FACULTAD DE CIENCIAS AGRARIAS
UNIVERSIDAD NACIONAL DE ROSARIO

ENSAMBLADO Y ANÁLISIS COMPARATIVO DE METAGENOMAS DE RUMEN VACUNO

LIC. LAURA LIS RICARDI

TRABAJO FINAL PARA OPTAR AL TÍTULO DE ESPECIALISTA EN BIOINFORMÁTICA

DIRECTOR: VÍCTOR BLANCATO

AÑO 2022

Ensamblado y análisis comparativo de metagenomas de rumen vacuno

Laura Lis Ricardi

Licenciada en Biotecnología - Facultad de cs. Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario (UNR)

Este Trabajo Final es presentado como parte de los requisitos para optar al grado académico de Especialista en Bioinformática, de la Universidad Nacional de Rosario y no ha sido previamente presentada para la obtención de otro título en ésta u otra Universidad. El mismo contiene los resultados obtenidos en investigaciones llevadas a cabo en el Instituto de Biología Molecular y Celular de Rosario (IBR-CONICET), durante el período comprendido entre 2020-2022, bajo la dirección de Víctor Blancato.

Nombre y firma del autor:

Laura Lis Ricardi 

Nombre y firma del Director:

Víctor Blancato 

Defendida: .....de 20__.

Presentaciones a congresos

Parte de los resultados obtenidos en este trabajo final han sido presentados en el siguiente evento nacional:

“Metagenome-assembled genomes from cow rumen”. Laura Ricardi, Christian Magni, Víctor Blancato. LVII Reunión Anual de la Sociedad Argentina de Investigación Bioquímica y Biología Molecular (SAIB) y XVI Congreso Anual de la Asociación Civil de Microbiología General (SAMIGE). Realizado de manera virtual del 1 al 5/11/2021.

Abreviaturas y símbolos

AA	Actividad Auxiliar
AAI	Identidad promedio de aminoácido
ANI	Identidad promedio de nucleótido
BBH	Método del mejor hit bidireccional
CAZymas	Enzimas activas sobre carbohidratos
CBM	Módulos de Unión a Carbohidratos
CE	Carbohidrato Esterasas
CGCs	*Clusters* de genes
GH	Glucósido Hidrolasas
GT	Glicosil Transferasas
HMM	*Hidden Markov Model*
KEGG	Enciclopedia Kioto de Genes y Genomas
KO	Identificador de ortología KEGG
MAGs	Genomas ensamblados de metagenomas
PL	Polisacárido Liasas
PPR	Reconocimiento de patrones peptídicos
SAGs	Genomas únicos amplificados
SBH	Método del mejor hit unidireccional
TCs	Proteínas transportadoras
TFs	Factores de transcripción

Resumen

Los rumiantes pueden transformar la energía almacenada en las plantas en productos alimenticios que pueden ser utilizados por los humanos, como la carne y la leche.

La microbiota del rumen está compuesta por protozoos, bacterias, hongos y arqueas, que son responsables de la degradación del material vegetal.

A pesar del fuerte interés industrial y científico, el rumen sigue siendo un hábitat poco conocido, con muchas especies y cepas microbianas no cultivadas. La secuenciación metagenómica del rumen produce secuencias muy novedosas, que pueden ser de gran interés para las industrias de biocombustibles, alimentos y biotecnología. En este trabajo, los metagenomas de muestras de rumen fueron obtenidos de vacas regionales jóvenes y adultas alimentadas con una dieta rica o pobre con el objetivo de ensamblar nuevos genomas. El ADN fue extraído y secuenciado por WGS. Luego, las lecturas se filtraron por calidad y se ensamblaron con Megahit. La calidad de los *contigs* se evaluó con el *software* QUAST, y BWA MEM se usó para asignar lecturas a los ensamblajes. El *binning* se realizó con Metabat2 usando los *contigs* obtenidos, y archivos BAM correspondientes a alineaciones de lecturas. Se recuperaron de 12 a 31 *bins* por muestra, y se evalúo su integridad y contaminación mediante CheckM. El filtrado de estos por completitud ≥80% y contaminación ≤10%, generó entre tres y cinco genomas ensamblados de metagenomas (MAGs) por muestra. La asignación taxonómica se llevó a cabo utilizando el servidor MiGA, lo cual permitió identificar organismos asociados al tracto gastrointestinal y la degradación de material vegetal. La predicción de genes se llevó a cabo mediante Prodigal. Con esta información, se determinó el perfil metabólico utilizando el programa Genomaple, el cual permitió obtener un análisis global de las principales vías presentes en los metagenomas.

Debido a la importancia de las enzimas activas sobre carbohidratos (CAZymas) en el rumen, se realizó un estudio de las CAZymas presentes en las muestras, identificándose posteriormente aquellas enzimas no ortólogas a la base de datos de proteínas ruminales RumiRef, lo cual sugeriría que estas proteínas son únicas en las muestras analizadas.

Palabras clave: Metagenómica, rumen, ensamblado

Abstract

Assembly and comparative analysis of cow rumen metagenomes

Ruminants can transform the energy stored in plants into food products that can be used by humans, such as meat and milk. The rumen microbiota is composed of protozoa, bacteria, fungi and archaea, which are responsible for plant material degradation. Despite strong industrial and scientific interest, the rumen remains a poorly understood habitat, with many uncultivated microbial species and strains. Metagenomic sequencing of the rumen still produces highly novel sequences, which can be of great interest for biofuels, food and biotechnology industries. In this work, the metagenomes of rumen samples of regional young and adult cows fed with a rich or poor diet were obtained aiming to assemble novel genomes. DNA was extracted and sequenced by WGS. The reads were then filtered for quality and assembled with Megahit. Quality of the contigs was assessed with QUAST software, and BWA MEM was used to map reads back to the assemblies. Binning was carried out with Metabat2 using the obtained contigs, and BAM files corresponding to reads alignments. 12 to 31 *bins* were recovered per sample, and their integrity and contamination were evaluated by CheckM. Filtering these for completeness ≥80% and contamination ≤10% generated between three and five metagenome assembled genomes (MAGs) per sample. The taxonomic assignment was carried out using MiGA server, which allowed the identification of organisms associated with gastrointestinal tract and plant material degradation. Gene prediction was carried out using Prodigal. With this information, the metabolic profile was determined using Genomaple program, which allowed obtaining a global analysis of the main pathways present in the metagenomes.

Due to the importance of carbohydrate-active enzymes (CAZymes) in the rumen, a study of the CAZymes present in the samples was carried out, subsequently identifying those enzymes not orthologous to the RumiRef ruminal protein database, which would suggest that these proteins are unique in the analyzed samples.

Índice

1. Introducción

1.1.Microbiota ruminal

Los rumiantes son un grupo de animales mamíferos herbívoros que comprenden alrededor de 200 especies en todo el mundo, representando unos 75 millones de individuos salvajes y 3.500 millones de individuos domesticados (Hackmann and Spain, 2010). Estos animales se caracterizan por su modo de digestión vegetal, ya que poseen un anteestómago, el rumen, en el cual se digieren parcialmente los alimentos antes de llegar al verdadero estómago.

El rumen es un ecosistema complejo compuesto por cientos de filotipos de bacterias, protozoos, hongos, metanógenos y bacteriófagos, con una concentración que varía entre los 10^6 y los 10^{11} células o partículas/ml (Morgavi *et al.*, 2013). Estos microorganismos interactúan entre sí y con su entorno, ejerciendo funciones no solo nutricionales, sino también protectivas e inmunológicas que benefician al animal huésped (Hooper, 2004). Los simbiontes microbianos del rumen son esenciales en el proceso de transformación de plantas forrajeras y otros alimentos no aptos para el consumo humano en productos de alta calidad; al tiempo que son responsables de la producción de metano y otros compuestos contaminantes para el medio ambiente.

Debido al crecimiento de la población a nivel mundial, la demanda de carne y leche de estos animales aumenta año tras año. Esto plantea un desafío para la industria agrícola ganadera debido a la disminución de los recursos naturales y el aumento de los costos de producción, como así también en el incremento de la toma de conciencia sobre el impacto ambiental de la cría intensiva de ganado. En este sentido, la caracterización de la microbiota ruminal juega un papel clave en la obtención de información acerca de las funciones que cumplen estos organismos en el rumen y la interacción entre sus miembros y con el huésped. Una mayor comprensión del microbioma ruminal a través de la construcción de catálogos de genes permite elaborar estrategias en torno a mejorar la digestibilidad del alimento y reducir la producción de metano; haciendo de la ganadería una práctica más sostenible. Una herramienta muy útil para lograr esto lo constituye la metagenómica.

1.2. Métodos de secuenciación

Actualmente, y a pesar de su importancia, la comprensión del funcionamiento y la ecología del microbioma ruminal no se ha dilucidado por completo.

Antes de la secuenciación del ADN, las técnicas basadas en el aislamiento de cultivos eran el único método de identificación de especies microbianas. Sin embargo, y similar a lo que ocurre en otros ecosistemas, el número de especies aisladas y caracterizadas en el rumen mediante estas técnicas es bajo, representando menos del 15% (Morgavi *et al.*, 2013).

La secuenciación del material genómico de un medioambiente, lo que se conoce como metagenómica, permitió tener una visión más completa de todas las comunidades microbianas presentes en él y no sólo de aquellos organismos que podían ser cultivables.

Existen dos métodos principales de secuenciación para el estudio de microbiomas: la secuenciación de amplicón de un gen objetivo, y la secuenciación "*shotgun*" del metagenoma entero (Scholz *et al.*, 2012). La primera estrategia secuencia genes marcadores filogenéticos o partes de ellos, generalmente el ARNr 16S (Langille *et al.*, 2013) y regiones del espaciador transcrito interno (ITS) (Unterseher *et al.*, 2011), utilizando *primers* específicos. Al tener grandes bases de datos de genes marcadores, el proceso de agrupamiento de secuencias se torna más simple y confiable (Ribeca and Valiente, 2011). Sin embargo, no se obtiene información adicional además de la riqueza y abundancia de especies. En este sentido, la secuenciación *shotgun* brinda una visión más profunda del metagenoma, ya que tiene la ventaja de que cada secuencia representa una parte aleatoria, incluidos genes desconocidos, de un genoma que se encuentra en el metagenoma (Sharpton, 2014). A través de esta técnica es posible determinar la composición taxonómica y el potencial funcional de las comunidades microbianas, y recuperar secuencias genómicas completas. Debido a la gran cantidad de datos que se generan y a la falta de disponibilidad de secuencias de referencia completas, el agrupamiento de lecturas provenientes de un mismo genoma se vuelve un mayor desafío.

El estudio de metagenómica por secuenciación *shotgun* se puede dividir en los siguientes pasos: 1) obtención de muestras, extracción de ADN, construcción de librerías y secuenciación, 2) preprocesado de las lecturas, 3) análisis bioinformático de las secuencias y 4) análisis estadístico e interpretación biológica de los datos.

Respecto al preprocesado de las lecturas ("*reads*") obtenidas en la secuenciación, lo primero que se realiza es un análisis de calidad, eliminándose las secuencias de baja calidad, así como

posibles contaminantes. Las lecturas que cumplen con los criterios de calidad pueden ser luego ensambladas en secuencias más largas denominadas *contigs*. En un paso posterior, los *contigs* se agrupan con el objetivo de reconstruir los genomas presentes la muestra, un método conocido como *binning*.

1.3. Ensamblado

El proceso de ensamblado consiste en la unión de lecturas cortas para formar secuencias más largas llamadas *contigs*, las cuales son representativas de los genomas originales que componen la muestra. El ensamblado de metagenomas puede realizarse mediante dos métodos: basado en una referencia (co- ensamblado) o ensamblado *de novo*. En el primer caso, se utiliza un genoma ensamblado previamente como referencia, alineando las secuencias en su posición más probable. Esto es útil cuando la muestra contiene secuencias que están estrechamente relacionadas con los genomas de referencia. Sin embargo, diferencias de la muestra con respecto a la referencia, como inserciones, deleciones o polimorfismos, pueden llevar a un ensamblaje fragmentado o a que las regiones divergentes no estén cubiertas (Thomas *et al.*, 2014).

El ensamblaje *de novo*, por otra parte, no utiliza genomas de referencia y generalmente requiere de mayores recursos computacionales. Dentro de las herramientas disponibles, aquellas basadas en gráficos de Bruijn son muy populares cuando se trata de metagenomas. En estos, cada lectura es dividida en subsecuencias superpuestas de una longitud fija k, denominadas “k-meros”, que definen los vértices y las aristas del gráfico de Bruijn. La tarea del ensamblador es encontrar luego un camino a través del gráfico que reconstruya los genomas de la muestra (Quince *et al.*, 2017).

1.3.1. Megahit

MEGAHIT v0.1 (Li *et al.*, 2016) ensambla conjuntos de datos metagenónicos grandes y complejos de manera eficiente en tiempo y memoria, mediante el uso de un gráfico sucinto de Bruijn. Este es una estructura de datos comprimida que representa un gráfico de Bruijn, el cual se basa en la lista ordenada de los elementos que definen sus aristas. El programa trabaja con diferentes tamaños de k-meros, discriminando aquellos que contienen errores de

secuenciación y una baja expresión, obteniéndose de esta manera una mayor calidad en los ensamblados.

1.3.2. Optimización del ensamblado: Quast

Si bien las tecnologías de secuenciación han avanzado, aún enfrentan muchas complicaciones que dificultan la reconstrucción de cromosomas completos, incluidos errores en las lecturas y grandes repeticiones en el genoma. Esto lleva a que los ensambladores generen *contigs* con grandes diferencias entre sí, lo que hace necesaria la evaluación y optimización del ensamblado a través de un análisis de calidad.

QUAST (Gurevich *et al.*, 2013) es una herramienta de evaluación de la calidad del ensamblaje que utiliza el alineador Nucmer de MUMmer v3.23 (Kurtz *et al.*, 2004) para alinear ensamblajes con un genoma de referencia y evaluar métricas según las alineaciones. Además permite calcular métricas sin requerir de genomas de referencia, útiles para el análisis de ensamblajes de especies no secuenciadas previamente.

Entre las métricas más importantes que evalúa QUAST se destaca el N50, el cual se obtiene al ordenar los *contigs* de acuerdo a su tamaño, del más grande al más pequeño, y sumar las bases que contienen. El tamaño del *contig* más corto que el 50% de la longitud total del genoma corresponde al N50, y se puede describir como una estadística de mediana ponderada, de modo que el 50% de todo el conjunto está contenido en *contigs* iguales o mayores que este valor. Otro parámetro de importancia es el L50, el cual se define como el número más pequeño de *contigs* cuya suma de longitud constituye la mitad del tamaño del genoma.

1.4. Binning

El *binning* permite reconstruir las especies presentes en la muestra (*bins*) mediante el agrupamiento de los *contigs*. Esto puede realizarse mediante métodos supervisados, los cuales utilizan bases de datos de genomas ya secuenciados para etiquetar *contigs* en clases taxonómicas; o mediante métodos no supervisados que buscan grupos naturales en los datos.

Ambos métodos poseen dos elementos principales: una métrica para definir la similitud entre un *contig* dado y un *bin*, y un algoritmo para convertir esas similitudes en asignaciones (Quince *et al.*, 2017).

Debido a que la mayoría de las especies microbianas aún no han sido secuenciadas, la estrategia no supervisada se vuelve más conveniente. Las métricas que emplean estos programas para agrupar *contigs* se basan en la composición de las secuencias, a través del análisis de las frecuencias de k-meros o tetrámeros; y/o su abundancia. Estos asumen que la abundancia de genes de una misma especie covaría para un mismo taxón, y que los genomas de diferentes especies microbianas contienen combinaciones particulares de bases que dan como resultado diferentes frecuencias de k-meros (Karlin and Campbell, 1997). Los tetrámeros, por su parte, se consideran los más informativos para agrupar datos metagenómicos (Dick *et al.*, 2009).

1.4.1. Metabat2

MetaBAT2 (Kang *et al.*, 2019) se utiliza para la construcción de los *bins*. Requiere como entradas un archivo de ensamblaje y archivos BAM correspondientes a los alineamientos de las lecturas de cada muestra con el metagenoma ensamblado. Para cada par de *contigs* en un ensamblaje, el programa calcula sus distancias probabilísticas basadas en la frecuencia de tetranucleótidos (TNF) y abundancia; las cuales se integran en una distancia compuesta. Todas las distancias por pares forman una matriz, que luego se agrupan a través de un método basado en gráficos.

