer volume and complexity of data that needs to be processed. This paper presents a bioinformatics pipeline and visualisation platform that facilitates deep analysis of microbial communities, under various conditions in cattle rumen, with the aim of leading to significant impact on probiotic supplement usage, methane production and feed conversion efficiency. This pipeline was developed as part of the EU H2020 MetaPlat project and will pave the way for a more optimal usage of metagenomic datasets, thus reducing the number of animals necessary to be engaged in such studies. This will ensure better and more economic animal welfare, better use of resources and lessen the impact of the dairy industry on climate change.**

Keywords—Metagenomics; cattle rumen; visualisation; classification, cloud architecture;

I. Introduction

Some of the most pressing global challenges are climate change and food security [1] and one of the biggest sectors that addresses both issues is the dairy production sector. It is noted that in many economies that dairy production is a major contributor to greenhouse gas emissions [2], [3]. However, there is significant economic, nutritional, and cultural value placed on the dairy industry, which stems from the ability of cattle to convert forages into quality, high protein products for human nutrition through fermentation by rumen microbiota, which have a vital role in cattle performance, productivity, health and immunity. Therefore, any strategy that aims to mitigate greenhouse gas emission also needs to maintain the efficiency of cattle in food production by investigating the action of rumen microbiota. This paper outlines our approach, which is centred on developing a reproducible and computationally scalable metagenomics platform, known as MetaPlat (www.metaplat.eu).

A. Cattle Rumen

The physiology of cattle is adapted to host such microorganisms through the rumen, which is the main chamber of the ruminant stomach, which contains symbiotic microbes that play a key role in the digestion of ingested food. *Bos taurus* is a member of the ruminant, a group of mammals which also include sheep and goats. Many studies have investigated the symbiotic microorganisms in the rumen because of their link to economically or environmentally important traits such as feed conversion efficiency, methane production, and more recently the discovery of microbes and enzymes that enable fermentation of biomass for biofuel production [4].

A key challenge is identifying rumen microbial profiles, which are associated, and potentially predictive of these traits. In typical rumen studies the emission and food production efficiency of cattle needs to be evaluated in a controlled experiment, where conditions such as food type and intake are evaluated against the composition of the gut microbiome. This can facilitate the creation of taxonomies and predictive models. To do so, both phenotype and genotype data need to be extracted, sequenced, cleaned, transformed, analysed and visualised in a reproducible traceable manner. Research into gut microbial community genomic composition is therefore crucial to provide knowledge on the functions of the microbiota to the physiological well-being of the host, insight into methane production, food production efficiency and meat/milk quality.

B. Metagenomics

A method of investigating gut microbiota genetic profiles is the use of metagenomics analysis, which is the study of genomic sequences extracted from microbiome samples. The growth of metagenomics stemmed from the evidence that as-yet-uncultured microorganisms represent the majority of organisms on earth [5]. These findings are distilled from analyses of 16S rRNA gene sequences and other sequences amplified directly from the ecosystem. This approach avoids the limitations imposed by culturing and can lead to the discovery of new lineages of microbial life.

While the characterisations of microbiota are still mainly focussed on the analysis of 16S rRNA genes, such studies yielded a phylogenetic description / profiling of community membership. Metagenomics profiling outside of 16S rRNA analysis enables the study of the presence of antibiotic resistance genes, metabolic pathways, and other important information to understand the dynamics of the gut microbiome. The number of projects or studies producing very large quantities of metagenomic data has increased in recent years, yet often the depth of analysis done is limited by the financial resources necessary to obtain samples and the cost of analysis and interpretation.

Metagenomics based on high-throughput sequencing offers unparalleled coverage and depth in determining microbial gut dynamics, but this is only feasible if the analytic computational resources are available. Thus, in order to investigate microbiota in the context of probiotic supplement usage, methane production and feed conversion efficiency, the careful development of a research software platform to fully analyse metagenomic data is necessary. A key step in building such an analysis is the provision of a cloud based research infrastructure that allows researchers to load and link controlled experimental data on cattle breed, supplement usage and feed correlated with methane production, and food production.

