Rumen-Centric Assembly of the Cattle Holobiont

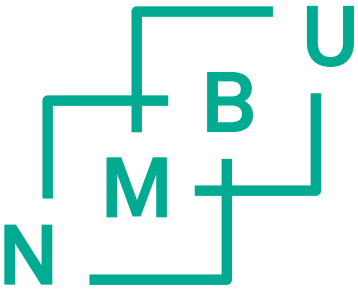
Vomsentrert Sammenstilling av Storfeholobionten

Philosophiae Doctor (PhD) Thesis
Carl Mathias Kobel, MSc

Norwegian University of Life Sciences
Faculty of Biosciences

Section for Ruminant Nutrition and Physiology &
The Microbial Ecology and Meta-Omics (MEMO) Group

Ås, Norway 2025



Thesis 2025:1
ISSN 1894-6402
ISBN 978-82-575-2204-9
version 1.0.0

Strictly no generative machine learning was used in the writing of this thesis.

Supervisory team

Main supervisor: Prof. Phillip B. Pope[1,2,3]
Co-supervisor: Assistant Prof. Ianina Altshuler[4]
Co-supervisor: Dr. Velma T. E. Aho[1]
Co-supervisor: Dr. Ove Øyås[1]

Evaluation committee

Member: Prof. Chris Creevey[5]
Member: Associate Prof. Manuel Kleiner[6]
Committee coordinator: Prof. Angela Schwarm[1]

Affiliations

1. Faculty of Biosciences,
   Norwegian University of Life Sciences, Norway
2. Faculty of Chemistry, Biotechnology and Food Science,
   Norwegian University of Life Sciences, Norway
3. School of Biomedical Sciences,
   Queensland University of Technology, Australia
4. Environmental Engineering Institute,
   École Polytechnique Fédérale de Lausanne, Switzerland
5. School of Biological Sciences,
   Queen's University Belfast, Northern Ireland
6. Department of Plant & Microbial Biology,
   North Carolina State University, United States of America

Acknowledgments

This work was undertaken at Norwegian University of Life Sciences in Ås, August 2021–October 2024. My colleagues and I were brought together in the Section for Ruminant Nutrition and Physiology (Faculty of Biosciences), The Protein Engineering and Proteomics Group (PEP, Faculty of Chemistry, Biotechnology and Food Science), and the Microbial Ecology and Meta-Omics group (MEMO) across both faculties.

I want to acknowledge the Novo Nordisk Foundation, NMBU, and the Norwegian Research School in Bioinformatics, Biostatistics, and Systems Biology (NORBIS), for the financial support that made this PhD possible.

Thanks to Johan Hjort Andersen for consultation on scientific computer systems designs, which have made possible the solution to severe computational challenges I encountered during my PhD.

For my overseas research stays I want to thank the people in several research groups. In Brisbane, Queensland, Australia: The Livestock group at Commonwealth Scientific and Industrial Research Organisation; the Australian Centre for Ecogenomics at Queensland University; and the Centre for Microbial Research at Queensland University of Technology—all for their hospitality and warm inclusion. Thanks to Thea for looking after me while traveling, making sure that I was always getting into an appropriate amount of trouble. Thanks to the Centre for Microbial Research (CMR) at Queensland University of Technology for joining us as partners on the SuPAcow project. Thanks to the Centre for Evolutionary Hologenomics (CEH) at University of Copenhagen for their uncompromised hospitality during the last months of my PhD, softening my re-entry to Denmark. What I have learnt during this PhD is that the culture of hospitality within research groups is simply amazing all around the world: Cheers to all the wonderful people I've met on my way!

Shout-outs to everyone in the MEMO group for being the coolest and most fun research group I could ever dream to be part of.

Most importantly, massive thanks to my supervisors, Phil, Ianina, Velma, & Ove, for taking me into the world of science, and to let me explore my own weird ideas on multiple continents. Thanks for all the help and the mind-expanding scientific discussions, and good humor.

Lastly, a shout-out to my family, especially Johanne, for taking care of me, keeping me sane, and showing me that there are greater things in life than work and career.

Til Sonja

Contents

Abbreviations & formulas

ATP	adenosine triphosphate
bp	basepair
CAZyme	carbohydrate-active enzyme
CO_2	carbon dioxide
DNA	deoxyribonucleic acid
H_2	dihydrogen
LC-MS/MS	liquid chromatography-tandem mass spectrometry
MAG	metagenome-assembled genome
NAD	nicotinamide adenine dinucleotide
PCA	principal component analysis
RCT	rumen community type
RNA	ribonucleic acid
VFA	volatile fatty acid
WGCNA	weighted gene co-expression network analysis

Included papers

These are referred to as **#1–4** within this thesis.

#1. Long-Read Metagenomics and CAZyme Discovery
Alessandra Ferrillo, Carl Mathias Kobel, Arturo Vera-Ponce de León, Sabina Leanti La Rosa, Benoît Josef Kunath, Phillip Byron Pope, Live Heldal Hagen
2023, Book chapter within Methods in Molecular Biology, Springer Nature
(*Reproduced with permission from Springer Nature*)
doi: 10.1007/978-1-0716-3151-5_19

#2. CompareM2 is a genomes-to-report pipeline for comparing microbial genomes
Carl M. Kobel, Velma T. E. Aho, Ove Øyås, Niels Nørskov-Lauritsen, Ben J. Woodcroft, Phillip B. Pope
2024, Preprint in Biorxiv, Cold Spring Harbor Laboratory
doi: 10.1101/2024.07.12.603264

#3. Integrating host and microbiome biology using holo-omics
Carl M. Kobel, Jenny Merkesvik, Idun Maria Tokvam Burgos, Wanxin Lai, Ove Øyås, Phillip B. Pope, Torgeir R. Hvidsten, Velma T. E. Aho
2024, Molecular Omics, Royal Society of Chemistry
doi: 10.1039/D4MO00017J

#4. Protozoal populations drive system-wide variation in the rumen microbiome
Carl M. Kobel, Andy Leu, Arturo V. P. de Leon, Ove Øyås, Wanxin Lai, Ianina Altshuler, Live H. Hagen, Rasmus D. Wollenberg, Cassie R. Bakshani, William G. T. Willats, Laura Nicoll, Simon J. McIlroy, Torgeir R. Hvidsten, Chris Greening, Gene W. Tyson, Rainer Roehe, Velma T. E. Aho*, Phillip B. Pope*
(* signifies equal contributions),
Manuscript, November 2024

In addition to the included papers, contributions were made to the following works.

