Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2025:60
Faculty of Veterinary Medicine and Animal Science

Seasonal variability of the rumen microbiome in indigenous African cattle: a bioinformatics approach

From pasture to pathways

Renaud Van Damme

Seasonal variability of the rumen microbiome in indigenous African cattle: a bioinformatics approach

From pasture to pathways

Renaud Van Damme
Faculty of Veterinary Medicine and Animal Science
Department of Animal Biosciences
Uppsala

SLU
SWEDISH UNIVERSITY
OF AGRICULTURAL
SCIENCES

DOCTORAL THESIS
Uppsala 2025

Acta Universitatis Agriculturae Sueciae
2025:60

Cover: An image generated from DALL-E 3 based on a photo taken by Renaud Van Damme, 2025, CC BY-SA

ISSN 1652-6880
ISBN (print version) 978-91-8124-044-3
ISBN (electronic version) 978-91-8124-090-0
https://doi.org/10.54612/a.77nq392jvl

Swedish University of Agricultural Sciences, Department of Animal Biosciences, Uppsala, Sweden

Print: SLU Grafisk service, Uppsala 2025

Seasonal variability of the rumen microbiome in indigenous African cattle: a bioinformatics approach.

Abstract

The Ethiopian climate causes cattle to frequently experience cycles of abundant feed availability during the rainy season and severe feed scarcity during the dry season. Using a bioinformatics-driven metagenomics approach, this study reveals how the rainy season favours fibre-degrading taxa and expanded methanogenic and biosynthetic potential. In contrast, the dry season is characterised by reduced metabolic diversity and an increase in opportunistic taxa. To achieve this, novel computational tools were developed and refined, including MUFFIN and PANKEGG. The results collected from extreme drought conditions provide valuable information to ensure that ongoing work to reduce methane emissions through breeding or feed additives does not negatively impact the ability of ruminants to adapt to drought or low-quality feed, which may threaten food security and animal welfare during heatwaves. The study revealed seasonal variations in microbial community structure and metabolic functions. During the rainy season, the microbiome exhibited an increase in acetoclastic methanogens and fibre-degraders. The dry season saw a rise in hydrogenotrophic methanogens. The latter increase reflects a decrease in acetate production, suggesting a decline in access to feed, as well as less efficient fibre breakdown. These findings bring critical insights into the metabolic adaptability and resilience of indigenous cattle microbiomes. The findings also suggest new potential targets for enhancing feed efficiency and promoting environmental sustainability. The microbiome harboured multiple antibiotic resistance genes (ARGs) throughout both seasons. The presence of ARGs exposes the risk of potential resistance spread to pathogens as well as broader ecological implications. In addition to the metagenomic study, a genomic study of Ethiopian cattle breeds highlighted significant genetic diversity and potential adaptive mechanisms to local environmental stressors. Those genetic markers open the possibility for future integration of host-genomic and microbiome analyses.

Keywords: Bioinformatics, Genomics, Metagenomics, Rumen, Cattle, Ethiopia, Seasonal changes

Säsongsvariation i våmmikrobiomet hos inhemska afrikanska nötkreatur: en bioinformatisk studie.

Sammanfattning

Skiftet från den årliga regnperioden till torrperiod innebär att boskap i Etiopien upplever cykler av riklig fodertillgång under regnperioden och allvarlig fodersbrist under torrperioden. Denna studie visar med hjälp av metagenomik hur regnperioden gynnar fiber-nedbrytande mikrober samt ökar den metanogena och biosyntetiska kapaciteten. Torrperioden kännetecknas däremot av minskad metabolisk mångfald och en ökning av opportunistiska arter. För att möjliggöra denna analys utvecklades och förfinades nya bioinformatiska verktyg. MUFFIN är ett analysflöde optimerat för hybridsekvenseringsdata, och PANKEGG, är ett interaktivt visualiseringsverktyg för studera metagenomiska data. Resultaten från extrema torkförhållanden ger viktig kunskap för att säkerställa att pågående arbete med att minska metanutsläpp via avel eller fodertillskott inte försämrar idisslares förmåga att anpassa sig till torka eller lågkvalitativt foder, något som annars kan hota både livsmedelssäkerhet och djurvälfärd under värmeböljor. Under regnperioden ökade förekomsten av acetoklastiska metanogener (t.ex. *CADBMS01*) och fiber-nedbrytare (t.ex. *Fibrobacter*). Torrperioden visade en ökning av hydrogenotrofiska metanogener (t.ex. *Methanosphaera*). Denna ökning återspeglar en minskning i acetatproduktion, vilket tyder på sämre tillgång till föda och mindre effektiv fiberomsättning. Dessa resultat ger viktiga insikter i våmmikrobiomets anpassningsförmåga. Fynden pekar även på nya möjliga mål för att förbättra fodereffektiviteten och främja hållbarhet. Mikrobiomet innehöll flera antibiotikaresistensgener (ARG) under båda säsongerna. Förekomsten av dessa gener innebär en risk för spridning av resistens till patogener samt vidare ekologiska effekter. Utöver metagenomstudien visade en genomisk kartläggning av etiopiska boskapsraser på betydande genetisk mångfald och möjliga anpassningsmekanismer till lokala miljöstressorer. Dessa genetiska markörer öppnar möjligheter för framtida integrering av värdgenomik och mikrobiomanalys.

Keywords: Bioinformatik, Genomik, Metagenomik, Våm, boskap, Etiopien, Säsongsvariationer

Variabilité saisonnière du microbiome du rumen chez les vaches africaines indigènes: une approche bioinformatique.

Résumé

L'Éthiopie possède plus de 70 millions de bovins, principalement des races indigènes. Le climat éthiopien fait subir aux bovins des cycles fréquents d'abondance alimentaire en saison des pluies et de pénurie sévère en saison sèche. Cette thèse explore les variations saisonnières du microbiome du rumen des bovins Boran éthiopiens grâce à une approche métagénomique pilotée par la bioinformatique. Pour cela, de nouveaux outils informatiques ont été développés, notamment MUFFIN, un pipeline optimisé pour les données de séquençage hybrides, et PANKEGG, un outil interactif pour visualiser les résultats de MUFFIN. L'étude révèle des variations saisonnières dans la structure des communautés microbiennes et leurs fonctions métaboliques. Durant la saison des pluies, le microbiome présente une augmentation des méthanogènes acétoclastes et des organismes dégradant les fibres. La saison sèche montre une augmentation des méthanogènes hydrogénotrophes. Cette hausse reflète une baisse de la production d'acétate, indiquant une diminution de l'accès à l'alimentation et une dégradation moins efficace des fibres. Ces résultats apportent des connaissances essentielles sur l'adaptabilité métabolique et la résilience des microbiomes des bovins indigènes. Ils suggèrent également de nouvelles cibles potentielles pour améliorer l'efficacité alimentaire et promouvoir une durabilité environnementale. Le microbiome arborait plusieurs gènes de résistance aux antibiotiques (ARG) durant les deux saisons. La présence d'ARG indique un risque potentiel de propagation de résistances vers les pathogènes ainsi que des conséquences écologiques plus larges. En complément de l'étude métagénomique, une étude génomique des races bovines éthiopiennes a révélé une diversité génétique significative et de potentiels mécanismes d'adaptation aux facteurs de stress environnementaux locaux. Ces marqueurs génétiques ouvrent la voie à une future intégration des analyses du génome hôte et du microbiome.

Keywords: Bioinformatique, Génomique, Métagénomique, Rumen, Bovin, Ethiopie, Changement saisonnier

Dedication

To all the farmers working tirelessly to feed us all.

A’m biesse pa , un limonadier qui a assouvi ma soif de connaissance, qu’a stî r’joint les ôtes.

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I. Renaud Van Damme*, Martin Hölzer, Adrian Viehweger, Bettina Müller, Erik Bongcam-Rudloff, Christian Brandt (2021). Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN). PLOS Computational Biology, 17(2): e1008716. https://doi.org/10.1371/journal.pcbi.1008716

II. Renaud Van Damme*, Arnaud Vanbelle, Juliette Hayer, Amrei Binzer-Panchal, Erik Bongcam-Rudloff (2025). PANKEGG: Integrative Visualisation and Comparison of Metagenome-Assembled Genomes Annotation, Taxonomy, and Quality. (submitted)

III. WondossenAyalew, WuXiaoyun, Getinet Mekuriaw Tarekegn*, Rakan Naboulsi, Tesfaye Sisay Tessema, Renaud Van Damme, Erik Bongcam-Rudlof, MinChu, Chunnian Liang, Zewdu Edea, Solomon Enquahone, Yan Ping* (2024). Whole genome sequences of 70 indigenous Ethiopian cattle. Nature Sci Data 11, 584. https://doi.org/10.1038/s41597-024-03342-9

IV. Renaud Van Damme*, Leila Nasirzadeh, Wondossen Ayalew, Tomas Klingström, Getinet Mekuriaw Tarekegn, Erik Bongcam-Rudloff (2025). Seasonal Dynamics of the Rumen Microbiota in Ethiopian Boran Cattle: a shotgun metagenomics study. (Manuscript)

 * Corresponding author

The contribution of Renaud Van Damme to the papers included in this thesis was as follows:

I. Renaud Van Damme developed the pipeline and wrote the paper.

II. Renaud Van Damme developed the software and wrote the paper.

III. Renaud Van Damme provided support for the bioinformatics analysis and was involved in the writing process.

IV. Renaud Van Damme planned the study, analysed the data and wrote the paper.

List of other publications

I. Michelle Wille*, Jonas Johansson Wensman, Simon Larsson, Renaud van Damme, Anna-Karin Theelke, Juliette Hayer, Maja Malmberg* (2020). Evolutionary genetics of canine respiratory coronavirus and recent introduction into Swedish dogs. Infection, Genetics and Evolution, Volume 82, 104290, https://doi.org/10.1016/j.meegid.2020.104290

II. Ayalew, Wondossen, Xiaoyun Wu, Getinet Mekuriaw Tarekegn, Tesfaye Sisay Tessema, Rakan Naboulsi, Renaud Van Damme, Erik Bongcam-Rudloff, Zewdu Edea, Solomon Enquahone, and Ping Yan* (2023). Whole-Genome Resequencing Reveals Selection Signatures of Abigar Cattle for Local Adaptation. Animals, 13, no. 20: 3269. https://doi.org/10.3390/ani13203269A

III. Andreas Gisel, Livia Stavolone, Temitayo Olagunju, Michael Landi, Renaud Van Damme, Adnan Niazi, Laurent Falquet, Trushar Shah, Erik Bongcam-Rudloff* (2023). EpiCass and CassavaNet4Dev Advanced Bioinformatics Workshop. EMBnet.journal, [S.l.], v. 29, p. E1045, https://doi.org/10.14806/ej.29.0.1045

IV. Ayalew, Wondossen, Xiaoyun Wu, Getinet Mekuriaw Tarekegn*, Tesfaye Sisay Tessema, Rakan Naboulsi, Renaud Van Damme, Erik Bongcam-Rudloff, Zewdu Edea, Min Chu, Solomon Enquahone, Chunnian Liang, and Ping Yan* (2024). Whole Genome Scan Uncovers Candidate Genes Related to Milk Production Traits in Barka Cattle. International Journal of Molecular Sciences 25, no. 11: 6142. https://doi.org/10.3390/ijms25116142

V. Wondossen Ayalew, Wu Xiaoyun*, Getinet Mekuriaw Tarekegn*, Tesfaye Sisay Tessema, Min Chu, Chunnian Liang, Rakan Naboulsi, Renaud Van Damme, Erik Bongcam-Rudloff, Yan Ping* (2024). Whole-genome sequencing of copy number variation analysis in Ethiopian cattle reveals adaptations to diverse environments. *BMC Genomics* 25, 1088. https://doi.org/10.1186/s12864-024-10936-5

VI. Elin Hermann, Renaud Van Damme, Erik Bongcam-Rudloff, Leila Nasirzadeh* (2024). Urban Pigeons as Reservoirs of Critical Pathogens: Improved protocol for sequencing pigeon faeces in disease monitoring. EMBnet.journal, [S.l.], v. 30, p. E1059, https://doi.org/10.14806/ej.30.0.1059

* Corresponding author

List of tables

List of figures

Abbreviations

ARG	Antibiotic resistance gene
ASV	Amplicon Sequence Variation
bp	base pair
BQSR	Base Quality Scores Recalibration
CAMP	Cationic antimicrobial peptide
CAZyme	Carbohydrate-Active Enzyme
CLR	Centre-Log Ratio
CNVR	copy number variation region
EBV	Estimated Breeding Values
EC	Enzyme Commission
FAIR	Findable Accessible Interoperable Reusable
GDP	Gross Domestic Product
GEBV	Genomically Estimated Breeding Value
GI	Gastrointestinal
GTDB	Genome Taxonomy DataBase
HPC	High Performance Computing
HTS	High-throughput Sequencing
kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Ortholog
LLM	Large Language Model
MAG	Metagenome-assembled genome
Meta-GEBV	Metagenomic/Genomic Estimated Breeding Value

MIMAG	Minimum Information about Metagenome-Assembled Genome
ML	Machine Learning
MQ	Mapping Quality
NGS	Next Generation Sequencing
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCA	Principal component analysis
PCR	Polymerase Chain Reaction
QD	Quality by Depth
qPCR	Quantitative Polymerase Chain Reaction
RFI	Residual Feed Intake
rRNA	ribosomal RNA
RUG	Rumen Uncultured genomes
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
VCF	Variant Call Format
VFA	Volatile Fatty Acid

1. Introduction

Ethiopia is home to one of Africa's largest cattle populations, with over 60 million head as of 2020, contributing more than 1 million tonnes of beef and more than 3.8 billion litres of milk annually, and accounting for roughly 45% of the agricultural GDP(Wakaso et al., 2025; Zewde et al., 2022). Indigenous breeds dominate the Ethiopian cattle sector, accounting for over 98.5% of the population in 2015/2016 (Sendeku et al., 2016). These animals are raised primarily under smallholder and pastoralist systems, often in challenging and variable environments. The reliance on local breeds is both a matter of necessity and adaptation: commercial high-yielding breeds usually cannot withstand the harsh climate, seasonal feed variability, and long walking distances required in pastoral systems.

However, adaptation is not immunity. In 2022, one of the worst droughts in recent history occurred, resulting in the death of over four million livestock and severely impacting rural livelihoods and food availability (Ethiopia - Situation Report, 10 Jan 2024 | OCHA, 2024). Among the 28 recognised Ethiopian cattle breeds, the Ethiopian Boran is particularly well-regarded for its heat tolerance, drought resistance, and ability to cover long distances. It is raised primarily in southern and eastern regions, which are also the most drought-prone. The breed is dual-purpose (meat and milk) and provides an opportunity to study the environmental resilience in livestock.

While the "core" rumen microbiome in cattle is relatively stable under controlled feeding and environmental conditions, evidence suggests that seasonality and environmental stress can induce marked changes in microbial diversity and functionality. In comparison, studies in temperate zones have found limited seasonal shifts in rumen composition under stable feed conditions (Noel et al., 2017). At the same time, other studies observed significant reductions in microbial diversity and *Ruminococcus* abundance during hot summers (Islam et al., 2021). These shifts are known to impair fibre digestion and reduce volatile fatty acid (VFA) production, which are critical for energy metabolism in ruminants. Research conducted in tropical Australia on grazing cattle further showed that microbial communities respond dynamically to feed quality and season, diverging significantly from those observed under higher-quality, supplemented diets (Martinez-Fernandez et al., 2020).

Despite these findings, Western commercial breeds are the primary focus of most metagenomic studies on the rumen microbiome, often conducted in controlled environments. Grazing systems, particularly in low- and middle-income countries, remain underexplored. This lack of knowledge is especially true for indigenous African cattle. We designed this PhD project to investigate the seasonal variability of the rumen microbiome in such animals, with an initial focus on the Ethiopian Boran. Additionally, we sought to investigate the potential interactions between the host genome and the microbiota. Although these host-microbe interactions are not the primary focus of this thesis, we have established the methodological groundwork to investigate them.

Understanding these microbial-host-environment interactions is not just an academic pursuit; It has a direct potential for improving cattle resilience in the face of climate change. Identifying microbiome features that support efficient digestion under dry and hot conditions could provide traits to inform future breeding programs. Enhanced microbiome profiles may serve as indirect selection criteria for heat and drought resistance, contributing to more sustainable livestock systems. This importance has recently been highlighted in the Global Methane Genetic Initiative, where the Bezos Earth Fund and the Global Methane Hub are investing $ 27.4 million to breed ruminants with lower methane emissions. Methane emissions from over 100,000 animals will be measured in the project, and rumen samples to study microbial variation will be collected from approximately 15,000 of these animals. Our approach aligns with several United Nations Sustainable Development Goals (SDGs) (*THE 17 GOALS | Sustainable Development*, n.d.): Goal 1 (No Poverty) and Goal 12 (Responsible Consumption and Production). Promoting livestock systems that are better suited to their local environments and more productive under environmental stress also intersects with Goal 2 (Zero Hunger), as more resilient animals can contribute to food security in vulnerable regions. Moreover, this work supports the One Health framework by addressing food safety and antimicrobial resistance through a better understanding of host-microbiome interactions.

Given the complexity of this challenge, the development of scalable, transparent, robust, FAIR-compliant and reproducible analytical workflows was a key part of this PhD. Accordingly, the thesis includes the creation of two bioinformatics tools:

- Paper I MUFFIN, a flexible pipeline for hybrid metagenomic assembly, binning, and functional analysis, and
- Paper II PANKEGG, a visual integration tool that enables comparison of metagenome-assembled genomes (MAGs) across samples and studies.

These tools were developed not only to support the analyses in this thesis but also to be reusable by others working on similar challenges. Note, they are not limited to rumen microbiome data; they are compatible with any short- and long-read metagenomic datasets. The ability to reconstruct MAGs and functionally annotate them gives researchers insight not only into "who is there," but also into the metabolic roles of microbial community members, particularly the degradation of plant fibres, synthesis of short-chain fatty acids (such as VFAs), and the presence of antibiotic resistance genes.

This thesis is organised to reflect the multidisciplinary nature of the work. It begins with a background chapter that provides readers with the necessary understanding of metagenomics, bioinformatics, rumen physiology, and the context of African cattle production. The methods chapter then details the experimental procedures and computational workflows used in Papers III and IV. After that, two chapters present the bioinformatics tools developed as part of this work (Papers I and II). The application chapters (Papers III and IV) demonstrate how these tools and methods were used to investigate the rumen microbiome and genome of Ethiopian cattle. The thesis concludes with a general discussion that integrates findings across chapters and outlines the future directions of this research, particularly the extension of these methods to new data from South African cattle and eventually Swedish livestock, with a focus on methane mitigation and sustainable agriculture.

2. Background

2.1 Metagenomics: Concepts and Methods

Genomics is the study of the genome of a single organism that we sequence and analyse. Metagenomics, on the other hand, is the study of a group of organisms that cannot be dissociated, which we sequence collectively and then attempt to discriminate and analyse individually (Riesenfeld et al., 2004).

Metagenomics as a field emerged from the challenge of studying unculturable microorganisms. A common claim suggests that over 99% of microorganisms cannot be cultivated in vitro (Amann et al., 1995). This claim is often quoted without nuance, yet it highlights a crucial truth: cultivation biases our understanding of microbial diversity (E. J. Stewart, 2012). When considering all microbial environments, it's indeed the case that fewer than 1% of microorganisms have been successfully cultured (Lloyd et al., 2018). However, much of microbiology focuses on anthropocentric environments, such as the human gut, wastewater, and livestock systems, where cultivation techniques have improved over time (Lagier et al., 2015, 2018; Lau et al., 2016). Although we have made progress in cultivating microbes from these environments, a significant fraction remains unexplored or unculturable (Lewis et al., 2021).

Thus, metagenomics remains essential, not only to study what we cannot culture, but also because we often don't know in advance what is present in a sample, how many different organisms there are, or their relative abundances (Nayfach et al., 2021). In this sense, metagenomics echoes Zeno's Dichotomy Paradox: our goal is to describe a community comprehensively, yet we only make incremental progress with little certainty of reaching the complete truth. Metagenomics studies began with a method that does not involve the genomes of the organisms and instead uses a selected gene. Yet it is often categorised as a metagenomics method: the metabarcoding sequencing (16S rRNA, 18S rRNA, ITS sequencing). The knowledge was then extended through whole metagenome shotgun sequencing and further deepened using cumulative discoveries (Handelsman, 2004).

Technological advances have played a significant role in this progress. The advent of next-generation sequencing (NGS, e.g. Illumina) enabled

deep, accurate short-read sequencing. Then, platforms like Oxford Nanopore and PacBio introduced long-read capabilities, with steadily improving accuracy (Almeida et al., 2019; Logsdon et al., 2020). The long-read technology, combined with the continued improvement of Illumina, led to the emergence of a new generation of sequencing, known as High-Throughput Sequencing (HTS). These developments led to the development of shotgun sequencing and ultimately enabled the reconstruction of metagenome-assembled genomes (MAGs), allowing for genome-resolved analyses of microbial communities (Tully et al., 2018).

Despite access to massive sequencing power, no method captures everything. DNA extraction biases, sequencing limitations, lab contamination (Salter et al., 2014) and uneven organism abundance all mean that some organisms may be missed, including those present at very low abundance or at a distance of two centimetres to the right of where the sample was collected.

In this PhD project, we apply both short-read (Illumina) and long-read (Oxford Nanopore) sequencing to maximise coverage and accuracy. Our goal is to recover high-quality MAGs, which provide taxonomic and functional insights. Functional annotation of MAGs allows us to explore microbial metabolism, interactions within the community, and interactions with the host environment, such as changes in feed composition and seasonality. We also studied the presence of antibiotic resistance genes.

2.2 Bioinformatics for Metagenomics

Bioinformatics is the process of transforming raw biological data into meaningful biological knowledge through computational analysis (Bayat, 2002; Marturano, 2012). This thesis primarily focuses on sequencing data, encompassing both genomic and metagenomic information.

Two primary analysis workflows were followed:

For the host's genomic studies: Reads are quality-checked and cleaned. High-quality reads are mapped to a reference genome. From the alignments, we extract single-nucleotide polymorphisms (SNPs) that differentiate our sample from the reference.

For the metagenomic studies of the rumen microbiome: Reads from each sample are first quality-controlled and cleaned. Initial taxonomic classification is conducted (e.g. via Kraken2) (Wood et al., 2019). Reads are

then assembled into contiguous sequences (contigs). These contigs are grouped into bins based on coverage and sequence characteristics. High-quality bins are retained as MAGs. These are then taxonomically classified and functionally annotated (*DRAM for Distilling Microbial Metabolism to Automate the Curation of Microbiome Function | Nucleic Acids Research | Oxford Academic*, n.d.).

While the general process may seem linear, numerous analytical and biological biases complicate interpretation. The choice of tools, parameter settings, reference databases, and sequencing platform all influence the results (Quince et al., 2017). That's why many new tools and pipelines continue to emerge (Scholz et al., 2016). To address specific needs and improve reproducibility, we developed a pipeline (MUFFIN) and a visualisation tool (PANKEGG) tailored to our workflow.

MAGs are central to this approach. They represent draft genomes reconstructed from metagenomic data. They often vary in completeness and contamination. A standard for assessing MAG quality comes from both the Genome Taxonomy Database (GTDB) and the MIMAG consortium, which specifies criteria for medium quality MAGs as follows (Bowers et al., 2017; *GTDB - Genome Taxonomy Database*, n.d.):

- CheckM completeness > 50%
- CheckM contamination < 10%
- Quality score (completeness - 5 × contamination) > 50
- 40% of marker genes present (bac120 or arc53)
- <1000 contigs, N50 > 5 kb
- <100,000 ambiguous bases

The criteria use quality information provided by previously by CheckM (Parks et al., 2015) and now by CheckM2 (Chklovski et al., 2023). CheckM first version was a quality assessment of genome bin using single copy gene marker set. The presence of those marker set was determining the completeness and contamination. CheckM2 is an assessment of genome bin quality using universally trained machine learning models. The default mode is using three different models, the first is a general gradient boost model to estimate the completeness of organisms not well represented in Genbank (Clark et al., 2016) and RefSeq (Goldfarb et al., 2025). The second is a specific neural network model, more accurate when predicting the completeness of bins closely related to known organisms that have been used

to train the model. The third model is a model specific for the contamination estimation.

As of 2024, the GTDB includes over 730,000 genomes, up from ~145,000 in 2019. This increase reflects the rapid accumulation of MAGs and the growing power of comparative metagenomics (Parks et al., 2017).

2.3 Cow Anatomy, Rumen Physiology, and Microbiome

Cattle are ruminants, meaning they possess a multi-compartment stomach specialised for fermenting fibrous plant material (see Figure 1). The rumen is the first and largest chamber, acting as a fermentation vat where ingested feed is mixed with a diverse microbiota (Perez et al., 2024). These microbes initiate the digestion process by breaking down plant polymers such as cellulose and hemicellulose (Russell et al., 2009).

The cow chews its feed and swallows it into the rumen, where microbial fermentation begins. It then regurgitates partially digested feed (cud) to rechew it, a process that increases the surface area and facilitates microbial action (Church, 1988). Fermentation in the rumen results in the production of volatile fatty acids (VFAs), primarily acetate, propionate, and butyrate, which are absorbed through the rumen wall and serve as the cow's primary energy source (Moharrery et al., 2014; Ungerfeld, 2020).

Rumen microbes also play a significant role in protein nutrition. Many of them grow and reproduce in the rumen; when they die, their biomass flows into the lower gastrointestinal tract, where microbial proteins are digested and absorbed by the cow (*Application of Biotechnology to Nutrition of Animals in Developing Countries*, n.d.). This process supports the production of meat and milk (Matthews et al., 2019; Owens et al., 1986).

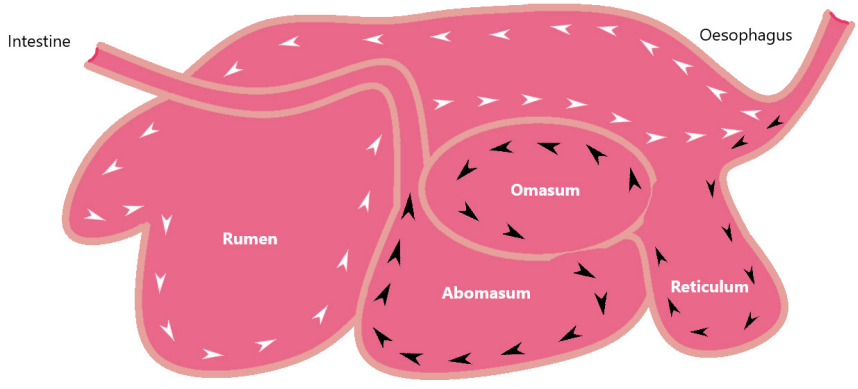

Figure 1: Schema of the Cattle stomach with the four compartments, Rumen, Reticulum, Omasum and Abomasum. Source Wikimedia (Millardcrystal, 2021)

The rumen is followed by the reticulum, which works in close coordination with the rumen to break down further and filter digesta. From there, the feed enters the omasum, a chamber characterised by ample folds that absorb water and nutrients (Church, 1988).

Finally, the digesta reaches the abomasum, the true acidic stomach, which kills the remaining microorganisms and further chemically degrades the feed (Church, 1988).

Then the digesta reaches the small intestine, where enzymatic digestion resumes and absorbs the last remaining nutrients (mostly remaining digested starches and amino acids) (Harmon & Swanson, 2020).

Microbial digestion is especially critical in ruminants; whose diets consist primarily of fibrous material that they cannot digest enzymatically. If one removes the microorganisms from the rumen, cattle would not be able to extract sufficient energy or protein from their feed to sustain themselves (Hook et al., 2010; Weimer, 2015).

2.4 Role of the Microbiome in Cattle Health and Productivity

The rumen microbiome is integral to the cow's overall health and productivity. An imbalanced microbial community, resulting from infection, stress, feed changes, heat stress, or antibiotic use, can lead to reduced feed

efficiency, nutrient malabsorption, and disease (Lopes et al., 2021; Monteiro et al., 2022).

One primary concern is methanogenesis, the production of methane (CH_4) by rumen archaea, particularly methanogens such as *Methanobrevibacter*. These microbes use different bacterial byproducts to reduce CO_2 into methane and produce ATP. While this process helps to some degree maintain the fermentation balance by removing excess components, it also contributes to greenhouse gas emissions (Greening et al., 2019; Hook et al., 2010). Three different pathways can occur.

The first is hydrogenotrophic methanogenesis, which is represented by the following orders of methanogens: Methanobacteriales, Methanococcales, Methanomicrobiales, Methanopyrales, and Methanosarcinales. This process reduces CO_2 into Methane using H_2.

The second is acetoclastic methanogenesis, which converts the acetate present in the environment into methane. The central clade to rely on this process is the Methanosarcinales.

The third is methylotrophic methanogenesis, which uses methanol and methylamines as substrates. This last pathway is the least common and only some Methanosarcinales and at least one member of the Methanomicrobiales are known to use it.

Methanogenesis also involves multiple coenzymes (B, F420, and M); without them, methanogenesis would not produce methane.

Another key role of the rumen microbiome is the degradation of plant fibre. Fibrolytic bacteria, such as *Ruminococcus*, *Fibrobacter*, and *Butyrivibrio*, degrade plant-fibres into fermentable sugars that are then metabolised into VFA(Comtet-Marre et al., 2017). A reduction in these populations, such as under heat stress or poor feed quality, can negatively affect digestion and energy production (Park et al., 2022).

The microbiota also facilitates carbohydrate fermentation, producing VFAs that supply up to 70% of the cow's energy requirements (Bergman, 1990; Ungerfeld, 2020). These VFAs are absorbed and converted into glucose or used directly in cellular metabolism. Disruption in these pathways (e.g., due to dietary shifts or disease) can lead to inefficiencies or metabolic disorders (Ungerfeld, 2020).

Thus, the rumen microbiome is both a productivity engine and a sustainability concern. To improve feed efficiency, reduce methane emissions, and develop more resilient livestock systems, it is essential to

understand the composition and function the rumen microbiome (Badhan et al., 2025).

2.5 Ethiopia's system and adaptation to the environment

2.5.1 Ethiopian cattle systems

Ethiopia's livestock sector is immense and predominantly rural, with around 70 million cattle, the largest cattle population in Africa (*Ethiopia - Situation Report, 10 Jan 2024 | OCHA*, 2024). Over 98% of these cattle are indigenous Bos indicus (zebu) breeds, which are kept by traditional smallholders and pastoralists (Y. Li et al., 2023). The country harbours a rich diversity of indigenous cattle breeds (28 recognised breeds), reflecting its varied agro-ecologies and long history as a gateway of cattle domestication into Africa (MEKURIAW & KEBEDE, 2015). These native breeds are well adapted to local conditions; for example, the Sheko (a rare taurine breed in southwest Ethiopia) is trypanotolerant (resistant to tsetse-borne disease) and noted for efficient milk production, while the Boran (a zebu breed of the southern rangelands) is renowned for drought hardiness (MEKURIAW & KEBEDE, 2015).

Ethiopian cattle production operates in distinct systems shaped by environment and culture. Smallholder farmers raise the vast majority (~78%) of cattle in mixed crop–livestock systems of the highlands (Y. Li et al., 2023). These smallholders typically keep a few zebu cattle for multiple purposes: draft power for ploughing, milk for home consumption or sale, manure for fertiliser, and as a form of savings (Y. Li et al., 2023). In contrast, pastoral systems in the arid and semi-arid lowlands (e.g., Somali, Afar, and Oromia regions) manage approximately 19% of the national herd in extensive, mobile herds (Y. Li et al., 2023). Pastoralists rely on communal rangelands and herd mobility, raising larger numbers of indigenous cattle (along with camels and goats) under open grazing. However, pastoral livelihoods are highly vulnerable to climate stress. Recurrent droughts and shifting rainfall patterns have led to water and pasture shortages, resulting in herd die-offs and increased heat stress in animals (Manyike et al., 2025). For example, in southern Ethiopia's Borana rangelands, the increasing frequency of drought has shrunk forage resources, making it difficult for large cattle like the Boran to maintain condition; herders have been forced to destock or

favour smaller livestock (goats, sheep) that cope better with prolonged dry conditions (Y. Li et al., 2023). Climate pressures, as well as disease and limited veterinary services, contribute to the low overall productivity of the livestock sector (Manyike et al., 2025).

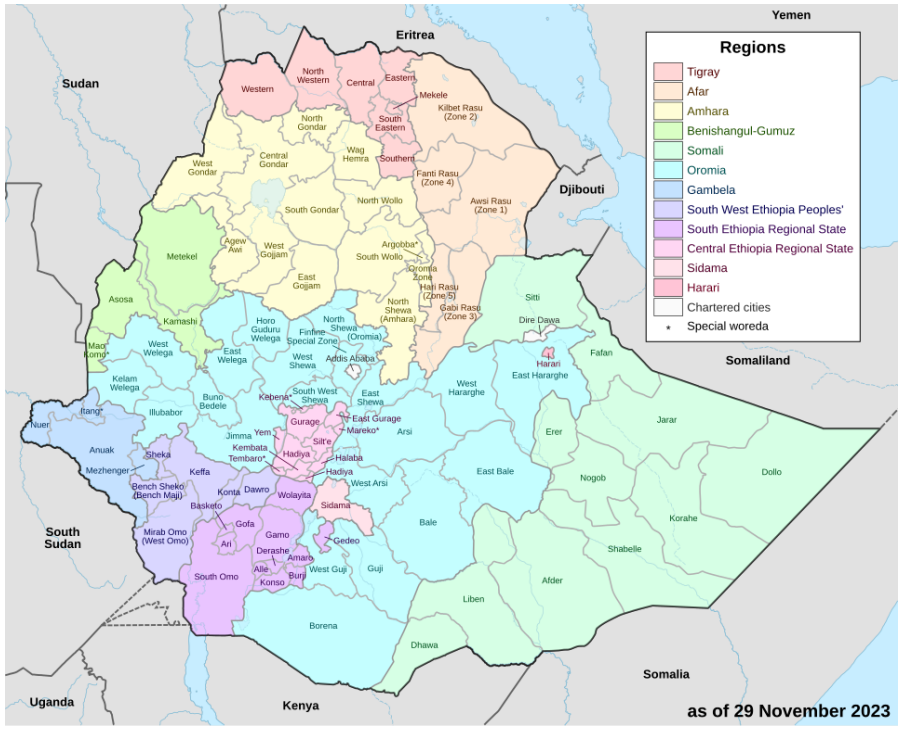

Figure 2: Cartography of Ethiopia with the region and zones. Source Wikimedia (User:SUM1, 2017)

2.5.2 Seasonal variation and adaptation strategies

Seasonal variability in Ethiopia's climate significantly impacts the livestock systems, with pronounced wet and dry seasons that create dramatic swings in feed and water availability. Ethiopia exhibits a diverse climatic pattern, but generally experiences three main seasons: the long rainy season ('Kiremt') from June to August, the dry season ('Bega') from October to February, and the short rainy season ('Belg') from March to May. In the southern lowland pastoral areas such as Borena, these patterns vary slightly, with the primary rainy season typically occurring from March to May and a short rainy period in October and November. But as the long dry season

progresses, pasture growth ceases and the quality of forage declines rapidly, leading to extended periods of feed scarcity (Duguma & Janssens, 2021). Cattle commonly experience poor body condition in late dry seasons, with markedly reduced productivity and higher susceptibility to diseases and parasites (Duguma & Janssens, 2021). In Janssens et al.'s study, all farmers reported inadequate feed during the dry season (versus ample feed during the wet season) and observed their animals losing weight and fertility during the drought months (Duguma & Janssens, 2021). Over the years, Ethiopian livestock keepers have developed a suite of adaptation strategies, both traditional and innovative, to help their animals cope with these environmental extremes.

Mobility and Grazing Rotation: Pastoralists practice strategic mobility, migrating with their herds to track seasonal water and pasture availability. They often rotate grazing areas (e.g. the traditional *seri* system in the Somali Region) to prevent the overuse of any one rangeland and allow vegetation to recover (Kebede et al., 2024; Manyike et al., 2025). This mobility exploits the patchy nature of rainfall, ensuring animals can find forage even in dry periods by moving to better-watered zones.

Water Harvesting: Communities employ water conservation methods, such as digging and maintaining *birkas* (traditional underground cisterns or ponds) to store rainwater for use during the dry season (Kebede et al., 2024). These stored water sources serve as critical lifelines for cattle when rivers and surface ponds dry up. Herders will also trek long distances to permanent water points during droughts, carefully rationing water to keep core breeding animals alive.

Feed Conservation and Supplementation: In highland mixed farming systems, farmers mitigate dry-season fodder gaps by conserving crop residues (such as straw from teff and maize stalks) and hay from the wet season (Duguma & Janssens, 2021). After harvest, cereal stover and other residues are collected and stored as feed for livestock. Many also purchase additional roughage (like straw or grass) if available. In times of extreme scarcity, livestock are supplemented with “non-conventional” feeds, such as leaves from indigenous browse trees, shrub foliage, and crop by-products, to augment the poor grazing (Duguma & Janssens, 2021; Manyike et al., 2025). These practices help bridge the nutritional gap until the next rains.

Herd Size Management: Herders commonly adjust their herd size and composition in response to climate forecasts and forage availability. During

severe droughts, destocking is used as a coping strategy; families sell off or slaughter weaker animals to reduce grazing pressure and generate income or food (Duguma & Janssens, 2021). Pastoralists may also split herds and send portions of the cattle to distant relatives or wetter areas, as a risk-spreading mechanism (Manyike et al., 2025). Additionally, many keep a mix of species (cattle, goats, camels) so that if pasture conditions favour one type (e.g. camels can browse drought-tolerant shrubs), the entire livelihood is not lost.