C. *MetaPlat*

To address these challenges, we have developed the MetaPlat platform, (www.metaplat.eu), which is a cloud based research infrastructure for metagenomics analysis of rumen microbiota. We illustrate how investigators can use MetaPlat to rapidly analyse and manage genome, phenome datasets along with related metadata. The system supports reproducibility and tracks the phenotypic information associated with each sequence, including its origin, quality, taxonomical position and associated biological genome and produces automated reports and visualisations. A metagenomic analysis of rumen microbiota is presented later in this paper.

A key issue that MetaPlat addresses is the lack of easy to use scalable parallel architectures and approaches to deal with the huge number of generated sequences that are produced in metagenomics. Reproducibility is also an issue and attributing a sequence to an organism is done by sequence similarity, which is not optimal if the sequences diverge greatly or not enough: e.g. if more than one sequence are identical, the assignment is often given by the first sequence name encountered by the algorithm, which can be ambiguous and even missing, producing very biased and non-reproducible results and statistics.

To specifically tackle the gut microbiota role in cattle performance, productivity, health, and immunity, then reproducible and easily scalable tools need to be developed and better analyses and practice must be developed for future and current studies. Hence there is a need for reproducible, scalable and time-efficient comparisons and analyses on large datasets producing phylogeny-aware classification and quantitative and functional analyses (when possible) for both 16S based and whole genome/transcriptome approach.

II. Methods

We address the objectives of the MetaPlat project in a reproducible fashion by implementing bespoke Docker technology, which facilitates reproducibility by encapsulating a complete environment with system tools, scripts, libraries, and tool dependencies [5]. In doing so this provides transparency in experimental methodology, observation, and collection of data to regulators through the development of web-based tools to facilitate collaboration, storage, and integration. It will also facilitate the management of large input raw data and reference data sets using technology such as Docker Volumes. Data is shared and managed using digital information objects interlinked with internal and external resources in a structured and machine-readable serialization mechanism, as measured against biocomputer objects BCO [1] by adding the required provenance and descriptive, i.e., domain information (metadata) for ensuring results may be shared and reproduced over long term data life cycle.

The MetaPlat system utilises highly scalable an asynchronous queueing system that lends itself to scaling up (making processing nodes more powerful) and scaling out (adding multiple processing nodes in parallel) as illustrated in Fig. 1. The platform uses a message-based producer/consumer system to manage the metagenomics analysis jobs. The API endpoint in the web application builds a job message and queues it in the Microsoft Azure Queue System (AQS). A message queue provides an asynchronous, non-blocking, decoupled message communication between two, or more, bits of code. The message contains information about what needs to be done and where the data can be retrieved in Azure SQL.

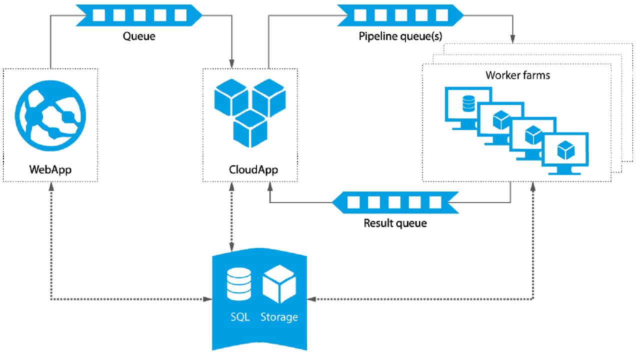

Fig. 1 MetaPlat Cloud Architecture and Queuing System.

Such queueing systems have many advantages. Firstly, their asynchronous nature means that resource usage is kept as efficient as possible: long-running jobs do not hold onto I/O resources and their related threads needlessly. Secondly, loose coupling between queues and their consumers permits the creation of multiple consumers without significant impact on the functioning of the queue itself. The queue does not need to 'know' about or manage its consumers, but rather processors nodes need only subscribe to the queue service. Scaling becomes a relatively simple matter of adding more processes on a multi-core node, or adding more nodes in a distributed system. Although some data processing is complex, in that it needs to recombine the results of parallel and distributed processes, certain architectures like the Actor Model (as exemplified by Akka or the Erlang language) can make this easier by effectively implementing a queueing system at a more fine-grained level.

