https://doi.org/10.22319/rmcp.v11i4.5202

Review

Molecular tools used for metagenomic analysis. Review

Nohemí Gabriela Cortés-López [a]

Perla Lucía Ordóñez-Baquera [a*]

Joel Domínguez-Viveros [a]

[a] Universidad Autónoma de Chihuahua, Facultad de Zootecnia y Ecología. Chihuahua, México.

*Corresponding author: plordonez@uach.mx

Abstract:

Metagenomics uses molecular biology techniques to analyze the diversity of microbial genomes (metagenomes). Metagenome diversity has been analyzed using molecular markers to classify bacteria and archaea into taxonomic groups at the genus level. Among the most widely used molecular markers are ribosomal genes, genes encoding subunits of cytochrome C, and certain constitutive genes (*gyrB*, *rpoB*, *rpoD*, *recA*, *atpD*, *infB*, *groEL*, *pmoA*, *sodA*). The most widely used marker for classifying bacteria and metagenomic samples is the 16S rRNA gene, although it does not allow certain sequences to be properly classified. However, all the sequences of the hypervariable regions can be identified with the sequencing of the complete 16S rRNA gene, and, therefore, this molecular marker has made it possible to classify them at the species taxonomic level. Next generation sequencing, also called mass sequencing or high throughput sequencing, has helped to describe complex metagenomes such as those of environmental samples, which have an ecological importance, as well as metagenomes growing in extreme environments. They have also proved helpful in studies related to animal and human health, and in the agro-food field. Specifically, both the 16S rRNA molecular marker and high throughput sequencing combined with bioinformatic tools for metagenomic analysis have been used to describe the ruminal metagenome, a microbial community of great importance because it is involved in animal production of meat and milk.

Despite the many studies that have been conducted in this field, some microorganisms still remain to be discovered and characterized.

Key words: Molecular Marker, 16S rRNA Gene, Metagenomics, Microbial diversity, High throughput sequencing.

Received: 17/12/2018

Accepted: 23/10/2019

Introduction

Metagenomics is based on the use of molecular biology techniques to analyze the diversity of microbial genomes, also called metagenomes, from environmental samples. The microbial diversity of metagenomes has been analyzed using the 16S rRNA gene, which encodes for the ribosomal RNA that forms the small subunit of the ribosomes. This gene comprises preserved and variable regions in bacteria and archaea. The 16S rRNA gene has been used as a molecular marker, since it allows the classification of bacteria and archaea into taxonomic groups according to families or genera.

The first studies of microbial diversity in environmental samples were carried out using culture-dependent methods, where only those microorganisms that could be isolated in the laboratory were studied. Through the advance of molecular biology techniques, it has been possible to analyze microbial diversity through the use of independent culture methods, obtaining more precise information about bacterial genomes. One of the most widely used methods is PCR amplification of 16S rRNA gene fragments, in some cases followed by denaturing gradient gel electrophoresis (DGGE). These techniques have been used to analyze ruminal bacterial diversity, changes in the microbial community, and gene expression after changes in the ruminants' diet[1,2]. Another advance that has allowed a broader analysis of microbial diversity in the rumen is the targeted sequencing of the variable regions of the 16S gene in order to differentiate microorganisms that are phylogenetically very close, analyze the genes and genomes that degrade the biomass in the rumen, characterize the rumen microbiota, and study the effects of yeasts on bacterial diversity in the rumen[3,4,5].

The recent development of metagenomics has allowed the study of microbial diversity in environmental samples by isolating and analyzing the total genetic material present in an environmental sample[6,7]. At the beginning, this strategy was used to search for new enzymes

with biotechnological potential, extracting the total DNA contained in an environmental sample, fragmenting it and cloning genes of different size in vectors such as plasmids (15 kb), phages (up to 20 kb), phosmids and cosmids (up to 40 kb), as well as Artificial Bacterial Chromosomes (for larger fragments). These vectors were inserted into different host strains, and fluorogenic substrates were used as expression indicators. However, in the functional search for genes through clones, protein expression and enzymatic activity were of a small magnitude[8,9,10].

A crucial part in the construction of metagenomic libraries is the extraction of the nucleic acids from the sample. There are two main strategies for metagenomic DNA extraction: chemical treatment and direct lysis with mechanical methods. Both methods have advantages and disadvantages. DNA of greater microbiological diversity is recovered with mechanical lysis than with chemical treatment; however, chemical treatment allows obtaining DNA of greater molecular weight. Regarding RNA extraction, the same extraction methods are used for any expression analysis in which RNAsase inhibitors are included, and it is recommended to freeze the samples at -80 ºC immediately after collection to avoid RNA degradation[9].

To select the ideal extraction method, the type of sample, the nucleic acid to be purified and the type of analysis to be performed must be taken into account. Different strategies have been used for metagenomic analysis. Within the mechanical methods, magnetic beads have been used for oral, dermal or fecal samples, as well as samples of soil and water, from which high-quality sequences have been obtained[11]. For the analysis of ruminal microbiomes, methods combining magnetic bead extraction (mechanical lysis) with extraction columns (chemical treatment) have been used to purify ruminal microbial DNA[11,12]. This combination increased extraction performance over the use of separate magnetic beads and extraction columns. Other identification methods use Stable Isotope Probing (SIP), which identify the microorganisms that incorporate these isotopes through the use of marked substrates. In particular, the nucleic acid stable isotope probe technique (Nucleic acids-SIP) uses substrates with ^{13}C and/or ^{15}N isotopes, which are incorporated into bacterial genomes and can thus be traced[13]. Other substrates with stable isotope probes are 13CH3-OH, 13C-phenol and 5-bromo-2-deoxyuridine. However, limitations of the use of substrates marked with stable isotopes include the crosslinking and recycling of the isotopes within the microbial community, resulting in the loss of specific enrichment of the analyzed microorganisms[13].

