MÁSTER INTERUNIVERSITARIO EN MEJORA GENÉTICA ANIMAL

Y BIOTECNOLOGÍA DE LA REPRODUCCIÓN

VARIANCE COMPONENTS ESTIMATION OF COMPLEX TRAITS INCLUDING MICROBIOTA INFORMATION

Tesis de Máster

Valencia, Julio, 2018

Alejandro Saborío Montero

Director:

Oscar González Recio

ACKNOWLEDGMENTS

This master thesis was developed at the Animal Breeding Department of the National Institute for Agricultural and Food Research and Technology (INIA), Madrid, Spain. Data was provided by the METALGEN project. The student was granted a scholarship from Universidad de Costa Rica (UCR) to course the two years master degree.

I would like to thank my advisor Dr. Oscar González Recio for been such a supportive leader who was always there when I needed, giving me the right direction to achieve little goals every day, with outstanding patience and kindness in the teaching process. He not only surpassed my expectations as a director with advanced technical knowledge, but as a human being. His enthusiastic nature and optimization in the use of time has inspired me not only in the professional area but in the way to combine a successful career with other meaningful things in life.

I thank to Agustin Blasco for organizing the Master in Animal Breeding and Reproduction Biotechnology, also to all the teachers that shared their knowledge with me and to my classmates that will always be my friends.

I would also like to thank to Latifa Ouatahar, Elisabeth Morales, Gabriel Molina, Beatriz Delgado, Adrián Lopez, Silvia Garcia-Ballesteros and María Saura for their friendship and for make labour days and coffee break were more pleasant.

Special thanks to my parents Alejandro and Marta and also to my sisters Gabriela, Priscilla, Rebeca and Graciela for all their support, and to my nephews Alejandro and Isaac for being a source of joy in my life

And finally I would like to give the most special thanks to my wife Sinaí Mariela for being always there for me and for her unconditional support, also to my children Andrés, Priscilla and David for making sense of all my efforts and being the engine of my will.

COMMUNICATIONS TO CONFERENCES

The results obtained in this thesis have led to two oral communications, the first one in the National Meeting of Animal Breeding at University of León and the second has been accepted and will be presented in the 69th EAAP Annual Meeting 2018, Dubrovnik, Croatia.

Saborío-Montero, A., Bach, A. and Gonzalez-Recio, O. 2018. Predicting complex traits using microbiome information: A comparison of metagenome distance matrices. XIX Reunión Nacional de Mejora Genética Animal. León, 14 y 15 de junio de 2018. España.

Saborío-Montero, A. and Gonzalez-Recio, O. 2018. Ordination methods to build microbiota similarity matrices for complex traits prediction. 69th EAAP Annual Meeting 2018, Dubrovnik, 27 – 31 August 2018. Croatia.

INDEX

INDEX OF FIGURES

INDEX OF TABLES

RESUMEN

1. RESUMEN

La influencia del microbioma sobre los rasgos complejos relevantes para el ganado lechero, tales como la eficiencia alimenticia o las emisiones de metano, ha sido bien establecida. Además, estudios recientes han publicado evidencias sobre el control de los antecedentes genéticos del animal sobre la composición de la microbiota. Sin embargo, hasta ahora la mayoría de los análisis se han centrado en enfoques de microorganismos únicos en lugar del conjunto de microbiomas que incluyen las relaciones subyacentes. El análisis conjunto de los antecedentes genéticos del huésped y su microbiota requiere tener en cuenta la distancia (o disimilitud) entre las comunidades de microorganismos en diferentes hospedadores. Por lo tanto, es necesario incorporar el microbioma completo en los modelos estadísticos para evaluar su asociación con rasgos complejos. La matriz de relación del microbioma (MRM) permite considerar la microbiota como un todo. Se han propuesto varios métodos para ordenar estas matrices; los cuales difieren en la métrica utilizada en la distancia (o disimilitud) entre las comunidades microbianas (por ejemplo, Euclidiana, Bray-Curtis, $\chi 2$). Estas distancias representan la diversidad alfa y beta de diferentes maneras. Aún no se ha llegado a un consenso sobre qué método es el más apropiado y podría depender de las singularidades de los datos y del propósito del estudio. El objetivo de este estudio fue comparar varias matrices de relación de microbiota, dentro de un marco de estimación de componentes de varianza. Se probaron cinco métodos de ordenación para construir la MRM: escalado multidimensional métrico (MDS), análisis de correspondencias sin tendencia (DCA), escalado multidimensional no métrico (NMDS), análisis de redundancia (RDA) y análisis de correspondencia restringida (CCA). La matriz de abundancias relativas log-transformada y estandarizada descrita en Ross et al. (2013) se utilizó como matriz de referencia.

Se utilizaron datos simulados (n = 1000) para estimar los componentes de la varianza, incluidos los fenotipos, los genotipos y la información de la microbiota del rumen. Los datos fueron analizados considerando dos posibles modelos. Primero, el efecto genómico y el efecto de la microbiota se incluyeron de forma independiente. Segundo, se agregó un efecto de interacción entre los efectos genómico y de microbiota. Todos los modelos se implementaron dentro de un marco Bayesiano utilizando el paquete BGLR en R. Se generaron un total de 100 repeticiones. Los datos reales se analizaron utilizando los mismos modelos.

Una estimación similar o ligeramente mejor en la simulación de h^2 (0,30) y m^2 (0,50) para los modelos de efectos independientes resultó de los métodos de ordenación MDS (0,307 y 0,493), RDA (0,307 y 0,501) y CCA (0,305 y 0,500) en comparación con la MRM de referencia (0,304 y 0,480), mientras que un pobre desempeño de los métodos DCA (0.249 y 0.349) y NMDS (0.217 y 0.266) se obtuvieron al estimar esos parámetros. Los coeficientes de correlación entre los valores genómicos de cría estimados (GEBV) y los valores verdaderos de cría (TBV), de mayor a menor, fueron: los obtenidos con la matriz de referencia ($\rho = 0.633$), CCA ($\rho = 0.631$), RDA ($\rho = 0.624$), DCA ($\rho = 0.598$), MDS ($\rho = 0.592$) y NMDS (0.557). Del mismo modo, las correlaciones para el efecto predicho de la microbiota en el mismo orden fueron: la matriz de referencia ($\rho = 0.975$), CCA ($\rho = 0.966$), RDA ($\rho = 0.949$), MDS ($\rho = 0.845$), DCA ($\rho = 0.807$) y NMDS ($\rho = 0.517$). Resultados similares, en términos de rendimiento de matrices, se obtuvieron para el modelo que incluyó interacción entre efectos. Un conjunto de datos reales (n = 70) también se analizó bajo los mismos estándares. Se observaron bajas estimas de heredabilidad para la eficiencia alimenticia (de 0.077 a 0.083) y la microbiabilidad (de 0.073 a 0.103); sin embargo, se obtuvieron valores consistentes para la microbiabilidad con las MRM que se desempeñaron mejor en las simulaciones (de 0,073 a 0,077). Además, se obtuvieron altas correlaciones ($\rho > 0.85$) entre el efecto genético del huésped y los fenotipos para todos los métodos, así como altas correlaciones entre el efecto de la microbiota y los fenotipos para las matrices RDA ($\rho = 0.91$) y CCA ($\rho = 0.91$).

Ambos modelos se compararon utilizando los criterios de información de desviación (DIC), número efectivo de parámetros (pD) y la media posterior del logaritmo de la verosimilitud (PostMeanLogLik), resultando en valores ligeramente inferiores para el modelo de efectos independientes (DIC: 183.9 a 189.3) en comparación con el modelo de efectos de interacción (DIC: 187.5 a 191.7), esos resultados indican que podría existir una relación que vincula genotipo-microbioma-fenotipo que podría usarse en la predicción de rasgos complejos.

Los análisis realizados en esta tesis sugieren que los métodos de ordenación canónica de RDA y CCA para crear MRM son preferidos cuando la información completa de la microbiota se incluye en los modelos estadísticos para el análisis de rasgos complejos.

Palabras clave: eficiencia alimentaria, microbiabilidad, heredabilidad, métodos de ordenación

ABSTRACT

2. ABSTRACT

The influence of the microbiome on relevant complex traits for dairy cattle, such as feed efficiency or methane emissions has been well established. Further, recent studies have released evidences on the control of the genetic background of the animal over the microbiota composition. However, until now most analyses have focused on single microorganism approaches instead of the joint microbiome as a whole, including underlying relationships. The joint analysis of the genetic background of the host and its microbiota requires accounting for the distance (or dissimilarity) between communities of microorganisms in different hosts. Therefore, it is necessary to incorporate the whole microbiome into the statistical models to assess its association with complex traits. Microbiome relationship matrix (MRM) allow considering the microbiota as a whole. Several methods have been proposed to ordinate these matrices; those differ on the metric used to account for the distance (or dissimilarities) between microbial communities (e.g. Euclidean, Bray-Curtis, χ^2). These distances account for alpha and beta diversity in different ways. Consensus on what method is the most appropriate hasn't been reached yet, and might depend on data singularities and the purpose of the study.

The aim of this study was to compare several microbiota relationship matrices, within a variance component estimation framework. Five ordination methods to build the MRM were tested: metric multidimensional scaling (MDS), detrended correspondence analysis (DCA), non-metric multidimensional scaling (NMDS), redundancy analysis (RDA) and constrained correspondence analysis (CCA). The log transformed and standardized relative abundances matrix described in Ross et al. (2013) was used as a benchmark matrix.

Simulated (n=1000) data were used to estimate variance components including phenotypes, genotypes and rumen microbiota information. Data were analysed considering two possible models. First, the genomic effect and the microbiota effect were included independently. Second, an interaction effect between the genomic and microbiota effects was added. All models were implemented within a Bayesian framework using the BGLR package in R. A total of 100 replicates were generated. Real data were analysed using the same models.

Similar or slightly better estimation of simulated h^2 (0.30) and m^2 (0.50) in the independent effects models resulted from ordination methods of MDS (0.307 and 0.493), RDA (0.307 and 0.501) and CCA (0.305 and 0.500) compared to the benchmark MRM

(0.304 and 0.480), while poor performance of the DCA (0.249 and 0.349) and NMDS (0.217 and 0.266) methods were obtained at estimating those parameters. The correlation coefficients between genomic estimated breeding values (GEBV) and true breeding values (TBV), from higher to lower, were: the obtained with the benchmark matrix (ρ = 0.633), CCA (ρ = 0.631), RDA (ρ = 0.624), DCA (ρ = 0.598), MDS (ρ =0.592) and NMDS (0.557). Likewise, correlations for predicted microbiota effect in the same order were: the benchmark matrix (ρ = 0.975), CCA (ρ = 0.966), RDA (ρ = 0.949), MDS (ρ = 0.845), DCA (ρ = 0.807) and NMDS (ρ = 0.517). Similar results, in terms of matrices performance, were obtained for the interaction effects model.

A real data set (n=70) was also analysed under the same frameworks. Low heritability estimates for feed efficiency (from 0.077 to 0.083) and microbiability (from 0.073 to 0.103) were observed; however, consistent values for the microbiability were obtained with the MRM that performed better in the simulations (from 0.073 to 0.077). Besides, high correlations (ρ > 0.85) between the genetic effect of the host and the phenotypes were obtained for all methods, as well as high correlations between the microbiota effect and the phenotypes for the RDA (ρ = 0.91) and CCA (ρ = 0.91) matrices.

Both models were compared using the deviance information criteria (DIC), the effective number of parameters (pD), and the posterior mean of the log likelihood (PostMeanLogLik), resulting in slightly lower values for the independent effects model (DIC: 183.9 to 189.3) than the interaction effects model (DIC: 187.5 to 191.7), those results indicate that it might be a relationship linking genotype-microbiome-phenotype which could be used in prediction of complex traits.

The analyses performed in this thesis suggest that canonical ordination methods of RDA and CCA to create MRM are preferred when whole microbiota information is included in the statistical models to analyse complex traits.

Keywords: feed efficiency, microbiability, heritability, ordination methods

INTRODUCTION

3. INTRODUCTION

3.1. *Background*

New discoveries in "omics" technologies have turned on the attention of the scientific community. Metagenomics has recently become a field of interest to worldwide researchers in human, animal and other biological systems. Whole genome shotgun metagenomics (**WGS**) has been used to analyse microbiome at a functional level. However, WGS approach is currently more expensive than amplicon-based sequencing methods like 16S ribosomal RNA (**16S rRNA**) analysis, which is commonly used to estimate relative abundance of microorganism taxa, and allows analysing microbiome at a compositional level (Malmuthuge and Guan, 2016). Both molecular approaches (WGS and amplicon-based sequencing methods) have increased the boundaries of microbial identification, previously limited to those, proportionally, few microbes identifiable by traditional laboratory cultures methods.

Microbiome effect on the estimation of host phenotypic complex traits has also been reported (e.g. Beaumont et al., 2016) leading to hypothesises that microbiome is a source of information to take into account when predicting difficult to measure traits.

There is evidence supporting that gut microbiome, is partially controlled by host genetic variation. Several species such as humans (Zoetendal et al., 2001; Blekhman et al., 2015), mice (Benson et al., 2010; McKnite et al., 2012), poultry (Zhao et al., 2013), pigs (Camarinha-Silva et al., 2017) and cattle (Roehe et al., 2016; Gonzalez-Recio et al., 2017) are among the studied superorganisms that consistently confirm the effect of host genetic variation controlling microbiome composition.

In livestock genetics, prediction of phenotypic complex traits of the host using its microbiome is a promising field. For instance, the microbiability or proportion of the phenotypic variance attributed to the microbiome variance (Difford et al. 2016) has previously shown higher values than narrow-sense heritability for feed efficiency and feed intake in pigs (Camarinha-Silva et al., 2017).

Microbiome is itself a holobiont organism affecting the complex trait (Benson, 2016), hence appropriate statistical fitted methods should be used to improve accuracy on the estimation of variance components, when involved on prediction of host phenotypic complex traits.

Two economic and environmentally relevant complex traits in cattle, that have previously been related to microbiome, are feed efficiency (Roehe et al., 2016) and methane emissions (Roehe et al., 2016; Tapio et al., 2017). Both traits are related to digestion efficiency; feed efficiency strongly determines the profitability of the productive system, whereas methane emissions also cause detrimental effects over environment by releasing a greenhouse gas 28 folds more harmful than carbon dioxide, in the global warming context.

A deeply review on this subjects will be addressed in this study, in order to understand the state of the art focusing on the development and evaluation of microbiome relationship matrices, innovative in prediction of complex traits.

3.2. *Definitions*

3.2.1. *Microbiota, metagenome and microbiome*

There are different definitions for the term "microbiome". This term was first coined, in humans, by Joshua Lederberg in 2001(Lederberg and McCray, 2001) to signify the "ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space". However, this meaning is often used interchangeably with "microbiota", which is confusing (Ursell et al., 2013).

According to the latter authors, there are differences between microbiota (the microbial taxa associated with humans) and microbiome (the catalog of these microbes and their genes), as supported by a more inclusive definition of microbiome (and microbiota) in Springer Nature (scientific journal) online portal, where is mentioned as: "all of the genetic material within a microbiota (the entire collection of microorganisms in a specific niche, such as the human gut). This can also be referred to as the metagenome of the microbiota." (Springer Nature, 2017).

The last definition of microbiome (and microbiota) leads to define metagenome, which would be the product of metagenomics analysis. Metagenomics is defined as the direct genetic analysis of genomes contained within an environmental sample (Thomas et al., 2012), the analysis of DNA from microbial communities in environmental samples without prior need for cultivating clonal cultures (Oulas et al., 2015), and has also been defined as the study of overall genetic material from a microbiome (Malmuthuge and Guan, 2016). Based on these definitions, microbiome and metagenome could be used as

synonyms, although other authors define microbiome as the microbiota, their genes and its surrounding environmental conditions while metagenome is defined strictly as the collection of genomes and genes from the members of a microbiota (Marchesi and Ravel, 2015).

In livestock, the definition of microbiome is applied like in Springer Nature's definition, but focused on animal hosts, however different adjustments are done to fit within specific areas of study, for instance, in gut microbiome studies, authors usually define gut microbiome as the total genome of microbiota present in the gut (Malmuthuge and Guan, 2016). For our interests, the definition of Marchesi and Ravel, (2015) will be used in this document.

3.2.2. *Operational taxonomic units (OTUs)*

Another relevant definition in microbiome research is operational taxonomic units (**OTUs**); in the context of microbiome studies, OTUs are obtained by comparing shared similarities of sequences from molecular analysis between each other (*de novo* sequences) or compared to previously established and freely available data frames. Some degree of sequence divergence is typically allowed (95%, 97%, or 99%) and the resulting cluster of nearly-identical tags (genomes are assumed as identical) is referred to as an Operational Taxonomic Unit or sometimes phylotype (Morgan and Huttenhower, 2012).

The degree of similarity between sequences is calculated as twice the number of base pairs that match, divided by the total number of bases in both sequences being compared. For instance, if 48 base pairs are equal when comparing 50 base pairs, the similarity between those sequences is 96% ((2*48)/100), sequences are usually compared between *de novo* sequences (Figure 1), or against a catalogue of sequences in widely recognised data frames, leading to taxonomic classification.

According to (Almeida et al., 2016), information of an extensive 16S rRNA catalogue compiling over two million distinct entries are available through three main databases RDP (Cole et al., 2009), Silva (Quast et al., 2013) and Greengenes (DeSantis et al., 2006). These OTUs are used to classify sequences into its corresponding taxonomic groups, previously established in the data bases.

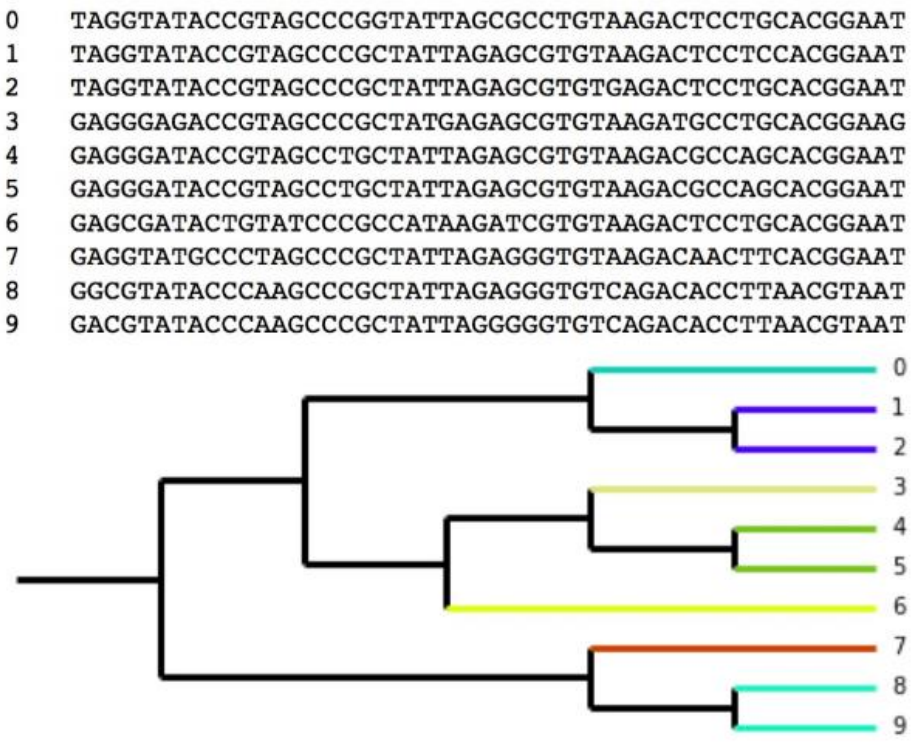

Figure 1.Ten de novo sequences clustered at 96% similarity, colours on the branches denote OTUs (seven OTUs total, three with two sequence representatives and four composed of a single sequence).
Taken from (Gibbons, 2015).