The rapid provision of reproducible compute containers (using technology such as Docker [5] and Kubernetes [6]) dovetails with the queue-based approach described above. In the last two years, Containerization has been widely adopted for its ability to provide isolated, reproducible, and scalable computing components in an elastic fashion. Containers define the runtime environment in which processes are deployed, including the 'flavour' of operation system, as well as the installed tools and libraries. This simplifies the job of distributing such computing resources over large numbers of physical nodes, especially with the introduction of 'orchestration' tools such as Kubernetes and Nomad. Such 'democratization' of distributed programming allows even modest development teams to create highly scalable, reliable and traceable systems. Cloud providers such as AWS, Azure, Google and others have been quick to realise this and now offer many tools and interfaces to further simplify the deployment of large-scale parallel systems.

To evaluate the performance of the platform, we implemented a microbiome analysis pipeline for 16S NGS deep-sequencing, relying on QIIME [7], along with visualizations generated in R to support the interpretation of the data. It includes bar charts, heatmaps and principal component analysis (PCA). The packages pheatmap [8], optparse [9], randomForest [10], klaR [11] ggplot2 [12] and ggfortify [13], [14] were used in the development of the necessary scripts.

The pipeline was tested on the *Bos taurus* rumen microbiota samples sequenced for MetaPlat project: - a new data set (SineadNEBQ5) and a data set previously described [15]. The meta-analysis for the categorization of binned metagenomic sequences, known as Operational Taxonomic Units (OTUs) into categories based on feed treated with oil or nitrate in the pipeline was achieved through application of supervised machine learning (ML) technique of classification. The related general workflow is represented in Fig. **2**.

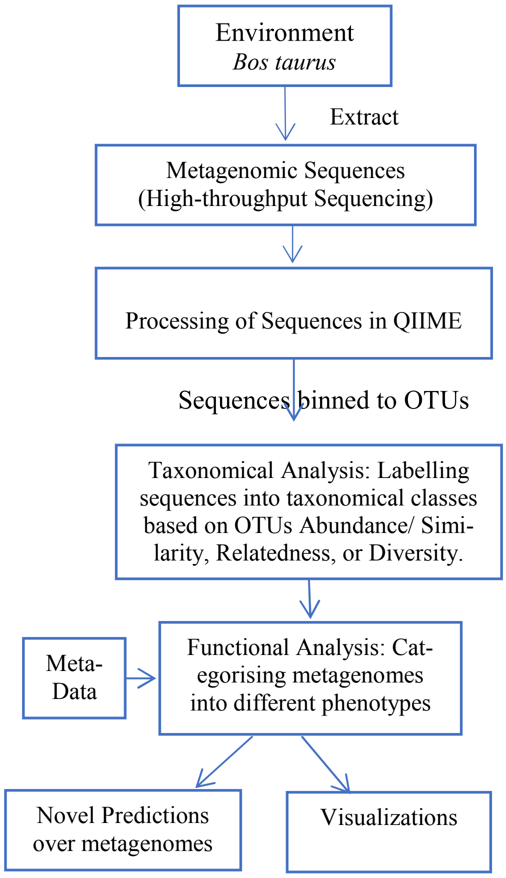
Environment
Bos taurus
Extract
Metagenomic Sequences
(High-throughput Sequencing)
Processing of Sequences in QIIME
Sequences binned to OTUs
Taxonomical Analysis: Labelling sequences into taxonomical classes based on OTUs Abundance/ Similarity, Relatedness, or Diversity.
Meta-Data
Functional Analysis: Categorising metagenomes into different phenotypes
Novel Predictions over metagenomes
Visualizations

Fig. 2 Metagenomic Analysis Workflow Collaborating Taxonomical and Functional Context

The data on the SineadNEBQ5 follows a 2 x 2 factorial design and a total of 40 samples. It consists of two breed types, Aberdeen Angus (AAx) and Limousin (LIMx), and two diets, an oil based and a nitrate based feeding. The nitrate diet was found to reduce methane emissions, based on data captured from animal studies.