- **Predicting microbial genome-scale metabolic networks directly from 16S rRNA gene sequences**
  Ove Øyås, Carl M. Kobel, Jan Olav Vik, Phillip B. Pope
  2024, Preprint in Biorxiv, Cold Spring Harbor Laboratory
  doi: 10.1101/2024.01.26.576649

- **Transkingdom network analysis across the host-microbiome nexus in the bovine rumen reveals associations to host traits**
  I. Altshuler, A. V.-P. de León, R. Roehe, M. Watson, C. M. Kobel, J. F. Firkins, Z. Yu, P. B. Pope, *Manuscript*

- **OmniCorr: An R-package for visualizing putative host-microbiota interactions using multi-omics data**
  Shashank Gupta, Wanxin Lai, Carl M. Kobel, Velma T. E. Aho, Arturo Vera-Ponce de León, Sabina Leanti La Rosa, Simen R. Sandve, Phillip B. Pope, Torgeir R. Hvidsten, *Manuscript*

- **Key Microbiota Species and Metabolic Capabilities Driving Methane Emission Levels of Cows**
  Wanxin Lai, Antton Alberdi, Andy Leu, Arturo V. P. de Leon, Carl M. Kobel, Velma T. E. Aho, Rainer Roehe, Phil B. Pope, Torgeir R. Hvidsten, *Manuscript*

- **Rumen microbiome reconstruction following rumen content exchange: Low methane emitters reconstitute while high emitters inherit**
  Puchun Niu, Carl M. Kobel, Velma T. E. Aho, Clementina Alvarez, Egil Prestløkken, Peter Lund, Bjørg Heringstad, Phil B. Pope, Angela Schwarm, *Manuscript*

Summary

The holobiont perspective takes into account the complete system of an animal including its microbiomes and to some extent the factors that form its external environment. This view has a scientific motivation: In the case that the microbiome affects the host, studying the host in isolation severely limits access to the complete information needed to understand the biology of the host. As a model for such a "holobiont" system, a starting point was taken in cattle and focused on the rumen and the interface that connects its inherent microbiome to the metabolism of the host. As the rumen microbiome produces a vital component of metabolites that the host budgets for energy and nutrient assimilation, it has a wide potential to impact the host. The association between the host and its rumen microbiome has made it a focal point for modulation strategies to improve health, nutrition and sustainable production of ruminants. However, the complexity of the rumen microbiome and its interactions with the host represent major challenges that must be overcome before microbiome-based approaches can be used in practice. To improve reconstruction of the rumen microbiome, a high-resolution dataset was generated for deep analysis from 80 cattle subjected to a feedlot trial. Here, the rumen microbiome was sampled over time, and host tissue (rumen wall and liver) samples were collected upon sacrifice, after rigorous measurement of the cattle's key performance traits (KPTs) and methane emissions.

To study the ruminant holobiont, molecular layers in both the host and its rumen microbiome were reconstructed. Multiple molecular layers (DNA, RNA, protein, metabolites, and glycans) as well as the host phenotype were analyzed, in order to track how potential interactions affected metabolism in 24 individual animals that exhibited the highest natural variation in measured methane yield. As most biological variation of an organism is encoded in the genome, DNA sequencing is central to forming a foundation upon which to assemble the holobiont. To further enhance our DNA analyses, long-read and high accuracy short-read shotgun metagenome sequencing was applied to reconstruct the microbial genomes of the rumen microbiome. To track which genes were expressed, transcriptomics was applied, and to analyze the presence of translated proteins and their derived metabolites, also proteomics and metabolomics. For complex eukaryotic populations that are unamenable to shotgun sequencing approaches, such as the protozoa and fungi, genomes were sourced from collaborative projects.

Analyzing a single molecular layer requires a specific set of technical tools. For this purpose, it is described how microbial genomes can be reconstructed, and how their relevant metabolic functions can be identified. Specifically in relation to the carbohydrate-active enzymes (CAZymes) that enable ruminants to assimilate carbon

and energy from plant fibers, representing the basis for the energy budget of the host. As it was not possible to identify a suitable pipeline with the tools necessary to analyze and compare the metabolic function of the archaeal and bacterial genomes generated in our datasets, an easy to use platform for analyzing metagenome-assembled genomes (MAGs) was developed. This pipeline is now distributed on Bioconda as CompareM2.

To apply the wide tool set that was put together and attempt to improve resolution and general understanding of how the ruminant host and its microbiome function as an integrated unit, an experimental cattle holobiont dataset was analyzed. The sampled molecular layers were refined to become biologically relevant representations on which integrative holo-omic methods could be applied to identify and investigate possible host-microbiome interactions. In practice, simpler computational dimensionality reduction methods may offer greater interpretability and allow more direct biological interpretation than more complex computational methods. Applying these methods to our experimental data led to the discovery that the protozoal fraction of the rumen seemingly drives two exclusive community types across the individual animals that were sampled, which have previously been described from micrography and 18S studies. These are referred to as RCT-A and -B (rumen community type). RCT-B is dominated by protozoa affiliated to *Epidinium* spp. that were observed to employ a plethora of fiber-degrading enzymes, which most likely provide favorable conditions for saccharolytic bacteria such as *Prevotella* spp. Conversely RCT-A is dominated by *Isotricha* and *Entodinium* protozoal species and harbors a wider representation of fiber, protein and amino acid fermenters. While no clear host effect for these rumen community types is found, there are signs in the more complex network analysis based computational methods that certain microbial populations of *Acutalibacteraceae* prevalent in RCT-A affect methionine metabolism in the rumen wall. This calls for further refinement of the holo-omic analyses and biological characterization.

Finally, our work highlights the need for *de facto* standards to refine individual molecular layers, and to find common methods for data integration across these molecular layers that represent the host-microbiome axis.

Norsk sammendrag

Holobiont-perspektivet tar hensyn til det komplette systemet til et dyr, inkludert dets mikrobiomer og, til en viss grad, faktorene som danner dets ytre miljø. Dette perspektivet har en vitenskapelig motivasjon: I tilfeller der mikrobiomene påvirker verten, begrenser det å studere verten i isolasjon tilgangen til den komplette informasjonen som er nødvendig for å forstå vertens biologi. Som en modell for et slikt holobiont-system har vi tatt utgangspunkt i storfe, med fokus på vomma og grensesnittet som forbinder dets iboende mikrobiom med vertens metabolisme, spesifikt vomveggen og leveren. Siden mikrobiomet i vomma produserer en viktig komponent av metabolitter som verten bruker til energibudsjettering og næringsopptak, har det et stort potensial til å påvirke verten. Sammenhengen mellom verten og dens iboende vommikrobiom er derfor et fokusområde for moduleringsstrategier for å forbedre helse, ernæring og bærekraftig storfeproduksjon. Vommikrobiomets kompleksitet og dets interaksjoner med verten representerer imidlertid store utfordringer som må takles før mikrobiom-baserte tilnærminger kan brukes i praksis. For å forbedre rekonstruksjonen av vommikrobiomet har vi generert et høyoppløselig datasett for dyp analyse fra 80 storfe som ble eksponert for en fôringsprøve. Her ble vommikrobiomet tatt prøver av over tid, og prøver fra vomveggen og leveren ble samlet ved avliving, kort tid etter grundige målinger av storfeets nøkkelprestasjonsegenskaper, inkludert metanutslipp.

For å studere holobionten hos drøvtyggere, rekonstruerte vi både molekylære lag hos verten og i dens iboende vommikrobiom. For å forstå hvordan potensielle interaksjoner påvirket metabolismen tok vi hensyn til flere molekylære lag (DNA, RNA, proteiner, metabolitter og glykans) samt vertens fenotype, i 24 individuelle storfe som viste den høyeste naturlige variasjonen i målt metanproduksjon. Siden praktisk talt all biologisk variasjon av en organisme er kodet i genomet, er DNA-sekvensering sentralt for å danne et grunnlag for å samle holobionten. For ytterligere å forbedre våre DNA-analyser, kombinerte vi lange sekvenser (long-read) og høy-presisjon shotgun metagenomsekvensering for å rekonstruere mikrobielle genomer i vommikrobiomet. For å kartlegge hvilke gener som ble uttrykt, anvendte vi transkriptomikk, og for å analysere tilstedeværelsen av translaterte proteiner og deres avledede metabolitter, ble det brukt proteomikk og metabolomikk. For komplekse eukaryote populasjoner som ikke lar seg analysere med shotgun-sekvenseringsmetoder, som protozoer og sopp, hentet vi genomer fra samarbeidende prosjekter.