Debido a que la TNF varía de acuerdo al genoma, ésta se usa como parámetro característico de la composición de secuencias. La diferenciación de las TNF de los distintos genomas se lleva a cabo calculando la probabilidad de distancia euclidiana entre especies e intraespecies utilizando 1.414 referencias genómicas completas y únicas de NCBI.

1.4.2. Evaluación de la calidad de los bins: CheckM

La calidad del *binning* ha sido evaluada tradicionalmente mediante la presencia y ausencia de genes marcadores de copia única, lo que permite estimar la integridad del genoma (*completitud*) (Rinke *et al.*, 2013; Wrighton *et al.*, 2012; Haroon *et al.*, 2013; Sharon *et al.*, 2014). Debido a que estos marcadores no están distribuidos de manera universal dentro del genoma y a su bajo número (menos del 10% del total) (Sharon and Banfield, 2013), se han

identificado aquellos genes que son ubicuos y simple copia dentro de un *phylum* específico. De esta forma se incrementa el número de genes marcadores usados en la estimación (Swan *et al.*, 2013), permitiendo además evaluar contaminación, ya que sólo debería haber una copia de cada gen marcador en cada genoma (Sekiguchi *et al.*, 2015; Albertsen *et al.*, 2013; Soo *et al*., 2014).

CheckM (Parks *et al.*, 2015) permite estimar la integridad y contaminación mediante la utilización de genes marcadores específicos del linaje inferido de un genoma dentro de un árbol de genoma de referencia. Para refinar las estimaciones de calidad, los genes marcadores se agrupan en conjuntos.

El *software* estima la integridad del genoma a partir del número de genes marcadores de copia única presentes en cada *bin*, mientras que la contaminación se obtiene calculando cuántos de estos genes marcadores están presentes en múltiples copias.

1.5.Asignación taxonómica: MiGA

Para identificar los organismos a los cuales pertenecen los *bins*, se recurre a distintos métodos de asignación taxonómica.

El gen del ARN ribosomal 16S (ARNr 16S) ha sido utilizado con éxito para catalogar y estudiar la diversidad de especies procariotas y sus comunidades. Sin embargo, posee una resolución limitada a nivel de especies y niveles más finos, y no puede representar la diversidad y fluidez del genoma completo.

Para resolver estos inconvenientes, se recurre a métodos basados en el genoma tales como la identidad promedio de nucleótido (ANI) (Konstantinidis and Tiedje, 2005; Goris *et al.*, 2007). Esta medida representa la identidad promedio de nucleótido de todos los genes ortólogos compartidos entre dos genomas; y ofrece una resolución robusta para cepas de la misma especie o altamente relacionadas.

The Microbial Genomes Atlas (MiGA) (Rodriguez-R *et al.*, 2018) es un servidor *web* que permite la clasificación taxonómica de secuencias genómicas desconocidas basado en los conceptos de ANI o de identidad promedio de aminoácido (AAI) (Konstantinidis and Tiedje, 2005) (en el caso de secuencias más divergentes) contra una base de datos de genomas de referencia.

MiGA identifica el genoma de referencia que mejor coincide con la secuencia de consulta en función de los valores ANI/AAI y, posteriormente, evalúa si la secuencia debe asignarse al mismo rango taxonómico (por ejemplo, especie, género, etc.) o si representa un taxón nuevo en ese rango. La secuencia de consulta puede compararse además contra genomas no clasificados, colecciones de genomas ensamblados de metagenomas (MAGs) y genomas únicos amplificados (SAGs) para identificar relaciones entre sí.

1.6. Predicción de genes: Prodigal

La identificación de genes que codifican proteínas y otros elementos regulatorios es uno de los pasos más importantes para su anotación estructural y funcional.

Los programas predictores de genes pueden ser categorizados en tres tipos: métodos *ab initio*, métodos basados en homología, y aquellos que combinan las características de los dos anteriores. Los métodos *ab initio* se centran en modelos estadísticos para identificar promotores, secuencias codificantes y no codificantes y uniones intrón-exón en la secuencia. Los métodos basados en homología, en cambio, alinean la secuencia con marcadores de secuencia expresada (EST), ADNc o evidencias de proteína, y utiliza las similitudes detectadas para la predicción de genes (Ejigu and Jung, 2020).

Prodigal (Hyatt *et al.*, 2010) es un algoritmo de *machine learning* no supervisado, el cual no necesita de la incorporación de genes conocidos externos como datos de entrenamiento. El programa genera un set de genes de entrenamiento propio recorriendo la secuencia y examinando el contenido de G y C en cada una de las tres posiciones del codón en cada ORF. La posición de codón de mayor contenido de GC para un ORF se considera la "ganadora" y se incrementa una suma acumulada para esa posición de codón. Una vez que todos los ORF se procesan, las sumas dan una medida aproximada de la preferencia de cada posición de codón por G y C. Usando esta información, Prodigal construye puntajes de codificación preliminares para cada gen en el genoma, y realiza luego métodos de programación dinámica con el fin de obtener estadísticas de codificación de hexámeros, motivos de sitios de unión a ribosoma y otras estadísticas que permiten finalmente identificar a los genes de la secuencia en cuestión.

1.7.Análisis funcional y perfil metabólico: Genomaple

Genomaple (Genomaple-2.3.2 *Genome Metabolic And Physiological potentiaL Evaluator*) es un sistema automatizado para inferir las funciones integrales potenciales que albergan los genomas y metagenomas, y para el cálculo de la relación de completitud del módulo (MCR) en cada módulo funcional definido por la Enciclopedia Kioto de Genes y Genomas (KEGG). El programa asigna un identificador de ortología KEGG (KO) al gen a analizar usando un servidor de anotación automática (KAAS) (el cual a su vez utiliza la herramienta BLAST o GHOSTX). Luego mapea los genes KO asignados en los módulos funcionales KEGG y calcula el MCR de cada módulo funcional y su abundancia cuando el módulo es completo. Existen dos métodos para la asignación de KO por KAAS: el método del mejor hit bidireccional (BBH) indicado para conjuntos de genes completos de genomas o *contigs*; y el método del mejor hit unidireccional (SBH), el cual se utiliza cuando se tienen secuencias cortas en metagenomas o genomas incompletos.

1.8.Enzimas de interés en el rumen vacuno

Existen varios ejemplos de enzimas presentes en el rumen vacuno con interés biotecnológico que han sido estudiados y reportados en la literatura biomédica. Algunos de ellos son:

- Lipasas: Estas enzimas catalizan la hidrólisis de grasas y aceites, y encuentran aplicaciones en la industria alimenticia, farmacéutica, de detergentes, entre otras.
- Proteasas: Estas enzimas rompen los enlaces peptídicos de las proteínas y son muy utilizadas en la industria alimentaria para ablandar la carne, así como en la industria de los detergentes y del cuero, entre otras.
- Amilasas: Catalizan la hidrólisis del almidón y se utilizan en la industria alimentaria para la producción de edulcorantes, entre otras aplicaciones.
- Celulasas: Participan en la descomposición de la celulosa y encuentran aplicaciones en las industrias de biocombustibles, textiles y papel.
- Pectinasas: estas enzimas descomponen la pectina, un polisacárido complejo que se encuentra en las paredes celulares de las plantas, y se utilizan en la industria alimentaria para la producción de jugos y vinos, entre otras aplicaciones.

- Xilanasas: Descomponen el xilano, un componente importante de las paredes celulares de las plantas, y se utilizan en las industrias del papel y los biocombustibles, entre otras.
- Lacasas: Estas enzimas están involucradas en la oxidación de compuestos fenólicos y tienen aplicaciones potenciales en la producción de biocombustibles y el tratamiento de aguas residuales.
- Esterasas: Existe un interés creciente en las feruloil esterasas (Wong *et al.* 2019), que catalizan la hidrólisis de ácidos ferúlicos unidos a éster presentes en los polisacáridos vegetales. Estas poseen aplicabilidad en las industrias de alimentos y farmacéutica, pulpa y papel y biocombustibles.
- Oxidasas: Catalizan reacciones de óxido-reducción (redox), y tienen una amplia variedad de especificidades de sustrato y mecanismos de reacción. Estas enzimas se utilizan en aplicaciones industriales como la decoloración de tintes, la biorremediación de suelos y aguas, y la biorrefinería (Ufarté, et. Al. 2018).

1.9. Base de datos CAZy

Entre las enzimas de interés nombradas anteriormente, el conjunto integrado por enzimas activas sobre carbohidratos (CAZymas), son de particular importancia. Estas son responsables de sintetizar, degradar y modificar carbohidratos complejos y glicoconjugados en todos los organismos; siendo particularmente abundantes en genomas de plantas y microorganismos que degradan plantas.

La base de datos CAZy se especializa en la visualización y análisis de información estructural, bioquímica y genómica de CAZymas, y representa actualmente la base de datos más integral de este tipo de proteínas, la cual consta a la fecha de 451 familias (Drula *et al.*, 2022).

CAZy clasifica a las familias de CAZymas en seis clases principales: glicosiltransferasas (GTs), glucósido hidrolasas (GHs), polisacárido liasas (PLs), módulo de unión a carbohidratos (CBM) y enzimas para actividades auxiliares (AAs).

1.9.1. Anotación de CAZymas: dbCAN2

dbCAN2 (Zhang *et al.*, 2018) es un servidor *web* para la anotación automática de CAZymas. El programa permite secuencias de entrada tanto proteicas como nucleotídicas. Estas últimas deben provenir de genomas o metagenomas procariotas, sobre las cuales se realiza luego la predicción de las secuencias proteicas utilizando Prodigal (Hyatt *et al.*, 2010) o FragGeneScan (Rho, Tang and Ye, 2010).

Para la anotación, dbCAN2 utiliza tres herramientas diferentes: HMMER para la determinación de dominios específicos de CAZymas de acuerdo con la base de datos *Hidden Markow Model* (HMM) generada por dbCAN (Yin *et al.*, 2012); DIAMOND para la detección rápida de coincidencias Blast contra la base de datos CAZy; y Hotpep para motivos cortos conservados en la biblioteca de reconocimiento de patrones peptídicos (PPR).

Los resultados de estos tres métodos se combinan, y las superposiciones se visualizan en un diagrama de Venn.

El programa permite además la identificación de factores de transcripción (TFs), transportadores (TCs) y *clusters* de genes (CGCs).

1.10. Base de datos RumiRef

Con el fin de hallar proteínas que fueran propias de los metagenomas analizados, se llevó a cabo en este trabajo la búsqueda de secuencias proteicas ortólogas contra la base de datos RumiRef.

RumiRef hace referencia a una base de datos de proteínas ruminales construida por Stewart et al. (Stewart *et al.*, 2019) a partir de la comparación de un conjunto de datos no redundantes de proteínas de 4941 genomas no cultivados de rumen (RUGs) y 460 genomas disponibles públicamente de la colección Hungate (10,69 millones de proteínas) siguiendo el modelo de UniRef (Suzek *et al.*, 2015). Las proteínas se agruparon de acuerdo a su porcentaje de identidad en 100 % (9,45 millones de grupos), 90 % (5,69 millones de conglomerados) y 50% (2,45 millones de conglomerados), conformando así las bases de datos RumiRef100, RumiRef90 y RumiRef50, respectivamente.

1.10.1.Búsqueda de secuencias proteicas ortólogas: Metaphor

Metaphor (Van der Veen *et al.*, 2014) determina relaciones ortólogas y parálogas basadas en *Bi-directional Best Hit* (BBHs). En forma general, el *software* crea una matriz donde todas las comparaciones posibles del metagenoma son representados. Sobre la base de esta matriz, las traducciones de proteínas de todos los genes predichos se buscan en BLAST contra los otros metagenomas, después de lo cual se analizan y conservan los mejores resultados.

2. Objetivos

2.1. Objetivo general

El objetivo general de este trabajo es contribuir a un mayor entendimiento de las funciones de la microbiota ruminal y su interacción con el animal huésped. En particular este proyecto se propone utilizar técnicas de secuenciación de alto rendimiento para estudiar la complejidad del rumen, enfocándose en el descubrimiento de enzimas adaptadas a los requerimientos degradativos de los rumiantes de la zona.

2.2. Objetivos específicos

I. Ensamblar metagenomas de rumen vacuno disponibles en el laboratorio, utilizando la herramienta Megahit con diferentes parámetros, y evaluar la calidad de los ensamblados.

II. Reconstruir genomas a partir de los *contigs* generados en el objetivo anterior, utilizando el *software* MetaBAT2.

III. Analizar el proteoma derivado de los metagenomas. Se llevará a cabo una predicción de las proteínas codificadas y perfil metabólico; se evaluará la presencia de proteínas activas sobre carbohidratos. Además, se detectarán proteínas únicas presentes en las muestras.

3. Materiales y métodos

3.1. Material biológico

Se trabajó con cuatro metagenomas de rumen de vaca jóvenes (Y) y adultas (A) alimentadas con una dieta pobre (P) o rica (R) en nutrientes, la cual consistió en pastura ryegrass o bien pastura suplementada con ensilaje de maíz, ensilaje de alfalfa, granos de maíz y soja, respectivamente.

Los metagenomas se identificaron como AP4 (Adulta con dieta pobre), AR5 (Adulta con dieta rica), YP1 (Joven con dieta pobre) e YR2 (Joven con dieta rica).

3.2.Secuenciación

Las bibliotecas se prepararon usando el *kit* de preparación de muestras de ADN Nextera (Illumina) siguiendo la guía del usuario del fabricante, y se secuenciaron en pares durante 300 ciclos con el sistema HiSeq (Illumina) de MR DNA (Shallowater, TX, EE. UU.). Se generaron lecturas emparejadas de 150 pb, lo que dio como resultado entre 2,3 y 2,7 Gb por muestra (entre 27 y 31,6 millones de lecturas emparejadas).

3.2.1. Procesamiento de las secuencias

Se usó Trimmomatic 0.32 (Bolger *et al.*, 2014) para realizar el recorte y el filtrado de calidad de las lecturas emparejadas de Illumina, usando los siguientes parámetros: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10; LEADING:20; TRAILING:20; CROP:148; HEADCROP:16; SLIDINGWINDOW:1:3; MAXINFO:40:0.5; MINLEN:40. Las lecturas recortadas obtenidas se cargaron, se controló la calidad y se analizaron mediante MG-RAST (Meyer *et al.*, 2008), y se puede acceder a ellas mediante el siguiente enlace http://v4-web.mg-rast.org/linkin.cgi?project =mgp20001.

3.3. Ensamblaje

Las lecturas con control de calidad se descargaron de MG-RAST y los metagenomas se ensamblaron *de novo* utilizando Megahit v1.1 (Li *et al.*, 2016). Se probaron varios parámetros, los mejores resultados se obtuvieron con --k-step 10.

*Job*1 (J1): parámetros por defecto

*Job*2 (J2): --presets meta-sensitive

*Job*3 (J3): --k-step 10

*Job*4 (J4): --min-count 1 --k-step 10

*Job*5 (J5): --presets meta-large

*Job*6 (J6): --k-min 21 --k-max 101 --k-step 10

- min-count: multiplicidad mínima para filtrar (k_min+1)-meros, por defecto es 2.
- k-min: mínimo tamaño de k-mero (<=127). Debe ser un número impar. Por defecto es 21.
- k-max: máximo tamaño de k-mero (<=127). Debe ser un número impar. Por defecto es 99.
- k-step: incremento del tamaño de k-mero de cada iteración (<=28). Debe ser un número par. Por defecto es 20.
- k-list: lista separada por comas de tamaños de k-meros. Todos deben ser impares, en el rango de 15-127, incremento <=28. Anula `--k-min', `--k-max' y `--k-step'
- presets: anula a un grupo de parámetros. Valores posibles:
    - meta: '--min-count 2 --k-list 21,41,61,81,99'. Metagenomas genéricos. Por defecto.
    - meta-sensitive: '--min-count 2 --k-list 21,31,41,51,61,71,81,91,99'. El ensamblaje del metagenoma es más sensible pero más lento.

- meta-large: '--min-count 2 --k-list 27,37,47,57,67,77,87'. Se usa para metagenomas grandes y complejos.