Techniques have also been developed to identify genes that change their expression levels during various biological processes. For example, Suppression Subtractive Hybridization (SSH) has been used to identify variations between complex DNA samples such as those in the ruminal environment[9,13]. Differential expression analysis allows to compare the gene expression profile of a microbial community before and after being exposed to a specific condition and/or substrate and, thus, to identify important genes that exhibit changes in gene

expression profiles due to the effect of such condition and/or substrate[13]. Another technique that is widely used in gene expression studies is microarrays, which offer the advantage of rapidly identifying and characterizing a large number of clones. Although microarrays can be used to identify a large number of conserved genes, they depend on known sequences previously reported in databases, thus eliminating the possibility of identifying new genes[8,10]. More recently, mass sequencing has been used to obtain as much information as possible about the metagenome present in a sample. One of the first works with massive sequencing was the identification of the metagenome of the Sargasso Sea, where 1,045 trillion base pairs of non-redundant sequences were generated, noted and analyzed in order to identify the genetic content, diversity and relative abundance of the microorganisms. It was estimated that the data obtained came from at least 1,800 genomic species that included 148 phylotypes of unknown bacteria and more than 782 genes never before described that code for rhodopsin-like photoreceptors[10,14].

The massive sequencing of the metagenome by "shotgun" has the characteristic of sequencing all the DNA present in the sample so that the microorganisms can be classified taxonomically up to the species level. Furthermore, with the sequences obtained by this type of sequencing, genes with functions never before described can be discovered, and even sequences belonging to the 16S rRNA gene can be selected for taxonomic annotation. These classifications are made with the use of bioinformatic tools that search for homology with the sequences analyzed in different existing databases[15]. Specifically in ruminal environments, metagenomic libraries have been analyzed in order to evaluate the effects of diets on ruminal microbiome by means of metagenomic profiles, and the 16S rRNA gene marker has been used to determine and classify the microbial diversity of the sequences[3,5]. However, some of the sequences of these samples have not been adequately classified; therefore, using at least one molecular phylogenetic marker other than the 16S rRNA gene may improve taxonomic classification[15].

In the present work it was reviewed the tools used for metagenome analysis, ranging from classical molecular markers to those used with data obtained from massive sequencing, with an emphasis on metagenomes from ruminal environments.

Molecular markers for metagenomic analysis

A molecular marker is a segment of DNA that corresponds to a non-coding gene or regions of the genome, these segments of DNA allow different variants (alleles) to be identified and are located at a particular site on the chromosomes (locus). The differences obtained in these DNA fragments are known as polymorphisms and can be detected by hybridization of nucleic

acid sequences, nucleotide sequencing, comparison of the length of the fragments produced by the polymerase chain reaction (PCR) and through sites recognized by restriction enzymes. Molecular markers can be used to classify taxonomic groups, populations, families or individuals in both eukaryotes and prokaryotes[16,17]. Various molecular markers have been used in genetic studies in domestic animals, in wildlife, in endangered species, and in forensic and paternity tests. The best known are RFLPs, mini-satellites, AFLPs, RAPDs, microsatellites and SNPs (Table 1).

The most relevant characteristics that molecular markers must have in order to optimize metagenomic studies include (1) that they are single copy genes (genes that have only one or two copies in the entire genome), as they provide less uncertainty than markers for genes with multiple copies (genes with repeated copies in the genome); (2) that the sequence of the marker gene is easily aligned to facilitate phylogenetic analysis; (3) that the proportion of the gene replacement region is sufficient to provide information needed for classification; (4) that primers are selective to amplify the marker gene, but not universal, in order to avoid false positives; (5) that there is no excessive variation in the marker sequence that limits the determination of ancestry. The genes that are used as molecular markers to classify microorganisms are described below.

Ribosomal genes

Ribosomal RNA genes are considered the ideal tool for taxonomic classification since they are highly conserved and evolutionarily stable genes, but they contain hypervariable regions. The sequencing of these regions has generated large databases that assist in the taxonomic classification[18]. Ribosomes of bacteria and archaea consist of two subunits: a small subunit containing a single type of RNA (16S) and a large subunit containing two types of RNA (5S and 23S)[17].

16S rRNA. This gene is also designated 16S rRNA, but the American Society for Microbiology (ASM) has decided to use the term "16S rRNA" in order to standardize the information**.** It has an approximate sequence length of 1,550 bp and contains variable and preserved regions with unique oligonucleotide sequences for each phylogenetic group[18,19]. The comparison of 16S rRNA gene sequences of unknown bacteria with known sequences in databases is of great help in classifying bacteria at the genus level and has even identified species in some cases[20,21].

5S rDNA. It is a gene of approximately 120 nucleotides in length and is found in virtually all ribosomes except mitochondria, some fungi, higher animals and most protists. Although

the sequence of this gene is highly conserved, the reliability of this gene as a marker is questioned because its length is very small and therefore does not offer sufficient resolution to contribute significantly to the understanding of phylogenetic relationships between taxa[17].

23S rDNA. It is a gene of approximately 3,000 nucleotides in length that is located in the large subunit of the ribosomes in prokaryotes. This gene has larger insertions and deletions than the 16S rRNA gene. Stable insertions and deletions of some bases in the 23S rDNA gene are common characteristics in some classes and subclasses of bacteria. These changes complicate the analyses, since the different positions cannot be considered for correct phylogenetic classifications[22]. The 23S rDNA gene has been used in conjunction with the 16S rRNA gene for the taxonomic classification of non-cultivable bacteria. The intergenic spacer (IGS) located in the 16S-23S region, which is very variable, has also been used to differentiate between two strains belonging to the same subspecies[22,23].

Genes encoding subunits of cytochrome C

Cytochrome Oxidase I/II (COI/II)**.** The cytochrome C oxidase enzyme is an electron transport chain protein found both in bacteria and in the mitochondria of eukaryotic organisms. The COI and COII genes encode for two of the seven polypeptide subunits of the cytochrome C oxidase complex. The COI gene evolves more slowly compared to other mitochondrial genes and is widely used in phylogenetic studies[17].

Genes encoding proteins with preserved functions

In studies that have found a greater diversity of microorganisms, molecular community analysis techniques based on the 16S rRNA gene have been used, supported by multilocus sequence analysis (MLSA) studies, which involve the sequencing of several genes encoding proteins with conserved functions (housekeeping genes) to evaluate the diversity in collections of isolated strains[24]. In these studies, the partial sequences of genes that encode for proteins with conserved functions are used to generate phylogenetic trees and, subsequently, to solve phylogenies. The main disadvantage of using the 16S rRNA gene as a phylogenetic marker is its insufficient resolution at the species level. However, the use of a complementary phylogenetic analysis based on protein coding genes[25] allows to increase the resolution of phylogenies at an infra-generic level and to determine new strains. Over 50

individual MLSA schemes are available, and MLSA databases (http://www.mlst.net/ and http://www.pubmlst.org) can also be used to identify microbial sequences not known at the species level[24,26].