3.3. *Main molecular technics of analysis in microbiome studies*

3.3.1. *16S rRNA amplicon*

New standards for identifying microbial isolates began to be developed in the 80's, prior taxonomic classification were performed by comparison of morphologic and phenotypic description of type strains. This new method showed that phylogenetic relationships of bacteria could be determined by comparison of a conserved region of the genetic code. The genes that code for 16S rRNA (in the small subunit) were among the best candidates because of its conservativeness in the bacterial genome (derivated from the importance of those genes in the cell function) which allow them to be used for taxonomic classification (Clarridge, 2004). The sequence analysis of 16S rRNA of distinct phylogenetic groups revealed the presence of one or more short specific sequences denominated "signature oligonucleotide" (Woese et al., 1985) present in all (or most) members of a determined phylogenetic group and never (or unfrequently) present in other groups, nor even in the nearest ones. These particularities of the 16S rRNA

analysis allowed to identify each bacteria within a given group (Rodicio and Mendoza, 2004).

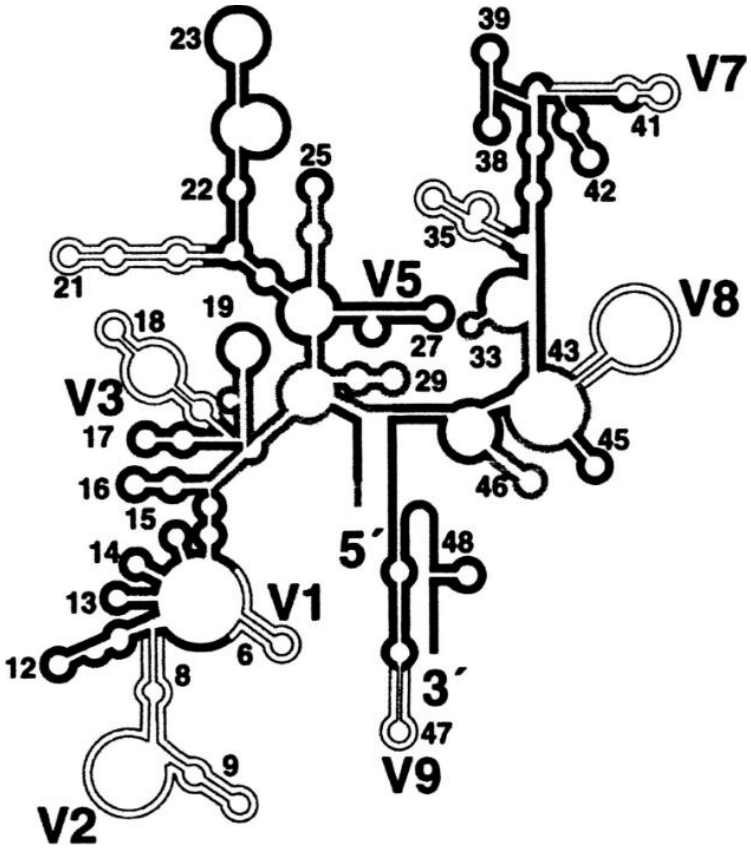

Figure 2. Secondary-structure model of the 16S rRNA (double lines indicate variable or hypervariable; black lines indicate highly conserved; V1 to V9 indicate major variable regions, numbers 6 to 48 indicate loops).
Taken from (Tortoli, 2003).

The 16S rRNA (Figure 2) genes are essential in prokaryotes and are in at least one copy in a genome (Wang and Qian, 2009), its extended presence is one of the main reasons why 16S rRNA gene sequence has been the most used DNA region for taxonomic purposes (Janda and Abbott, 2007). There exist up to nine known hypervariable regions (Figure 3) in the 16S rRNA gene (Chakravorty et al., 2007), which allow a comprehensive classification of the microbiome composition and thus of the relative abundance of all taxa.

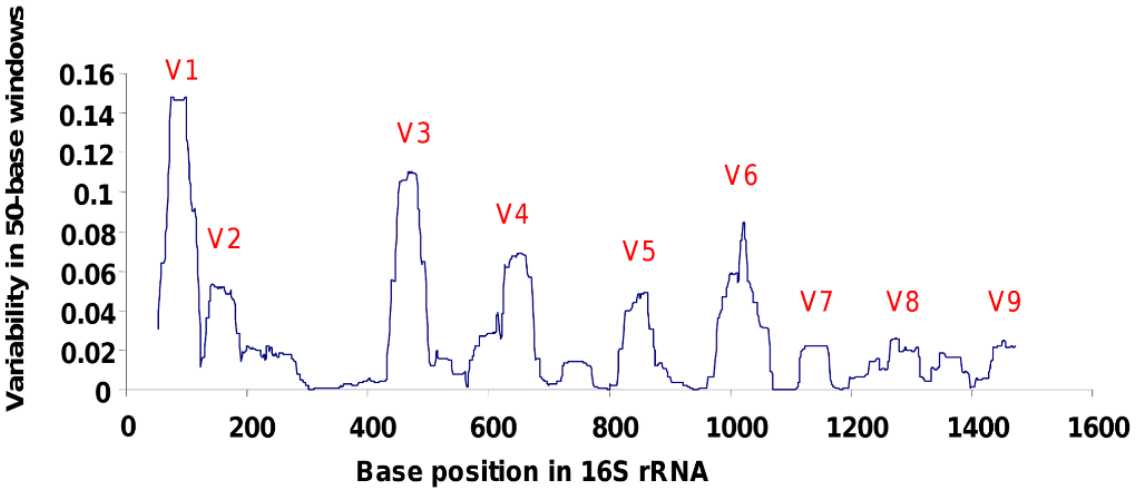

Figure 3. Hypervariable regions (V1 to V9) within the 16S rRNA gene in *Pseudomonas*. Taken from (Bodilis et al., 2012).

Microbiota identification using 16S rRNA methodology is often restricted to bacteria and archaea (Janda and Abbott, 2007), and provided relatively high accuracy at genus level identification, but it provides common misclassification at species level (Poretsky et al., 2014).

Although 16S rRNA amplicon is not the gold standard to assess functional capabilities of the microbiome, one way to infer functional capabilities of a microbiome from amplicons involves the correlation between phylogenetic trees and clusters of genes shared between taxa (Langille et al., 2013). A comprehensive explanation of this association is described in Jovel et al. (2016). A software (PICRUSt) was developed by these authors to construct a phylogenetic tree from a gene database (e.g. Greengenes), then genes are assigned to nodes in the previously constructed tree if sequenced genomes are available, or predicted by algorithms if not available. After that, sequences from OTUs which are associated with the gene database identifiers are normalized by 16S rRNA gene copy number and mapped to the associated gene database identified in the reference tree. The final product is an annotated table of gene counts per sample that can be linked to the Kyoto encyclopedia of genes and genomes (**KEGG**) orthology (**KO**) numbers or other orthologous protein catalog.

The analysis of the 16S rRNA amplicon gene has been the most used microbiome analysis by the Human Microbiome Project (**HMP**) to compile most of the data (Ranjan et al., 2016).

The HMP is a consortium established in 2008 and funded by the common fund of the National Institutes of Health (United States of America). Its goals are: "(1) to take

advantage of new, high-throughput technologies to characterize the human microbiome more fully by studying samples from multiple body sites from each of at least 250 ‘‘normal’’ volunteers; (2) to determine whether there are associations between changes in the microbiome and health/disease by studying several different medical conditions; and (3) to provide both a standardized data resource and new technological approaches to enable such studies to be undertaken broadly in the scientific community.” (The NIH HMP Working Group, 2009).

3.3.2. *Whole genome shotgun metagenomics*

Whole genome shotgun metagenomics (**WGS**) allows assessing both taxonomic composition and diversity of microbial communities without the limitation of target and amplify a specific gene, giving robust estimates (Poretsky et al., 2014) and accurately defining taxa at the species level (Ranjan et al., 2016), if the genome has previously been entered in the databases. Moreover WGS provides a quickly microbiome analysis without the cultivation bias or variation related to PCR amplification anomalies or primer selection (Tapio et al., 2017). There is evidence supporting that WGS has multiple advantages like increased detection of diversity, increased prediction of genes and improved accuracy of species detection, compared to 16S amplicon method (Ranjan et al., 2016). The WGS analysis is also a more accurate method to elucidate functional capabilities of the microbial community than 16S rRNA technique, however WGS is a less cost-effective method than 16S rRNA sequencing for taxonomic purposes, therefore, several software have been developed to predict functional features of the microbiome from 16S rRNA outputs (Aßhauer et al., 2015).

A schematic workflow of WGS steps and tools are shown in Figure 4.

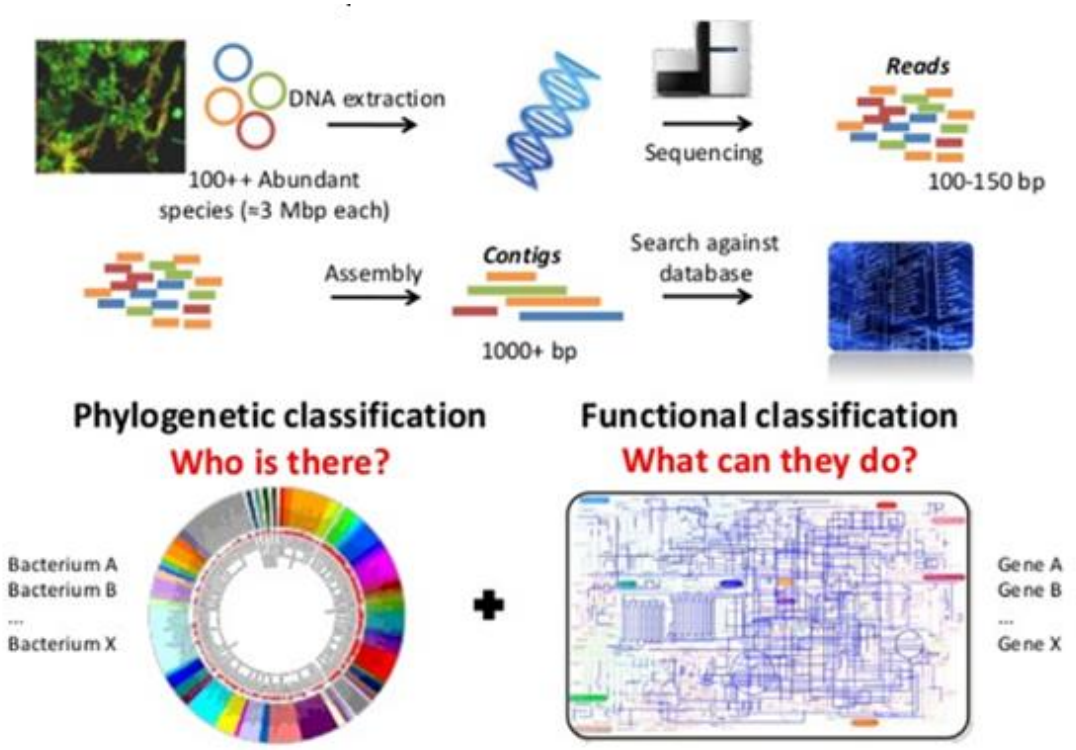

Figure 4. Workflow diagram of basic metagenomics steps and tools currently in practice for Whole Genome Shotgun Metagenomics (WGS).
Taken from: https://www.cd-genomics.com/Metagenomic-Shotgun-Sequencing.

When performing WGS, information about functionality of the community can be obtained from the complete sequences of protein coding genes in the sequenced genomes. An assembly of shorter reads into genomic *contigs* and orientation of these into *scaffolds* is often performed to provide a more compact and concise view of the sequenced community under investigation (Oulas et al., 2015).

Direct assessment of the functional attributes of the microbiome can be done using WGS approach (Knight et al., 2012). Regarding patterns of KO abundance, the concordance between 16S rRNA and WGS depends on the pathway under consideration, a clear example of this phenomenon is shown in Jovel et al. (2016) who compared two KEGG reference pathways (at the KO level), glycolysis and fatty acid biosynthesis using both techniques, and obtained high ($r = 0.88$) and medium ($r = 0.52$) associations values, for correlations between 16S rRNA and WGS methods, respectively (Figure 5). However, it is known that a more reliable assessment of functional profiling of the microbiome is obtained with WGS, compared to 16S rRNA (Franzosa et al., 2015).

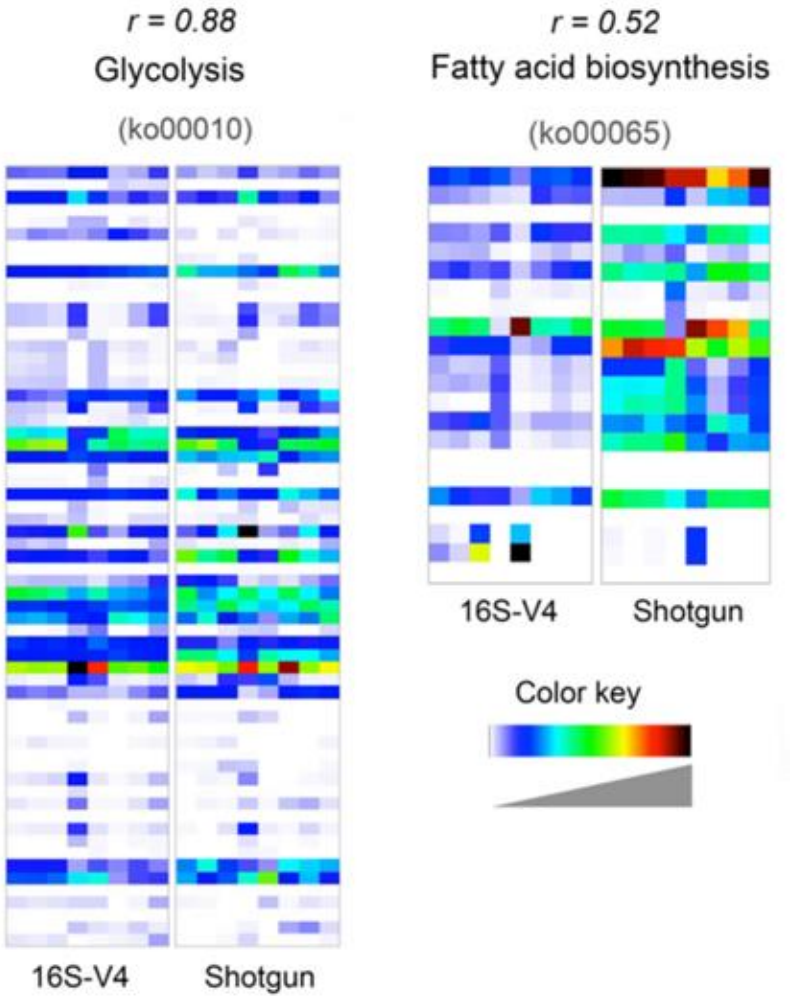

Figure 5. Two KEGG pathways at the gene (KEGG orthology, KO, groups) level, comparing 16S rRNA (16S-V4) vs WGS (Shotgun). On top of each heatmap pair, the Pearson correlation coefficient for relative abundance of KOs derived with each method.
Taken from Jovel et al. (2016).

3.4. *Microbiome effect on host phenotypic complex traits*

Research in human microbiome has revealed associations between imbalances in the gut microbiome and a variety of host phenotypes (Jovel et al., 2016), including obesity (Turnbaugh and Gordon, 2009; Turnbaugh et al., 2009), inflammatory bowel disease (Norman et al., 2015; Imhann et al., 2016), type II diabetes (Hartstra et al., 2015), fatty liver disease (Arslan, 2014), among other disorders.

Beaumont et al. (2016) found that heritable components of the human fecal microbiome were significantly associated with visceral fat, which is a cardio-metabolic disease risk factor. They state that their findings, association of *Oscillospira* and *Blautia* to visceral fat mass (VFM), support the hypothesis that microbiome plays a role as a

biomarker of cardio-metabolic disease risk. They also mentioned that the microbiome may be one potential source contributing to missing heritability in obesity, although this has not been proved yet. Hall et al. (2017) also mentioned *Blautia* as an over-represented microbial specie in human obesity (Figure 6).

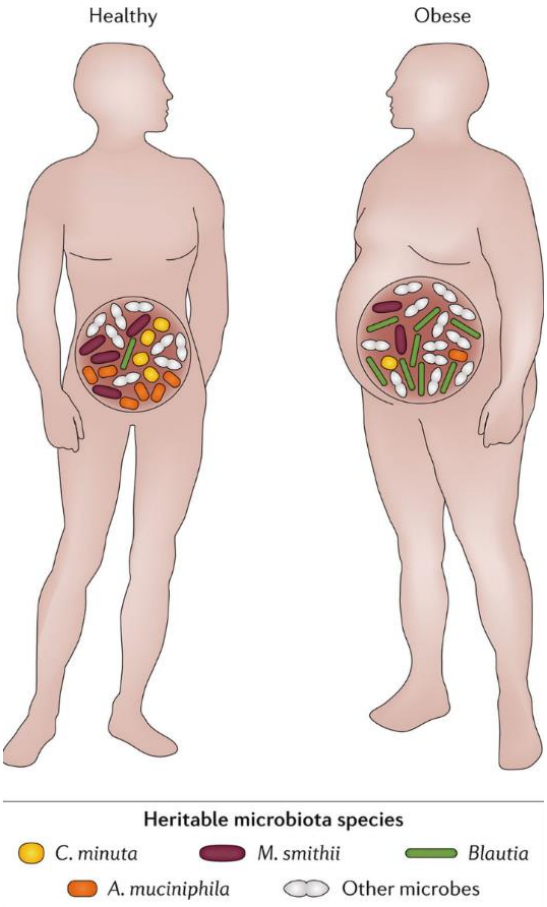

Figure 6. Heritable species partially responsible for microbiome composition in obesity. Relative abundance of Christensenella minuta, Akkermansia muciniphila and Methanobrevibacter smithii, are consistently under-represented in obesity, and Blautia, is over-represented in obesity.
Taken from (Hall et al., 2017).

(Bonder et al., 2016) identified several associations of genetic variants and the human gut microbiome composition and their function were identified. For instance, they observed association between the abundance of *Lactococcus* bacteria and a single nucleotide polymorphism (**SNP**) associated with body fat distribution affecting the expression of the nearby ZNRF3 gene, which is a gene that acts as a tumour suppressor in gastric cancer. Association of the microbiome with several GWAS SNPs suggests that microbiome could mediate some link between host genetics and immunological and metabolic phenotypes.

In dairy cattle, microbiome has also been associated to several phenotypic traits. Lima et al. (2015) used multivariable models to identify associations between bacterial taxa and milk production and composition. They used microbiome information to develop a metagenome-based prediction, and obtained a high association with milk yield and milk composition. Li and Guan (2017) analysed rumen microbiome to associate relative abundance of specific taxon to feed efficiency in beef cattle. They argue that their results obtained by comparative analysis of extreme phenotypes for feed efficiency (efficient, n=10; inefficient, n=10) showed that "three bacterial families (*Lachnospiraceae*, *Lactobacillaceae*, and *Veillonellaceae*) tended to be more abundant in inefficient animals ($P < 0.10$), and one archaeal taxon (*Methanomassiliicoccales*) tended to be more abundant in high-feed-efficiency (efficient) cattle ($P < 0.10$)". However due to the low number of animals and the high *P* value they used, those results are to be supported by studies with larger sample size. A meta-analysis by Guyader et al. (2014) from 28 different experiments in ruminants, studying the relationship between methane emissions and protozoal numbers, showed a strong linear significant relationship ($R^2 = 0.90$). In ruminants these findings also provide evidence for a potential interaction between microbiome and host phenotypic traits.