Genetic and Reproductive Strategies: Over generations, Ethiopian herders have leveraged indigenous knowledge to select hardy breeds and manage breeding timing to suit the climate. Pastoral communities favour cattle with traits such as heat tolerance, drought resistance, and disease resilience, for instance, the Masai and Boran zebu strains, which can withstand sporadic access to water (Manyike et al., 2025). Herders also control mating seasons: it is a common practice to ensure cows become pregnant so that calving coincides with the rainy season, when pasture and water are sufficient for lactating mothers and newborn calves (Duguma & Janssens, 2021). By avoiding births in the peak of the dry season, they reduce calf mortality and stress on the cows.

Ethiopia's farmers demonstrate a resilient interplay between environment, genetics, and management. Indigenous cattle breeds and traditional practices have evolved to allow communities to endure cycles of abundance and hardship. However, with climate change intensifying drought frequency and feed shortages, sustaining these systems will require enhancing adaptation strategies, from improving feed storage and water infrastructure to conserving genetic diversity, so that Ethiopia's cattle sector can continue to support livelihoods and food security in the coming decades (Duguma & Janssens, 2021).

2.6 Cattle metagenomics: From 2000 to now

2.6.1 First metagenomics studies in livestock

Early explorations of the rumen microbiome (2000s) relied on 16S rRNA gene clone libraries and low-throughput sequencing. These studies revealed that rumen bacterial communities are dominated by a few major phyla, chiefly Firmicutes and Bacteroidetes, with Proteobacteria and others in lower

abundance (L. Wang et al., 2019). For example, a 2005 study constructed a metagenomic DNA library from a dairy cow's rumen to screen for novel fibre-degrading enzymes, uncovering previously unknown hydrolases from the rumen microbiome (Ferrer et al., 2005). Around the same time, clone library surveys in cattle and yak rumen identified numerous novel 16S sequences, highlighting the vast diversity of rumen bacteria and archaea even in these early efforts (An et al., 2005). However, these pioneering studies were limited in scope; they often focused on a few Holstein dairy cows under controlled diets, reflecting primarily European/North American cattle with limited breed diversity. As a result, early rumen microbiome profiles lacked representation of the broader genetic and dietary diversity found in global cattle populations (Conteville et al., 2024). The emphasis then was on cataloguing "who's there" in the rumen using Sanger sequencing or 454 pyrosequencing of marker genes. These approaches established baseline knowledge (e.g., confirming a core set of prevalent rumen genera, such as *Prevotella* and *Ruminococcus*). Still, they provided little functional insight and were unable to capture less abundant or unculturable microbes.

The next step was the application of the newly developed shotgun metagenomics to the rumen around 2009–2011. In a landmark study, Brulc *et al.* (2009) performed gene-centric metagenome sequencing of fibre-adherent rumen microbes, uncovering numerous novel glycoside hydrolase genes linked to plant fibre breakdown (Brulc et al., 2009). Subsequently, Hess *et al.* (2011) performed deep sequencing of the rumen contents of a switchgrass-fed cow, assembling partial genomes of previously uncultured bacteria and identifying tens of thousands of carbohydrate-active enzymes (Hess et al., 2011). This *Science* 2011 study demonstrated the rumen's rich reservoir of biomass-degrading enzymes and marked one of the first instances of assembling draft genomes (metagenome-assembled genomes, MAGs) from rumen microbes. Together, these early metagenomic works showed that the rumen harbours vast functional diversity (e.g., novel cellulases, xylanases) and hinted at specific microbial lineages (such as *Prevotella* spp. or *Fibrobacter succinogenes*) as key fibre degraders. Still, they were typically small-scale (involving single or a few animals). They predominantly examined high-producing dairy cattle, which constrained the generalizability of the findings to other breeds or management systems. The limitations in sequencing depth and computational methods enabled the

studies to detect a large number of genes. Yet the studies struggled to assemble complete genomes or link genes to specific microbial taxa.

2.6.2 Rise of MAG-based studies and comparative frameworks

High-throughput sequencing advances in the mid-2010s (Illumina HiSeq, etc.) enabled more comprehensive rumen metagenomic surveys and led to the emergence of genome-centric metagenomics in cattle. Researchers began assembling large numbers of metagenome-assembled genomes (MAGs) from rumen samples, which provided much higher taxonomic and functional resolution than earlier gene catalogues. A watershed project by Stewart et al. (2018) assembled 913 draft bacterial/archaeal genomes from rumen metagenomes of 43 Scottish cattle (R. D. Stewart et al., 2018). This study introduced the term "rumen uncultured genomes (RUGs)" for these MAGs and demonstrated that genome-resolved metagenomics could retrieve new rumen species never cultured before. An even larger effort soon followed it: Stewart et al. (2019) generated a compendium of 4,941 non-redundant MAGs from the rumen microbiomes of 283 cattle (R. D. Stewart et al., 2019). This extensive catalogue, often referred to as the Rumen Genome Catalog, was constructed by aggregating and dereplicating tens of thousands of bins, setting a new benchmark for rumen microbial genomics. Notably, incorporating these MAGs into reference databases significantly increased the read classification rate for rumen metagenomes from ~15% (with previous references) to over 50–70% (R. D. Stewart et al., 2019). In other words, more than half of the DNA sequences from rumen samples could now be assigned to a known genome or MAGs, whereas previously most sequences had no match (R. D. Stewart et al., 2019). This increase highlights how under-represented rumen microbes were in earlier databases and how genome-centric approaches have filled significant knowledge gaps.

Alongside MAG collection, researchers developed comparative frameworks and databases to organise the deluge of new data. For example, the Hungate1000 project took a complementary approach by culturing and sequencing 410 rumen microbial isolates (reference genomes) [9]. While cultured isolates are high-quality genomes, adding the Hungate1000 genomes improved rumen metagenomic read mapping by only ~10%, indicating that most rumen microbes remained uncultured, hence the value of MAGs (Seshadri et al., 2018). Large gene catalogues were also compiled. In 2020, Li et al. published a bovine rumen microbial gene catalogue

containing millions of unique genes from 77 metagenomes (J. Li et al., 2020), enabling cross-study comparisons of functional potential. These resources, MAG compendia, isolate genome libraries, and gene catalogues, formed a foundation for comparative analysis. Researchers can now ask questions like: How do rumen microbial genomes differ between dairy and beef cattle, or between different continents? Or how do diet or host genetics shape the presence/absence of specific microbial genes?

Another emerging framework was the creation of global reference databases and consortia. The Global Rumen Census (2015) surveyed rumen microbiota from cattle worldwide (via 16S sequencing) and identified a core set of abundant taxa shared across herds (*Global Rumen Census*, 2025; Henderson et al., 2015). Building on this, post-2018 MAG studies started to include diverse breeds and geographies. For instance, a 2020 study generated a catalogue of 1,200 MAGs from African Boran cattle rumen, expanding representation beyond the European breeds (Wilkinson et al., 2020). By 2022–2024, international efforts had assembled thousands of genomes from zebu cattle in Brazil (Conteville et al., 2024), as well as dairy buffalo, sheep, goats, and wild ruminants, underscoring a move toward pan-ruminant microbiome frameworks. These comparative resources enable scientists to benchmark new findings, such as verifying whether a MAG from one study is novel or already present in the database, or comparing enzyme repertoires between datasets. They also facilitated functional analyses; researchers can map metagenomic reads to the reference gene catalogue to quantify abundances of genes/pathways, or use the MAG genomes to study metabolic pathways (e.g., methanogenesis genes) across different animals. In summary, the late 2010s witnessed a shift in the rumen microbiome field from small-scale descriptive studies to large-scale, data-rich resources that enable systematic comparisons and meta-analyses. The emphasis moved toward cataloguing "who is there and what they can do" on a global scale, which was a necessary step before tackling more profound biological questions.

2.6.3 Metagenomics studies Until 2020

Before 2020, most cattle metagenomics studies fell into a few significant themes, centred on improving livestock production and understanding rumen function:

Diet and nutrition effects: A large body of work examined how different diets or feed additives alter the rumen microbiome. For example, high-grain

vs. high-forage diets were compared to see shifts in microbial populations and fermentation end-products. These studies often used 16S rDNA profiling or moderate-depth metagenomes. A consistent finding was that high-forage (fibre-rich) diets enrich fibre-degrading bacteria (such as *Fibrobacter* and *Ruminococcus*). In contrast, high-grain diets favour amylolytic and lactic acid-producing bacteria, which can sometimes lead to reduced diversity or the proliferation of microbes associated with acidosis (L. Wang et al., 2019). Feed additives, such as fats, oils, or plant secondary compounds, were tested for their ability to suppress methanogens or protozoa. For instance, adding lauric acid drastically reduced protozoal counts and shifted the bacterial community structure in cows (Hristov et al., 2012). These nutrition-oriented studies aimed to manipulate the microbiome to improve feed efficiency or animal health (e.g. preventing bloat or acidosis). Up to 2020, methods included amplicon sequencing for studying community shifts, as well as occasionally metagenomic or metatranscriptomic analyses to link diet to functional genes (e.g., fibre enzyme profiles). Wang et al. (2019) is an example, where metagenomes of cows on different forage-to-concentrate ratios revealed diet-dependent abundance of carbohydrate-active enzyme genes (CAZymes) involved in fibre breakdown (L. Wang et al., 2019). Such studies addressed questions like "Which microbes flourish on a high-starch feed?" or "Do fibre-rich diets increase cellulase gene abundance?", linking microbiome composition to feed utilisation.

Methane mitigation and host emissions: Ruminant enteric methane, produced by archaea in the rumen, has been a topic of significant concern due to its substantial climate impact; many pre-2020 studies have profiled rumen archaeal communities (mostly methanogens of the family Methanobacteriaceae) under various conditions. For example, comparisons of high-methane versus low-methane cattle revealed higher Methanobrevibacter abundance, which correlated with increased methane production (Wallace et al., 2015). Some other interventions (like certain dietary fats, 3-nitrooxypropanol additives, or even selecting low-methane sheep/cattle lines) were tested and their microbiomes analysed to see how methane production dropped. These works were typically smaller scale, using qPCR or 16S/ITS amplicons to quantify methanogen populations. By 2019, several studies had combined host genetics and microbiome analysis, finding, for instance, that some methane emission traits are moderately heritable and linked to the composition of the rumen microbiome (Difford et

al., 2018). Overall, by 2020, it was recognised that "methanogen abundance and community structure are key determinants of methane output", and that altering the microbiome (by diet or possibly breeding) could reduce emissions. The progress in the established knowledge set the stage for deeper metagenomic inquiries into why specific microbiomes produce less methane (e.g. differences in hydrogen-utilisation pathways)(Danielsson et al., 2017).

Feed efficiency and production traits: Another focus was the link between the rumen microbiome and feed conversion efficiency, milk yield, or growth rate. Earlier efforts often characterised microbiomes of cows with high vs. low feed efficiency (measured by residual feed intake, RFI) using 16S sequencing. Some consistent patterns emerged, e.g., more efficient animals tend to have lower overall diversity and a distinct bacterial community composition, although the results varied. Protein- and fat-yield differences in milk were also correlated to specific microbes or fermentation profiles in some studies. Methods included both amplicon surveys and metatranscriptomics; for instance, Shi et al. (2014) used metatranscriptomes to demonstrate in sheep that low-methane versus high-methane had different active microbial populations and fermentation pathways (Shi et al., 2014). By 2020, evidence was mounting that certain bacterial families (like Ruminococcaceae) and functions (like better fibre degradation) associate with superior feed efficiency or milk production (Monteiro et al., 2024). These studies aimed at identifying microbial biomarkers for feed efficiency that could be harnessed in selective breeding or dietary interventions.

Microbial biogeography and core microbiome: Several studies have surveyed differences in microbiomes across breeds, hosts, or geographic regions. For instance, Henderson et al. (2015) sampled cattle from Asia, Europe, and North America, finding a remarkably similar core microbiome dominated by Prevotella, Succinivibrionaceae, and Ruminococcaceae, among others, despite geographical differences (Henderson et al., 2015). Others compared dairy breeds (Holstein vs. Jersey) under identical diets to determine if breed genetics influence the microbiome; the results were mixed, with some differences in minor taxa but largely overlapping communities (Roehe et al., 2016). Comparisons of rumen vs. faecal communities, or different gut compartments (rumen, omasum, colon), also began to appear, showing that while the rumen is unique in hosting fibre degraders, some taxa carry through the GI tract. These broad surveys built an understanding of which microbes are ubiquitous (e.g. Prevotella is often

~20–50% of sequences) and which are variable. By establishing the "normal" rumen microbiota structure, they provided context to interpret results from diet or treatment studies.

Methodological studies: As next-generation sequencing gained momentum, some papers in the 2010s focused on the methodology for rumen metagenomics. For example, evaluating different DNA extraction protocols for this fibre-rich, tough sample, or benchmarking bioinformatic tools for assembly and binning on rumen datasets. CheckM (2015) became a standard for assessing MAG completeness/contamination (Parks et al., 2015), and pipelines for metagenomic assembly got democratised (Kieser et al., 2020; Tamames & Puente-Sánchez, 2019; Uritskiy et al., 2018).

Additionally, the first attempts at rumen viromics appeared (e.g., sequencing rumen bacteriophages and pathogens) to catalogue viral diversity in the rumen (Anderson et al., 2017). Though not as numerous as diet studies, these methodological papers were crucial for enabling the larger studies that followed.

In summary, up to 2020, cattle rumen microbiome research had transitioned from simple descriptive surveys to more hypothesis-driven studies targeting productivity and environmental outcomes. They employed increasingly complex methods, from 16S rRNA profiling to shotgun metagenomics and metatranscriptomics, but sample sizes were still modest in most cases (dozens of animals at most). The typical study might ask, "How does X intervention change the rumen community or gene abundance, and what does that imply for fermentation or animal performance?" What changed near 2020 was the ability to assemble genomes and integrate data across studies, but the full exploitation of those advances was just beginning.

2.6.4 Metagenomics studies since 2020

Since 2020, rumen metagenomics research in cattle has grown in scale and adopted new perspectives and technologies, leading to notable shifts in the types of studies, questions, and methods:

Far larger and more diverse cohorts: Post-2020 studies often include hundreds or even thousands of samples, whereas pre-2020 works had tens. There has been a conscious effort to cover greater genetic and geographic diversity in cattle, for example, recent projects in 2021–2023 surveyed cattle from multiple continents (Europe, Asia, Africa, Americas) in one study, or

included breeds beyond Holsteins, such as Zebu (Bos indicus) cattle, buffaloes, and indigenous breeds, to capture how microbiomes differ with host genetics. The 2024 study by Conteville et al. assembled genomes from 52 Nelore (Brazilian Zebu) cattle, specifically to address the prior underrepresentation of tropical breeds in rumen metagenomics (Conteville et al., 2024). Similarly, Scientific Data (2025) by Legrand et al. expanded the sampling to Australian Angus beef cattle, examining not only the rumen but also the oral and nasal microbiomes (Legrand et al., 2025). These broader surveys revealed that while core microbial phyla remain consistent (Firmicutes, Bacteroidota still dominate in bovines), there are significant differences in community composition and MAGs linked to breed, diet, and environment when you look across diverse herds. In short, microbiome research has become less focused on high production breeds and farming system , acknowledging that farms throughout the world have themselves a diversity of breeds, feed and farm system. The questions now include comparative analyses: e.g., "Do tropical cattle harbour unique microbial species or gene patterns that differ from temperate cattle?" or "How does the rumen microbiome of high-methane vs. low-methane cattle compare across different breeds?" This trend enhances the robustness of microbiome-derived solutions (such as probiotics or dietary recommendations) by ensuring they're effective across various cattle genetics and management systems.

Integration of multi-omics approaches: Recent studies often go beyond DNA sequencing alone. There's an uptick in metatranscriptomics analyses (to see which genes are actively expressed in the rumen under different conditions) and metaproteomics or metabolomics to link microbiome function with fermentation end-products.

Several studies in 2021–2022 integrated both shotgun metagenomics and metatranscriptomics to distinguish between genes present and those actively expressed in cows on different diets. For instance, Xue et al. demonstrated that metatranscriptomic analysis uncovered stronger associations between rumen microbial functions and host feed efficiency compared to metagenomics alone; this included CAZyme expression linked to carbohydrate degradation pathways active in efficient cows (Xue et al., 2022).

Combining microbial profiling with VFA and metabolomics

Other research has paired microbial sequencing (16S or shotgun) with the quantification of volatile fatty acids (VFAs) or broader metabolomic

profiling to relate microbial changes to fermentation chemistry directly. A study of Sanhe heifers highlighted how dietary regimes altered both taxonomic and functional microbial profiles alongside metabolite shifts in rumen fluid, connecting microbiome structure with fermentation outputs (Zhang et al., 2024).

Together, these multi-omics approaches shift the focus from the mere presence of microbial taxa toward a more refined understanding of metabolic activity, identifying which microbes are expressing carbohydrate-active enzymes, producing metabolites such as butyrate or propionate, and influencing fermentation patterns that affect host performance, including methane production or energy harvest.

Consequently, research questions have evolved to "Which microbial pathways are upregulated in more efficient animals?" or "How do microbiome metabolites signal to the host and influence feed utilisation or immunity?" These multi-omic studies are more complex but yield a systems-level view of the rumen as an ecosystem.

Focus on previously neglected community members: While bacteria have dominated past research, there has been a growing interest in rumen archaea, fungi, protozoa, and viruses since 2020. For example, a 2024 study constructed a comprehensive catalogue of 998 archaeal genomes from ruminant guts (including cattle), shedding light on the diversity of methanogens and their relatives across species (Mi et al., 2024). This work demonstrated that archaea exhibit variation by host breed and gut compartment, and even identified new archaeal strains with unusual metabolic genes. Rumen anaerobic fungi (Neocallimastigales) and ciliate protozoa, critical fibre digesters and hydrogen integrators, are notoriously hard to sequence due to large genomes and eukaryotic DNA. Still, recent efforts have applied long-read sequencing or targeted approaches to assemble draft genomes of these organisms (Hanafy et al., 2022). Moreover, the rumen virome (bacteriophages, pathogens and eukaryotes viruses) is being actively studied with metagenomics; for instance, in 2022–2023, researchers described diverse novel rumen viruses and CRISPR arrays in rumen bacteria, some of which may influence bacterial population dynamics or gene transfer. Thus, post-2020 studies pose new questions, such as "How do bacteriophages contribute to rumen microbial turnover or gene flow?" or "Can we manipulate protozoal populations to reduce methanogens indirectly?" The rumen microbiome is now viewed holistically as comprising

bacteria, archaea, eukaryotes, and viruses, rather than just bacteria/archaea. This broadened perspective is crucial for a genuine understanding of ecosystem function (e.g., protozoa synergise with methanogens, and viruses may carry essential genes).

Applied and translational research: In the post-2020 period, there has been a striking shift from descriptive studies to applied and translational research that directly targets livestock productivity and environmental impact. Several studies now focus on leveraging microbiome data in cattle selection programs, identifying rumen microbial biomarkers associated with high feed efficiency (often defined by residual feed intake, RFI) or low methane emission phenotypes alongside conventional genetic breeding values (Fonseca et al., 2023; Fregulia et al., 2024; Peraza et al., 2024). For example, Xie et al. (2022) linked specific rumen taxa and metabolic pathways (e.g., *Prevotella, Fibrobacter succinogenes*, CAZymes, and methanogenesis pathways) to feed-inefficient cattle, highlighting microbial signatures that may serve as predictive markers or targets for intervention (Xie et al., 2022). Adrade et al. (2022) identified amplicon sequence variants (ASVs) in rumen and stool samples of Brazilian Nelore bulls. They showed a strong association with both RFI and methane output. These results suggest the potential of using ASVs as biomarkers for selection or dietary modulation (Andrade et al., 2022).

New multi-omics efforts (e.g., metagenomics, metatranscriptomics, and metabolomics) have also revealed microbial functional networks and metabolic markers (e.g., *Selenomonas, Succinivibrionaceae*, and specific carbohydrate metabolites) that consistently differentiate high- from low-efficiency dairy cows, offering targets for the development of feed additives or precision microbiome modulation (Xue et al., 2022).

Concurrently, dietary interventions informed by metagenomics are increasingly used. Researchers mine extensive rumen gene catalogues for fibre-degrading enzymes or secondary metabolite gene clusters that can be turned into feed additives, enzymes, or engineered probiotics tailored to boost digestive functions or suppress methanogens. Recent studies have demonstrated how metagenomic and metabolomic data can inform the development of feed additives designed to reduce methane emissions and enhance feed efficiency. For example, supplementation with red seaweed (Asparagopsis taxiformis) has shown methane reduction effects exceeding 80%, still with an impact on rumen fermentation and animal productivity

despite being administered in very low dosage (Angellotti et al., 2025). Other interventions study the use of synthetic inhibitors targeting methanogenesis pathways. They achieve significant reductions in enteric methane while maintaining rumen function and animal performance (Krizsan et al., 2023). These novel approaches are increasingly supported by functional omics analyses, metagenomics, metatranscriptomics, and metabolomics, which help elucidate shifts in microbial taxa and fermentation patterns under additive use. These approaches integrate microbial and host response. The data allows researchers to assess not just methane outcomes but also broader sustainability impacts, such as feed conversion efficiency, nitrogen use, and overall productivity (Ramin et al., 2023).

Precision livestock farming tools are also integrating microbiome data, with portable sequencers, methane sensors, and sensor networks now offering the potential for on-farm real-time monitoring of microbial activity or proxy measures (e.g., methane flux), enabling immediate adjustments to diet or management. While full implementation is still in its early stages, pilot studies highlight the feasibility and the beginning of decision-support systems informed by microbial signals.

In summary, since 2020, the field has progressed, and the framing has moved decisively to:

"How can we manipulate or utilise the rumen microbiome to achieve a specific outcome (better growth, lower methane, improved feed efficiency)?" marking a clear shift from earlier, primarily descriptive ecological studies toward strategic, outcome-driven research.

Improved analytical tools and pipelines: Methodologically, since 2020, the field has adopted cutting-edge tools for analysing metagenomic data (elaborated in the next section). As a result, newer studies can assemble more complete genomes with greater confidence. It's now common to retrieve near-complete (>90% complete) MAGs and even closed genomes for some rumen bacteria. This increase in completeness has improved functional annotations and the discovery of biosynthetic gene clusters, CRISPR elements, and other features that were missed in drafts. Quality control is stricter; modern studies often report only high-quality MAGs (completeness $\geq 90\%$, contamination $\leq 5\%$) for analysis (Sáenz et al., 2025). There has also been an uptick in statistical rigour and the application of machine learning to microbiome data (discussed below), ensuring that findings (such as microbial biomarkers) are robust and generalizable.

In essence, the post-2020 period in cattle metagenomics is characterised by scaling up (more samples, more data types) and drilling down (previously under-studied microbial groups and mechanistic insights). The research has become more interdisciplinary, merging animal science, microbiology, data science, and even climate science, to tackle complex questions about host-microbe interactions in the rumen. The ultimate vision is to precisely modulate the rumen microbiome to benefit both the animal (improved feed conversion and health) and the planet (reduced greenhouse gases), and the advancements made since 2020 are rapidly moving us closer to that goal.

2.6.5 Machine learning and deep learning revolution

Machine learning and deep learning have revolutionised the application of metagenomics. Along with progress in sequencing technologies like Oxford Nanopore Technologies (ONT), new chemistry, and PacBio Revio, these innovations have transformed data processing and analysis:

Long-read sequencing (ONT/PacBio): Before 2020, nearly all rumen metagenomes utilised short reads (Illumina), which presented challenges in assembling complete genomes from complex communities. Now, long-read sequencers are being applied to rumen samples, sometimes even generating metagenomes that are long-read-only. The long reads (tens of kb in length) can span repetitive regions and plasmids, enabling assembly of entire microbial chromosomes from the rumen. In Stewart *et al.* (2019), the addition of a subset of Nanopore reads enabled the assembly of at least three complete, single-contig bacterial genomes from the rumen microbiome (R. D. Stewart et al., 2019), a feat previously impossible with short reads alone. Since then, specialised pipelines like NanoPhase (2021; https://github.com/Hydro3639/NanoPhase) have been developed. The new sequencing technologies and methods demonstrated that reference-quality genomes can be reconstructed using only Nanopore reads from metagenomic DNA (Moss et al., 2020). The impact on cattle microbiome research is huge: complete MAGs mean we can identify genes like virulence factors or antibiotic resistance elements that often reside on plasmids, and we get complete rRNA operons and other genomic regions that short-read assemblies miss. Long reads also enhance the assembly of eukaryotic genomes (such as those of rumen fungi and protozoa), which are too large and repetitive for short-read assembly.

Additionally, real-time sequencing with ONT opens the door to on-farm or on-site analysis of rumen microbes in the future. The caveat had been higher error rates, but steady improvements in basecalling accuracy (thanks to deep learning models in ONT's basecallers) have made long reads more accurate and thus more widely adopted. In summary, the way we process reads has changed; modern studies often employ *hybrid assembly* (short + long reads) or even long-read assembly to achieve high-contiguity genomes, drastically improving the quality of microbiome data from cattle.

Advanced binning algorithms (ML-powered): Binning contigs into MAGs previously relied on simple compositional patterns and coverage clustering with tools such as MetaBAT2 (Kang et al., 2019) and CONCOCT (Alneberg et al., 2014). Since 2020, deep learning and embedding techniques have revolutionised metagenomic binning. For example, VAMB (Variational Autoencoder for Metagenomic Binning) (Nissen et al., 2021) introduced an autoencoder neural network that learns latent representations of contigs (based on oligonucleotide frequency and coverage across samples) to group them more accurately than manual thresholds. Other tools, e.g. Semibin2 (Pan et al., 2023) or Comebin (Z. Wang et al., 2024) utilise contrastive learning, an approach that can better resolve complex genomes (e.g., differentiate strains) by generating multiple views for each contigs. These ML-based binning methods reduce contamination and recover more complete genomes, particularly for *high-coverage yet complex communities, such as the rumen*. In practice, many recent rumen studies employ a combination of algorithms and then refine the bins by automating what was previously a laborious, manual curation process. Deep learning is also used in taxonomic classification: instead of BLASTing each contig, models can classify sequences based on learned features, improving speed and sometimes accuracy for novel sequences. As an example, a 2022 review successfully applied to metagenomic binning tasks, convolutional neural networks and autoencoders, highlighting that these approaches capture sequence patterns beyond the scope of traditional methods (Elhassani et al., 2021). The net effect is that more high-quality MAGs are recovered from a given dataset than was possible a few years ago, and this is particularly beneficial for rumen samples, where genomes of interest may be closely related or occur at low abundance.

Genome quality assessment, CheckM2 (Chklovski et al., 2023)**:** A significant development in computational tools is CheckM2, released in

2022–2023, which utilises, in part, machine learning to evaluate MAG quality. The original CheckM (2015) identified single-copy marker genes to estimate completeness and contamination of a genome bin. CheckM2 instead was trained on a wide range of genomes to predict completeness more accurately, even for novel lineages, using patterns learned via ML . It can adjust to new reference genomes on the fly and correct some biases (for instance, tiny genomes of specific symbionts or very GC-rich genomes that confounded CheckM1). For rumen studies, CheckM2 provides greater confidence in MAG quality, mainly since many rumen microbes belong to lineages with no close reference. In this scenario, machine learning predictions help avoid under- or overestimating completeness . The most recent cattle MAG studies report CheckM2 scores to substantiate that their recovered genomes are of genuinely high quality. In practical terms, this means downstream biological analyses (like estimating pan-genomes or metabolic capacities) rest on a more solid foundation of genome quality.

Machine learning for predictive modelling: Beyond data processing, ML and deep learning are increasingly used to derive insights from the rumen microbiome. One primary application is in predicting host phenotypes from microbiome data. To predict traits such as feed efficiency, methane yield, or dairy production metrics, approaches like random forests, support vector machines, and deep neural networks have been applied to 16S or shotgun profiles. For example, Difford *et al.* (2018) employed statistical learning to predict methane emissions from rumen microbial profiles, achieving moderate accuracy, which suggests a microbial signature for distinguishing between low- and high-methane cattle (Difford et al., 2018). More recently, Monteiro *et al.* (2024) applied an ensemble of machine learning models to rumen metagenomic data from 454 Holstein cows and were able to predict feed efficiency (using Residual Feed Intake) with approximately 36% of the variance explained by microbiome features (Monteiro et al., 2024). They also identified key microbial genera (like *Ruminococcus* and *Butyrivibrio* groups) that were most influential in the predictions, providing biologically interpretable results. Likewise, deep learning models (e.g., neural networks with attention mechanisms) have been used to predict dairy cows' milk yield or composition from the rumen microbiome, and to identify microbial indicators of health issues (such as subacute rumen acidosis or mastitis risk) through changes in gut microbiome. These models can handle high-dimensional data and complex interactions more effectively than traditional

statistics. Machine Learning considers the whole community and its relationship with host performance. The outcome is a move toward data-driven, deep-learning predictive understanding, rather than studying one microbe at a time. As these models become more interpretable (through techniques that highlight which microbial features are most important), they also generate hypotheses: for instance, if a model consistently flags *Prevotella* abundance as a top predictor of efficiency, researchers can investigate how *Prevotella*-driven fermentation might confer that benefit.

Deep learning in functional annotation: The application of deep learning to functional annotation is transforming how we interpret complex metagenomic datasets such as those from the rumen. Modern deep learning architectures, including convolutional and recurrent neural networks, have shown promising results in predicting protein functions, such as enzyme classes or antibiotic resistance genes, directly from raw sequence data. In the context of rumen metagenomes, these models can help uncover the roles of previously uncharacterized genes, highlighting candidates involved in glycoside hydrolysis or methanogenesis.

Deep learning has played a crucial role in identifying biosynthetic gene clusters from metagenome-assembled genomes (MAGs). Additionally, it points to novel antimicrobial compounds or metabolites with potential biotechnological applications. However, as noted by Mathieu et al. (2022), the rapid evolution of model architectures, training strategies, and reference datasets poses a significant challenge: models and benchmarks that were state-of-the-art just a few years ago may no longer be adequate today. Keeping the pace with these developments requires continual adaptation and critical evaluation of both methods and biological interpretations (Mathieu et al., 2022).

In summary, machine learning and deep learning have become integral to state-of-the-art metagenomics in cattle. They enable researchers to handle the complexity and scale of modern datasets, from efficiently assembling genomes with long, noisy reads to robustly binning thousands of contigs, and making sense of how hundreds of microbial species collectively influence a cow's phenotype. The revolution in computational tools since 2020 enables us to extract significantly more information from rumen samples than previously possible. A single study today might sequence a cow's rumen, assemble dozens of genomes, evaluate them using ML-based quality control, and then utilise an ML model to predict that cow's methane output from its

microbiome. These approaches are accelerating discovery and pushing towards practical applications, such as microbiome-assisted breeding or precision nutrition, which were once merely dreams a few years ago.

2.7 Conventional Breeding vs. Microbiome-Informed Breeding

2.7.1 Traditional selection traits (growth rate, milk yield, etc.)

Cattle breeding involves selecting animals based on traits of interest such as growth rate, milk yield, fertility, and disease resistance, and tailored to specific production goals (beef or milk) and the environment (e.g. hot or temperate climate). Breeding aims to maximise desirable outcomes by evaluating phenotypic performance not only of the individual animals but also considering their pedigree and progeny (Pryce & Daetwyler, 2011).

Historically, selection relied on subjective assessments, but modern approaches have significantly improved objectivity. Estimated Breeding Values (EBVs) marked the initial shift towards more systematic breeding practices. The aim is to quantify the genetic merit of animals for specific traits by considering both individual phenotypes and the performance of related individuals (parents and offspring). This approach separates the genetic effects from the environmental influences. It allows breeders to select based on genetic potential rather than potentially biased phenotypic observations (*Interpreting EBVs and Indexes*, n.d.; Martín et al., 2021).

The Genomic Estimated Breeding Values (GEBVs) represent a further advancement. GEBVs aim to estimate the value of the animal early in its development. For this, it integrates direct genomic information (e.g., SNP markers) into the evaluation. This method relies on associating genetic markers with observed traits, thus enhancing accuracy as the volume of genetic and phenotypic data increases through ongoing research and data accumulation (Hayes et al., 2009; Wiggans & Carrillo, 2022).

2.7.2 Emerging interest in the microbiome as a trait

Increasing recognition of the microbiome's impact on animal productivity and welfare is prompting researchers to investigate its role as a potential breeding target. The rumen microbiome has a significant effect on host

productivity, methane emissions, resilience to environmental stress, and resistance to diseases (Wallace et al., 2019; Waters et al., 2025). These microbial communities influence fermentation efficiency, feed utilisation, and energy extraction. Thereby shaping crucial production traits and environmental impacts (Wallace et al., 2019).

Furthermore, recent studies have begun exploring the interplay between host genetics and the microbiome, revealing potential heritable components. For instance, a study by Li et al. (2019) showed that the host genotype influences the rumen microbial composition, suggesting the feasibility of microbiome-informed genetic selection (F. Li et al., 2019). Wallace et al. (2019) highlighted evidence that specific microbial taxa in the rumen may be influenced by host genetics. Those taxa could thus be associated with traits of economic relevance. Again, underscoring the potential for integrating microbiome data into livestock breeding programs (Difford et al., 2018).

The adoption of microbiome-informed breeding approaches remains limited in cattle breeding practices. Recent research has proposed the concept of integrating microbial data into breeding evaluations, referred to as metagenomic/genomic Estimated Breeding Values (Meta-GEBV). Ross & Hayes et al. (2022) propose using the microbiome for predicting phenotypes (Ross & Hayes, 2022). Still, we could potentially further refine this approach by incorporating both genomic and microbiome data to enhance the accuracy of selection. However, the practical application of Meta-GEBVs is still in its early stages, requiring substantial validation and integration efforts before it can be implemented systematically.

3. Aims of the project

3.1 Aim 1: Seasonal Dynamics in Indigenous Cattle

We aimed to explore whether the microbiome composition shifts under seasonal stress in the rumen microbiome of Ethiopian indigenous cattle.

We hypothesise that the dry season results in an increased abundance of fibre degraders (e.g., *Fibrobacter* and certain *Bacteroidetes*) as the feed is harder to digest and requires better degraders, accompanied by a decline in methanogen and other genera populations due to the lack of feed and competition for it. We also expect to see tetracycline and aminoglycoside antibiotic resistance genes (ARGs), as their use is highly spread and not regulated. The use is more dependent on the farmer than the veterinarian (Gemeda et al., 2020).

Beyond taxonomic shifts, the long-term vision is to investigate the potential interactions between the host genome and microbial composition.

Although our sample size is limited, this project is the first step toward identifying genomic markers associated with microbiome-driven phenotypes. These exploratory findings should enable more in-depth studies and broader population-scale research in the future.

3.2 Aim 2: FAIR-Compliant Method Development

One of the major challenges in scientific research is ensuring the reproducibility of results. To address this, the scientific community has developed the FAIR principles. The first principle is to have the research Findable, the data must be easily found through common databases search engine and with immutable identifiers. The second is to have the data Accessible, in addition to finding the data anyone should be able to get it. The third the data must be Interoperable, this means that the data must be able to be used by different tools, methods and people. The last is Reusability, any FAIR data should have enough metadata and information to yield results when analysed by someone.

A key objective of this work was therefore to strictly follow the FAIR principles throughout every stage of the data lifecycle. However, the project did not stop at compliance; it actively contributed to improving existing

standards. To achieve this, we developed in-house bioinformatics pipelines and software tools specifically designed for analysing rumen metagenomic data. These tools are broadly applicable and can be used by other researchers across diverse metagenomic studies.

Among the tools developed are the MUFFIN pipeline and the PANKEGG visualisation platform, both of which were built with a strong emphasis on transparency, reproducibility, and modularity.

4. Methods

4.1 Paper III: Whole Genome Sequences of 70 Indigenous Ethiopian Cattle

4.1.1 Sampling and DNA Extraction

For this study, we collected blood samples from 70 individuals across seven indigenous Ethiopian cattle breeds. The seven breeds were Afar, Arsi, Barka, Fogera, Horro, Sheko, and Begait. Animals were sampled from different agroecological zones to capture environmental and genetic diversity relevant to local adaptation. The blood was collected by jugular venipuncture into EDTA tubes, immediately placed on ice, and transported to the laboratory for storage at -20°C.

For the DNA extraction, we use a standard phenol-chloroform protocol. Then, the integrity of the DNA was checked using 1% agarose gel electrophoresis. The purity and concentration were assessed with a Nanodrop spectrophotometer and Qubit fluorometry.

4.1.2 Library Preparation and Sequencing

Sequencing libraries were prepared using the NEBNext Ultra DNA Library Prep Kit following the manufacturer's protocol. DNA was fragmented to an average size of 350 bp using sonication. After end-repair, A-tailing, and adapter ligation, libraries were purified and amplified by PCR.

Libraries were pooled and sequenced on the Illumina NovaSeq 6000 platform, generating paired-end reads of 150 bp. Each sample achieved a mean genome coverage of approximately 15x, ensuring sufficient depth for reliable variant detection.

4.1.3 Data Processing and SNP Calling

Raw reads were quality-checked using FastQC and trimmed with Trimmomatic to remove low-quality bases and adapter sequences. Cleaned reads were then aligned to the Bos taurus reference genome (ARS-UCD1.2) using BWA-MEM. SAMtools was used to convert, sort, and index the aligned reads.

PCR duplicates were marked and removed using Picard tools. The Genome Analysis Toolkit (GATK, v4.4.0.0) was used following best practices for variant calling. The workflow included recalibration of base quality scores (BQSR), indel realignment, and SNP calling using HaplotypeCaller in GVCF mode.

We performed joint genotyping through all 70 samples, followed by variant quality filtering using thresholds based on depth, quality by depth (QD), mapping quality (MQ), and strand bias. The resulting SNP dataset was used for downstream analysis, including evaluation of population structure, detection of selective sweeps, and investigation of potential trait-associated loci (Figure 3).