3.4.Calidad del ensamblado: Quast

Se aplicó un *loop* para descomprimir los *contigs* y correr el chequeo de parámetros para cada metagenoma y para cada *job*:

```
Metag.txt:
AP4
AR5
YR2
YP1

Job.txt:
J1
J2
J3
J4
J5
J6

for i in $(cat metag.txt)
do
    for j in $(cat Job.txt)
    do tar -xzvf ${i}.final.contigs.${j}.fa.gz
    mv final.contigs.fa ${i}.final.contigs.${j}.fa
    /home/Soft/quast-5.0.2/quast.py  ${i}.final.contigs.${j}.fa  -t
16 -o quast_${i}_${j}
    rm ${i}.final.contigs.${j}.fa
    done
done
```

Para calcular la cobertura se procedió al empleo de BWA MEM para mapear las lecturas al ensamblaje filtrado y Samtools para convertir al formato BAM. Luego se utilizó el *script* jgi_summarize_bam_contig_depths del paquete MetaBAT2 para calcular la cobertura del archivo BAM resultante.

```
for i in $(cat metag.txt)
do
    for j in $(cat Job.txt)
    do tar -xzvf ${i}.final.contigs.${j}.fa.gz
    mv final.contigs.fa ${i}.final.contigs.${j}.fa
    bwa index ${i}.final.contigs.${j}.fa
    bwa  mem  -t  16  ${i}.final.contigs.${j}.fa  ${i}_reads_used.gz
>aln_${i}_${j}.sam
```

```
    samtools view -@ 16 -S -b aln_${i}_${j}.sam > aln_${i}_${j}.bam
    samtools sort -m 2G -@ 12 aln_${i}_${j}.bam -o
aln_${i}_${j}.sorted.bam
    /home/Soft/metabat/bin/jgi_summarize_bam_contig_depths --
outputDepth depth_${i}_${j}.txt aln_${i}_${j}.sorted.bam
    rm ${i}.final.contigs.${j}.fa
    rm aln_${i}_${j}.sam
    rm aln_${i}_${j}.bam
    rm aln_${i}_${j}.sorted.bam
    rm *.amb
    rm *.ann
    rm *.bwt
    rm *.pac
    rm *.sa
    done
done
```

3.5. Binning

La obtención de los *bins* se llevó a cabo mediante el *software* Metabat2 (Kang *et al.*, 2019). La cobertura se calculó tal como se describió en la sección anterior. Se usó BWA MEM para mapear las lecturas al ensamblaje filtrado y Samtools para convertir al formato BAM; y posteriormente se utilizó el *script* jgi_summarize_bam_contig_depths del paquete MetaBAT2 para calcular la cobertura del archivo BAM resultante.

Ejemplo para un metagenoma

```
#crear un indice
bbmap.sh ref=final.contigs.AR5.fa.gz
#mapear reads
bbmap.sh in=AR5-mgm4718365.3.150.dereplication.passed.fna
out=megahit.sam threads=15

#convertir sam to bam
samtools view -S -b megahit.sam > megahit.bam
#ordenar el archivo bam
#samtools sort -m memoria por thread -@ nro de threads ….
samtools sort -m 2G -@ 12 megahit.bam -o megahit.sorted.bam

#generar el archivo depth
jgi_summarize_bam_contig_depths --outputDepth depth.txt
*.sorted.bam

#correr metabat
metabat2 -t 16 -m 2000 -i final.contigs.YR2.fa.gz -a depth.txt -o
bins_dir/bin
```

Cobertura:

Ejemplo para 1 bin. Se generó un *loop* para todos los *bins* obtenidos.

```
#indice
bwa index bin.1.fa

#aling
bwa mem -t 16 bin.1.fa mgm4718367.3.150.dereplication.passed.fna.gz
>aln_bin1.sam

#convertir sam to bam
samtools view -@ 16 -S -b aln_bin1.sam > aln_bin.bam
#ordenar
samtools sort -m 2G -@ 12 aln_bin.bam -o aln_bin.sorted.bam

#generar el archivo depth
jgi_summarize_bam_contig_depths       --outputDepth       depth.txt
*.sorted.bam
```

3.5.1. Evaluación de calidad de los bins

La evaluación de calidad de todos los *bins* obtenidos en el apartado anterior se llevó a cabo con el programa CheckM (Parks *et al.*, 2015), utilizando las opciones lineage_wf, -t 16, -x fa; filtrando luego aquellos genomas con completitud ≥ 80% y contaminación ≤ 10% (Stewart *et al.*, 2018).

Para cada *bin* se calculó una cobertura promedio ponderada por la longitud de los *contigs* presentes en cada uno de ellos; y esta información se incluyó en la tabla de calidad obtenida con CheckM.

```
checkm lineage_wf -t 16 -x fa --reduced_tree ./bins_dir ./checkm_out
>out_terminal.txt
```

3.6.MiGA

Para la clasificación taxonómica, los *bins* obtenidos en el apartado anterior se cargaron al servidor online http://microbial-genomes.org/ seleccionando la opción MAGs (Metagenome-assembled genomes).

3.7. Prodigal

Se llevó a cabo la predicción de genes y traducción utilizando la siguiente línea de comando para cada metagenoma:

```
prodigal -i my.genome.fna -o my.genes -a my.proteins.faa
```

3.8. Genomaple

Mediante el servidor Genomaple se realizó la asignación de KO de las proteínas presentes en las muestras, y se calculó el MCR de los principales módulos KEGG relacionados a la nutrición y salud de los animales. Los módulos se consideraron biológicamente factibles cuando el *Q value* fue < 0,5.

Dado que Genomaple tiene un límite de secuencias que pueden ser cargadas, los proteomas se dividieron en tres partes cada uno utilizando pyfasta. La predicción se realizó por separado sobre cada archivo, y luego se juntaron, realizando una comparación de todos contra todos.

3.9. Dbcan2

Para la predicción de CAZymas se utilizó Dbcan2 de forma local, en el cual las proteínas predichas se compararon con los HMM provistos por HMMER para las familias y subfamilias CAZymas obtenidas de la base de datos CAZy. Los resultados se filtraron con un valor de cobertura de 0,35 y un *e-value* de 1E-15.

Para ello se siguieron los siguientes pasos:

1. Se descargó dbCAN-fam-HMMs.txt, hmmscan-parser.sh
2. Se descargó el paquete HMMER 3.0 [hmmer.org] y se instaló apropiadamente.
3. Se formateó HMM db: hmmpress dbCAN-fam-HMMs.txt
4. hmmscan --domtblout yourfile.out.dm dbCAN-fam-HMMs.txt yourfile > yourfile.out
5. sh hmmscan-parser.sh yourfile.out.dm > yourfile.out.dm.ps (si el alineamiento > 80aa, usar *E-value* < 1e-5, sino usar *E-value* < 1e-3; fracción de cobertura de HMM > 0,3)

6. cat yourfile.out.dm.ps | awk '$5<1e-15&&$10>0.35' > yourfile.out.dm.ps.stringent

Ejemplo para un metagenoma:

```
hmmscan --cpu 15 --domtblout AP4_clean.out.dm dbCAN-HMMdb-V9.txt
AP4_clean.fasta>AP4_clean.out

python hmmscan-parser.sh AP4_clean.out.dm > AP4_clean.out.dm.ps

cat AP4_clean.out.dm.ps | awk '$5<1e-15&&$10>0.35' >
AP4_clean.out.dm.ps.stringent
```

Para contar el total de CAZymas en cada muestra, se utilizó un *script* R ad hoc. Los gráficos se generaron con R usando el paquete ggplot2 (Wickham, H. and Wickham, M., 2016).

```
#Leer archivo y generar columna con numero de proteinas que aparecen
en él
archivo<-read.table("AP4_clean.out.dm.ps.stringent", header=FALSE,
dec=".", sep="\t")
archivo<-as.data.frame(table(archivo$V1))

#Sacar el ".hmm" a la columna de proteinas
library("tidyverse")
AP4<- as.data.frame(str_replace_all(archivo$Var1, ".hmm", "")

#Unir columnas de prot y frecuencias
AP4<-cbind(AP4, archivo$Freq)

#cambiar nombres de columnas
names(AP4)<- c("CH","AP-4")

#Archivo AR-5
datos<-read.table("AR5_clean.out.dm.ps.stringent", header=FALSE,
dec=".", sep="\t")
datos<-as.data.frame(table(datos$V1))
AR5<-as.data.frame(str_replace_all(datos$Var1, ".hmm",""))
AR5<-cbind(AR5, datos$Freq)
names(AR5)<- c("CH","AR-5")

#Unir los dos dataframes por la columna "CH" y que aparezcan los
valores nulos
datos_completos<-merge (AP4, AR5, by = "CH", all=TRUE)

#Generar tabla
#replaces the NA values with 0
write.csv(datos_completos,"tabla.csv",row.names=F,na = "0")
```

3.10. Metaphor

3.10.1. Búsqueda de CAZymas no ortólogas

Para la detección de CAZymas no ortólogas a las reportadas en RumiRef, se empleó el *software* Metaphor, el cual permitió en un principio obtener una lista de las proteínas ortólogas a esta base de datos con un % Identidad >30% y *alignment lenght* >80%. Mediante pyfasta se extrajeron luego aquellas que no estaban presentes en la lista (no ortólogas).

3.10.2. GH no ortólogas

Debido a la importancia de las proteínas GH en la digestión ruminal, se analizaron las GH no ortólogas a la base de datos RumiRef respecto al total de GH halladas por Dbcan2 en cada uno de los metagenomas.

Para evitar trabajar con proteínas truncas que pudieran haberse generado durante el ensamblado, se seleccionaron aquellas secuencias de GH no ortólogas obtenidas mediante Metaphor que fueran mayores o iguales al tamaño mínimo de las GH. Este valor se seteó en 213 AA, correspondiente a la Endo-1,4 beta xilanasa A de *B. subtilis* (https://www.uniprot.org/).

3.11. Análisis estadístico

Para determinar si existe correlación entre la edad y la dieta de los rumiantes con la composición de CAZymas, se utilizó el software STAMP v2.1.3 (Parks *et al.* 2014). La significación estadística de las diferencias entre los grupos se evaluó mediante la prueba de Welch de dos colas (Thomas *et al.* 2017), mientras que la prueba estadística G (w + yates) + Fisher de dos colas se utilizó para examinar las diferencias entre las muestras según lo recomendado por Jose *et al.* 2017a, y Parks *et al.* 2014. Para ambos análisis, se aplicó el método de corrección de prueba múltiple de tasa de descubrimiento falso (FDR) de Storey. Se asignó significación estadística a las características con un valor de *p-value* corregido < 0,05 (Thomas *et al.* 2017).

4. Resultados y discusión

El objetivo de este trabajo fue caracterizar el potencial metabólico y biotecnológico de la microbiota de rumen vacuno, enfocándose en el descubrimiento de enzimas relacionadas con la alimentación de la región. Para llevar a cabo este estudio, se procedió al ensamblado de las secuencias previamente depuradas, de forma de obtener secuencias más largas que permitan analizar la secuencia completa de enzimas, y predecir así con mayor precisión su función pudiendo recurrir al contexto génico de ser necesario.

4.1. Ensamblado y evaluación de calidad

La obtención de los ensamblados se llevó a cabo utilizando el programa Megahit tal como se describe en la sección 3.3. La optimización de los ensamblados se evaluó empleando diferentes parámetros del programa, generándose así seis *scripts* (*Jobs*) para cada metagenoma, de los resultados arrojados por el *software* se extrajo la información sobre el número y la longitud de los *contigs* obtenidos. Para completar el análisis de calidad de los ensamblados, se determinaron los valores de N50 y cobertura promedio mediante Quast y BWA respectivamente.

Los resultados se muestran en la Tabla 1.

Tabla 1: Optimización de los ensamblados utilizando distintos parámetros de Megahit. El valor de N50 y la cobertura promedio se obtuvieron mediante Quast y BWA respectivamente.

Jobs	Metagenoma	Cobertura promedio	Número de contigs	Longitud total de contigs (pb)	Número de contigs ≥1000 bp	Longitud total de contigs ≥1000 bp	Longitud máxima de contig (pb)	N50 (pb)
J1	AP4	3,55	638.163	350.457.885	41.258	82.987.877	187.542	899
J1	AR5	3,33	637.680	346.833.294	39.895	66.872.494	26.803	782
J1	YP1	4,07	703.288	422.796.693	60.646	122.730.393	682.556	963
J1	YR2	4,39	617.642	397.828.108	60.534	140.603.709	231.178	1.157
J2	AP4	2,84	1.091.565	545.062.013	49.036	94.786.377	205.179	817
J2	AR5	2,72	1.057.805	526.801.306	49.223	79.203.696	26.804	763
J2	YP1	3,32	1.104.138	607.542.065	74.380	143.552.505	682.556	898
J2	YR2	1,48	1.001.448	570.709.687	70.286	157.891.432	223.525	1.019
J3	AP4	3,54	639.251	353.090.431	42.321	85.513.846	161.546	911
J3	AR5	3,32	640.470	352.442.301	42.551	72.585.769	32.543	801
J3	YP1	4,06	703.350	427.697.815	63.550	129.133.365	556.226	988
J3	YR2	4,38	617.761	399.507.423	61.125	142.724.870	240.016	1.173
J4	AP4	2,86	1.073.236	536.224.260	48.341	93.795.739	154.255	820
J4	AR5	2,74	1.044.413	519.280.803	48.128	77.587.383	26.802	762
J4	YP1	3,34	1.094.052	601.784.611	73.574	142.358.143	556.226	899
J4	YR2	3,56	991.165	564.875.621	69.592	156.534.738	302.510	1.022
J5	AP4	2,99	997.684	500.703.643	45.944	90.121.764	231.562	831
J5	AR5	2,86	982.233	487.712.791	46.312	74.695.165	30.601	772
J5	YP1	3,44	1.054.176	581.126.868	71.587	139.299.712	682.556	907
J5	YR2	3,69	938.179	537.873.370	67.023	152.260.654	265.413	1.042
J6	AP4	3,55	641.715	352.374.811	41.854	84.049.311	121.058	907
J6	AR5	3,32	648.350	351.545.846	41.302	69.796.103	26.803	796
J6	YP1	4,06	711.432	426.697.625	62.062	125.491.535	556.226	974
J6	YR2	4,38	619.742	398.666.810	60.826	141.131.312	240.016	1.165

Basándose en el valor de N50, el cual brinda información acerca del tamaño de los *contigs* en la muestra, puede verse que el *Job* 3 es el que arrojó un mayor valor para todos los metagenomas analizados. Este parámetro indica que el 50% de los *contigs* poseen un tamaño igual o mayor al valor calculado. Respecto a la cobertura, definida como el número de veces que una porción del genoma es secuenciada, se obtuvieron en promedio mayores valores en el *Job* 3 respecto al *Job* 2, 4 y 5, y estos fueron similares al *Job* 1.

Observando estos resultados, se eligieron los ensamblados generados en el *Job* 3 para proseguir con el posterior análisis de los metagenomas.

La Figura 1 muestra los gráficos de los principales parámetros evaluados. En cuanto al tamaño del ensamblado, longitud máxima de *contig* y número de *contigs*, el metagenoma YP1 fue el que mostró los mayores valores respecto a los demás. La cobertura promedio, por su parte, arrojó valores similares para todos los metagenomas.

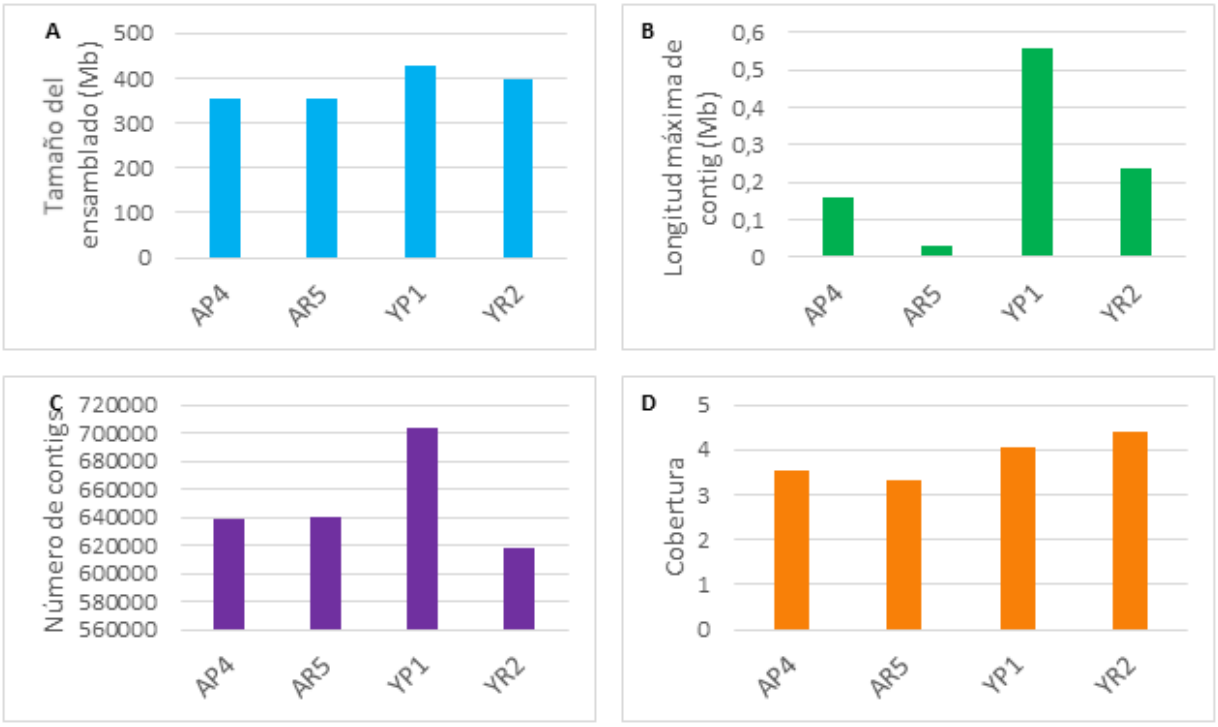

Figura 1: Ensamblados optimizados por el parámetro N50. Se muestra el tamaño de cada ensamblado (A), la longitud máxima de contig (B), el número de contigs (C) y los valores de cobertura (D) para cada uno de los metagenomas.