Analyse av et enkelt molekylært lag krever et spesifikt sett med tekniske verktøy. I denne sammenhengen beskriver vi hvordan mikrobielle genomer kan rekonstrueres og hvordan deres relevante metabolske funksjoner kan identifiseres. Mer spesifikt,

fokuserer vi på karbohydrataktive enzymer i vomma som gjør det mulig for drøvtyggere å tilegne seg karbon og energi fra plantefibre, som utgjør grunnlaget for vertens energibudsjett. Ettersom vi ikke klarte å identifisere en passende pipeline med verktøyene som var nødvendige for å analysere og sammenligne den metabolske funksjonen til de arkeiske og bakterielle genomene som ble generert i våre datasett, utviklet vi en brukervennlig plattform for å analysere metagenom-assembled genomes (MAGs). Denne pipelinen distribueres nå på Bioconda som CompareM2.

For å kunne utnytte det brede verktøysettet vi har satt sammen, og dermed forsøke å forbedre oppløsningen og den generelle forståelsen av hvordan verten og dens mikrobiom fungerer som en integrert enhet, satte vi oss fore å analysere eksperimentelle data fra storfe-holobiontet. Vi forbedret allerede testede molekylære lag til å bli biologisk relevante representasjoner, og anvendte integrerte holo-omiske metoder for å identifisere og undersøke mulige interaksjoner mellom vert og mikrobiom. I praksis finner vi at de enklere beregningsmetodene for dimensjonsreduksjon tilbyr større tolkbarhet og gir en mer direkte biologisk tolkning enn mer komplekse beregningsmetoder. Ved å bruke disse metodene på våre eksperimentelle data oppdaget vi at protozoa-fraksjonen av vomma tilsynelatende driver to eksklusive samfunnstyper på tvers av individuelle dyr, som tidligere har blitt beskrevet i mikrografi og 18S-studier. Vi refererer til disse som RCT-A og -B (vomsamfunnstyper). RCT-B domineres av protozoer tilknyttet *Epidinium* spp. som ble observert å utnytte et mangfold av fiber-nedbrytende enzymer, noe vi mener gir gunstige forhold for andre sukker-nedbrytende bakterier som Prevotella spp. Motsatt domineres RCT-A av protozoiske arter fra *Isotricha* og *Entodinium*, og har i tillegg en bredere representasjon av fiber-, protein-, og aminosyrenedbrytende organismer. Selv om vi ikke finner en klar vertseffekt for disse vomsamfunnstypene, indikerer de mer komplekse nettverksanalysebaserte beregningsmetodene at visse mikrobiologiske populasjoner av Acutalibacteraceae som er utbredt i RCT-A påvirker metioninmetabolismen i vomveggen. Dette krever ytterligere forbedring av de holo-omiske analysene og biologisk karakterisering.

Avslutningsvis fremhever arbeidet vårt behovet for å sette standarder for å forbedre resolusjonen av individuelle molekylære lag, og for å finne gullstandarden for data-integrasjon på tvers av de molekylære lagene som representerer vert-mikrobiom-aksene.

(*Translated to Norwegian by Thea Os Andersen*)

Synopsis

1. Introduction

The biology of the cattle holobiont

Origin of the holobiont

Animals live in a microbial world. The endosymbiotic theory states that eukaryotes themselves arose from the symbiosis of microorganisms 1.8–2.7 gigayears ago, when an Asgard-related archaeon engulfed a Rickettsiales-related bacterium[1,2]. The archaeon assimilated the metabolic function of the bacterium by engulfing the complete cell. The assimilated bacterium "mitochondrion" performed respiration for the host, and in return was provided with a protected environment where it could thrive. Over time, coevolution between the host and the assimilated partner gradually intertwined the two into becoming mutually interdependent through their shared biology. In the case of the mitochondrion, several of its key genes now reside in the host genome, rendering an escape of the assimilated cell line unlikely[1,2]. In an analogous model of co-dependence but on a grander and more complex scale, ancestors of cattle have succeeded to co-evolve with and assimilate the metabolic function of a diverse complex of microorganisms within their rumen[3,4]. Herein, cattle maintain a protected environment where microorganisms can thrive to break down plant fibers for the microbiome itself, and the host can harness the energy and nutrition extracted from these. Many of the microorganisms observed are endemic to the rumen of cattle, which means that they are not observe to live outside this niche. Theoretically, only a limited number of pathways are necessary to convert plant fibers into energy for the host[5]. However, the rumen contains thousands of species representing a wide radiation of the tree of life. Archaea, bacteria, protozoa and fungi establish an intricate network, where the plant fibers in the diet are refined to become a nutritious food source for the cattle. Despite decades of research focus, many of the interactions within this microbiome and between it and the host are still largely uncharacterized.

Metabolic cascade of plant fiber deconstruction in the rumen

As the rumen is an anoxic or at least microoxic compartment within the gastrointestinal tract of cattle, the microorganisms inhabiting this environment must rely on fermentation to obtain energy and carbon[4]. To address this, the dominant pathways that convert plant fibers into volatile fatty acids (VFA) will be highlighted, as these represent the most important energy source for the host. The first step of this metabolic cascade of the rumen starts with ingestion of plant fiber polysaccharides such as

cellulose, hemicelluloses, and pectins. Catabolically, these are first degraded by microbially encoded carbohydrate active enzymes (CAZyme) into oligosaccharides and ultimately to their monosaccharide components (e.g. glucose), that fuel microbial populations and processes, mainly by entering glycolysis[4]. Then, glycolysis converts glucose into pyruvate, yielding ATP (adenosine triphosphate) and reduced nicotinamide adenine dinucleotide (NADH), which can be used to drive independent biological processes and anabolism[4]. Most importantly for the cascade, pyruvate is converted into VFAs such as acetate, propionate, and butyrate[6]. Acetate and butyrate are produced via acetyl-CoA, whereas propionate can be produced via succinate and propanoyl-CoA[7]. These metabolic actions can be grouped into three trophic layers[8]: (I) recalcitrant plant polymer degradation, (II) mixed polymer degradation and sugar fermentation, and (III) exclusive sugar fermentation.

Hydrogen is produced as a by-product along most fermentation pathways, including glycolysis, but can feed into various sinks. As dihydrogen (H_2) has a high potential for oxidation, it represents an energy loss if not incorporated into a compound that the host can assimilate[9]. From this perspective, the best-case scenario is to produce propionate by the reduction of fumarate and a succinate intermediary, or to produce acetate. Unfortunately, fumarate concentrations in the rumen may be limiting and the production of acetate may not be energetically favorable, although this depends on the hydrogen partial pressure of the substrate[10]. The most common fate of fermentation-derived H_2 is methanogenesis. In terms of free energy, this process is the most favorable[11], but it represents a loss for the animal as the methane leaves the rumen as eructations, known colloquially as burps or belching. An exemplar alternative hydrogen sink is the Wood-Ljungdahl pathway, where hydrogen acts as electron acceptor for the assimilation of carbon from carbon dioxide (CO_2) resulting in the production of acetate[12]. Hydrogen sinks play an important role in maintaining homeostasis of the pathways that produce it. Which sinks are energetically favorable depends on the hydrogen partial pressure. At lower hydrogen partial pressures methanogenesis is the most favorable option, but at higher pressures alternatives like the Wood-Ljungdahl pathway may become energetically favorable[13].