III. Results

The pipeline generated 5 output tables with taxonomic resolution varying from phylum to genus. The number of microbial features used to investigate the influence of diet (oil/nitrate) on cattle rumen microbial metabolites are highlighted in Table I.

Table 1 Count of OTUs as Features at 5 Taxonomical Levels

# of Sample rows = 40 in each table	**Column Attributes/Features in OTU Tables**	**Taxonomic Level of Classification**
OTU Table 1	**27**	Phylum (L2)
OTU Table 2	**52**	Class (L3)
OTU Table 3	**101**	Order (L4)
OTU Table 4	**194**	Family (L5)
OTU Table 5	**386**	Genus (L6)

The algorithms used in experiments for supervised classification [16], [17] were: Naïve Bayes(NB), Neural Networks (NN), Support Vector machine(SVM), Random Forest(RF), Adaptive Boosting (AdaB), Nearest Neighbor(K-NN), Ensemble of Zero-R(Voting), NN, k-NN, LWL classifiers (E-ZNNL) and Logistic Regression(LR), as listed in Table II. The results presented in Fig. **3**, are obtained after the application of listed classifiers, and tuning their learning parameters to yield optimum possible output in terms of Accuracy (Acc.), Precision (Pr.), Specificity (Sp.) and Sensitivity (Se.) [18]. In literature, RF has been proposed as the best model for metagenomic analysis [19], [20]. Our analysis over the sampled metagenomic cattle data (Fig. **3**.), supports “No lunch free theorem” illustrating no single computational model is best for metagenomic analysis at different level of taxonomies [21].

Table II Classifiers and Optimizing Parameter Settings

Classifier/Model	**Optimizing Parameter Settings**
LR	batch size = 100, ridge estimator for log likelihood: 1.0E-8, Conjugate Descent algorithm: True/false, max no of iterations: -1 to 100
NN	batch size =100, hidden layers=01/02/no=(attr+classes)/2, value used to seed random number generator (seed)= 1-10, Training Time = 100-500, validation threshold =20, model learning Rate = 0.3, momentum =0.2
SVM	batch size 100, calibrator: Logistic, Kernel = PolyKernel/, RBF kernel, random seed = 1 to 10, complexity parameter c = 1/4/8/ (= 4 steps), Optimization: sequential minimum optimization algorithm
RF	batch size 100, maxDepth: 0 to 6, seed: 1-10, no: of iterations: 100,
NB	batch size =100, use kernel estimator = false/true
k-NN	batch size =100, No: of neighbours (KNN) = 1-4, Search algorithm: Linear NN search, Window size =0, distance Weighting: false, mean squared: false
AdaB	batch size =100, classifier: NN, no: of iterations: 10, seed: 1-10, weight threshold: 100, resampling: false
Stacking/Ensemble (E-ZNNL)	batch size: 100, num-Folds: 10, seed: 1, classifiers: 4, metaclassifiers: Zero-R, NN, k-NN, LWL