4.2. Binning y evaluación de calidad

A continuación, se procedió a ensamblar genomas a partir de los *contigs* obtenidos para cada metagenoma en la sección anterior, a fin determinar si las secuencias permitían obtener MAGs (*metagenome-assembled genomes*) de calidad o de interés biotecnológico.

Mediante el método de *binning* llevado a cabo por Metabat2 se generaron un total de 23 *bins* para el metagenoma AP4, 12 para AR5, 31 para YP1 y 40 para YR2, a los cuales se les calculó su cobertura promedio mediante BWA tal como se describe en la sección 3.5. Los parámetros de calidad para todos los *bins* fueron obtenidos con el programa CheckM. Con esta información, se construyeron las Tabla 6, 7, 8 y 9 del anexo, y se seleccionaron luego aquellos *bins* con *completitud* ≥ al 80% y contaminación ≤ al 10% (Stewart *et al.*, 2018a).

La Figura 2 muestra una comparación entre el número de *bins* total y el número de *bins* filtrado de acuerdo a los criterios de calidad antes mencionados para cada metagenoma. Los

resultados muestran la obtención de entre 3 y 5 genomas ensamblados a partir de metagenomas (MAGs) para AP4, YP1 y YR2, los cuales se detallan en la Tabla 2. Respecto al metagenoma AR5, no fue posible obtener MAGs con estos criterios de selección.

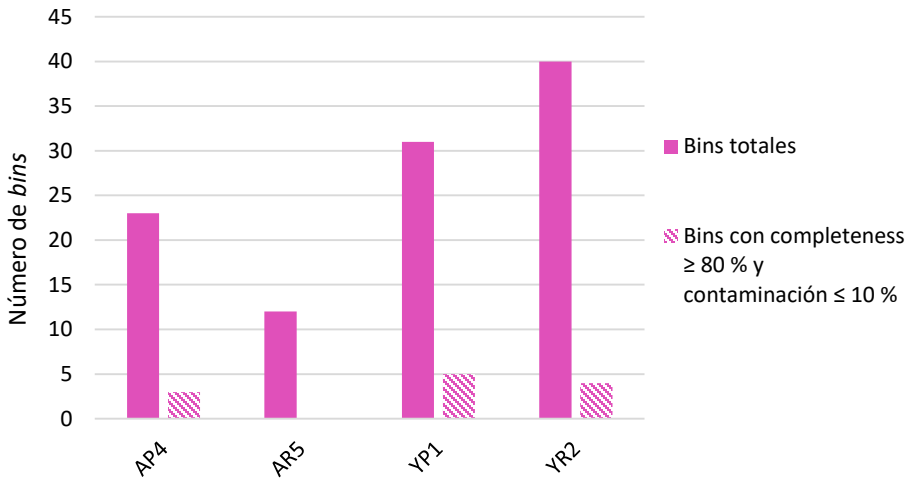

Figura 2: Comparación entre el número de bins total y el número de bins con completitud ≥80% y contaminación ≤10%

Tabla 2: Parámetros obtenidos para los bins con completitud ≥80% y contaminación ≤10%.

Metagenoma	Bin	Marcador de linaje	Genomas	Completitud (%)	Contaminación (%)	Heterogeneidad (%)	Cobertura promedio	Contig más largo (pb)	N50 (pb)
AP4	**1**	k__Bacteria (UID2569)	434	83,57	2,08	66,67	7,46	33.478	10.360
AP4	**17**	k__Bacteria (UID2495)	2.993	87,71	0,1	0	14,44	121.058	44.077
AP4	**23**	p__Bacteroidetes (UID2605)	350	89,81	2,63	71,43	12,34	81.217	24.794
YP1	**1**	f__Lachnospiraceae (UID1286)	57	93,67	0,5	0	13,40	556.226	195.201
YP1	**6**	o__Clostridiales (UID1120)	304	92,67	2,48	60	11,32	141.583	13.231
YP1	**18**	o__Bacteroidales (UID2657)	160	94,5	0,09	0	22,13	137.797	44.119
YP1	**23**	o__Bacteroidales (UID2657)	160	81,07	3,34	71,43	14,86	103.819	16.402
YP1	**28**	c__Gammaproteobacteria (UID4202)	67	91,49	0,54	50	6,13	21.776	7.363
YR2	**7**	k__Bacteria (UID2569)	434	82,26	0,07	75	15,40	135.585	38.521
YR2	**9**	o__Bacteroidales (UID2617)	213	81,75	4,85	14,29	6,64	43.647	8.324
YR2	**12**	c__Clostridia (UID1118)	387	88,06	2,02	100	15,62	39.992	11.028
YR2	**29**	p__Bacteroidetes (UID2605)	350	94,84	0,95	50	18,18	240.016	108.343

Además de la completitud y contaminación, CheckM brinda distintas estadísticas, de las cuales se muestran el marcador de linaje, genomas, heterogeneidad de cepa, *contig* más largo y N50.

El marcador de linaje hace referencia al rango taxonómico del conjunto de marcadores específicos del linaje utilizado para estimar la integridad del genoma, la contaminación y la heterogeneidad de la cepa (Tabla 2).

La heterogeneidad de cepa es determinada a partir del número de pares de marcadores de múltiples copias que superan un umbral de identidad de aminoácidos específico (por *default* = 90%). Un valor alto sugiere que la mayoría de la contaminación proviene de uno o más organismos estrechamente relacionados (es decir, potencialmente de la misma especie), mientras que un valor bajo sugiere que la mayoría de la contaminación proviene de fuentes filogenéticamente más diversas. Los resultados de heterogeneidad de cepa en los MAGs analizados arrojaron valores que variaron entre 0 y 100%.

Por otra parte, la contaminación fue entre el 0,07 y el 4,85%; mientras que los valores de cobertura promedio fueron entre 6,13 y 22,13. Como referencia en cuanto a valores de cobertura aceptables, Stewart y col reportaron valores entre 8,9 y 1189 (Stewart *et al.*, 2018). Se puede decir que, si bien los valores no se asemejan a los máximos reportados, la cobertura obtenida en este caso está dentro del rango publicado.

4.3.Asignación taxonómica

La asignación taxonómica de los *bins* obtenidos en el apartado anterior fue realizada como se describe en la sección 3.6 utilizando el servidor *web* MiGA.

Los organismos identificados para los MAGs de AP4 y YP1 se muestran en la Tabla 3. En esta se resaltan los niveles taxonómicos predichos por el servidor que tuvieron un *p-value* ≤0,5; y se indican además los posibles taxones restantes a los que pertenecerían los organismos.

El *bin* 23 de YP1, así como los *bins* provenientes del metagenoma YR2, si bien habían sido seleccionados en cuanto a sus valores de contaminación y completitud, no lograron cumplir con los criterios de calidad internos de MiGA. Por este motivo, no fue posible su asignación taxonómica.

Tabla 3: Organismos identificados con MiGA. Los taxones resaltados corresponden a organismos clasificados por el servidor con un p-value ≤0,5.

Metagenoma	Bin	Filo	Clase	Orden	Familia	Género	Especie	Calidad MiGA (%)
AP4	1	**Bacteroidetes**	***Bacteroidea***	***Bacteroidales***	*Tannerellaceae*	*Parabacteroides*	*Parabacteroides distasonis*	67,9
AP4	17	**Proteobacteria**	***Gammaproteobacteria***	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*	*Pseudomonas citronellolis*	82,3
AP4	23	**Bacteroidetes**	***Bacteroidea***	***Bacteroidales***	*Rikenellaceae*	*Alistipes sp. dk3624*	*Alistipes sp*	81,1
YP1	1	**Firmicutes**	***Clostridia***	***Clostridiales***	***Lachnospiraceae***	*Roseburia*	*Roseburia intestinalis*	83,0
YP1	6	**Firmicutes**	***Clostridia***	*Clostridiales*	*Hungateiclostridiaceae*	*Mageeibacillus*	*Mageeibacillus indolicus*	85,8
YP1	18	**Bacteroidetes**	***Bacteroidea***	***Bacteroidales***	***Prevotellaceae***	***Alloprevotella***	*Alloprevotella sp*	87,0
YP1	28	**Proteobacteria**	***Gammaproteobacteria***	***Xanthomonadales***	***Xanthomonadaceae***	***Xanthomonas***	***Xanthomonas vasicola***	94,3

Los organismos hallados en los metagenomas que cumplieron con los criterios de calidad de MiGA pertenecieron a los filos Proteobacteria, Firmicutes y Bacteroidetes. Estos filos bacterianos comprenden varios taxones capaces de catabolizar una amplia gama de componentes de los alimentos; y han sido reportados en muestras de rumen vacuno (Wang *et al.* 2019; Hernández, R. *et al.*, 2022). De ellos, los órdenes *Clostridiales* y *Bacteroidales*, están especialmente relacionados a la degradación de fibras y polisacáridos en el rumen bovino (Wright and Klieve, 2011; Asma *et al.*, 2013).

La importancia de identificar MAGs radica en un mayor entendimiento de la microbiota ruminal.

La variación en la composición de la microbiota puede deberse a diversos factores, entre los cuales pueden destacarse la edad y la dieta del animal.

La colonización del rumen por microorganismos comienza después del nacimiento y continúa desarrollándose a medida que el ternero en crecimiento se expone a los forrajes. La comunidad microbiana del rumen de los terneros jóvenes es heterogénea, pero a medida que las vacas se acercan a la madurez (alrededor de los 2 años) se vuelve más homogénea (Jami *et al.*, 2013). Se ha demostrado que la dieta resultó ser la principal fuerza impulsora del cambio en la composición microbiana (Xue *et al*., 2018).

Al tener poca diferencia de abundancia entre las muestras y solo dos muestras por grupo, los resultados obtenidos en este trabajo no permiten una estadística confiable. Sin embargo, se

logró la identificación de MAGs asociados al tracto gastrointestinal y a la degradación de material vegetal, lo cual resulta de interés en la comprensión de la función de estos organismos en el animal húesped sometido a determinadas condiciones nutricionales.

4.4.Análisis funcional y perfil metabólico mediante Genomaple

Con el objetivo de obtener una primera descripción de las funciones y del perfil metabólico de los metagenomas, se llevó a cabo un análisis utilizando el programa Genomaple v2.3.2, tal como se describe en la sección 3.8.

A continuación, se describen los porcentajes de completitud de algunos módulos KEGG seleccionados por su relación con el bienestar animal (nutrición, salud) y potencial biotecnológico.

Metabolismo de carbohidratos

Los porcentajes de completitud calculados para cada módulo (MCR) indicaron que los metagenomas poseen los metabolismos centrales de carbohidratos completos o casi completos, siendo la mayoría de los módulos biológicamente factibles (Q < 0,5). Sólo tres módulos no fueron biológicamente factibles: la vía Semifosforilativa Entner-Doudoroff (gluconato => glicerato-3P, Módulo KEGG M00308) en las muestras AP-4 y AR-5 (MCR 80%, Q = 0,875); la vía Semi-fosforilativa Entner-Doudoroff (gluconato/galactonato => glicerato-3P, Módulo KEGG M00633) en todas las muestras (MCR < 50%, Q > 0,881; y la vía No-fosforilativa Entner-Doudoroff (gluconato/galactonato => glicerato, Módulo KEGG 00309) en las muestras AP-4 y AR-5 (MCR 33%, Q = 0,982).

Con respecto a otros metabolismos de carbohidratos, la degradación de glucógeno (Módulo KEGG 00855) fue biológicamente factible en todas las muestras. La degradación de pectina (Módulo KEGG 00081) no fue factible en la muestra YR-2 (MCR 66% Q = 0,571), debido a la ausencia de la enzima galacturano 1,4 alfa-galactouronidasa [EC:3.2.1.67]. El D-Galacturonato es el principal constituyente monomérico de la pectina, el cual puede ser degradado por el módulo KEGG M00631 en todas las muestras. En plantas, el D-glucuronato participa en el metabolismo de nucleótidos y es un constituyente del tejido vegetal (Reiter

and Vanzin 2001), siendo su degradación llevada a cabo por el módulo KEGG M00061 presente en todas las muestras.

Fijación de carbono

Entre los módulos KEGG para la fijación de carbono, el ciclo del ácido C4-dicarboxilico (NADP – enzima málica, módulo KEGG M00172) (MCR > 75%, Q < 0,5), la vía reductiva acetil-CoA (Vía Wood-Ljungdahl, Módulo KEGG M00377) (MCR > 85,7%, Q < 0,5) y la vía Fosfato acetiltransferasa-acetato quinasa (acetil-CoA => acetato, Módulo KEGG M00579) (MCR 100%, Q = 0) fueron biológicamente factibles en todas las muestras. El ciclo reductivo de las pentosas fosfato (ribulosa-5P => gliceraldehído-3P, Módulo KEGG M00166) (MCR 100%) fue factible en YP-1; reportándose en las demás muestras (MCR 75%, Q = 0,75) la falta de la enzima fosforibuloquinasa [EC:2.7.1.19]. El ciclo reductivo incompleto del citrato (acetil-CoA => oxoglutarato, Módulo KEGG M00620) (MCR 100%, Q = 0), fue factible en la muestra YR-2. El resto de las muestras carecieron de la subunidad A de la enzima piruvato carboxilasa [EC:6.4.1.1] (MCR 85.7%, Q = 0,75). En la muestra YP-1, además, estuvieron ausentes las subunidades A y B de la enzima fumarato reductasa (CoM/CoB) [EC:1.3.4.1] (MCR 71,4%, Q = 0,938).

Vías relacionadas con el metabolismo del metano

Tres de las cuatro vías de metanogénesis reportadas por genomaple fueron biológicamente factibles en las muestras analizadas. La vía (metilamina/dimetilamina/trimetilamina => metano, Módulo KEGG M00563) fue factible en todas las muestras (MCR 75%, Q = 0,5). Por otra parte, la vía metanogénica (CO_2 => metano, Módulo KEGG M00567) (MCR 100%, Q = 0) así como la metanogénesis (metanol => metano, Módulo KEGG M00356) (MCR > 66%, Q < 0,5) fue factible en las muestras YR-2, AP-4 y AR-5.

La metanogénesis (acetato => metano, módulo KEGG M00357) no fue factible en ninguna de las muestras (MCR < 83,3; Q > 0,75).

Ácidos grasos volátiles

Entre las vías que conducen a AGV en bacterias de rumen, sólo se halló la conversión de acetilCoA a acetato (módulo KEGG M00579, descripto arriba). El resto de las vías que

producen acetato, butirato o propionato no han sido descriptas en los resultados obtenidos con genomaple.

Sistema de transporte de lípidos, polioles y sacáridos

Se reportaron cuarenta y dos módulos asociados al sistema de transporte de lípidos, polioles y sacáridos, de los cuales trece fueron hallados completos en todas las muestras.

Entre los módulos KEGG informados, el sistema de transporte maltosa/maltodextrina (Módulo KEGG M00194) fue biológicamente factible en todas las muestras (MCR 100% Q=0).

El sistema de transporte de celobiosa, por otra parte, (Módulo KEGG M00206) fue factible sólo en la muestra YP-1 (MCR 75% Q=0,5), debido a que las demás muestras carecieron de la proteína permeasa cebG, además de la permeasa cebF en la muestra AR-5.

Sistema de transporte de aminoácidos y fosfato

Dentro de esta categoría se encontraron veinticinco módulos funcionales, de los cuales nueve fueron hallados completos en todas las muestras analizadas.