Species richness and trophic layers of the rumen

While the foundational microbial functions that convert complex plant material into energy-yielding nutrients are well-characterized in the rumen, the elaborate web of microbial populations that perform these essential processes are not wholly understood. Here follows an overview of the currently surveyed diversity of the microorganisms and their central functions that give them a competitive advantage warranting their residence in the rumen microbiome.

Overview and biases

Protozoa, archaea, bacteria, fungi and their associated viruses inhabit the rumen[4,14]. The taxonomic variety of the rumen microbiota might be justified by their ability to express and utilize a diverse set of metabolic functions leading into the digesta-wall-liver axis that ultimately drives the host. For archaea and bacteria, metagenomic assembly ideally recovers high-quality population-level genomes. For eukaryotes, represented by fungi (class Neocallimastigomycetes) and ciliates (class Litostomatea), metagenomic assembly requires unfeasibly deep sequencing, because their genomes are large and complex. For the protozoa, mostly single-amplified genomes (SAG) are available. In this approach, individual cells are isolated from solution and their genomes are amplified using random DNA primers[15,16]. The fungi are also manually isolated, but can be cultured on grass substrates thus not requiring genome amplification[17]. Another factor biasing the presence of genome representations of various species has been dubbed "the great plate count anomaly"[18,19]. This refers to the fact that microbial species can be hard to cultivate in isolation despite their ecological prevalence. This can be for a number of reasons: Some microorganisms might require specific substrates and essential nutrients to enable their growth. These compounds may be unknown for many species or their synthesis may be technically challenging to formulate[18]. Many microorganisms form symbiotic relationships characterized by synthrophy[4], where two species populations are mutually interdependent on metabolites that they each produce. If a substrate does not support such an obligate beneficial interaction between species, or if one species is cultured in isolation, cultivation will be futile. After all, it is probably unnatural for microbes sourced from complex environments to live alone in a static environment. Modern high-throughput cultivation technologies combine single-cell sorting and permutations of many substrate combinations to make it possible to culture many uncharacterized species[20]. In many cases, genomics on isolates leads to higher-quality genomes as contamination and erroneous assembly are the major factors limiting genomic representations[21].

Taxonomic domains of the rumen

The kingdom Bacteria encompasses a broad range of taxa with upper estimates in the thousands of species[22], with many distant phyla. Bacteria can vary overwhelmingly in shape and motility as well as in their metabolic features. Within the rumen of herbivores, predominant cellulose degraders include *Fibrobacter succinogenes* of phylum Fibrobacterota and *Ruminococcus flavefaciens* and *R. albus* of phylum Bacillota[4,23,24]. Exemplar hemicellulose degraders include *Prevotella* spp. of the phylum Bacteroidota as well as *Butyrivibrio* spp. and *Pseudobutyrivibrio*, both from the same family *Lachnospiraceae* and phylum Bacillota[25]. *Prevotella* spp. have been described

as keystone species in the rumen as they perform a wide range of essential degradative processes that influence carbon flow and amino acid metabolism[26]. The products from plant fiber degradation are predominantly fermented into VFAs as well as other intermediates such as lactate, succinate, formate, and H_2 by many different species[27–29].

The protozoa of the rumen are large (40-400μm), complex, single-celled eukaryotic microorganisms of the class Litostomatea (subclass Trichostomatia)[30]. They are also known as ciliates for their cilia-covered outer membrane, which provides a locomotive force enabling them to roam, forage, and even hunt for prey[31]. The protozoa have a relatively large biomass in the rumen, up to 50%, and are found in many ruminant species[14]. They have complex organelles and physiological features such as a mouth-like adoral opening that leads to a tongue-like extrusible peristome, which ingests feed particles into their esophagus. To add to their versatility, they express a broad array of CAZymes, many of which are acquired via horizontal gene transfer from bacteria[15], enabling them to feed on plant fibers. Especially *Epidinium* of Entodiniomorphida encode a vast array of glycoside hydrolase CAZymes[15,32] and are described as actively attaching to and breaking down plant material[30]. Because of their diverse ways of life, protozoa can carry out both bottom-up and top-down[33] control of their environment. As they take part in plant fiber degradation they liberate the constituent fibers for other species to use from the bottom of the food chain—while potentially preying on other microorganisms from the top of the same system. The protozoa may play a significant role in pruning the bacterial populations[33] and this activity represents an important cause of bacterial protein turnover in the rumen[34,35].

Rumen protozoa are known to play a part in forming distinct community profiles. In 1962, Eadie[36] initially described community profile "types" existing in exclusivity, which can be defined in terms of the key protozoal species that dominate them. Along with the gradual refinement of the nomenclature describing the rumen protozoa, these profile types have been continuously updated and still seem to be ubiquitous among rumen ecosystems decades later[30,37]. The most prevalent community profile types are historically named "type-A" and "type-B"[30,36,37]. Type A is defined with the presence of *Polyplastron multivesiculatum* with or without *Diploplastron affine*. The B-type is defined with the presence of *Epidinium spp.* with or without *Eudiplodinium maggii*[30].

Another property of protozoa that comes from their ciliate coat is that they provide a microhabitat for other microorganisms to hitchhike[38]. *Methanobacteriaceae* are known such hitchhikers[31]. As hydrogen is a common side product from the fermentation performed by protozoa, it has been speculated that hitchhiking methanogens utilize this hydrogen as an electron donor in their methanogenesis. Protozoa living in low oxygen environments often do not possess mitochondria but rather hydrogenosomes which are membrane-bound organelles where pyruvate is fermented to acetate and hydrogen[39].

Pyruvate enters the hydrogenosome and interfaces with a pyruvate:ferredoxin oxidoreductase that catalyzes oxidative decarboxylation of pyruvate, forming acetyl-CoA[40]. The ferredoxin is oxidized by hydrogenases utilizing H^+ which is the source of the H_2 being produced. This H_2 finally leaves the hydrogenosome and is available for other metabolic functions like methanogenesis and reductive acetogenesis.

An entirely different taxonomic domain is the one of archaea, which are infamous for their extremophile characteristics, in some occurrences being able to fill environments with high temperatures or salt concentrations[41]. This is enabled by their distinct physiology having cell walls constructed with a combination of ester- and ether-linked lipids, the latter enabling the formation of monolayer lipids spanning the entire membrane of their cell[41]. Archaea are ubiquitous in the rumen of cattle where they act as obligate methanogens[42], reducing CO2, CO or formate to methane[4,43] using H_2 as an electron donor (hydrogenotrophic). Other substrates for methanogenesis may be acetate (acetoclastic) and methylated compounds such as methylamines and methanol (methylotrophic)[44]. The hydrogenotrophic methanogenesis makes archaea metabolically closely linked with many taxonomic groups that produce hydrogen via fermentation. As hydrogenotrophs, archaeal metabolism of H_2 facilitates fermentation homeostasis, which directly affects digestion efficiency. The rumen ecosystem is receiving considerable attention due to the methanogenic activity alone, as it explains a large fraction of the total atmospheric greenhouse gas load contributed by agriculture[4]. The most abundant clades in the rumen are *Methanobrevibacter and Methanosphaera* of the class methanobacteria, as well as *Methanomassiliicoccaceae* which according to Henderson et al. 2015 make up 90% of the archaeal clades observed in the rumen[45].