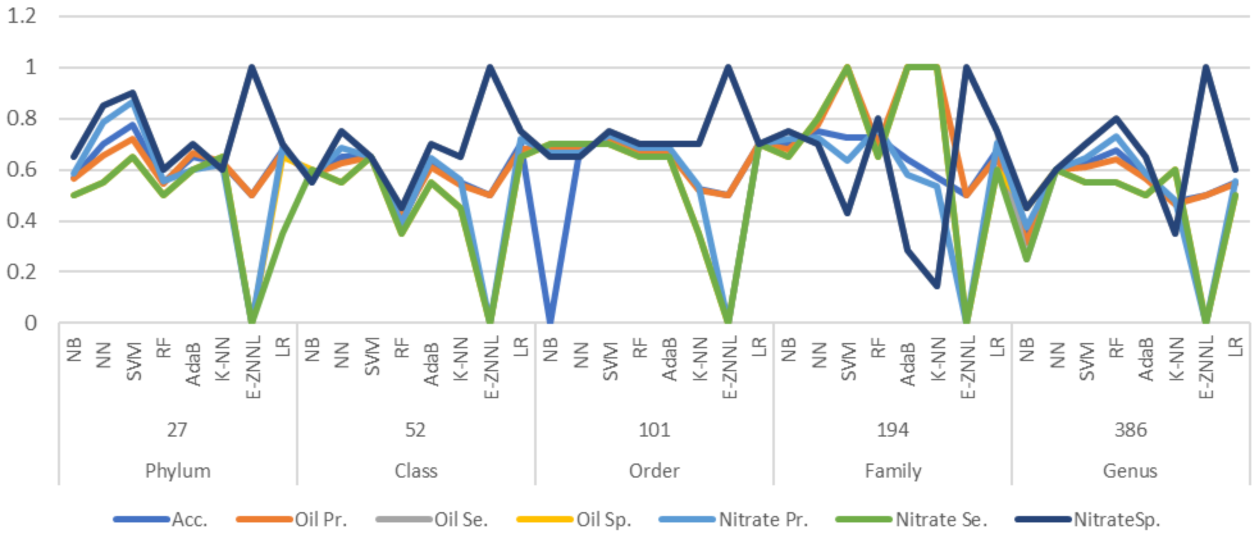

Fig. 4 Analysis with Supervised ML models over *Bos taurus* microbiome

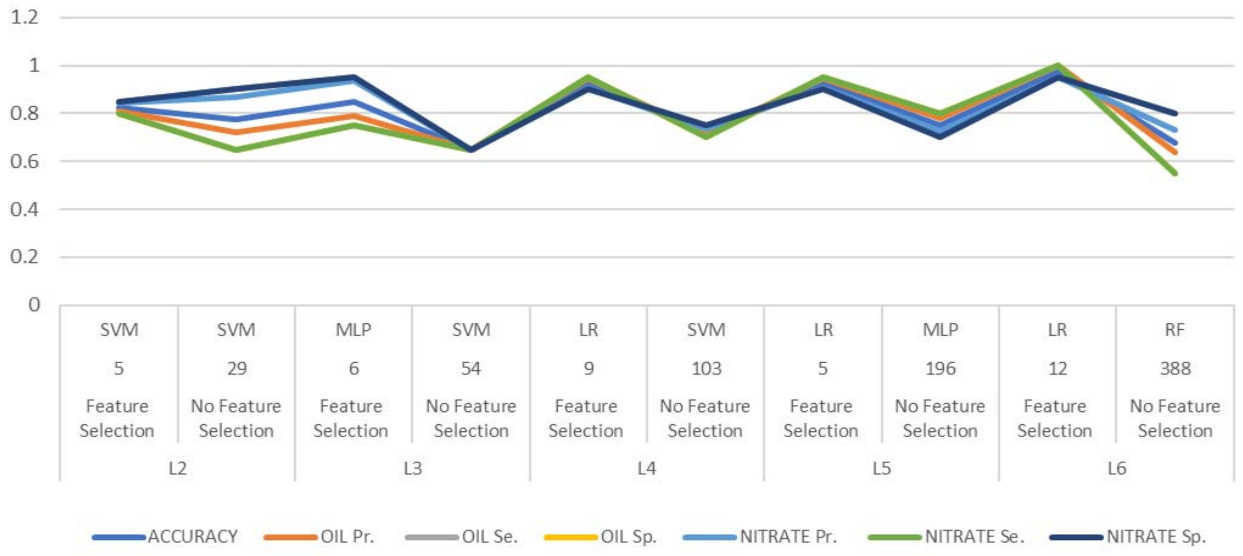

Fig. 3 Analysis with Supervised ML models over *Bos taurus* microbiome: - with and without feature selection (using Wrapper based feature selection strategy with Logistic Regression)

Also, the study for predictive modelling over the use case of *Bos taurus* was enhanced using feature selection strategies [15]. Wrapper based feature selection strategy with Logistic Regression proved to best in terms of accuracy for analyzing the cattle rumen microbiome at genus level of study [15]. The improvements achieved with feature selection are indicated in Fig. 4. The further validations in terms of different sizes of test and train sets for the proposed ML model [15] are indicated in Fig. **5**.