The genes that have been used in MLSA are those that encode ubiquitous enzyme subunits, such as the of DNA gyrase subunit β (gyrB), the RNA polymerase subunit β (rpoB), the sigma 70 factor (sigma D) of RNA polymerase (RpoD) , the recombinase A (recA), the β subunit of ATP synthase F0F1 (atpD), the translation initiation factor IF-2 (infB), the tRNA modification GTPase ThdF or TrmE (thdF) and the chaperonin GroEL (groEL)[24,26].

The particulate methane-monooxygenase subunit β (pmoA) has been used as a functional marker for the detection of aerobic methanotrophs. Methane-monooxygenase is the enzyme responsible for the initial conversion stage from methane to methanol. Two forms of this enzyme are known, soluble methane-monoxygenase (sMMO) and a membrane-bound enzyme, particulate methane-monoxygenase (pmoA). The pmoA gene is the most frequently used marker, as it is present in most methanotrophic aerobic bacteria. It is also present in anaerobic denitrifying bacteria[27]. Another marker that can be used for the detection of methanotrophs is the mxaF gene that encodes the major subunit of methanol dehydrogenase[27,28].

As an example of this approach, can be cited the work of Sánchez-Herrera *et al*[26], who have used the 16S rRNA gene as a molecular reference marker to identify and classify strains of the genus *Nocardia* at the genus level. However, being a gene with multiple copies generates problems in the identification of isolated strains of clinical cases. After testing other genes through PCR amplification of their segments: sodA (gene encoding the enzyme superoxide dismutase), hsp65 (heat shock protein), secA1 (preprotein translocase subunit secA), gyrB (DNA gyrase subunit β), rpoB (RNA polymerase subunit β) and the 16S-23S intergenic spacer, the authors were able to discriminate only between closely related species of *Nocardia* using the sodA gene. The 386 bp fragment of the sodA gene includes variable regions, with 4 and 5 bp segments, and has the potential to be used as a molecular marker. In conclusion, although there is a great diversity of molecular markers to analyze microbial communities, so far, the gold standard for the classification of sequences obtained from samples remains the 16S rRNA gene.

The use of mass sequencing in metagenomics

Although metagenomic analysis started with the use of different molecular markers such as AFLP, RAPDs, 16S rRNA etc. (Table 1), some of these markers have been observed to

improve their efficiency when the technique used to identify them includes their sequencing instead of characterizing them by means of reactions with restriction enzymes and/or amplification by PCR. From its inception, DNA sequencing with Sanger's technology has had a major impact on virtually every branch of the biological sciences, including microbial community studies. Currently, the use of Sanger sequencing can generate up to 96 sequences per run with an average length of 650 bp, which may be sufficient for phylogenetic marker analysis[15]. This type of study is known as first generation sequencing and results in high quality sequences of a length between 500 and 1,000 bp. However, its disadvantage is that the proportion of molecular markers that can be sequenced in a run, compared to the total number of microorganisms present in a metagenomic sample, is very low[11].

Table 1: Molecular methods used in genetic studies

Molecular Marker	**Characteristics**	**Reference**
RFLP (restriction fragment length polymorphism)	It is based on nucleotide changes in a genome that occur at a restriction enzyme recognition site. In forensic science it has been used to prove whether tissues from crime scenes (blood, skin, sperm, etc.) belong to a suspect. In the management of animal breeds, it is used to track progeny, as well as for paternity testing and disease diagnosis.	Khlestkina[16] Wakchaure *et al*[50]
Minisatellites or VNTR (variations in the number of tandem repeats)	They are short sequences of 10 to 60 bp, repeated in variable number at one or more sites of the genome. They have been used to identify paternal lineages in individuals and to assess genetic diversity in domestic animal, wildlife and grass populations.	Kumar *et al*[51] Lang *et al*[52]
AFLP (amplified fragment length polymorphism)	It is the amplification of digested genomic fragments with restriction enzymes that recognize sequences dispersed throughout the genome. It has been used for "fingerprinting" DNA studies, to clone and map specific DNA sequences and to make genetic maps.	Khlestkina[16] Kumar *et al*[51]

RAPD (randomly amplified polymorphic DNA)	They use short, arbitrarily sequenced primers to direct an amplification reaction in discrete regions of the genome, resulting in fragments of various sizes. They have been used for fingerprinting DNA studies, to relate close species, in genetic mapping, in population genetics, in molecular evolutionary genetics, and in genetic breed studies in animals and plants.	Beuzen *et al*[53] Vignal *et al*[54] Wakchaure *et al*[50]
Microsatellites or SSR (simple sequence repeats)	They are sequences of 2 to 6 bp repeated in tandem throughout the genome and have a high polymorphism depending on the number of repetitions found in non-coding gene regions. They have been used in animal identification studies, genetic resource evaluation, paternity testing, disease research, determination of genetic variation within and between races, population genetics, gene and genome mapping migration, and the detection polymorphisms even *in silico* studies.	Khlestkina[16] Beuzen *et al*[53] Kumar *et al*[51] Duran *et al*[55]
SNP (single nucleotide polymorphism)	These are regions of DNA in which the substitution of one nucleotide by another, or the addition or removal of one or a few nucleotides, is observed. It has been used in the analysis of biparental inheritance genes and in the analysis of genetic differences, to make genetic maps and to detect genetic variations within species.	Khlestkina[16] Yu *et al*[56] Beuzen *et al*[53] Kumar *et al*[51]

With the emergence of mass sequencing technologies, known as "Next Generation Sequencing technologies (NGS)" millions of DNA molecules can be sequenced simultaneously, which greatly facilitates the study of microbial diversity[15]. One of the first high-throughput sequencing technologies was 454 pyrosequencing, which was used for targeted sequencing of ribosomal RNA gene amplicons[29]. This technique had the advantage of allowing the obtainment of sequences of up to 1,200 bp, albeit with a significantly higher error than other sequencing platforms (1%) and at a higher cost[15]. Second generation sequencing, also known as short reading sequencing (50 to 400 bp) uses mainly the Illumina platform[11]. Among its advantages, it is worth mentioning that it allows a greater number of readings, with an approximate error rate of 0.1% and at a comparatively low cost[15]. It is currently the most popular technology, but it requires a more complex bioinformatic analysis phase than other platforms.