Las diferencias más notables se encontraron en los sistemas de transporte lisina/arginina/ornitina (Módulo KEGG M00225), histidina (Módulo KEGG M00226), y arginina (Módulo KEGG M00229), los cuales fueron biológicamente factibles sólo en la muestra YP-1 (MCR 100% Q=0); así como en el caso del sistema de transporte S-metilcisteína putativo (Módulo KEGG M00586), el cual fue biológicamente factible sólo en la muestra YR-2 (MCR 100% Q=0).

Sistema fosfotransferasa (PTS)

Se encontraron veintinueve módulos asociados al sistema PTS, once de los cuales se hallaron completos en todas las muestras.

El componente II beta-glucosido (arbutina/salicina/celobiosa)-específico (Módulo KEGG M00272) así como el componente II trealosa- específico (Módulo KEGG M00270) y el componente II celobiosa- específico (Módulo KEGG M00275) fueron biológicamente factibles y completos en las muestras YP-1, YR-2 y AR5.

Por otra parte, el componente II alfa-glucosido-específico (Módulo KEGG M00268), así como el componente II ascorbato-específico (Módulo KEGG M00283) fueron factibles y completos para todas las muestras excepto para YP-1.

El componente II 2-O-A-manosil-D-glicerato-específico (Módulo KEGG M00305), el componente II lactosa-específico (Módulo KEGG M00281) y el componente II galactitol-específico (Módulo KEGG M00279) fueron factibles sólo en las muestras AP-4, YP-1 e YR-2 respectivamente (MCR 100% Q=0).

Producción y Resistencia a antibióticos

Las vías metabólicas asociadas a la síntesis de policétidos o antibióticos beta-Lactámicos no se encontraron utilizando genomaple.

Los complejos transportadores de eflujo de drogas, tales como la bomba VexEF-TolC (resistencia a eritromicina, novobiocina, (Rahman *et al.*, 2007) Módulo KEGG M00720) se encontraron como biólogicamente factibles en las muestras YR-2 y AP-4 (MCR 66% Q =0,5). La bomba de eflujo MexMN-OprM (resistencia a macrólidos, fluoroquinolones (Mima *et al.*, 2005), Módulo KEGG M00821) fue hallada como probablemente activa en YP-1, YR-2 y AR-5 (MCR > 75% Q < 0,5).

El transportador EfrAB (resistencia a norfloxacina, ciprofloxacina, doxiciclina, entre otras (Lee et al. 2003), Módulo KEGG M00706) se mostró completo en las muestras YP-1 e YR-2 (MCR 100% Q=0). Por su parte, el transportador MdlAB/SmdAB (resistencia a norfloxacina, tetraciclina (Matsuo *et al.*, 2008), Módulo KEGG M00707) fue reportado completo en todas las muestras (MCR 100% Q=0); y el transportador PatAB (el cual confiere resistencia a la fluoroquinolona (Alvarado *et al*., 2017), Módulo KEGG M00708) fue hallado completo sólo en la muestra AP-4 (MCR 100% Q=0).

Las bombas de eflujo resistentes a multidrogas AcrAD-TolC (relacionadas con la resistencia a aminoglucósidos), AcrAB-TolC/SmeDEF (resistencia a tetraciclina), y MdtABC (involucrada en la resistencia a b-lactámicos y novobiocina) estuvieron presentes en todas las muestras (MCR > 66,7% Q < 0,5). La bomba de eflujo OqxAB (activa en fluoroquinolonas) fue hallada en las muestras YP-1 y AP-4 (MCR 100% Q=0), mientras que la bomba de eflujo NorA (también actuando con fluoroquinolonas) se halló completa en la muestra AR-5.

El transportador de resistencia a macrólidos MacAB-TolC se encontró completo en las muestras YP-1 e YR-2 (MCR 100% Q=0)

En cuanto a la resistencia a aminoglucósidos, las proteasas FtsH y HtpX (Módulos KEGG M00742 y M00743, respectivamente) se hallaron completas en todas las muestras. El sistema AmpC (involucrado en la resistencia a beta-lactámicos Módulo KEGG M00628) fue biológicamente factible en YP-1 e YR-2, y la resistencia a fluoroquinolona mediada por la proteína Qnr protectora de girasa (Módulo KEGG M00729) se encontró completa sólo en YP-1.

El módulo *Signature* descripto por Genomaple hace referencia a unidades funcionales de conjuntos de genes que describen características fenotípicas. Dentro de esta categoría se reportaron veinticinco módulos KEGG relacionados con la resistencia a drogas, hallándose dos de estos completos en todas las muestras analizadas: la resistencia a beta-lactámicos, el sistema Bla (Módulo KEGG M00627) y la resistencia a Imipenem (llevado a cabo por la represión de la porina OprD, módulo KEGG M00745).

Es de destacar que, la resistencia a Vancomicina tipo D-Ala-DLac (Módulo KEGG M00651) fue hallada con un MCR de 100% en las tres muestras (YP-1, YR-2 y AP-4). Por el contrario, la bomba de eflujo de Tetraciclina Tet38 fue hallada sólo en la muestra AR-5.

La bomba de eflujo de resistencia a multidrogas AdeABC (resistencia a aminoglucosidos, beta-lactámicos y tetraciclina, Módulo KEGG M00649) fue encontrada completa sólo en la muestra YP-1.

Por otra parte, la bomba de eflujo de resistencia a multidrogas AbcA (que confiere resistencia a meticilina, daptomicina, cefotaxima, y moenomicina, Módulo KEGG M00700) fue encontrada completa en YR-2 y AR-5

Degradación de xenobióticos

La presencia de vías de degradación de Xenobióticos fue escasa; sólo el clivaje de orto-catecol (catecol => 3-oxoadipato, Módulo KEGG M00568) fue biológicamente factible en la muestra AP-4 (MCR 100% Q = 0). Las vías de clivaje de tolueno, xileno y meta-catecol se mostraron con un MCR > 60% y Q> 0,5 en algunas de las muestras.

Patogenicidad

La firma de patogenicidad Pertussis T1SS (Modulo KEGG M0575) fue hallada casi completa y biológicamente factible (MCR 80% y Q =0,5) en todas las muestras, sugiriendo la presencia del patógeno en el rumen de las vacas estudiadas.

En general, el perfil metabólico de la microbiota del rumen vacuno descrito sugiere que los procesos de degradación de carbohidratos son completos o casi completos.
En cuanto a los procesos específicos, los módulos para la degradación de glucógeno y la mayoría de los carbohidratos son biológicamente factibles en todas las muestras, aunque la degradación de pectina no fue posible en una muestra debido a la ausencia de una enzima específica.
Este análisis sugiere que la microbiota es altamente especializada para la degradación de carbohidratos y la producción de AGV, lo que contribuye a la salud y nutrición del animal.

4.5. Enzimas activas sobre carbohidratos (CAZymas)

Debido a que el metabolismo de carbohidratos es conocido por ser la actividad metabólica más significativa en el rumen de vacas, se investigó el potencial de la descomposición de la lignocelulosa a través de la identificación de secuencias génicas putativas activas sobre carbohidratos desde bibliotecas metagenómicas. Las proteínas putativas de las muestras fueron alineadas con las familias CAZy usando Dbcan2 de forma local (sección 3.9).

La búsqueda contra la base de datos CAZy completa reveló que estaban presentes en las cuatro muestras, 83 familias Glucósido Hidrolasas (GH), 28 Módulos de Unión a Carbohidratos (CBM), 10 Polisacárido Liasas (PL), 13 Carbohidrato Esterasas (CE), 29 Glicosil Transferasas (GT) y una Actividad Auxiliar (AA). Con respecto a las CBMs, las funciones detectadas fueron las uniones a celulosa, quitina, galactano, galactosa, glucógeno, pectina, almidón y xilano, entre otras (Figura 3). Las familias CBM dominantes (Figura 4) fueron CBM6 (unión a celulosa) y 67 (unión a L-ramnosa) en todas las muestras. CBM32 (unión a galactosa y lactosa), 35 (unión a xilano), 50 (unión a quitina) y 20 (unión a almidón) fueron los siguientes más abundantes.

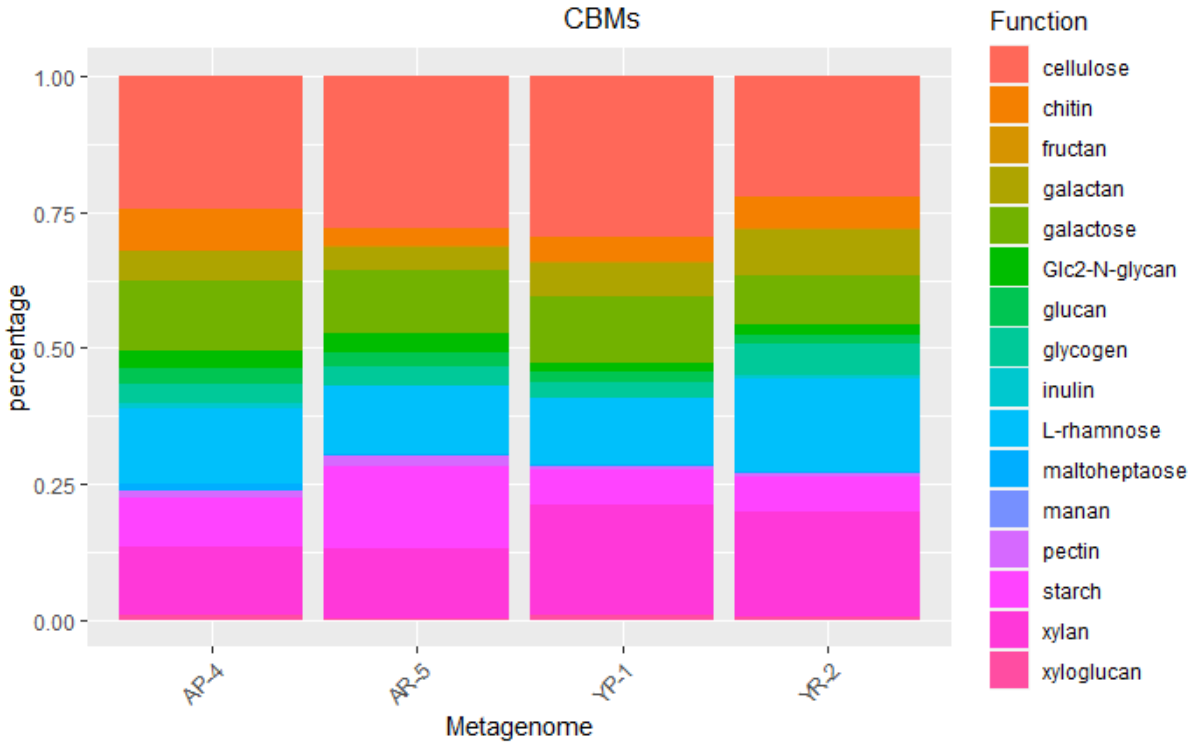

Figura 3: Distribución de CBMs en los metagenomas de acuerdo a su función.

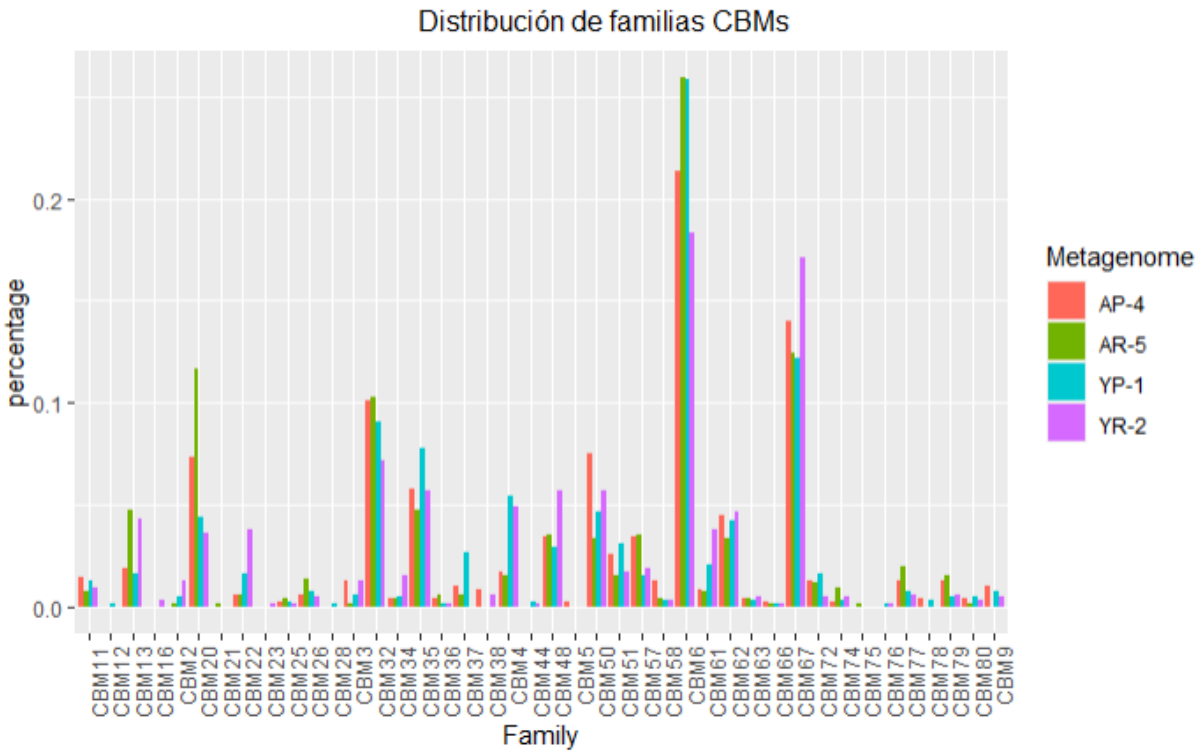

Figura 4: Distribución de las familias CBMs en cada uno de los metagenomas.

Las familias GH fueron agrupadas en cinco categorías (celulasas, endohemicelulasas, xiloglucanasas, enzimas desramificadoras y enzimas de degradación de oligo-sacáridos). Las más abundantes fueron las enzimas de degradación de oligosacáridos (66%), las enzimas desramificadoras fueron las segundas más abundantes (14%), seguidas por endohemicelulasas (11%), celulasas (6,5%) y finalmente xiloglucanasas (1,8%) (Tabla 4). Estas cuatro primeras tuvieron abundancias similares a lo reportado previamente por Wang *et al.* 2019.

Tabla 4: Abundancia de enzimas GH en los cuatro metagenomas.

Función	**Abundancia (%)**
Enzimas de degradación de oligosacáridos	66,25
Desramificadoras	13,88
Endo-hemicelulasas	11,53
Celulasas	6,5
Xiloglucanasas	1,85

Las familias más abundantes halladas fueron GH43, GH3 y GH13 que pertenecen a la categoría de enzimas degradadoras de oligosacáridos, seguido por las familias GH5 y GH25 que codifican celulasas y enzimas desramificadoras respectivamente (Figura 5 y Figura 6). Las mayores actividades reportadas por la familia GH43 son α-L-arabinofuranosidasas, endo-α-L-arabinanasas (o arabinanasas endo-procesivas) y β-D-xilosidasas que remueven cadenas laterales de arabinofuranosa de xilanos (The CAZypedia Consortium, 2018). En los metagenomas analizados, la mayor diferencia en cuanto al porcentaje de esta familia se observó en YP-2, donde la abundancia de GH43 fue menor respecto a los otros tres metagenomas. (Figura 6).

La familia de Glucósido Hidrolasa 3 agrupa actualmente β-D-glucosidasas, α-L-arabinofuranosidasas, β-D-xilopiranosidasas, N-acetil-β-D-glucosaminidasas (glucósido hidrolasas) y N-acetil-β-D- glucosaminida fosforilasas (glucósido hidrolasas), y N-acetil-β-D-glucosaminida fosforilasas. Las enzimas GH3 llevan a cabo una serie de funciones incluyendo la degradación de biomasa celulósica, remodelamiento de la pared celular vegetal y bacteriana, metabolismo energético y defensa contra patógenos (The CAZypedia Consortium, 2018).

GH13 actúa en sustratos conteniendo uniones α-glucósido. Las α-amilasas, pululanasas, α-glucosidasas, neopululanasas, por ejemplo, pertenecen a esta familia. Respecto a la abundancia de GH13, los metagenomas AR-5 y YR-2 tuvieron mayores porcentajes de esta familia en relación a AP-4 y YP-1 (Figura 6).

GH5 fue previamente conocida como “celulasa familia A”, donde actividades como endoglucanasa (celulasa) y endomananasa, tanto como exoglucanasas, exomananasas y β-glucosidasas y β-manosidasas están asociadas a esta familia.