The following examples were generated based on the analysis of SineadNEBQ5 files, for which we identified 29 phyla, 61 classes, 117 orders, 221 families and 470 genera. Due to the large number of organisms assigned in the higher resolu-

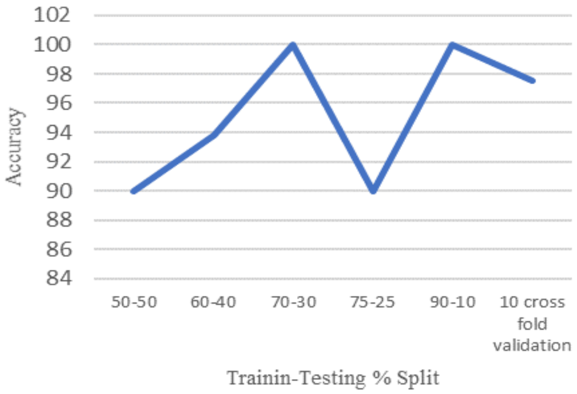

Fig. 5 Performance over different test sets using Wrapper based feature selection strategy with Logistic Regression for classifying *Bos taurus* microbiomes

tion results (family and genera), we had to implement an in-house color palette.

The bar chart summarizes the samples' metagenome relative composition, separated according to specified features, and has a complete list of taxa (Fig. 6). The graphics presented here were made using the Class level output (level 3 output table). To support the data interpretation, we also provide the average composition for each relevant feature, accompanied by a caption highlighting the most abundant organisms (Fig. 7).

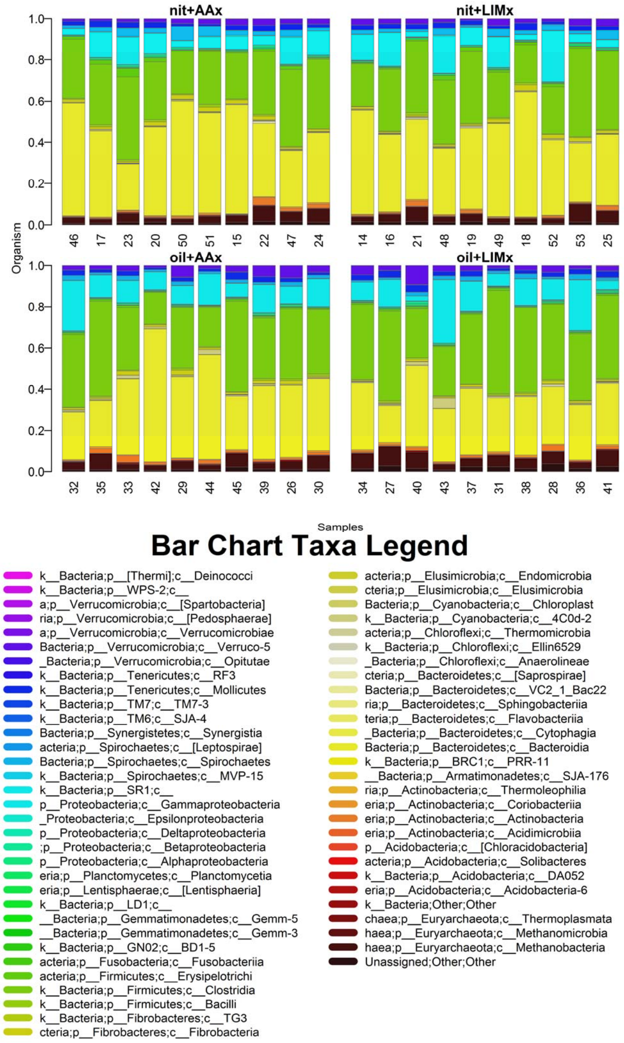

Fig. 6 Bar chart illustrating the relative taxa composition of each sample

In these examples, we separated the samples based on breed and diet. In this study-case, 60 bacteria classes where identified (Fig. 6), whereas only 14 had an average presence above 0.5%. On average, the most abundant class is bacteroidia, followed by clostridia, but the samples from the breed Limousin treated with oil-based feed, the clostridia are the most predominant class on average (Fig. 7).