Traditionally, when these two platforms (454 pyrosequencing and Illumina) are used for metagenomic analysis with the 16S rRNA marker, a previous amplification step by PCR is

performed, limiting the identified species to bacteria and archaea only, since the primers will always be used for amplifying fragments of the 16S rRNA gene. If the population also includes eukaryotic microorganisms such as yeasts and protozoa, they cannot be detected. On the other hand, this step of amplification by PCR entails an enrichment of the DNA which produces a bias towards the species that are found in greater proportion causing that the species that are found in smaller percentage to become hard to detect. Finally, this type of analysis identifies microorganisms down to the gender level[29].

An alternative for increasing resolution at the taxonomic level lies in the metagenomic study with the mass sequencing techniques called "Whole-Genome Shotgun sequencing" (WGS) and "Shotgun metagenomics sequencing (SMS)", in which the total metagenomic DNA is sequenced[30,31]. The major advantage of these methods is that microorganisms can be classified down to the species level and that not only prokaryotes but also eukaryotes can be identified; also, it does not require the previous amplification step by PCR, and therefore the bias is eliminated. Another advantage of these sequences is that by having sequences of all the DNA present in the sample, those corresponding to the 16S rRNA gene can be selected for use as taxonomic molecular markers; sequences of genes of other constituent polymorphic markers (MLSA) can also be sought in order to achieve a better classification of the microorganisms. The main disadvantages are that it has a higher cost than targeted sequencing of the 16S rRNA gene and requires more complex bioinformatic data analysis[32]. Several studies have been conducted to identify metagenomes in a wide range of population environments, using both 16S rRNA gene targeted sequencing and full metagenome sequencing with WGS and/or SMS.

Bioinformatic tools for metagenomic analysis

It is important to point out that bioinformatic tools must be used to analyze data obtained from massive sequencing. The greater the amount of data generated, the greater the need for bioinformatics resources[15], both for applications implementing analysis algorithms and for databases with information on microbial genomes (Table 2).

Table 2: Bioinformatic software for the analysis of metagenomic sequences

Bioinformatic application	Method of analysis	Reference
MG-RAST	Assigns structural and functional annotations according to nucleotide and protein databases by homology.	Meyer *et al*[33]
MOTHUR	Analyzes 16S rRNA gene sequences, quantifies ecological parameters to measure α and β diversity; visualizes the analysis using Venn diagrams, heat maps and dendrograms; selects sequence collections based on their quality, and calculates the sequence distance in pairs.	Schloss *et al*[34]
QUIIME	Analyzes microbial sequences of the 16S rRNA gene, performs taxonomic and phylogenetic profiles, and compares between samples.	Kuczynski *et al*[35]
PhaME	Performs SNP-based comparisons of entire genomes, assembled sequences, and processed sequences for phylogenetic and molecular evolutionary analysis.	Ahmed *et al*[36]
VITCOMIC1	Analyzes the 16S rRNA gene and high throughput sequences to visualize the phylogenetic composition of metagenomic samples.	Mori *et al*[37]
16SPIP	Rapidly detects pathogenic microorganisms in clinical samples based on metagenomic sequences of the 16S rRNA gene.	Miao *et al*[38]
PICRUSt	Algorithm with a predictive metagenomics approach based on 16S rRNA gene data and a reference genome database.	Langille *et al*[39]
CowPI	Uses PICRUSt to Analyze 16S rRNA Gene Data from Rumen Microbiome.	Wilkinson *et al*[57]
Kraken	Assigns taxonomic tags on metagenomic DNA sequences using k-mers alignment achieving more accurate classification compared to BLAST.	Wood *et al*[58]
Kaiju	Metagenome classifier that finds maximum matches at the protein level using the Burrows-Wheeler transformation; classifies readings with similar sensitivity and accuracy compared to k-mers based classifiers, especially in genera that are underrepresented in reference databases.	Menzel *et al*[59]

One of the most used applications since its launch is the MG-RAST[33] server, which assigns functional annotations to the analyzed sequences by comparing them with protein and nucleotide homology databases, in addition to allowing phylogenetic analysis. This tool is free and easily accessible, and it is fed with information provided by researchers; therefore, it helps to end the main bottleneck in metagenome sequence analysis, which lies in the availability of information to assign genomic annotations[33]. Two other widely used bioinformatic tools in metagenomics are MOTHUR[34], which is also freely accessible and which feeds on metagenomic information that users add to a database with monthly updates, and QUIIME[35], which is used for the analysis of microbial communities from bacterial and archaeal data.

Another software widely used for metagenome analysis is PhaME[36] (Phylogenetic and Molecular Evolutionary), which uses whole genome SNPs to measure interspecific diversity by phylogenetic analysis. PhaME[36] can be used to measure inter-species and inter-strain divergence and minimize errors in sequencing and assembly. Comparative genomics, including phylogenetic analysis based on ortho genes and SNPs, requires assembled or finished genomes. PhaME uses the SNP-based approach of complete genomes available in the databases, assembled sequences (contigs) and raw sequences to perform phylogenetic and molecular evolutionary analysis. This software combines algorithms for genome-wide alignment, reading mapping, and phylogenetic construction; it uses internal commands to infer the main genome and SNP, infer trees, and perform other molecular evolution analysis. PhaME is especially useful for the analysis and detection of organisms that are not very abundant in metagenome samples and has been used in data on bacterial samples, viruses, such as Ebola in Zaire, and yeasts, among others[36].

Other tools focus on the analysis of the hypervariable regions of the 16S rRNA gene, such as VITCOMIC1[37], which combines the information obtained from the targeted sequencing of the 16S rRNA gene as well as from the massive WGS or SMS sequencing to better visualize the phylogenetic composition of metagenomic samples, in addition to generating a more accurate record of the microbial community. Similarly, the 16SPIP[38] application has also been used for rapid detection of pathogenic microorganisms in clinical samples based on 16S rRNA metagenomic sequence data.