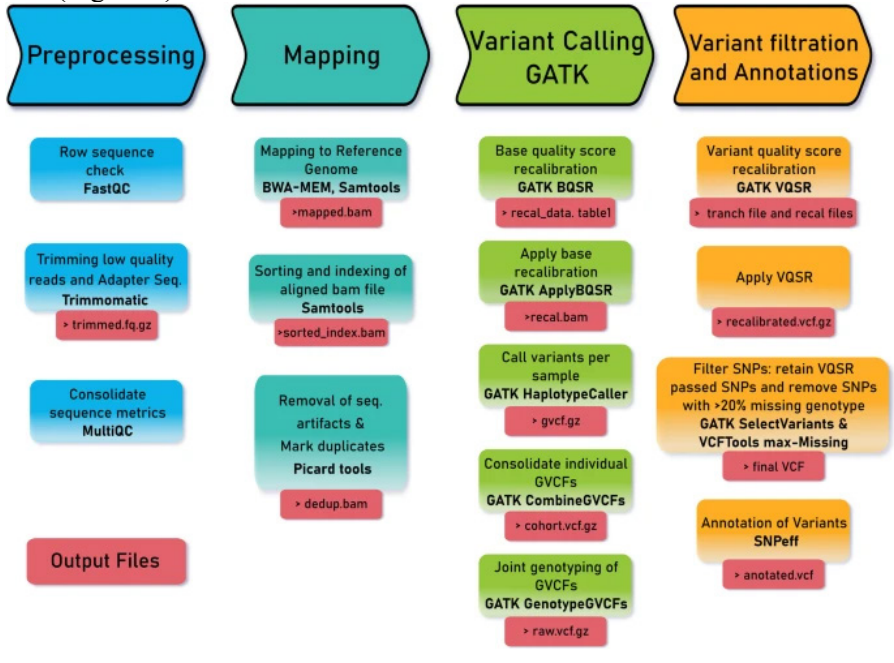

Figure 3: Overview of raw data quality control, sequence mapping, variant calling, and variant filtration pipeline. The pipeline follows GATK's best practice protocol for germline short variant discovery. Source: Paper III (Ayalew, Xiaoyun, Tarekegn, Naboulsi, et al., 2024)

4.1.4 Relevance to the PhD Project

This high-quality variant dataset provided a genomic backbone for investigating local adaptation and environmental resilience in Ethiopian cattle. Although the associated phenotypic studies extended beyond the

scope of this thesis, the pipeline established here serves as a foundation for integrating host genomic markers with rumen microbiome features in future studies.

Although not explicitly applied to the Ethiopian Boran breed that was used for the rumen metagenome study (paper IV), this dataset has been instrumental in advancing studies on environmental adaptation and production traits in Ethiopian cattle.

4.2 Paper IV: Seasonal Dynamics of the Rumen Microbiota in Ethiopian Boran Cattle

4.2.1 Sampling

Ethiopia exhibits a diverse climatic pattern, but generally experiences three main seasons: the long rainy season (‘Kiremt’) from June to August, the dry season (‘Bega’) from October to February, and the short rainy season (‘Belg’) from March to May. In the southern lowland pastoral areas such as Borena, these patterns vary slightly, with the primary rainy season typically occurring from March to May and a short rainy period in October and November. Sample collections for this metagenomic study was conducted during February, aligning with the end of the dry season (Bega) and middle of April, aligning with the middle of the primary rainy season (Belg) in the Borena area to allow the time for the grass to grow and the time for the rumen microbiome to stabilise under the new conditions. The sampling period was also constrained on the availability and feasibility of the sampling due to external constraint rather than the scientifically ideal time. This strategic timing enables the capture of microbial communities under contrasting ecological conditions, including dry and early wet periods, which can influence host-microbiome interactions, water and forage availability, and overall microbial diversity. This seasonal context is crucial for understanding the observed metagenomic profiles related to environmental and ecological dynamics in pastoral systems. The Sample collection was carried out in February 2022 (dry season) and April 2022 (rainy season) at the Boran cattle Dida Tuyera ranch in Yabelo town, Oromia Regional State, Ethiopia. In February, rumen content was collected from 20 Boran cattle, with the same individuals re-collected in April. Due to an extreme drought in Ethiopia in

2022, only 7 of the initial individuals could be sampled again in April as the other died from the drought..

The rumen content was obtained via oro-gastric intubation. The initial 10 mL of rumen fluid was discarded to avoid contamination with saliva, and subsequent solid and liquid phases were collected in sterile 50 mL Falcon tubes. Samples were kept on ice during transportation to Addis Ababa University, where they were stored at -20°C. Later, they were shipped on dry ice to the Swedish University of Agricultural Sciences and stored again at -20°C.

All biological samples imported from Ethiopia to Sweden for this study were collected and transported in full compliance with the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from Their Utilisation. Prior Informed Consent (PIC) and Mutually Agreed Terms (MAT) were obtained between the Institute of Biotechnology, Addis Ababa University, and the Swedish University of Agricultural Sciences through the Ethiopian Biodiversity Institute, ensuring that the acquisition and use of genetic resources complied with national and international legal frameworks. Necessary export and import permits, including Material Transfer Agreements (MTAs), were secured before shipment. The research team also fulfilled all due diligence obligations under EU Regulation (No. 511/2014), ensuring transparent and lawful use of the samples in accordance with both Ethiopian and Swedish regulations.

The samples are denominated with the month of sampling and the cow ID. The cow IDs are 0199, 0350, 0407, 0428, 0446, 0476, 0667. So the sample for the cow 0199 in February is Feb0199, and in April is April0199.

4.2.2 DNA Isolation, Library Preparation, and Sequencing

DNA was extracted using the ZYMO RESEARCH Quick-DNA Faecal Microbe MiniPrep Kit, in combination with non-kit-based protocols to improve yield and reduce fragmentation. Extraction quality was assessed using Nanodrop, Qubit, and Tapestation.

Two separate DNA extractions were performed for each sample, one for Illumina sequencing and another for Oxford Nanopore sequencing.

Illumina Sequencing

Library preparation and sequencing were performed at SciLifeLab using the TruSeq PCR-Free DNA Library Preparation Kit on 25 μL of input DNA.

Libraries were multiplexed and sequenced using paired-end 150 bp reads on a NovaSeq 6000 S4 lane.

Nanopore Sequencing

Before library preparation, DNA was improved using the NEB Blunt/TA Ligase Master Mix (M0367), NEBNext FFPE Repair Mix (M6630), NEBNext Ultra II End repair/dA-tailing Module (E7546), and NEBNext Quick Ligation Module (E6056). Libraries were prepared using the Native Barcoding Kit 24 V14 (SQK-NBD114.24). Sequencing was conducted on FLO-MIN114 flow cells (R10.4.1), with 2 to 3 samples per flow cell and run times of 48 hours.

The Flowcells were subsequently washed using the Wash Kit (EXP-WSH004). If quality and pore count were sufficient, samples were resequenced. Only samples 199April, 199Feb, and 667Feb were sequenced just once due to poor flow cell performance.

4.2.3 Bioinformatics

Quality Control

Quality assessment was performed using FastQC (v0.12.1)(Andrews et al., 2012), PycoQC (v2.5.2)(Leger, 2017/2025), and MultiQC (v1.15)(Ewels et al., 2016). Quality filtering was performed using fastp(v0.20.0)(Chen, 2023) for Illumina reads and Chopper (v0.9.0)(De Coster & Rademakers, 2023) for Nanopore reads, both of which were integrated through the MUFFIN pipeline. The reads were also mapped to the Bos taurus reference genome ARS-UCD2.0, using Bowtie2 (v2.5.3)(Langmead et al., 2019) for the Illumina reads and Minimap2 (v2.17)(H. Li, 2018) for the Oxford Nanopore reads. We removed the reads mapping to the reference using samtools (v1.21)(Danecek et al., 2021).

Reads Classification

Illumina and Nanopore reads were classified using Kraken2 (v2.1.3)(Wood et al., 2019), with the PlusPF database (release date 2024-04-09) [https://benlangmead.github.io/aws-indexes/k2]. The PlusPF database contains the RefSeq database for archaea, bacteria, viral, plasmid, human[1], UniVec_Core, protozoa and fungi formatted for Kraken2.

The results were computed in an R Markdown script (https://github.com/RVanDamme/Boran_kraken_study). The script worked as follows:

Raw Kraken2 reports from both Illumina- and Oxford Nanopore–sequenced Ethiopian Boran samples were imported into R (R Core Team, 2021) and merged into a single data frame, keyed by season (February or April), organism ID, and sequencing platform. Reads assigned to clades were filtered to remove low-confidence hits (a minimum of 1,000 reads per clade for Illumina and 100 reads per clade for Nanopore). To reduce platform-specific noise, only taxa detected by Nanopore in a given sample were retained from the corresponding Illumina dataset ("true hits"). For each taxonomic rank of interest (e.g., phylum, class, order, family, genus, species), at the clade level, we use the read counts to calculate the relative abundances (number of reads for the clade divided by the total number of reads classified); any missing values were set to zero. The Relative-abundance matrices underwent centred log-ratio (CLR) transformation to address compositionality. Temporal shifts between February and April were quantified as differences in CLR-transformed abundances (ΔCLR = April CLR - February CLR), with the most variable taxa highlighted by bar plots.

All data wrangling, statistical analyses, and visualisations were performed in R version 4.4.1 using the following libraries: tidyverse, tidyr, vegan, phyloseq, reshape2, ggplot2, ape, stringr, dplyr, pheatmap, ggstream, patchwork, viridis and purrr.

MUFFIN Pipeline

The metagenome-assembled genomes reconstruction, as well as their functional annotation, were performed using MUFFIN, a hybrid metagenomics pipeline designed for modularity and high-quality genome recovery. MUFFIN integrates long-read and short-read sequencing, assembly, binning, and annotation using updated tools and configurations.

Table 1 details the parameters and configurations applied during the MUFFIN pipeline run, including specific tools, modes, and justifications for their use.

Table 1: Configurations Used in the MUFFIN Pipeline

Parameter	Option	Description
-profile	dardel	Utilized Singularity and Nextflow parameters tailored for Dardel requirements (SLURM, etc.).
--mode	hybrid	Used both Nanopore and Illumina reads for hybrid assembly.
--assembler	metaspades	Selected SPAdes hybrid assembly with metagenomic option instead of Flye assembly.
--modular	full	Executed all steps of the pipeline (assemble, classify, annotate).
--bintool	metabat2	Chosen as the other binning methods are still being integrated into MUFFIN version 2.
--skip_bad_reads_recovery	enable	Skipped recovery of unused reads due to resource and time constraints.

For a more detailed presentation of MUFFIN, see chapter 5 (Paper I).

We then used GTDB-TK to classify the MAGs again using the GTDB-TK 2.4.1 with the latest database (release 226). As sourmash is accurate and fast, but it can sometimes lack sensitivity.

Downstream Analysis with PANKEGG

MAG-level annotations and comparisons were visualised and analysed using PANKEGG, a lightweight platform for integrating taxonomy, quality scores, and functional annotation. Through the visualisation, we enabled high-resolution comparison of MAGs across individuals and time points.
For a more detailed presentation, see chapter 6 (Paper II).

5. Paper I Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN)

Metagenomic analysis, especially from complex environments such as the rumen, presents multiple computational challenges, including managing vast amounts of short- and long-read sequencing data, as well as accurately assembling genomes and annotating their functions. MUFFIN (Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics, and pathway analysis) was developed to provide a reproducible, modular, and scalable solution for such tasks. The pipeline is designed to incorporate best-in-class tools while facilitating integration of both short-read (Illumina) and long-read (Nanopore) sequencing data to improve genome assembly quality and downstream functional analysis.

The MUFFIN pipeline was released in 2021 as a modular workflow for metagenomic data analysis, integrating hybrid assembly of short- and long-read data, binning, and functional annotation. It was designed with transparency and flexibility in mind, allowing users to customise steps and scale up analyses across multiple datasets. MUFFIN was built using Nextflow and Conda environments or Docker/Singularity containers to ensure reproducibility.

5.1 Initial Design and Workflow

The original pipeline follows a five-stage structure (See Figure 4):

I. **Read Preprocessing:** Short reads (typically Illumina) are quality-trimmed with fastp, while long reads (Nanopore or PacBio) are filtered using chopper.

II. **Hybrid Assembly:** The trimmed reads are assembled using metaSPAdes (both short and long reads together) or Flye (use only the long reads). For Flye it is followed by polishing with the short reads.

III. **Binning and Refinement:** Contigs are binned using a combination of MetaBAT2, MaxBin2, and CONCOCT. Metawrap refine module then integrates the binning results to yield high-quality MAGs performing an automatic bin refinement.

IV. **Quality Assessment:** CheckM is used to estimate completeness and contamination, and Sourmash, using the GTDB, is used for the taxonomic classification of the MAGs

V. **Annotation and Pathway Analysis:** Functional annotation is performed using eggNOG-mapper, then the Quality and Annotation of the MAGs are parsed into PANKEGG.

VI. **Optional Transcriptomics analysis:** If provided, RNA-seq data can be assembled using Trinity and then quantified with Salmon. The last part is to annotate the transcripts to see the functional expression at the time of sampling.

MUFFIN proved valuable for studies that combine the precision of Illumina with the structural benefits of long reads, especially in complex environments such as the rumen.

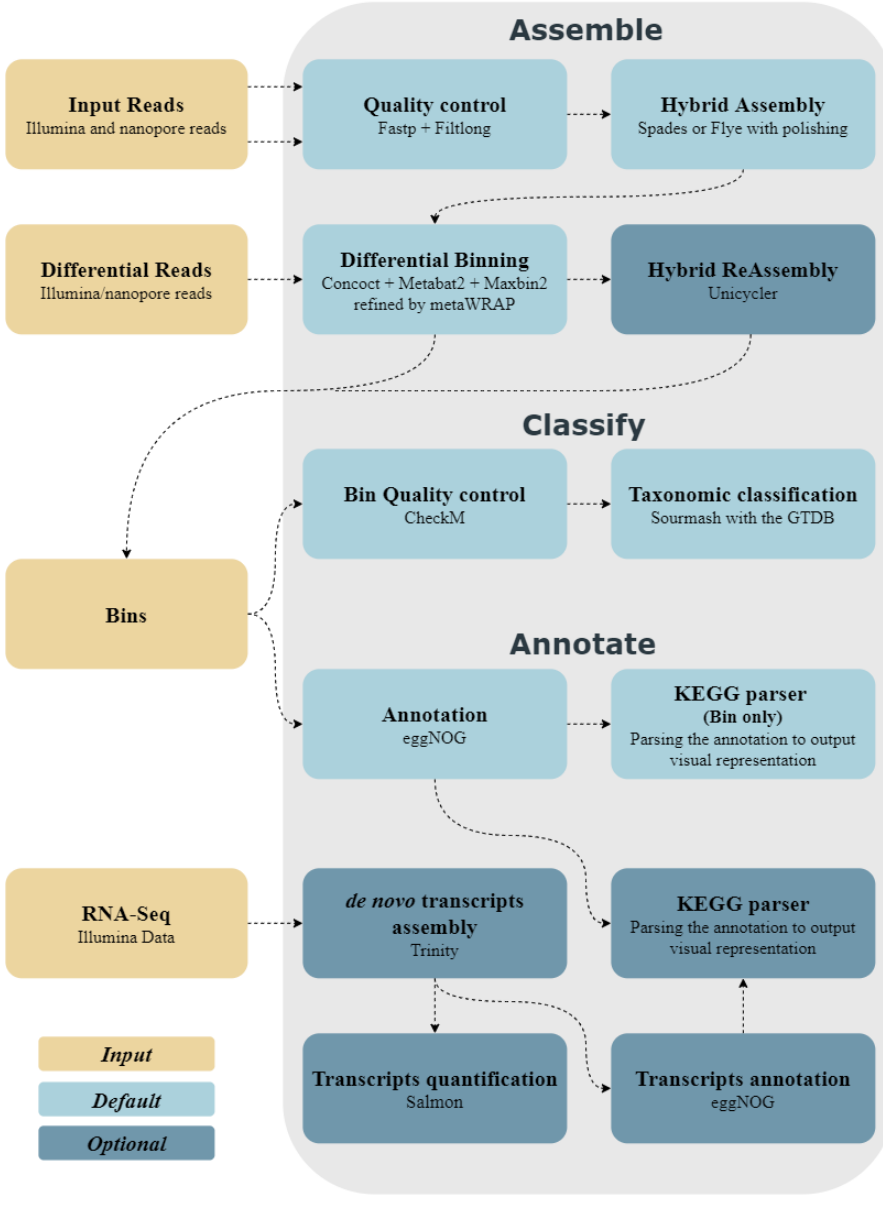

Figure 4: Simplified overview of the MUFFIN workflow. All three steps (Assemble, Classify, Annotate) are shown from top to bottom. The RNA-Seq data for Step 3 (Annotate) is optional.

5.2 MUFFIN Version 2: Responding to Changes in State of the Art

Since the original publication, the field has undergone significant evolution. New tools have emerged, and existing tools have improved in both speed and accuracy. During this PhD, MUFFIN was updated to maintain adherence to best practices in metagenomics. The Table 2 outlines the significant improvements and new features introduced in MUFFIN Version 2, including workflow updates, preprocessing, assembly, binning, quality assessment, annotation, and enhancements to the user experience.

Table 2: Key Changes in MUFFIN Version 2

Category	**Changes/Improvements**
Workflow and Containerization	Transition from conda to fully containerized Docker/Singularity system for better reproducibility and HPC compatibility. Improved container implementation compared to Version 1. Integration of Nextflow DSL2 modules. Addition of short-read-only and long-read-only modes.
Read Preprocessing	Enhanced adapter removal and trimming with updated fastp. Replaced Chopper with Filtlong for long-read preprocessing, improving ONT error handling.
Assembly	Reworked Pilon and Medaka usage for polishing hybrid assemblies, improving consensus accuracy.
Binning	Replaced MaxBin2 and CONCOCT with more efficient tools: ComeBin (multi-view contrastive learning) and SemiBin (semi-supervised binning using MAG databases).
Quality Assessment	Shifted from CheckM v1 to CheckM2, enabling faster and more accurate genome completeness and contamination estimation.
Annotation and Visualisation	Updated annotation with eggNOG-mapper v2 (faster, extended functional coverage). Export formats prepared for downstream analysis in PANKEGG.
User Experience and Scalability	Expanded logging and reporting with MultiQC and Nextflow trace reports. Enhanced support for large-scale studies with batched sample analysis.

5.3 Installation and Reproducibility

MUFFIN supports installation via git, Nextflow or through a download. It is tested across multiple platforms (Linux and HPC environments), ensuring broad compatibility.

Documentation and installation instructions are hosted on GitHub (https://github.com/RVanDamme/MUFFIN), making it easily accessible to the broader community.

We recommend that users wait for the release of MUFFIN version 2, as the current version is under development.

5.4 Role in This PhD

The initial version of MUFFIN was planned for the analysis phases of Paper IV, providing solid hybrid assembly and annotation results for our rumen microbiome samples. However, due to the delays in sample delivery and the evolving state of metagenomics tools, a second version of MUFFIN was developed, which was then used for Paper IV.

MUFFIN Version 2 made the analysis of Ethiopian cattle microbiomes more efficient and comprehensive. The pipeline now retrieves MAGs of higher completeness and lower contamination, leverages more powerful binning algorithms, and includes expanded options for annotation and pathway reconstruction. In practice, the updated pipeline improved the consistency and reproducibility of MAGs across individuals and seasons. Thereby enhances our ability to identify taxonomic and functional shifts in the rumen microbiota.

Moreover, the use of containerization and the modularity of the workflow enable greater flexibility and adaptability, allowing other research groups to easily implement the pipeline in different computational environments and for other host-microbiome systems. This positions MUFFIN not only as a tool for this PhD but also as a general-purpose, community-accessible workflow to support microbiome research aligned with FAIR principles.

The ongoing development of MUFFIN underscores the need for adaptive tools in the rapidly evolving field of metagenomics. While MUFFIN was

designed with cattle microbiomes in mind, its architecture and logic apply to a wide variety of environments and host-associated communities.

6. Paper II PANKEGG: Integrative Visualisation and Comparison of Metagenome-Assembled Genomes Annotation, Taxonomy, and Quality

With the evolution of technology and increased sequencing depth, the complexity and scale of metagenomic data, particularly those derived from shotgun sequencing of host-associated microbiomes, have also increased. It is now essential to integrate taxonomy, quality assessment, and functional annotation into coherent and accessible formats.

We developed PANKEGG to address this need, offering a streamlined and intuitive web-based application for visualising and comparing metagenome-assembled genomes (MAGs). PANKEGG is a two-tool software, the first of which parses different output information from metagenomic pipelines and compiles it into a central SQL database. The database then serves as the backbone of the PANKEGG application, enabling users to explore microbial diversity and functionality at the genomic level through a local web browser interface.

6.1 Tool Architecture and Function

PANKEGG is written in Python and uses a lightweight SQL database (SQLite) as its data store. It supports the integration of outputs from CheckM2 for genome quality metrics (completeness and contamination), GTDB-Tk or Sourmash for taxonomic classification, and eggNOG-mapper for KEGG ortholog annotations.

The tool consists of two main components:

- **PANKEGG_make_db.py**: This tool parses the input files and stores the information into a structured SQL database. The input is configured through a simple CSV file.
- **PANKEGG_app.py**: This tool runs a local server accessible through a browser to visualise the database and enables real-time filtering, sorting, and plotting of results.

6.2 Key Features

PANKEGG's web interface provides multiple views:

The Bin page displays the metadata for each bin, including sample ID, quality metrics, and classification. The table can be filtered to discard all bins/MAGs below the medium-quality MIMAG standard or filtered using the search bar for a sample name, bin name, or classification (Figure 5).

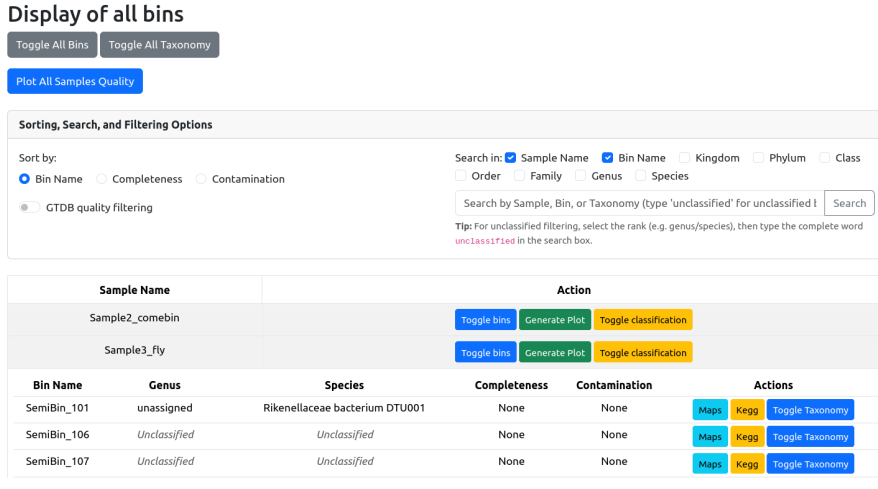

Figure 5: Screenshot of an example of the bin page displayed by PANKEGG.

The Pathway page (also known as the Map page) displays the KEGG pathway completeness. The information can be filtered to display the pathways of one or more bins/MAGs or samples, enabling comparison across datasets. An additional search bar is available to filter pathways and search for specific keywords (Figure 6).

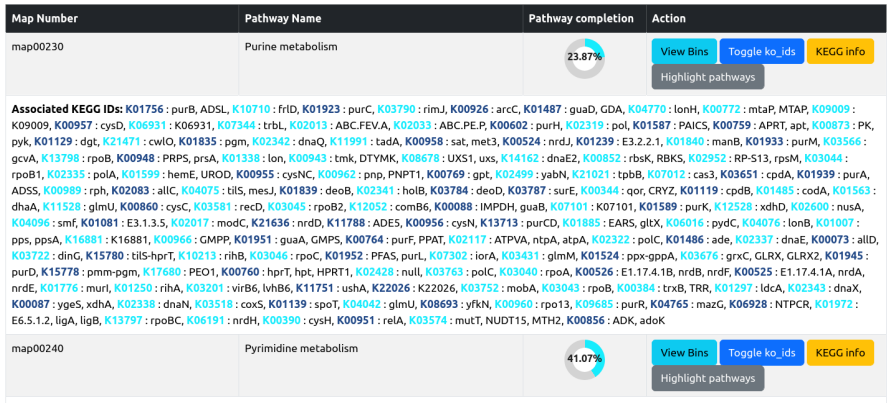

Figure 6: Screenshot of an example of the map page (pathway) displayed by PANKEGG.

The KEGG Ortholog page lists each identified KEGG ortholog, its associated functions, and the bins/MAGs in which they are located (Figure 7).

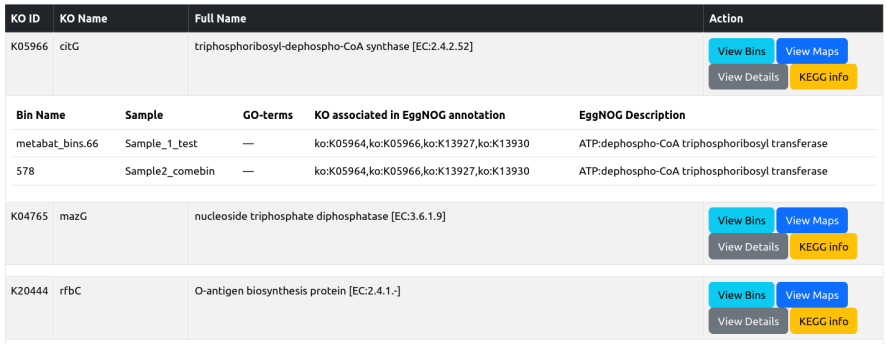

Figure 7: Screenshot of an example of the KEGG page displayed by PANKEGG.

The Taxonomy page summarises the taxonomic composition across all bins/MAGs. A selector at the top allows the user to view a specific rank (Figure 8).

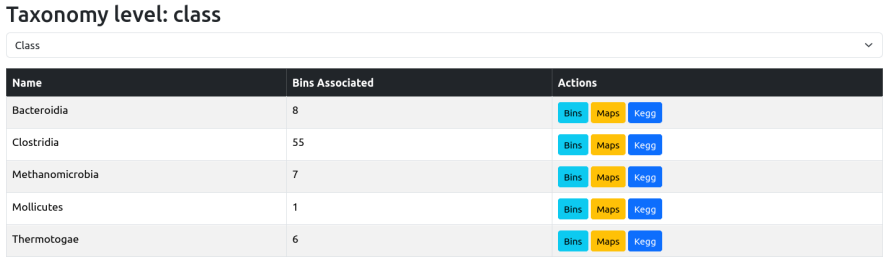

Figure 8: Screenshot of an example of the taxonomy page displayed by PANKEGG.

The "Sample vs. Sample" and "Bin vs. Bin" pages enable users to cross-compare samples or bins/MAGs based on quality and functional annotation.

On both pages, a user can select two elements to compare, and then can visualise different information. On the Sample vs Sample page, users can choose one or more pathway groups from the KEGG database to view a heatmap of the pathways present in each group for their sample. The colour is based on the completion of the pathway for that bin (Figure 9). The user can also see a scatterplot of the completeness vs contamination of the MAGs present in the samples (Figure 10). On both the "Sample vs. Sample" and "Bin vs. Bin" pages, the user can view a table and a plot comparing the number of orthologs found in the first, the second, or both elements being compared (Figure 11).

Figure 9: Screenshot of an example of the heatmap of the “Sample vs Sample” page displayed by PANKEGG.

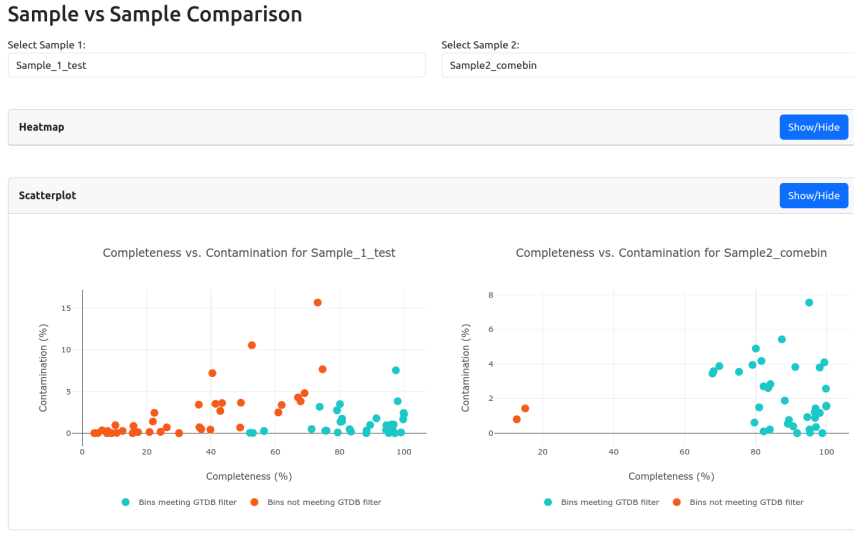

Figure 10: Screenshot of an example of the scatterplot of the “Sample vs Sample” page displayed by PANKEGG.

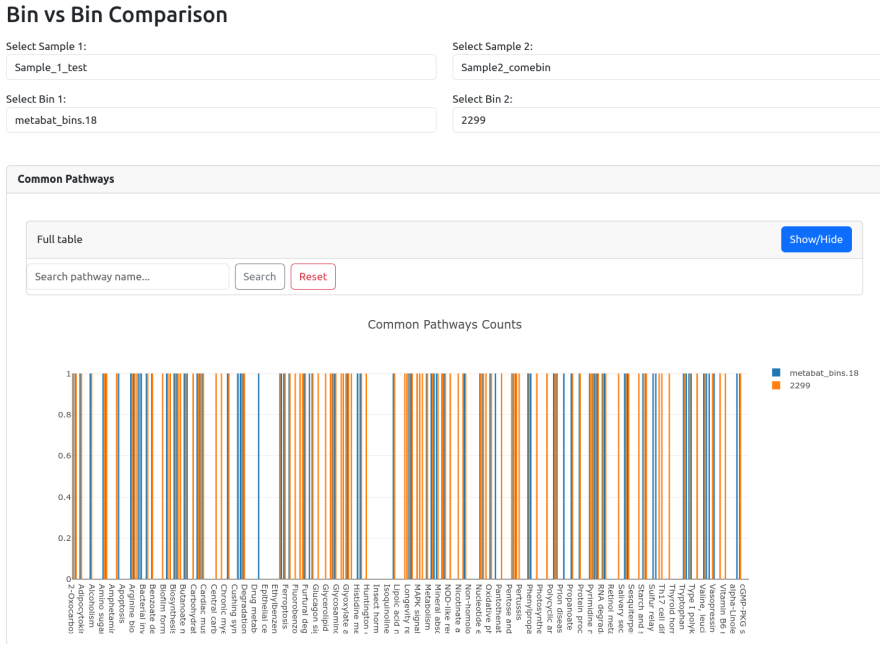

Figure 11: Screenshot of an example of the common pathways table and plot from the "Bin vs Bin" page displayed by PANKEGG.

On **the PCA page**, a Principal Component Analysis is performed based on the functional or taxonomic results, allowing the user to visualise how the samples or bins/MAGs cluster together or not. Beware that PCA interpretation is only valid with enough data, we recommend at least 40 MAGs (Shaukat et al., 2016).

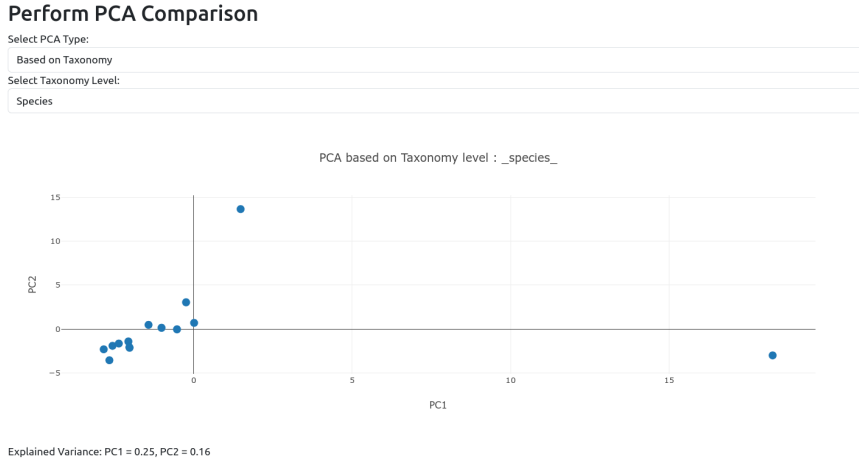

Figure 12: Screenshot of an example of the PCA page displayed by PANKEGG. The explained variance is detailed below the graph.

Interactive elements such as sortable tables, linked views, and colour-coded pathway maps enhance the user experience. Users can export filtered tables for downstream statistical analyses.

6.3 Motivation and Innovation

PANKEGG does not attempt to manage the complete metagenomics process, like pipelines such as Anvi'o or MAGFlow would do. Instead, it focuses on one of the most challenging aspects of post-assembly analysis: integration and interpretation. It fills a niche by enabling easy navigation across thousands of MAGs and their associated metadata with minimal setup.

PANKEGG's contribution lies in KEGG pathway-centric exploration, which enables the assessment of metabolic capacities of MAGs, comparison of metabolic profiles across samples, and identification of key organisms for specific functional roles.

By organising annotations around KEGG orthologs and visualising completeness of pathways, PANKEGG facilitates hypothesis generation about community function and ecological adaptation.

6.4 Installation and Reproducibility

PANKEGG supports installation via pip, Conda, or pixi. It was tested across multiple platforms (Linux, Windows via WSL, macOS, and HPC environments), ensuring broad compatibility.
It uses lightweight dependencies for Python (with Flask, Pandas, Numpy, SciKit-Learn, SciPy) and also uses SQLite3 (embedded, so no external server is required).

Documentation and installation instructions are hosted on GitHub (https://github.com/RVanDamme/PANKEGG) to make it easily accessible to the broader community.

PANKEGG adheres to the FAIR principles by offering transparency, simplicity, and long-term reproducibility for metagenomic data interpretation (Reusable). The databases generated are lightweight, also for easy storage and sharing between people (Findable, Accessible and Interoperable)

6.5 Role in This PhD

PANKEGG was crucial to the downstream analysis of Papers IV. It allowed for structured, interpretable visualisation of:

- Pathway-level functional differences between samples from dry vs. wet seasons
- Quality curation and validation of MAGs based on GTDB criteria
- Detection of low-quality or potentially contaminated bins

This tool was invaluable when evaluating the rumen samples, as MUFFIN generated hundreds of bins across all samples. Instead of manually cross-referencing quality, taxonomy, and function, PANKEGG integrated the information, allowing for visual confirmation of patterns in microbial community structure.

PANKEGG's flexible design ensures it can also be reused in future metagenomic studies, especially for microbiomes involving environmental gradients or host physiological changes.

7. Paper III. Whole genome sequences of 70 indigenous Ethiopian cattle

7.1 The study

The work presented in this chapter was conducted as part of a collaborative effort to sequence and analyse the genomes of 70 individuals from seven indigenous Ethiopian cattle breeds. While not directly tied to the metagenomic study on Boran cattle (Paper IV), this genomic work is essential to the broader aims of this PhD: enabling integrative microbiome–host interaction research in future stages.

The study generated a high-quality dataset containing over 18 million high-confidence SNPs. These SNPs provide valuable insights into the genetic diversity, structure, and potential adaptive traits of indigenous cattle breeds (Abigar, Barka, Boran, Fellata, Fogera, Gojjam-Highland, and Horro). Population structure analyses confirmed the genetic uniqueness of each breed. In contrast, selection signature analyses identified genomic regions potentially involved in local adaptation to harsh Ethiopian environments, including those related to heat tolerance and resilience to feed scarcity.

Although the additional studies from the obtained dataset focused on breeds other than Boran, it played a foundational role in the PhD project. It allowed the development and validation of a robust, reproducible SNP calling and filtering pipeline suited to African indigenous cattle, which will now be directly applied to Boran individuals in future work. In particular, it prepares the next step of the Boran study: sequencing the host genome and exploring correlations between host genotype and microbiome composition and function.

In this way, the project bridges genomic and metagenomic methodologies, establishing one of the first large-scale genomic resources for Ethiopian Boran cattle. The approach is generalisable and will support future integration efforts with microbiome datasets in this PhD and beyond.

The resulting variant dataset contained high-confidence SNPs and indels distributed across the autosomes and sex chromosomes. The per-individual mean genome coverage ranged from 9.2 to 14.7 times. After filtering, the final VCF contained over 18 million variants.

Principal Component Analysis (PCA) confirmed breed-level genetic clustering, validating sample selection and highlighting population structure. Additional downstream analyses, including signature detection for selection and candidate gene identification, were performed by collaborators using this dataset.

7.2 Three different applied research studies

From the initial study, in which we retrieved over 18 million variants while working with seven different breeds, three applied studies have emerged. Dr. Wondossen Ayalew led these studies as part of his doctoral research. We collaborated on the research and developed methods of the 3 following papers. Those methods can be applied in the future to our study of host genome–microbiome interactions, using the same cattle used for the metagenomic research (Paper IV). These applied studies already yielded valuable insights into indigenous Ethiopian cattle breeds.