GH25 son lisozimas que clivan el enlace β-1,4 glucosídico entre el ácido N-acetilmurámico (NAM) y N-acetilglucosamina (NAG) en el esqueleto de carbohidrato del peptidoglicano bacteriano. Ayudan al remodelado de peptidoglicano en procesos celulares tales como la división, y en hongos pueden ser agentes selectivos contra bacterias. La abundancia de esta familia fue mayor en el metagenoma AR-5 respecto a los demás, que tuvieron porcentajes de GH25 similares entre sí (Figura 6).

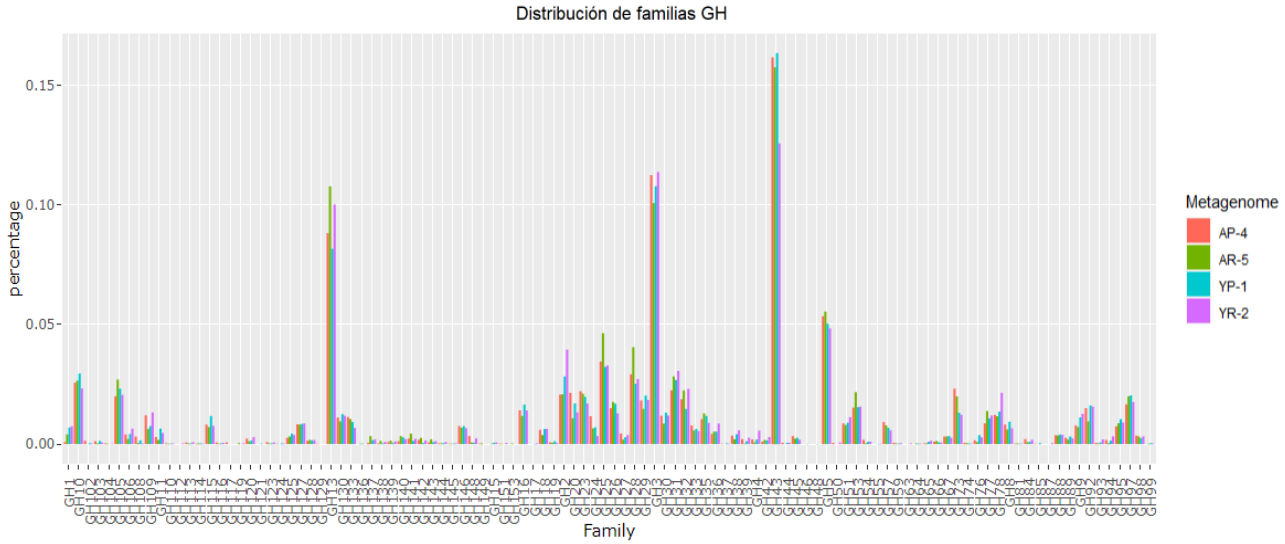

Figura 5: Distribución de familias GH en cada uno de los metagenomas.

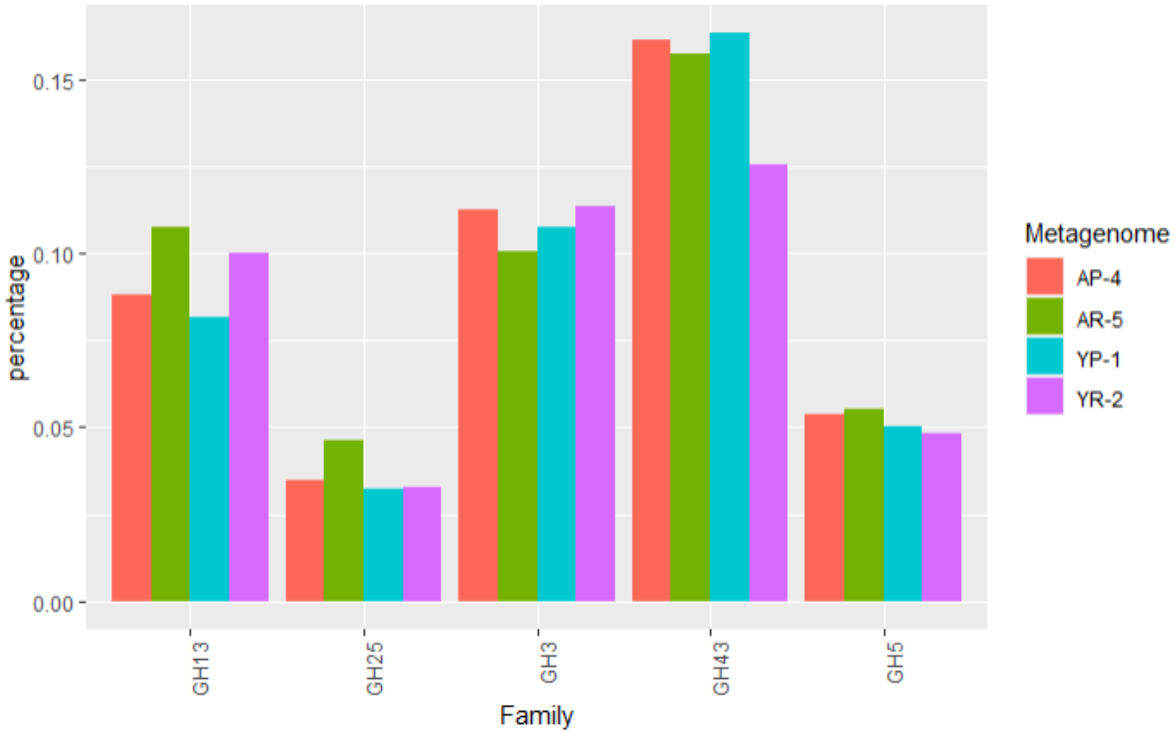

Figura 6: Familias de GH más abundantes en los metagenomas. Se observa la distribución de las familias GH3, GH13, GH43, GH5 y GH25.

Entre las polisacárido liasas (PL) las familias más abundantes halladas fueron PL1, 11, 9, 10 involucradas en actividades liasa contra pectina, pectato y ramnogalacturonano (

Figura 7).

Las carboxilesterasas (CE) de las familias 1,10,12 y 8 fueron halladas como las más abundantes, su actividad involucra acetil xilano esterasa, arilesterasa, pectina acetilesterasa y pectina metilesterasa, respectivamente (Figura 8).

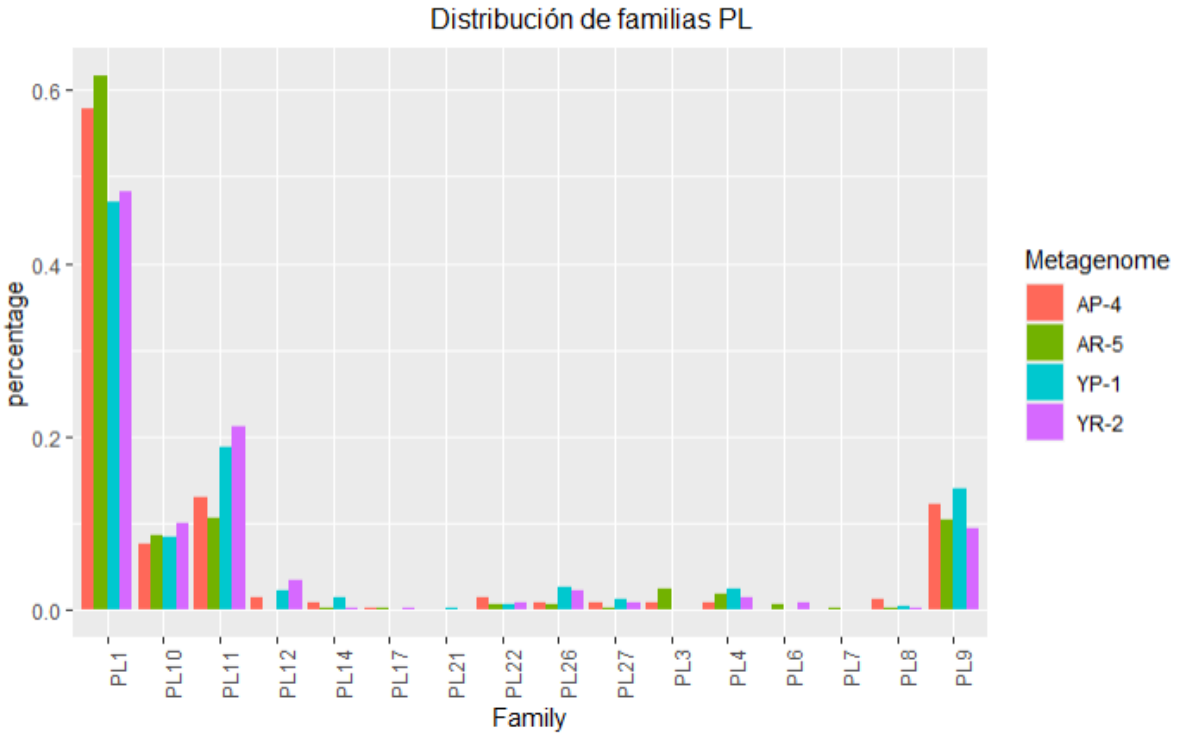

Figura 7: Distribución de familias PL en los metagenomas.

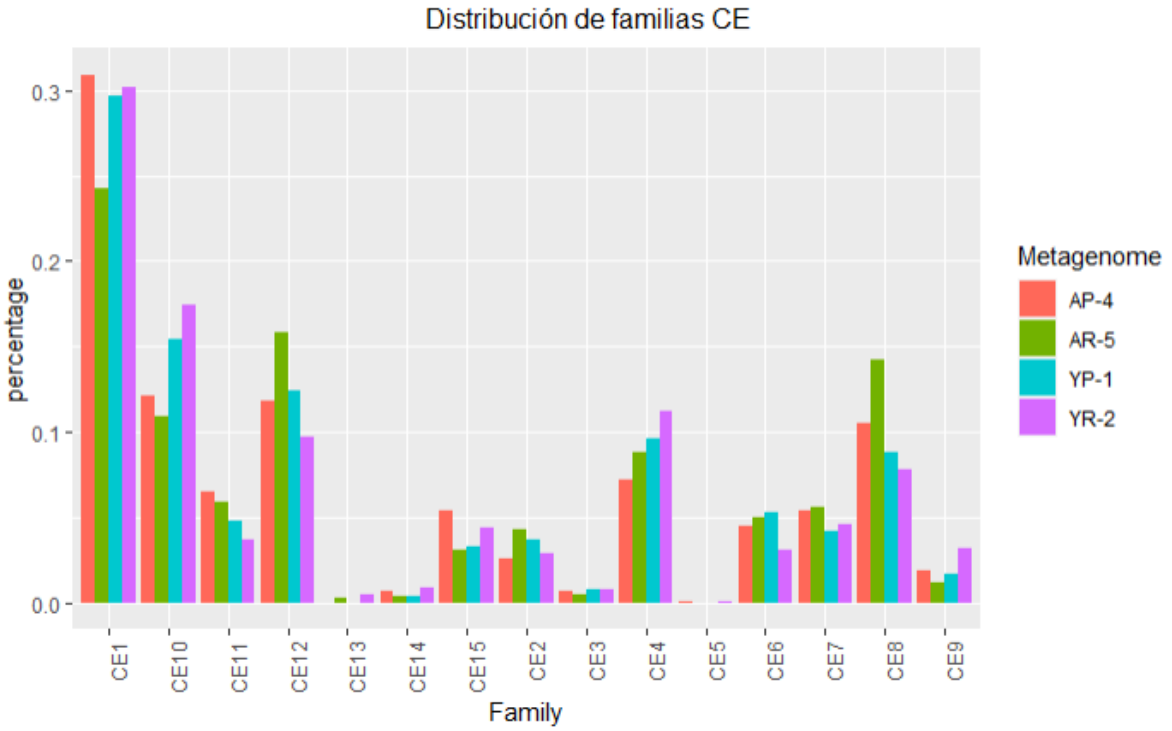

Figura 8: Distribución de familias CE en los metagenomas.

La capacidad de los rumiantes para descomponer los alimentos poco nutritivos en azúcares fermentables se atribuye por completo a las enzimas hidrolíticas de polisacáridos producidas por los microorganismos del rumen (Huws *et al.*, 2018). Las GH son el conjunto de enzimas más abundante y diversificado responsable de la ruptura de los enlaces glucosídicos en los polisacáridos vegetales, y representan el 50 % de las enzimas categorizadas en la base de datos CAZy (The CAZypedia Consortium, 2018). La mitad de las CAZymas detectadas en las muestras pertenecían a la clase GH, tal como lo reportado en la bibliografía (Wang *et al.* 2019; Li *et al.* 2022); siendo las actividades más abundantes encontradas la degradación de oligosacáridos (familias GH43, GH3 y GH13), desramificación (familia GH25), endo-hemicelulasas y celulasas (familia GH5).

Las muestras analizadas en este trabajo arrojaron un total de 112 familias diferentes de GH. Publicaciones previas informaron la presencia de 31 a 155 familias de GH, (Patel *et al.*, 2014; Pitta *et al.*, 2015; Jose *et al.*, 2017b) relacionando un mayor número de miembros de GH con un proceso más complejo de descomposición de la lignocelulosa (Bohra, Dafale and Purohit, 2019).

La diversidad en las CAZymas encontradas sugiere una hidrólisis más eficiente de la biomasa compleja. Una fracción más alta de enzimas degradadoras de oligosacáridos conduce a la producción de azúcares simples que dan como resultado la producción de ácidos grasos volátiles (AGV). Estos satisfacen aproximadamente el 80 % de las necesidades energéticas del animal, por lo que una alta producción de AGV es una característica deseada en la producción (Patel *et al.*, 2014).

En cuanto a la relación de la composición de CAZymas con la dieta y la edad de los animales, se utilizó el *software* estadístico STAMP para determinar si había una correlación entre estas variables. A nivel de clase o familia, no se encontraron diferencias estadísticas. Cuando se compararon muestras de vacas adultas entre sí, la familia GH13 fue más abundante en AR5, y cuando se compararon muestras jóvenes, GH13 resultó nuevamente más abundante en la muestra rica en alimento (YR2) y la familia GH43 fue más abundante en la muestra con alimentación pobre (YP1), lo que coincide con lo reportado por estudios en muestras de vacas alimentadas con bajo contenido de forraje respecto a un alto porcentaje del mismo (Wang *et al.* 2019).

Cuando se compararon muestras con alimentación pobre no se encontraron diferencias, pero cuando se compararon muestras con alimentación rica, GH43 y PL1 fueron más abundantes en la muestra adulta (AR5), mientras que GH2 fue más abundante en la muestra joven (YR2). A pesar de estas diferencias, la estadística tomando una sola muestra por grupo no es confiable, por lo que se debería aumentar el número de animales para poder confirmar los resultados.

4.6. Enzimas no ortólogas totales

A modo de aproximación y con el objetivo de detectar proteínas únicas en las muestras analizadas, se procedió a la detección de proteínas no ortólogas a las reportadas en la base de datos RumiRef. El procedimiento se llevó a cabo tal como se describe en la sección 3.10.1 mediante el *software* Metaphor. Debido a la importancia para el metabolismo del rumen y a la complejidad del análisis posterior, el estudio se centró en las CAZymas.

La comparación entre la abundancia de CAZymas totales y las no ortólogas respecto a la base de datos RumiRef se muestra en la Tabla 5.

Tabla 5: Abundancia de CAZymas totales y no ortólogas (NO) respecto a la base de datos RumiRef.

Familia	AP4		AR5		YP1		YR2	
	Total	NO	Total	NO	Total	NO	Total	NO
AA	0,03	**0,00**	0,03	**0,01**	0,06	**0,06**	0,05	**0,02**
CBM	4,32	**2,60**	4,01	**2,23**	5,08	**4,87**	4,92	**2,04**
CE	13,43	**4,28**	13,24	**3,37**	13,02	**11,00**	12,72	**3,05**
GH	52,56	**22,42**	54,30	**19,16**	53,97	**53,92**	55,60	**15,19**
GT	26,55	**10,30**	24,42	**8,05**	25,22	**25,10**	24,22	**8,42**
PL	3,12	**1,69**	4,01	**2,20**	2,66	**2,65**	2,49	**0,83**

Los resultados mostraron que los porcentajes de proteínas no ortólogas más abundantes descriptas a continuación fueron mayores en YP1 comparado con los demás metagenomas.