The PCA supports the understanding of which organisms are defining the differences between samples and helps to understand how they relate to the treatments (Fig. 8). In this use-case, there are three classes that are predominant (bacteroidia, clostridia and gamma proteobacteria) and they are the ones defining the PCA plot structure.

These new tools and techniques help to provide added insight and analysis to metagenomics and we will continue to build on this work in subsequent months.

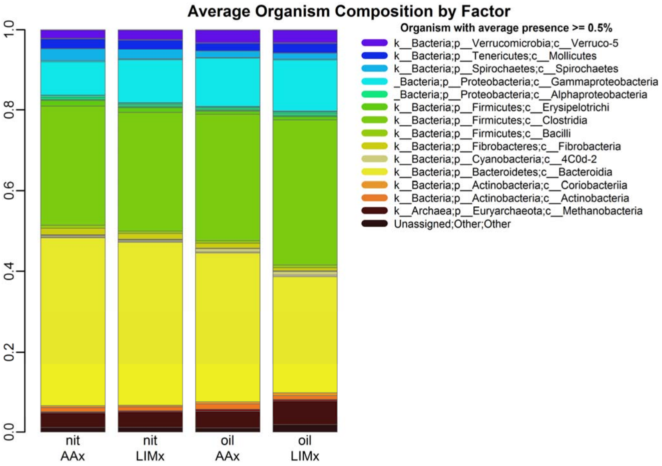

Fig. 7 MetaPlat visualization on metagenome average relative composition for breed versus treatment.

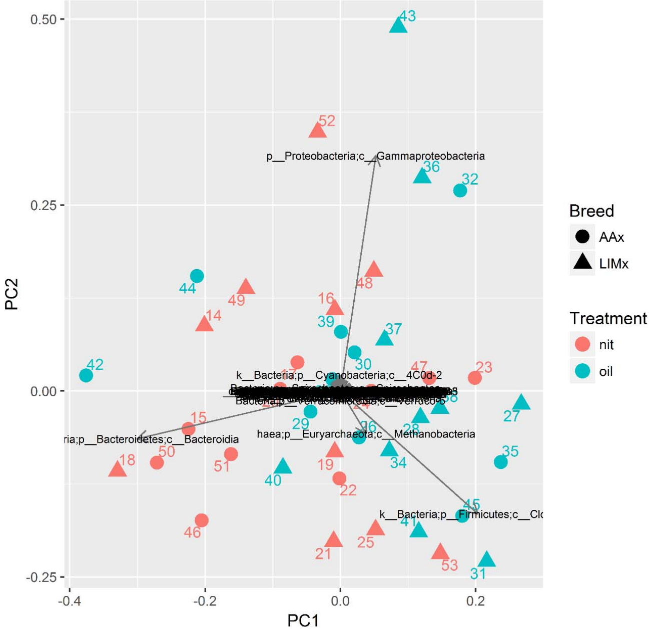

Fig. 8 Principle component analysis for MetaPlat.

IV. Summary

This paper provides an overview of the MetaPlat metagenomics platform and reports on progress to date. The system is based around a highly scalable asynchronous queueing system that lends itself to scaling up and scaling out. It also facilitates reproducibility by use of container technology. A sample use case was demonstrated on data from a 2 × 2 factorial design originals from MetaPlat. The study involved the analysis of rumen microbiome using two breed types, Aberdeen Angus and Limousin and two diets involving concentrate and forage. Future work will focus on extending the platform and optimizing its features.

The system generated graphical summaries of the metagenome composition of the samples and highlighted the most abundant classes and a complete list of taxa. The outputs provide understanding of which organisms contribute to the differences between samples and helps to understand how they relate to the treatments.