7.2.1 Abigar, Fellata, and Gojjam-Highland copy number variations reveal adaptation to diverse environments (Ayalew, Xiaoyun, Tarekegn, Tessema, et al., 2024)

This study identified 3,893 copy number variation regions (CNVRs) spanning 19.15 Mb (0.71% of the cattle genome). These CNVRs, ranging from 1.60 kb to 488.0 kb, included 1,713 deletions, 1,929 duplications, and 251 mixed events, with significant breed-specific differences. Validation by qPCR confirmed four of five randomly selected CNVRs. Key candidate genes associated with high-altitude adaptation (GBE1, SOD1), heat stress tolerance (HSPA13, DNAJC18, DNAJC8), and tick infestation resistance (BoLA, KRT33A) were identified. Variance stabilising transformation (VST) statistics underscored population-specific CNVRs, highlighting unique adaptive signatures in Gojjam-Highland cattle. Notably, 4.93% of CNVRs overlapped with quantitative trait loci (QTLs), implicating them in economically important traits such as growth and disease resistance.

7.2.2 Candidate genes related to milk production discovered in Barka cattle (Ayalew, Wu, et al., 2024)

The Barka cattle breed is renowned for its milk production in semi-arid conditions. We used three selective sweep methods (ZFST, $\theta\pi$ ratio, ZHp).

In this study, we identified three candidate genes consistently associated with milk production and composition traits: ACAA1, P4HTM, and SLC4A4.

The candidate genes show roles in critical biological pathways, including fatty acid metabolism, mammary gland development, and milk protein synthesis. The candidate genes are new potential genetic targets for selective breeding strategies aimed at improving milk productivity in tropical dairy cattle. However, we must validate through genome-wide association studies and transcriptomic analyses for practical breeding applications.

7.2.3 Selection signatures for local adaptation identified in Abigar cattle (Ayalew et al., 2023)

The Abigar cattle are highly adapted to the hot and humid climates of southwestern Ethiopia, significantly contributing to local livelihoods. This study presented the first whole-genome sequencing analysis for Abigar cattle, uncovering genes associated with heat tolerance (HOXC13, DNAJC18, RXFP2), immune responses (IRAK3, MZB1, STING1), and oxidative stress management (SLC23A1).

The genetic diversity assessments revealed high nucleotide diversity and heterozygosity, coupled with low inbreeding indicators. Those assessments demonstrate robust genetic health.

We also identified 83 shared genes linked to environmental adaptation that provide crucial insights for future breeding programs aimed at enhancing resilience to climate challenges. These findings provide a crucial foundation for understanding and harnessing adaptive genetic mechanisms in tropical cattle breeds.

8. Paper IV. Seasonal Dynamics of the Rumen Microbiota in Ethiopian Boran Cattle: a shotgun metagenomics study

This study investigated the seasonal dynamics of the rumen microbiome in Ethiopian Boran cattle by comparing samples collected during the dry season (Bega, February) and the rainy season (Kiremt, April). Using a hybrid metagenomic approach combining Illumina and Oxford Nanopore sequencing, we performed taxonomic profiling with Kraken2 and genome-resolved analyses through the MUFFIN pipeline to recover metagenome-assembled genomes (MAGs) and reconstruct metabolic pathways. Our hypothesis was that the microbiome would shift in composition and function between seasons, with fibre-degrading bacteria (e.g., *Fibrobacter*) becoming more abundant during the dry season as they are the bacteria able to digest the drier and sturdier composition of the feed. Methanogens and other clades would decrease due to the lack of feed to maintain the complex populations. Tetracycline and aminoglycoside antibiotic resistance genes (ARGs) should also be high, as the use of these antibiotics is widespread and not regulated.

8.1 Quality control

For the Illumina reads, the mean quality score was Q36 across all samples, with a minimum of 150 million reads per sample. Less than 1% of the reads mapped to the host genome, ARS-UCD2.0, and were removed.

For the Oxford Nanopore reads, the mean quality score was Q25 across all samples, with a number of reads ranging from 0.4 million to 2 million and an average read length between 2500 bp and 3500 bp. Around 10% of the reads mapped to ARS-UCD2.0 and were removed.

8.2 Kraken Classification Overview

Table 3 summarises classification rates and the number of taxa detected before and after filtering for Illumina sequencing reads.

Table 3: Illumina Reads Classification Statistics

Metric	Minimum (Sample)	Maximum (Sample)	Median
Classification rate (%)	18.23 (Feb_0350)	19.56 (Feb_0428)	18.94
Taxa detected (before filtering)	23,103 (April_0407)	29,277 (Feb_0350)	–
Taxa detected (after filtering)	2,202 (April_0407)	10,324 (Feb_0350)	–

Table 4 summarises classification rates and the number of taxa detected before and after filtering for Nanopore sequencing reads.

Table 4: Nanopore Reads Classification Statistics

Metric	Minimum (Sample)	Maximum (Sample)	Median
Classification rate (%)	72.61 (Feb_0350)	80.26 (April_0350)	76.00
Taxa detected (before filtering)	17,738 (Feb_0667)	20,762 (April_0350)	–
Taxa detected (after filtering)	1,032 (Feb_0667)	2,944 (April_0350)	–

Shared Taxa Between Platforms

When considering only taxa retained by both Illumina and Nanopore classifications for the same sample, counts ranged from 1,028 (Feb_0667) to 2,919 (April_0350), with a median of 1,894 taxa.

Human DNA Contamination

Human DNA was detected in every sample, with relative abundances ranging from 0.93% (April_0476) to 6.31% (Feb_0428). A minor contamination likely occurred during sample handling, which was removed for the MUFFIN analyses following the host removal protocol.

Taxonomic Resolution and Over-Classification

Table 5 summarises the number of taxa shared across all samples and those unique to each season. Notably, more genera than species are retained post-filtering. This trend reflects a common tendency for over-classification at the species level, where our read-count thresholds (at least 1000 reads classified for Illumina and 100 reads classified for ONT) remove many low-confidence species calls.

Table 5: Taxa shared across all samples and for each season

Taxonomical Rank	**Number of taxa common to all samples**	**Number of Taxa common to February**	**Number of Taxa common to April**
All ranks	906	916	1120
Phylum	29	29	32
Class	49	49	56
Order	103	103	115
Family	186	187	219
Genus	356	359	462
Species	183	189	236

8.2.1 Genus level

When we examine the top 10 most abundant genera between February and April, we see the disappearance of *Butyrivibrio* and *Paenibacillus* in favour of *Fibrobacter* and *Unclassified Methanobrevibacter* (Figure 13).

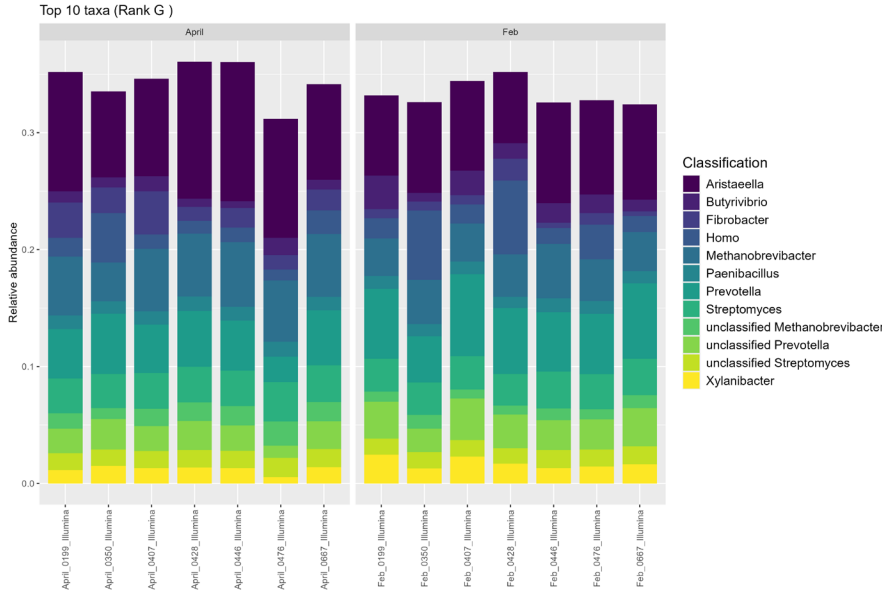

Figure 13: Top 10 genera found in the February and April samples based on their mean relative abundance.

This change is also reflected when we look at the genera with the highest variation in their centred Log-Ratio transformed relative abundance. In figure 14 we can see that from February to April, we gain in *Fibrobacter* and *Methanobrevibacter*. We lose the relative abundance in *Butyrivibrio* and human contamination and on a smaller degree we lose abundance in *Paenibacillus, Streptomyces* and *Prevotella*.

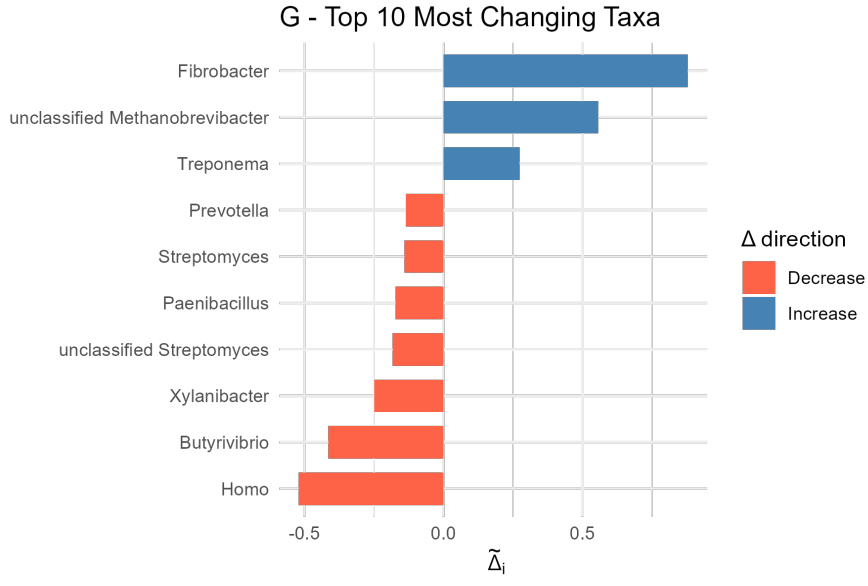

Figure 14: Top 10 phyla with the highest ΔCLR (where ΔCLR = April CLR - February CLR).

8.2.2 Species level

At the Species level, the depth of rank makes the changes in relative abundance more interesting. In figure 15, we see the disappearance of *Butyrivibrio fibrisolvens* in favour of *Methanobrevibacter millerae*.

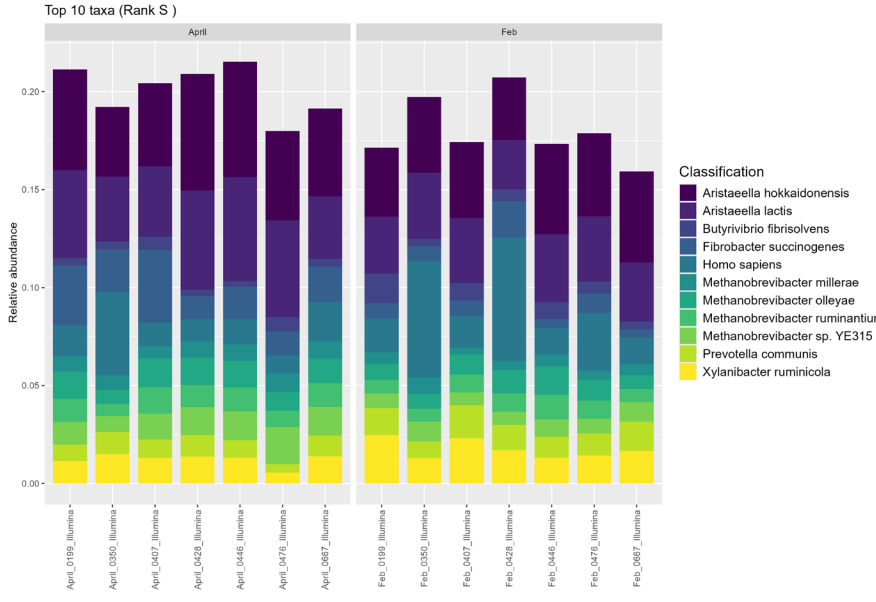

Figure 15: Top 10 species found in the February and April samples based on their mean relative abundance.

Figure 16 highlights the increase in *Fibrobacter succinogenes* and *Methanobrevibacter* sp. YE315, but also a decrease in Human contamination, and to a minor degree, two other *Methanobrevibacter species*.

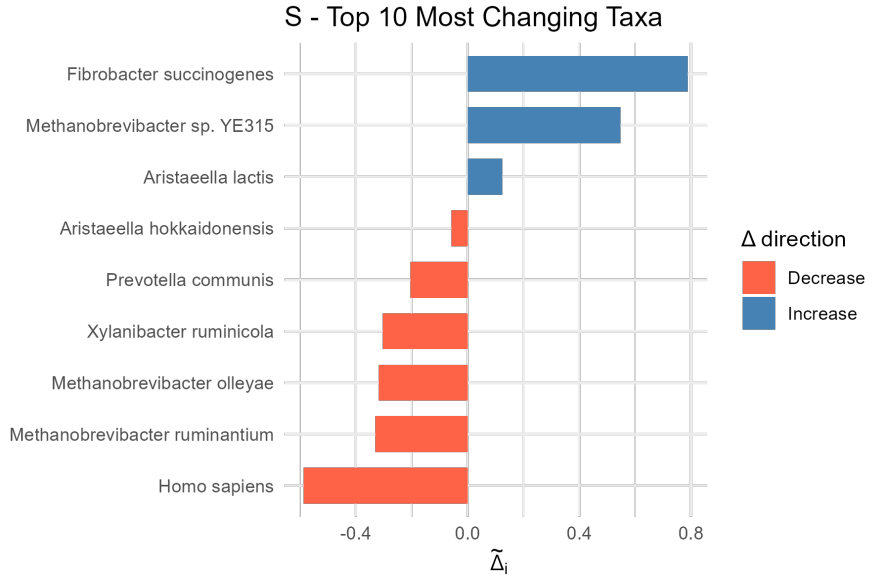

Figure 16: Top 10 species with the highest ΔCLR (where ΔCLR = April CLR - February CLR).

Throughout all the ranks, we see the same trend of changes, primarily the consequent increase of *Fibrobacter* and *Methanobrevibacter* and their higher classifications in the rainy season (April). Another change is the clear reduction of human contamination compared to the dry season (February) The results of the species rank should remain only an indication, as due to all the filtering steps, the data used for it represent only up to 45% of the total relative abundance calculated by the Kraken2 classification. Yet the results concur with what we see in higher ranks, where, after filtering, we kept 75% or more of the relative abundance. Two other interesting findings are the decrease in *Butyrivibrio* and *Prevotella* in the rainy season.

8.3 MUFFIN Results

Across all 14 samples, MUFFIN recovered a total of **656 MAGs**: 202 from February and 454 from April. Of these, 14 were classified as archaea, and 642 were classified as bacteria. Binning quality met the medium-quality MIMAG standard (completeness > 50%, contamination < 10%, and completeness − 5×contamination > 50%) (Bowers et al., 2017). Individual

bin completeness ranged from 50.7% to 100%, and contamination from 0% to 7.85%.

8.3.1 Archaeal MAGs

The archeal MAGs retrieved are exclusively methanogens and were not found in all the individuals. In Table 6, we list the archaeal MAGs identified in the dataset, indicating the samples and seasons where each was detected.

Table 6: Archaeal MAGs Detected Across Samples and Seasons.

Archaeal MAG	**Samples Detected**	**Seasons**
Methanosarcina mazei	0199, 0667	Both seasons
Methanoprimaticola sp. 015063165	0476, April0199	Both seasons (0476)
Methanosphaera sp. 016282985	Feb0407, Feb0476	February
Unclassified Methanobrevibacter	Feb0428	February
CADBMS01	April0199, April0350, April0407, April0667	April

8.3.2 Bacterial MAGs

Unresolved to Species

Four hundred thirty MAGs lacked species-level assignments. We therefore focus on genus and above.

Family-Level Only

Six MAGs had no genus classification. Four in the **Clostridia** class (order Christensenellales), two families: Christensenellaceae (3 MAGs) and Aristaeellaceae (1 bin). And one in **Anaerofustaceae** (Clostridia, order Eubacteriales) and **JAIRKQ01** (Verrucomicrobia, order Opitutales).

Candidate and "CAG/ISDG/RGIG/RUG/UBA" Genera

Those genera are a compendium of uncultured organisms found through metagenomics and belong to various classes and orders. Just for the *Rumen Uncultured Genomes* (RUG) the orders found in our samples are Lachnospirales, Oscillospirales, Selenomonadales, Christensenellales, Erysipelotrichales, Peptostreptococcales, Erysipelotrichales, Pirellulales, Bacteroidales, RF39, RFN20, Coriobacteriales.

Table 7 summarises the number of MAGs associated with candidate or uncultured clades, including the detailed counts for specific subgroups.

Table 7: MAGs Associated with Candidate and Uncultured Genera.

Candidate/Uncultured Genera	**Number of MAGs**	**Notes**
CAG	43	30 MAGs from CAG-791
ISDG	19	–
RGIG	38	15 MAGs from RGIG5612
RUG	108	14 from RUG11783, 15 from RUG11977, 11 from RUG14130
UBA	93	11 from UBA1258, 10 from UBA3857
UMGS1696	9	–
Cryptobacteroides	21	–
Pseudobutyrivibrio	5	–

Well-Characterized Genera

Through the analysis, 10 "well-characterized" genera were found. Those genera are defined by cultivated organisms and scientific literature in their functions.

The distribution in Table 8 reveals the high diversity of uncultured candidate lineages in the rumen microbiome, as well as the presence of numerous well-characterised genera. This diversity will enable us to conduct downstream functional analyses and study host–microbe interactions.

Table 8: Number of MAGs linked to well-known microbial genera.

Genus	**Number of MAGs**
Bulleidia	29
Prevotella	28
Chordicoccus	16
Ruminococcoides	14
Nanosynococcus	11
Fibrobacter	9
Saccharofermentans	7
Sodaliphilus	10
Streptococcus	10
Eubacterium (R, Q, G, S subclades)	10 (in total)

Seasonal Differences in Genus-Level Binning

To identify "genera unique to each season", we required that a genus appear in at least three samples from that season and only that season.

In February, only *Candidatus Liminaster* (3 samples) was found to follow those restrictions.

In April we found multiple genera to be unique to the season. Table 9 lists the microbial genera that were uniquely found in the April samples, along with the number of samples in which they appeared in

Table 9: Genera Unique to April Samples.

Genus	**Number of April Samples**
Fibrobacter	7 (all April samples)
Pseudobutyrivibrio	3
CADBMS01	4
G11	5
Physcousia	3
Porcincola	3
RGIG7111	4
RGIG7949	4
RUG420	4
RUG754	6
UBA2834	4
UBA2912	3
UBA3766	3

Predominant but Not Exclusive Genera

Some genera appeared in both seasons but were far more common in one than the other. Table 10 shows the number of samples in April and February in which specific genera were detected, highlighting those that were more common in one season.

Table 10: Genera Present in Both Seasons with Skewed Prevalence.

Genus	**April Samples**	**February Samples**
Chordicoccus	6	2
ISDG	7	1
Ruminococcoides	6	3
Sodaliphilus	6	1
Streptococcus	1	7

Several candidates and uncultured clades also showed seasonal bias. Genera more abundant in April than in February include *CAG-177*, *Limivicinus*, *RUG14130*, *RUG842*, *SFMI01*, *UBA1258*, *UBA3857*, and *UBA7702*. Conversely, *GA6A1*, *RGIG1955*, *TWA4*, and *Ventricola* were more common in February.

Equally Common Genera in Both Seasons

Several core rumen taxa maintained similar prevalence across seasons. Table 11 presents core rumen taxa that maintained relatively similar prevalence across February and April samples, as measured by the number of MAGs detected and the number of samples containing those MAGs.

Table 11: Genera Equally Common in Both Seasons.

Genus	**February (MAGs)**	**February (Samples)**	**April (MAGs)**	**April (Samples)**
Prevotella	11	7	17	7
Bulleidia	11	7	18	7
Methanosarcina	2	2	2	2
Saccharofermentans	3	3	4	3
Cryptobacteroides	6	5	15	7
Nanosynococcus	3	3	8	3
RGIG5612	6	4	9	7
CAG-791	11	6	19	7
RUG11783	7	7	7	7
RUG12372	5	5	4	4
UBA1367	3	2	2	2
UMGS1696	5	5	4	4

Notable Observation Streptococcus

Among Among well-characterised genera, Streptococcus stood out: it was present in every February sample, suggesting either a widespread colonisation at that time. Streptococcus equinus_B was found in all the February samples. Feb0350 also possessed two Streptococci that were unclassified at the species level.

Only one animal (April0446) continued to harbour Streptococcus equinus_B in the April sampling. At the time of sampling, none of the animals exhibited signs of disease; this suggests that the strain of Streptococcus equinus found is unlikely to play a significant role in the rumen microbiota or animal health. This strain is an interesting example of a seasonal contaminant possessing a risk to both animal and human health.

.

8.3.3 Metabolic Pathway Analysis

Using PANKEGG, we not only classify genome MAGs but also map their metabolic potential. For each bin (or sample), we identify all associated KEGG Orthologs (KOs) and compute a pathway completion score. This score equals the number of KOs detected in the bin divided by the total KOs known for that pathway (as defined in the KEGG database).

Across our dataset, we detected 399 distinct KEGG pathway maps. However, many of these are tangential or incidental. For example, pathways like “Morphine addiction,” “Thyroid cancer,” or “Parkinson's disease” may appear simply because they share one or two common KOs with more relevant metabolic routes. Such outliers are expected when screening the entire KEGG repertoire, since broad or niche pathways can overlap via shared orthologs.

For clarity, we will concentrate our discussion on four key categories of pathways to capture the core functional shifts and ecological impacts revealed by our metagenomic and MAG-based analyses. The Degradation (breakdown of complex molecules), the Biosynthesis (production of essential biomolecules), the Methanogenesis (methane production pathways) and the Antibiotic resistance mechanisms.

Starch and sucrose metabolism

The completeness of the starch and sucrose metabolism pathway within individual samples ranged from 46.23% to 62.26%. In February, most

samples showed completeness levels of 50% or below, except for two samples (Feb0350 at 62.26% and Feb0476 at 53.77%). In contrast, all samples from April had completeness above 55.66%.

No single bin contributed more than 25% to the pathway on its own. In February, the major contributors included *Prevotella* (~20%), *Streptococcus* (~20%), *Ventricola* (~20%), and bacteria from the Lachnospiraceae and Atopobiaceae families (~20%).

In April, the main contributors remained similar, with *Prevotella* (~18%) and bacteria from the Lachnospiraceae and Atopobiaceae families (up to ~25%), and *Fibrobacter* emerging as an additional significant contributor (~25%).

Most essential functions within the starch and sucrose metabolism were consistently maintained in both seasons. Specifically, critical processes such as the degradation of cellulose into D-glucose and the conversion of amylose and dextrin into D-glucose remained active across both seasons.

Biosynthesis of amino acids

The amino acid biosynthesis pathway showed apparent seasonal variation. Pathway completeness in February ranged from 59.83% (Feb0446) to 68.62% (Feb0350), while in April it was higher, ranging from 66.95% (April0428) to 74.90% (April0476).

No single bin stood out prominently, with most MAGs, from various taxa, contributing evenly between 20% and 40% to the overall pathway completeness.

When comparing samples with the lowest and highest completion rates, the primary differences observed were in the branches of lysine, tyrosine, and phenylalanine transformation.

Phenylalanine, tyrosine and tryptophan biosynthesis:

This biosynthesis pathway also exhibited seasonal differences. In February, pathway completeness ranged from 41.89% (samples Feb0446 and Feb0476) to 52.70% (sample Feb0199). In April, completeness was higher, ranging from 50.00% (April0350) to 62.16% (April0667).

A notable finding is the consistent appearance of the metabolic branch involving D-Fructose-1-phosphate and D-Fructose-1,6-diphosphate whenever *Methanosarcina mazei* was present, irrespective of the season.

Lysine biosynthesis

In this pathway, the variation in completeness does not follow a clear seasonal pattern and appears to vary significantly across samples. The lowest completeness observed was 41.67% (Feb0446), and the highest was 64.58% (April0199).

The primary variation lies in the processing of 2-Oxoglutarate within the Citrate pathway, as demonstrated in samples Feb0446 and April0199. This branch involves the transformation of acetyl-CoA and 2-oxoglutarate into L-lysine and Pyrrolysine. The branch is specifically populated by *Methanosarcina mazei*, *Methanoprimaticola* sp. 015063165, and the genus *UBA3766* (absent from Feb0446).

Valine, leucine and isoleucine biosynthesis

This pathway showed no noticeable seasonal variation. Pathway completeness ranged from 63.16% (sample April0407) to 73.68% (samples April0199 and April0476). Additionally, eleven samples (all seven from February and four from April) consistently had a completeness level of 68.42%.

Samples April0407, April0667, and all February samples lacked the K00263 enzyme (leucine dehydrogenase), which was present exclusively in the RUG754 bin within other April samples. Consequently, samples without leucine dehydrogenase depended entirely on the K00826 enzyme (branched-chain amino acid aminotransferase) to facilitate the biosynthesis reactions.

Arginine biosynthesis

Pathway completeness ranged from 36.67% in sample Feb0407 to 60.00% in sample April0476, with April consistently showing higher completeness.

Table 12 highlights the four enzymes identified as the main drivers of seasonal differences, including their functions and occurrence across samples.

Table 12: Arginine biosynthesis Key Enzymes Driving Seasonal Differences.

Enzyme (KEGG ID)	**Function**	**Occurrence February**	**Occurrence April**
Urease (K01427)	Converts urea into ammonia (NH_3) and bicarbonate (HCO_3^-)	0/7	7/7
Allophanate Hydrolase (K01457)	Hydrolyses urea-1-carboxylate into CO_2	0/7	1/7
Glutaminase (K01425)	Converts glutamine to NH_3	2/7	7/7
Glutamin-(asparagin)-ase (K05597)	Produces NH_3 from glutamine or asparagine	1/7	7/7

Valine, leucine and isoleucine degradation

Pathway completeness spanned from 26.09% (samples Feb0446 and Feb0667) up to 49.28% (sample April0476), revealing apparent seasonal variation. Table 13 lists the primary enzymes influencing metabolic differences, including their KEGG identifiers, EC numbers, roles, and occurrence patterns.

Table 13: Valine, leucine and isoleucine biosynthesis Key Enzymatic Drivers Identified in Samples.

Enzyme (KEGG ID; EC Number)	**Function**	**Occurrence February**	**Occurrence April**
2-Oxoisovalerate Dehydrogenase E1 Component (K11381; EC 1.2.4.4)	Catalyses the first step in breaking down valine, leucine, and isoleucine.	5/7	7/7
Acyl-CoA Dehydrogenase (K00249; EC 1.3.8.7)	Functions as an alternative to butyryl-CoA dehydrogenase (EC 1.3.8.1) during the dehydrogenation step	1/7	7/7
Leucine Dehydrogenase (K00263; EC 1.4.1.9)	Responsible for the reductive deamination of leucine.	5/7	0/7

Lysine degradation

Pathway completeness ranged from just 4.08% in sample Feb0428 up to 19.39% in sample Feb0350. In the lowest-completeness samples (Feb0428, Feb0476, Feb0667), the key orthologs required to convert L-β-lysine into acetyl-CoA were entirely missing.

In contrast, all other samples possessed at least a partial or complete degradation branch. A remarkably diverse set of taxa contributed to these branches, including (but not limited to): *UBA1205, UBA3857, UBA3792, UBA1217, UBA6987, RUG754, Colimorpha, Cacconaster, Sodaliphilus, Lentihominibacter, Egerieousia, CAG-791, Limimorpha, RGIG7150, RGIG5612, RGIG7949, Hornefia, F23-*

D06, Prevotella, Alectryocaccobium, Bilifractor, HGM13006, Chordicoccus, Faecousia, and Alloscillospira.

Fatty acid biosynthesis

This pathway displayed variability, with completeness ranging from 35.90% (sample April0407) to 46.15% (samples Feb0350 and April0476).

Table 14 lists the four enzymes identified as key factors influencing the observed differences in the data.

Table 14: Fatty acid biosynthesis Key Enzymes Underlying Observed Differences.

Enzyme Code (KEGG ID)	**Enzyme Name**	**Occurrence February**	**Occurrence April**
K00208	Enoyl-[acyl-carrier protein] reductase I	6/7	5/7
K10780	Enoyl-[acyl-carrier protein] reductase III	2/7	2/7
K15013	Long-chain-fatty-acid-CoA ligase	1/7	0/7
K18660	Malonyl-CoA/methylmalonyl-CoA synthetase	1/7	0/7

Fatty acid degradation

This pathway varied from a low completeness of 13.56% in sample Feb0667 to a higher completeness of 33.90% in sample April0476. Excluding Feb0667, all other samples achieved at least 18.64% completeness. On average, April samples exhibited slightly higher completeness than February, likely because *Fibrobacter*, one of the most significant individual contributors, accounted for around 10% of the pathway on its own.

In addition to *Fibrobacter*, several taxa played significant roles in β-oxidation of fatty acids (converting fatty acids to acetyl-CoA) and in the upstream processing of fatty alcohols and aldehydes into fatty acids. Key

contributors include *RGIG5612*, *Limivicinus*, *RUG754*, *Sodaliphilus*, and *Lentihominibacter*.

Methane metabolism

Pathway completeness varied widely, from a low of 24.10% in sample Feb0446 to a high of 60.51% in sample April0199. As expected, the presence and identity of methanogenic archaea strongly influenced these differences.

Table 15 summarises the relative contributions (in percentage) of different microbial taxa to pathway steps in various samples collected in April and February.

Table 15: Relative Contribution of Microbial Taxa to Methane Metabolism Pathway Steps Across Samples.

Taxa	**Contribution (%)**	**Samples February**	**Samples April**
Methanosarcina mazei	32–46	2/7	2/7
Methanosphaera sp. 016282985	~33	2/7	0/7
Unclassified Methanobrevibacter	~21	1/7	0/7
Methanoprimaticola sp. 015063165	~15	1/7	2/7
CADBMS01	~25	0/7	4/7
Fibrobacter	7-10	0/7	7/7

Coenzyme F_{420} biosynthesis

The methanogens (archaea) harbour the orthologs required for coenzyme F_{420} biosynthesis, a critical cofactor in methanogenesis. A bacterium (*Fibrobacter*) provides redundancy in the two following orthologs, K11780 and K11781.

Coenzyme M Biosynthesis

The initial, third, and fourth steps of coenzyme M production were catalysed by *Methanosphaera*, *Methanobrevibacter*, *CADBMS01*, and *Methanosarcina*. The second step's enzyme orthologs appeared not only in *Methanosphaera* and *CADBMS01* but also in a diverse set of bacterial taxa, including *UBA3857, UBA4181, RUG369, RUG695, RUG11795, RUG11797, CAG-791, Chordicoccus, Porcincola, Colimorpha, Aristaeella, Anaerobutyricum, Scatonaster, Colinaster, Bulleida, and Curtobacterium*.

Coenzyme B and Methanofuran Biosynthesis

Coenzyme B (from lysine) was biosynthesised by *Methanosarcina, Methanosphaera, Methanobrevibacter, Methanoprimaticola*, and *CADBMS01*.

Methanofuran steps were similarly handled by *Methanosarcina, Methanosphaera, Methanobrevibacter*, and *CADBMS01*.

Acetate-Based Methanogenesis and Serine Biosynthesis

Fibrobacter uniquely completed the **acetoclastic methanogenesis** reaction (module M00357) and was also responsible for serine biosynthesis in all April samples.

Sample-Specific Pathway Visualisations

Below are the full methanogenesis pathway maps for the samples with the lowest and highest completion.

In Feb0446, despite having the lowest completion rate, we can identify some of the key orthologs required for acetoclastic methanogenesis. What we observe is the presence of the orthologs necessary to transform acetate into acetyl-CoA (see Figure 17).

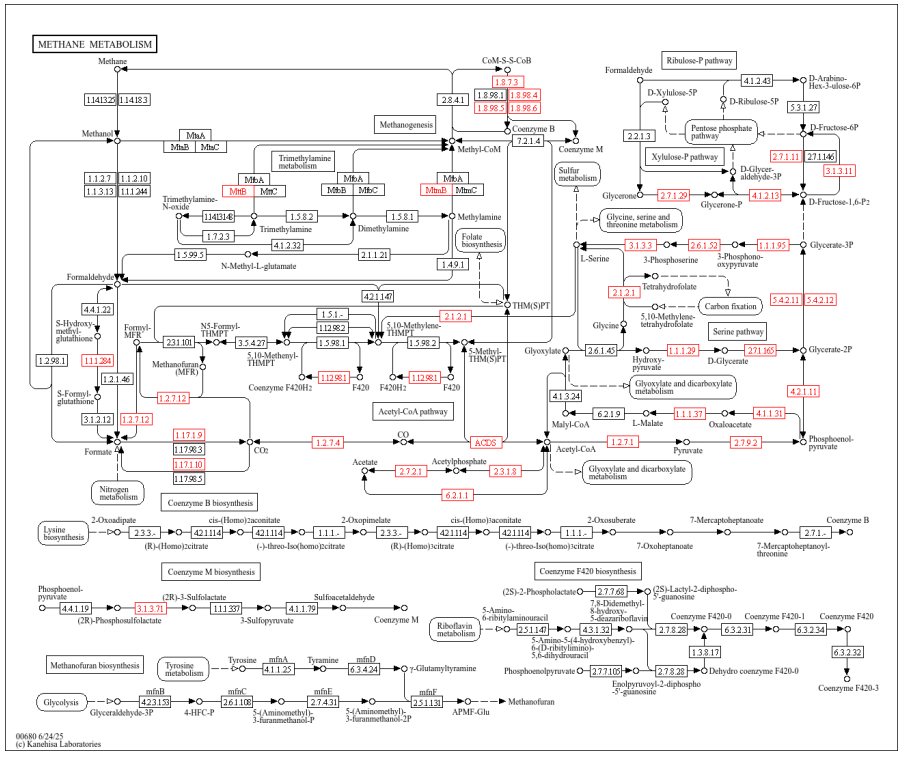

Figure 17: Methane metabolism pathway. In red, the KEGG orthologs found in all MAGs of Feb0446 are highlighted. Feb0446 is the sample with the lowest completion rate (24.10%).

In April0199, we observed the presence of all the necessary coenzymes for methanogenesis, and in addition, we found the orthologs required for the three different types of methanogenesis. *Methanosarcina mazei* alone covers most of the orthologs for the three different methanogenesis pathways, only complemented by *Methanoprimaticola* sp. 015063165 and some unknown species from *CADBMS01* (see Figure 18).

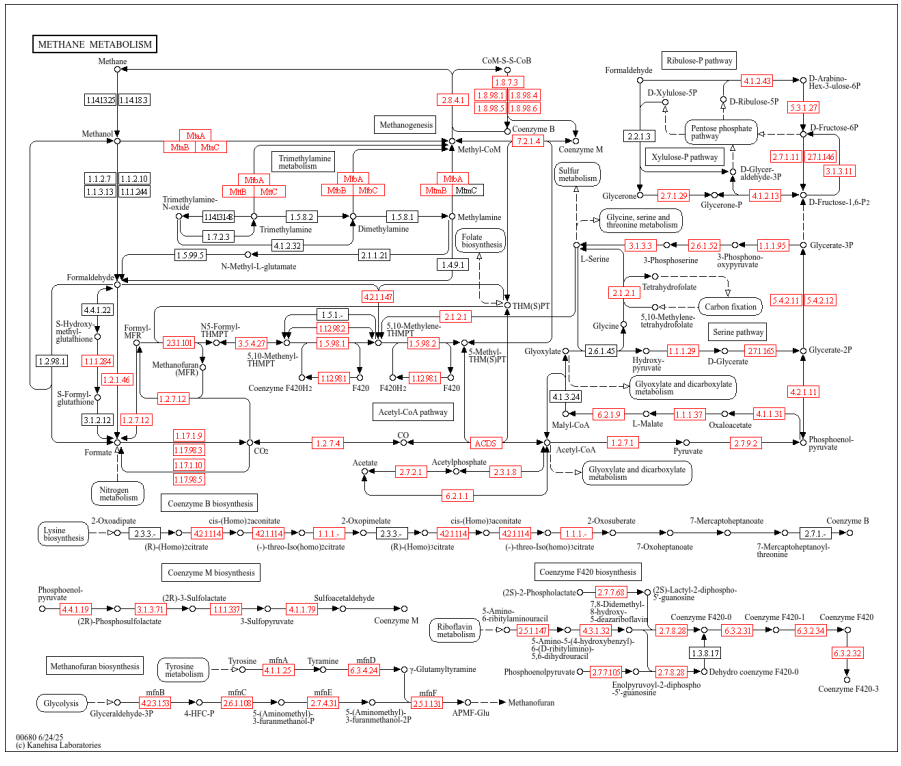

Figure 18: Methane metabolism pathway. In red, the KEGG orthologs found in all MAGs of April0199 are highlighted. April0199 is the sample with the highest completion rate (60.51%).

Antibiotics and resistance

Streptomycin biosynthesis

Streptomycin pathway completeness shows only a minor variation in April0350 with 47.62% complete where all the other samples are 52.38% complete

In April0350, the sole missing ortholog is K00844 (hexokinase), which is present in every other sample. Notably, none of the canonical streptomycin

biosynthesis genes were detected. Instead, the detected reactions convert D-glucose into scyllo-inosose and d-TDP-L-rhamnose.

Monobactam biosynthesis

Pathway completeness varied from a low of 17.86% in sample Feb0350 to a high of 35.71% in multiple samples (Feb0199, Feb0428, April0199, April0350, April0446, and April0476).