While less commonly reported in rumen microbiome surveys, anaerobic gut fungi can play instrumental roles in feed digestion. Anaerobic fungi are specialized degraders of plant polysaccharides and often grow and become embedded into the lignin that forms the outer sheath of plant fibers[46]. They produce a highly branched rhizoid thallus which mechanically decomposes the plant fibers[47] on which they attach zoospores to spread on the fiber fragments. These fungi of the rumen belong to the class Neocallimastigomycetes, family *Neocallimastigaceae*, and are distributed among six genera. The Neocallimastigomycetes are only found in ruminants and possess hydrogenosomes which oxidize pyruvate and produce acetate and H_2. They organize their broad range of CAZymes within cellulosomes, which are structurally unique from those known in bacteria[47], and are able to break down complete plant fibers. Currently, limited numbers of rumen fungal species are known[48]. Moreover, their genomes are not well characterized, but recently genomic representations of 12 species, with genome sizes ranging from 56 to 210 megabases, have been made publicly available at the JGI

Mycocosm platform[48]. Because of the way the rhizoids of fungi are embedded in the plant fibers, the isolation of their genetic and metabolic biomolecules is not straightforward, necessitating specific isolation techniques. This leaves the fungi with an undiscovered potential of their diversity and metabolic capacity, to impose undiscovered effects on the rumen environment.

Viruses are obligate intracellular parasites affecting all forms of cell-based life[49]. They inject their genetic material into the cell and manipulate the translational or transcriptional machinery of the living organisms to reproduce their genomes and synthesize viral proteins, before assembling and leaving the cell through lysis or budding to spread to other susceptible cells[50]. Viruses are exceptionally genetically diverse, with genomes consisting of single-stranded or double-stranded DNA, or single-stranded or double-stranded RNA, each type associated with a variety of unique infection dynamics and impacts on their host. The vast majority of double-stranded DNA (dsDNA) viruses, are known as bacteriophages, and exclusively infect bacteria, making them especially relevant in a metagenomic context[50]. The numbers of viral particles in the metazoan gastrointestinal tracts are reported to outnumber bacterial cells by 10-100 times, thereby indicating a dynamic relationship between the bacteria and their viruses, with potential to affect the composition and diversity of microbial populations[50,51]. A recent study on the moose rumen[8] indicates that viruses impose top down control on microbial species that play a key role in carbohydrate degradation.

Plant fibers and their decomposition

Plant fibers

Plant fibers largely function as a structural component. Their tensile strength works in conjunction with the cell turgor pressure, allowing plants to grow tall, thin, and lightweight structures. For this, the necessary properties are strength and resistance against environmental and biological degradation. These properties may not immediately make plant fibers an obvious choice as a main feed ingredient. Nevertheless, as plant fibers are some of the most abundant organic compounds on earth composed of energy-yielding sugars, an animal thriving off their decomposition gains a virtually limitless source of food. **Box 1** presents an overview of the diverse fibers that exist in the plants that are relevant for ruminant nutrition.

Cellulose is a polymer of linked glucose monomers. Adjacent polymers form hydrogen bonds and make up microfibrils[52].

Lignin is the most recalcitrant polymer of the plant cell wall and works as a cement that holds these cellulose microfibrils together. Lignin is a branched polymer consisting of cross-linked phenols[52]. It protects the embedded cellulose microfibrils as it is not susceptible to hydrolysis.

Hemicellulose is a diverse class of polysaccharides characterized as branched and typically shorter than cellulose. Hemicellulose consists of many different saccharides which are linked with glycosidic bonds, and often plays a role in crosslinking cellulose microfibrils[52]. A few key examples include:

- Xylans are one form of hemicelluloses containing xylose saccharides. Homoxylose forms linear chains reminiscent of cellulose, whereas heteroxylose can form branched chains[52]. The difference between these two types is on which atomic positions the polymeric glycosidic links are formed.
- Mannans are another type of hemicellulose with various subtypes: linear mannan, galactomannan, glucomannan, and galactoglucomannan. Mannans have a backbone consisting mostly of mannose and occasionally glucose[53]. Galactomannans specifically can be decorated with branches of galactose.
- Beta-glucans represent a generalization of polysaccharides with a glucose backbone with many beta-glycosidic bonds in various different positions. Overall these are ubiquitous amongst distant species as they are found in the cell walls of fungi, bacteria, yeast as well as in cereal grains such as oat[54].
- Xyloglucans have a backbone consisting of glucose, reminiscent of cellulose. What sets them apart is that the backbone is decorated with side chains consisting of various forms of xylans. Xyloglucans are mainly used for their structural properties in the cell wall, but also show roles in cell signaling and as an energy source in certain seeds[55].

Pectins are structurally complex saccharide polymers with structural functions. The backbone of pectin consists of galacturonic acid, which is an oxidized form of galactose[56]. Homogalacturonans form a linear chain whereas substituted galacturonans form side chains consisting of xylose or apiose.

Box 1: Walkthrough of the common plant fibers and their structures, which are relevant to the rumen.

Carbohydrate-active enzymes and polysaccharide utilization loci

One central dogma of the rumen is that carbohydrate polymers are degraded into shorter chains by carbohydrate-active enzymes (CAZymes) located on polysaccharide utilization loci. As plant fiber polysaccharides are highly diverse in their molecular composition, a diverse set of enzymes is necessary for their monomerization. Highly similar orthologs of CAZymes can be found between distantly related microorganisms and fungi, indicating that these genes have been horizontally transferred[15,17]. Without this mode of adaptation, the high taxonomic diversity seen in the modern rumen would not have been possible[57]. Each CAZyme group catalyzes a unique biochemical function to carry out the intended reaction. **Box 2** presents an overview of the general classifications used in the CAZy database[58].

Glycoside hydrolases (GHs) or glycosidases catalyze hydrolysis of the glycosidic bond within a glycan. These can be classified into exo- or endo acting categories. Exo-acting hydrolases act specifically on the glycosidic bond leading to the terminal saccharide in the polymer, whereas endo-acting can attack glycosidic bonds in the middle of the polymer.

Glycosyltransferases (GTs) use phosphate-activated compounds that donate a glycosyl and thereby catalyze breaking up longer polymers. The phosphate-activated compounds can be saccharide mono- or diphosphonucleotides, alternatively polyprenol pyrophosphates. In the latter case, some glycosyltransferases are classified as glycoside hydrolases.

Polysaccharide lyases (PLs) act on uronic acids typically found in the backbone of glycosaminoglycans where they facilitate beta-elimination of the carboxylic acid to produce an hexenuronic acid, and a reducing end at the point of cleavage.

Carbohydrate esterases (CEs) catalyze de-O or de-N-acylation of substituted saccharides. Enzymes of auxiliary activity represent a large class of redox active enzymes e.g. utilizing lytic polysaccharide monooxygenases (LPMO) which play a role in degradation of lignin[59]. Carbohydrate-binding modules do not cleave polysaccharides but rather bind to these to direct the carbohydrate-active enzymes to enhance their efficiency of the cleavage. They most often co-occur with glycoside hydrolases.

Box 2: Definition of CAZyme classes as defined by CAZy[60].