As for "predictive metagenomics" approaches, the PICRUSt[39] algorithm, which uses evolutionary models to predict metagenomes from 16S rRNA gene data and a reference genome database, should be highlighted. This tool has been used with data from soil microbiome samples, mammalian intestines, microbial mats, and humans[39], such as the human oral microbiota study which analyzed 6,431 samples of the 16S rRNA gene from the Human Microbiome Project[39,40].

Examples of metagenomic characterization with high throughput methodologies

Several metagenomic characterization works have been carried out to identify microorganisms living in environments of interest due to their great variability and ecological importance (Table 3). The following are a few examples of these works, without being exhaustive. For example, a massive sequencing of 29 metagenomes from samples from three marine stations that are part of the global Tara expedition was performed[29]. The taxonomic analysis carried out with the sequence data corresponding to the 16S rRNA gene made it possible to identify all the variable regions of the gene (V1 to V9). Targeted sequencing of the 16S rRNA gene was also performed for comparative purposes. The results obtained indicated that the efficiency in taxonomic classification with the use of ribosomal database RDP (Ribosomal Database Project) is similar for both types of sequencing. However, massive sequencing offers two major advantages: it reduces the error caused in amplicon PCR and it generates a large amount of functional data that can be analyzed along with the taxonomic analysis.

Table 3: Examples of metagenomic characterization

Sample	Type of analysis	Reference
Marine Plankton from Tara Oceans Expedition marine stations	Taxonomic profiles and structure of prokaryotic communities through massive 16S rRNA directed sequencing	Logares *et al*[29]
Sundarban Mangrove Sediments	Analysis of bacterial diversity and distribution through targeted sequencing of 16S rRNA	Basak *et al*[41]
Sediments from the Arabian Sea	Analysis of bacterial structure and diversity based on the sequencing of a 16S rRNA library	Nair *et al*[42]
Malaysia Sungai Klah Hot Springs	Diversity analysis through 16S rRNA V3-V4 region targeted sequencing	Chan *et al*[43]
Mushroom Spring in Yellowstone National Park	Microbial diversity based on 16S rRNA gene targeted sequencing and metagenomic sequencing.	Thiel *et al*[44]

Basal ice of Matanuska Glacier, Alaska	16S rRNA gene directed sequencing microbial diversity analysis and metagenomic sequencing	Kayani *et al*[45]
Blood from healthy donors	Analysis of the microbiome by PCR amplification and directed sequencing of 16S rRNA	Païsse *et al*[46]
Human Fecal Microbiome	Comparative study of the entire genome by massive and targeted sequencing of 16S rRNA	Ranjan *et al*[32]
Pasteurized and un-pasteurized Gouda cheese	Diversity analysis through targeted sequencing of the 16S rRNA gene	Salazar *et al*[47]
Ileal and cecal microbiota from broilers	Diversity analysis by amplification of the V3 region of the 16S rRNA gene	Mohd-Shaufi *et al*[48]
Microbiota attached to fiber in bovine rumen	Characterization of genes and genomes of metagenomic DNA	Hess *et al*[3]
Rumen of dairy and beef cattle	Taxonomic analysis of the rumen microbiome through directed pyrosequencing of the 16S rRNA gene	Wu *et al*[20]
Rumen microbiota in cattle supplemented with yeast	Analysis of rumen microbial diversity through pyrosequencing	Pinloche *et al*[5]
Rumen microbiota in cattle supplemented with thiamine	Analysis of bacterial diversity through targeted sequencing of the 16S rRNA gene	Pan *et al*[49]
Microbiome of healthy skin and with digital bovine dermatitis	Microbial characterization and functional gene composition of healthy skin or skin in active and inactive lesion stages by massive sequencing of the entire genome and annotation of the samples by MG-RAST	Zinicola *et al*[30]
Rumen fluid from three fractions of the bovine rumen	Metagenomic profiling of the rumen by non-directed parallel mass sequencing in metagenomic DNA	Ross *et al*[31]

Another metagenomic work in the field of mass sequencing focused on analyzing the diversity and bacterial distribution present in sediments of the tropical mangrove of Sundarban[41]. For this identification, it was used the 16S rRNA directed sequencing through 454 pyrosequencing, obtaining a total of 153,926 sequences. The analysis with MG-RAST software made possible the identification of 56,547 species belonging to 44 different

phylotypes, being the most dominant the phylotype *Proteobacteria*. On the other hand, metagenomic analysis of sediments from the Arabian Sea[42] with Sanger 16S rRNA sequencing classified the sequences obtained into seven different phylotypes where the phylotype *Proteobacteria* also predominated.

A large number of papers have focused on the characterization of metagenomes from extreme environments. For example, sequencing of 16S rRNA and complete genomes has been used to identify the diversity of thermophilic bacteria present in thermal waters in Malaysia whose temperature varies between 50 and 110 ºC[43]. An analysis of the 16S rRNA data identified approximately 35 phylotypes, of which *Firmicutes* and *Proteobacteria* represented 57 % of the microbiome. As for thermophiles, 70 % of those detected were strictly anaerobic; however, *Hydrogenobacter* spp. (forced chemolithotrophic thermophilotypes) constituted one of the most frequent taxa, and a large number of thermophilic photosynthetic microorganisms were found as well. Most of the identified phylotypes coincided with the findings of the sequencing of complete genomes. Thanks to this type of analysis, it was possible to identify and classify extreme microorganisms, such as thermophilotypes, anaerobes and chemolithophytes, that would have been difficult to characterize with classic microbiological methods[43].

Another study for identifying microbiota from extreme environments was conducted from samples of microorganisms that grow in the fungi that inhabit Yellowstone Park through the directed sequencing of 16S rRNA[44]. Over the years, the study of microorganisms in this habitat has focused on chlorphototrophic bacteria belonging to the *Cyanobacteria* and *Chloroflexi*. However, the results of the study revealed that microbial variation is dominated by a single taxon: *Roseiflexus* spp. which belongs to the group of anoxigenic phototrophic microorganisms[44]. Targeted 16S rRNA sequencing, along with full genome sequencing, has equally been used in glaciers, for which microbial information is also very limited. The first reported metagenomic study of glaciers[45] identified nine different genomes, including *Anaerolinea*, *Synthrophus* and *Thiobacillus*, and metabolic pathways involved in sulfur oxidation and nitrification were identified.