Sample Feb0350 lacks all orthologs required to convert sulfate into adenylyl sulfate. In contrast, this transformation is supported by a diverse set of MAGs in the other samples, including *Desulfovibrio, UMGS1696, Ga6A1, UBA4181, UBA3054, Physcousia, HGM12619, Pseudobutyrivibrio, Eubacterium_Q, Lentihominibacter, Porcincola, Anaerobutyricum, Bilifractor, RGIG8048, RUG11977, F23-D06, RUG708, RUG695, RUG369, W1P20-047*, and an unclassified *Fibrobacter* in April0667.

Additionally, samples April0350 and April0476 both possess the enzyme 4-hydroxymandelate oxidase (K16422; EC 1.1.3.46), which is present in *RGIG5612* for April0350 and in *RUG159* and *Blautia_A* for April0476. K16422 is absent in the other datasets, including samples with higher completeness.

In all samples, the pathway is missing the orthologs for synthetising nocardicin, monobactam, or β-lactam antibiotics. What we detect is the overlap with other metabolic networks. The L-aspartate to L-2,3,4,5-tetrahydrodipicolinate transformation is a key step in lysine biosynthesis. The sulfate activation and 4-hydroxymandelate oxidation intersect with sulfur and pyruvate metabolism, respectively. This indicate that there are no synthetisation of antibiotics.

beta-Lactam resistance

This pathway's completeness ranged from 18.75% in Feb0428 to 25.89% in Feb0350. The overall pathway completion across all samples is 28.57%. Despite the "low" coverage, multiple resistance genes were already detected.

Table 16 highlights the presence of key antibiotic resistance genes, their associated modules, and the genera or samples where they were detected.

Table 16: Methicillin and Beta-Lactam Resistance Genes Detected Across Samples.

Resistance module	**Gene (KEGG ID; EC Number)**	**Occurrence February**	**Occurrence April**	**Associated genera**
Methicillin resistance (MD: M00625)	mecA (K02545; penicillin-binding protein 2′, EC 3.4.16.4)	1/7	2/7	Unclassified *Christensenellaceae*
β-lactam resistance, Bla system (MD: M00627)	blaR1 (K02172; regulator protein BlaR1)	4/7	7/7	*RGIG7949*, *Saccharofermentans*, *RUG756*, *Anaerobutyricum*
	blaI (K02171; transcriptional regulator BlaI) penP (K17836; β-lactamase class A, EC 3.5.2.6)	7/7	6/7	*streptococcus equinus_B, Anaerobutyricum, RGIG7949*, *Evtepia*

Vancomycin resistance

Pathway completeness ranged from 54.55% (samples Feb0667 and Feb0428) up to 68.18% (samples Feb0350, April0199, April0407, April0446, April0476, and April0667). In the lower-completeness samples, several resistance operons were only partially present, whereas in the higher-completeness samples, most operons were intact. VanG, a less common S. coelicolor–type operon, and one accessory gene in the VanA cluster remained incomplete throughout all samples. Even so, every sample harboured at least seven known vancomycin-resistance genes, and some carried up to eleven.

Table 17 and Table 18 summarise the findings for different categories of resistance genes.

Table 17: Key Resistance Determinants – D-Ala-D-Lac Type (MD:M00651).

Gene (KEGG ID; EC Number)	**Occurrence February**	**Occurrence April**	**Host MAGs**
vanSB/S/D (K18345; EC 2.7.13.3)	1/7	7/7	Diverse MAGs
vanY (K07260; EC 3.4.17.14)	7/7	7/7	Diverse MAGs
vanW (K18346)	7/7	7/7	Diverse MAGs
vanH (K18347; EC 1.1.1.–)	1/7	0/7	*Colivicinus*
vanB/A/D (K15739; EC 6.1.2.1)	7/7	7/7	Diverse MAGs
vanX (K08641; EC 3.4.13.22)	7/7	7/7	Diverse MAGs

Table 18: Key Resistance Determinants – D-Ala-D-Ser Type (MD:M00652).

Gene (KEGG ID; EC Number)	Occurrence February	Occurrence April	Host MAGs
VanSC/E/G (K18350)	5/7	7/7	*Streptococcus equinus_B* and Diverse MAGs
vanRC/E/G (K18349)	4/7	7/7	Diverse MAGs
vanT (K18348; EC 5.1.1.18/5.1.1.1)	4/7	7/7	Diverse MAGs
vanC/E/G (K18856; EC 6.3.2.35)	7/7	7/7	Diverse MAGs
vanXY (K18866; EC 3.4.13.22/3.4.17.14)	6/7	7/7	Diverse MAGs

Cationic antimicrobial peptide (CAMP) resistance

Pathway completeness ranged from 22.22% in the lowest-scoring samples (Feb0476 and Feb0667) up to 46.30% in the highest-scoring sample (April0476). In February, completeness was consistently low (22.22–27.76%), whereas April samples showed higher variability (31.48–46.30%).

A key driver of increased completeness, and by extension, more resistance genes, in April was *Fibrobacter*. It contributed strongly to the Gram-negative resistance modules. Among Gram-positive contributors, *Sodaliphilus, Colivivens*, and *UBA1258* featured prominently in several samples.

Below are the different categories of CAMP resistance genes:

- dltABCD Operon (MD:M00725)

This operon mediates D-alanylation of cell-wall teichoic acids and is found in every sample. Table 19 lists the genes of the dlt operon, their functions, occurrence patterns, and the main taxa carrying them.

Table 19: D-Alanylation Operon (dltABCD) Distribution Across Samples.

Gene (KEGG ID; EC Number)	**Function**	**Occurrence February**	**Occurrence April**	**Associated Taxa**
dltA (K03367; EC 6.1.1.13)	D-alanine–poly(phosphoribitol) ligase subunit 1	7/7	7/7	*Streptococcus and diverse MAGs*
dltB (K03739)	Membrane protein for D-alanine export	4/7	7/7	*Streptococcus, Porcincola, Ornithomonoglobus, Evtepia, UBA9715, Faecousia, RGIG5952*
dltC (K14188; EC 6.1.1.13)	D-alanine–poly(phosphoribitol) ligase subunit 2	6/7	4/7	*Streptococcus, UBA1367, UBA1258, AC2028, UBA9715, Porcincola, Ruminococcoides*
dltD (K03740)	D-alanine transfer protein	7/7	2/7	*Streptococcus, UBA9715*

- Lysyl-Phosphatidylglycerol Synthase MprF (MD: M00726)

This enzyme modifies membrane lipids to repel CAMPs. mprF/fmtC (K14205; phosphatidylglycerol lysyltransferase, EC 2.3.2.3) was detected in Feb0199, Feb0428, Feb0667, April0199, April0350, April0407, April0428, April0476, and April0667. Hosts include *Desulfovibrio*, *Methanosarcina*, *Chordicoccus*, *Ruminimicrobiellum*, *UBA1367*, and *UBA2912*.

- VraFG Transporter (MD: M00730)

This ABC transporter pumps out cationic peptides. vraF (K19079; CAMP transport system ATP-binding protein) was only found in *UBA1258*

sp.016291405 in two April samples (April0199 and April0476).

Together, these resistance modules illustrate the combined roles of key taxa, especially *Fibrobacter* in April, in bolstering CAMP resistance across the rumen microbiome.

Multi-Drug Resistance Genes

The multidrug efflux systems collectively illustrate a robust network of transporters in the rumen microbiome, capable of extruding a wide array of toxic compounds and antibiotics, with varying representation across seasons and sample types. Table 20 summarizes the key multidrug efflux systems (MexAB–OprM, AdeABC, AcrEF–TolC, and AbcA/BmrA) detected across samples, including the main components, their occurrence, and associated genera.

Table 20: Efflux Pump Systems and Their Distribution.

Efflux System (Module)	**Gene (KEGG ID; Function)**	**Occurrence February**	**Occurrence April**	**Associated Taxa**
MexAB–OprM (MD: M00718)	acrA/mexA/adeI/smeD/mtrC/cmeA (K03585) – Membrane fusion component	7/7	7/7	*Prevotella, Fibrobacter, Cryptobacteroides, Desulfovibrio, Liminaster, Colivivens*, and others
	acrB/mexB/adeJ/smeE/mtrD/cmeB (K18138) – Inner-membrane transporter	4/7	7/7	*Cryptobacteroides, Fibrobacter, Desulfovibrio*
	oprM/emhC/ttgC/cusC/adeK/smeF/mtrE/cmeC/gesC (K18139) – Outer-membrane channel	1/7	0/7	*Desulfovibrio* sp. 016284885

AdeABC (MD: M00649)	adeR (K18144) – OmpR-family response regulator	2/7	0/7	*Limivicinus, Hominimerdicola*
	adeA (K18145) – Membrane fusion component	0/7	1/7	*Physcousia* sp. 902779315
AcrEF–TolC (MD: M00696)	tolC/bepC/cyaE/raxC/sapF/rsaF/hasF (K12340) – Outer-membrane channel	7/7	7/7	*Prevotella, Fibrobacter*, and diverse MAGs
	AcrEF–TolC homolog	0/7	1/7	*Physcousia* sp. 902779315
AbcA/BmrA (MD: M00700)	abcA/bmrA (K18104; EC 7.6.2.2) – ABC-type efflux pump	1/7	4/7	*Streptococcus, RUG754, RUG521, RUG13615*

9. Discussions

9.1 MUFFIN and PANKEGG

In 2021, metagenomic studies predominantly relied on Illumina sequencing, which necessitated considerable sequencing depth to produce reliable results. MUFFIN aimed to be an alternative by supplementing short-read Illumina sequencing with long-read nanopore sequencing, effectively addressing the inherent limitations of short-read assembly while preserving high per-base accuracy.

As detailed in Chapter 5 (Paper I), significant methodological advancements have occurred since the initial creation of MUFFIN.

The fundamental concepts of assembly and binning have remained consistent. But numerous components required changes or updates to align with the evolution of the bioinformatics practices. Notably, outdated tools were replaced, and inaccuracies identified through community feedback and validation studies were corrected.

The most substantial revision arose from issues associated with CheckM version 1, which depended on an outdated (2015) database to estimate genome completeness and contamination through marker genes. Although CheckM was initially considered a standard tool in 2020, the anticipated database update was delayed until 2023, with the release of CheckM version 2. The updated version, in addition to an update to the database, also significantly improved marker gene sets and estimation accuracy.

Another significant challenge involved the bin refinement step, which was initially implemented using the Metawrap pipeline. The idea was to leverage existing high-quality tools rather than creating new ones entirely. Unfortunately, the Metawrap refinement module became problematic due to its dependency on the obsolete CheckM database and the developers' subsequent abandonment of it. The lack of updates for over five years made the module unsuitable for integration in MUFFIN version 2. Consequently, we decided to remove bin refinement from MUFFIN altogether. Developing an in-house refinement module remains a possibility, but is not currently prioritised.

MUFFIN version 2 is still under development, prioritising the removal of outdated components (e.g., bin refinement, re-assembly steps, and obsolete binning methods) and incorporating essential updates. The upgrades include

compatibility enhancements allowing MUFFIN to process nanopore-only and Illumina-only datasets. Thus, MUFFIN is evolving into a versatile, broadly applicable pipeline, beyond its initial hybrid sequencing scope. Despite ongoing refinement, MUFFIN version 2 is functional in one of its analysis paths (Hybrid mode using metabat2) and was successfully employed in this thesis.

Parallel improvements were also implemented in PANKEGG. Initially, PANKEGG was a basic, inefficient parser generating minimalistic HTML outputs. The transition to MUFFIN version 2 was the spark that led to the overhaul of PANKEGG, transforming it into a fast and efficient parsing tool capable of generating portable servers for dynamic data visualisation. PANKEGG now facilitates easy navigation and comprehensive cross-referencing of results generated at various MUFFIN pipeline stages. Recognising its broader utility, we further expanded PANKEGG into a standalone visualisation tool supporting standard outputs from other common metagenomics tools, such as gtdb-tk classification.

Designing PANKEGG required careful consideration of user needs, informed by consultations with multiple researchers working in the field of metagenomics. It resulted in the current PANKEGG structure, which encompasses dedicated pages for MAGs, metabolic pathways, KEGG entries, taxonomic classifications, sample comparisons, and bin comparisons.

The web interface was optimised to minimise the number of clicks required to access essential data. Any information can be accessed by three to five clicks from the main page, thereby facilitating efficient and user-friendly data exploration.

Both MUFFIN and PANKEGG illustrate critical aspects of bioinformatics pipeline development, including the complexity of automating advanced analytical workflows, maintaining compatibility with and up-to-date methodologies, and refining existing pipelines to incorporate technological innovations without compromising the original objectives. These tools demonstrate a robust approach to clear, insightful visualisation and management of extensive metagenomic datasets.

The developments outlined in Papers I and II directly fulfil the second aim of this thesis, creating open-source, reproducible tools that enhance metagenomic analyses.

9.2 Genomic analysis

Chapter 7 (Paper III) emerged from a collaborative genomic sequencing project. Despite originating externally, this work aligns closely with the objectives of our Boran cattle microbiome study. Collaboration with the same research partners involved in sample collection for our microbiome analyses facilitated method validation and refinement. Although the genomic sequencing of the Boran cattle was delayed and thus not included in this thesis (see "Future Perspectives"), the genomic methodologies developed through this project laid essential groundwork for future host genome-microbiome correlation studies planned for the Boran breed. This initial collaboration has already inspired three applied studies exploring genetic adaptation and productivity traits in Ethiopian cattle breeds(Ayalew et al., 2023; Ayalew, Wu, et al., 2024; Ayalew, Xiaoyun, Tarekegn, Tessema, et al., 2024).

9.3 Metagenomic analysis

9.3.1 Kraken2 analysis

Using Nanopore reads to filter out the Kraken assignments helps overcome the low species- and strain-level specificity that arises from attempting to classify environmental sequences against an ever-incomplete reference database. Because many natural taxa aren't represented in Kraken's database, reads often "scatter" across the closest available relatives, especially within large and diverse clades, resulting in over-dispersed, low-confidence calls. By imposing read-count thresholds (1,000 reads for Illumina, 100 reads for Nanopore), we first collapse spurious low-support taxa and sharpen our classifications. Then, by retaining only those taxa seen by both platforms, where Nanopore's longer reads provide independent, higher-resolution evidence (in terms of the number of k-mers per read), we further boost precision. Of course, applying these filters reduces overall sensitivity (we discard some true but low-abundance taxa), but it greatly increases our confidence in the taxa we do report.

Notably, after filtering, the cumulative relative abundance of retained taxa remains high at broad ranks (Phylum through Genus). In each sample, after the filtering, we maintained at least 75% of the total number of classified reads. However, at the species level, this drops to around 40% (and

only 35% for sample April0667), reflecting both the inherent limitations of the reference database and the more stringent filtering required to achieve reliable species-level assignments.

We performed all quantitative abundance analyses using the Illumina data because every read is the same length, which avoids length-driven biases in count-based estimates. Although we explored incorporating Nanopore reads into our quantitative pipeline, we were unable to identify a robust approach that simultaneously accounts for both read count and read length. In principle, a weighting scheme that integrates the number of classified reads with their varying lengths could work, but no satisfactory method currently exists.

When comparing samples from February (the dry season) and April (the rainy season), several clear shifts become apparent. Notably, Fibrobacter and ruminococcoides, key fibre-degrading genera, appears only or majoritarly in April, reflecting improved (if still limited due to the drought) feed availability once the rains began. Methanobacteriales also increase in abundance during the rainy season, likely due to increased availability of substrate. Conversely, the Streptococcaceae decline sharply from February to April, revealing an asymptomatic bloom earlier in the year.

We likewise observe a drop in human DNA contamination between seasons, suggesting either changes in handling protocols or simply lower background contamination levels in the April sampling.

9.3.2 MUFFIN analysis

Although our MUFFIN binning recovered 656 total MAGs, only 14 were archaeal versus 642 bacterial. Despite, for example, *Methanobrevibacter* ranking among the top 20 taxa by relative abundance in our read-level classifications (∼5% of reads) for all samples, we did not retrieve it as MAGs in all samples. This discrepancy suggests archaeal MAGs are being under-recovered. Future work should explore newer, deep–learning–based binning tools (e.g., SemiBin2, ComeBin) trained specifically on archaeal genomes. We were unable to evaluate these methods here due to resource and compatibility constraints; however, an in-depth comparison is warranted.

Bin quality met the medium-quality MIMAG standard (completeness >50%, contamination <10%), but only 103 MAGs would qualify as high-quality (completeness >90%, contamination <5%). On a per-sample

basis, high-quality MAGs would number just 2 to 19 each, whereas including medium-quality MAGs raises that to 19 to 122 per sample.

A fascinating finding is how different methanogenic archaea partition not just seasonally but also functionally, reflecting shifts in substrate availability and community stability. *Methanosphaera* sp. 016282985 was detected solely in February's samples, and its MAGs encode exclusively the hydrogenotrophic pathway ($CO_2 + H_2 \rightarrow CH_4$). During the dry season, when plant-derived substrates may be more recalcitrant, hydrogen concentrations from fibre fermentation are perhaps relatively higher, while the concentration of acetate is lower. *Methanosphaera* fills a niche specialised for scavenging H_2. In contrast, the uncharacterized "*CADBMS01*" clade appears only in April and carries orthologs for both hydrogenotrophic and acetoclastic pathways ($CH_3COO^- \rightarrow CH_4 + CO_2$). The ability to harness acetate directly likely gives *CADBMS01* an advantage when wetter conditions boost the breakdown of complex carbohydrates into acetate, providing a richer pool of substrates than in February. Both these seasonally restricted taxa nonetheless carry the full suite of coenzyme biosynthesis genes necessary to drive their respective methanogenesis modules, underscoring their metabolic self-sufficiency.

By comparison, *Methanosarcina mazei* emerges as a "fixed" methanogen, detected in two samples and remaining present in the next season. Unlike the seasonally restricted taxa, *Methanosarcina mazei* encodes the full complement of hydrogenotrophic, acetoclastic, and methylotrophic (e.g., methanol- or methylamine-driven) methanogenesis pathways. This metabolic versatility likely underpins its stability in the face of environmental fluctuations. When feed input and fermentation products shift with the rains, *Methanosarcina mazei* can toggle among substrates, H_2, acetate, or simple methylated compounds to maintain methane production. The flexibility may also be an advantage compared to the competition, enabling persistence even when other methanogens dominate specific niches.

Together, these patterns indicate a dynamic methanogen landscape where specialist taxa emerge under distinct seasonal regimes, while generalists, such as *Methanosarcina mazei,* provide functional resilience. Future work could integrate metatranscriptomic or activity assays to quantify actual pathway usage in February versus April, validate which substrates drive in situ methane fluxes, and explore whether feed additives or management strategies could selectively suppress high-yield methanogens (e.g.,

Methanosarcina mazei) or displace acetoclastic specialists, such as *CADBMS01,* to mitigate greenhouse gas emissions.

A striking outcome of our MAG-based study is the sheer abundance of genomic MAGs assigned to uncultured, candidate, or entirely novel clades. These lineages sit as "dark matter" on the microbial tree of life. These candidate groups (the *CAG*s, *RUG*s, *RGIG*s, *UBA*s, and others) lack any cultured representatives, meaning we cannot yet validate their actual physiology or ecological roles by traditional isolation and laboratory experiments. In many cases, their inferred metabolisms are based solely on *in silico* annotations of draft genomes, leaving open the possibility that some assemblies may reflect chimeric MAGs or database biases.

Yet the repeated recovery of the same candidate clades across geographically and environmentally distinct rumen studies lends weight to their genuine existence. Even more compelling is the observation that specific candidate lineages, such as *CAG-791*, *UBA3857*, and *RUG754*, exhibit seasonally consistent patterns, suggesting they respond to genuine ecological drivers rather than random assembly artefacts.

We stand at a tipping point, however. Current metagenomic and binning pipelines, regardless of their sophistication, are still subject to limitations in read length, assembly algorithms, and taxonomic reference gaps. As single-cell genomics and long-read platforms continue to mature, we will soon be able to verify whether these candidate MAGs correspond to *bona fide* organisms by recovering complete genomes from individual cells, and also characterising their metabolic capabilities in isolation, or capturing their transcripts in environmental RNA studies. Until then, we should treat these novel clades as both a glimpse of hidden diversity and a reminder of the biases inherent in our methods. Their consistent detection across multiple studies, however, suggests that we are indeed beginning to chart the contours of microbial "terra incognita," setting the stage for targeted cultivation efforts and functional assays that could ultimately transform our understanding of rumen ecology and microbial ecology more broadly.

9.3.3 Well-characterised genera

Well-characterised bacterial genera like *Prevotella, Fibrobacter,* and *Butyrivibrio* serve as internal "controls" against which we can gauge the fidelity of both our Kraken2 read-level classifications and our MUFFIN binning. A few points to draw out:

Prevotella was among the top taxa in Kraken2 (often >10% relative abundance) and yielded the most significant number of high-quality MAGs (28 MAGs), reinforcing that these classifications reflect real, abundant populations.
Fibrobacter, which only appears in April, exhibits the same seasonal spike in Kraken2 counts and MUFFIN MAGs, providing excellent cross-validation of both methods.

Kraken2 often "splits" reads among multiple reference species within a genus, leading to dispersed species-level calls; yet at the genus rank, assignments for well-characterised clades remain robust (retaining >75% of reads).

For *Butyrivibrio*, Kraken2 indicated it as one of the top 10 most abundant species in February; however, MUFFIN produced only one *Butyrivibrio* MAG. This disparity suggests some degree of limitations in what MUFFIN can recover.

On the other hand, in the April samples, MUFFIN produced multiple *Pseudobutyrivibrio*. Yet, in Kraken2, *Pseudobutyrivibrio xylanivorans* represented less than 0.5% of the relative abundance in all samples and was the only specie identified. The gap between some of the Kraken2 findings and MUFFIN MAGs indicates a limitation in the capacity of Kraken2 to discover new species.

By enforcing ≥1,000 Illumina-reads and ≥100 Nanopore-reads cutoffs, nearly all well-characterised genera stay above threshold in both datasets. In contrast, spurious or very low-abundance taxa drop out, the continued presence of core clades after filtering highlights their genuine ecological importance.

When functional pathways are mapped back to MAGs, the fact that well-characterised clades (e.g., *Prevotella* in starch metabolism, *Fibrobacter* in cellulose breakdown) track neatly from Kraken2 abundance to MAG-derived gene content gives us extra confidence in linking taxa to function.

9.3.4 Cross-method validation & limitations

One of the most powerful checks on our taxonomic assignments comes from comparing read-level classifications (via Kraken2) to the genome-resolved MAGs generated by MUFFIN. As already stated, for many abundant, well-characterised genera, including *Prevotella*, *Fibrobacter*, and *Streptococcus*, we observe concordance: high relative abundance in Kraken2 corresponds to

numerous high-quality MAGs, and seasonal trends (e.g., the April emergence of *Fibrobacter*) appear in both datasets. This cross-validation bolsters our confidence that these signals reflect true community dynamics rather than method-specific artefacts. Conversely, discrepancies hint at limitations: for example, *Streptococcus*'s large Kraken2 footprint in February yielded relatively few MAGs per sample, suggesting either that short Illumina reads misassign ambiguously among related reference genomes or that the *Streptococcus* population is so genetically homogenous that assembly and binning collapse multiple strains into one. For the under-characterised genera, the analysis is much more challenging, as the Kraken2 database (PlusPF) and the GTDB database employ different nomenclature, which prevents an accurate comparison.

Applying minimum-read thresholds (1,000 reads for Illumina, 100 for Nanopore) shrinks the long tail of low-abundance, low-confidence taxa. Before filtering, raw Kraken2 classifications detect tens of thousands of taxa per sample; after filtering, that number falls to a few thousand. The filtering also reduces noise from spurious hits, especially at the species level, where reference gaps scatter reads among many near matches, but inevitably sacrifices sensitivity to rare taxa. In practice, these thresholds preserve over 75% of classified reads at broad ranks (Phylum through Genus), yet drop species-level coverage to ~40%. Balancing the desire for comprehensive community profiling against the risk of over-interpreting artefacts, for functional studies or ecological modelling, higher precision through stringent filtering would be preferable.

When mapping MAG-derived KEGG Orthologs, we initially identified nearly 400 pathway maps, ranging from core metabolic routes to seemingly bizarre "disease" pathways (e.g., Parkinson's, morphine addiction). These outliers arise because many KO entries are shared among diverse pathways, resulting in incidental hits that lack ecological relevance in the rumen. Without careful curation, one might misinterpret these signals as novel functions when they simply reflect overlapping enzyme annotations. Focusing on key functional categories (degradation, biosynthesis, methanogenesis, and antibiotic resistance) helped avoid such pitfalls; however, vigilance is still warranted, as incidental pathway detection remains a persistent caveat in large-scale annotation workflows.

Several methodological constraints can skew our binning results. Assemblies may fragment highly repetitive or GC-rich genomes, leading to underrepresentation of particular taxa in the final MAGs, even if Kraken2 detects their reads. Binning algorithms differ in their sensitivity to coverage variation and k-mer composition; tools not optimised for archaeal genomes (e.g., MetaBAT2) can under-recover methanogens despite their read-level abundance. Enzyme annotation pipelines likewise vary in how permissively they assign KO terms. Finally, uneven sequencing depth or DNA extraction biases (e.g., cell lysis efficiency) can distort perceived community composition before any computational step.

Using hybrid reads (Illumina and ONT) circumvents some of those constraints. The first is that, as the assembly uses long reads, we prevent fragmentation of highly repetitive or GC-rich genomes. Then, the short reads ensure that the quality at the base level remains satisfactory. During binning with Metabat2, using reads from the two different sequencing methods yields slightly different coverage depths for the contigs due to variations in sequencing depth and bias. That is key to improving the binning by providing different coverage information for the same sample.

To mitigate these limitations further, in our planned future studies, we should:

I. **Benchmark multiple binning tools**, especially those leveraging deep learning or tailored to archaeal genomes (SemiBin2, CoMetBin), to maximise MAG recovery across domains.

II. **Incorporate complementary data types** (e.g., metatranscriptomes, metabolomics) to confirm which detected pathways are actively expressed or realised in situ.

III. **Calibrate filtering strategies** using mock communities or spike-in standards, allowing for a quantitative assessment of sensitivity versus precision trade-offs.

By systematically addressing these biases, we will sharpen both taxonomic and functional insights, ensuring that our interpretations are grounded in a robust, multi-layered foundation.

9.3.5 Metabolic pathway analyses

I focused on the metabolic reconstructions on four major functional categories: degradation, biosynthesis, methanogenesis, and antibiotic resistance.

Degradation pathways

Starch & sucrose metabolism

Pathway completeness jumps from ~46–50% in February to ~56–62% in April, mirroring the rains' boost to readily fermentable carbohydrates. *Prevotella* and *Lachnospiraceae/Atopobiaceae* drive most of this activity year-round. Still, *Fibrobacter* emerges in April as a new contributor, highlighting its role in breaking down plant polymers when fresh forage increases. *Fibrobacter* on its own explain ~25% of the pathway across the samples.

No single genome bin contributes more than 25%, emphasising a distributed community effort rather than a single "super-degrader."

Fatty acid degradation

Completeness edges up from ~18–29% in February to ~20–34% in April. *Fibrobacter* again stands out in April, supplying ~10% of β-oxidation capacity, while *RGIG5612, Sodaliphilus,* and *Lentihominibacter* round out the core degraders.

The absence of key dehydrogenases (e.g., 2-Oxoisovalerate dehydrogenase) in low-completeness samples reveals a potential bottleneck that could limit energy harvesting from specific substrates. A lack of depth could also explain the failure to capture it.

Biosynthesis pathways

Amino acid biosynthesis

Overall pathway completeness rises from ~60–69% in February to ~67–75% in April.
The most significant seasonal delta occurs in lysine, tyrosine, and phenylalanine branches, likely reflecting shifts in nitrogen availability and microbial demand.

No individual MAG dominates, indicating functional redundancy and complementarity among taxa for these core anabolic routes, as multiple taxa cover the identical orthologs but also complement each other.

Branched-chain amino acid (Val/Leu/Ile) biosynthesis

Completeness is high (~63–74%), with no discernible seasonal pattern, suggesting that these essential pathways are core functions that are maintained across environmental fluctuations.

The absence of leucine dehydrogenase in all February samples (and two April samples) necessitates reliance on aminotransferases. Yet the completeness has only a little variation. This variation in pathways also implies how the different MAGs complement each other to maintain the functioning of the metabolic pathways in the rumen.

Arginine biosynthesis

Rises from ~37% in February to ~60% in April.
April-specific *Fibrobacter* MAGs carry urease and allophanate hydrolase, enabling them to tap urea and funnel it into arginine, highlighting niche specialisation under wetter conditions.

Methanogenesis

Module completeness categories (complete, incomplete, missing) vary widely across samples, ranging from ~24% to ~60%.
Methanosarcina mazei (32–46%) and *Methanosphaera* (33%) dominate hydrogenotrophic and methylotrophic steps; *CADBMS01* covers ~25% when present.

The acetoclastic module (M00357) is uniquely completed by *Fibrobacter* in April, revealing cross-domain cooperation between bacteria and archaea. Seasonal patterns track substrate availability: acetate-driven methanogenesis is stronger in April, while hydrogenotrophic routes persist in February.

Antibiotic biosynthesis & resistance

Streptomycin & monobactam biosynthesis show only minor variation (~47–53% completeness), but both pathways lack the canonical gene clusters. This lack of gene cluster suggests that the detected reactions likely reflect cross-pathway overlaps rather than true secondary metabolism in the rumen.

β-Lactam and vancomycin resistance reach ~19–68% completeness, with core determinants (mecA, blaR1, vanA/B/C/X) found in both seasons.

Streptococcus equinus B and *Anaerobutyricum* carry key β-lactamase genes in February, whereas *RGIG7949* and *Saccharofermentans* dominate in April, indicating shifts in the taxon composition of resistance reservoirs.

The CAMP resistance pathway increases from ~22–28% in February to ~31–46% in April, primarily due to Fibrobacter-mediated Gram-negative modules.

Multidrug efflux systems (MexAB–OprM, AcrEF–TolC, AdeABC, AbcA/BmrA) are ubiquitous but vary in carrier taxa, reflecting a baseline level of intrinsic resilience against antibiotics in the rumen microbiome.

The detection of antibiotic resistance genes (ARGs) across all rumen samples is a clear red flag for animal health. Moreover, it is also a concern for food safety and broader economic stability, especially in a country like Ethiopia, where livestock underpin both subsistence and commercial agriculture.

In our datasets, key resistance determinants were found in every sample, often carried by both well-characterised genera (e.g., *Streptococcus equinus*, *Anaerobutyricum*, *Ruminococcoides*) and uncultured candidate lineages. This ubiquity suggests that rumen microbiomes act as **reservoirs** for ARGs, which can transfer horizontally to potential pathogens (e.g., enteric bacteria), leading to treatment failures in common livestock infections (mastitis, respiratory or gastrointestinal diseases).
A 2020 survey of smallholder systems found that 86.7% of pastoralists and 24–95% of mixed-crop farmers routinely keep and use antibiotics, predominantly tetracyclines (36.4%), aminoglycosides (31.3%), and trimethoprim–sulfonamides, often under suboptimal storage conditions, with off-label human formulations shared between people and animals. MexAB–OprM and AcrEF–TolC found in our samples can directly reduce susceptibility to tetracycline and aminoglycosides. The use of other antibiotics would also be limited, as we identified β-lactamase (mecA, blaR1, blaI and penP) ARG and Vancomycin-resistance operons (vanSB/S/D,

vanY, vanW, vanH, vanB/A/D, vanX, vanSC/E/G, vanRC/E/G, vanT, vanC/E/G, and vanXY)(Gemeda et al., 2020).

In Ethiopia, it is estimated that ~84% of livestock farmers administer antibiotics to sick animals (cattle, sheep, goats, poultry), with widespread self-medication and inadequate veterinary oversight(Odey et al., 2024).

Ethiopia's livestock sector contributes approximately 19% of the country's GDP and supplies a significant share of protein through meat and dairy products. Rising ARG prevalence threatens herd productivity by limiting effective treatments, increasing morbidity or mortality, and forcing farmers to cull more animals. Resistant bacteria can also enter the food chain via milk or meat, or spread through environmental run-off, posing zoonotic risks to handlers and consumers(Kumar et al., 2020; Odey et al., 2024).

Reduced herd health directly translates into lower milk yields and weight gains, undermining household incomes and national export potential. With 36% of GDP tied to agriculture (including livestock), any systemic drop in productivity can ripple through Ethiopia's economy, impacting employment, trade balance, and food prices.

10. Conclusions

Severe drought conditions in Ethiopia resulted in the death of a significant number of sampled animals, while COVID-19 restrictions led to a two-year delay in obtaining research samples. Early work in the project was therefore focused on developing robust tools to ensure that once data became available, the analyses could be performed efficiently and reproducibly. This proactive work included the creation of MUFFIN and PANKEGG, both of which have now been validated and provide value not only for this study but also for broader metagenomic research, aligning with FAIR data principles.

The analysis phase also highlighted the resource-intensive and time-sensitive nature of bioinformatics. High-performance computing (HPC) was critical, particularly for memory-demanding tasks such as SPAdes assemblies (1–1.5 TB RAM per sample). The closure of UPPMAX added complexity, but computational needs were successfully met through the NAISS Dardel HPC and the IRD Itrop HPC infrastructures. These resources enabled the completion of all computational tasks, but the work necessary to complete the analysis with these resources emphasises the need for flexible and well-supported HPC infrastructures to handle fluctuating bioinformatics workloads.

MUFFIN is an innovative, versatile metagenomics pipeline that represents a valuable contribution to the scientific community, with broad utility, as exemplified by the research projects that utilise it. PANKEGG is a highly intuitive visualisation platform that significantly simplifies complex metagenomic data analysis. Both tools are open-source, user-friendly, and designed with reproducibility and accessibility in mind. I hope the tools will foster broader adoption and facilitate scientific discoveries.

The genomic study of Ethiopian cattle is a robust foundation for future integrative research efforts aimed at enhancing the understanding and breeding strategies of indigenous breeds. These studies not only advance scientific knowledge but also hold tangible promise for improving livestock welfare, productivity, and resilience to climate change in Ethiopia. The effective outreach and knowledge dissemination derived from this work could have a profound benefit for local farming communities, ensuring that animals not only survive but also thrive under increasingly challenging environmental conditions.

Our hypothesis about the fibre degraders was incorrect; the increased abundance of fibre degraders (e.g., *Fibrobacter, Ruminococcus*) occurred during the rainy season. This contradiction might indicate that the lack of feed and the heat stress during the drought influenced greatly the microbiome. Our hypothesis about the decline in methanogens in the dry season was partially correct; we indeed saw a lower proportion of methanogens in February (dry) compared to April (rainy), but through the binning, we also noticed that some methanogens were rare (2/7 samples) but a "fixed" (in both season) part of the microbiota (*Methanosarcina mazei*). We also observed that hydrogenotrophic specific methanogens appeared during the dry season (*Methanosphaera* sp. 016282985), as the feed produced less acetate; this organism would therefore thrive over acetoclastic methanogens. Through the pathway analysis, we also saw an increase in all key degradation and biosynthesis pathways in the rainy season compared to the dry season, indicating a "reactivation" of the digestive function of the microbiome when feed became more available.

Due to the extensive unregulated use of antibiotics in Ethiopia, we expected to see some ARGs for the antibiotics commonly used in Ethiopia (tetracycline, aminoglycosides). Unfortunately, we found a total of 31 ARGs showing resistance to β-lactam, Vancomycin, CAMP, and Multidrug efflux systems (more information is found in Chapter 8.2.3 - Antibiotics and Resistance).

In this metagenomic investigation of Ethiopian Boran cattle, we observed pronounced seasonal shifts in rumen microbial communities and their functional capabilities, patterns that likely influence animal productivity, resilience to environmental stress, and methane emissions. While our relatively small cohort prevents generalisations, our findings demonstrate the value of examining microbiome dynamics across different grazing seasons. Moving forward, we plan to expand this work to larger herds and integrate host performance data and methane levels to further enhance our understanding. Ultimately, harnessing these microbial insights in breeding and management programs will provide a tangible path toward more sustainable and efficient livestock systems.

Throughout my PhD journey, I have gained substantial expertise in both scientific methodologies and managing unforeseen research challenges. Equipped with these experiences, I feel prepared and resilient, ready to tackle similar obstacles in future scientific endeavours effectively.

11. Further perspective

As outlined earlier in this thesis, the presented work constitutes only one component of a more extensive, long-term research plan. Throughout this project, many unexpected challenges necessitated continual adaptation and methodological refinements. Despite these setbacks, numerous exciting opportunities and future directions remain to be explored.

At the time of writing this thesis, Ethiopian Boran cattle blood samples are queued for genomic sequencing. With this, we will do a detailed analysis of genetic variants and explore the potential interactions between these host genomic variants and the microbiome data. Establishing such integrative host-microbiome networks will significantly enhance our understanding of microbiome-driven traits, thereby aiding future selection strategies for resilience, productivity, and mitigating environmental impact.

In parallel to the Ethiopian project, another study is underway. It involves 60 cattle samples from three breeds (Afrikaner, Hereford, and Bonsmara) in South Africa, which are analysed for both rumen microbiome seasonal variation and their genome. This project mirrors the Ethiopian research design but expands its scope. The broader scope allows for a comparative assessment across breeds and geographic environments.

Metagenomic sequencing for these samples has been completed, and the genomic sequencing is in the queue alongside the Ethiopian samples. This larger-scale initiative presents an opportunity to validate and extend the analytical frameworks and tools developed during this PhD, facilitating broader applicability and scientific impact. The lessons learned from this thesis will also reinforce the methods, such as benchmarking other binning methods, incorporating complementary data, and systematically addressing known biases.