A polysaccharide utilization locus (PUL) is a genomically linked co-regulated cluster of genes that was originally characterized as the starch utilization system (SUS), but has since been found to engage with a vast diversity of polysaccharides[61,62]. PULs

encodes CAZymes as well as factors relevant for binding, regulation, and uptake of a target polysaccharide. For example, SusD-like, SusC-like and TonB are membrane lipoproteins that work in conjunction to transport a polysaccharide into the periplasm of its encoding cell. SusD-like proteins bind to a free polysaccharide substrate and bring it to the adjacent SusC-like membrane transporter while working as a lid, keeping other compounds from entering the transporter. SusC makes contact with TonB, which is embedded in the inner membrane, to release the substrate into the periplasm where it can be further processed by CAZymes that correspond to the binding affinity of the SusD-like lipoprotein that first attached to the free substrate[61]. This ingenious mechanism has in instances been described as selfish because it allows a cell to gain the full potential of breaking down a polysaccharide, without sharing the cleaved products with others[61].

As carbohydrate polymers can reach lengths in the µm range[55], they cannot be readily transported inside the cells of the microorganisms that are of a similar size as the substrates that they try to eat. As a solution, some species of bacteria use complex structures called “cellulosomes” that are attached on the external cell wall[63]. Cellulosomes use a scaffolding system that can link many different cellulose-binding modules, as well as enzymatic subunits such as CAZymes[60]. Cellulosomes are used to break down the polysaccharides in close proximity to any cell surface protein that take up the shortened polysaccharides. Within the terminology of cellulosomes, a “scaffoldin” unit with incorporated cellulose-binding modules interfaces with exchangeable enzymatic subunits via a system of cohesin-dockerin units. The scaffoldin unit is attached to an anchoring protein via a similar cohesin-dockerin system. The anchoring protein is finally attached to the bacterial cell that encodes the cellulosome. Cellulosomes are observed in distant bacterial classes such as Bacteroidia and Clostridia, notably species *Ruminococcus albus* and *Ruminococcus flavefaciens*, which are ubiquitous amongst ruminants[4].

Physiology of the cattle rumen

For the purpose of converting dietary plant material into the requirements of the host animal, there is a biological axis from the animal's feed, its rumen digesta, through the rumen wall and finally the liver. Cattle utilize a series of auxiliary forestomachs to allow pre-processing and fermentation of the plant fibers. These consist of the reticulum, rumen, omasum and abomasum[64]. The reticulum and rumen together form the reticulorumen[64]. From the esophagus, ingested feed is deposited into the reticulum which is led by the reticulorumen fold into the rumen. The cattle host takes part in breaking down the plant fibers by ruminating: Continuous regurgitation and chewing of ruminal contents, increasing the exposed surface area of the plant fibers, and mixing to

accelerate its decomposition. In this process, the reticulum plays a role in selecting the digesta for rumination via the esophagus[65]. The rumen gains the most attention as it is the largest of the forestomachs and harbors a high microbial diversity. It is non-functional in newborn calves but subsequently develops to constitute 80% of the total stomach volume[66]. Fermentation begins within weeks after birth, when a composition of microorganisms encompassing archaea, bacteria, fungi and protozoa has been established. Peristaltic movements work in conjunction with rumination to process the digesta. The epithelial surface of the rumen wall is covered with papillae, which are small millimeter-range protrusions that drastically increase the effective surface area for absorption of fermentation products into the host[67]. The rumen wall epithelium is able to directly take up VFAs[68], which are reported to make up 70% of the total energy budget of the host animal[4]. To facilitate the fermentation of supplied nutrients, the rumen maintains a supportive environment for the microbial organisms with a temperature of 39.4 ℃ and a pH ranging 5.5-7.0[66]. The digesta proceeds into the omasum which plays a role in removing water from the ingesta[66]. The omasum leads further into the abomasum which is a glandular compartment secreting hydrochloric acids. This environment together with enzymes hydrolyzes proteins and likely inactivates microorganisms. The degraded digesta is retained for a few hours in the abomasum before being transferred into the small intestine[66] where metabolites and proteins, especially, are assimilated[68,69].

Link to the liver

The digesta-rumen wall axis for dietary nutrients continues further into the host via the liver. All veins leading from the gastrointestinal tract, including the forestomachs, make their way through the portal vein into the liver. As this blood is devoid of oxygenated erythrocytes, the liver is supplied with oxygen from the hepatic arteries. As the liver processes the metabolites carried from the portal vein, these are transported via the central vein, which will eventually lead the metabolites to any organ within the host animal. The liver modulates the blood composition by taking up and secreting excess nutrients. For example glucose, which by the use of a bidirectional glucose transporter, facilitates being stored as glycogen within the hepatocytes of the liver[70]. Similarly, the liver responds to blood VFA levels by taking up VFAs, which the liver can either oxidize for the production of ATP or use in the synthesis of lipids. In a rumen-centric holo-omic context, the liver is important, as it represents the final stage of processing in the host before the metabolites from the processed food become available for the host animal.

Multi-omics: From DNA to metabolites

Gaining a mechanistic understanding of how the vast diversity of rumen microorganisms interact and collectively perform the digestive processes that are essential to the host animals' health and nutrition, requires a suite of molecular tools that span both traditional laboratory and *in silico* approaches. In particular, “multi-omic” and “holo-omic” technologies require the analysis of many molecular layers within the samples. Here follows a presentation of the motivation of approaching individual layers with specific technologies, as well as discussing their advantages and pitfalls in the context of their application.

Sequencing of DNA from complex samples

The cornerstone of multi-omics

Genomics is the cornerstone of any multi-omic study as information about genes encoded within a living organism is stored and propagated evolutionarily as double-stranded DNA. Transcribing gene potential, RNA polymerase produces transcripts from the coding regions of the genomes. Curiously, these transcripts confer the chemical ability, as structural ribozymes, to perform the translation of their own kind into proteins. The ribozymes can act in diverse ways: As subunits of the ribosome and even as the transfer vessels that put into place the amino acids corresponding to the triplet nucleic acid codons encoded in the transcripts themselves. These amino acids form peptide bonds that hold together the backbone that defines the possibilities for folding and further post-translational modifications. Finally, a mature protein with any imaginable biochemical function is formed. All of the aforementioned molecules are important fields of study on their own and will be discussed below.

Isolation of DNA and other biomolecules

Before it is possible to investigate a given holobiont at multiple molecular layers that tell the story of biological interactions and their metabolic consequences, the corresponding biomolecules need to be collected. Environmental samples containing microorganisms often carry diverse features such as enzymatic inhibitors, humic acids, or biofilms (**#1**), produced as part of the metabolism that enables life of the organisms themselves. Current sequencing and mass spectrometry technologies are highly sensitive to contamination of such organic compounds, which means that in order to sequence these biomolecules, they must first be isolated chemically. Inside living cells, genomic DNA is archived in megabase-range chromosomes, which are potentially wound into complex chromatin structures[71]. In order to access this DNA for

sequencing, the cells must lysed to release the DNA without fragmenting it excessively. This is because every time a DNA strand is needlessly fragmented, a bioinformatic algorithm must later take the responsibility to stitch together those fragments, which is not a trivial task[72].