3.5. *Host genetic variation effect on gut microbiome*

In the last years, there is increasing evidences that support the hypothesis of host genetic effect over gut microbiome. For instances, in humans, Zoetendal et al. (2001) showed that the microbiome of monozygotic twins were more similar than those of marital couples or unrelated individuals, which suggest that host genetic variation partially control gut microbiome. Likewise, a study comparing concordance rate between monozygotic (MZ) and dizygotic (DZ) twins, for *Methanobrevibacter*, showed higher concordance in MZ twins (Hansen et al., 2011). Another twin study also found grater similarities for MZ compared to DZ in *Lachnospiraceae* and *Ruminococcaceae* (Firmicutes) bacterial families, not so for the *Bacteroidaceae* bacterial family (Goodrich et al., 2014). They showed heritability estimates for *Blautia* and *Methanobacteriaceae* of 0.34 and 0.22 respectively, which also indicates a heritable component of the microbial composition. In this study, the abundance of specific taxa were more highly correlated within MZ compared to DZ twins, corroborating that host genetics influence the human gut microbiome composition. A reanalysis of two previously published data sets of twins

population from Missouri, USA (Turnbaugh et al., 2009; Yatsunenko et al., 2012), also showed significantly greater mean twin-pair intraclass correlation coefficients of OTU abundances for MZ compared to DZ.

Another study estimating heritability in the human fecal microbiome using twins can be found in Beaumont et al. (2016). These authors showed differences in heritability (Wilcoxon rank test, $P < 2.2 \times 10^{-16}$) between specific phenotypic-associated OTUs and overall OTUs: heritability estimates from 97 adiposity-associated OTUs was 0.16, while overall average heritability over all OTUs was 0.07. Furthermore, the average unique environmental component of the 97 adiposity-associated OTUs was 0.79, while overall average for all OTUs was 0.93 also differing by Wilcoxon rank test ($P < 2.2 \times 10^{-16}$), suggesting that host genetics impacts fecal microbiome. Based on these results, these authors inferred that host genetics affects the variation of microbes associated to obesity.

There is also evidence of host genetic variation affecting gut microbiome in cattle. For instance, Weimer et al. (2010) found that bacterial community composition of the rumen differed between individual cows, under the same diet, using correspondence analysis and automated ribosomal intergenic spacer analysis. Another study using the same statistical analysis found differences between dairy cows for the bacterial community composition (Welkie et al., 2010). King et al. (2011) compared microbiome from Holstein and Jersey cows located within the same herd and managed under the same diet and environmental conditions. They found twenty OTUs common in both breeds while 23 and 18 OTUs were found only in Holstein and Jersey, respectively. These authors concluded that the differences they observed may be due to differences in host breed genetics. Recently, another study assessing host genetic influence on rumen microbial methane production and feed conversion efficiency, used sire progeny groups and found consistent ranking of the sire progeny groups (overall and within diet) based on methane emissions or relative archaeal abundance, proposing a genetic control of the host for these traits, and suggesting that rumen microbial gene abundance could be used as a predictor for complex traits (Roehe et al., 2016).

3.6. *Statistical approaches*

3.6.1. *Common statistical analysis on host-microbiome studies*

Simple statistical tools for the analysis of the relationship between complex traits and the host microbiome have been used frequently. For instance, Pearson correlation analysis were implemented in a study associating productive and physiological

parameters in dairy cows and its ruminal microbiome (Jami et al., 2014), authors found positive correlations between the ratio of Firmicutes-to-Bacteroidetes and daily milk fat yield (ρ = 0.72, P = 2x10^{-3}), and between *Prevotella* genus and milk fat yield (ρ = -0.69, P = 5x10^{-3}), among other Pearson correlation results. The correlation coefficient (ρ) is used to measure the strength of the linear relationship between two variables and is used when it is not clear which one is the independent variable, its values range between -1 and 1. For $\rho > 0$, the two variables have a positive correlation, whereas the two variables have a negative correlation for $\rho < 0$. The value $\rho = 1$ or $\rho = -1$ indicates an ideal or perfect linear relationship, and $\rho = 0$ means that there is no linear association (Kaps and Lamberson, 2004). The coefficient of correlation (ρ) is defined in equation 1 as:

$$\boldsymbol{\rho} = \frac{\boldsymbol{\sigma_{xy}}}{\sqrt{\boldsymbol{\sigma_x^2 \sigma_y^2}}} \qquad [1]$$

where:

σ_{xy} = covariance between x and y

σ_x^2 = variance of x

σ_y^2 = variance of y

Variables x and y are assumed to be random normal variables jointly distributed with a bivariate normal distribution (Kaps and Lamberson, 2004).

Another common statistical approach, which has also been used to associate microbiome to complex traits, is linear regression. Roehe et al. (2016) regressed methane emission on the relative abundance of different microbial genes of cows under the same diet and found that the slope of the regression were similar among diets. Linear regression comprises procedures designed to analyse statistical associations among variables defining one variable as dependent while others are defined as independent variables. Simple linear regression is defined when changes of the dependent variable are described by linear relationship of only one independent variable, while multiple linear regression procedures are used when changes in two or more independent variables explain the change of the dependent variable (Kaps and Lamberson, 2004). Multiple linear regression is another frequent approach to analyse microbiome association with complex traits.

A t-test can be used when two small samples are compared and the number of observations in both samples is the same. This statistic test is described in equation 2:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} \quad [2]$$

For instance, the study by (Mao et al., 2015) used an independent t-test to compare the sample type (mucosal tissue vs. digesta) effects on bacterial prevalence and found that the abundance of certain OTU in the ruminal digesta were significantly higher than in the rumen epithelium.

Some authors emphasise that appropriate correction for multiple hypothesis testing (e.g. Bonferroni correction for t-test based analyses) are to be applied in metagenomics studies. They advise so because the number of species or gene functions are often more than the number of samples taken (Thomas et al., 2012). Bonferroni correction is frequently used in multiple testing to compensate the likelihood of commit type I error by testing each individual hypothesis at an alpha level of α/n, where α is the desired overall alpha level and n is the number of hypothesis. For instance consider a case where you have chosen an overall alfa value of 0.05 and 30 hypothesis test are to be performed, then the significance cut-off value for each hypothesis would be 0.05/30 = 0.0017 (Glickman et al., 2014).

Bonferroni correction is suitable when not many hypothesis are involved (e.g. fewer than 50) and when variables are independent. One limitations of this correction is that is overly conservative when much more hypothesis are to be tested. Other correction methods like false discovery rate are more appropriate when there are dependencies among variables, and more hypothesis are to be tested (Khatri and Draghici, 2005).

According to (Xia and Sun, 2017), classical statistical test are available for hypothesis testing in microbial taxa, which can be conducted by comparing alpha diversity (within samples) and beta diversity (between samples) indices. Those authors mentioned that two-sample t-test and its nonparametric counterpart Wilcoxon rank-sum test (also called Mann-Whitney U) were widely used in microbiome studies to comparing continuous variables between two groups. They also mentioned that when comparing more than two groups, the one-way ANOVA or its non-parametric equivalent, the Kruskal-Wallis test are appropriate, depending on fulfil or not of normality assumption, respectively.

Zero inflated models (e.g. Zero-Inflated Poisson (**ZIP**), Zero-Inflated Negative Binomial (**ZINB**), hurdle model or Zero-Inflated Gaussian distribution mixture model

(**ZIG**) and negative binomial models are often used in microbiome studies because of presence of many zeros and overdispersed data, respectively (Xia and Sun, 2017).

Regarding multivariate statistical tools in microbiome studies, Xia and Sun (2017) mentioned multivariate analysis of variance with permutation (**PERMANOVA**), analysis of group similarities (**ANOSIM**), multi-response permutation procedures (**MRPP**), and Mantel's test (**MANTEL**), as tests of among-group differences for analysing microbiome data. Also mention Dirichlet multinomial model (based on difference between mean comparison and variance comparison/dispersion) and UniFrac distance metric family (weighted or unweighted by relative abundance of taxa) which is based in phylogenetic distances, as recently developed statistical methods more suitable for microbiome studies. Those authors also mentioned ANOVA-like differential express (**ALDEx** and **ALDEx2**) and analysis of composition of microbiomes (**ANCOM**) for compositional analysis of microbiome data, and remark that currently, microbiome researchers are shifting their emphasis from correlation to causality, as a better approach of host-microbiome studies.

3.7. *Opportunities*

Regarding greenhouse gas production from agriculture, some authors (Ross et al., 2013) argue that it is possible that metagenomic predictions could aid in its reduction, if increased accuracy in the prediction of enteric methane production level is achieved.

Microbiome-host interactions and prediction of phenotypic complex traits of the host using its microbiome and genetic parameters are among the topics that remain under study, but still to be deciphered. Genetic selection using estimation of additive breeding values are potentially the most sustainable way of reducing enteric methane emission from ruminant (Pickering et al., 2015)

Complex traits like feed efficiency and methane emissions can be included in genetic evaluations of ruminants in order to obtain more cost effective animals while diminish environmental impact. Because of difficulty to measure methane emissions, some correlated traits are to be evaluated.

Residual feed intake (**RFI**) is an associated trait to methane emissions, and is estimated as the difference between net energy intake and calculated net energy requirements for maintenance, this latter takes into accounts body weight and fat and protein corrected milk yield; both methane emissions and RFI are heritable traits with narrow sense heritabilities estimated of 0.35 and 0.40, respectively (de Haas et al., 2011).

Positive genetic correlations (0.44) between RFI and methane output has been reported in beef cattle (Nkrumah et al., 2006) and microbiome has been also associated to methane emissions and RFI (Roehe et al., 2016)

If these results are consistent among livestock populations and the microbiome is persistently different between animals and also heritable in some extent, it could represent an opportunity to use genetic selection of the livestock to select animals with desirable microbiomes for these traits (Tapio et al., 2017). Consequently, application of statistical methods that might elucidate causative relationships between genotype, microbiome and these relevant complex traits are to be tested as a challenging and ambitious gold that might have a pertinent impact in livestock and environmental issues.

This study aims to develop and test some statistical approaches to evaluate feed efficiency in dairy cattle accounting for the host genotype and the microbiome simultaneously.

OBJECTIVES

4. OBJECTIVES

4.1. *Main Objective*

The objective of this research was to develop and test some statistical approaches to evaluate feed efficiency in dairy cattle accounting for the host genotype and the microbiome simultaneously.

4.2. *Specific objectives*

1. Compare ordination methods to construct microbiome distance (or dissimilarity) matrices through simulation.
2. Estimate the proportion of phenotypic variance for feed efficiency explained by microbiome, considering the interaction between the microbiome and the host genetics.

MATERIALS AND METHODS

5. MATERIALS AND METHODS

This study was developed in the Department of Animal Breeding and Genetics of the National Agricultural and Food Research and Technology Institute (INIA), Madrid, Spain.

5.1. *Data*

Variance components, heritability and microbiability, regarding feed efficiency in dairy cows, was estimated using two data subsets: either simulated or real data.

5.1.1. *Real data*

All cows belonged to "BLANCA de los Pirineos" located in Lleida, Spain. Three data frames containing phenotypic performance data, genotypic information and relative abundances of OTUs from 70 Holstein cows were used. The phenotypic data set enclosed information for each cow regarding: ID, sire, born date, calving date, days in milk, parturition, pen, dry matter intake, residual feed intake, feed efficiency, milk yield, fat yield, protein yield and body weight. Microbiota information was obtained from independent samples of the rumen content of 70 Holstein cows, previously collected via intra-oesophagic hose extraction, and sequenced using Illumina Miseq for the hypervariable region V3-V4 of the16S rRNA amplicon.

The final data frame of relative abundances of OTUs contained information about the 92 OTUs in the core (RA > 0.1%) rumen microbiome in rumen content for each cow (host).

5.1.2. *Simulated data*

Simulations were generated using the observed data structure in the real data set. A data frame of 1000 genotyped Holstein animals with allelic variants for 9244 SNPs was used. Additive genetic effects were determined by 1000 QTL that were simulated as randomly distributed along the genome. QTL effects were generated based on a normal distribution (N ~ (0, 1)). True breeding values ($\boldsymbol{u}$) were calculated by summing all QTL effects and were subsequently scaled to a realized genetic variance of σ_u^2.

Simulation of the OTU effects started with a symmetric co(variance) matrix obtained from the real relative abundance of the 92 OTUs from the rumen content samples of the 70 animals. The correlation elements lower than 0.20 were set to 0. The resulting

symmetric matrix was then converted to the nearest positive-definite matrix to ensure it was a valid covariance matrix.

Then, 1000 vectors of 92 OTUs were generated using the variance covariance matrix described above, using Cholesky factorization: First, relative abundance of OTUs were sampled from a normal distribution (N~(0.5, 0.1)). This matrix was then multiplied by the cross product of the Cholezky factorized matrix of the positive-definite matrix of (co)variances calculated before. This creates a final matrix of 1000 simulated microbiotas with 92 OTUs each. Any OTU resulting in negative values for the RA was set to zero.

Once the relative abundance of the simulated OTUs was generated, the microbiome effect ($\boldsymbol{m}$) for each animal was simulated as follows: 50 OTUs were randomly selected out of the 92 OTUs. Then, an effect ($\boldsymbol{\beta_j}$) was sampled from a normal distribution (N ~ (0, 1)) and assigned to each of the selected 50 OTUs. The m_i was then simulated as follow:

$$\boldsymbol{m_i} = \sum_{\boldsymbol{j}} \boldsymbol{\beta_j} \times \boldsymbol{OTU_{ij}} \quad \textbf{[3]}$$

Where $\boldsymbol{\beta_j}$ is the effect of OTU_j sampled from a N ~ (0, 1) and $\boldsymbol{OTU_{ij}}$ is the relative abundance of OTU *j* in animal *i*. The resulting $\{m_i\}$ were scaled to have a variance of σ_m^2.

Phenotypes were finally simulated assigning a residual variance to obtain a heritability and a microbiability of 0.30 and 0.50, respectively. Phenotypes were respectively simulated for an independent effects model and for an interaction effect model as:

$$\boldsymbol{y_i} = \boldsymbol{\mu} + \boldsymbol{u_i} + \boldsymbol{m_i} + \boldsymbol{e_i} \quad \textbf{[4]}$$

And

$$\boldsymbol{y_i} = \boldsymbol{\mu} + \boldsymbol{u_i} + \boldsymbol{m_i} + \boldsymbol{u_i} \times \boldsymbol{m_i} + \boldsymbol{e_i} \quad \textbf{[5]}$$

Where $\boldsymbol{\mu}$ is the population mean, $\boldsymbol{u_i}$ is the genomic effect, $\boldsymbol{m_i}$ is the microbiome effect, $\boldsymbol{u} \times \boldsymbol{m_i}$ is a genomic-microbiome interaction effect and $\boldsymbol{e_i}$ is the residual error.

5.2. *Genomic relationship matrix*

A genomic relationship matrix (**GRM**), between individuals *j* and *k* was constructed following method 2 of VanRaden (2008) and Yang et al. (2010) with the following formula:

$$\boldsymbol{GRM_{jk} = \frac{1}{N}\sum_{i=1}^{L} GRM_{ijk} = \frac{1}{N}\sum_{i=1}^{L} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}} \quad \textbf{[6]}$$

Where $\boldsymbol{x_{ij}}$ refers to the AA, Aa and aa SNP genotypes, coded as 2, 1, and 0, respectively, of individual *j* or *k* at locus i (*i* = 1, …, *L*), with *L* being the number of SNP (9244) and p_i being the allele frequency in the whole genotyped population. This matrix depicted genetic relationships within individuals (diagonal elements) and between individuals (out-diagonal elements). The Gmatrix.f90 code (Legarra, A. personal communication) in Fortran® was used for its construction, resulting in a 70 x 70 square matrix for real data and a 1000 x 1000 square matrix for the simulated set, where elements in the diagonal were close to the value one (describing within relationship) whereas elements out of the diagonal variated according to genomic relationship between individuals.

5.3. *Microbiome relationship matrix (MRM)*

As the genomic relationship matrix is constructed, to establish the relationship between genotypes of different individuals, likewise a microbiome relationship matrix (**MRM**) can be built to associate microbiome of a given community of microorganisms between hosts. There are many methods to build a MRM matrix, for instance, Ross et al. (2013) mentioned that "the relationship between samples can be described by a matrix G = XX /m", where "metagenomic profiles in a group of samples are defined as an *n* x *m* matrix X with elements $\{x_{ij}\}$, being the log tansformed and standarised count for sample *i* for contig *j*, with *n* samples and *m* contigs". There are other ways to ordinate a MRM matrix, understanding ordination as the arrangement of units in some order (Goodall, 1954), to represent objects as points along one or several axes of reference (Legendre and Legendre, 1998). There are packages that allow ordination of matrices, for instances, the Phyloseq Package can be used to perform ordination in R software.

Several methods of ordination in R (MDS, DCA, NMDS, RDA and CCA) were compared to previously published method of Ross et al., (2013) in order to disentangle the most appropriate MRM to be used in further statistical analysis according to our interests.

5.3.1. Metric Multidimensional Scaling (MDS)

The MDS, also known as Principal Coordinate Analysis (PCoA), uses a linear (Euclidean) mapping of the distance or dissimilarities between objects onto the ordination space (i.e. projection in a Cartesian space), and the algorithm attempts to explain most of the variance in the original data set. This method uses any dissimilarity measure and thus specific association coefficients that better deal with the problem of the presence of many double zeros in data sets. This should be consistent to analyse data of many OTUs only present in few samples leading to value of cero in many samples. MDS does not provide a direct link between the components and the original variables making more difficult the interpretation of variable contribution (Ramette, 2007).

MDS components are complex functions of the original variables depending on the selected dissimilarity measure, and not on linear combinations of the original variables as in PCA. The selection of the distance measure is thus very important, and subsequent transformation of the data to correct for negative eigenvalues is sometimes necessary. Although there is no direct, linear relationship between the components and the original variables, it is still possible to correlate object scores on the main axis (or axes) with the original variables to assess their contribution to the ordination (Ramette, 2007).

MDS takes a symmetric matrix of distances (D) of any type among replicates and produces corresponding Cartesian (Euclidean) coordinates for each replicate which, in the full-dimensional principal coordinate space, preserve the original distances calculated among replicates (Gower, 1966).

The procedure of MDS is summarize from Legendre and Anderson (1998) as follow:

1) Transform the symmetric matrix of distances D of elements $\{d_{ij}\}$, $i = 1, \ldots, N$ and $h = 1, \ldots, N$, where N = total number of replicates, into a new matrix A of elements $\{a_{ih}\}$ by means of the following equation:

$$a_{ih} = -\frac{1}{2} d_{ih}^2 \quad [7]$$

2) Center the values in matrix A by its rows and columns, transforming it into matrix Δ_1 of elements $\{\delta_{ih}\}$ by means of the following equation:

$$\delta_{ih} = a_{ih} - \bar{a}_{\imath} - \overline{a_h} + \bar{a} \quad [8]$$

where $\bar{a}_i$ = average of row i, $\bar{a}_h$ = average of column h and $\bar{a}$ = average of entire matrix A.

3) Compute the eigenvalues and eigenvectors of matrix Δ_1.