V. Acknowledgments

The MetaPlat (http://www.meta plat.eu) project is founded by the European Horizon H2020-MSCA-RISE-2015 and the research presented here was also supported by the Irish Research Council.

VI. References

[1] S. Henson, “Global Challenges: Food, Agriculture and Nutrition,” *Global Challenges,* vol. Volume 1, no. Issue 1, 2015.

[2] J. W. &. H. N. M. Casey, “Analysis of greenhouse gas emissions from the average Irish milk production system.,” *Agricultural systems,* vol. 86, no. 1, pp. 97-114, 2005.

[3] H. W. J. D. E. &. S. A. F. Phetteplace, “Greenhouse gas emissions from simulated beef and dairy livestock systems in the United States,” *Nutrient cycling in agroecosystems,* vol. 60, no. 1, pp. 99-102, 2001.

[4] M, Hess; A, Sczyrba; R, Egan; TW, Kim; H, Chokhawala; G, Schroth; S, Luo; DS, Clark; F, Chen; T, Zhang; RI, Mackie; LA, Pennacchio; SG, Tringe; A, Visel; T, Woyke; Z, Wang; EM, Rubin, “Metagenomic discovery of biomassdegrading genes and genomes from cow rumen.,” *Science,* pp. 331(6016):463-7, 2011.

[5] C. Boettiger, “An introduction to Docker for reproducible research.,” *ACM SIGOPS Operating Systems Review 49.1,* pp. 71-79, 2015.

[6] B. Burns, B. Grant, D. Oppenheimer, E. Brewer and J. Wilkes, “Borg, Omega, and Kubernetes,” *ACM Queue,* vol. 14, pp. 70-93, 2016.

[7] J.G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K Costello, et al, “QIIME allows analysis of high-throughput community sequencing data.,” *Nature Methods,* vol. 10.1038/nmeth.f.303, 2010.

[8] R. Kolde, “pheatmap: Pretty Heatmaps,” R package version 1.0.8., 2015.

[9] T. L Davis, “optparse: Command Line Option Parser,” R package version 1.4.4., 2017.

[10] A. Liaw;M.Wiener, “Classification and Regression by randomForest.,” R News 2(3), 18--22., 2002.

[11] C. Weihs, U. Ligges, K. Luebke, N. Raabe, “klaR Analyzing German Business Cycles,” In Baier, D., Decker, R. and Schmidt-Thieme, L., *Data Analysis and Decision Support,* pp. 335-343, 2005.

[12] H. Wickham, ggplot2: Elegant Graphics for Data Analysis., New York: Springer-Verlag, 2009.

[13] M. Horikoshi and Y. Tang, “ggfortify: Data Visualization,” *Tools for Statistical Analysis Results,* 2016.

[14] Y. Tang, M. Horikoshi, W. Li, “ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages.,” *The R Journal,* 2016.

[15] J. T. Wassan, HY. Wang, F. Browne, P. Walsh, B. Kelly, C. Palu, N. Konstantinidou, R. Roehe, R. Dewhurst and H. Zhen, "An Integrative Approach for the Functional Analysis of Metagenomic Studies," in *Intelligent Computing Theories and Application. ICIC 2017. Lecture Notes in Computer Science*, Springer, Cham, 2017.

[16] S. David, A. Saeb, K. A. Rubeaan, “Comparative Analysis of Data Mining Tools and Classification,” *comput, Eng.Intell,* pp. 4,28-39, 2013.

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

[18] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manag.,* vol. 45, no. 4, pp. 427–437, 2009.

[19] D. Knights, E.K. Costello, R. Knight, “Supervised classification of human microbiota” FEMS Microbiology Reviews, 35(2), pp. 343-359, 2011

[20] A. Statnikov, M. Henaff, V. Narendra, K. Konganti, Z. Li, L. Yang, et al., “A comprehensive evaluation of multicategory classification methods for microbiomic data,” *Microbiome,* 2013, 11:1

[21] D.H. Wolpert and W.G. Macready, “No free lunch theorems for optimization,,", *IEEE Transactions on Evolutionary Computation,* vol. 1, no. 1, pp. 67–82, 1997.