Looking even further ahead, we have submitted an application to the Swedish Farmers' Foundation for Agricultural Research (SLF) for a comprehensive metagenomic study involving Swedish cattle. This study aims to investigate the influence of the microbiome on methane emissions, leveraging an extensive dataset comprising 10,000 dairy cattle from farms equipped with methane sniffers for measuring emissions. Beyond the methane metrics, detailed phenotypic measurements will also be incorporated, allowing the integration of phenotype, genotype, and

metagenomic data from 500 animals, from which rumen samples will be sequenced.

This effort could represent a critical step toward incorporating metagenomic data into the estimation of breeding values (EBVs), potentially revolutionising cattle breeding strategies to optimise environmental sustainability, productivity, and animal welfare.

Collectively, these future projects represent a strategic progression in my research career, building upon the foundations laid during this PhD and pushing toward innovative, integrative approaches in livestock genetics and metagenomics. Through continued collaborations and methodological advancements, I look forward to contributing to a deeper scientific understanding and practical applications that enhance cattle productivity and sustainability globally.

12. Usage of Artificial Intelligence in the thesis

Given the rapid technological advancements in artificial intelligence (AI), it is essential to transparently articulate its applications in scientific research, emphasising cautious, informed usage rather than blind reliance. Indeed, the concept of Maslow's Hammer, where every problem seems like a nail to someone equipped only with a hammer, is particularly pertinent here. While powerful, AI and Large Language Models (LLMs) must be applied judiciously, recognising their strengths and inherent limitations.

In my work, AI and related technologies were employed thoughtfully in various contexts, detailed as follows:

Machine learning and deep learning in bioinformatics: Within the analytical pipelines used in this thesis, particularly MUFFIN, machine learning (ML) and deep learning (DL) methods have become indispensable. These technologies have significantly advanced critical bioinformatics tasks, notably Oxford Nanopore sequencing basecalling, quality control, and data refinement processes. Moreover, ML and DL have transformed downstream analytical steps such as binning and classification, improving accuracy, efficiency, and the depth of biological insights obtained.

LLM (ChatGPT) for code refinement: ChatGPT was employed as a coding assistant primarily for refining and restructuring existing code. Rather than generating code from scratch, ChatGPT's utility was in optimising code readability, debugging assistance, and suggesting best practices. Nonetheless, all suggestions were carefully vetted and manually validated to ensure their accuracy and suitability for the project's specific context.

LLM (ChatGPT) for writing structure and clarity: Similarly, ChatGPT provided valuable assistance in structuring my thesis. It served as an additional perspective, helping to restructure long or convoluted sentences and clarify logical flows. Importantly, I never directly requested ChatGPT to compose original text. Instead, its role resembled that of a critical reader, offering suggestions and improvements to pre-existing drafts, thus enhancing clarity and coherence without compromising originality.

LLM (ChatGPT) as a citation manager: Attempts to use ChatGPT as a citation manager proved ineffective due to its tendency to produce inaccurate references, mixing details or entirely fabricating DOIs. Consequently, I

reverted to reliable and dedicated citation management software, specifically Zotero, to ensure accurate and reliable references.

LLM (ChatGPT) for literature summarisation: Conversely, ChatGPT excelled at summarising research articles, significantly streamlining literature review tasks. It enabled efficient sorting through my extensive collections of papers accumulated over several years, quickly identifying relevant studies and summarising their key findings.

DALL·E 3 for image generation: AI-based image generation via DALL·E 3 was utilised creatively in my thesis and associated outputs. The thesis cover integrates a picture I took and edited with DALL·E 3. Additionally, the PANKEGG logo was fully designed using DALL·E 3. However, based on the original design of MUFFIN, created by Tanguy Desmarez. Beyond these instances, DALL·E 3 was also employed to generate backgrounds for presentations or serve as inspiration. Notably, aside from the specified instances (cover and PANKEGG logo), all other visuals, including plots and scientific illustrations, were either personally created or appropriately credited if sourced externally.

References

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., Lawley, T. D., & Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. *Nature*, *568*(7753), 499–504. https://doi.org/10.1038/s41586-019-0965-1

Alneberg, J., Bjarnason, B. S., Bruijn, I. de, Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), Article 11. https://doi.org/10.1038/nmeth.3103

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, *59*(1), 143–169. https://doi.org/10.1128/mr.59.1.143-169.1995

An, D., Dong, X., & Dong, Z. (2005). Prokaryote diversity in the rumen of yak (Bos grunniens) and Jinnan cattle (Bos taurus) estimated by 16S rDNA homology analyses. *Anaerobe*, *11*(4), Article 4. https://doi.org/10.1016/j.anaerobe.2005.02.001

Anderson, C. L., Sullivan, M. B., & Fernando, S. C. (2017). Dietary energy drives the dynamic response of bovine rumen viral communities. *Microbiome*, *5*(1), 155. https://doi.org/10.1186/s40168-017-0374-3

Andrade, B. G. N., Bressani, F. A., Cuadrat, R. R. C., Cardoso, T. F., Malheiros, J. M., de Oliveira, P. S. N., Petrini, J., Mourão, G. B., Coutinho, L. L., Reecy, J. M., Koltes, J. E., Neto, A. Z., R. de Medeiros, S., Berndt, A., Palhares, J. C. P., Afli, H., & Regitano, L. C. A. (2022). Stool and Ruminal Microbiome Components Associated With Methane Emission and Feed Efficiency in Nelore Beef Cattle. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.812828

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., & Wingett, S. (2012). *FastQC*.

Angellotti, M., Lindberg, M., Ramin, M., Krizsan, S. J., & Danielsson, R. (2025). Asparagopsis taxiformis supplementation to mitigate enteric methane emissions in dairy cows—Effects on performance and metabolism. *Journal of Dairy Science*, *108*(3), 2503–2516. https://doi.org/10.3168/jds.2024-25258

Application of biotechnology to nutrition of animals in developing countries. (n.d.). Retrieved 30 July 2025, from https://www.fao.org/4/t0423e/t0423e03.htm

Ayalew, W., Wu, X., Tarekegn, G. M., Sisay Tessema, T., Naboulsi, R., Van Damme, R., Bongcam-Rudloff, E., Edea, Z., Chu, M., Enquahone, S., Liang, C., & Yan, P. (2024). Whole Genome Scan Uncovers Candidate Genes Related to Milk Production Traits in Barka Cattle. *International Journal of Molecular Sciences*, *25*(11), Article 11. https://doi.org/10.3390/ijms25116142

Ayalew, W., Wu, X., Tarekegn, G. M., Sisay Tessema, T., Naboulsi, R., Van Damme, R., Bongcam-Rudloff, E., Edea, Z., Enquahone, S., & Yan, P. (2023). Whole-Genome Resequencing Reveals Selection Signatures of Abigar Cattle for Local Adaptation. *Animals*, *13*(20), Article 20. https://doi.org/10.3390/ani13203269

Ayalew, W., Xiaoyun, W., Tarekegn, G. M., Naboulsi, R., Sisay Tessema, T., Van Damme, R., Bongcam-Rudloff, E., Chu, M., Liang, C., Edea, Z., Enquahone, S., & Ping, Y. (2024). Whole genome sequences of 70 indigenous Ethiopian cattle. *Scientific Data*, *11*(1), 584. https://doi.org/10.1038/s41597-024-03342-9

Ayalew, W., Xiaoyun, W., Tarekegn, G. M., Tessema, T. S., Chu, M., Liang, C., Naboulsi, R., Van Damme, R., Bongcam-Rudloff, E., & Ping, Y. (2024). Whole-genome sequencing of copy number variation analysis in Ethiopian cattle reveals adaptations to diverse environments. *BMC Genomics*, *25*(1), 1088. https://doi.org/10.1186/s12864-024-10936-5

Badhan, A., Wang, Y., Terry, S., Gruninger, R., Guan, L. L., & McAllister, T. A. (2025). *Invited review*: Interplay of rumen microbiome and the cattle host in modulating feed efficiency and methane emissions. *Journal of Dairy Science*, *108*(6), 5489–5501. https://doi.org/10.3168/jds.2024-26063

Bayat, A. (2002). Bioinformatics. *BMJ*, *324*(7344), 1018–1022. https://doi.org/10.1136/bmj.324.7344.1018

Bergman, E. N. (1990). Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiological Reviews*, *70*(2), 567–590. https://doi.org/10.1152/physrev.1990.70.2.567

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., … Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, *35*(8), 725–731. https://doi.org/10.1038/nbt.3893

Brulc, J. M., Antonopoulos, D. A., Berg Miller, M. E., Wilson, M. K., Yannarell, A. C., Dinsdale, E. A., Edwards, R. E., Frank, E. D., Emerson, J. B., Wacklin, P., Coutinho, P. M., Henrissat, B., Nelson, K. E., & White, B. A. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(6), Article 6. https://doi.org/10.1073/pnas.0806191105

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta*, *2*(2), e107. https://doi.org/10.1002/imt2.107

Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2023). CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, *20*(8), 1203–1212. https://doi.org/10.1038/s41592-023-01940-w

Church, D. C. (1988). *The ruminant animal: Digestive physiology and nutrition*. Englewood Cliffs, NJ : Prentice-Hall.

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, *44*(D1), Article D1. https://doi.org/10.1093/nar/gkv1276

Comtet-Marre, S., Parisot, N., Lepercq, P., Chaucheyras-Durand, F., Mosoni, P., Peyretaillade, E., Bayat, A. R., Shingfield, K. J., Peyret, P., & Forano, E. (2017). Metatranscriptomics Reveals the Active Bacterial and Eukaryotic Fibrolytic Communities in the Rumen of Dairy Cow Fed a Mixed Diet. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.00067

Conteville, L. C., Silva, J. V. da, Andrade, B. G. N., Coutinho, L. L., Palhares, J. C. P., & Regitano, L. C. de A. (2024). Recovery of metagenome-assembled genomes from the rumen and fecal microbiomes of Bos indicus beef cattle. *Scientific Data*, *11*(1), 1385. https://doi.org/10.1038/s41597-024-04271-3

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Danielsson, R., Dicksved, J., Sun, L., Gonda, H., Müller, B., Schnürer, A., & Bertilsson, J. (2017). Methane Production in Dairy Cows Correlates with Rumen Methanogenic and Bacterial Community Structure. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.00226

De Coster, W., & Rademakers, R. (2023). NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics*, *39*(5), btad311. https://doi.org/10.1093/bioinformatics/btad311

Difford, G. F., Plichta, D. R., Løvendahl, P., Lassen, J., Noel, S. J., Højberg, O., Wright, A.-D. G., Zhu, Z., Kristensen, L., Nielsen, H. B., Guldbrandtsen, B., & Sahana, G. (2018). Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. *PLOS Genetics*, *14*(10), e1007580. https://doi.org/10.1371/journal.pgen.1007580

DRAM for distilling microbial metabolism to automate the curation of microbiome function | Nucleic Acids Research | Oxford Academic. (n.d.). Retrieved 30 July 2025, from https://academic.oup.com/nar/article/48/16/8883/5884738?login=false

Duguma, B., & Janssens, G. P. J. (2021). Assessment of Livestock Feed Resources and Coping Strategies with Dry Season Feed Scarcity in Mixed Crop–Livestock Farming Systems around the Gilgel Gibe Catchment, Southwest Ethiopia. *Sustainability*, *13*(19), Article 19. https://doi.org/10.3390/su131910713

Elhassani, M. E., Maisonnasse, L., Olgiati, A., Jerome, R., Rehali, M., Duroux, P., Giudicelli, V., & Kossida, S. (2021). Deep Learning concepts for genomics: An overview. *EMBnet.Journal*, *27*(0), Article 0. https://doi.org/10.14806/ej.27.0.990

Ethiopia—Situation Report, 10 Jan 2024 | OCHA. (2024, January 10). https://www.unocha.org/publications/report/ethiopia/ethiopia-situation-report-10-jan-2024

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Ferrer, M., Golyshina, O. V., Chernikova, T. N., Khachane, A. N., Reyes-Duarte, D., Santos, V. A. P. M. D., Strompl, C., Elborough, K., Jarvis, G., Neef, A., Yakimov, M. M., Timmis, K. N., & Golyshin, P. N. (2005). Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. *Environmental Microbiology*, *7*(12), Article 12. https://doi.org/10.1111/j.1462-2920.2005.00920.x

Fonseca, P. a. S., Lam, S., Chen, Y., Waters, S. M., Guan, L. L., & Cánovas, A. (2023). Multi-breed host rumen epithelium transcriptome and microbiome associations and their relationship with beef cattle feed efficiency. *Scientific Reports*, *13*(1), 16209. https://doi.org/10.1038/s41598-023-43097-8

Fregulia, P., Campos, M. M., Dhakal, R., Dias, R. J. P., & Neves, A. L. A. (2024). Feed efficiency and enteric methane emissions indices are inconsistent with the outcomes of the rumen microbiome composition. *Science of The Total Environment*, *949*, 175263. https://doi.org/10.1016/j.scitotenv.2024.175263

Gemeda, B. A., Amenu, K., Magnusson, U., Dohoo, I., Hallenberg, G. S., Alemayehu, G., Desta, H., & Wieland, B. (2020). Antimicrobial Use in Extensive Smallholder Livestock Farming Systems in Ethiopia: Knowledge, Attitudes, and Practices of Livestock Keepers. *Frontiers in Veterinary Science*, *7*. https://doi.org/10.3389/fvets.2020.00055

Global Rumen Census. (2025, July 31). https://globalresearchalliance.org/research/livestock/collaborative-activities/global-rumen-census/

Goldfarb, T., Kodali, V. K., Pujar, S., Brover, V., Robbertse, B., Farrell, C. M., Oh, D.-H., Astashyn, A., Ermolaeva, O., Haddad, D., Hlavina, W., Hoffman, J., Jackson, J. D., Joardar, V. S., Kristensen, D., Masterson, P., McGarvey, K. M., McVeigh, R., Mozes, E., … Murphy, T. D. (2025). NCBI RefSeq: Reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Research*, *53*(D1), D243–D257. https://doi.org/10.1093/nar/gkae1038

Greening, C., Geier, R., Wang, C., Woods, L. C., Morales, S. E., McDonald, M. J., Rushton-Green, R., Morgan, X. C., Koike, S., Leahy, S. C., Kelly, W. J., Cann, I., Attwood, G. T., Cook, G. M., & Mackie, R. I. (2019). Diverse hydrogen production and consumption pathways influence methane production in ruminants. *The ISME Journal*, *13*(10), 2617–2632. https://doi.org/10.1038/s41396-019-0464-2

GTDB - Genome Taxonomy Database. (n.d.). Retrieved 30 July 2025, from https://gtdb.ecogenomic.org/

Hanafy, R. A., Dagar, S. S., Griffith, G. W., Pratt, C. J., Youssef, N. H., & Elshahed, M. S. (2022). Taxonomy of the anaerobic gut fungi (Neocallimastigomycota): A review of classification criteria and description of current taxa. *International Journal of Systematic and Evolutionary Microbiology*, *72*(7), 005322. https://doi.org/10.1099/ijsem.0.005322

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669–685. https://doi.org/10.1128/mmbr.68.4.669-685.2004

Harmon, D. L., & Swanson, K. C. (2020). Review: Nutritional regulation of intestinal starch and protein assimilation in ruminants. *Animal*, *14*, s17–s28. https://doi.org/10.1017/S1751731119003136

Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., & Goddard, M. E. (2009). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution*, *41*(1), 51. https://doi.org/10.1186/1297-9686-41-51

Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., & Janssen, P. H. (2015). Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Scientific Reports*, *5*. https://doi.org/10.1038/srep14567

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., Chen, F., Zhang, T., Mackie, R. I., Pennacchio, L. A., Tringe, S. G., Visel, A., Woyke, T., Wang, Z., & Rubin, E. M. (2011). Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science*, *331*(6016), Article 6016. https://doi.org/10.1126/science.1200387

Hook, S. E., Wright, A.-D. G., & McBride, B. W. (2010). Methanogens: Methane Producers of the Rumen and Mitigation Strategies. *Archaea*, *2010*(1), 945785. https://doi.org/10.1155/2010/945785

Hristov, A. N., Callaway, T. R., Lee, C., & Dowd, S. E. (2012). Rumen bacterial, archaeal, and fungal diversity of dairy cows in response to ingestion of lauric or myristic acid1. *Journal of Animal Science*, *90*(12), 4449–4457. https://doi.org/10.2527/jas.2011-4624

Interpreting EBVs and Indexes. (n.d.). Signet Breeding. Retrieved 30 July 2025, from https://signetdata.com/technical/ebvs-for-commercial-herds/interpreting-ebvs-and-indexes/

Islam, M., Kim, S.-H., Son, A.-R., Ramos, S. C., Jeong, C.-D., Yu, Z., Kang, S. H., Cho, Y.-I., Lee, S.-S., Cho, K.-K., & Lee, S.-S. (2021). Seasonal Influence on Rumen Microbiota, Rumen Fermentation, and Enteric Methane Emissions of Holstein and Jersey Steers under the Same Total Mixed Ration. *Animals*, *11*(4), Article 4. https://doi.org/10.3390/ani11041184

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*, e7359. https://doi.org/10.7717/peerj.7359

Kebede, H. Y., Mekonnen, A. B., Emiru, N. C., Mekuyie, M., & Ayal, D. Y. (2024). Climate variability and indigenous adaptation strategies by Somali pastoralists in Ethiopia. *Theoretical and Applied Climatology*, *155*(8), 7259–7273. https://doi.org/10.1007/s00704-024-04993-9

Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M., & McCue, L. A. (2020). ATLAS: A Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinformatics*, *21*(1), 257. https://doi.org/10.1186/s12859-020-03585-4

Krizsan, S. J., Ramin, M., Chagas, J. C. C., Halmemies-Beauchet-Filleau, A., Singh, A., Schnürer, A., & Danielsson, R. (2023). Effects on rumen microbiome and milk quality of dairy cows fed a grass silage-based diet supplemented with the macroalga Asparagopsis taxiformis. *Frontiers in Animal Science*, *4*. https://doi.org/10.3389/fanim.2023.1112969

Kumar, S. B., Arnipalli, S. R., & Ziouzenkova, O. (2020). Antibiotics in Food Chain: The Consequences for Antibiotic Resistance. *Antibiotics*, *9*(10), Article 10. https://doi.org/10.3390/antibiotics9100688

Lagier, J.-C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., Levasseur, A., Rolain, J.-M., Fournier, P.-E., & Raoult, D. (2018). Culturing the human microbiota and culturomics. *Nature Reviews Microbiology*, *16*(9), 540–550. https://doi.org/10.1038/s41579-018-0041-0

Lagier, J.-C., Hugon, P., Khelaifia, S., Fournier, P.-E., La Scola, B., & Raoult, D. (2015). The Rebirth of Culture in Microbiology through the Example of Culturomics To Study Human Gut Microbiota. *Clinical Microbiology Reviews*, *28*(1), 237–264. https://doi.org/10.1128/cmr.00014-14

Langmead, B., Wilks, C., Antonescu, V., & Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, *35*(3), 421–432. https://doi.org/10.1093/bioinformatics/bty648

Lau, J. T., Whelan, F. J., Herath, I., Lee, C. H., Collins, S. M., Bercik, P., & Surette, M. G. (2016). Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome Medicine*, *8*(1), 72. https://doi.org/10.1186/s13073-016-0327-7

Leger, A. (2025). *A-slide/pycoQC* [Python]. https://github.com/a-slide/pycoQC (Original work published 2017)

Legrand, T. P. R. A., Alexandre, P. A., Wilson, A., Farr, R. J., Reverter, A., & Denman, S. E. (2025). Genome-centric metagenomics reveals uncharacterised microbiomes in Angus cattle. *Scientific Data*, *12*(1), 547. https://doi.org/10.1038/s41597-025-04919-8

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., & Ettema, T. J. G. (2021). Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology*, *19*(4), 225–240. https://doi.org/10.1038/s41579-020-00458-8

Li, F., Li, C., Chen, Y., Liu, J., Zhang, C., Irving, B., Fitzsimmons, C., Plastow, G., & Guan, L. L. (2019). Host genetics influence the rumen microbiota and heritable rumen microbial features associate with feed efficiency in cattle. *Microbiome*, *7*(1), 92. https://doi.org/10.1186/s40168-019-0699-1

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), Article 18. https://doi.org/10.1093/bioinformatics/bty191

Li, J., Zhong, H., Ramayo-Caldas, Y., Terrapon, N., Lombard, V., Potocki-Veronese, G., Estellé, J., Popova, M., Yang, Z., Zhang, H., Li, F., Tang, S., Yang, F., Chen, W., Chen, B., Li, J., Guo, J., Martin, C., Maguin, E., … Morgavi, D. P. (2020). A catalog of microbial genes from the bovine rumen unveils a specialized and diverse biomass-degrading environment. *GigaScience*, *9*(6), Article 6. https://doi.org/10.1093/gigascience/giaa057

Li, Y., Mayberry, D., Jemberu, W., Schrobback, P., Herrero, M., Chaters, G., Knight-Jones, T., & Rushton, J. (2023). Characterizing Ethiopian cattle production systems for disease burden analysis. *Frontiers in Veterinary Science*, *10*. https://doi.org/10.3389/fvets.2023.1233474

Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J., & Crosby, L. (2018). Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems*, *3*(5), 10.1128/msystems.00055-18. https://doi.org/10.1128/msystems.00055-18

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Lopes, D. R. G., de Souza Duarte, M., La Reau, A. J., Chaves, I. Z., de Oliveira Mendes, T. A., Detmann, E., Bento, C. B. P., Mercadante, M. E. Z., Bonilha, S. F. M., Suen, G., & Mantovani, H. C. (2021). Assessing the relationship between the rumen microbiota and feed efficiency in Nellore steers. *Journal of Animal Science and Biotechnology*, *12*(1), 79. https://doi.org/10.1186/s40104-021-00599-7

Manyike, J. Z., Taruvinga, A., & Akinyemi, B. E. (2025). Mapping the research landscape of livestock adaptation to climate change: A bibliometric review using Scopus database (1994–2023). *Frontiers in Climate*, *7*. https://doi.org/10.3389/fclim.2025.1567674

Martín, N., Coleman, L., López-Villalobos, N., Schreurs, N., Morris, S., Blair, H., McDade, J., Back, P., & Hickson, R. (2021). Estimated Breeding Values of Beef Sires Can Predict Performance of Beef-Cross-Dairy Progeny. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.712715

Martinez-Fernandez, G., Jiao, J., Padmanabha, J., Denman, S. E., & McSweeney, C. S. (2020). Seasonal and Nutrient Supplement Responses in Rumen Microbiota Structure and Metabolites of Tropical Rangeland Cattle. *Microorganisms*, *8*(10), Article 10. https://doi.org/10.3390/microorganisms8101550

Marturano, A. (2012). Bioinformatics and Ethics. In R. Chadwick (Ed.), *Encyclopedia of Applied Ethics (Second Edition)* (pp. 278–285). Academic Press. https://doi.org/10.1016/B978-0-12-373932-2.00023-5

Mathieu, A., Leclercq, M., Sanabria, M., Perin, O., & Droit, A. (2022). Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.811495

Matthews, C., Crispie, F., Lewis, E., Reid, M., O'Toole, P. W., & Cotter, P. D. (2019). The rumen microbiome: A crucial consideration when optimising milk and meat production and nitrogen utilisation efficiency. *Gut Microbes*, *10*(2), 115–132. https://doi.org/10.1080/19490976.2018.1505176

MEKURIAW, G., & KEBEDE, A. (2015). *A review on indigenous cattle genetic resources in Ethiopia: Adaptation, status and survival. Online J. Anim. Feed Res., 5(5): 125-137. 5*(5:125-137).

Mi, J., Jing, X., Ma, C., Shi, F., Cao, Z., Yang, X., Yang, Y., Kakade, A., Wang, W., & Long, R. (2024). A metagenomic catalogue of the ruminant gut archaeome. *Nature Communications*, *15*(1), 9609. https://doi.org/10.1038/s41467-024-54025-3

Millardcrystal. (2021). *English: Ruminant Digestive System* [Graphic]. Own work. https://commons.wikimedia.org/wiki/File:Ruminant_digestive_system.png

Moharrery, A., Larsen, M., & Weisbjerg, M. R. (2014). Starch digestion in the rumen, small intestine, and hind gut of dairy cows – A meta-analysis. *Animal Feed Science and Technology*, *192*, 1–14. https://doi.org/10.1016/j.anifeedsci.2014.03.001

Monteiro, H. F., Figueiredo, C. C., Mion, B., Santos, J. E. P., Bisinotto, R. S., Peñagaricano, F., Ribeiro, E. S., Marinho, M. N., Zimpel, R., da Silva, A. C., Oyebade, A., Lobo, R. R., Coelho Jr, W. M., Peixoto, P. M. G., Ugarte Marin, M. B., Umaña-Sedó, S. G., Rojas, T. D. G., Elvir-Hernandez, M., Schenkel, F. S., … Lima, F. S. (2024). An artificial intelligence approach of feature engineering and ensemble methods depicts the rumen microbiome contribution to feed efficiency in dairy cows. *Animal Microbiome*, *6*(1), 5. https://doi.org/10.1186/s42523-024-00289-5

Monteiro, H. F., Zhou, Z., Gomes, M. S., Peixoto, P. M. G., Bonsaglia, E. C. R., Canisso, I. F., Weimer, B. C., & Lima, F. S. (2022). Rumen and lower gut microbiomes relationship with feed efficiency and production traits throughout the lactation of Holstein dairy cows. *Scientific Reports*, *12*(1), 4904. https://doi.org/10.1038/s41598-022-08761-5

Moss, E. L., Maghini, D. G., & Bhatt, A. S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature Biotechnology*, *38*(6), 701–707. https://doi.org/10.1038/s41587-020-0422-6

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T. B. K., Nielsen, T., Kirton, E., Faria, J. P., Edirisinghe, J. N., Henry, C. S., … Eloe-Fadrosh, E. A. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), 499–509. https://doi.org/10.1038/s41587-020-0718-6

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., & Rasmussen, S. (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*, *39*(5), 555–560. https://doi.org/10.1038/s41587-020-00777-4

Noel, S. J., Attwood, G. T., Rakonjac, J., Moon, C. D., Waghorn, G. C., & Janssen, P. H. (2017). Seasonal changes in the digesta-adherent rumen bacterial communities of dairy cattle grazing pasture. *PLOS ONE*, *12*(3), e0173819. https://doi.org/10.1371/journal.pone.0173819

Odey, T. O. J., Tanimowo, W. O., Afolabi, K. O., Jahid, I. K., & Reuben, R. C. (2024). Antimicrobial use and resistance in food animal production: Food safety and associated concerns in Sub-Saharan Africa. *International Microbiology*, *27*(1), 1–23. https://doi.org/10.1007/s10123-023-00462-x

Owens, F. N., Zinn, R. A., & Kim, Y. K. (1986). Limits to Starch Digestion in the Ruminant Small Intestine1,2. *Journal of Animal Science*, *63*(5), 1634–1648. https://doi.org/10.2527/jas1986.6351634x

Pan, S., Zhao, X.-M., & Coelho, L. P. (2023). SemiBin2: Self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics*, *39*(Supplement_1), i21–i29. https://doi.org/10.1093/bioinformatics/btad209

Park, T., Ma, L., Gao, S., Bu, D., & Yu, Z. (2022). Heat stress impacts the multi-domain ruminal microbiota and some of the functional features independent of its effect on feed intake in lactating dairy cows. *Journal of Animal Science and Biotechnology*, *13*(1), 71. https://doi.org/10.1186/s40104-022-00717-z

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), Article 7. https://doi.org/10.1101/gr.186072.114

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, *2*(11), Article 11. https://doi.org/10.1038/s41564-017-0012-7

Peraza, P., Fernández-Calero, T., Naya, H., Sotelo-Silveira, J., & Navajas, E. A. (2024). Exploring the Linkage Between Ruminal Microbial Communities on Postweaning and Finishing Diets and Their Relation to Residual Feed Intake in Beef Cattle. *Microorganisms*, *12*(12), Article 12. https://doi.org/10.3390/microorganisms12122437

Perez, H. G., Stevenson, C. K., Lourenco, J. M., & Callaway, T. R. (2024). Understanding Rumen Microbiology: An Overview. *Encyclopedia*, *4*(1), Article 1. https://doi.org/10.3390/encyclopedia4010013

Pryce, J. E., & Daetwyler, H. D. (2011). Designing dairy cattle breeding schemes under genomic selection: A review of international research. *Animal Production Science*, *52*(3), 107–114. https://doi.org/10.1071/AN11098

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), Article 9. https://doi.org/10.1038/nbt.3935

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria* [Computer software]. https://www.R-project.org/

Ramin, M., Chagas, J. C. C., Pal, Y., Danielsson, R., Fant, P., & Krizsan, S. J. (2023). Reducing methane production from stored feces of dairy cows by Asparagopsis taxiformis. *Frontiers in Sustainable Food Systems*, *7*. https://doi.org/10.3389/fsufs.2023.1187838

Riesenfeld, C. S., Schloss, P. D., & Handelsman, J. (2004). Metagenomics: Genomic Analysis of Microbial Communities. *Annual Review of Genetics*, *38*(Volume 38, 2004), 525–552. https://doi.org/10.1146/annurev.genet.38.072902.091216

Roehe, R., Dewhurst, R. J., Duthie, C.-A., Rooke, J. A., McKain, N., Ross, D. W., Hyslop, J. J., Waterhouse, A., Freeman, T. C., Watson, M., & Wallace, R. J. (2016). Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance. *PLOS Genetics*, *12*(2), e1005846. https://doi.org/10.1371/journal.pgen.1005846

Ross, E. M., & Hayes, B. J. (2022). Metagenomic Predictions: A Review 10 years on. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.865765

Russell, J. B., Muck, R. E., & Weimer, P. J. (2009). Quantitative analysis of cellulose degradation and growth of cellulolytic bacteria in the rumen. *FEMS Microbiology Ecology*, *67*(2), 183–197. https://doi.org/10.1111/j.1574-6941.2008.00633.x

Sáenz, J. S., Rios-Galicia, B., & Seifert, J. (2025). Antiviral defense systems in the rumen microbiome. *mSystems*, *10*(2), e01521-24. https://doi.org/10.1128/msystems.01521-24

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*(1), 87. https://doi.org/10.1186/s12915-014-0087-z

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D. T., Tett, A., Morrow, A. L., & Segata, N. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods*, *13*(5), 435–438. https://doi.org/10.1038/nmeth.3802

Sendeku, A. T., Kumar, D., Abegaz, S., & Mekuriaw, G. (2016). *Evaluations of Reproductive Performances of Fogera Cattle Breed in Selected Districts of Amhara Region, Ethiopia*. *5*(1).

Seshadri, R., Leahy, S. C., Attwood, G. T., Teh, K. H., Lambie, S. C., Cookson, A. L., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Hadjithomas, M., Varghese, N. J., Paez-Espino, D., Perry, R., Henderson, G., Creevey, C. J., Terrapon, N., Lapebie, P., Drula, E., Lombard, V., Rubin, E., … Kelly, W. J. (2018). Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nature Biotechnology*, *36*(4), 359–367. https://doi.org/10.1038/nbt.4110

Shaukat, S. S., Rao, T. A., & Khan, M. A. (2016). Impact of sample size on principal component analysis ordination of an environmental data set: Effects on eigenstructure. *Ekológia (Bratislava)*, *35*(2), 173–190. https://doi.org/10.1515/eko-2016-0014

Shi, W., Moon, C. D., Leahy, S. C., Kang, D., Froula, J., Kittelmann, S., Fan, C., Deutsch, S., Gagic, D., Seedorf, H., Kelly, W. J., Atua, R., Sang, C., Soni, P., Li, D., Pinares-Patiño, C. S., McEwan, J. C., Janssen, P. H., Chen, F., … Rubin, E. M. (2014). Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. *Genome Research*, *24*(9), 1517–1525. https://doi.org/10.1101/gr.168245.113

Stewart, E. J. (2012). Growing Unculturable Bacteria. *Journal of Bacteriology*, *194*(16), 4151–4160. https://doi.org/10.1128/jb.00345-12

Stewart, R. D., Auffret, M. D., Warr, A., Walker, A. W., Roehe, R., & Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnology*, *37*(8), Article 8. https://doi.org/10.1038/s41587-019-0202-3

Stewart, R. D., Auffret, M. D., Warr, A., Wiser, A. H., Press, M. O., Langford, K. W., Liachko, I., Snelling, T. J., Dewhurst, R. J., Walker, A. W., Roehe, R., & Watson, M. (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-03317-6

Tamames, J., & Puente-Sánchez, F. (2019). SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.03349

THE 17 GOALS | Sustainable Development. (n.d.). Retrieved 30 July 2025, from https://sdgs.un.org/goals

Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, *5*(1), 170203. https://doi.org/10.1038/sdata.2017.203

Ungerfeld, E. M. (2020). Metabolic Hydrogen Flows in Rumen Fermentation: Principles and Possibilities of Interventions. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.00589

Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, *6*(1), Article 1. https://doi.org/10.1186/s40168-018-0541-1

User:SUM1, F. adm location map svg: U. work: (2017). *English: Clickable map of the regions and zones of Ethiopia.* [Graphic]. Created from File:Ethiopia adm location map.svg by User:NordNordWest. https://commons.wikimedia.org/wiki/File:Map_of_zones_of_Ethiopia.svg

Wakaso, A. A., Mummed, Y. Y., & Yesuf, Y. K. (2025). Examining Ethiopia's live animal and meat value chain. *Heliyon*, *11*(1), e41752. https://doi.org/10.1016/j.heliyon.2025.e41752

Wallace, R. J., Rooke, J. A., McKain, N., Duthie, C.-A., Hyslop, J. J., Ross, D. W., Waterhouse, A., Watson, M., & Roehe, R. (2015). The rumen microbial metagenome associated with high methane production in cattle. *BMC Genomics*, *16*(1), 839. https://doi.org/10.1186/s12864-015-2032-0

Wallace, R. J., Sasson, G., Garnsworthy, P. C., Tapio, I., Gregson, E., Bani, P., Huhtanen, P., Bayat, A. R., Strozzi, F., Biscarini, F., Snelling, T. J., Saunders, N., Potterton, S. L., Craigon, J., Minuti, A., Trevisi, E., Callegari, M. L., Cappelli, F. P., Cabezas-Garcia, E. H., … Mizrahi, I. (2019). A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. *Science Advances*, *5*(7), eaav8391. https://doi.org/10.1126/sciadv.aav8391

Wang, L., Zhang, G., Xu, H., Xin, H., & Zhang, Y. (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*. https://doi.org/10.3389/fmicb.2019.00649

Wang, Z., You, R., Han, H., Liu, W., Sun, F., & Zhu, S. (2024). Effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nature Communications*, *15*(1), 585. https://doi.org/10.1038/s41467-023-44290-z

Waters, S. M., Roskam, E., Smith, P. E., Kenny, D. A., Popova, M., Eugène, M., & Morgavi, D. P. (2025). International Symposium on Ruminant Physiology: The role of rumen microbiome in the development of methane mitigation strategies for ruminant livestock *. *Journal of Dairy Science*, *108*(7), 7591–7606. https://doi.org/10.3168/jds.2024-25778

Weimer, P. J. (2015). Redundancy, resilience, and host specificity of the ruminal microbiota: Implications for engineering improved ruminal fermentations. *Frontiers in Microbiology*, *6*. https://www.frontiersin.org/article/10.3389/fmicb.2015.00296

Wiggans, G. R., & Carrillo, J. A. (2022). Genomic selection in United States dairy cattle. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.994466

Wilkinson, T., Korir, D., Ogugo, M., Stewart, R. D., Watson, M., Paxton, E., Goopy, J., & Robert, C. (2020). 1200 high-quality metagenome-assembled genomes from the rumen of African cattle and their relevance in the context of sub-optimal feeding. *Genome Biology*, *21*(1), 229. https://doi.org/10.1186/s13059-020-02144-7

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Xie, Y., Sun, H., Xue, M., & Liu, J. (2022). Metagenomics reveals differences in microbial composition and metabolic functions in the rumen of dairy cows with different residual feed intake. *Animal Microbiome*, *4*(1), 19. https://doi.org/10.1186/s42523-022-00170-3

Xue, M.-Y., Xie, Y.-Y., Zhong, Y., Ma, X.-J., Sun, H.-Z., & Liu, J.-X. (2022). Integrated meta-omics reveals new ruminal microbial features associated with feed efficiency in dairy cattle. *Microbiome*, *10*(1), 32. https://doi.org/10.1186/s40168-022-01228-9

Zewde, M. M., Mustefa, W. S., Zewde, M. M., & Mustefa, W. S. (2022). Review on milk production performance, challenges, and opportunities of dairy cows production in oromia regional state, Ethiopia. *International Journal of Veterinary Science and Research*, *8*(3), 080–085. https://doi.org/10.17352/ijvsr.000118

Zhang, X., Wang, W., Wang, Y., Cao, Z., Yang, H., & Li, S. (2024). Metagenomic and metabolomic analyses reveal differences in rumen microbiota between grass- and grain-fed Sanhe heifers. *Frontiers in Microbiology*, *15*. https://doi.org/10.3389/fmicb.2024.1336278

Popular science summary

Cattle are central to Ethiopian livelihoods, with approximately 70 million heads, mostly on small farms, utilising indigenous breeds that are adapted to the country's dramatic dry and rainy seasons. Inside each cow's rumen lives a bustling community of microbes that break down grass into energy for milk and meat, but some of these tiny organisms also produce methane, a greenhouse gas.

In this work, we tracked how the rumen microbiome shifts between the dry season and the lush rainy season. One standout discovery was *Fibrobacter*, a champion fibre-eater that only shows up when the rains bring fresh forage. Its arrival boosts the rumen's capacity to digest starches and sugars, potentially helping cows gain more weight and produce more milk. On the other hand, we found that antibiotic-resistance genes are present, potentially conferring resistance to penicillin, vancomycin, cationic peptides, and even multidrug. Their presence raises red flags: if resistant bacteria spread from cows to people or contaminate meat and milk, common infections in livestock could become much harder to treat.