4) To obtain principal coordinate axes, scale the eigenvectors to the square root of their respective eigenvalues. For the special case of the Euclidean distance, if there are fewer variables (species) than there are replicates in Y, then (1) the maximum number of principal coordinates is the number of variables in the original matrix, and (2) the principal coordinates are the same as principal components. For metric distance measures (such as Euclidean or chi-square distances), axes determined using MDS will preserve all of the original distances, D.

5.3.2. Detrended Correspondence Analysis (DCA)

The correspondence analysis was proposed by Hirschfeld and Wishart (1935) in a paper entitled "Connection between correlation and contingency", as a multivariate statistical technic to analyse categorical data similarly as continuous variables are analysed in Principal Component Analysis. The correspondence analysis method is widely used for analysing cross tabular data in the form of numerical frequencies, resulting in a plot which allows interpretation and understanding of the data (Greenacre, 2017).

Correspondence analysis, also called reciprocal averaging, can be calculated beginning with a matrix of n rows of samples and *p* columns of taxa, following a reciprocal averaging approach as follow:

$$y_j = \frac{\sum_{i=1}^{n} a_{ij} x_i}{a_{+j}} \quad \textbf{[9]}$$

Where y_j is the score for taxon *j*, a_{ij} the abundance of a taxon *j* in sample *I*, x_i is an arbitrarily chosen score for each sample and a_{+j} is the total abundance for the taxon across all samples. Then the taxon scores are used to calculate a new set of sample scores using the same procedure as follow:

$$x_i = \frac{\sum_{j=1}^{p} a_{ij} y_j}{a_{i+}} \quad \textbf{[10]}$$

Sample scores are centered and standardized such that their mean is zero and their variance is one

$$\sum_{i=1}^{n} a_i + x_i = 0 \quad \textbf{[11]}$$

and

$$\sum_{i=1}^{n} a_i + x_i^2 = 1 \quad \textbf{[12]}$$

This procedure of alternately calculating sample and taxon scores is repeated until the scores stabilize, producing the correspondence analysis axis I scores for both samples and taxa. The DCA is a methodology developed by Hill and Gauch (1980) to perform an ordination method that corrects for the two major problems of "Correspondence Analysis", the "arch effect" and the distortion of relative distances between samples (and species) on its axes. In DCA, after a correspondence analysis is performed, there are several approaches to detrend and rescale axes. One approach is to divide the axis into an arbitrary number of equal length segments and within each segment, the scores on the next higher order axis are re-centered such that the mean is zero these procedure causes, if there is arch present, flattened of the lower order axis. The process to detrend is sensitive to the number of segments used, usually a value of 26 has produced acceptable results, a sliding moving average window is the method performed by the algorithm to detrend and

rescale axes. The rescaling of an axis is accomplished by equalizing the weighted variance of taxon scores along the axis segments (Holland, 2008).

The "arch effect" in correspondence analysis is a mathematical artifact, corresponding to no real structure in the data, which arises because the second axis of correspondence analysis is constrained to be uncorrelated with the first axis, but not constrained to be independent of it, and for the axes to be separately interpretable, they need to be independent. Distortion of relative distances between samples (and species) on its axes, occurs when samples differing ecologically by an identical amount show different distances in the ordination. Correspondence analysis with detrending, followed by standardization to unit within-sample variance combine to characterize the method of DCA (Hill and Gauch, 1980). Even DCA corrects the "arch effect", results obtained with DCA vary depending on the number of segments used to remove the arch effect and should be avoided when analysing data that represent complex ecological gradients (Legendre and Legendre, 1998).

5.3.3. Non-metric Multidimensional Scaling (NMDS)

The NMDS algorithm ranks distances between objects, and uses these ranks to map the objects nonlinearly onto a simplified, two dimensional ordination space preserving their ranked differences, instead of the original distances (Shepard, 1966). Therefore in NMDS ordination, the proximity between objects corresponds to their similarity, but the ordination distances do not correspond to the original distances among objects. The NMDS procedure works randomly placing objects in the ordination space, previous definition of the desired number of dimensions, and their distances in this configuration are compared by monotonic regression with the distances in the original data matrix based on a stress function (values between 0 and 1). The stress function establishes how different the ranks on the ordination configuration are from the ranks in the original distance matrix. This is repeated in an iterative procedure until the lowest stress possible value is obtained (i.e. best fitness) using different random initial positions of the objects in the ordination space. NMDS has higher computer requirements than eigen analysis such as PcoA, PCA (Principal Component Analysis), or CA (Correspondence Analysis) because of the involved iterative procedure, nevertheless it is not considered a problem for small to medium size matrices because of constant improvements in computer power (Ramette, 2007).

The Bray-Curtis distance was used here. However, in the mathematical strict sense it is not a distance because it doesn't accomplish one of the three properties (the 3rd) of a true measure of distance which are the metric axioms described below:

1. $d_{ab} = d_{ba}$
2. $d_{ab} \geq 0$ and $= 0$ if and only if a = b
3. $d_{ab} \leq d_{ac} + d_{cb}$ (also called triangle inequality)

Where d_{ab} denotes the distance between objects a and b.

The formula for the estimation of Bray-Curtis dissimilarity between two samples *i* and *h* is as follows (McCune et al., 2002):

$$\boldsymbol{D_{i,h} = \frac{\sum_{j=1}^{p} a_{i,j} - a_{h,j}}{\sum_{j=1}^{p} a_{ij} + \sum_{j=1}^{p} a_{h,j}}} \qquad \textbf{[13]}$$

Where $\boldsymbol{a_{i,j}}$ is a matrix with each of its elements being the abundance of species *j* in sample unit *i*.

5.3.4. Redundancy Analysis (RDA)

The RDA is a canonical ordination method, thus some similarities are seen between RDA and Constrained Correspondence Analysis (CCA) which are common between canonical ordination technics. RDA is considered the canonical version of principal component analysis (PCA), (Ruokolainen and Blanchet, 2014). RDA is similar to CCA, the main difference is that RDA preserves the Euclidean distance instead of the $\chi 2$ distance among objects. Euclidean distance between two *p*-dimensional vectors **i** and **h** is calculated as follow (McCune et al., 2002):

$$\boldsymbol{ED_{\mathbf{i,h}} = \sqrt{\sum_{j=1}^{p} (a_{i,j} - a_{h,j})^2}} \qquad \textbf{[14]}$$

Where $a_{i,j}$ is a matrix with each of its elements being the abundance of species *j* in sample unit *i*.

In many cases the explanatory variables are not dimensionally homogeneous, thus, canonical ordinations are usually carried out using standardized explanatory variables.

This heterogeneity of dimensions don't affect the choice between running the analysis on a covariance or a correlation matrix in RDA, however; depending on the algorithm used, optimal linear combinations of explanatory variables is done either sequentially (axis by axis, using iterative algorithm) or in one step (direct algorithm). This latter method uses four steps, the first one is to regress each dependent variable separately on the explanatory variables and to compute both the fitted and residual values of the regressions. The second step is to run a principal component analysis (PCA) of the matrix of fitted values of these regressions. The third one is to use the matrix of canonical eigenvectors to compute two sorts of ordinations (for response and explanatory variables). And the last one is to use the matrix of residuals from the multiple regressions to compute a principal component analysis ordination (Ruokolainen and Blanchet, 2014).

5.3.5. Constrained Correspondence Analysis (CCA)

The CCA is another canonical (constrained) ordination method, this one uses the χ^2 distance among objects to ordinate (Legendre and Legendre, 1998). In CCA, χ^2 distance is calculated as follow:

$$\boldsymbol{x^2_{i,h}} = \sqrt{\sum_{j=1}^{p} \frac{(\boldsymbol{b_{hj}} - \boldsymbol{b_{ij}})^2}{\boldsymbol{a_{+j}}}} \qquad \textbf{[15]}$$

Where $x^2_{\mathbf{i,h}}$ is the chi-square distance between two samples of p species/taxa with profiles **i** = [$i_1, i_2, \ldots, i_p$] and **h** = [$h_1, h_2, \ldots, h_p$] (Greenacre, 2017), b_{hj} and b_{ij} denotes the prerelativized sample units for a_{hj}/a_{h+} and a_{ij}/a_{i+}, respectively, with a_{hj} depicting a matrix with each of its elements being the abundance of species *j* in sample unit *h* and a_{+j} being the total abundance of that species/taxa in all the samples. In this equation, the numerator is the square difference in relative abundance, expressed as the proportion of the species total and summed over all species.

Constrained ordination puts into relationship two matrices, one matrix of response variables (e.g. community matrix) and one matrix of explanatory variables. In CCA the ordination seeks the axes that are best explained by a linear combination of explanatory variables. In other words, CCA method seek the combination of explanatory

variables that best explain the variation of the response matrix. A constrained ordination produces as many canonical axes as there are explanatory variables, but each of these axes is a linear combination (a multiple regression model) of the explanatory variables. The canonical coefficients (i.e., the regression coefficients of the models) of the explanatory variables on each axis gives information about which variables are most important to explain the first, second,…*k*, axis. CCA is the canonical version of correspondence analysis (Ruokolainen and Blanchet, 2014). According to R documentation, when no formula is specified for CCA in Phyloseq package in R, and only the community data matrix is given, data is analysed by ordinary correspondence analysis and an unconstrained correspondence analysis ordination is obtained (Legendre and Legendre, 2012)(Legendre and Legendre, 2012).

5.3.6. Relevant characteristics regarding ordination methods

Some relevant characteristics regarding ordination methods are summarized in Table 1.

Table 1. Relevant metrics, procedures and miscellaneous characteristics regarding ordination methods[1].

Trait	MDS	DCA	NMDS	RDA	CCA
Distance	Euclidean	Bray-Curtis	Bray-Curtis	Euclidean	Chi Square
Indirect gradient	Yes	Yes	Yes	No	No
Arch effect correction	No	Yes	No	No	No
Canonical analysis	No	No	No	Yes	Yes
Computation time	Low	Low	High	Low	Low

[1]MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

The analysed methods to ordinate MRM matrices were performed by Phyloseq package in R.

5.4. *Variance component analysis and effects estimation*

5.4.1. *Meta-genomic BLUP*

This approach assumes independent effects of genotype and microbiome. Mixed models were used in a Bayesian framework, using the following independent effect model in linear notation:

$$\mathbf{y} = \mathbf{1}'\boldsymbol{\mu} + \mathbf{Zu} + \mathbf{Wm} + \mathbf{e} \qquad \textbf{[16]}$$

Where: **y** = feed efficiency, μ = population mean, **1** = vector of ones of *n*x*1* dimensions, **u** = genetic background, **m** being the microbiota effect, **Z** and **W** the corresponding incidence matrices for the genetic and the microbiota effects, respectively, and **e** = residual error, with **u** ~**N** (0, **GRM**σ_u^2), **m** ~**N** (0, **MRM**σ_m^2) and **e** ~**N** (0,σ_e^2), where **GRM** is the genomic relationship matrix and **MRM** the microbiome relationship matrix between cows.

5.4.2. *Meta-genomic BLUP with host interaction*

Another model accounting for the interaction between the genetic and the microbiota effects was tested:

$$\mathbf{y} = \mathbf{1}'\boldsymbol{\mu} + \mathbf{Zu} + \mathbf{Wm} + \mathbf{Tu}x\mathbf{m} + \mathbf{e} \qquad \textbf{[17]}$$

Where: **y**, ***μ*, Zu**, **Wm** and **e** are the same as in the previous model and **u***x***m** stands for the interaction between genetic background of the host and her microbiome, **T** represent the corresponding incidence matrix.

Models were solved in a Bayesian framework using the BGLR package in R (De Los Campos and Perez Rodriguez, 2016). The means and standard error of 100 replicates for the parameters of interest were obtained. Real data was analysed using the same models.

Six ordination methods were independently used to build the microbiota distance (or dissimilarity) matrices between cows. The six methods were: the one reported in (Ross et al., 2013) from now on identified as “Ross”, Metric Multidimensional Scaling (MDS), Detrended Correspondence Analysis (DCA), Non-Metric Multidimensional Scaling (NMDS), Redundancy Analysis (RDA) and Constrained Correspondence Analysis (CCA).

RESULTS AND DISCUSSION

6. RESULTS AND DISCUSION

6.1. *Simulated data*

6.1.1. *Comparison between diagonal elements of distance matrices*

The correlation between diagonal elements of the microbiome distance (or dissimilarity) matrices obtained with the ordination methods of Ross, MDS, DCA, NMDS, RDA and CCA is depicted in Figure 7. This correlation represent the association between values for alpha diversity (within cow) obtained with different methods of ordination. Those associations can also be represented for each element of the diagonal in all matrices with a scatterplot (Figure 8).

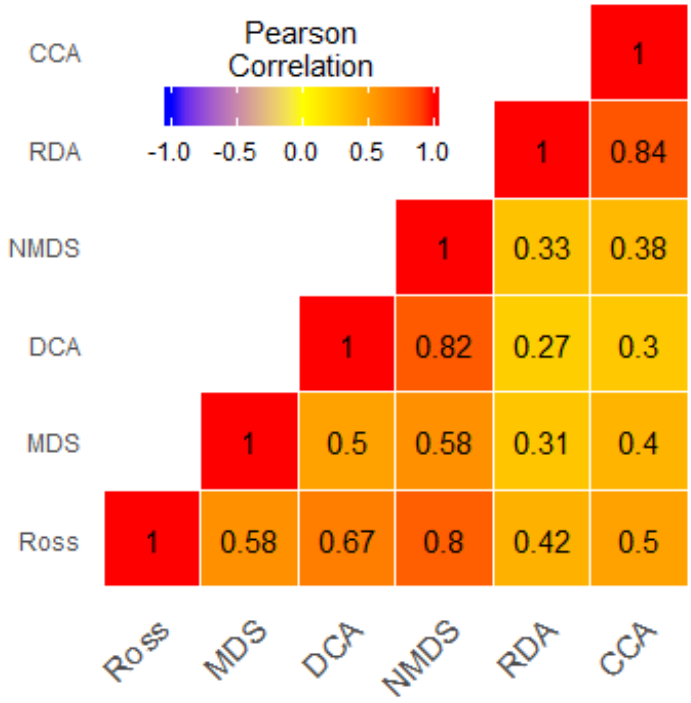

Figure 7. Pearson correlation between diagonal elements of 1000 x 1000 simulated microbiome distance (or dissimilarity) matrices according to ordination method of Ross et al. 2013 (Ross), Multidimensional Scaling (MDS), Detrended Correspondence Analysis (DCA), Non-Metric Multidimensional Scaling (NMDS), Redundancy Analysis (RDA) and Constrained Correspondence Analysis (CCA). The strength of the correlation is also represented with the intensity of the colour.

It can be inferred that all matrices had different grades of similitude between diagonal elements, ranging from mid to high associations. The highest correlation (ρ = 0.84) was obtained between diagonal elements of CCA and RDA matrices, which are both canonical technics that use similar ordination processes and differ mainly in the distance used, where CCA uses $\chi 2$ distance while RDA uses Euclidean distance

(Legendre and Legendre, 1998). The lowest Pearson correlation (ρ = 0.27) was obtained between diagonal elements of RDA and DCA matrices; this could be caused by both, the ordination process and the distance used. RDA and DCA use different ordination procedures and distances, as explained before.

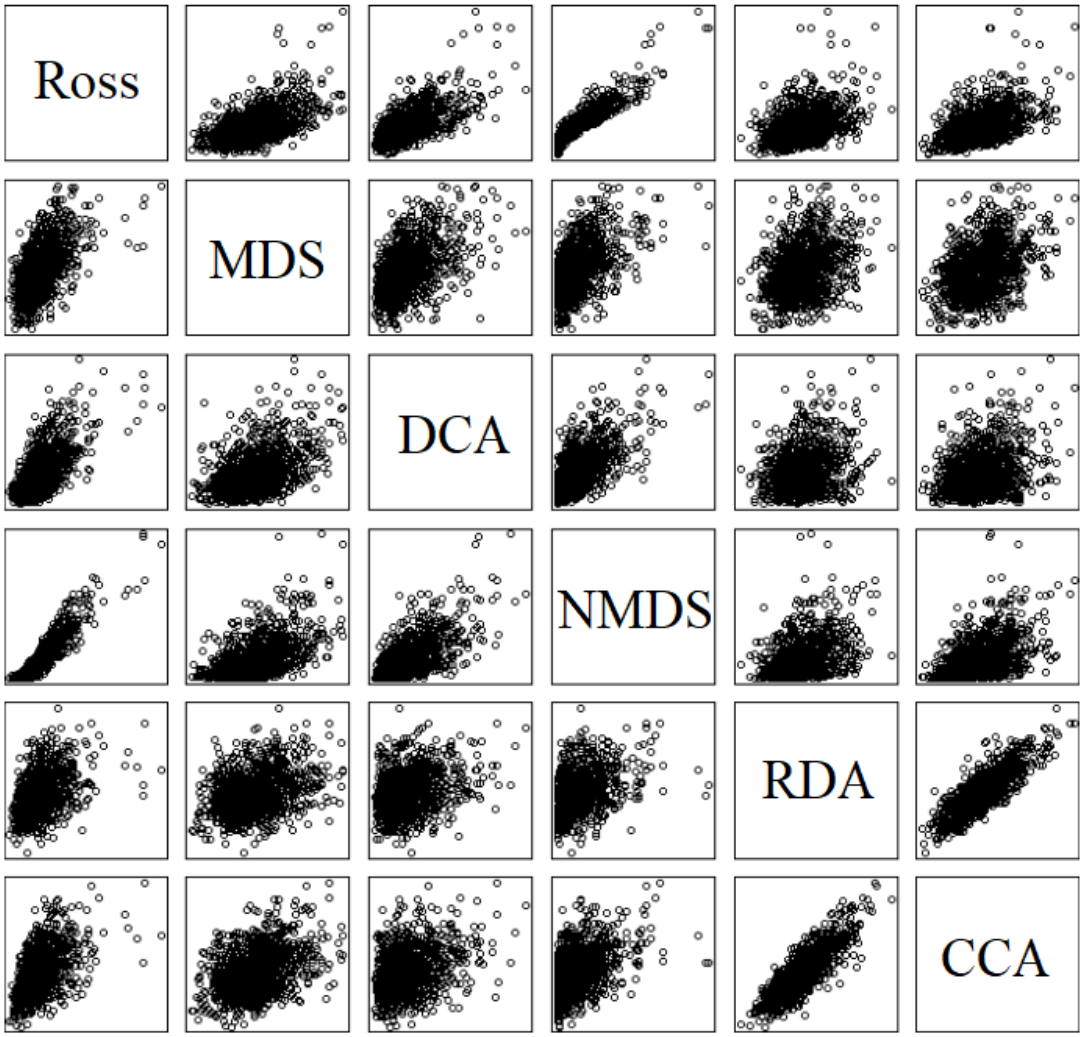

Figure 8. Association between diagonal elements of a 1000 x 1000 simulated matrix using ordination method of Ross et al. 2013 (Ross), Multidimensional Scaling (MDS), Detrended Correspondence Analysis (DCA), Non-Metric Multidimensional Scaling (NMDS), Redundancy Analysis (RDA) and Constrained Correspondence Analysis (CCA).

6.1.2. *Comparison between out-diagonal elements of the distance matrices*

The correlation between the out-diagonal elements of the microbiome distance (or dissimilarity) matrices obtained with the ordination methods of Ross, MDS, DCA, NMDS, RDA and CCA is depicted in Figure 9. This correlation represent the association between values for beta diversity (between cows) obtained with different methods of

ordination. Association for all out-diagonal elements can be depicted with scatterplots for all matrices (Figure 10).