There are examples of the use of mass sequencing in metagenomic populations within the health and agro-food sectors. As an example within the field of human health, studies of directed sequencing of 16S rRNA to describe the microbiota present in the blood of healthy individuals have shown that this body fluid is not a sterile tissue[46]. At the phylotype level, more than 80% of the microorganisms present in the blood belonged to *Proteobacteria*, although phylotypes of *Actinobacteria*, *Firmicutes* and *Bacteriodetes* were also found. Ranjan *et al.*[32] used different strategies to characterize the human fecal microbiome. From a single sample they obtained 194.1 x106 readings from different sequencing strategies (16S rRNA directed sequencing, Illumina HiSeq, Illumina MiSeq). When comparing these,

especially the 16S rRNA gene directed sequencing with the WGS sequencing, they concluded that the latter has more advantages, as it increases the ability to identify bacterial species and the detection of diversity and gene prediction, and it also improves the accuracy of species detection by increasing the length of the sequences.

In the agro-food field, directed sequencing of 16S rRNA has also been used to identify microorganisms present in Gouda cheese[47] whether prepared with pasteurized or unpasteurized milk, and to evaluate changes due to the effect of aging. This study identified 120 genera in unpasteurized cheese and 92 in pasteurized cheese. In addition, depending on the aging time, it had a significant influence on the presence of microbiota. The most abundant genera in all samples were *Bacillaceae*, *Lactococcus*, *Lactobacillus*, *Streptococcus* and *Staphylococcus*.

In the case of growing broilers, the variation of ileal and cecal microbiota through time has been studied[48]. In order to do this, the hypervariable V3 region of the 16S rRNA gene was amplified and sequenced. The results showed that the cecal microbial communities were more diverse than the ileal ones. In addition, the presence of (potentially pathogenic) *Clostridium* bacteria was observed to increase as the animals grew and that the population of beneficial microorganisms such as *Lactobacillus* was low in all intervals[48].

In the case of ruminal metagenomes, it should be noted that one of the first sequencing studies was conducted to search for cellulolytic enzymes never before described[3]. In this study, 454 pyrosequencing was performed, obtaining 268 gigabases of metagenomic DNA information. From this information, 27,755 supposed genes of carbohydrate-active enzymes were identified, of which 90 codified for possible proteins, and 57% of them were enzymatically activated by cellulosic substrates. Another study focusing on ruminal metagenome in dairy calves and beef cattle steers[20] used 16S rRNA targeted pyrosequencing to assess population variation according to the type of livestock. This study found 8 phylotypes, 11 classes, 15 families and 17 different genera, and differences in the abundance of phylotypes found between dairy and beef cattle. The most abundant phylotypes were *Bacteriodetes*, *Firmicutes*, *Proteobacteria*, *Fibrobacteres* and *Spirochaetes* in both types of cattle, but with a lower abundance of *Bacteriodetes* and *Proteobacteria* in beef cattle. The use of yeast as a nutritional additive in cattle is known to improve milk production and weight gain. However, there is no knowledge of whether the effect caused by yeasts is a general stimulus to all microbial species or only affects some of the ruminal environment. Due to the above, a study was conducted to evaluate the changes in the rumen microbiota when the animals were fed with a yeast additive compared to when they consumed only the basal diet[5]. In this work, 454 pyrosequencing of the V1 region of the 16S rRNA gene was used to identify the population of ruminal microorganisms. The results showed that a change was observed in the main fibrolytic bacteria (*Fibrobacter* and *Ruminococcus*) and in lactate-using bacteria (*Megasphaera* and *Selenomonas*) when the yeast additive was added. Targeted sequencing

of the 16S rRNA gene in the adult dairy cattle ruminal microorganism population when combining thiamine with high grain diets has been used to evaluate its effect as an additive in animal nutrition[49]. The results confirmed that thiamine supplementation can improve ruminal function, as the number of cellulolytic bacteria increased when this amino acid was administered.

In the field of animal health, the sequencing of complete metagenomes has also been used. For example, skin metagenome with active and receding bovine digital dermatitis has been compared with the skin of healthy cattle to see if pathogens involved in the pathogenesis of the disease were detected[30]. The sequences obtained were analyzed with MG-RAST and six main phylotypes were identified, among which *Firmicutes* and *Actinobacteria* predominated in the microbiome of healthy patients, while *Spirochetes*, *Bacteroidetes* and *Proteobacteria* were the most abundant in active and recession patients; this confirms that the presence of the disease changes the population of the metagenome.

Rumen metagenomic profiles have been obtained by sequencing complete metagenomes from samples of ruminal fluid from three different cattle and between different locations in the rumen[31]. In addition to comparing with the metagenome from feces of the same animals, the results indicated that the variation in metagenomic profiles was less among samples taken from the same animal, even if they were taken from different regions of the rumen. Contrary to expectations, no relationship was found with the metagenomic profile of faeces and ruminal fluid from the same animal.

Conclusions

Traditionally, metagenomic analysis used laborious methodologies, such as denaturing gradient gel electrophoresis, the digestion of genomes with restriction enzymes, and their visualization by means of agarose and/or acrylamide gels. The development of nucleic acid sequencing methodologies, especially new mass sequencing technologies, has helped to reduce this problem.

The 16S rRNA gene has traditionally been considered the gold standard for classifying prokaryotic microorganisms (bacteria and archaea), as it meets all the characteristics required to be a molecular marker. However, despite the large number of works that have used the sequencing of the hypervariable regions of this marker, it has the disadvantage of not being able to determine taxa at an infra-generic level. A strategy used to improve taxonomic classification has been the combination of the 16S rRNA marker with some other constitutive expression genes such as the genes sodA, hps65, gyrB, among others, and even genes

encoding for subunits of the cytochrome c enzyme complex have been used to classify microorganisms into species.

In the last decade, mass sequencing technologies have made it possible for microbial populations to be analyzed in greater depth, either by sequencing the entire 16S rRNA gene, thus increasing the resolution of that marker, or by combining the information of that gene with the sequencing of complete metagenomes. In this last type of analysis, sequences of all the genomic material present in the sample are obtained, which offers the great advantage that in addition to making the taxonomic classification it is also possible to obtain functional information of the detected genes. Thus, despite the limitations of the required bioinformatic analysis, the use of these methodologies allows for more complete analyses.