We also saw the rumen's metabolic "toolbox" reconfiguring with the seasons. During the rainy period, pathways that build amino acids and fatty acids run at peak efficiency, reflecting the richer diet. In the dry season, methane-making routes powered by hydrogen dominate, suggesting cows might emit more methane per bite of feed. It also means that for the same amount of feed between dry and rainy seasons, more energy would be lost in the production of methane during the dry season. It also means that total emissions are higher during the rainy season, as more feed is consumed; however, emissions per kilogram of milk or meat produced are lower, as feed efficiency is higher, enabling the animals to produce more milk and build their bodies.

Why does this matter? By understanding which microbes rise and fall with the seasons and which genes they carry, we can begin to develop more effective feeding and breeding strategies. Imagine designing diets that favour fibre-digesters without fuelling methane spikes, or monitoring antibiotic use to curb resistance? That's the promise of a microbiome-informed approach: healthier cows, safer food, and lower environmental impact for Ethiopia's vital cattle sector.

Populärvetenskaplig sammanfattning

Kor är avgörande för människors försörjning i Etiopien – landet har omkring 70 miljoner djur, mestadels på små gårdar, där inhemska raser är anpassade till landets dramatiska torr- och regnperioder. I varje kos våm finns ett mylleråv mikrober som bryter ner gräs till energi för mjölk och kött, men vissa av dessa mikroskopiska organismer bildar också metan, en kraftig växthusgas.

I denna studie följde vi hur våmmens mikrobiom förändras mellan torrperioden och den frodiga regntiden. En viktig upptäckt är att *Fibrobacter*, vilka är bakterier som ofta är specialiserade på att bryta ned fibrer, bara dyker upp när regnet ger färskt foder. Deras närvaro ökar våmmens förmåga att bryta ner stärkelse och socker, vilket kan hjälpa kor att gå upp i vikt och ge mer mjölk. Under både regn- och torrperiod fann vi även gener för antibiotikaresistens mot penicilliner, vankomycin, katjoniska peptider och till och med så kallade multidrug-pumpar. Detta är oroväckande: om resistenta bakterier sprids från kor till människor eller förorenar kött och mjölk kan vanliga infektioner hos djur bli mycket svårare att behandla.

Vi såg även att våmmens metabola "verktygslåda" ställs om med årstiderna. Under regnperioden går processer som bygger aminosyror och fettsyror på högvarv, vilket speglar en näringsrikare kost. Under torrperioden dominerar metanbildande processer som är beroende av tillgång på väte, det innebär att kornas fodereffektivitet minskar under torrperioden. Under regnperioden producerar kor mer metangas då korna konsumerar mer foder men samtidigt innebär högre fodereffektivitet sannolikt att utsläppen per kilogram producerad mjölk eller kött blir lägre.

Varför är detta viktigt? Genom att förstå vilka mikrober och gener som växlar med säsongerna kan vi utveckla smartare foderscheman och avelsstrategier. Tänk dig dieter som gynnar fibernedbrytare utan att driva upp metan, eller noggrann kontroll av antibiotikaanvändning för att bromsa resistens. Det är löftet med en mikrobiomstyrd strategi: friskare kor, säkrare mat och mindre miljöpåverkan för Etiopiens livsviktiga boskapsnäring.

Acknowledgements

This thesis was carried out at the Department of Animal Biosciences (HBIO) at the Swedish University of Agricultural Sciences (SLU). Over 7 years ago, I arrived in Sweden for a Bachelor's internship, and that would have been it if not for the many people I will acknowledge for supporting me one way or another during my PhD studies:

Monsieur Professeur Erik Bongcam-Rudloff, you have been doing something extraordinary for many years now, welcoming Belgian students for their bachelor's internship and transforming them into skilled bioinformaticians in just 5 months. I was fortunate to be one of them, and then you gave me the opportunity to stay and learn even more from you, which has continued to this day, as I am now completing my PhD studies. I don't know what you saw in me that led you to make this initial proposal, but I can only hope I was good enough to satisfy you with this decision.

Tomas Klingström, we're not done, but thanks for everything so far and for what we already have ahead. You have taught me a great deal about what makes people good humans and good scientists. Your perspective is an invaluable lesson that I strive to follow.

Juliette Hayer, while you had to leave for France early in my PhD, thank you for our amazing discussions, your positivity and great feedback. Although I couldn't knock on your door as I used to, I could always reach you when I needed to.

Getinet Mekuriaw Tarekegn, thank you very much for your support, feedback and help through my PhD. You and everyone above know how challenging things were, but you made that a non-issue and always brought great solutions when I only saw the problem.

Leila, thank you for your support and help in the lab. I learned a great deal from you in the lab, and I hope I was able to teach you some bioinformatics in return.

I would also like to thank everyone at SLU who made my PhD experience great: **Lotta, Cano, Valeriia, Laingshun, Anahit, Adrien, Claire, Leonie, Nanxing, Hector, Kristina, Julie, Juan, Christian, Thomas Eliasson, Eliska, and Sallam.**

I would also like to express my gratitude to the **MEDBIOINFO** graduate school, the **Epicass/CassavaNet4Dev team,** and **SLUBI** for their support, the work we accomplished, and the knowledge I gained from them.

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

I acknowledge the ISO 9001 certified IRD ITrop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that contributed to the research results reported in this thesis. URL: https://bioinfo.ird.fr/- http://www.southgreen.fr

I would like to thank **Melania, Gabriele, Anna, and George** for their friendship and great DnD games. **Lea** and **Markos,** thank you for being amazing friends from the beginning of my adventure in Sweden and for years to come.

Merci **Charlotte** d'avoir vécu le début de cette aventure avec moi.

Merci **Benjamin** et **Antoine** de l'avoir continuée.

Merci **Thomas** de finir cette aventure avec moi ; je te souhaite tout le meilleur de ce que j'ai vécu.

À **David, Aline** et **Françoise**, merci pour toutes les opportunités que vous m'avez données et pour tout ce que vous m'avez enseigné. Sans vous, c'est sûr, je ne serais pas là.

À **Doris** et **Frédéric**, merci de m'avoir guidé dans cette voie depuis le début.

À **Philippe, Dominique** et **Mathilde**, merci d'avoir toujours été là.

Tout n'a pas toujours été tout rose, mais on a toujours fait avec, ensemble, et grâce à vous, je suis toujours vivant. Rassurez-vous : toujours la banane, toujours debout.

I

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN)

Renaud Van Damme[1,2]*, **Martin Hölzer**[3], **Adrian Viehweger**[3,4], **Bettina Müller**[1], **Erik Bongcam-Rudloff**[2], **Christian Brandt**[2,5]

1 Department of Molecular Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden, **2** Department Animal Breeding and Genetics, Bioinformatics section, Swedish University of Agricultural Sciences, Uppsala, Sweden, **3** RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Jena, Germany, **4** Department of Medical Microbiology, University Hospital Leipzig, Leipzig Germany, **5** Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany

* renaud.van.damme@slu.se

Citation: Van Damme R, Hölzer M, Viehweger A, Müller B, Bongcam-Rudloff E, Brandt C (2021) Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN). PLoS Comput Biol 17(2): e1008716. https://doi.org/10.1371/journal.pcbi.1008716

Editor: Mihaela Pertea, Johns Hopkins University, UNITED STATES

Received: July 20, 2020

Accepted: January 17, 2021

Published: February 9, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008716

Data Availability Statement: All subset files for testing the pipeline are available from https://osf.io/m5czv/ MUFFIN is available at https://github.com/

Abstract

Metagenomics has redefined many areas of microbiology. However, metagenome-assembled genomes (MAGs) are often fragmented, primarily when sequencing was performed with short reads. Recent long-read sequencing technologies promise to improve genome reconstruction. However, the integration of two different sequencing modalities makes downstream analyses complex. We, therefore, developed MUFFIN, a complete metagenomic workflow that uses short and long reads to produce high-quality bins and their annotations. The workflow is written by using Nextflow, a workflow orchestration software, to achieve high reproducibility and fast and straightforward use. This workflow also produces the taxonomic classification and KEGG pathways of the bins and can be further used for quantification and annotation by providing RNA-Seq data (optionally). We tested the workflow using twenty biogas reactor samples and assessed the capacity of MUFFIN to process and output relevant files needed to analyze the microbial community and their function. MUFFIN produces functional pathway predictions and, if provided *de novo* metatranscript annotations across the metagenomic sample and for each bin. MUFFIN is available on github under GNUv3 licence: https://github.com/RVanDamme/MUFFIN.

Author summary

Determining the entire DNA of environmental samples (sequencing) is a fundamental approach to gain deep insights into complex bacterial communities and their functions. However, this approach produces enormous amounts of data, which makes analysis time intense and complicated. We developed the Software "MUFFIN," which effortlessly untangle the complex sequencing data to reconstruct individual bacterial species and determine their functions. Our software is performing multiple complicated steps in

Funding: This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – BR 5692/1-1 and BR 5692/1-2. This material is based upon work supported by Google Cloud. BM was funded by FORMAS, grant number 942-2015-1008. MH is supported by the Collaborative Research Centre AquaDiva (CRC 1076 AquaDiva) of the Friedrich Schiller University Jena, funded by the DFG. MH appreciates the support of the Joachim Herz Foundation by the add-on fellowship for interdisciplinary life science. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

parallel, automatically allowing everyone with only basic informatics skills to analyze complex microbial communities.

For this, we combine two sequencing technologies: "long-sequences" (nanopore, better reconstruction) and "short-sequences" (Illumina, higher accuracy). After the reconstruction, we group the fragments that belong together ("binning") via multiple approaches and refinement steps while also utilizing the information from other bacterial communities ("differential binning"). This process creates hundreds of "bins" whereas each represents a different bacterial species with a unique function. We automatically determine their species, assess each genome's completeness, and attribute their biological functions and activity ("transcriptomics and pathways"). Our Software is entirely freely available to everyone and runs on a good computer, compute cluster, or via cloud.

This is a *PLOS Computational Biology* Software paper.

Introduction

Metagenomics is widely used to analyze the composition, structure, and dynamics of microbial communities, as it provides deep insights into uncultivatable organisms and their relationship to each other [1–5]. In this context, whole metagenome sequencing is mainly performed using short-read sequencing technologies, predominantly provided by Illumina. Not surprisingly, the vast majority of tools and workflows for the analysis of metagenomic samples are designed around short reads. However, long-read sequencing technologies, as provided by PacBio or Oxford Nanopore Technologies (ONT), retrieve genomes from metagenomic datasets with higher completeness and less contamination [6]. The long-read information bridges gaps in a short-read-only assembly that often occur due to intra- and interspecies repeats [6]. Complete viral genomes can be already identified from environmental samples without any assembly step via nanopore-based sequencing [7]. Combined with a reduction in cost per gigabase [8] and an increase in data output, the technologies for sequencing long reads quickly became suitable for metagenomic analysis [9–12]. In particular, with the MinION, ONT offers mobile and cost-effective sequencing device for long reads that paves the way for the real-time analysis of metagenomic samples. Currently, the combination of both worlds (long reads and high-precision short reads) allows the reconstruction of more complete and more accurate metagenome-assembled genomes (MAGs) [6].

One of the main challenges and bottlenecks of current metagenome sequencing studies is the orchestration of various computational tools into stable and reproducible workflows to analyze the data. A recent study from 2019 involving 24,490 bioinformatics software resources showed that 26% of all these resources are not currently online accessible [13]. Among 99 randomly selected tools, 49% were deemed 'difficult to install,' and 28% ultimately failed the installation procedure. For a large-scale metagenomics study, various tools are needed to analyze the data comprehensively. Thus, already during the installation procedure, various issues arise related to missing system libraries, conflicting dependencies and environments, or operating system incompatibilities. Even more complicating, metagenomic workflows are computing intense and need to be compatible with high-performance compute clusters (HPCs), and thus different workload managers such as SLURM or LSF. We combined the workflow manager Nextflow [14] with virtualization software (so-called 'containers') to generate reproducible results in various working environments and allow full parallelization of the workload to a higher degree.

Several workflows for metagenomic analyses have been published, including MetaWRAP (v1.2.1) [15], Anvi'o [16], SAMSA2 [17], Humann [18], MG-Rast [19], ATLAS [20], or Sunbeam [21]. Unlike those, MUFFIN allows for a hybrid metagenomic approach combining the strengths of short and long reads. It ensures reproducibility through the use of a workflow manager and reliance on either install-recipes (Conda [22]) or containers (Docker [23], Singularity).

Design and implementation

MUFFIN integrates state-of-the-art bioinformatic tools via Conda recipes or Docker/Singularity containers for the processing of metagenomic sequences in a Nextflow workflow environment (Fig 1). MUFFIN executes three steps subsequently or separately if intermediate results, such as MAGs, are available. As a result, a more flexible workflow execution is possible. The three steps represent common metagenomic analysis tasks and are summarized in Fig 1:

1. Assemble: Hybrid assembly and binning
2. Classify: Bin quality control and taxonomic assessment
3. Annotate: Bin annotation and KEGG pathway summary

The workflow takes paired-end Illumina reads (short reads) and nanopore-based reads (long reads) as input for the assembly and binning and allows for additional user-provided read sets for differential coverage binning. Differential coverage binning facilitates genome bins with higher completeness than other currently used methods [24]. Step 2 will be executed automatically after the assembly and binning procedure or can be executed independently by providing MUFFIN a directory containing MAGs in FASTA format. In step 3, paired-end RNA-Seq data can be optionally supplemented to improve the annotation of bins.

On completion, MUFFIN provides various outputs such as the MAGs, KEGG pathways, and bin quality/annotations. Additionally, all mandatory databases are automatically downloaded and stored in the working directory or can be alternatively provided via an input flag.

Step 1—Assemble: Hybrid assembly and binning. The first step (**Assembly and binning**) uses metagenomic nanopore-based long reads and Illumina paired-end short reads to obtain high-quality and highly complete bins. The short-read quality control is operated using fastp (v0.20.0) [25]. Optionally, Filtlong (v0.2.0) [26] can be used to discard long reads below a length of 1000 bp. The hybrid assembly can be performed according to two principles, which differ substantially in the read set to begin with. The default approach starts from a short-read assembly where contigs are bridged via the long reads using metaSPAdes (v3.13.2) [27–29]. Alternatively, MUFFIN can be executed starting from a long-read-only assembly using metaFlye (v2.8) [30,31] followed by polishing the assembly with the long reads using Racon (v1.4.13) [32] and medaka (v1.0.3) [33] and finalizing the error correction by incorporating the short reads using multiple rounds of Pilon (v1.23) [34]. Both approaches should be chosen based on the available amount of raw read data available to users. E.g., if more short read data is available, meta-spades should be the choice (long reads are "supplemental"). If more long-read data is available, e.g.,> 15 Gigabases (corresponds to a full MinION or GridION flow cell) [35] flye should be used as the assembly approach.

Binning is one of the most crucial steps during metagenomic analysis besides assembly. Therefore, MUFFIN combines three different binning software tools, respectively CONCOCT (v1.1.0) [36], MaxBin2 (v2.2.7) [37], and MetaBAT2 (v2.13) [38] and refine the obtained bins via MetaWRAP (v1.3) [15]. The user can provide additional read data sets (short or long

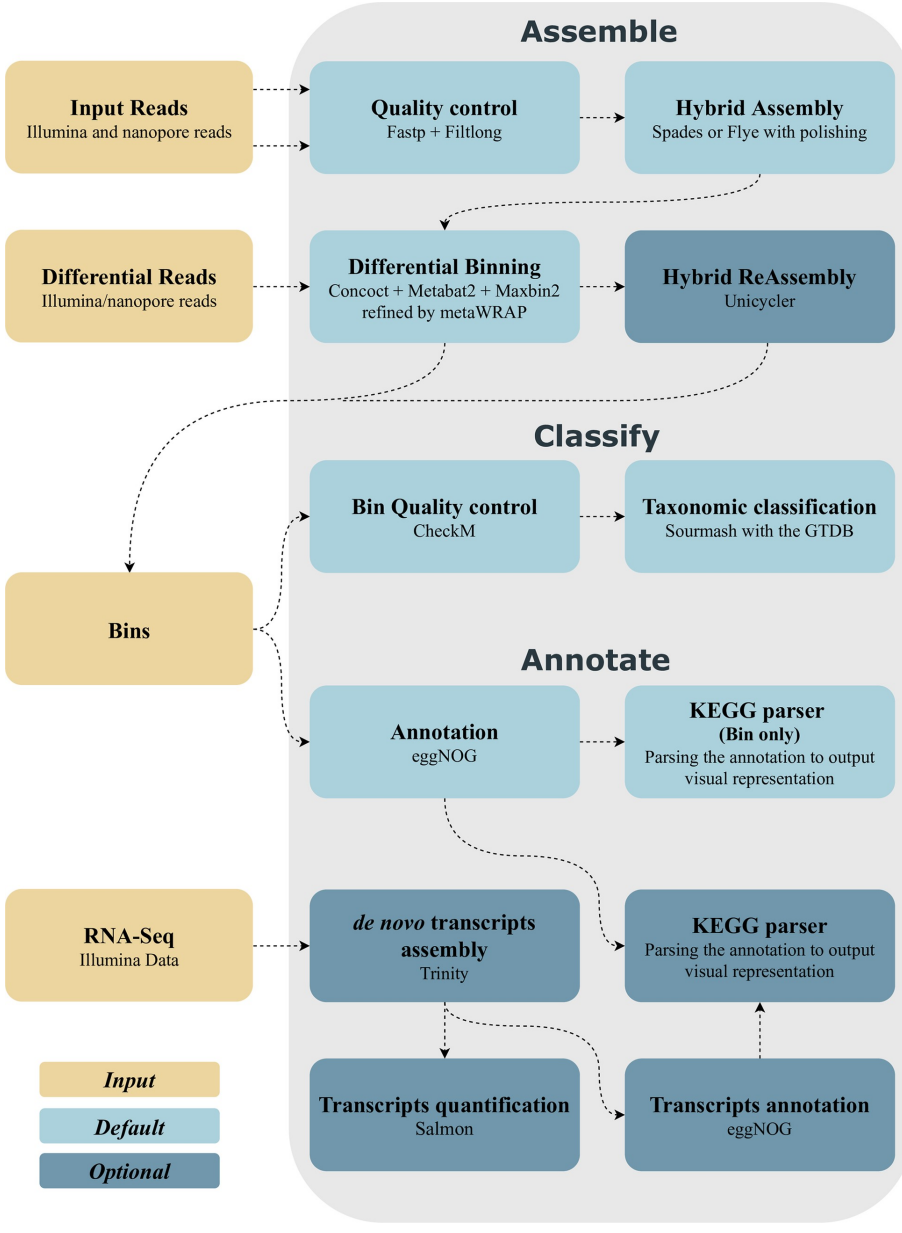
Assemble
Input Reads
Illumina and nanopore reads
Quality control
Fastp + Filtlong
Hybrid Assembly
Spades or Flye with polishing
Differential Reads
Illumina/nanopore reads
Differential Binning
Concoct + Metabat2 + Maxbin2
refined by metaWRAP
Hybrid ReAssembly
Unicycler
Classify
Bin Quality control
CheckM
Taxonomic classification
Sourmash with the GTDB
Bins
Annotate
Annotation
eggNOG
KEGG parser
(Bin only)
Parsing the annotation to output visual representation
RNA-Seq
Illumina Data
de novo transcripts assembly
Trinity
KEGG parser
Parsing the annotation to output visual representation
Transcripts quantification
Salmon
Transcripts annotation
eggNOG
Input
Default
Optional

Fig 1. Simplified overview of the MUFFIN workflow. All three steps (Assemble, Classify, Annotate) from top to bottom are shown. The RNA-Seq data for Step 3 (Annotate) is optional. Differential reads are other read data sets that are solely used for "differential coverage binning" to improve the overall binning performance.

https://doi.org/10.1371/journal.pcbi.1008716.g001

reads) to perform automatically differential coverage binning to assign contigs to their bins better.

Moreover, an additional reassembly of bins has shown the capacity to increase the completeness and N50 while decreasing the contamination of some bins [15]. Therefore, MUFFIN allows for an optional reassembly to improve the continuity of the MAGs further. This reassembly is performed by retrieving the reads belonging to one bin and doing an assembly with Unicycler (v0.4.7) [39]. As each reassembly might improve or worsen each bin, this process is optional and therefore deactivated by default. Individual manual curation is necessary by the user to compare each bin before and after reassembly, as described by Uritskiy *et al.* [15].

To support a transparent and reproducible metagenomics workflow, all reads that cannot be mapped back to the existing high-quality bins (after the refinement) are available as an output for further analysis. These "unused" reads could be further analyzed by other tools such as Kraken2 [40], Kaiju [41], or centrifuge [42] for read classification, "What the Phage" [43] to search for phages, mi-faser [44] for functional annotation of the reads or even use these reads as a new input to run MUFFIN.

Step 2—Classify: Bin quality control and taxonomic assessment. In the second step (**Bin quality control and taxonomic assessment**), the quality of the bins is evaluated with CheckM (v1.1.3) [45] followed by assigning a taxonomic classification to the bins using sourmash (v2.0.1) [46] and the Genome Taxonomy Database (GTDB release r89) [47]. The GTDB was chosen as it contains many unculturable bacteria and archaea–this allows for monophyletic species assignments, which other databases do not assure [35,48]. Moreover, the coherent taxonomic classifications and more accurate taxonomic boundaries (e.g., for class, genus, etc.) proposed by GTDB substantially increases the general classification accuracy [48]. The user can also analyze other bin sets in this step regardless of their origin by providing a directory with multiple FASTA files (bins).

Step 3—Annotate: Bin annotation and KEGG pathway summary. The last step of MUFFIN (**Bin annotation and output summary**) comprises the annotation of the bins using eggNOG-mapper (v2.0.1) [49] and the eggNOG database (v5) [50]. If RNA-Seq data of the metagenome sample is provided (Illumina, paired-end), quality control using fastp (v0.20.0) [25] and a *de novo* metatranscript assembly using Trinity (v2.9.1) [51] followed by quantification of the metatranscripts by mapping of the RNA-seq reads using Salmon (v1.0) [52] are performed. Lastly, the metatranscripts are annotated using eggNOG-mapper (v2.0.1) [49]. Again, the annotation by eggnog-mapper provides a wide array of annotation information such as the GO terms, the NOG terms, the BiGG reaction, CAZy, KEGG orthology, and pathways.

These gene annotations are parsed and visualized in KEGG pathways for each sample and bin. The expression of low and high abundant genes present in the bins is shown. If only bin sets are provided without any RNA-Seq data, the pathways of all the bins are created based on gene presence alone. The KEGG pathway results are summarized in detail as interactive HTML files (example snippet: Fig 2).

Like step 2, this step can be directly performed with a bin set created via another workflow.

Running MUFFIN and version control

MUFFIN (V1.0.3, 10.5281/zenodo.4296623) requires only two dependencies, which allows an easy and user-friendly workflow execution. One of them is the workflow management system

Sample overview
Summary of the pathways and orthologs on a sample level

Pathway highlight 1	...	Pathway highlight 5	Bins Compositions
Pathway name + link to the highlight 1	...	Pathway name + link to the highlight 5	Bin 1 [X orthologs identical to RNAseq, Y orthologs not found in RNAseq]; Bin 2 [X , Y]; ..... Bin n° [X , Y]

Bin overview
Summary of the pathways and orthologs for each bin

Pathway highlight 1 to 4	List of orthologs present in both RNAseq and the bin	List of orthologs only present in the bin
Pathway name + link to the highlight	Ortholog name +link Ortholog name +link ..... Ortholog name +link	Ortholog name +link Ortholog name +link .... Ortholog name +link

Pathway figure
contains a highlight of the orthologs present.

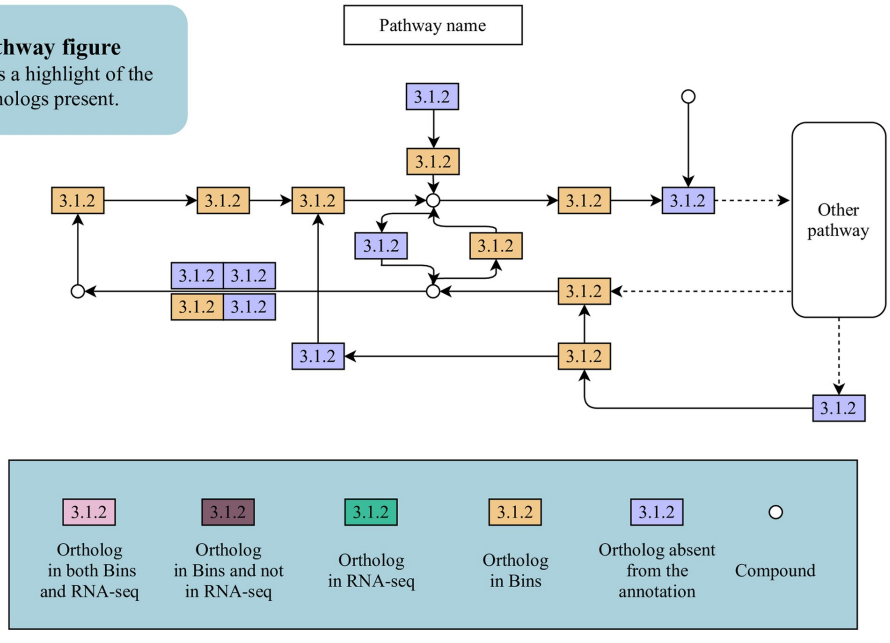

Fig 2. Example snippets of the sub-workflow results of step 3 (Annotate).

https://doi.org/10.1371/journal.pcbi.1008716.g002

Nextflow [14] (version 20.07+), and the other can be either Conda [20][22] as a package manager or Docker [23] / Singularity to use containerized tools. A detailed installation process is available on https://github.com/RVanDamme/MUFFIN. Each MUFFIN release specifies the Nextflow version it was tested on, but any version of MUFFIN V1.0.2+ will work with nextflow version 20.07+. A Nextflow-specific version can always be directly downloaded as an executable file from https://github.com/nextflow-io/nextflow/releases, which can then be paired with a compatible MUFFIN version via the -r flag.

Results

We chose Nextflow for the development of our metagenomic workflow because of its direct cloud computing support (Amazon AWS, Google Life Science, Kubernetes), various ready-to-use batch schedulers (SGE, SLURM, LSF), state-of-the-art container support (Docker, Singularity), and accessibility of a widely used software package manager (Conda). Moreover, Nextflow [14] provides a practical and straightforward intermediary file handling with process-specific work directories and the possibility to resume failed executions where the work ceased. Additionally, the workflow code itself is separated from the 'profile' code (which contains Docker, Conda, or cluster related code), which allows for a convenient and fast workflow adaptation to different computing clusters without touching or changing the actual workflow code.

The entire MUFFIN workflow was executed on 20 samples from the Bioproject PRJEB34573 (available at ENA or NCBI) using the Cloud Life Sciences API (google cloud) with docker containers. This metagenomic bioreactor study provides paired-end Illumina and nanopore-based data for each sample [35]. We used five different Illumina read sets of the same project for differential coverage binning, and the workflow runtime was less than two days for all samples. MUFFIN was able to retrieve 1122 MAGs with genome completeness of at least 70% and contamination of less than 10% (Fig 3). In total, MUFFIN retrieved 654 MAGs with genome completeness of over 90%, of which 456 have less than 2% contamination out of the 20 datasets. For comparison, a recent study was using 134 publicly available datasets from different biogas reactors and retrieved 1,635 metagenome-assembled genomes with genome completeness of over 50% [53].

Exemplarily, we investigated the impact of additional reassembly of each bin for five samples (Fig 3). The N50 was increased by an average of 6–7 fold across all samples. Twenty-six bins of the five samples had an N50 ranging between 1 to 3 Mbases. Reassembly of bins has shown the capacity to increase the completeness and N50 while decreasing the contamination of some bins [15]. This is in line with our samples as some bins benefit more from this step than others. In general, while we observed a general increase in N50 for most bins, the genome quality based on checkM metrics (completeness, contamination) was slightly increasing or decreasing for individual bins.

Discussion

The analysis of metagenomic sequencing data evolved as an emerging and promising research field to retrieve, characterize, and analyze organisms that are difficult to cultivate. There are numerous tools available for individual metagenomics analysis tasks, but they are mainly developed independently and are often difficult to install and run. The MUFFIN workflow gathers the different steps of a metagenomics analysis in an easy-to-install, highly reproducible, and scalable workflow using Nextflow, which makes them easily accessible to researchers.

MUFFIN utilizes the advantages of both sequencing technologies. Short-reads provide a better representation of low abundant species due to their higher coverage based on read count. Long-reads are utilized to resolve repeats for better genome continuity. This aspect is

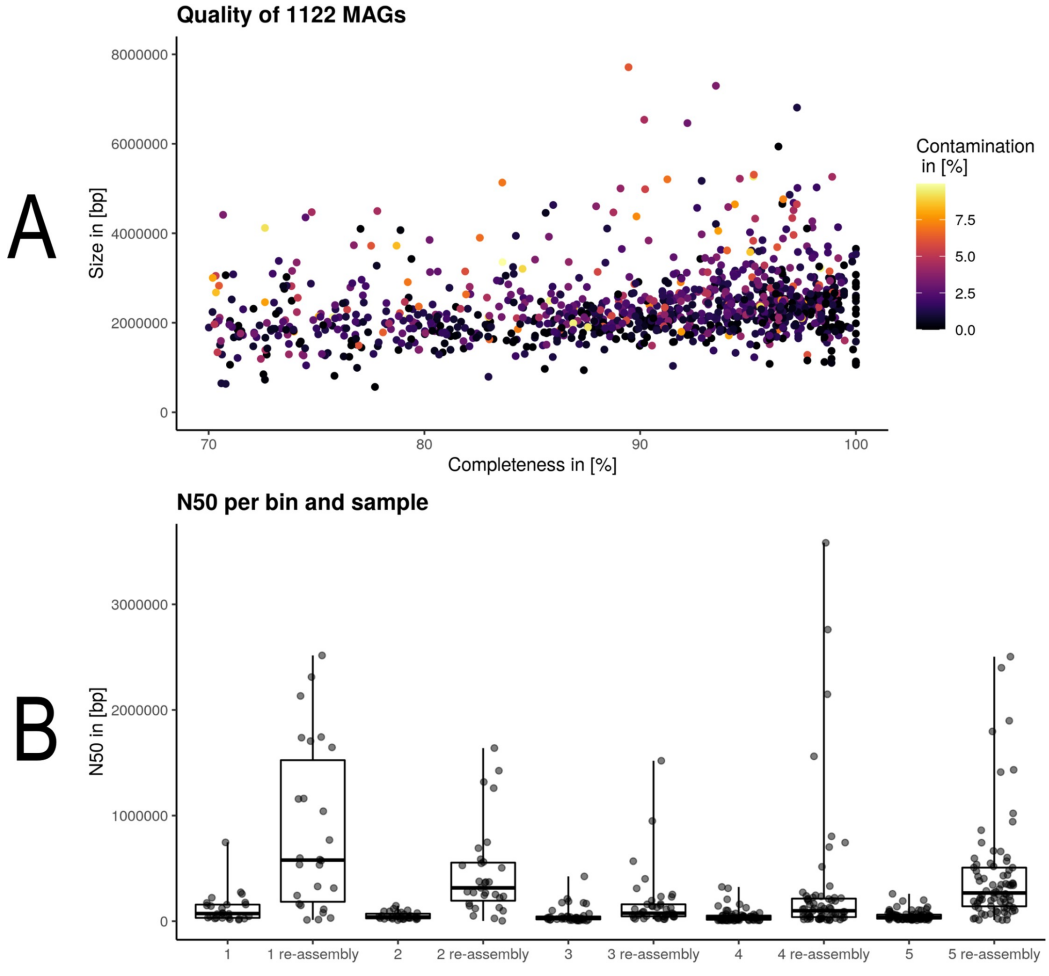

Fig 3. Quality of meta-assembled genomes (MAGs). [A] Quality overview of 1122 MAGs by plotting size to completeness and coloring based on contamination level. [B] N50 comparison between each bin of five selected samples from the Bioproject PRJEB34573 before and after individual bin reassembly.

https://doi.org/10.1371/journal.pcbi.1008716.g003

further utilized via the final reassembly step after binning, which is an optional step due to the additional computational burden which solely aims to improve genome continuity.

Another critical aspect is the full support of differential binning, for both long and short reads, via a single input option. The additional coverage information from other read sets of similar habitats allows for the generation of more concise bins with higher completeness and less contamination because more coverage information is available for each binning tool to decide which bin each contig belongs to.

With supplied RNA-Seq data, MUFFIN is capable of enhancing the pathway results present in the metagenomic sample by incorporating this data as well as the general expression level of the genes. Such information is essential to further analyze metagenomic data sets in-depth, for example, to define the origin of a sample or to improve environmental parameters for production reactors such as biogas reactors. Knowing whether an organism expresses a gene is a crucial element in deciding whether more detailed analysis of that organism in the biotope where the sample was taken is necessary or not.

MUFFIN utilizes a large number of tools to provide a comprehensive analysis of metagenomics samples. The associated tools were mainly chosen based on benchmark performance, e.g., assembly [29,31,54–56], polishing [55], binning [15], annotation for pathways [49], taxonomic classification [47], however stability and workflow compatibility was also an important factor to consider. Due to the modular coding structure of nextflow DSL2 language, MUFFIN can quickly adapt towards better tools or improved versions if necessary, in the future.

MUFFIN executes a de novo assembly of the RNA-seq reads instead of a mapping of the reads against the MAGs to avoid bias and error during the mapping. Indeed, not all the DNA reads were assembled or binned and present in the last step (annotation). Thus we might miss transcripts on the sample level. In addition, for similar genes, it's impossible to know to which organism the reads should map to. By using metatranscripts and comparing the annotations of the metatranscripts to the annotation of the MAGs, we avoid those issues.

Availability and future directions

MUFFIN is an ongoing workflow project that gets further improved and adjusted. The modular workflow setup of MUFFIN using Nextflow allows for fast adjustments as soon as future developments in hybrid metagenomics arise, including the pre-configuration for other workload managers. MUFFIN can directly benefit from the addition of new bioinformatics software such as for differential expression analysis and short-read assembly that can be easily plugged into the modular system of the workflow. Another improvement is the creation of an advanced user and wizard user configuration file, allowing experienced users to tweak the different parameters of the different software as desired.

MUFFIN will further benefit from different improvements, in particular by graphically comparing the generated MAGs via a phylogenetic tree. Furthermore, a convenient approach to include negative controls is under development to allow the reliable analysis of super-low abundant organisms in metagenomic samples.

MUFFIN is publicly available at https://github.com/RVanDamme/MUFFIN under the GNU general public license v3.0. Detailed information about the program versions used and additional information can be found in the GitHub repository. All tools used by MUFFIN are listed in the S1 Table. The Docker images used in MUFFIN are prebuilt and publicly available at https://hub.docker.com/u/nanozoo, and the GTDB formatted for sourmash (v2.0.1)[46] usage is publicly available at https://osf.io/m5czv/. The MAGs produced by the 20 samples; the template of the output of MUFFIN (README_output.txt); the subset data use in the test profile of MUFFIN (subset_data.tar.gz); and the results of MUFFIN on the subset data with and without RNA using both flye and spades are also available at https://osf.io/m5czv/. The Version of MUFFIN presented in this paper is (V1.0.3, 10.5281/zenodo.4296623).

Supporting information

S1 Table. List of the MUFFIN task, the softwares and versions.
(XLSX)

Acknowledgments

We want to thank Hadrien Gourlé and Moritz Buck for the valuable insights into metagenomic analysis and annotation.

Author Contributions

Conceptualization: Renaud Van Damme, Christian Brandt.

Data curation: Renaud Van Damme, Bettina Müller.

Formal analysis: Renaud Van Damme, Christian Brandt.

Funding acquisition: Christian Brandt.

Investigation: Renaud Van Damme.

Methodology: Renaud Van Damme, Martin Hölzer, Adrian Viehweger, Christian Brandt.

Project administration: Christian Brandt.

Resources: Bettina Müller, Erik Bongcam-Rudloff.

Software: Renaud Van Damme, Martin Hölzer, Adrian Viehweger, Erik Bongcam-Rudloff, Christian Brandt.

Supervision: Christian Brandt.

Validation: Renaud Van Damme, Bettina Müller.

Visualization: Renaud Van Damme.

Writing – original draft: Renaud Van Damme.

Writing – review & editing: Martin Hölzer, Adrian Viehweger, Erik Bongcam-Rudloff, Christian Brandt.