Respecto a los CBMs, la abundancia relativa promedio de estas se redujo un 64%. Puede decirse que hay gran proporción de CBM no ortólogos en los metagenomas estudiados. El más abundante fue el CBM6, de unión a celulosa, seguido por CBM67, de unión a L-rhamnosa, CBM32, de unión a galactosa, y CBM50, de unión a quitina.

La abundancia relativa promedio de las enzimas PL, por otro lado, se redujo un 60%, lo que mostraría que la mayoría de las proteínas PL son únicas en las muestras analizadas. De estas, las familias más abundantes fueron la PL1, la cual tiene función pectina liasa; y PL11 y PL9, las cuales tienen función rhamnogalacturonano endoliasa y pectato liasa respectivamente.

Las enzimas CE no ortólogas, por otra parte, representaron en promedio un 41% respecto del total de CE. La familia más abundante hallada fue CE1 con función acetil xilano esterasa, seguida por CE8 y CE12 con funciones pectina metilesterasa y arilesterasa respectivamente.

En cuanto a las GT, la abundancia relativa de enzimas pertenecientes a este grupo bajó un 52%, siendo las más abundantes GT2, GT4 y GT51.

Respecto a las enzimas GH, se obtuvo en promedio un 51% de GH no ortólogas respecto al total de GH. De estas, fueron dominantes las enzimas degradadoras de oligosacáridos, seguidas por las desramificadoras, endo- hemicelulasas, celulasas y xiloglucanasas. Las familias más abundantes fueron las GH43, GH3 y GH13, pertenecientes al grupo de enzimas degradadoras de oligosacáridos, seguidas por GH5 y GH25, las cuales están dentro de la categoría celulasas y desramificadoras, respectivamente.

4.7.Enzimas no ortólogas GH

Debido a la importancia de las enzimas GH en la digestión de rumiantes, y a la alta proporción de proteínas no ortólogas pertenecientes a este grupo en los cuatro metagenomas, se analizó la proporción de GH no ortólogas, respecto al total de GH halladas en cada muestra; tal como se describe en la sección 3.10.2 (Figura 9).

En promedio, el porcentaje relativo de estas proteínas fue del 0,5%; siendo dominantes las GH con función de degradación de oligosacáridos (0,3%). A estas les siguieron en abundancia las enzimas GH desramificadoras (0,09%), endo-hemicelulasas (0,05%), celulasas (0,03%) y xiloglucanasas (0,03%).

La familia con mayor abundancia fue la GH13, la cual se encontró en una proporción superior en los metagenomas AR5 e YR2 respecto de AP4 e YP1. El segundo y tercer lugar en abundancia lo ocuparon las familias con función de degradación de oligosacáridos GH43 y GH31 respectivamente, hallándose un porcentaje elevado de esta última en el metagenoma YR2.

Otras proteínas con abundancias altas correspondieron a las familias GH25 y GH5, con función desramificadora y celulasa respectivamente, las cuales mostraron porcentajes mayores en el metagenoma AP4 en comparación con los demás.

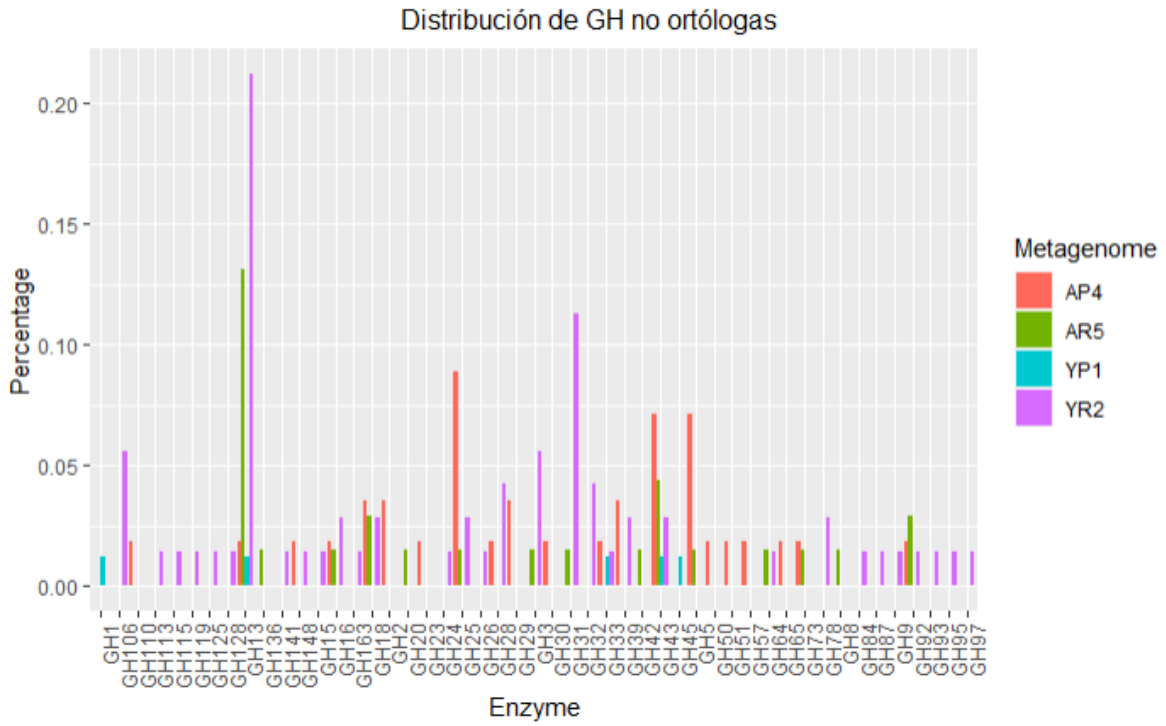

Figura 9: Distribución de familias GH no ortólogas en los metagenomas

Como puede verse, la naturaleza diversa y dinámica del microbioma del rumen lo convierte en un reservorio prometedor para la identificación de nuevas proteínas microbianas con una potente aplicación en el sector de la biotecnología, más notablemente aquellas involucradas en la hidrólisis de la biomasa lignocelulósica.

El análisis de las enzimas halladas en las muestras podría complementarse a futuro comparando las secuencias de estas proteínas contra la base de datos CAZy, con el objetivo de hallar enzimas no reportadas. Estas proteínas únicas estarían adaptadas a los requerimientos degradativos de la zona, y por lo tanto serían específicas de estos animales.

5. Conclusiones

En este trabajo ha sido posible ensamblar los cuatro metagenomas mediante el *software* Megahit, identificándose como mejor condición la utilización del parámetro --k-step 10.

Se lograron obtener siete genomas ensamblados a partir de metagenomas (MAGs) con porcentajes de completitud ≥ 80 % y contaminación $\leq 10\%$. Su asignación taxonómica indicó que mayoritariamente están relacionados a la degradación de fibras y polisacáridos.

En cuanto a la determinación del perfil metabólico de los metagenomas, el programa Genomaple permitió un análisis global de las vías presentes. Los módulos para la degradación de la mayoría de los carbohidratos fueron biológicamente factibles en todas las muestras, lo que demuestra la especialización de la microbiota en estas vías, así como también en la producción de AGV.

La diversidad en las CAZymas encontradas sugiere que los metagenomas son capaces de realizar una hidrólisis eficiente de la biomasa. En particular, la mitad de estas pertenecieron a la clase GH, hallándose un total de 112 familias diferentes. Las actividades más abundantes fueron la degradación de oligosacáridos (familias GH43, GH3 y GH13), desramificación (familia GH25), endo-hemicelulasas y celulasas (familia GH5).

Fue posible además la identificación de CAZymas no ortólogas a las presentes en la base de datos RumiRef, lo que sugeriría que estas proteínas son únicas en las muestras analizadas.

Para determinar si la composición de CAZymas tiene correlación con la dieta y edad de los rumiantes, se utilizó el *software* estadístico STAMP. Sólo se hallaron diferencias cuando se compararon muestras de a pares; al analizar vacas adultas, jóvenes y con dieta rica entre sí respectivamente. Si bien es necesario en un futuro incrementar el número de animales que permitan una estadística confiable, estos hallazgos son relevantes en cuanto a la identificación de proteínas adaptadas a diferentes condiciones nutricionales, de gran interés en el sector biotecnológico.

Anexo

Tabla 6: Parámetros de calidad de los bins totales correspondientes al metagenoma AP4.

Bin	Marcador de linaje	Genomas	Completitud (%)	Contaminación (%)	Heterogeneidad (%)	Cobertura promedio	Contig más largo (pb)	N50 (pb)
bin.1	k__Bacteria (UID2569)	434	83,57	2,08	66,67	7,46	33.478	10.360
bin.2	k__Bacteria (UID203)	5.449	41,38	0	0	26,56	69.445	38.169
bin.3	k__Bacteria (UID203)	5.449	42,76	0,86	100	22,89	95.119	33.827
bin.4	k__Bacteria (UID203)	5.449	15,52	0	0	5,63	10.499	3.284
bin.5	k__Bacteria (UID2569)	434	79,62	64,96	7,98	5,39	17.761	3.374
bin.6	root (UID1)	5.656	0	0	0	4,60	11.195	2.627
bin.7	k__Bacteria (UID203)	5.449	19,59	1,72	0	4,69	7.491	2.932
bin.8	root (UID1)	5.656	0	0	0	7,25	65.305	10.200
bin.9	k__Bacteria (UID2982)	88	68,93	3,78	10,53	6,61	17.795	5.152
bin.10	root (UID1)	5.656	0	0	0	30,67	161.546	70.789
bin.11	root (UID1)	5.656	0	0	0	10,61	82.258	38.547
bin.12	root (UID1)	5.656	0	0	0	12,89	51.677	11.877
bin.13	k__Bacteria (UID203)	5.449	0,86	0	0	24,32	54.674	17.169
bin.14	k__Bacteria (UID203)	5.449	55,82	12,07	5,88	12,88	39.602	6.418
bin.15	o__Bacteroidales (UID2657)	160	41,62	14,44	2,44	4,98	24.973	2.918
bin.16	k__Bacteria (UID203)	5.449	18,18	0	0	4,49	6.325	2.580
bin.17	k__Bacteria (UID2495)	2.993	87,71	0,1	0	14,44	121.058	44.077
bin.18	k__Bacteria (UID203)	5.449	45,52	5,52	0	5,93	13.836	3.040
bin.19	o__Selenomonadales (UID1024)	64	42,18	1,2	33,33	5,94	9.712	3.114
bin.20	k__Bacteria (UID203)	5.449	14,66	0	0	5,04	8.063	2.981
bin.21	root (UID1)	5.656	0	0	0	3,52	5.253	2.312
bin.22	root (UID1)	5.656	0	0	0	5,78	12.229	5.595
bin.23	p__Bacteroidetes (UID2605)	350	89,81	2,63	71,43	12,34	81.217	24.794

Tabla 7: Parámetros de calidad de los bins totales correspondientes al metagenoma AR5.

Bin	Marcador de linaje	Genomas	Completitud (%)	Contaminación (%)	Heterogeneidad (%)	Cobertura promedio	Contig más largo (pb)	N50 (pb)
bin.1	k__Bacteria (UID203)	5.449	18,34	0,16	0	5,84	7.633	2.926
bin.2	f__Lachnospiraceae (UID1255)	90	20,15	0	0	4,26	9.977	2.774
bin.3	root (UID1)	5.656	0	0	0	5,30	7.247	2.744
bin.4	g__Prevotella (UID2724)	55	65,02	10,93	17,65	6,10	11.115	3.166
bin.5	root (UID1)	5.656	0	0	0	6,85	25.786	11.820
bin.6	o__Bacteroidales (UID2716)	92	89,86	170,38	21,7	8,63	23.381	3.003
bin.7	k__Bacteria (UID203)	5.449	65,28	33,86	86,21	7,70	21.021	4.965
bin.8	k__Bacteria (UID203)	5.449	29,23	9,48	40	4,71	14.685	3.354
bin.9	k__Bacteria (UID203)	5.449	3,51	0	0	9,27	7.609	2.769
bin.10	k__Bacteria (UID203)	5.449	45,69	1,72	0	7,63	12.639	3.622
bin.11	o__Selenomonadales (UID1024)	64	48,17	1,26	71,43	7,59	15.100	3.229
bin.12	g__Prevotella (UID2722)	64	65,78	1,65	60	10,65	20.310	5.240

Tabla 8: Parámetros de calidad de los bins totales correspondientes al metagenoma YP1.

Bin	Marcador de linaje	Genomas	Completetitud (%)	Contaminación (%)	Heterogeneidad (%)	Cobertura promedio	Contig más largo (pb)	N50 (pb)
bin.1	f__Lachnospiraceae (UID1286)	57	93,67	0,5	0	13,40	556.226	195.201
bin.2	k__Bacteria (UID203)	5.449	11,7	0	0	4,39	6.648	2.773
bin.3	k__Bacteria (UID203)	5.449	45,69	0	0	5,05	21.934	3.731
bin.4	k__Bacteria (UID203)	5.449	2,59	0	0	15,22	89.178	22.459
bin.5	k__Bacteria (UID2329)	174	58,11	19,5	2,38	5,87	34.305	4.356
bin.6	o__Clostridiales (UID1120)	304	92,67	2,48	60	11,32	141.583	13.231
bin.7	root (UID1)	5.656	0	0	0	7,47	50.681	16.110
bin.8	k__Bacteria (UID203)	5.449	76,77	101,72	43,27	7,78	29.843	3.604
bin.9	k__Bacteria (UID2495)	2.993	78,85	0,43	100	9,21	48.483	16.862
bin.10	k__Bacteria (UID203)	5.449	69,04	75,34	53,42	11,44	54.794	5.125
bin.11	k__Bacteria (UID203)	5.449	30,14	0	0	4,21	7.544	2.608
bin.12	k__Bacteria (UID203)	5.449	29,23	0	0	4,55	7.192	2.621
bin.13	k__Bacteria (UID203)	5.449	46,72	1,72	100	15,04	39.545	7.489
bin.14	k__Bacteria (UID2569)	434	59,56	9,51	42,86	9,09	20.055	5.136
bin.15	root (UID1)	5.656	0	0	0	25,57	51.331	43.196
bin.16	root (UID1)	5.656	0	0	0	9,32	53.905	16.853
bin.17	k__Archaea (UID2)	207	3,98	0	0	17,68	80.037	75.527
bin.18	o__Bacteroidales (UID2657)	160	94,5	0,09	0	22,13	137.797	44.119
bin.19	root (UID1)	5.656	0	0	0	6,32	20.538	9.314
bin.20	root (UID1)	5.656	0	0	0	6,81	19.424	5.889
bin.21	k__Bacteria (UID203)	5.449	21,55	3,45	33,33	5,35	16.874	2.650
bin.22	k__Bacteria (UID203)	5.449	18,97	0	0	5,19	9.888	2.628
bin.23	o__Bacteroidales (UID2657)	160	81,07	3,34	71,43	14,86	103.819	16.402
bin.24	k__Bacteria (UID203)	5.449	0,16	0	0	14,57	154.893	95.374
bin.25	root (UID1)	5.656	0	0	0	14,11	292.046	292.046
bin.26	k__Bacteria (UID203)	5.449	18,53	1,72	0	3,96	8.154	2.442
bin.27	k__Bacteria (UID203)	5.449	22,49	0	0	13,87	21.206	6.636
bin.28	c__Gammaproteobacteria (UID4202)	67	91,49	0,54	50	6,13	21.776	7.363
bin.29	root (UID1)	5.656	4,17	0	0	5,47	34.633	2.989
bin.30	root (UID1)	5.656	0	0	0	13,40	50.784	6.972
bin.31	root (UID1)	5.656	0	0	0	10,60	152.381	152.381

Tabla 9: Parámetros de calidad de los bins totales correspondientes al metagenoma YR2.