However, despite the development of high-performance sequencing techniques, the targeted sequencing of 16S rRNA on the Sanger platform is not entirely obsolete, and the selection of the analysis strategy will depend on the objectives of the study, the degree of precision desired, the sample size and the financial resources that can be allocated by the research team. For example, if you are looking for the presence and/or absence of a single bacterial genus, Sanger sequencing would be ideal because it has the ability to sequence relatively large fragments with greater precision than any mass sequencing platform. If what is wanted is to discriminate between species of a single bacterial genus, two strategies can be utilized: the sequencing of some hypervariable region of the 16S rRNA together with some other constitutive gene (MLSA), or the sequencing of the whole gene in order to obtain the information of all the hypervariable regions.

Today, metagenomics faces numerous challenges arising from the large amount of information generated, its storage and the way in which it must be treated. Although many tools and applications have been designed for bioinformatic analysis of metagenomes, there is no single "protocol" of analysis; therefore, each study must be adapted to the nature of the samples and the objectives of the experiment.

In conclusion, microbial diversity studies will always use the 16S rRNA molecular marker to make taxonomic classifications, either through the sequencing of one or two of its hypervariable regions or through that of the whole gene, and it can even be combined with the use of another constitutive gene as a molecular marker to achieve a better taxonomic classification. On the other hand, mass sequencing technologies have greatly improved the study capacity and speed of metagenome analysis. This has occurred particularly in environmental samples with ecological importance, in both human and animal health, in studies on symbiosis of plants with endophytic fungi, and in the evaluation of ruminal metagenomes, to mention a few.

Literature cited:

1. Li W, Huan X, Zhou Y, Ma Q, Chen Y. Simultaneous cloning and expression of two cellulase genes from Bacillus subtilis newly isolated from Golden Takin (*Budorcas taxicolor* Bedfordi). Biochem Biophys Res Commun 2009;383(4):397-400.

2. Sadet S, Martin C, Meunier B, Morgavi DP. PCR-DGGE analysis reveals a distinct diversity in the bacterial population attached to the rumen epithelium. Animal 2007;1(7):939-944.

3. Hess M, Sczyrba A, Egan R, *et al*. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 2011;331(6016):463-467.

4. Li RW, Connor EE, Li C, Ransom L, Baldwin VI, Sparks ME. Characterization of the rumen microbiota of pre-ruminant calves using metagenomic tools. Environ Microbiol 2012;14(1):129-139.

5. Pinloche E, McEwan N, Marden JP, Bayourthe C, Auclair E, Newbold CJ. The Effects of a probiotic yeast on the bacterial diversity and population structure in the rumen of cattle. PLoS One 2013;8(7):e67824.

6. Kumar M, Shrivastava N, Teotia P, *et al*. Omics: Tools for assessing environmental microbial diversity and composition. In: Varma A. SA, ed. Modern tools and techniques to understand microbes. Springer, Cham; 2017:273-283.

7. Marshall IPG, Karst SM, Nielsen PH, Jørgensen BB. Metagenomes from deep Baltic Sea sediments reveal how past and present environmental conditions determine microbial community composition. Mar Genomics 2018;37:58-68.

8. Simon C, Daniel R. Metagenomic analyses: Past and future trends. Appl Environ Microbiol 2011;77(4):1153-1161.

9. Cowan D, Meyer Q, Stafford W, Muyanga S, Cameron R, Wittwer P. Metagenomic gene discovery: Past, present and future. Trends Biotechnol 2005;23(6):321-329.

10. Streit WR, Schmitz RA. Metagenomics - The key to the uncultured microbes. Curr Opin Microbiol 2004;7(5):492-498.

11. Pacheco-Arjona JR, Sandoval-Castro CA. Tecnologías de secuenciación del metagenoma del rumen. Trop Subtrop Agroecosyt 2018;21:587-598.

12. Yu Z, Morrison M. Improved extraction of PCR-quality community DNA from digesta and fecal samples. BioTechniques 2004;36:808-812.

13. Singh B, Bhat TK, Kurade NP, Sharma OP. Metagenomics in animal gastrointestinal ecosystem: a microbiological and biotechnological perspective. Indian J Microbiol 2008;48:216-227.

14. Venter JC, Remington K, Heidelberg JF, *et al*. Environmental genome shotgun sequencing of the Sargasso sea. Science 2004;304(5667):66-74.

15. Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A. The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. Front Genet 2015;6:348.

16. Khlestkina EK. Molecular markers in genetic studies and breeding. Russ J Genet Appl Res 2014;4(3):236-244.

17. Patwardhan A, Samit R, Roy A. Molecular markers in phylogenetic studies-A Review. J Phylogenetics Evol Biol 2014;02(02).

18. D´Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, *et al*. A comprenhensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics. 2016;17:55.

19. Clarridge JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev 2004;17(4):840-862.

20. Wu S, Baldwin RL, Li W, Li C, Connor EE, Li RW. The bacterial community composition of the bovine rumen detected using pyrosequencing of 16S rRNA genes. Metagenomics 2012;1:235571.

21. Valenzuela-Gonzalez F, Martínez-Porchas M, Villalpando-Canchola E, Vargas-Albores F. Studying long 16S rDNA sequences with ultrafast-metagenomic sequence classification using exact alignments (Kraken). J Microbiol Methods 2016;122:38-42.

22. Ludwig W, Schleifer KH. Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol Rev 1994;15(2-3):155-173.

23. Osorio CR, Collins MD, Romalde JL, Toranzo AE. Variation in 16S-23S rRNA intergenic spacer regions in Photobacterium damsalae: a Mosaic-Like structure. Appl Environ Microbiol 2005;71(2):636-645.

24. Glaeser SP, Kämpfer P. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Syst Appl Microbiol 2015;38(4):237-245.

25. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol 2007;73(1):278-288.

26. Sánchez-Herrera K, Sandoval H, Mouniee D, *et al*. Molecular identification of Nocardia species using the sodA gene: Identificación molecular de especies de Nocardia utilizando el gen sodA. New Microbes New Infect 2017;19:96-116.