Bead beating represents one approach to expose the biochemical compounds of complex samples: Violent shaking at several meters per second, with the addition of glass beads that physically break apart the cells and any cell-enclosing capsules. This frees the contained DNA from the individual organismal cells, but also potentially physically fragments the DNA in the process. The tuning of this bead beating is challenging because different lifeforms may use different cell-enclosing capsule structures. One striking variability within bacteria lies in the composition and physiology of their cytoplasmic membrane. Where monoderm bacteria lack an outer membrane and compensate with a peptidoglycan capsule[41], diderms have an outer membrane forming a periplasmic space between it and the inner membrane[73]. Protozoa are large and fragile, thus more susceptible to external stressors[74]. Fungi of the rumen grow into the plant fibers that they degrade[17]. This means that to access the DNA of such cells, the plant fibers must become disassembled first. This systematic difference in DNA availability calls for multi-layered or fractional extraction techniques where individual batches are performed to obtain a good representation of individual taxonomic groups. Failure to take this into account will lead to a bias in the relative abundances of these groups which is especially problematic in microbiome studies because DNA of all lifeforms is desired concurrently. These issues are also relevant for extraction of genetic and metabolic biomolecules of other omic layers (RNA, protein, etc.) calling for taxonomically stratified extraction of these as well.

DNA sequencing

Several fundamentally different technologies are currently widespread for sequencing of genomic DNA. One popular technology is sequencing-by-synthesis, where a fluorescent terminating nucleotide is paired to a single-stranded DNA molecule. Before this terminating nucleotide is swapped with an equivalent but non-fluorescent replacement, its color is recorded. As the fluorophore has a specific color for the four types of dNTP necessary for DNA synthesis, it allows reading the complementary nucleic acids one by one. In Illumina's implementation of sequencing-by-synthesis, DNA is fragmented to a desired length. These fragments are attached on a flow cell and grown into clusters with a polymerase chain reaction (PCR) bridge amplification process. This allows the sequencing of both ends of the clusters leading to paired-end reads. Typical sequencing on the Illumina Novaseq platform produces 80-6000 Gbp reads with an N50 of 250 bp[75] and an error rate of 0.1%[76].

Recently, long-read sequencing technologies have become more accessible and are increasingly being applied to microbiome studies. PacBio is another implementation of sequencing-by-synthesis, however, their process is described as single-molecule real-time sequencing. Here a circularized DNA fragment is captured within nanometer-range wells referred to as zero-mode wave-guides, where anchored DNA polymerases perform the incorporation of fluorophores with dNTP-specific wavelengths that allow recording of the incorporated complementary bases. Typical sequencing on the PacBio HiFi platform produces 15 Gbp reads with an N50 of 15.4 kbp and an error rate of 99.93%[77]. This type of sequencing offers the low-error-rate advantage of sequencing-by-synthesis but avoids the length limitation in Illumina's implementation.

A different sequencing concept altogether is nanopore single stranded DNA sequencing. Here, DNA is passed through a pore of nanometric dimensions while a circuit encircling the pore reads the fluctuations of current induced by the passing of a single stranded DNA molecule. In Oxford Nanopore Technologies' (ONT) implementation, adaptors are ligated to DNA fragments. These adaptors bind to DNA-unwinding enzymes that sit on top of nanopores which are embedded on a flow cell. As the correlation between current flow and nucleotide is not straightforward, a considerable part of the advancements in this technology is the development of better basecalling algorithms which are currently based on recurrent neural networks[78]. A typical sequencing run on the ONT flow cell R10.4 platform produces 14 Gbp reads with an N50 of 5.6 kbp and an error rate of 98.11%[77].

Metagenomics

Metagenomics is an approach where genomes of microorganisms are analyzed directly from environmental samples, which is crucial for understanding their overall dynamics in terms of abundances and genomic repertoires. Due to the enormous variation of microorganisms, it is beneficial to reconstruct their genomes *de novo* without relying on previous knowledge such as reference genomes. In *de novo* assembly, sequence reads are overlapped by using various alignment algorithms and heuristics[72]. By overlapping the reads it is possible to construct a graph that follows one or more paths through the assembled reads. When one such contiguous path is identified, it is referred to as a "contig", which ideally represents a complete chromosome or plasmid, although in practice such contigs are often fragmented. Contig fragmentation can be due to low coverage when there are not enough reads to resolve a complex region, or when it cannot be distinguished whether sequence variation comes from within or between clonal populations. Short reads may not be able to resolve repetitive regions when the

repetitive region is longer than the reads which is largely the reason why long reads have recently become popular in metagenomics[77].

The goal of genomic binning is to bin the assembled contigs into groups representing individual species or strains. To achieve this, binning takes into account several statistics like k-mer frequencies, differential abundance[79] and in the future possibly even species-specific DNA methylation patterns[80]. After successful binning, the bins are referred to as MAGs. A common approach to quantify MAGs that have been recovered from a metagenomic context is to map the original metagenome sequences back onto these. This gives an indication of the relative abundances of the individual microbial species or populations. As MAGs represent ideally complete genomes of a specific species or population, it is in principle possible to characterize these biologically with bioinformatic analysis. In practice, genes are annotated on the genomes by searching gene databases for similar sequences. Following recent developments in protein folding[81,82], it is possible that in the future structure-based search will play a large role in how genes are annotated, avoiding the alignment of sequence representations[83].

Genomes and genes of microorganisms form a central pillar in multi-omic studies. In order to study RNA (i.e., transcriptomics) and proteins (i.e., proteomics), the only way to definitively identify the origin of a specific transcript or protein is to map it to a complete genome that has been taxonomically identified. A similar point goes for metabolites, which can be correlated via pathways inferred from gene sets in the genomes. This is why, in any metatranscriptomic, metaproteomic, or metabolomic ecological study, metagenomics should be included and highly prioritized.

Metatranscriptomics

Transcriptomics enables measuring which genes are being expressed in a biological system, whether investigating an individual cell or an ecological sample. The latter case is referred to as metatranscriptomics. Transcriptomic analysis is achievable with the sequencing of messenger RNAs (mRNA) which communicate genetic instructions between DNA and protein. As mRNAs are usually short-lived relative to the proteins they encode[84,85], a qualitative interpretation of transcriptomics is that it gives a view into what is being changed in the system, rather than what biologically active factors (often proteins) are present. This is an interpretation of the often negligible correlation between transcripts and proteins. (Meta)transcriptomics is usually sequenced with a DNA sequencing platform, so the isolated RNAs are typically reverse transcribed into complementary DNA (cDNA) at first. Within cells, ribosomes play a key role in translating mRNA into protein, and they are numerous in microbial community

samples. Since ribosomes themselves contain RNA, the default is that much of the sequencing budget will be wasted on covering these redundant sequences. Therefore ribosomal depletion plays a key role in transcriptomic workflows[86].

The taxonomic domain-specific architecture of genes enables chemical selection of specific clades. One such example is polyadenylation, which is part of mRNA maturation in eukaryotes, where a tail of adenosine monophosphates is added to the end of an mRNA. By chemically enriching for these while isolating mRNA, it is possible to narrow down the sequencing coverage budget to a eukaryotic population, or to deplete it for a non-eukaryotic focus[87].

Metaproteomics

Proteins, particularly enzymes, are the major chemical drivers responsible for carrying out metabolic processes within all biological systems[82]. Metaproteomics enables the detection and quantification of proteins within complex environments, offering insights into the ongoing biology[88]. Extraction of proteins may be susceptible to biases, especially because many proteins are membrane-embedded or -bound, which may necessitate special extraction techniques. In addition, the differences in cell characteristics that exist across different taxonomic clades, for example the thickness of the bacterial cell wall (Gram positive or negative), potentially affects amenability to protein extraction techniques[89], and even more so for fungi and protists present in the sample.