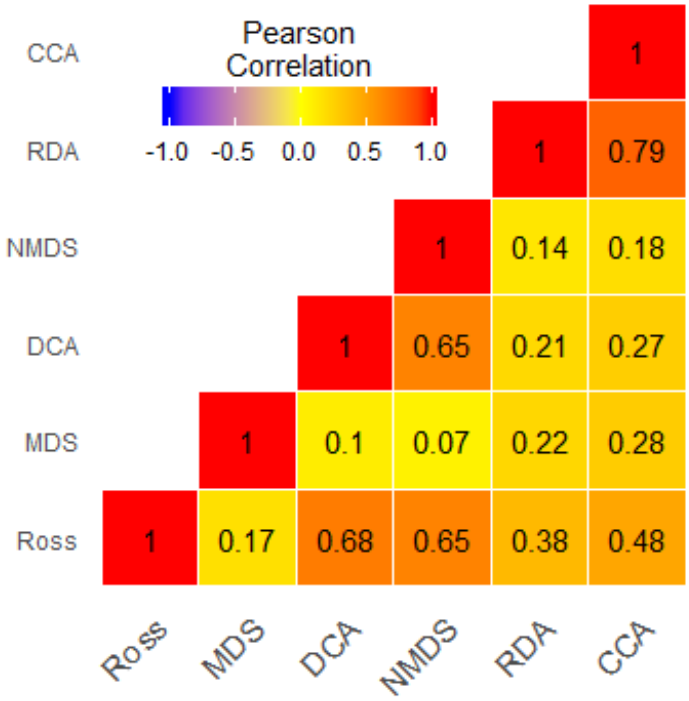

Figure 9. Pearson correlation between out-diagonal elements of 1000 x 1000 simulated microbiome distance (or dissimilarity) matrices according to ordination method of Ross et al. 2013 (Ross), Multidimensional Scaling (MDS), Detrended Correspondence Analysis (DCA), Non-Metric Multidimensional Scaling (NMDS), Redundancy Analysis (RDA) and Constrained Correspondence Analysis (CCA). The strength of the correlation is also represented with the intensity of the colour.

In general, out-diagonal elements had lower correlations between ordination methods than diagonal elements. There were correlations coefficients ranging from low to high between out-diagonal elements of microbiome similarity matrices, the highest correlation (ρ = 0.79) was again between CCA and RDA ordination methodologies. Both methodologies are canonical ordination technics (similarities and differences between those methods were mentioned before) which might partially explain its high correlation. The lowest correlation (ρ = 0.07) between out-diagonal elements of the microbiome similarity matrices was between NMDS and MDS technics. As mentioned before, the NMDS methodology uses Bray-Curtis distance in its estimation while MDS uses Euclidean distance for that propose, the ordination procedure also differs between those methods, with NMDS using a non-linear ranking between objects (Shepard, 1966), while MDS uses a linear mapping of its distances (Ramette, 2007), which could lead to such a low correlation between the out-diagonal elements of the matrices. Matrices obtained

with methods that use the same distance but different ordination processes had low correlations (i.e. MDS vs RDA, ρ = 0.22).

In general, there were two main factors affecting the correlation between all elements: the ordination procedure and the distances used. The ordination procedure had a larger impact than the distances used (except for the methods that used Bray-Curtis distance), obtaining the highest correlations when similar ordination procedure and different distances were used (i.e. RDA vs CCA, ρ = 0.79). Lower correlations were observed when the ordination procedures were different and within distance metrics (i.e. MDS vs RDA, ρ = 0.22), yielding the lowest correlations when ordination procedure and distances used were completely different (i.e. MDS vs NMDS, ρ = 0.07).

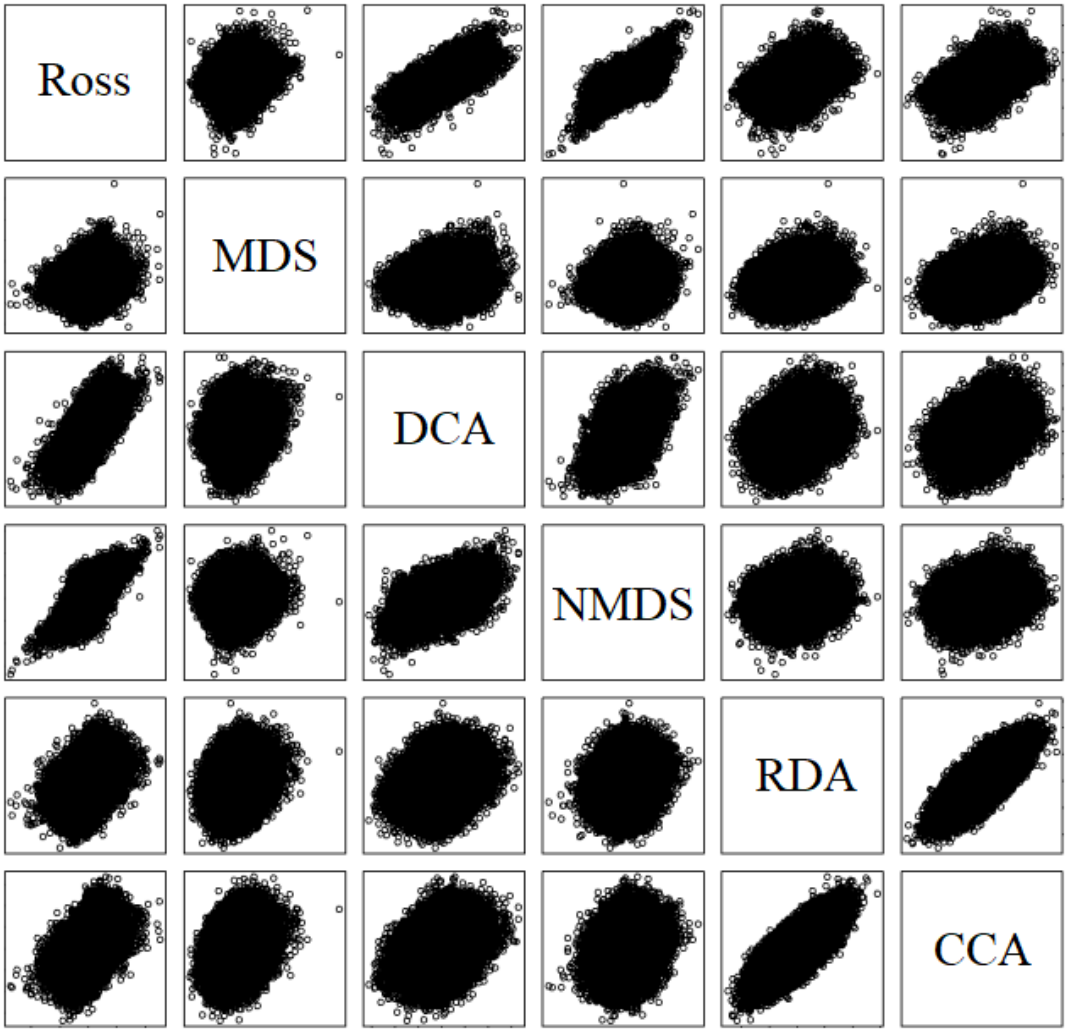

Figure 10. Association between out-diagonal elements of a 1000 x 1000 simulated matrix using ordination method of Ross et al. 2013 (Ross), Multidimensional Scaling (MDS), Detrended Correspondence Analysis (DCA), Non-Metric Multidimensional Scaling (NMDS), Redundancy Analysis (RDA) and Constrained Correspondence Analysis (CCA).

6.1.3. *Estimation of variance components using the independent effects model*

Average genetic variance (σ_u^2), microbiome variance (σ_m^2) and residual variance (σ_e^2) for the genomic and microbiome independent effects model and its corresponding standard error for 100 replicates (SEM) are presented in Table 2.

Table 2. Variance components for feed efficiency according to ordination method for the ruminal microbiota distance matrix, from a model independently including genomic and microbiome effects, using simulated data for 1000 cows and 100 replicates[1].

	Simulated value	Ross	MDS	DCA	NMDS	RDA	CCA
σ_u^2	370.34	374.2	383.4	378.3	397.1	379.1	375.8
SEM σ_u^2	---	7.256	7.425	7.853	8.763	7.582	7.346
σ_m^2	864.12	811.8	840.2	626.9	526.2	856.1	855.8
SEM σ_m^2	---	22.34	8.652	20.11	16.62	8.702	9.459
σ_e^2	864.12	852.6	863.4	1151	1436	850.4	852.9
SEM σ_e^2	---	7.548	8.521	17.55	18.73	7.608	7.326

[1]σ_u^2 = Genetic variance, σ_m^2 = Microbiome variance, σ_e^2 = Residual variance, SEM = Standard error of the means for 100 replicates, Ross = ordination method of Ross et al. 2013, MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

From Table 2 it can be inferred that all ordination methods achieved relatively good estimations of σ_u^2; although all methods slightly overestimated it. On the other hand, σ_m^2 was underestimated by DCA and NMDS ordination methods, while σ_e^2 was over estimated by the same two methods. DCA and NMDS methods are the only techniques evaluated in this analysis that use Bray-Curtis dissimilarity in the ordination procedure of the microbiota relationship matrix.

A visual representation of these variance components estimation is shown in Figure 11. The σ_u^2 was uniformly estimated by all methods (Figure 11A), while there was a larger accuracy for σ_m^2 for Ross, MDS, RDA and CCA methods compared with DCA and NMDS; looking at the methods that yielded the most exact estimations for microbiome variance, a higher variability in the estimation of Ross method is observed when compared to MDS, RDA and CCA, indicating better precision of the latter methods (Figure 11B). Likewise, σ_e^2 was best estimated by Ross, MDS, RDA and CCA than DCA and NMDS methods (Figure 11C).

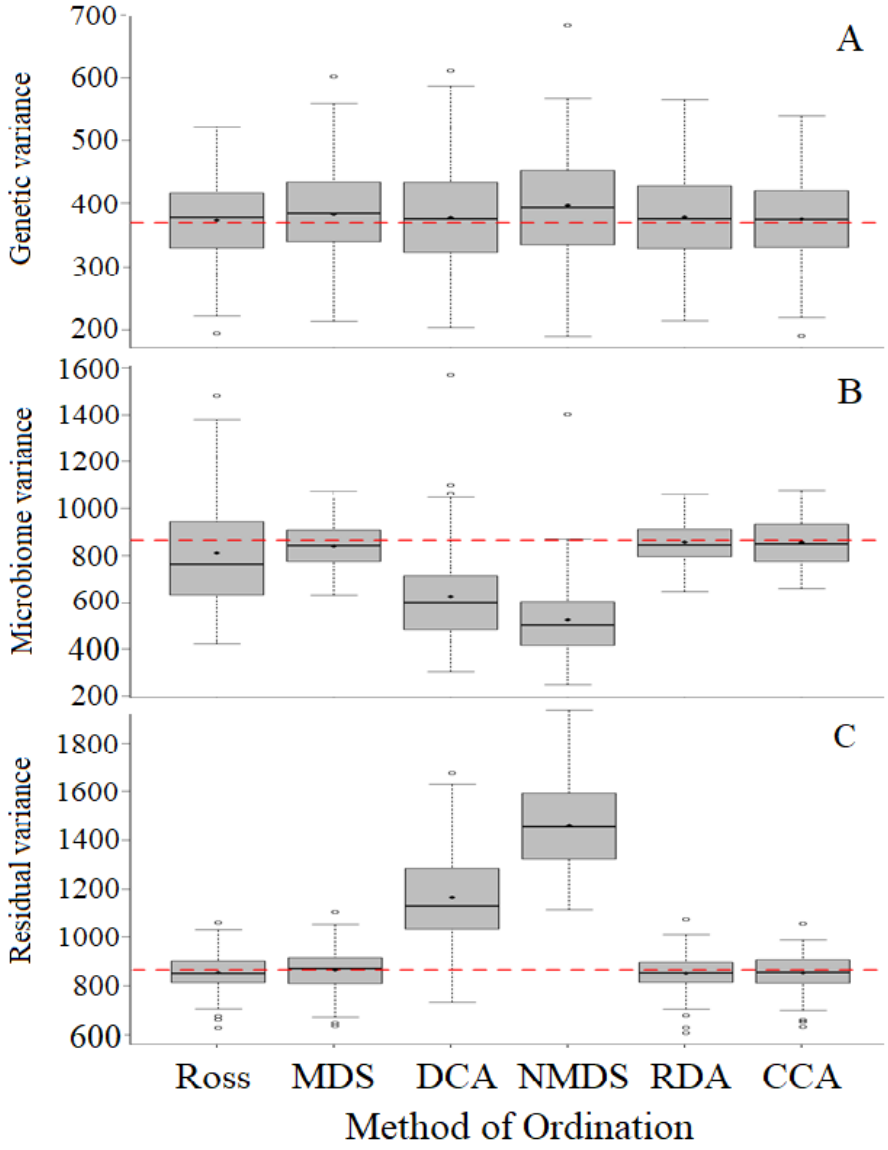

Figure 11. Genetic variance (A), microbiome variance (B) and residual variance (C) for feed efficiency according to ordination method for the ruminal microbiota distance matrix, from genomic and microbiome independent effects model using simulated data for 1000 cows. Ross = ordination method of Ross et al. 2013, MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

6.1.4. *Heritability and microbiability using the independent effects model*

The corresponding average heritability and microbiability estimates for the independent effects model are shown in Table 3. As expected from variance components previously mentioned, estimated heritability was slightly higher than simulated by method of Ross, MDS, RDA and CCA while DCA and NMDS methods underestimated the simulated heritability (Figure 12A). Microbiability estimates showed the same pattern as microbiome variance estimation, with methods of Ross, MDS, RDA and CCA

performing better than DCA and NMDS; also MDS, RDA and CCA methods were more precise than Ross method (Figure 12B).

Table 3. Heritability (h^2) and microbiability (m^2) estimates for feed efficiency according to ordination method for the ruminal microbiota distance matrix, from a model independently including genomic and microbiome effects using simulated data on 1000 cows and 100 replicates[1].

	Simulated value	Ross	MDS	DCA	NMDS	RDA	CCA
h^2	0.30	0.304	0.307	0.249	0.217	0.307	0.305
SEM h^2	---	0.005	0.005	0.005	0.004	0.005	0.005
m^2	0.50	0.480	0.493	0.349	0.266	0.501	0.500
SEM m^2	---	0.006	0.003	0.008	0.007	0.003	0.003

[1]SEM = Standard error of the means for 100 replicates, Ross = ordination method of Ross et al. 2013, MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

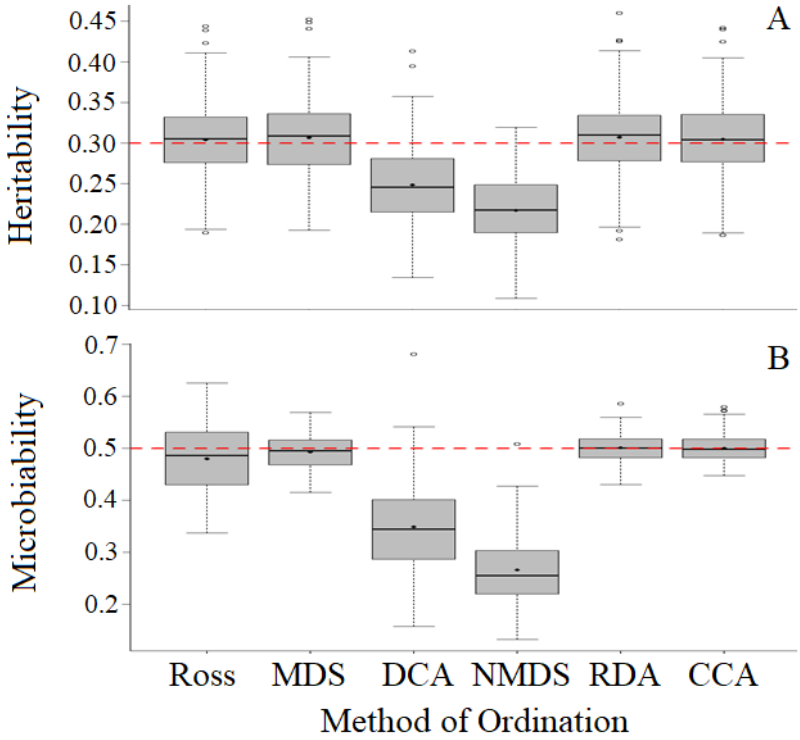

Figure 12. Heritability (A) and microbiability (B) for feed efficiency according to ordination method for the ruminal microbiota distance matrix, from genomic and microbiome independent effects model using simulated data for 1000 cows. Ross = ordination method of Ross et al. 2013, MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

6.1.5. *Correlation between GEBV and TBV using the independent effects model*

Correlations (standard error within brackets) between genomic estimated breeding values (**GEBV**) and true breeding values (**TBV**) were similar for all ordination methods and were: 0.633(±0.003), 0.592(±0.004), 0.598(±0.004), 0.557(±0.004), 0.624(±0.003) and 0.631(±0.003) for ordination procedures of Ross, MDS, DCA, NMDS, RDA and CCA, respectively. A scatter plot between GEBV and TBV is depicted in Figure 13. Showing a similar pattern for all ordination methods.

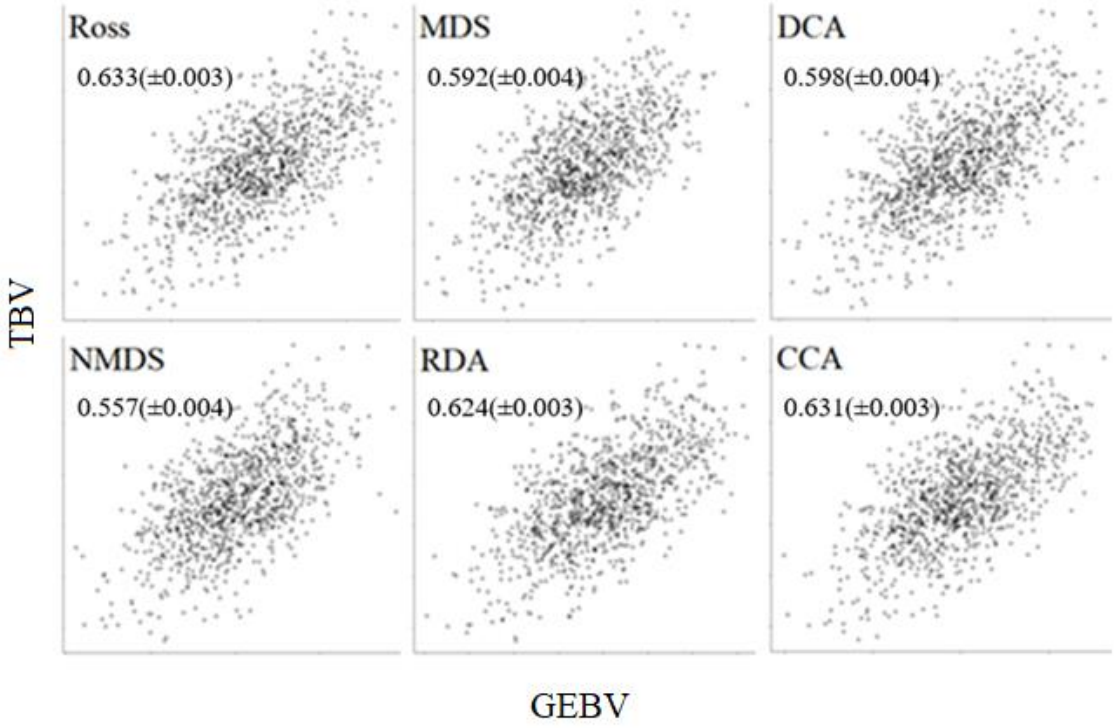

Figure 13. Association between genomic estimated breeding values (GEBV) and true breeding values (TBV) according to method of ordination from genomic and microbiome independent effects model using simulated data for 1000 cows. Ordination method of Ross et al. 2013 (Ross), Multidimensional Scaling (MDS), Detrended Correspondence Analysis (DCA), Non-Metric Multidimensional Scaling (NMDS), Redundancy Analysis (RDA) and Constrained Correspondence Analysis (CCA).