27. Dumont MG. Primers: Functional marker genes for Methylotrophs and Methanotrophs. In: McGenity T, Timmis KNB, editors. Hydrocarbon and lipid microbiology protocols - Springer Protocols Handbooks. Berlin: Springer Protocols Handbooks; 2014.

28. Kolb S, Stacheter A. Prerequisites for amplicon pyrosequencing of microbial methanol utilizers in the environment. Front Microbiol 2013;4(SEP):1-12.

29. Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, *et al*. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ Microbiol. 2014;16(9):2659-2671.

30. Zinicola M, Higgins H, Lima S, Machado V, Guard C, Bicalho R. Shotgun metagenomics sequencing reveals functional genes and microbiome associated with bovine digital dermatitis. PLoS ONE 2015;10(7)e0133674.

31. Ross EM, Moate PJ, Bath CR, Davidson SE, Sawbridge TI, Guthridge KM, Cocks BG, Hayes BJ. High throughput whole rumen metagenome profiling using untargeted massively parallel sequencing. BMC Genetics 2012;13:53.

32. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of microbiome: Advantages of whole genome shotgun *versus* 16S amplicon sequencing. Biochem Biophys Res Commun 2016;469:967-977.

33. Meyer F, Paarmann D, D`Souza M, Olson R, Glass EM, Kubal M, *et al*. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 2008;9:386.

34. Schloss PD, Westcott S. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75(23):7537-7541.

35. Kuczynski J, Stombauhg, Walters WA, Gonzalez A, Caporaso JG, Knight R. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Bioinformatics 2011;10:7.

36. Ahmed SA, Lo CC, Li PE, Davenport KW, Chain PSG. From raw reads to trees: Whole genome SNP phylogenetics across the tree of life. bioRxiv 2015:032250.

37. Mori H, Maruyama T, Yano M, Yamada T, Kurokawa K. VITCOMIC2: visualization tool for the phylogenetic composition of microbial communites based on 16S rRNA gene amplicons and metagenomic shotgun sequencing. BMC Systems Biol 2018;12(2):30.

38. Miao J, Han N, Qiang Y, Zhang T, Li X, Zhang W. 16SPIP: a comprehensive analysis pipeline for rapid pathogen detection in clinical samples based on 16S metagenomic sequencing. BMC Bioinformatics 2017;18(16):568.

39. Langille MGI, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 2013;31(9):814-82.

40. Human Microbiome Project Consortium. A framework for human microbiome research. Nature 2012;13;486(7402):215-21. doi: 10.1038/nature11209.

41. Basak P, Pramanik A, Sengupta S, Nag S, Bhattacharyya A, Roy D, Pattanayak R, Ghosh A, Chattopadhyay D, Bhattachryya M. Bacterial diversity assessment of pristine mangrove microbial community from Dhulibhashani, Sundarbans using 16S rRNA gene tag sequencing. Genomics Data 2016;7:76-78.

42. Nair HP, Puthusseri RM, Vincent H, Bhat SG. 16S rDNA-based bacterial diversity analysis of Arabian Sea sediments: A metagenomic approach. Ecol Genet Genomics 2017;3(5):47-51.

43. Chan CS, Chan KG, Tay YL, Chua TH, Goh KM. Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. Front Microbiol 2015;6:177.

44. Thiel V, Wood JM, Olsen MT, Tank M, Katt CG, Ward DM, Bryant DA. The dark side of the mushroom spring microbial mat: Life in the shadow of chlorophototrophs. I. Microbial diversity based on 16S rRNA gene amplicons and metagenomic sequencing. Front Microbiol 2016;7:919.

45. Kayani MR, Doyle SM, Sangwan N, Wang G, Gilbert JA, Christner BC, Zhu TF. Metagenomic analysis of basal ice from an Alaskan glacier. Microbiome 2018;6:123.

46. Païsse S, Valle C, Servant F, Courtney M, Burcelin R, Amar J, Lelouvier B. Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. Transfusion 2016;56:1138-1147.

47. Salazar JK, Carstens CK, Ramachandran P, Shazer AG, Narula SS, Reed E, Ottesen A, Schill KM. Metagenomics of pasteurized and unpasteurized gouda cheese using targeted 16S rDNA sequencing. BMC Microbiology 2018;18:189.

48. Mohd-Shaufi MA, Sieo CC, Chong CW, Gan HM, Ho YW. Deciphering chicken gut microbial dynamics based on high-throughput 16S rRNA metagenomics analyses. Gut Pathog 2015;7:4

49. Pan X, Xue F, Nan X, Tang Z, Wang K, Beckers Y, Jiang L, Xiong B. Illumina sequencing approach to characterize thiamine metabolism related bacteria and the impacts of thiamine supplementation on ruminal microbiota in dairy cows fed high-grain diets. Front Microbiol 2017;8:1818.

50. Wakchaure R, Ganguly S, Para PA, Praveen PK, Qadri K. Molecular markers and their applications in farm animals : A Review. Int J Recent Biotechnol 2015;3(January 2016):23-29.

51. Kumar A, Tomar SS, Kumar A, Singh J. Importance of molecular markers in livestock improvement: a review. Int J Agric Res Innov Technol 2017;5(4):614-621.

52. Lang T, Li G, Yu Z, Ma J, Chen Q, Yang E, Yang Z. Genome-wide distribution of novel Ta-3A1 mini-satellite repeats and its use for chromosome identification in wheat and related species. Agronomy 2019;9(2):60.

53. Beuzen ND, Stear MJ, Chang KC. Molecular markers and their use in animal breeding. Vet J 2000;160(1):42-52.

54. Vignal A, Milan D, San Cristobal M, Eggen A. A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 2002;34(2002):275-305.

55. Duran C, Singhania R, Raman H, Batley J, Edwards D. Predicting polymorphic EST-SSRs in silico. Mol Ecol Resour 2013;13(3):538-545.

56. Yu H, Xie W, Wang J, *et al*. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. PLoS One 2011;6(3).

57. Wilkinson TJ, Huws SA, Edwards JE, Kingston-Smith A, Siu Ting K, *et al*. CowPI: a rumen microbiome focused version of the PICRUSt functional inference software. Frontiers in Microbiol 2018;(9):1095.

58. Wood DE, Salzberg SL. Kraken: ultrafast metagenomics sequence classification using exact alignments. Genome Biology. 2014;15(3):R46.

59. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature communications 2016;7:11257.