References

1. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998; 5: R245–R249. https://doi.org/10.1016/s1074-5521(98)90108-9 PMID: 9818143
2. De R. Metagenomics: aid to combat antimicrobial resistance in diarrhea. Gut Pathog. 2019; 11: 47. https://doi.org/10.1186/s13099-019-0331-8 PMID: 31636714
3. Mukherjee A, Reddy MS. Metatranscriptomics: an approach for retrieving novel eukaryotic genes from polluted and related environments. 3 Biotech. 2020; 10: 71. https://doi.org/10.1007/s13205-020-2057-1 PMID: 32030340
4. Grossart H-P, Massana R, McMahon KD, Walsh DA. Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. Limnol Oceanogr. 2020; 65: S2–S20. https://doi.org/10.1002/lno.11382
5. Carabeo-Pérez A, Guerra-Rivera G, Ramos-Leal M, Jiménez-Hernández J. Metagenomic approaches: effective tools for monitoring the structure and functionality of microbiomes in anaerobic digestion systems. Appl Microbiol Biotechnol. 2019; 103: 9379–9390. https://doi.org/10.1007/s00253-019-10052-5 PMID: 31420693
6. Overholt WA, Hölzer M, Geesink P, Diezel C, Marz M, Küsel K. Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. Environ Microbiol. 2020; 22: 4000–4013. https://doi.org/10.1111/1462-2920.15186 PMID: 32761733
7. Assembly-free single-molecule nanopore sequencing recovers complete virus genomes from natural microbial communities | bioRxiv. [cited 3 Dec 2020]. Available: https://www.biorxiv.org/content/10.1101/619684v1 PMID: 32075851
8. Wetterstrand KA. DNA Sequencing Costs: Data. In: www.genome.gov/sequencingcostsdata [Internet]. 5 Feb 2020 [cited 5 Feb 2020]. Available: www.genome.gov/sequencingcostsdata

9. Somerville V, Lutz S, Schmid M, Frei D, Moser A, Irmler S, et al. Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. BMC Microbiol. 2019; 19: 143. https://doi.org/10.1186/s12866-019-1500-0 PMID: 31238873
10. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, et al. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ. 2019;7. https://doi.org/10.7717/peerj.6800 PMID: 31086738
11. Driscoll CB, Otten TG, Brown NM, Dreher TW. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. Stand Genomic Sci. 2017;12. https://doi.org/10.1186/s40793-017-0232-8 PMID: 28138356
12. Suzuki Y, Nishijima S, Furuta Y, Yoshimura J, Suda W, Oshima K, et al. Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. Microbiome. 2019; 7: 119. https://doi.org/10.1186/s40168-019-0737-z PMID: 31455406
13. Mangul S, Martin LS, Eskin E, Blekhman R. Improving the usability and archival stability of bioinformatics software. Genome Biol. 2019; 20: 47. https://doi.org/10.1186/s13059-019-1649-8 PMID: 30813962
14. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017; 35: 316–319. https://doi.org/10.1038/nbt.3820 PMID: 28398311
15. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018; 6: 158. https://doi.org/10.1186/s40168-018-0541-1 PMID: 30219103
16. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015; 3: e1319. https://doi.org/10.7717/peerj.1319 PMID: 26500826
17. Westreich ST, Treiber ML, Mills DA, Korf I, Lemay DG. SAMSA2: a standalone metatranscriptome analysis pipeline. BMC Bioinformatics. 2018; 19: 175. https://doi.org/10.1186/s12859-018-2189-z PMID: 29783945
18. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. PLOS Comput Biol. 2012; 8: e1002358. https://doi.org/10.1371/journal.pcbi.1002358 PMID: 22719234
19. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, et al. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008; 9: 386. https://doi.org/10.1186/1471-2105-9-386 PMID: 18803844
20. Kieser S, Brown J, Zdobnov EM, Trajkovski M, McCue LA. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. BMC Bioinformatics. 2020; 21: 257. https://doi.org/10.1186/s12859-020-03585-4 PMID: 32571209
21. Clarke EL, Taylor LJ, Zhao C, Connell A, Lee J-J, Fett B, et al. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. Microbiome. 2019; 7: 46. https://doi.org/10.1186/s40168-019-0658-x PMID: 30902113
22. Anaconda Software distribution. Anaconda | The World's Most Popular Data Science Platform. In: https://anaconda.com [Internet]. 5 Feb 2020 [cited 5 Feb 2020]. Available: https://www.anaconda.com/
23. Boettiger C. An introduction to Docker for reproducible research. ACM SIGOPS Oper Syst Rev. 2015; 49: 71–79. https://doi.org/10.1145/2723872.2723882
24. Albertsen M, Philip H, Skarshewski A, Nielsen K, Tyson G, Nielsen P. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31. https://doi.org/10.1038/nbt.2480 PMID: 23302930
25. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018; 34: i884–i890. https://doi.org/10.1093/bioinformatics/bty560 PMID: 30423086
26. Wick R. rrwick/Filtlong. 2020. Available: https://github.com/rrwick/Filtlong
27. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012; 19: 455–477. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599
28. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinforma Oxf Engl. 2016; 32: 1009–1015. https://doi.org/10.1093/bioinformatics/btv688 PMID: 26589280
29. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017; 27: 824–834. https://doi.org/10.1101/gr.213959.116 PMID: 28298430
30. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019; 37: 540–546. https://doi.org/10.1038/s41587-019-0072-8 PMID: 30936562

31. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods. 2020; 17: 1103–1110. https://doi.org/10.1038/s41592-020-00971-x PMID: 33020656

32. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017; 27: 737–746. https://doi.org/10.1101/gr.214270.116 PMID: 28100585

33. nanoporetech/medaka. Oxford Nanopore Technologies; 2020. Available: https://github.com/nanoporetech/medaka

34. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS One. 2014; 9: e112963. https://doi.org/10.1371/journal.pone.0112963 PMID: 25409509

35. Brandt C, Bongcam-Rudloff E, Müller B. Abundance Tracking by Long-Read Nanopore Sequencing of Complex Microbial Communities in Samples from 20 Different Biogas/Wastewater Plants. Appl Sci. 2020; 10: 7518. https://doi.org/10.3390/app10217518

36. Alneberg J, Bjarnason BS, Bruijn I de, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014; 11: 1144–1146. https://doi.org/10.1038/nmeth.3103 PMID: 25218180

37. Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, Singer SW. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome. 2014; 2: 26. https://doi.org/10.1186/2049-2618-2-26 PMID: 25136443

38. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015; 3: e1165. https://doi.org/10.7717/peerj.1165 PMID: 26336640

39. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13. https://doi.org/10.1371/journal.pcbi.1005595 PMID: 28594827

40. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15: R46. https://doi.org/10.1186/gb-2014-15-3-r46 PMID: 24580807

41. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016; 7: 11257. https://doi.org/10.1038/ncomms11257 PMID: 27071849

42. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016 [cited 3 Dec 2020]. https://doi.org/10.1101/gr.210641.116 PMID: 27852649

43. Marquet M, Hölzer M, Pletz MW, Viehweger A, Makarewicz O, Ehricht R, et al. What the Phage: A scalable workflow for the identification and analysis of phage sequences. bioRxiv. 2020. https://doi.org/10.1101/2020.07.24.219899

44. Zhu C, Miller M, Marpaka S, Vaysberg P, Rühlemann MC, Wu G, et al. Functional sequencing read annotation for high precision microbiome analysis. Nucleic Acids Res. 2018; 46: e23. https://doi.org/10.1093/nar/gkx1209 PMID: 29194524

45. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015; 25: 1043–1055. https://doi.org/10.1101/gr.186072.114 PMID: 25977477

46. Brown C, Irber L. sourmash: a library for MinHash sketching of DNA. In: Journal of Open Source Software [Internet]. 14 Sep 2016 [cited 18 Nov 2019]. https://doi.org/10.21105/joss.00027

47. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018; 36: 996–1004. https://doi.org/10.1038/nbt.4229 PMID: 30148503

48. Méric G, Wick RR, Watts SC, Holt KE, Inouye M. Correcting index databases improves metagenomic studies. bioRxiv. 2019; 712166. https://doi.org/10.1101/712166

49. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol Biol Evol. 2017; 34: 2115–2122. https://doi.org/10.1093/molbev/msx148 PMID: 28460117

50. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019; 47: D309–D314. https://doi.org/10.1093/nar/gky1085 PMID: 30418610

51. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013; 8: 1494–1512. https://doi.org/10.1038/nprot.2013.084 PMID: 23845962

52. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017; 14: 417–419. https://doi.org/10.1038/nmeth.4197 PMID: 28263959

53. Campanaro S, Treu L, Rodriguez-R LM, Kovalovszki A, Ziels RM, Maus I, et al. The anaerobic digestion microbiome: a collection of 1600 metagenome-assembled genomes shows high species diversity related to methane production. bioRxiv. 2019; 680553. https://doi.org/10.1101/680553

54. Wick RR, Holt KE. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. F1000Research. 2020; 8: 2138. https://doi.org/10.12688/f1000research.21782.3 PMID: 31984131

55. Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. GigaScience. 2019;8. https://doi.org/10.1093/gigascience/giz043 PMID: 31089679

56. Lau MCY, Harris RL, Oh Y, Yi MJ, Behmard A, Onstott TC. Taxonomic and Functional Compositions Impacted by the Quality of Metatranscriptomic Assemblies. Front Microbiol. 2018;9. https://doi.org/10.3389/fmicb.2018.00009 PMID: 29387050

III

scientific data

OPEN

DATA DESCRIPTOR

Whole genome sequences of 70 indigenous Ethiopian cattle

Wondossen Ayalew[1,2,7], Wu Xiaoyun[1,7], Getinet Mekuriaw Tarekegn[2,3] ✉, Rakan Naboulsi[4], Tesfaye Sisay Tessema[2], Renaud Van Damme[5], Erik Bongcam-Rudloff[5], Min Chu[1], Chunnian Liang[1], Zewdu Edea[6], Solomon Enquahone[3] & Yan Ping[1] ✉

Indigenous animal genetic resources play a crucial role in preserving global genetic diversity and supporting the livelihoods of millions of people. In Ethiopia, the majority of the cattle population consists of indigenous breeds. Understanding the genetic architecture of these cattle breeds is essential for effective management and conservation efforts. In this study, we sequenced DNA samples from 70 animals from seven indigenous cattle breeds, generating about two terabytes of pair-end reads with an average coverage of 14X. The sequencing data were pre-processed and mapped to the cattle reference genome (ARS-UCD1.2) with an alignment rate of 99.2%. Finally, the variant calling process produced approximately 35 million high-quality SNPs. These data provide a deeper understanding of the genetic landscape, facilitate the identification of causal mutations, and enable the exploration of evolutionary patterns to assist cattle improvement and sustainable utilization, particularly in the face of unpredictable climate changes.

Background & Summary

Indigenous animal genetic resources, primarily found in developing countries, are known to contain a significant portion of the world's genetic diversity. Millions of people rely directly on these resources for their livelihoods[1]. Ethiopia, in particular, is considered a gateway for cattle migrations in Africa[2]. Presently, the cattle population in Ethiopia exceeds 70 million heads[3], with 98.5% of them being indigenous cattle. These indigenous cattle are often named based on their appearance, morphological structure, the ethnic group of the herder, and their geographical location[4,5]. Over time, these cattle have developed unique adaptive traits that enable them to withstand challenges such as limited feed availability, high environmental temperatures, and a high prevalence of internal and external parasites and diseases. These adaptive features have been shaped through natural and human selection processes[6,7].

By far, cattle production in Ethiopia is an integral part of almost all farming systems in the crop-livestock mixed farming systems of highlanders and mid attitudes, and the main occupation in the lowland pastoralists, and still promising to rally around the country's economic development. Despite multiple functions and significant phenotypic variations of indigenous cattle populations, little attention was paid to the livestock sector, which threatened the country's cattle diversity and population size. These are mainly associated with complex and interrelated factors such as indiscriminate crossbreeding and interbreeding between adjacent indigenous breeds due to herders' migrations and socio-cultural interactions[8,9]. Furthermore, recurrent drought, the prevalence of disease, ethnic conflicts, and the illegal cross-border market hasten the decline in cattle numbers. Thus, a comprehensive understanding of breed characteristics, including population size, genetic landscape, and geographical distribution, is crucial for effectively managing farm animal genetic resources[1,10]. It also serves as a guiding framework for breed development programs, enabling them to align with specific production needs in diverse environments.

[1]Key Laboratory of Animal Genetics and Breeding on Tibetan Plateau, Ministry of Agriculture and Rural Affairs, Key Laboratory of Yak Breeding Engineering, Lanzhou Institute of Husbandry and Pharmaceutical Sciences, Chinese Academy of Agricultural Sciences, Lanzhou, 730050, P.R. China. [2]Institute of Biotechnology, Addis Ababa University, Addis Ababa P.O. Box 1176, Addis Ababa, Ethiopia. [3]Scotland's Rural College (SRUC), Roslin Institute Building, University of Edinburgh, Edinburgh, EH25 9RG, UK. [4]Childhood Cancer Research Unit, Department of Women's and Children's Health, Karolinska Institute, Tomtebodavägen 18A, 17177, Stockholm, Sweden. [5]Department of Animal Biosciences, Swedish University of Agricultural Sciences, 75007, Uppsala, Sweden. [6]Ethiopian Bio and Emerging Technology Institute, Addis Ababa, Ethiopia. [7]These authors contributed equally: Wondossen Ayalew, Wu Xiaoyun. ✉e-mail: Getinet.Tarekegn@sruc.ac.uk; pingyanlz@163.com

Breeds	No. of samples	Geographic region	Altitude	Latitude	Longitude	Agro-Ecology
Abigar	10	Gambela	523	8.123469	34.30687	Hot, humid, and low-altitude
Barka	10	Amhara	895	14.18467	36.89087	Hot, humid, and low-altitude
Boran	10	Oromiya	1368	4.978936	38.27516	Hot, humid, and low-altitude
Felata	10	Amhara	552	12.40733	35.87573	Hot, humid, and low-altitude
Fogera	10	Amhara	1735	11.86045	37.81373	Humid and mid altitude
Gojjam-Highland	10	Amhara	3410	10.72113	37.85988	Cold, humid, and high-altitude
Horro	10	Oromiya	1722	9.672949	37.07545	Humid and mid-altitude

Table 1. Ethiopian cattle breeds and their respective sampling locations.

Quantitative genetic analysis has historically been characterized as a black box due to the intricate nature of gene action, which involves multiple loci with unknown effects and their interactions in shaping quantitative traits[11]. This complexity has posed challenges in understanding the underlying mechanisms and unraveling the genetic architecture of these traits. As a result, researchers have faced difficulties replicating the results of selective breeding across different spatial and temporal scales, making it essential to explore further and elucidate these complex genetic processes. Advancements in genome sequencing, SNP genotyping technologies, and statistical analysis tools have shifted research focus from analyzing neutral variation to exploring functional variation[12]. Notably, the advent of whole-genome sequencing (WGS) in domestic animals has revolutionized our understanding of their genetic makeup. It has allowed for the identification of causal variants that have significant implications for animal production, health, welfare, and evolutionary studies within livestock species and breeds[13]. While WGS has become a standard tool in various biological sciences, including animal breeding, its application for genetic characterization and routine evaluation of livestock genetic resources in developing countries is still limited. This study presents the whole-genome sequencing data from 70 indigenous cattle originating from seven distinct Ethiopian cattle populations sampled from various agro-ecological and climatic settings (Table 1; Ayalew *et al.*[14]). Thus, our WGS data will serve as a valuable resource for conducting further in-depth studies and investigations in tropical cattle. This sequence dataset will facilitate a deeper understanding of the genetic landscape, allowing for the identification and validation of causal mutations that contribute to essential traits and the exploration of evolutionary patterns.

Moreover, the detailed analytical procedures offer significant advantages for researchers, such as ease of management of similar WGS and implementation of global cattle meta-assemblies at a broader scale. The meta-assembly, which combines multiple genetic or genomic data assemblies into a single, comprehensive assembly, will enable the accurate validation of regions under selection reported by various researchers, ensuring the identification of actual signals while minimizing false positives and supporting future breed improvement and conservation efforts.

Methods

Cattle sampling and collection. We specifically selected seven indigenous cattle populations (Abigar, Barka, Boran, Fellata, Fogera, Gojjam-Highland, and Horro) for our study, with ten unrelated samples collected from each population. These cattle populations inhabit distinct agro-climatic regions, representing Ethiopia's diverse environments (Table 1). We selected these particular populations based on their relevance to agricultural practices, providing insights into desirable production traits, environmental adaptation, and regional livestock farming systems. Blood samples were drawn from the jugular vein of the cattle under sterile conditions, using 10 ml EDTA tubes. The samples were carefully transported to the laboratory in an ice box and stored at −20 °C until DNA extraction.

Extraction and quality control of genomic DNA. The blood samples were thawed for 30 minutes at room temperature and underwent DNA extraction using the Tiangen genomic DNA extraction kit based on the manufacturer's protocols (TIANGEN Biotech, Beijing, China). We conducted 0.8% agarose gel electrophoresis to assess DNA integrity and visualized the resulting DNA bands using a gel imaging apparatus. Each sample's DNA concentration and quality were determined using a Nanodrop Spectrophotometer (ND-2000, Thermo Scientific, Massachusetts, USA) at a wavelength of A260/A280. Samples with DNA concentrations above 50 μg/μl were then sent to Wuhan Frasergen Bioinformatics Co. Ltd in China for whole-genome sequencing (WGS).

Sequence library preparation and sequencing. The VAHTS Universal DNA Library Prep Kit for MGI (Vazyme, Nanjing, China) was employed to generate sequencing libraries of each sample, targeting fragments of approximately 500 bp in length using one microgram of DNA as input material. Adapter sequences were ligated to each sample. Library size and quantification were assessed using Qubit 3.0 Fluorometers and Bioanalyzer 2100 systems (Agilent Technologies, CA, USA). Finally, the sequencing process was conducted by Frasergen Bioinformatics Co., Ltd. (Wuhan, China) on an MGI-SEQ 2000 platform, resulting in a 150 bp sequence length for each sample.

Sequence data pre-processing and mapping. The demultiplexed 70 individual samples (forward and reverse reads) were received and checked for their quality metrics using FastQC v0.11.8[15]. The raw reads were subjected to initial quality control by Trimmomatic v0.39 using default settings[16]. After removing adapter sequences and low-quality reads, MultiQC v1.14 was run on the clean reads, and standard sequence quality metrics were confirmed for subsequent analysis. BWA-MEM 0.7.17-r1188[17] was employed to align individual reads

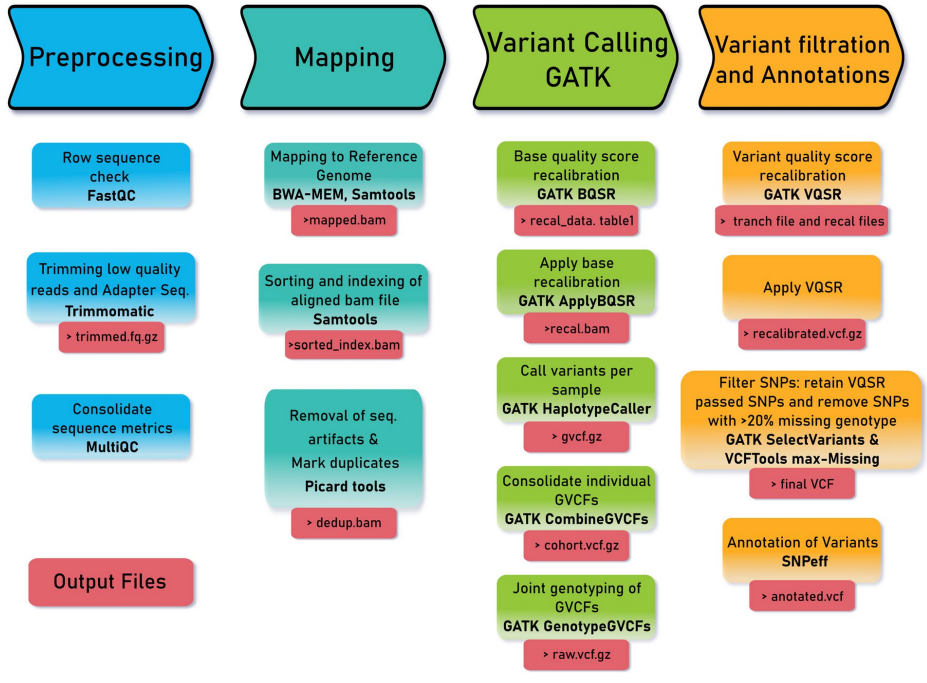

Fig. 1 Overview of raw data quality control, sequence mapping, variant calling, and variant filtration pipeline. The pipeline follows GATK's best practice protocol for germline short variant discovery.

to the latest bovine reference genome ARS-UCD1.2[18]. The aligned reads were converted to binary alignment map (BAM) format, sorted by coordinates, and indexed using SAMtools version 1.6[19]. Finally, the duplicate sequences were marked using the MarkDuplicates function of Picard 2.27.4 (https://broadinstitute.github.io/picard/) to produce a non-duplicated bam file for variant calling.

Variant calling and filtration. High-quality variant calling and filtration are vital in genomic research. The Genome Analysis Toolkit best practices pipeline (https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant discovery) was employed for SNPs discoveries (Fig. 1). First, the marked duplicate bam files were used as input to generate Base Quality Score Recalibration (BQSR) tables using GATK 4.3.0.0. The "Apply BQSR" argument of the same software was then employed to create recalibrated BAM files. The HaplotypCaller method, followed by joint genotyping of all samples and VQSR procedures for SNP recalibrations, was performed using validated SNPs provided by the 1000 bull genome project. In the Variant Quality Score Recalibration (VQSR) procedure, SNP recalibrations utilized different variant annotators, including Quality of Depth (Q.D.), Fisher Strand Test (F.S.), Mapping Quality Score (M.Q.), Mapping Quality Rank Sum Test (MQRankSum), Read Position Rank Sum Test Statistic (ReadPosRankSum), and StrandOddsRatio Test (SOR). Subsequently, the ApplyVQSR procedure was employed to select variants with a true sensitivity of 99.0%. Finally, the 'SelectVariant' procedure from the same software was used, and the final SNPs were used for annotations (refer to the Code availability section).

Data Records

The 70 Ethiopian indigenous cattle pair-end raw sequencing data (in fastq.gz format) were available at NCBI under Sequence Read Archive (SRA) accession numbers SRP478348[20] and SRP480803[21] (Supplementary file 1). The VCF file can be available in the European Variation Archive (EVA) with the accession number for Project PRJEB75238 (https://identifiers.org/ena.embl:ERP159827)[22].

Technical Validation

Quality control for raw reads and alignments. In next-generation sequencing (NGS) data analysis, quality control of raw sequence reads is a standard preliminary procedure before further analysis. This crucial pre-processing step enhances the overall data quality and reliability before conducting downstream analyses[23]. Some essential quality measures used to make choices for the downstream analysis are the base quality, nucleotide

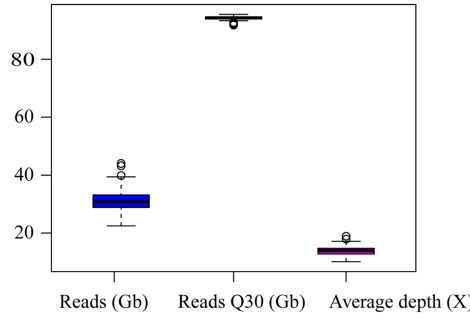

Fig. 2 Boxplot presentation of 70 Ethiopian cattle sequencing yield, yield Q30 and estimated sequence coverage.

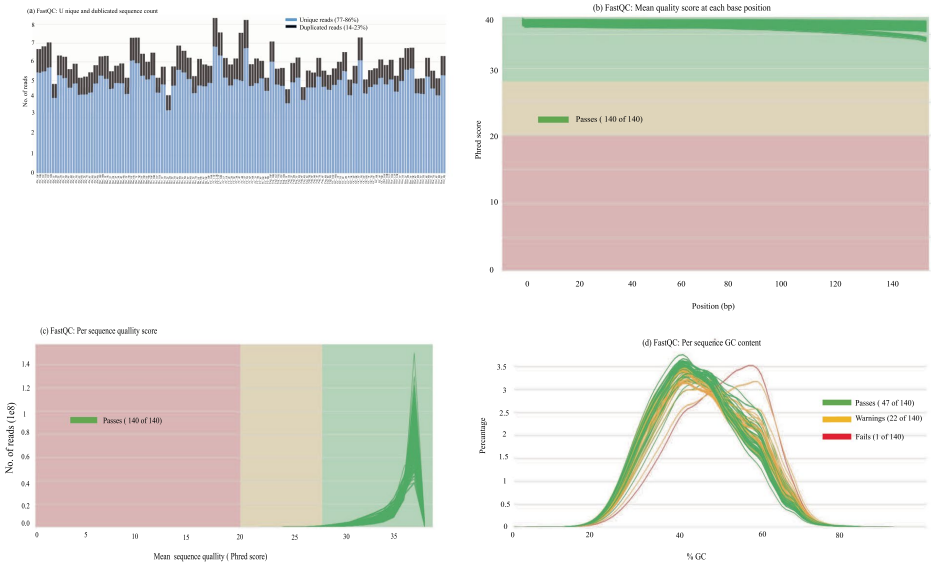

Fig. 3 The quality control metrics from FastQC analysis of 70 cattle sequences. The metrics from all FASTQ files are consolidated using the MultiQC package.

distribution, G.C. content, and duplication rate of the raw sequences[24]. Sequencing of each individual yielded between 13.61 gigabases to 25.45 gigabases, of which 91.8–95.5% of the reads fell above Phred scaled quality score of 30, which proves the bases were called with 99.9% accuracy (Fig. 2). To elucidate all types of variants (including SNVs, indels, and CNVs), a high-depth WGS (30X) is the 'gold standard'[25]. Due to budget constraints, it is common practice to sequence fewer samples at high coverage (20 to 30X). However, this approach may result in a poor representation of a population's genetic variation. The smaller dataset may not adequately capture the full range of genetic diversity present, leading to potential biases or incomplete insights[23]. Recently, Jiang *et al*. suggested 4X as the lowest boundary and 10X as an ideal depth for achieving greater than 99% genome coverage in pigs[26]. The average estimated coverage for each of the 70 Ethiopian cattle samples was above the threshold with an average depth of 14X (Fig. 2). The relatively moderate depth of coverage in our study enhances the resolution and reliability of downstream analyses, leading to more robust findings and insights into the genetic basis of various traits and population dynamics[26,27].

The MultiQC software[28] was employed to generate a pooled sequence quality metrics report (Fig. 3). The MultiQC reports for 70 paired-end Ethiopian cattle sequences confirm that the mean quality scores and per-sequence metrics fell within the high sequence standard range for downstream analysis (Fig. 3b,c). Although there is no universal threshold for duplication levels in WGS data, FastQC flagged a warning for

Annotation categories	Count	% of total
Downstream	2,563,798	4.51%
Exon	513,998	0.90%
Intergenic	23,537,404	41.41%
Intron	27,406,871	48.22%
Splice_site_acceptor	613	0.00%
Splice_site_donor	966	0.00%
Splice_site_region	49,852	0.09%
Transcript	551	0.00%
Upstream	2,507,622	4.41%
UTR_3_prime	176,834	0.31%
UTR_5_prime	75,531	0.13%

Table 2. Single Nucleotide Polymorphisms (SNPs) across various annotation categories.

BTA	CHR Length	SNP count	Density/kb
1	158534110	2225913	14.04
2	136231102	1835540	13.47
3	121005158	1571987	12.99
4	120000601	1692789	14.11
5	120089316	1578815	13.15
6	117806340	1653802	14.04
7	110682743	1467141	13.26
8	113319770	1509341	13.32
9	105454467	1442407	13.68
10	103308737	1391180	13.47
11	106982474	1437389	13.44
12	87216183	1312516	15.05
13	83472345	1092309	13.09
14	82403003	1126064	13.67
15	85007780	1265285	14.88
16	81013979	1116393	13.78
17	73167244	1042961	14.25
18	65820629	874411	13.28
19	63449741	847878	13.36
20	71974595	1041114	14.47
21	69862954	968519	13.86
22	60773035	836115	13.76
23	52498615	874180	16.65
24	62317253	916025	14.70
25	42350435	605379	14.29
26	51992305	747549	14.38
27	45612108	723378	15.86
28	45940150	726207	15.81
29	51098607	803719	15.73
X	139009144	919141	6.61
Unplaced	76654434	213849	2.79

Table 3. Summary of SNPs density in each chromosome.

sequences with more than 20% duplicates[15]. Unlike PCR-free methods, PCR-based sequencing introduces bias in sequencing data by causing uneven amplification of genomic regions and generating duplicate reads, which can impact the accuracy of the sequencing data[29]. Intriguingly, we found an average duplication rate of 17% (Fig. 3a), and this relatively low level of duplication observed in our data can mitigate challenges in variant calling and uneven distribution of coverage across the genome and enhance the efficiency and speed of analysis pipelines[30].

A uniform G.C. content among reads indicates high-quality sequencing, suggesting minimal artifacts or contaminants[24]. However, in our dataset comprising 70 forward and 70 reverse sequencing files (140 files), all sequenced in the same lane and on the same instrument, Fig. 3d reveals some deviations from the expected

distribution of G.C. content in a subset of 23 files (16.43%). These deviations may be attributed to challenges during library preparations[15]. Notably, despite deviations observed in the G.C. content distribution of some sequencing files, a warning message is acceptable for fewer than 30% of the reads, indicating that the overall data quality remains suitable for subsequent analysis[15].

While the quality control for aligned reads is not routinely conducted, it is a valuable tool for gaining additional insights into sample quality. It can help identify problematic samples that might pass the initial raw data quality control checks[24]. In our data, 99.2% of the reads were successfully mapped to the *Bos taurus* (ARS-UCD1.2) reference genome (Supplementary file 2). It suggests that most reads were mapped correctly to their corresponding genomic locations.

Quality control for SNP data. After consolidating individual sample VCF files, the joint genotyping analysis yielded 39 million SNPs. To ensure the reliability of these variants and filter out false-positive calls for downstream analyses, we employed a robust machine-learning model called VQSR (https://gatk.broadinstitute.org/hc/en-us/articles/360035531612-Variant-Quality-Score-Recalibration-VQSR). VQSR is a two-step process that involves training a machine learning model using a training dataset and then applying this model to recalibrate the variant quality scores in the primary dataset. VQSR offers several advantages, including improved accuracy, adaptability, comprehensive assessment, and reduced false positives compared to traditional filtering methods. By incorporating VQSR, we optimized the quality control process and enhanced the validity of our variant calls. Specifically, threshold values of 99% retained about 35 million true variants and excluded four million variants as poor/false positive calls. We also computed the transition/transversion (Ti/Tv) ratio and the heterozygosity-to-homozygosity (het/hom) ratio for SNPs passing the 99% threshold. The observed Ti/Tv and het/hom ratios were 2.35 and 1.17, respectively. These metrics are consistent with values reported for other African zebu cattle breeds[31].

To investigate the genomic distribution and functional impact of genetic variants, we used the SNPeff variant annotation tool. A significant portion of variants (over 89%) were annotated within intronic and intragenic regions (Table 2). Notably, while the number of SNPs per chromosome correlated with chromosome length[32], our study revealed varying SNP densities across chromosomes. For instance, Chromosome 23 showed the highest SNP density (16.65), whereas the X chromosome had the lowest (6.61). These variations are likely attributed to multiple factors, including differences in recombination and mutation rates, genetic drift, demographic influences, selective pressures, and population history[33]. Despite containing more repetitive regions, the X chromosome experiences heightened selection pressure against genetic variants, driven by hemizygosity in males and X-chromosome inactivation in females. As a result, the X chromosome exhibits a lower SNP density than autosomes. These unique genetic mechanisms and evolutionary dynamics significantly shape the distinct SNP profiles observed between the X chromosome and autosomes[34] Table 3.

Code availability

Data analyses were primarily conducted using standard bioinformatics tools on the Linux operating system. We provide detailed information about the versions and code parameters of the software tools used at https://github.com/WondossenA/WGS_Ethiopian_cattle/blob/main/code_explanation.md.

Received: 14 February 2024; Accepted: 2 May 2024;
Published online: 05 June 2024

References

1. Rege, J. E. O. & Gibson, J. P. Animal genetic resources and economic development: issues in relation to economic valuation. *Ecol. Econ.* **45**, 319–330 (2003).
2. Hanotte, O. *et al*. African pastoralism: genetic imprints of origins and migrations. *Science* **296**, 336–339 (2002).
3. CSA. Federal Democratic Republic of Ethiopia Central Statistical Agency Agricultural Sample Survey 2021/[2013E.C.], Volume II, Report on Livestock and Livestock Characteristics. 1–199 (2021).
4. Ethiopian Institute of Biodiversity (EIB). Ethiopia's Fifth National Report to the Convention on Biological Diversity. Ethiopian Biodiversity Institute, Addis Ababa, Ethiopia. (2016).
5. Domestic Animal Diversity Information System (DADIS). Number of breeds by species and country. http://dad.fao.org/ (2021).
6. Mwai, O., Hanotte, O., Kwon, Y. J. & Cho, S. African indigenous cattle: unique genetic resources in a rapidly changing world. *Asian Australas J. Anim. Sci.* **28**, 911–921 (2015).
7. Taye, M. *et al*. Whole genome scan reveals the genetic signature of African Ankole cattle breed and potential for higher quality beef. *BMC Genet.* **18**, 1–14 (2017).
8. Hassen, F., Bekele, E., Ayalew, W. & Dessie, T. Genetic variability of five indigenous Ethiopian cattle breeds using RAPD markers. *Afr. J. Biotechnol.* **6**, 19 (2007).
9. Hanotte, O., Dessie, T. & Kemp, S. Time to tap Africa's livestock genomes. *Science* **328**, 1640–1641 (2010).
10. FAO. Global Plan of Action for Animal Genetic Resources and the Interlaken Declaration. http://www.fao.org/docrep/010/a1404e/a1404e00 (2007).
11. Hill, W. G. Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society B: Biol. Sci.* **365**, 73–85 (2010).
12. Mrode, R., Ojango, J. M. K., Okeyo, A. M. & Mwacharo, J. M. Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: Current status and future prospects. *Front. Genet.* **9**, 694 (2019).
13. Sharma, A. *et al*. Next generation sequencing in livestock species: A review. *J. Anim. Breed. Genom.* **1**, 23–30 (2017).
14. Ayalew, W. *et al*. Whole-Genome Resequencing Reveals Selection Signatures of Abigar Cattle for Local Adaptation. Animals **13**, p.3269 (2023).
15. Andrews, S. FastQC: a quality control tool for high throughput sequence data, Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010).
16. Bolger, A.M., Lohse, M., & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* btu170 (2014).
17. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
18. Rosen, B. D. *et al*. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **9**, p.giaa021 (2020).

19. Li, H. SAMtools 1.6: a toolkit for DNA sequence analysis. *Bioinformatics* **34**, 3313–3314 (2017).
20. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP478348 (2024).
21. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP480803 (2024).
22. *European Variation Archive*. https://identifiers.org/ena.embl:ERP159827 (2024).
23. Pfeifer, S. From next-generation resequencing reads to a high-quality variant data set. *Heredity* **118**, 111–124 (2017).
24. Guo, Y., Ye, F., Sheng, Q., Clark, T. & Samuels, D. C. Three-stage quality control strategies for DNA resequencing data. *Brief Bioinform.* **15**, 879–89 (2014).
25. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–32 (2014).
26. Jiang, Y. *et al*. Optimal sequencing depth design for whole genome resequencing in pigs. *BMC Bioinform.* **20**, 556 (2019).
27. Rashkin, S., Jun, G., Chen, S. & Abecasis, G. R. Optimal sequencing strategies for identifying disease-associated singletons. *PLoS Genet.* **13**(6), e1006811 (2017).
28. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
29. Van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* **322**, 12–20 (2014).
30. Ebbert, M. T. *et al*. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinform.* **25**(17 Suppl 7), 239 (2016).
31. Tijjani, A. *et al*. Genomic signatures for drylands adaptation at gene-rich regions in African zebu cattle. *Genomics* **114**, 110423 (2022).
32. Zhao, Z., Fu, Y. X., Hewett-Emmett, D. & Boerwinkle, E. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene* **312**, 207–213 (2003).
33. Czech, B., Guldbrandtsen, B. & Szyda, J. Patterns of DNA variation between the autosomes, the X chromosome and the Y chromosome in *Bos taurus* genome. *Sci. Rep.* **10**, 13641 (2020).
34. Gorlov, I. P. & Amos, C. I. Why does the X chromosome lag behind autosomes in GWAS findings? *PLoS Genet.* **19**, e1010472 (2023).

Acknowledgements

The authors express their gratitude for the financial assistance received from the Innovation Project of the Chinese Academy of Agricultural Sciences (Project No. 25-LZIHPS-01) and the China Agriculture Research System of the Ministry of Finance and the Ministry of Agriculture and Rural Affairs (Project No. CARS-37). Additionally, the authors thank the Swedish University of Agricultural Sciences (SLU) in Uppsala, Sweden, for generously providing access to the SLU Bioinformatics Infrastructure (SLUBI) and other valuable support throughout this research. Lastly, the authors would like to acknowledge the Institute of Biotechnology at Addis Ababa University, Ethiopia, for graciously providing access to laboratory facilities and other essential resources.

Author contributions

W.A., G.M.T. and W.X. conceived the research project. W.A. collected the blood sample. W.A. and S.E. participated in laboratory work. W.A. did bioinformatics analysis and got inputs from G.M.T., X.W., R.N., R.V. and Z.E. W.A., X.W., E.B., T.S.T., Z.E., R.N., R.V., C.L., M.C. and Y.P. were involved in the review and writing process, and Y.P., T.S.T. and E.B. provided resources and managed project administration. All authors made critical contributions to the manuscript drafts.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03342-9.

Correspondence and requests for materials should be addressed to G.M.T. or Y.P.

Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2025:60

Ethiopia experiences significant seasonal changes, which impact the living conditions of its cattle. This thesis explored the changes in the rumen microbiome to understand their effects. Bioinformatics tools were developed: MUFFIN and PANKEGG. The results have shown an increase in fibre-degrading and methane-producing microbes during the rainy season, while hydrogenotrophic methanogens increase in the dry season, indicating a loss of feed efficiency during periods of limited feed availability. Antibiotic resistance genes were detected in various bacteria across both seasons.

Renaud Van Damme Damme received his doctoral education at the Department of Animal Biosciences, Swedish University of Agricultural Sciences (SLU). He received his undergraduate degree from the Haute Ecole en Hainaut (HEH) and his MSc from the SLU.

Acta Universitatis Agriculturae Sueciae presents doctoral theses from the Swedish University of Agricultural Sciences (SLU).

SLU generates knowledge for the sustainable use of biological natural resources. Research, education, extension, as well as environmental monitoring and assessment are used to achieve this goal.

ISSN 1652-6880
ISBN (print version) 978-91-8124-044-3
ISBN (electronic version) 978-91-8124-090-0