6.1.6. *Correlation between EMV and TMV using the independent effects model*

Correlations (standard error within brackets) between estimated microbiome values (**EMV**) and true microbiome values (**TMV**) varied according to the ordination method and were: 0.975(±0.001), 0.845(±0.001), 0.807(±0.011), 0.517(±0.019),

0.949(±0.001) and 0.966(±0.001) for ordination procedures of Ross, MDS, DCA, NMDS, RDA and CCA, respectively.

A scatter plot between EMV and TMV is depicted in Figure 14, showing different accuracies in the patterns of association according to the ordination method used.

It must be point out that the simulated effect of the microbiome favoured the benchmark method of Ross, because a linear and independent effect was applied over the OTUs. In this sense, the similar results obtained from the RDA and CCA matrices.

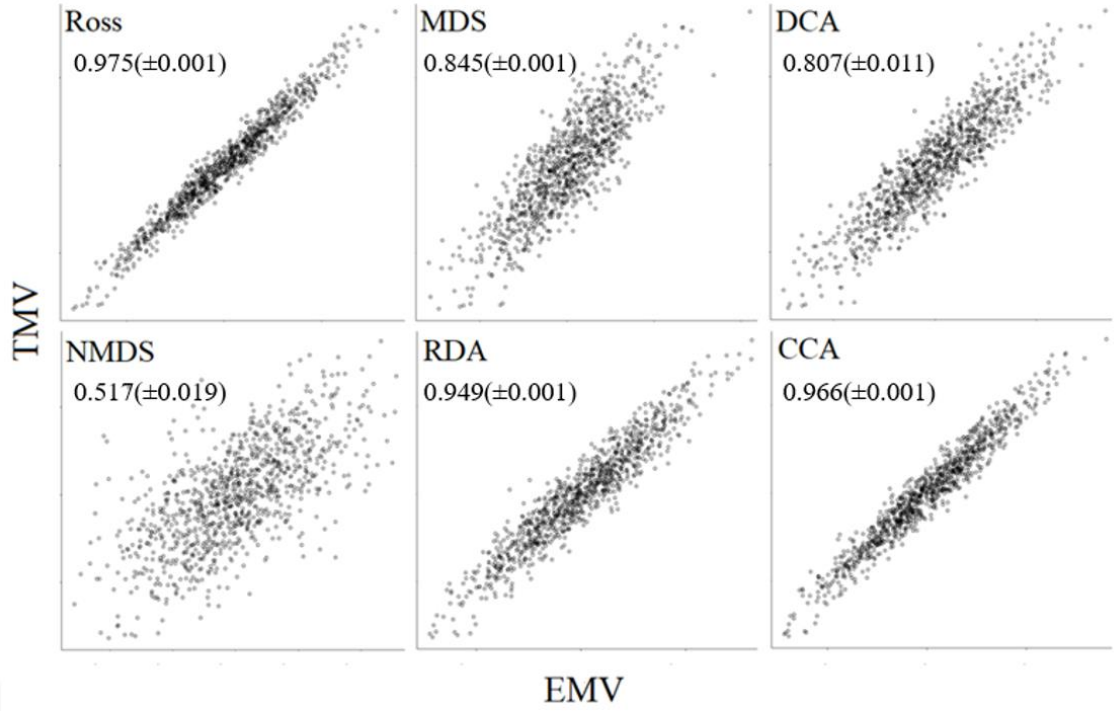

Figure 14. Association between estimated microbiome values (EMV) and true microbiome values (TMV) according to method of ordination from genomic and microbiome independent effects model using simulated data for 1000 cows. Ordination method of Ross et al. 2013 (Ross), Multidimensional Scaling (MDS), Detrended Correspondence Analysis (DCA), Non-Metric Multidimensional Scaling (NMDS), Redundancy Analysis (RDA) and Constrained Correspondence Analysis (CCA).

6.1.7. *Estimation of variance components using an interaction effect model*

Average genetic variance (σ_u^2), microbiome variance (σ_m^2) and residual variance (σ_e^2) for the genomic and microbiome interaction effects model and its corresponding standard error for 100 replicates (SEM) are presented in Table 4.

Table 4. Variance component estimation according to ordination method for the ruminal microbiota distance matrix, using a model that included the interaction between genomic and microbiome effects from simulated data for 1000 cows and 100 replicates[1].

	Simulated value	Ross	MDS	DCA	NMDS	RDA	CCA
σ_u^2	367.2	358.6	341.0	381.2	394.5	362.5	365.1
SEM σ_u^2	—	6.5	6.0	8.5	10.5	6.6	7.3
σ_m^2	856.7	1017	1196	654.1	573.2	1121	1111
SEM σ_m^2	—	74.7	89.3	40.3	40.4	85.4	84.5
σ_{uxm}^2	353.2	218.5	266.7	215.8	249.6	262.1	251.2
SEM σ_{uxm}^2	—	5.2	4.7	7.1	12.2	5.9	6.1
σ_e^2	856.7	652.1	621.4	1100	1321	689.3	696.8
SEM σ_e^2	—	9.5	8.5	33.3	50.5	8.4	7.9

[1]σ_u^2 = Genetic variance, σ_m^2 = Microbiome variance, σ_{uxm}^2 = Genetic x Microbiome interaction variance, σ_e^2 = Residual variance, SEM = Standard error of the means for 100 replicates, Ross = ordination method of Ross et al. 2013, MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

As in the previous model, all methods achieved good estimation of σ_u^2, Whereas DCA and NMDS methods underestimated σ_m^2 and Ross, MDS, RDA and CCA overestimated it. An underestimation of σ_{uxm}^2 was observed for ordination matrices. Ross, MDS, RDA and CCA underestimated σ_e^2 while DCA and NMDS overestimated it. The visual representation of these results is shown in Figure 15.

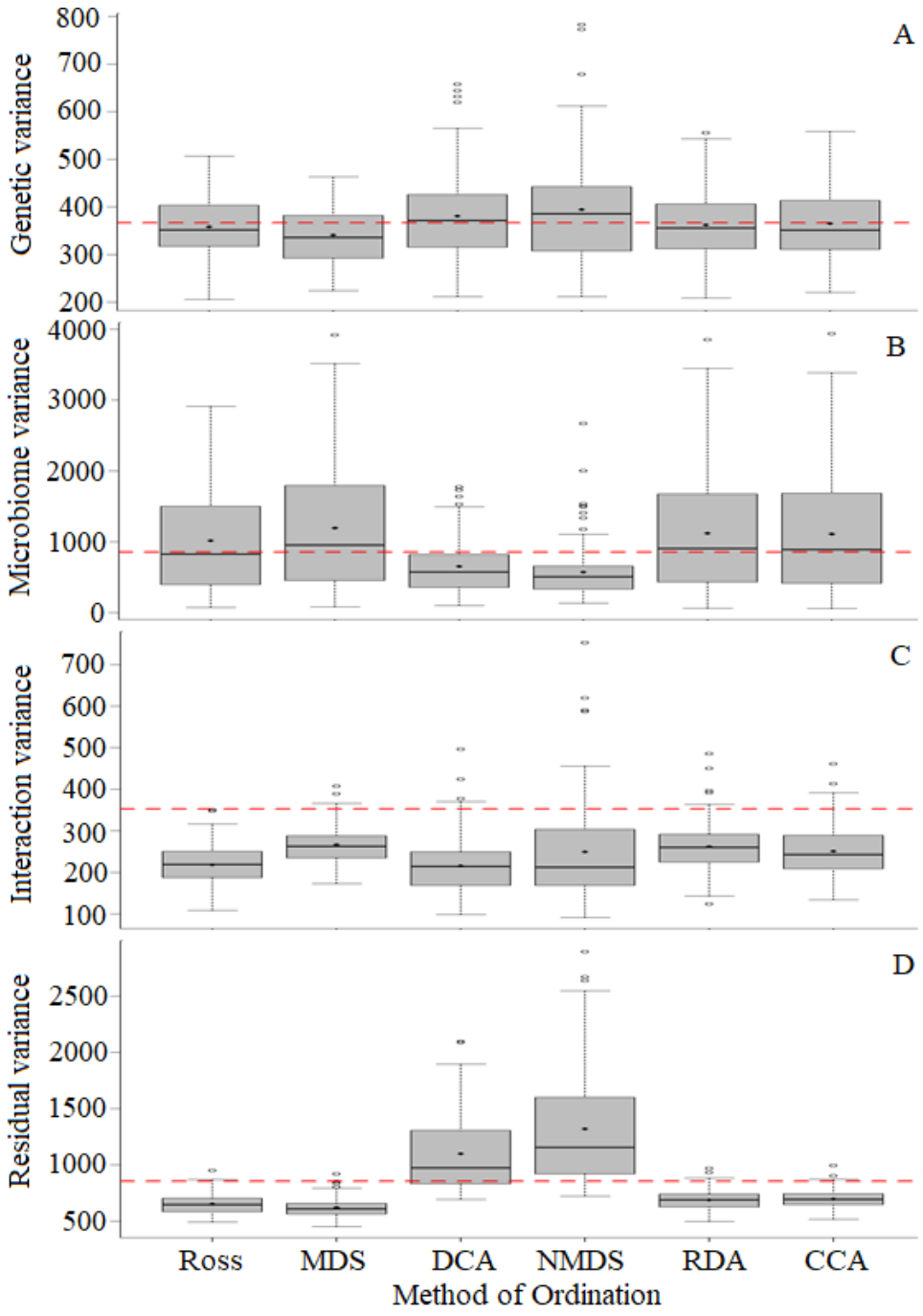

Figure 15. Genetic variance (A), microbiome variance (B), Interaction variance (C) and residual variance (D) for feed efficiency according to ordination method for the ruminal microbiota distance matrix, from genomic and microbiome interaction effects model using simulated data for 1000 cows. Ross = ordination method of Ross et al. 2013, MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

6.1.8. *Estimation of heritability and microbiability using an interaction effect*

The corresponding average heritability and microbiability estimates for the interaction effects model are shown in Table 5. As expected from variance components previously mentioned, estimated heritability was larger with matrices of Ross, MDS, RDA and CCA than simulated heritability while DCA and NMDS methods underestimated the simulated heritability (Figure 16A). Microbiability estimates showed the same pattern than microbiome variance estimation, where matrices of Ross, MDS, RDA and CCA performed better than DCA and NMDS (Figure 16B). The Interaction model was less accurate at estimating the heritability and microbiability than the independent effects model, suggesting that part of this interaction is captured by the independent effects.

Table 5. Heritability (h^2) and microbiability (m^2) estimates for feed efficiency according to ordination method for the ruminal microbiota distance matrix, using a model that included the interaction between genomic and microbiome effects from simulated data for 1000 cows and 100 replicates[1].

	Simulated value	Ross	MDS	DCA	NMDS	RDA	CCA
h^2	0.30	0.36	0.36	0.27	0.24	0.35	0.34
SEM h^2	—	0.01	0.01	0.01	0.01	0.01	0.01
m^2	0.50	0.53	0.57	0.35	0.29	0.53	0.53
SEM m^2	—	0.02	0.02	0.01	0.01	0.02	0.02

[1]SEM = Standard error of the means for 100 replicates, Ross = ordination method of Ross et al. 2013, MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

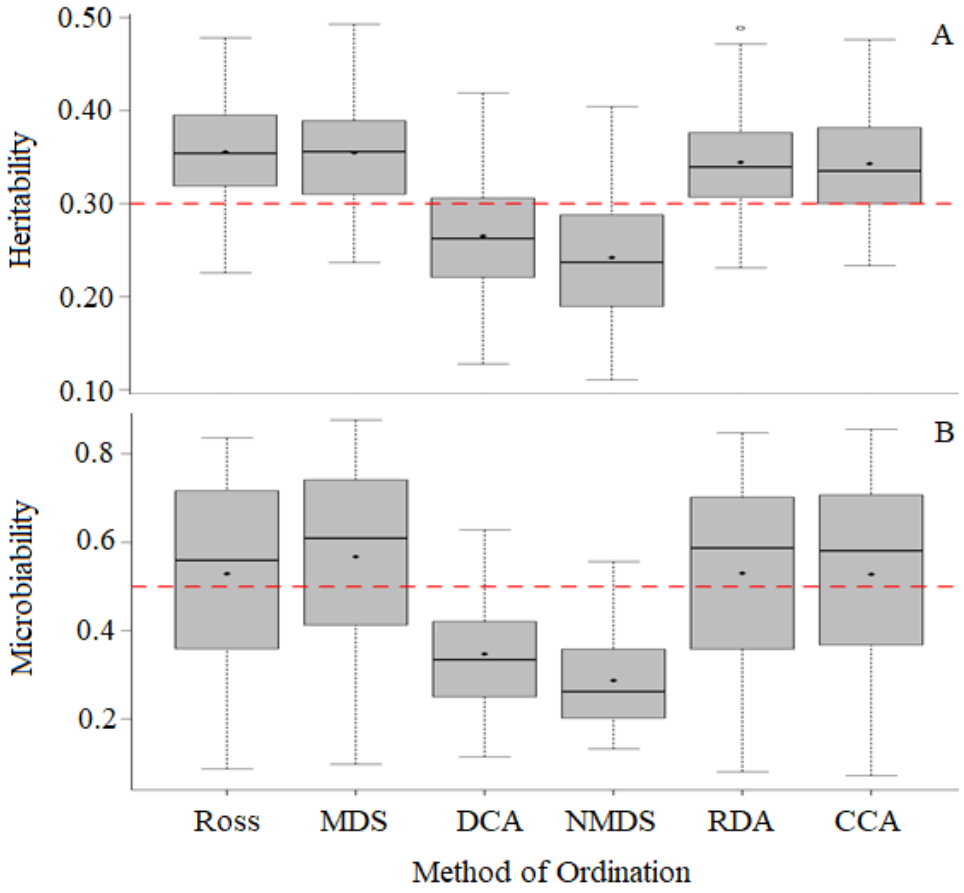

Figure 16. Heritability (A) and microbiability (B) for feed efficiency according to ordination method for the ruminal microbiota distance matrix, from genomic and microbiome interaction effect model using simulated data for 1000 cows. Ross = ordination method of Ross et al. 2013, MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

6.1.9. *Correlation between GEBV and TBV using an interaction effect model*

Correlations (standard error within brackets) between GEBV and TBV were also similar for all ordination methods in the interaction effects model and were: 0.629(±0.003), 0.594(±0.004), 0.582(±0.005), 0.560(±0.006), 0.623(±0.003) and 0.627(±0.003) for ordination procedures of Ross, MDS, DCA, NMDS, RDA and CCA, respectively.

6.1.10. *Correlation between EMV and TMV using an interaction effect model*

As in the independent effects model, correlations (standard error within brackets) between EMV and TMV for the interaction effects model varied according to ordination

method and were: 0.893(±0.030), 0.753(±0.026), 0.739(±0.031), 0.609(±0.027), 0.859(±0.028) and 0.880(±0.029) for ordination procedures of Ross, MDS, DCA, NMDS, RDA and CCA, respectively.

6.1.11. *Interaction effect*

The variance for simulated interaction effect (σ^2_{uxm}= 353.2) was underestimated by all methods and were: 218.5(±5.2), 266.7(±4.7), 215.8(±7.1), 249.6(±12.2), 262.1(±5.8) and 251.2(±6.1) for procedure of Ross, MDS, DCA, NMDS, RDA and CCA, respectively. Correlations between the estimated interaction effect and its corresponding simulated value were generally low: 0.177(±0.007), 0.172(±0.017), 0.242(±0.015), 0.260(±0.020), 0.155(±0.009) and 0.158(±0.008) in the same order. From these results it can be inferred that interaction effects model performed poorly compared to independent effects model and that also was inefficient estimating the interaction effect.

6.2. ***Real data***

6.2.1. *Relative abundance of OTUs*

OTUs were classified to the genus level and their relative abundance was computed; however if genus was not possible to be determined, the previous classification (*family*) was set to the OTU, and so on for order, class, phylum and kingdom.

The relative abundance of 92 OTUs for each sample (microbiome of each animal) is depicted in Figure 17.

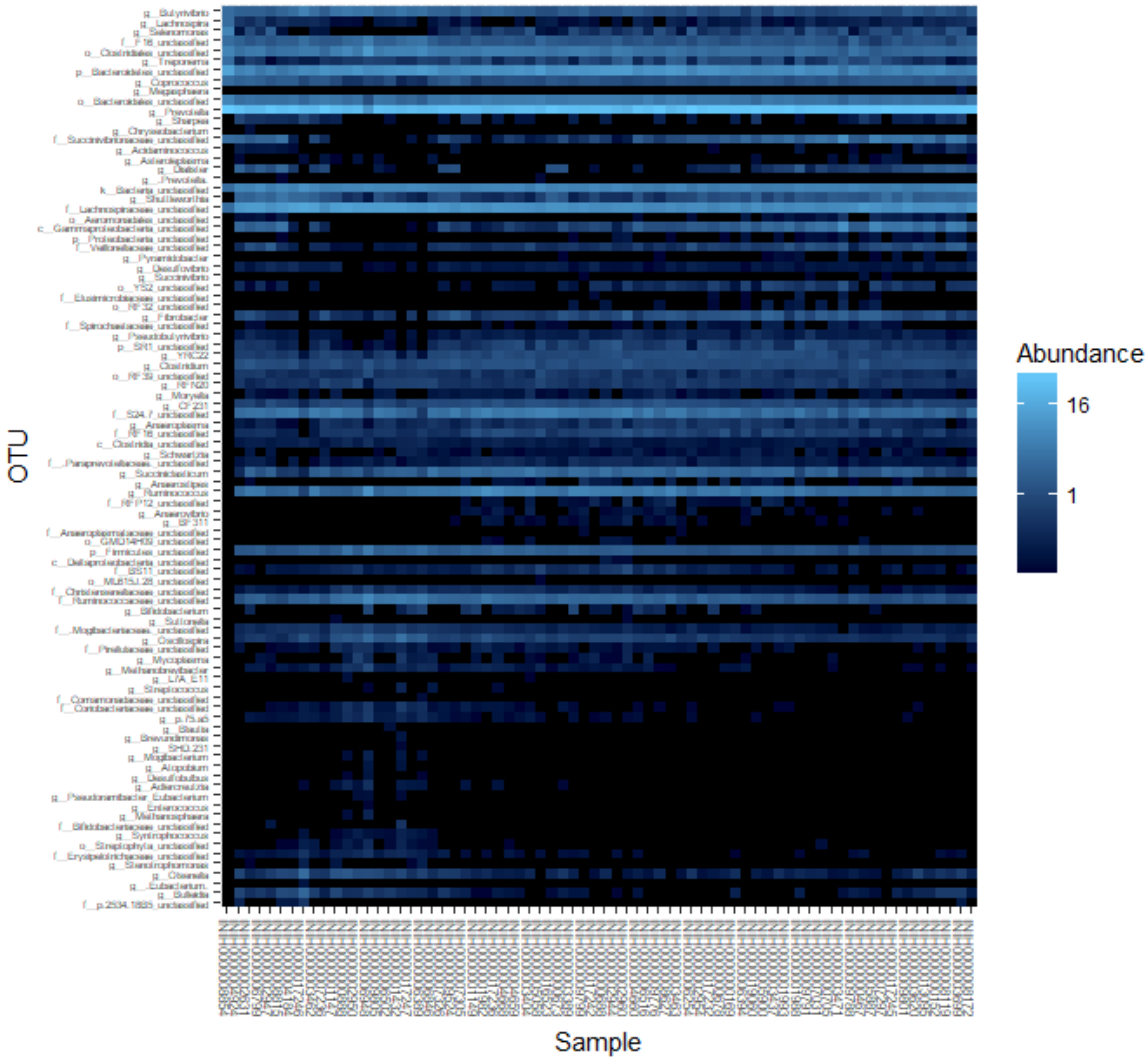

Figure 17. Relative abundance of OTUs in microbiome sequenced using Illumina Miseq for the hypervariable region V3-V4 of the 16S rRNA amplicon from 70 lactating Spanish Holstein cows according to sample.