Bin	Marcador de linaje	Genomas	Completetitud (%)	Contaminación (%)	Heterogeneidad (%)	Cobertura promedio	Contig más largo (pb)	N50 (pb)
bin.1	k__Bacteria (UID203)	5.449	5,64	0	0	6,16	38.406	4.180
bin.2	k__Bacteria (UID203)	5.449	29,66	0	0	5,05	12.950	3.594
bin.3	k__Bacteria (UID2569)	434	66,38	5,91	15,38	11,05	44.436	13.306
bin.4	k__Bacteria (UID203)	5.449	13,79	0	0	26,47	43.759	29.479
bin.5	k__Bacteria (UID203)	5.449	15,67	0	0	4,59	9.237	2.824
bin.6	root (UID1)	5.656	0	0	0	14,03	82.705	17.564
bin.7	k__Bacteria (UID2569)	434	82,26	0,07	75	15,40	135.585	38.521
bin.8	o__Bacteroidales (UID2657)	160	63,6	8,51	7,89	5,77	23.035	4.349
bin.9	o__Bacteroidales (UID2617)	213	81,75	4,85	14,29	6,64	43.647	8.324
bin.10	root (UID1)	5.656	0	0	0	9,00	111.202	81.461
bin.11	k__Bacteria (UID203)	5.449	30,5	0	0	4,62	12.818	3.253
bin.12	c__Clostridia (UID1118)	387	88,06	2,02	100	15,62	39.992	11.028
bin.13	k__Bacteria (UID203)	5.449	12,07	0	0	4,46	5.473	2.513
bin.14	c__Clostridia (UID1118)	387	70,76	14,59	5,71	5,81	15.508	3.250
bin.15	o__Bacteroidales (UID2657)	160	51,93	17,85	2,02	5,00	17.241	3.126
bin.16	k__Bacteria (UID2328)	3.167	51,88	4	0	5,79	17.699	3.666
bin.17	o__Bacteroidales (UID2716)	92	73,42	0,56	50	30,78	67.936	21.472
bin.18	k__Bacteria (UID203)	5.449	56,03	0	0	5,54	21.610	4.714
bin.19	k__Bacteria (UID203)	5.449	25,86	0	0	8,87	74.684	11.799
bin.20	k__Bacteria (UID203)	5.449	20,41	1,75	0	4,81	6.709	2.586
bin.21	root (UID1)	5.656	0	0	0	30,35	130.484	19.989
bin.22	root (UID1)	5.656	0	0	0	3,91	3.753	2.519
bin.23	root (UID1)	5.656	0	0	0	16,84	74.031	27.994
bin.24	root (UID1)	5.656	0	0	0	7,13	64.516	14.047
bin.25	f__Lachnospiraceae (UID1286)	57	63,48	0,86	0	5,71	20.072	4.739
bin.26	o__Bacteroidales (UID2716)	92	43,45	4,51	5,26	5,27	12.258	3.612
bin.27	k__Bacteria (UID203)	5.449	15,86	0	0	5,05	9.049	3.528
bin.28	k__Bacteria (UID1453)	901	62,11	11,11	0	6,06	116.281	3.940
bin.29	p__Bacteroidetes (UID2605)	350	94,84	0,95	50	18,18	240.016	108.343
bin.30	k__Archaea (UID2)	207	0,14	0,07	0	9,12	125.676	125.676
bin.31	k__Bacteria (UID203)	5.449	51,14	0	0	9,34	59.790	11.156
bin.32	k__Bacteria (UID203)	5.449	10,33	0	0	6,17	37.774	3.901
bin.33	root (UID1)	5.656	0	0	0	37,18	47.346	33.534
bin.34	root (UID1)	5.656	0	0	0	6,12	18.299	4.803
bin.35	k__Bacteria (UID203)	5.449	91,44	101,1	23,4	7,45	39.519	5.018
bin.36	k__Bacteria (UID203)	5.449	91,17	84,09	57,14	8,33	72.970	15.563
bin.37	k__Bacteria (UID203)	5.449	13,24	0	0	4,44	7.443	3.014
bin.38	k__Bacteria (UID2982)	88	47,8	6,37	0	4,72	8.724	2.818
bin.39	k__Bacteria (UID203)	5.449	3,45	0	0	30,12	79.722	19.246
bin.40	root (UID1)	5.656	4,17	0	0	4,50	5.243	2.630

6. Referencias

Albertsen, M. *et al.* (2013) 'Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes', *nature.com*, 31(6). pp 533-538. doi: 10.1038/nbt.2579.

Asma, Z. *et al.* (2013) 'Microbial ecology of the rumen evaluated by 454 GS FLX pyrosequencing is affected by starch and oil supplementation of diets', FEMS microbiology ecology, 83(2), pp 504-514. doi: 10.1111/1574-6941.12011

Bohra, V., Dafale, N. A. and Purohit, H. J. (2019) 'Understanding the alteration in rumen microbiome and CAZymes profile with diet and host through comparative metagenomic approach', *Archives of Microbiology*. Springer Verlag, 201(10), pp. 1385–1397. doi: 10.1007/S00203-019-01706-Z.

Bolger, A. *et al.* (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp 2114-2120. doi:10.1093/bioinformatics/btu170

Dick, G. J. *et al.* (2009) 'Community-wide analysis of microbial genome sequence signatures', *Genome Biology*, 10(8). doi: 10.1186/GB-2009-10-8-R85.

Drula, E. *et al.* (2022) 'The carbohydrate-active enzyme database: functions and literature', N*ucleic acids research*, 50(D1), D571-D577. doi: 10.1093/nar/gkab1045

Ejigu, G. F. and Jung, J. (2020) 'Review on the computational genome annotation of sequences obtained by next-generation sequencing', *Biology*, 9(9), pp. 1–27. doi: 10.3390/biology9090295.

Genomaple-2.3.2 - Genome Metabolic And Physiological potentiaL Evaluator. Available at: https://maple.jamstec.go.jp/maple/maple-2.3.1/index.html.

Goris, J *et al.* (2007) 'DNA–DNA hybridization values and their relationship to whole-genome sequence similarities', *microbiologyresearch.org*, 2(1), p. 194. doi: 10.1099/ijs.0.64483-0.

Gurevich, A. *et al.* (2013) 'QUAST: Quality assessment tool for genome assemblies', *Bioinformatics*, 29(8), pp. 1072–1075. doi: 10.1093/bioinformatics/btt086.

Hackmann, T. J. and Spain, J.N (2010), 'Invited review: ruminant ecology and evolution: perspectives useful to ruminant livestock research and production', *Journal of dairy science*, 93(4), 1320-1334. doi: 10.3168/jds.2009-2071

Haroon MF, Hu S, Shi Y, Imelfort M, Keller J, Hugenholtz P, Yuan Z, T. and GW (2013) 'Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage', *nature.com*. doi: 10.1038/nature12375

Hernández, R. *et al.* (2022). 'Functional and phylogenetic characterization of bacteria in bovine rumen using fractionation of ruminal fluid'. *Frontiers in Microbiology*, 13, 813002. doi: 10.3389/fmicb.2022.813002

Huws, S. A. *et al.* (2018) 'Addressing global ruminant agricultural challenges through understanding the rumen microbiome: Past, present, and future', *Frontiers in Microbiology*. Frontiers Media S.A., 9(SEP). doi: 10.3389/FMICB.2018.02161/FULL.

Hyatt, D. *et al.* (2010) 'Prodigal: Prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics*, 11. doi: 10.1186/1471-2105-11-119.

Jami, E. *et al.* (2013) 'Exploring the bovine rumen bacterial community from birth to adulthood', *ISME Journal*, 7(6), 1069–1079.

Jose, V. L. *et al.* (2017 a) 'In depth analysis of rumen microbial and carbohydrate-active enzymes profile in Indian crossbred cattle'. *Syst Appl Microbiol*. ;40: 160–170. doi:10.1016/j.syapm.2017.02.003.

Jose, V. L. *et al.* (2017 b) 'Metagenomic insights into the rumen microbial fibrolytic enzymes in Indian crossbred cattle fed finger millet straw', *AMB Express*. Springer Verlag, 7(1). doi: 10.1186/S13568-016-0310-0.

Kang, D. D. *et al.* (2019) 'MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies', *PeerJ*, 2019(7), pp. 1–13. doi: 10.7717/peerj.7359.

Karlin, S., Mrázek, J. and Campbell, A. M. (1997) 'Compositional biases of bacterial genomes and evolutionary implications', *Journal of Bacteriology*. American Society for Microbiology, 179(12), pp. 3899–3913. doi: 10.1128/JB.179.12.3899-3913.1997.

Konstantinidis, K. T. and Tiedje, J. M. (2005) 'Towards a genome-based taxonomy for prokaryotes', *Journal of Bacteriology*, 187(18), pp. 6258–6264. doi: 10.1128/JB.187.18.6258-6264.2005.

Kurtz, S. *et al.* (2004) 'Versatile and open software for comparing large genomes.', *Genome biology*, 5(2). doi: 10.1186/GB-2004-5-2-R12.

Langille, M. *et al.* (2013) 'Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences', *Nature biotechnology*, 31(9), pp 814-821.

Li, D. *et al.* (2016) 'MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices', *Methods*. Elsevier Inc., 102(2016), pp. 3–11. doi: 10.1016/j.ymeth.2016.02.020.

Li, L. *et al.* (2022). 'An Age Effect of Rumen Microbiome in Dairy Buffaloes Revealed by Metagenomics'. *Microorganisms*, 10(8), 1491.doi: 10.3390/microorganisms10081491

Meyer, F. *et al.* (2008) 'The metagenomics RAST server - A public resource for the automatic phylogenetic and functional analysis of metagenomes', *BMC Bioinformatics*, 9. doi: 10.1186/1471-2105-9-386.

Morgavi, D. P. *et al.* (2013) 'Rumen microbial (meta)genomics and its application to ruminant production', *Animal*, 7(SUPPL.1), pp. 184–201. doi: 10.1017/S1751731112000419.

Parks, D. H et al. (2014) 'STAMP: statistical analysis of taxonomic and functional profiles'. *Bioinformatics*, 30: 3123–3124. doi:10.1093/bioinformatics/btu494

Parks, D. H. *et al.* (2015) 'CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome Research*, 25(7), pp. 1043–1055. doi: 10.1101/gr.186072.114.

Patel, D. D. *et al.* (2014) 'Microbial and Carbohydrate Active Enzyme profile of buffalo rumen metagenome and their alteration in response to variation in the diet', *Gene*. Elsevier B.V., 545(1), pp. 88–94. doi: 10.1016/j.gene.2014.05.003.

Pitta, D. *et al.* (2015) 'Metagenomic assessment of the functional potential of the rumen microbiome in Holstein dairy cows', *Elsevier*. doi: 10.1016/j.anaerobe.2015.12.003.

Quince, C. *et al.* (2017) 'Shotgun metagenomics, from sampling to analysis', *Nature Biotechnology*, 35(9), pp. 833–844. doi: 10.1038/nbt.3935.

Rahman, M. M. *et al.* (2007) 'Molecular cloning and characterization of all RND-type efflux transporters in Vibrio cholerae non-O1', *Microbiology and Immunology*. Center for Academic

Publications Japan, 51(11), pp. 1061–1070. doi: 10.1111/j.1348-0421.2007.tb04001.x.

Rho, M., Tang, H. and Ye, Y. (2010) 'FragGeneScan: predicting genes in short and error-prone reads', *Nucleic Acids Research*. Oxford Academic, 38(20), pp. e191–e191. doi: 10.1093/NAR/GKQ747.

Ribeca, P., and Valiente, G. (2011). 'Computational challenges of sequence classification in microbiomic data', *Briefings in bioinformatics*, 12(6), pp 614-625. doi: 10.1093/bib/bbr019

Rinke, C. *et al.* (2013) 'Insights into the phylogeny and coding potential of microbial dark matter', *Nature*, 499(7459), pp 431-437.

Rodriguez-R, L. M. *et al.* (2018) 'The Microbial Genomes Atlas (MiGA) webserver: Taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level', *Nucleic Acids Research*, 46(W1), pp. W282–W288. doi: 10.1093/nar/gky467.

Scholz, M. B., Lo, C. C. and Chain, P. S. G. (2012) 'Next generation sequencing and bioinformatic bottlenecks: The current state of metagenomic data analysis', *Current Opinion in Biotechnology*, 23(1), pp. 9–15. doi: 10.1016/J.COPBIO.2011.11.013.

Sekiguchi, Y. *et al.* (2015) 'First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking', *PeerJ*, 3, e740.

Sharon, I. and Banfield, J. F. (2013) 'Genomes from metagenomics', *Science*, 342(6162), pp. 1057–1058. doi: 10.1126/SCIENCE.1247023.

Sharon, I., Morowitz, M. and Thomas, B. (2014) 'Time series community genomics analysis reveals rapid shifts in', *researchgate.net*. doi: 10.1101/gr.142315.112.

Sharpton, T. J. (2014) 'An introduction to the analysis of shotgun metagenomic data', *Frontiers in Plant Science*. Frontiers Research Foundation, 5(JUN), p. 209. doi: 10.3389/FPLS.2014.00209/BIBTEX.

Soo RM, Skennerton CT, Sekiguchi Y, Imelfort M, Paech SJ, D. P. and Steen JA, Parks DH, Tyson GW, H. P. (2014) 'An expanded genomic representation of the phylum Cyanobacteria', *academic.oup.com*, (6), pp. 1031–1045.

Stewart, R. D. *et al.* (2018a) 'Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen', *Nature Communications 2018 9:1*. Nature Publishing Group, 9(1), pp. 1–11. doi: 10.1038/s41467-018-03317-6.

Stewart, R. D. *et al.* (2018b) 'Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen', *Nature Communications*. Nature Publishing Group, 9(1), p. 870. doi: 10.1038/s41467-018-03317-6.

Stewart, R. D. *et al.* (2019) 'Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery', *Nature Biotechnology*. Springer Science and Business Media LLC, 37(8), pp. 953–961. doi: 10.1038/s41587-019-0202-3.

Suzek, B. *et al.* (2015) 'UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches', *Bioinformatics*, 31(6), pp 926-932.

Swan, B. *et al.* (2013) 'Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean', *National Acad Sciences*, 110. doi: 10.1073/pnas.1304246110.

The CAZypedia Consortium (2018) 'Ten years of CAZypedia: a living encyclopedia of carbohydrate-active enzymes', *Glycobiology*, 28(1), pp. 3–8. doi: 10.1093/glycob/cwx089.

Thomas, T., Gilbert, J. and Meyer, F. (2014) ‘Metagenomics: A guide from sampling to data analysis’, *The Role of Bioinformatics in Agriculture*, (Figure 1), pp. 357–383. doi: 10.1201/b16568.

Ufarté, L., Potocki-Veronese, G., Cecchini, D., Tauzin, A. S., Rizzo, A., Morgavi, D. P., ... & Laville, E. (2018). 'Highly promiscuous oxidases discovered in the bovine rumen microbiome'. *Frontiers in Microbiology*, 9, 861.doi: 10.3389/fmicb.2018.00861

Unterseher, M. *et al.* (2011) ‘Species abundance distributions and richness estimations in fungal metagenomics – lessons learned from community ecology’, *Molecular Ecology*. John Wiley & Sons, Ltd, 20(2), pp. 275–285. doi: 10.1111/J.1365-294X.2010.04948.X.

Van der Veen, B. E. *et al.* (2014) ‘Metaphor: Finding Bi-directional Best Hit homology relationships in (meta)genomic datasets’, *Genomics*. Academic Press Inc., 104(6), pp. 459–463. doi: 10.1016/j.ygeno.2014.10.008.

Wickham, H., Chang, W. and Wickham, M.H. (2016) ‘Package “ggplot2”’, Create elegant data visualisations using the grammar of graphics. Version, 2(1), pp 1-189.

Wang, L., *et al*. (2019). ‘Metagenomic analyses of microbial and carbohydrate-active enzymes in the rumen of holstein cows fed different forage-to-concentrate ratios’. *Frontiers in microbiology*, 10, 649. doi:

Wong, D. W., Chan, V. J., & Liao, H. (2019). Metagenomic discovery of feruloyl esterases from rumen microflora. *Applied microbiology and biotechnology*, 103, 8449-8457. doi: 10.1007/s00253-019-10102-y

Wright, A. D. G. and Klieve, A. V. (2011) ‘Does the complexity of the rumen microbial ecology preclude methane mitigation?’, *Animal Feed Science and Technology*. Elsevier, 166–167, pp. 248–253. doi: 10.1016/J.ANIFEEDSCI.2011.04.015.

Wrighton, K. C. *et al.* (2012) ‘Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla’, *Science*. American Association for the Advancement of Science, 337(6102), pp. 1661–1665. doi: 10.1126/SCIENCE.1224041.

Xue, M., Sun, H., Wu, X., Guan, L.L. and Liu, J. (2018) 'Assessment of rumen microbiota from a large dairy cattle cohort reveals the pan and core bacteriomes contributing to varied phenotypes, applied and environmental microbiology', *Applied and Environmental Microbiology*, 84(19), e00970-18.

Yin, Y. *et al.* (2012) ‘DbCAN: A web resource for automated carbohydrate-active enzyme annotation’, *Nucleic Acids Research*, 40(W1), pp. 445–451. doi: 10.1093/nar/gks479.

Zhang, H. *et al.* (2018) ‘dbCAN2: a meta server for automated carbohydrate-active enzyme annotation.’, *Nucleic acids research*, 46(W1), pp. W95–W101. doi: 10.1093/nar/gky418.