From Figure 17 it can be inferred that there is variation between animals for the relative abundance of the OTU, it is also evident that some specific OTU are more frequent than others within an animal; in general the most abundant genus was *Prevotella* with 32.32 % (SEM = 0.69) followed by a family of Lachnospiraceae with a mean frequency of 8.87 % (SEM = 0.27) and an uncultured microorganism from the phylum Bacteroidetes, with a mean relative abundance of 7.91 % (SEM = 0.26).

Most of the OTUs were classified in the bacteria kingdom; however some cows presented a low proportion of archaea within their microbiome. Regarding relative abundance of phyla, there were an evident higher proportion of Bacteroidetes and Firmicutes than any other phylum (Figure 18). This is consistent with other studies regarding relative abundance of phylum in rumen microbiota, for instance: Gonzalez-

Recio et al., (2017) found that Bacteroidetes and Firmicutes were the most abundant phyla in Brown Swiss cows, with a contribution of 48 and 32%, respectively. Likewise, Deusch et al., (2017) determined that overall bacterial composition was dominated by the Bacteroidetes phylum followed by Firmicutes when analysed microbiota of three rumen cannulated lactating Jersey cows.

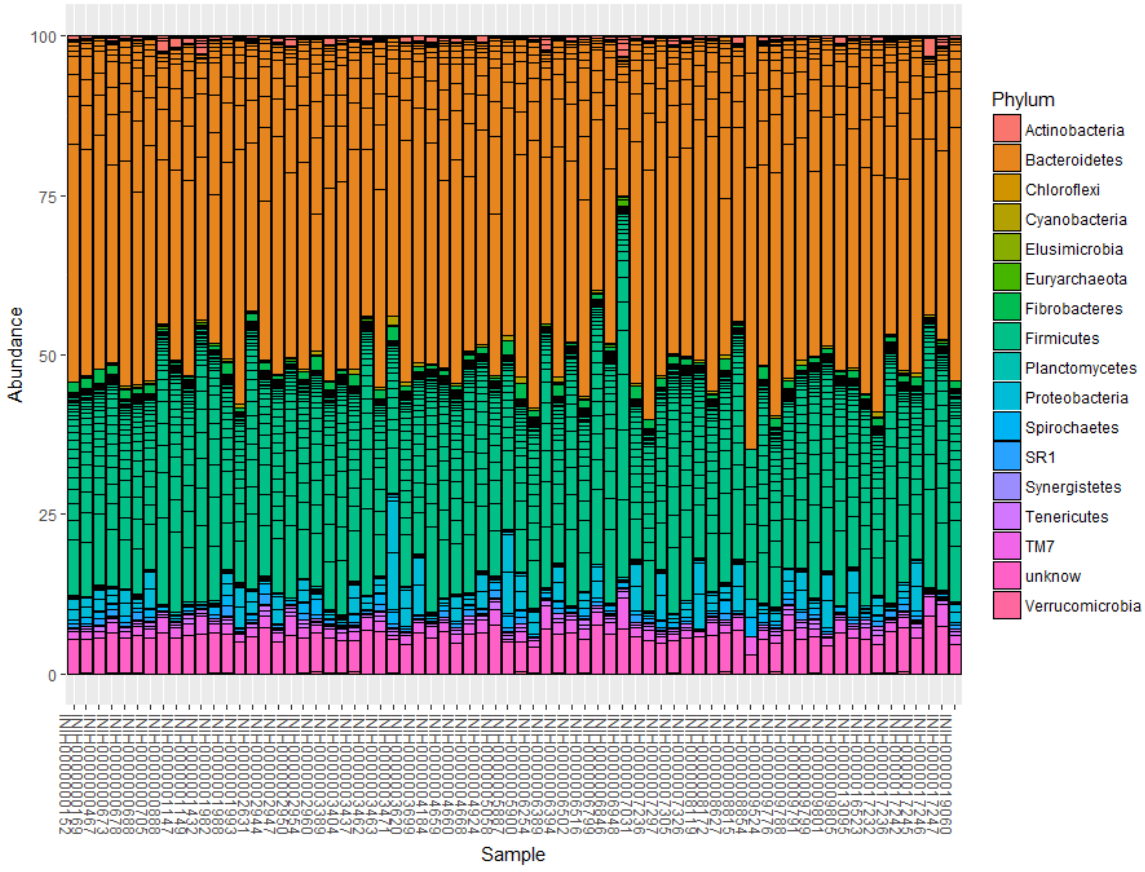

Figure 18. Relative abundance of taxonomic phyla according to microbiome sequenced using Illumina Miseq for the hypervariable region V3-V4 of the 16S rRNA amplicon according to sample for 70 lactating Spanish Holstein cows.

6.2.2. *Estimation of variance components from the independent effects model using real data.*

Genetic variance (σ_u^2), microbiome variance (σ_m^2) and residual variance (σ_e^2) for feed efficiency were estimated from model of independent genetic and microbiome effects using real data (Table 6).

Table 6. Variance components for feed efficiency according to ordination method for the ruminal microbiota distance matrix, from a model independently including genomic and microbiome independent effects using real data for 70 cows[1].

Variance Components	Ross	MDS	DCA	NMDS	RDA	CCA
$\mathbf{y = 1'\mu + Zu + Wm + e}$						
Genetic variance	0.107	0.107	0.110	0.111	0.117	0.112
Microbiome variance	0.104	0.101	0.139	0.142	0.107	0.106
Residual variance	1.281	1.273	1.234	1.233	1.299	1.270

[1]Ross = ordination method of Ross et al. 2013, MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

Similar values were obtained for σ_u^2, σ_m^2 and σ_e^2 across the ordination methods that performed better in the simulation (Ross, MDS, RDA and CCA). Whereas DCA and NMDS methods estimated similar σ_u^2 than the other methods, but lower σ_e^2, which might be due to on overestimation of σ_m^2.

6.2.3. *Heritability, microbiability and correlations between GEBV and EMV with phenotype for the independent effects model*

Results of h^2 estimates ranged from 0.077 (Ross and MDS) to 0.083 (NMDS), m^2 estimates ranged from 0.073 (MDS) to 0.103 (NMDS). Correlations between posterior means for GEBV and the phenotype ranged from 0.857 (DCA) to 0.912 (NMDS), while the correlation between the posterior means of the EMV and the phenotype ranged from 0.210 (NMDS) to 0.910 (RDA) (Table 7).

Table 7. Heritability, microbiability and correlations between GEBV and phenotype; and between EMV and phenotype for feed efficiency, estimated using a model independently including genomic and microbiome effects according to method of ordination for microbiota using real data from 70 cows[1].

Parameters	Ross	MDS	DCA	NMDS	RDA	CCA
$\mathbf{y = 1'\mu + Zu + Wm + e}$						
Heritability	0.077	0.077	0.082	0.083	0.083	0.081
Microbiability	0.075	0.073	0.102	0.103	0.076	0.077
Correlation GEBV vs Phenotype	0.865	0.862	0.857	0.912	0.892	0.879
Correlation EMV vs Phenotype	0.483	0.666	0.360	0.210	0.910	0.906

[1]GEBV= Genomic Estimated Breeding Values, EMV= Estimated Microbiome Values, Ross = ordination method of Ross et al. 2013, MDS = Multidimensional Scaling, DCA = Detrended Correspondence Analysis, NMDS = Non-Metric Multidimensional Scaling, RDA = Redundancy Analysis, CCA = Constrained Correspondence Analysis.

Heritability estimates for feed efficiency from a meta-analysis of up to 11 estimates revealed estimates for gross feed efficiency of 0.06 ± 0.010 (Berry and Crowley, 2013) which are slightly lower than values found in the present study. Those estimates of heritability differ from the one reported in (Spurlock et al., 2012) which inform of higher values of h^2 (0.32) for gross feed efficiency.

6.2.4. *Estimation of variance components including an interaction effect in the model*

As in the independent effects model, similar values were obtained for σ_u^2 (0.09-0.11), σ_m^2 (0.08-0.10) and σ_e^2 (1.26-1.30) across the ordination methods that performed better in the simulation (Ross, MDS, RDA and CCA). While as in the previous model, DCA and NMDS methods estimated similar σ_u^2 than the other methods, but higher σ_m^2 and lower σ_e^2 (Table 8).

Table 8. Variance components for feed efficiency according to ordination method for the ruminal microbiota distance matrix, using a model with an interaction between genomic and microbiome effects from real data for 70 cows[1].

Variance Components	Ross	MDS	DCA	NMDS	RDA	CCA
$\mathbf{y = 1'\mu + Zu + Wm + Tu}x\mathbf{m + e}$						
Genetic variance	0.089	0.095	0.078	0.088	0.110	0.085
Microbiome variance	0.099	0.078	0.116	0.133	0.076	0.079
Residual variance	1.284	1.290	1.234	1.261	1.299	1.257

[1]Ross = ordination method of Ross et al. 2013, MDS = Multidimensional scaling, DCA = Detrended correspondence analysis, NMDS = Non-metric multidimensional scaling, RDA = Redundancy analysis, CCA = Constrained correspondence analysis.

6.2.5. *Heritability, microbiability and correlations between GEBV and feed efficiency, microbiome effect and feed efficiency in the model with an interaction effect.*

The *h2* estimates ranged from 0.059 (DCA) to 0.078 (RDA), m^2 estimates ranged from 0.056 (RDA) to 0.096 (NMDS). Correlations between the posterior means of GEBV and feed efficiency ranged from 0.799 (CCA) to 0.889 (Ross and MDS), while the correlation between the posterior means of EMV and the phenotype ranged from 0.211 (NMDS) to 0.906 (CCA) for the interaction effects model (Table 9).

Table 9. Heritability, microbiability and correlations between GEBV and phenotype; and between EMV and phenotype for feed efficiency, estimated using a model with interaction between genomic and microbiome effects according to method of ordination for microbiota using real data from 70 cows[1].

Parameters	Ross	MDS	DCA	NMDS	RDA	CCA
$\mathbf{y = 1'\mu + Zu + Wm + Tu}x\mathbf{m + e}$						
Heritability	0.065	0.069	0.059	0.065	0.078	0.063
Microbiability	0.072	0.057	0.086	0.096	0.056	0.059
Correlation GEBV vs Phenotype	0.889	0.889	0.839	0.802	0.861	0.799
Correlation EMV vs Phenotype	0.506	0.639	0.360	0.211	0.899	0.906

[1]GEBV= Genomic estimated breeding values, EMV= Estimated microbiome values, Ross = ordination method of Ross et al. 2013, MDS = Multidimensional scaling, DCA = Detrended correspondence analysis, NMDS = Non-metric multidimensional scaling, RDA = Redundancy analysis, CCA = Constrained correspondence analysis.

6.2.6. *Model comparison*

Different criteria that could assist on selecting the best model can be obtained from the analysis: log-likelihood evaluated at posterior mean (**logLikAtPostMean**), the posterior mean of the Log-Likelihood (**PostMeanLogLik**), estimated effective number of parameters (**pD**), as well as the deviance information criteria (**DIC**). Results from those criteria were obtained for a model accounting for independent genetic and microbiome effects, for a model including an interaction between those effects, as well as for a model accounting only for the genomic effect (GBLUP), without microbiome effect (Table 10).

Results regarding logLikAtPostMean, PostMeanLogLik and pD for the three models showed minor differences between them, with values for those criteria overlapping between methods of the independent effects model and the interaction effect model, as well as with the GBLUP model.

Results obtained from the DIC were slightly lower for the GBLUP model, followed by the model of independent genetic and microbiome effects, while the model including the genetic-microbiome interaction effect showed higher values for these criteria. The DCA and NMDS methods had slightly lower DIC values than the other methods of ordination.

Table 10. Information criteria estimated for models with and without interaction between genetic and microbiome effect according to method of ordination for the microbiota relationship matrix for real data from 70 cows, a GBLUP model* is included as reference [1].

	Ross	MDS	DCA	NMDS	RDA	CCA	$\mathbf{y = 1'\mu + Zu + e}^*$
$\mathbf{y = 1'\mu + Zu + Wm + e}$							
LogLikAtPostMean	-84.2	-84.9	-84.0	-85.4	-84.1	-84.2	-85.1
PostMeanLogLik	-88.8	-89.8	-88.0	-88.9	-89.1	-89.2	-88.7
pD	9.3	9.8	7.9	7.1	10.1	10.1	7.1
DIC	187.0	189.3	183.9	185.0	188.3	188.5	184.4
$\mathbf{y = 1'\mu + Zu + Wm + Tu}x\mathbf{m + e}$							
LogLikAtPostMean	-84.1	-84.2	-83.9	-85.4	-83.6	-83.7	-85.1
PostMeanLogLik	-89.5	-90.0	-88.8	-89.6	-89.5	-89.7	-88.7
pD	10.9	11.6	9.8	8.3	11.9	11.9	7.1
DIC	190.0	191.7	187.5	187.5	191.0	191.3	184.4

[1]Ross = ordination method of Ross et al. 2013. MDS = Multidimensional scaling. DCA = Detrended correspondence analysis. NMDS = Non-metric multidimensional scaling. RDA = Redundancy analysis. CCA = Constrained correspondence analysis. logLikAtPostMean = log-likelihood evaluated at posterior mean. pD = estimated effective number of parameters. PostMeanLogLik = posterior mean of the Log-Likelihood. DIC = deviance information criteria.* $\mathbf{y = 1'\mu + Zu + e}$ = GBLUP model accounting for the genomic effect, not including microbiome effect.

The DIC is a hierarchical modelling generalization of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), which consists of two components, a term that measures goodness-of-fit and a penalty term for increasing model complexity, in which models with smaller DIC should be preferred to models with larger DIC, because this point to a better fit and a lower degree of model complexity; however, some authors (Sorensen and Gianola, 2002) make emphasis in consider DIC as a preliminary device for screening alternative models. DIC calculation is as follow:

$$\boldsymbol{DIC = \bar{D} + p_D} \qquad \textbf{[18]}$$

The first term ($\boldsymbol{\bar{D}}$), is a Bayesian measure of model fit, defined as the posterior expectation of the deviance, while the second component ($\boldsymbol{p_D}$) measures the complexity

of the model by the effective number of parameters, also defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters (Berg et al., 2004)

DIC is considered particularly useful for Bayesian model selection where the posterior distribution of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation. When prior information is negligible, DIC result in an equivalent approximation to Akaike's criterion (Spiegelhalter et al., 2002), but the DIC uses the posterior expectation of the log likelihood as a measure of model fit (Sorensen, 2004).

The DIC can be used to decide adequacy of a model; however, the deference between DIC of two given models should be enough to make a good choice. There are values that roughly guide that decision, for instance Stevens (2004) indicate that it is difficult to affirm what is an important difference in DIC and suggests that 10 definitively exclude the model with the highest DIC, differences from 5 to 10 are still substantial, but to choose a model that differs only by a value below 5 in DIC could be misleading.

Based on the obtained results, there is not enough difference between DIC to consider that the model without interaction should be chosen over model with interaction.

CONCLUSIONS

7. CONCLUSIONS

1. I inferred that microbiota relationship matrices obtained with MDS, RDA and CCA are as suitable as, or even better than the previously reported by Ross et al. (2013) in terms of the estimation of variance components, heritability and microbiability using simulation analysis.

2. Matrices using Bray-Curtis distance (NMDS and DCA) underperformed regarding other methods with biased estimates of the simulated variance components, heritability and microbiability.

3. The microbiome-genotype interaction is not properly accounted for using simple interaction methods, more sophisticated statistical methods must be developed.

4. The genomic breeding values were accurately predicted when a microbiome effect was accounted for; however, the benchmark matrix and the canonical ordination methods of CCA and RDA showed higher accuracies than MDS, DCA and NMDS.

5. The benchmark matrix and canonical ordination methods of CCA and RDA were preferred at estimating the microbiome effect in the simulations.

6. Microbiability estimates for feed efficiency were low although consistent between ordination methods that performed better in the simulation of microbiability (Ross, MDS, RDA, CCA), indicating that it may be possible to include a microbiome effect in the statistical analysis of complex traits.

7. Deviance information criteria (DIC), log-likelihood evaluated at posterior mean, and the estimated effective number of parameters was slightly lower for a model ignoring the microbiome-genotype interaction; however differences between both models were not substantial, consequently there is not enough evidence to recommend one of either models.

8. Microbiota relationship matrices obtained with CCA and RDA were suitable both for variance components estimation as well as for estimated breeding and microbiome values, hence based on results obtained from this master thesis, I recommend to use those canonical ordination methods to build the microbiota relationship matrix to predict complex traits.

REFERENCES

8. REFERENCES

Almeida, M., M. Pop, E. Le Chatelier, E. Prifti, N. Pons, A. Ghozlane, and S.D. Ehrlich. 2016. Capturing the most wanted taxa through cross-sample correlations. ISME J. 10:2459–2467. doi:10.1038/ismej.2016.35.

Arslan, N. 2014. Obesity, fatty liver disease and intestinal microbiota. World J Gastroenterol 20:16452–16463. doi:10.3748/wjg.v20.i44.16452.

Aßhauer, K.P., B. Wemheuer, R. Daniel, and P. Meinicke. 2015. Tax4Fun: Predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics 31:2882–2884. doi:10.1093/bioinformatics/btv287.

Beaumont, M., M.A. Jackson, C.J. Steves, T.D. Spector, and J.T. Bell. 2016. Heritable components of the human fecal microbiome are associated with visceral fat. Gut Microbes 1–7. doi:10.1080/19490976.2017.1356556.

Benson, A.K. 2016. The gut microbiome—an emerging complex trait. Nat. Genet. 48:1301–1302. doi:10.1038/ng.3707.

Benson, A.K., S.A. Kelly, R. Legge, F. Ma, S.J. Low, J. Kim, M. Zhang, P.L. Oh, D. Nehrenberg, K. Hua, S.D. Kachman, E.N. Moriyama, J. Walter, D.A. Peterson, and D. Pomp. 2010. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc. Natl. Acad. Sci. 107:18933–18938. doi:10.1073/pnas.1007028107.

Berg, A., R. Meyer, and J. Yu. 2004. Deviance Information Criterion for Comparing Stochastic Volatility Models. J. Bus. Econ. Stat. 22:107–120. doi:10.1198/073500103288619430.

Berry, D.P., and J.J. Crowley. 2013. CELL BIOLOGY SYMPOSIUM: Genetics of feed efficiency in dairy and beef cattle. J. Anim. Sci 91:1594–1613. doi:10.2527/jas2012-5862.

Blekhman, R., J.K. Goodrich, K. Huang, Q. Sun, R. Bukowski, J.T. Bell, T.D. Spector, A. Keinan, R.E. Ley, D. Gevers, and A.G. Clark. 2015. Host genetic variation impacts microbiome composition across human body sites. Genome Biol. 16:191. doi:10.1186/s13059-015-0759-1.

Bodilis, J., S. Nsigue-Meilo, L. Besaury, and L. Quillet. 2012. Variable Copy Number, Intra-Genomic Heterogeneities and Lateral Transfers of the 16S rRNA Gene in Pseudomonas. PLoS One 7:e35647. doi:10.1371/journal.pone.0035647.

Bonder, M.J., A. Kurilshikov, E.F. Tigchelaar, Z. Mujagic, F. Imhann, A.V. Vila, P.

Deelen, T. Vatanen, M. Schirmer, S.P. Smeekens, D. V Zhernakova, S.A. Jankipersadsing, M. Jaeger, M. Oosting, M.C. Cenit, A.A.M. Masclee, M.A. Swertz, Y. Li, V. Kumar, L. Joosten, H. Harmsen, R.K. Weersma, L. Franke, M.H. Hofker, R.J. Xavier, D. Jonkers, M.G. Netea, C. Wijmenga, J. Fu, and A. Zhernakova. 2016. The effect of host genetics on the gut microbiome. Nat. Genet. 48:1407–1412. doi:10.1038/ng.3663.

Camarinha-Silva, A., M. Maushammer, R. Wellmann, M. Vital, S. Preuss, and J. Bennewitz. 2017. Host Genome Influence on Gut Microbial Composition and Microbial Prediction of Complex Traits in Pigs. Genetics 206:1637–1644. doi:10.1534/genetics.117.200782.

Chakravorty, S., D. Helb, M. Burday, and N. Connell. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods 69:330–339. doi:10.1016/j.mimet.2007.02.005.A.

Clarridge, J.E. 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin. Microbiol. Rev. 17:840–862. doi:10.1128/CMR.17.4.840.

Cole, J.R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R.J. Farris, A.S. Kulam-Syed-Mohideen, D.M. McGarrell, T. Marsh, G.M. Garrity, and J.M. Tiedje. 2009. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37:141–145. doi:10.1093/nar/gkn879.

DeSantis, T.Z., P. Hugenholtz, N. Larsen, M. Rojas, E.L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G.L. Andersen. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72:5069–5072. doi:10.1128/AEM.03006-05.

Deusch, S., A. Camarinha-Silva, J. Conrad, U. Beifuss, M. Rodehutscord, and J. Seifert. 2017. A Structural and Functional Elucidation of the Rumen Microbiome Influenced by Various Diets and Microenvironments. Front. Microbiol. 8:1605. doi:10.3389/fmicb.2017.01605.

Franzosa, E.A., T. Hsu, A. Sirota-Madi, A. Shafquat, G. Abu-Ali, X.C. Morgan, and C. Huttenhower. 2015. Sequencing and beyond: integrating molecular “omics” for microbial community profiling.. Nat. Rev. Microbiol. 13:360–72. doi:10.1038/nrmicro3451.

Gibbons, S. 2015. Protocol: 16S Amplicon Sequencing | Method. Accessed September 26, 2017. http://www.methodquarterly.com/2015/11/protocol-16s-amplicon-

sequencing/.

Glickman, M.E., S.R. Rao, and M.R. Schultz. 2014. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J. Clin. Epidemiol. 67:850–857. doi:10.1016/j.jclinepi.2014.03.012.

Gonzalez-Recio, O., I. Zubiria, A. García-Rodríguez, A. Hurtado, and R. Atxaerandio. 2017. Short communication: Signs of host genetic regulation in the microbiome composition in 2 dairy breeds: Holstein and Brown Swiss.. J. Dairy Sci. 101:1–8. doi:10.3168/jds.2017-13179.

Goodall, D. 1954. Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. Aust. J. Bot. 2:304. doi:10.1071/BT9540304.

Goodrich, J.K., J.L. Waters, A.C. Poole, J.L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, J.T. Bell, T.D. Spector, A.G. Clark, and R.E. Ley. 2014. Human genetics shape the gut microbiome. Cell 159:789–799. doi:10.1016/j.cell.2014.09.053.

Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325.

Greenacre, M.J. 2017. Correspondence Analysis in Practice. Third Edit. CRC Press, Taylor and Francis Group.

Guyader, J., M. Eugène, P. Nozière, D.P. Morgavi, M. Doreau, and C. Martin. 2014. Influence of rumen protozoa on methane emission in ruminants: a meta-analysis approach. Animal 8:1816–1825. doi:10.1017/S1751731114001852.

de Haas, Y., J. Windig, M. Calus, J. Dijkstra, M. de Haan, A. Bannink, and R. Veerkamp. 2011. Genetic parameters for predicted methane production and potential for reducing enteric emissions through genomic selection. J. Dairy Sci. 94:6122–6134. doi:10.3168/jds.2011-4439.

Hall, A.B., A.C. Tolonen, and R.J. Xavier. 2017. Human genetic variation and the gut microbiome in disease. Nat. Rev. Genet. 18:690–699. doi:10.1038/nrg.2017.63.

Hansen, E.E., C.A. Lozupone, F.E. Rey, M. Wu, J.L. Guruge, A. Narra, J. Goodfellow, J.R. Zaneveld, D.T. McDonald, J.A. Goodrich, A.C. Heath, R. Knight, and J.I. Gordon. 2011. Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. Proc. Natl. Acad. Sci. 108:4599–4606. doi:10.1073/pnas.1000071108.

Hartstra, A. V., K.E.C. Bouter, F. Bäckhed, and M. Nieuwdorp. 2015. Insights Into the Role of the Microbiome in Obesity and Type 2 Diabetes. Diabetes Care 38:159–165.

doi:10.2337/dc14-0769.

Hill, M.O., and H.G. Gauch. 1980. Detrended correspondence analysis: An improved ordination technique. Vegetatio 42:47–58. doi:10.1007/BF00048870.

Hirschfeld, H.O., and J. Wishart. 1935. A Connection between Correlation and Contingency. Math. Proc. Cambridge Philos. Soc. 31:520. doi:10.1017/S0305004100013517.

Holland, S.M. 2008. DETRENDED CORRESPONDENCE ANALYSIS (DCA). Dep. Geol. Univ. Georg. Athens, GA 30602-2501.

Imhann, F., A. Vich Vila, M.J. Bonder, J. Fu, D. Gevers, M.C. Visschedijk, L.M. Spekhorst, R. Alberts, L. Franke, H.M. van Dullemen, R.W.F. Ter Steege, C. Huttenhower, G. Dijkstra, R.J. Xavier, E.A.M. Festen, C. Wijmenga, A. Zhernakova, and R.K. Weersma. 2016. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. Gut. doi:10.1136/gutjnl-2016-312135.

Jami, E., B.A. White, and I. Mizrahi. 2014. Potential role of the bovine rumen microbiome in modulating milk composition and feed efficiency. PLoS One 9. doi:10.1371/journal.pone.0085423.

Janda, J.M., and S.L. Abbott. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. J. Clin. Microbiol. 45:2761–2764. doi:10.1128/JCM.01228-07.

Jovel, J., J. Patterson, W. Wang, N. Hotte, S. O'Keefe, T. Mitchel, T. Perry, D. Kao, A.L. Mason, K.L. Madsen, and G.K.S. Wong. 2016. Characterization of the gut microbiome using 16S or shotgun metagenomics. Front. Microbiol. 7:1–17. doi:10.3389/fmicb.2016.00459.

Kaps, M., and W. Lamberson. 2004. Biostatistics for Animal Science. 1st ed. Cambridge, MA 02139, USA; Oxfordshire OX10 8DE UK CABI.

Khatri, P., and S. Draghici. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinforma. Rev. 21:3587–3595. doi:10.1093/bioinformatics/bti565.

King, E.E., R.P. Smith, B. St-Pierre, and A.-D.G. Wright. 2011. Differences in the rumen methanogen populations of lactating Jersey and Holstein dairy cows under the same diet regimen.. Appl. Environ. Microbiol. 77:5682–7. doi:10.1128/AEM.05130-11.

Knight, R., J. Jansson, D. Field, N. Fierer, N. Desai, J.A. Fuhrman, P. Hugenholtz, D. van der Lelie, F. Meyer, R. Stevens, M.J. Bailey, J.I. Gordon, G.A. Kowalchuk, and J.A.

Gilbert. 2012. Unlocking the potential of metagenomics through replicated experimental design. Nat. Biotechnol. 30:513–520. doi:10.1038/nbt.2235.

Langille, M.G.I., J. Zaneveld, J.G. Caporaso, D. McDonald, D. Knights, J.A. Reyes, J.C. Clemente, D.E. Burkepile, R.L. Vega Thurber, R. Knight, R.G. Beiko, and C. Huttenhower. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. 31:814–821. doi:10.1038/nbt.2676.

Lederberg, B.J., and A.T. McCray. 2001. ' Ome Sweet ' Omics-- A Genealogical Treasury of Words. Sci. 15:8.

Legendre, P., and M.J. Anderson. 1999. Distance-Based Redundancy Analysis: Testing Multispecies Responses in Multifactorial Ecological Experiments. Ecol. Monogr. 69:1. doi:10.2307/2657192.

Legendre, P., and L. Legendre. 1998. Numerical Ecology. 2nd ed. Elsevier, Amsterdam.

Legendre, P., and L. Legendre. 2012. Numerical Ecology. 3rd ed. Elsevier, Amsterdam.

Li, F., and L.L. Guan. 2017. Metatranscriptomic Profiling Reveals Linkages between the Active Rumen Microbiome and Feed Efficiency in Beef Cattle.. Appl. Environ. Microbiol. 83:e00061-17. doi:10.1128/AEM.00061-17.

Lima, F.S., G. Oikonomou, S.F. Lima, M.L.S. Bicalho, E.K. Ganda, J.C. de O. Filho, G. Lorenzo, P. Trojacanec, and R.C. Bicalhoa. 2015. Prepartum and postpartum rumen fluid microbiomes: characterization and correlation with production traits in dairy cows.. Appl. Environ. Microbiol. 81:1327–37. doi:10.1128/AEM.03138-14.

De Los Campos, G., and P. Perez Rodriguez. 2016. Bayesian Generalized Linear Regression. R package.

Malmuthuge, N., and L.L. Guan. 2016. Gut microbiome and omics: a new definition to ruminant production and health. Anim. Front.. doi:10.2527/af.2016-0017.

Mao, S., M. Zhang, J. Liu, and W. Zhu. 2015. Characterising the bacterial microbiota across the gastrointestinal tracts of dairy cattle: membership and potential function. Sci. Rep. 5:16116. doi:10.1038/srep16116.

Marchesi, J.R., and J. Ravel. 2015. The vocabulary of microbiome research: a proposal. Microbiome 3. doi:10.1186/s40168-015-0094-5.

McCune, B., J.B. Grace, and D.L. Urban. 2002. Analysis of ecological communities. J. Exp. Mar. Bio. Ecol. 289. doi:10.1016/S0022-0981(03)00091-1.

McKnite, A.M., M.E. Perez-Munoz, L. Lu, E.G. Williams, S. Brewer, P.A. Andreux, J.W.M. Bastiaansen, X. Wang, S.D. Kachman, J. Auwerx, R.W. Williams, A.K.

Benson, D.A. Peterson, and D.C. Ciobanu. 2012. Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. PLoS One 7. doi:10.1371/journal.pone.0039191.

Morgan, X.C., and C. Huttenhower. 2012. Chapter 12: Human Microbiome Analysis. PLoS Comput. Biol. 8:e1002808. doi:10.1371/journal.pcbi.1002808.

Nkrumah, J.D., E.K. Okine, G.W. Mathison, K. Schmid, C. Li, J.A. Basarab, M.A. Price, Z. Wang, and S.S. Moore. 2006. Relationships of feedlot feed efficiency, performance, and feeding behavior with metabolic rate, methane production, and energy partitioning in beef cattle.. J. Anim. Sci. 84:145–53.

Norman, J.M., S.A. Handley, M.T. Baldridge, L. Droit, C.Y. Liu, B.C. Keller, A. Kambal, C.L. Monaco, G. Zhao, P. Fleshner, T.S. Stappenbeck, D.P.B. Mcgovern, A. Keshavarzian, E.A. Mutlu, J. Sauk, D. Gevers, R.J. Xavier, D. Wang, M. Parkes, and H.W. Virgin. 2015. Disease-specific Alterations in the Enteric Virome in Inflammatory Bowel Disease. Cell. January 29:447–460. doi:10.1016/j.cell.2015.01.002.

Oulas, A., C. Pavloudi, P. Polymenakou, G.A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis, and I. Iliopoulos. 2015. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies.. Bioinform. Biol. Insights 9:75–88. doi:10.4137/BBI.S12462.

Pickering, N.K., V.H. Oddy, J. Basarab, K. Cammack, B. Hayes, R.S. Hegarty, J. Lassen, J.C. Mcewan, S. Miller, C.S. Pinares-Patiño, and Y. De Haas. 2015. Animal board invited review: genetic possibilities to reduce enteric methane emissions from ruminants. Animal 9:1431–1440. doi:10.1017/S1751731115000968.

Poretsky, R., L.M. Rodriguez-R, C. Luo, D. Tsementzi, and K.T. Konstantinidis. 2014. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One 9. doi:10.1371/journal.pone.0093827.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F.O. Glöckner. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41:590–596. doi:10.1093/nar/gks1219.

Ramette, A. 2007. Multivariate analyses in microbial ecology. FEMS Microbiol. Ecol. 62:142–160. doi:10.1111/j.1574-6941.2007.00375.x.

Ranjan, R., A. Rani, A. Metwally, H.S. McGee, and D.L. Perkins. 2016. Analysis of the

microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem. Biophys. Res. Commun. 469:967–977. doi:10.1016/j.bbrc.2015.12.083.

Rodicio, M.D.R., and M.D.C. Mendoza. 2004. Identificación bacteriana mediante secuenciación del ARNr 16S: fundamento, metodología y aplicaciones en microbiología clínica. Enferm. Infecc. Microbiol. Clin. 22:238–245. doi:10.1157/13059055.

Roehe, R., R.J. Dewhurst, C.A. Duthie, J.A. Rooke, N. McKain, D.W. Ross, J.J. Hyslop, A. Waterhouse, T.C. Freeman, M. Watson, and R.J. Wallace. 2016. 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.. doi:10.1371/journal.pgen.1005846.

Ross, E.M., P.J. Moate, L.C. Marett, B.G. Cocks, and B.J. Hayes. 2013. Metagenomic Predictions: From Microbiome to Complex Health and Environmental Phenotypes in Humans and Cattle. PLoS One 8:e73056. doi:10.1371/journal.pone.0073056.

Ruokolainen, L., and G. Blanchet. 2014. Introduction to Ecological Multivariate Analysis. Course notes 1–79.

Shepard, R.N. 1966. Metric structures in ordinal data. J. Math. Psychol. 3:287–315. doi:10.1016/0022-2496(66)90017-4.

Sorensen, D. 2004. An Introductory Overwiew of Model Comparison and Related Topics. Course notes Model choice, Model Assess. Relat. Top. from a likelihood a Bayesian Perspect. 47.

Sorensen, D., and D. Gianola. 2002. Likelihood of Bayesian, and MCMC Methods in Quantitative Genetics. 1st ed. K. Dietz, M. Gail, K. Krickeberg, J. Samet, and A. Tsiatis, ed. Springer-Verlag New York, Inc, New York.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B (Statistical Methodol. 64:583–639. doi:10.1111/1467-9868.00353.

Spurlock, D.M., J.C.M. Dekkers, R. Fernando, D.A. Koltes, and A. Wolc. 2012. Genetic parameters for energy balance, feed efficiency, and related traits in Holstein cattle.. J. Dairy Sci. 95:5393–5402. doi:10.3168/jds.2012-5407.

Stevens, A. 2004. DIC: Deviance Information Criterion. MRC Biostat. Unit, Cambridge, UK.

Tapio, I., T.J. Snelling, F. Strozzi, and R.J. Wallace. 2017. The ruminal microbiome associated with methane emissions from ruminant livestock.. J. Anim. Sci. Biotechnol. 8:7. doi:10.1186/s40104-017-0141-0.

The NIH HMP Working Group. 2009. The NIH Human Microbiome Project. Genome Res. 19:2317–2323. doi:10.1101/gr.096651.109.

Thomas, T., J. Gilbert, and F. Meyer. 2012. Metagenomics - a guide from sampling to data analysis. Microb. Inform. Exp. 2:3. doi:10.1186/2042-5783-2-3.

Tortoli, E. 2003. Impact of genotypic studies on mycobacterial taxonomy: The new mycobacteria of the 1990s. Clin. Microbiol. Rev. 16:319–354. doi:10.1128/CMR.16.2.319-354.2003.

Turnbaugh, P.J., and J.I. Gordon. 2009. The core gut microbiome, energy balance and obesity. J Physiol 58717:4153–4158. doi:10.1113/jphysiol.2009.174136.

Turnbaugh, P.J., M. Hamady, T. Yatsunenko, B.L. Cantarel, R.E. Ley, M.L. Sogin, W.J. Jones, B. a Roe, P. Jason, M. Egholm, B. Henrissat, A.C. Heath, R. Knight, and J.I. Gordon. 2009. A core gut microbiom in obese and lean twins. Nature 457:480–484. doi:10.1038/nature07540.A.

Ursell, L.K., J.L. Metcalf, L. Wegener Parfrey, and R. Knight. 2013. Definig the Human Microbiome. NIH Manuscripts 70:1–12. doi:10.1111/j.1753-4887.2012.00493.x.Defining.

Vanraden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980.

Wang, Y., and P.Y. Qian. 2009. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PLoS One 4. doi:10.1371/journal.pone.0007401.

Weimer, P.J., D.M. Stevenson, and D.R. Mertens. 2010. Shifts in bacterial community composition in the rumen of lactating dairy cows under milk fat - depressing conditions. J. Dairy Sci. 93:265–278. doi:10.3168/jds.2009-2206.

Welkie, D.G., D.M. Stevenson, and P.J. Weimer. 2010. ARISA analysis of ruminal bacterial community dynamics in lactating dairy cows during the feeding cycle. Anaerobe 16:94–100. doi:10.1016/J.ANAEROBE.2009.07.002.

Woese, C.R., E. Stackebrandt, T.J. Macke, and G.E. Fox. 1985. A phylogenetic definition of the major eubacterial taxa.. Syst. Appl. Microbiol. 6:143–51.

Xia, Y., and J. Sun. 2017. Hypothesis testing and statistical analysis of microbiome. Genes Dis. 4:138–148. doi:10.1016/J.GENDIS.2017.06.001.

Yang, J., B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.E. Goddard, and P.M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565–569. doi:10.1038/ng.608.

Yatsunenko, T., F. Rey, M. Manary, I. Trehan, M. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, and R. Baldassano. 2012. Human gut microbiome viewed across age and geography. Nat. 2012 486:222–227. doi:10.1038/nature11053.Human.

Zhao, L., G. Wang, P. Siegel, C. He, H. Wang, W. Zhao, Z. Zhai, F. Tian, J. Zhao, H. Zhang, Z. Sun, W. Chen, Y. Zhang, and H. Meng. 2013. Quantitative Genetic Background of the Host Influences Gut Microbiomes in Chickens. Sci. Rep. 3:1163. doi:10.1038/srep01163.

Zoetendal, E.G., A.D.L. Akkermans, W.M.A. Vliet, J.A.G.M. De Visser, W.M. De Vos, E.G. Zoetendal, A.D.L. Akkermans, W.M. Akkermans-van, J.A.G.M. De Visser, W.M. De Vos, T. Host, G. Affects, E.G. Zoetendal, A.D.L. Akkermans, and W.M.A. Vliet. 2001. Microbial Ecology in Health and Disease The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract. Microb. Ecol. Health Dis. 13:129–134.