A common approach to metaproteomics is bottom-up analysis, where the extracted proteins are first cleaved into peptides using a protease (e.g. trypsin). The peptides, typically a mix of hundred thousands, are then separated based on their hydrophobicity using liquid chromatography and analyzed by mass spectrometry. Within this process, the peptides transition from a liquid phase to the gas phase and acquire a charge (e.g. using electrospray ionization) and can then be guided into the mass spectrometer. Various types of mass analyzers exist that can be part of a mass spectrometer, for example a Time-of-Flight (ToF) analyzer, where peptides are accelerated up until the entrance of the ToF, followed by a movement through a so-called flight tube having a defined length. The time that the peptide ions spend in this flight tube is proportional to the mass to charge (m/z) ratio of each ion, which is recorded on a sensor, producing a mass spectrum. The use of tandem mass analyzers allows the selection and fragmentation of peptides (producing so called MS/MS spectra), and newer techniques such as trapped ion mobility spectrometry (TIMS), which allows additional gas-phase separation of ions, adding one more dimension to the m/z spectra, ultimately enable a

deeper proteomics analysis, i.e., identifying more proteins from a complex sample such as a metaproteome.

Unfortunately, a recorded spectrum can not be trivially mapped to a protein[90]. To identify the protein of origin, the observed *m/z* spectra are matched with theoretical spectra. These can be predicted *in silico* from a reference database of proteins expected to largely cover the proteome of the sample in question. The design of this reference database can have profound effects on the output, as missing proteins will not be identifiable. A common issue when matching observed and predicted spectra is to control the number of false matches. One solution, known as the target-decoy method, is to reverse the sequences of the proteins in the reference database, and adjust the matching threshold until a targeted false discovery rate (FDR) of these reverse sequence spectra is achieved, e.g. 1%[91].

Performing metaproteomics comes with some limitations. Because of the huge diversity and dynamic found in ruminal samples, metaproteomics often contain a multitude of taxa and many homologous proteins between species. This is further complicated by initial protein extraction and the challenges of matching spectra. Oftentimes, due to the high number of similar proteins between species, metaproteomics identifies protein groups instead of unique singular proteins, and these groups often consist of proteins from several species, making the inference of origin difficult. A biological artifact of protein abundances is that they are typically skewed, calling for transformation to approximate a symmetrical distribution which is necessary to perform parametric statistical tests. Missing values can be a concern in metaproteomics, especially when the metaproteomic data is scrutinized with statistical methods that do not handle sparse data well. In a perfect world the number of sample replicates could be increased to fill these missing values. A cheaper solution to the missing-value problem is to use imputation, either to naively replace the missing values with samples from a distribution around the detection level of the instrument or to make sample-specific statistical models for each protein where information about correlated proteins in other samples are used to infer a missing value[92,93]. Although the mass-spectrometry-based method has limitations both regarding missing values and indistinguishable protein groups, it still offers a unique view into the biological activity of complex samples. Moreover, it is less prone to technical biases arising from sample storage and molecular degradation and can capture biological information from all domains of life simultaneously, which is important in transdomain ecosystems such as the herbivore rumen.

Metabolomics

Even with perfect information on upstream molecular layers (DNA, transcripts, and proteins), it is currently not possible to infer whether expressed and present pathways are regulated to be active in time and space. Metabolomics offers insight into the intermediates and products that follow from the active pathways within a biological system. In a holobiont setting, a unique feature of metabolomics is that it allows testing hypotheses on how metabolites transfer across the host-microbiome boundary. Overall, several completely different technologies can be used: Untargeted metabolomics is reminiscent of the procedure used in metaproteomics. Here a gas chromatography (GC) or LC-MS/MS system separates and records spectra for extracted metabolites[94]. The extraction method used has profound effects on the distribution of the extracted metabolites, and the method should be selected based on the type of metabolites that are most relevant for the greater picture. A reference library of known metabolites is used to match spectra, which means that there will be similar issues of missing values and redundancy, as discussed for metaproteomics. Despite these challenges, mass-spectrometry-based metabolomics still represents a useful approach to investigate complex interactions in a holobiont setting. In targeted metabolomics, only one or more carefully selected metabolites are isolated and quantified by using highly specific approaches. Typically, these are products that are relevant for the study system, for instance plant fibers[95], VFAs, and gasses[96].

Holo-omics: Integrating layers across the host-microbiome boundary

Collecting samples from the host and its microbiome and performing multi-omics yields two separate datasets which must be integrated. In this context, holo-omics can be considered a special case of multi-omics, and here it will be presented in context to a holobiont application.

The general challenge that underpins holo-omics is that the datasets are large, which in turn means that they contain a lot of background variation which is not relevant for specific host-microbiome interactions. To make matters worse, as multi-omics on both host and microbiome sides is in itself expensive, the sample sizes are usually low. This leads to a problem where a few samples are used to infer the relationship between a high number of factors. This can produce spurious results (false positives) if multiple testing is not taken into account, but also hinder the identification of biological relationships (false negatives) when too much regularization is applied.

Several methods have been developed to counteract the discrepancy between small sample size and large number of features. Dimensionality reduction is motivated by the need to bring data with many independent features (i.e., dimensions) into a low number of projections, each of which represents a weighted average of the original features. A classic dimensionality reduction method is principal component analysis[97] (PCA). This method linearly projects the data into orthogonal projections. Multidimensional scaling (classic) is reminiscent of PCA but allows the use of a dissimilarity matrix with pairwise dissimilarities between samples instead of raw data[97]. Among other methods, many complex algorithms exist[97]. To name a few, popular methods include t-distributed stochastic neighbor embedding (t-SNE)[98,99], and uniform manifold approximation and projection[100] (UMAP). These are powerful methods to project high-dimensional manifolds in more manageable projections, but their complexity may challenge the interpretation of biological relationships.

An alternative to dimensionality reduction is to use correlation networks and extract connections within these that reflect the underlying biological interactions[101]. Omic datasets may contain large numbers of genes, and many of these genes may be expressed together when they perform a linked activity (e.g. encoding the genes for a specific pathway). To make complex data more interpretable, it may be useful to cluster these co-expressed genes together, and this is where correlation-network-based methods can be useful. One example of a correlation-network-based framework is weighted gene co-expression network analysis[102] (WGCNA). Here, a series of algorithms process normalized and filtered gene expression data from multiple samples into clusters, each represented by idealized representative eigengenes. First, pairwise correlations are calculated between all genes across the samples. These correlations are soft-thresholded by raising them to a power that allows the distribution of distances to follow a power law, resulting in the network to become scale-free[103]. This soft-thresholding step also works as a means of filtering, as it amplifies stronger correlations. On the soft-thresholded correlations, a network proximity measure is applied, typically the topological overlap measure (TOM). The TOM takes into account shared connections among correlated genes. On these measures, hierarchical clustering, using average linkage clustering, constructs a binary tree of genes whose branches can be cut to yield exclusive clusters. Finally, the genes inside these clusters are represented with cluster eigengenes, which captures the expression pattern of that cluster. The eigengenes are calculated as the first principal component (via PCA) of the gene expression within their cluster. These eigengenes can then be correlated to phenotypic traits of the original samples, or cross-correlated across the host-microbiome boundary to locate potentially interacting clusters of genes between

these two partners. WGCNA has been used to characterize roles of important genes in many systems[104,105].

Networks can also represent chemical rates within and between pathways of complex biological systems[106]. By modeling these, it is possible to gain insight into how larger systems interact upon stimuli. Metabolic networks can be represented by a stoichiometric matrix with metabolites and reactions and a corresponding reaction rate (flux) vector, and the product of these is the mass balances of the metabolites. This can be used to model the complete pathways of entire genomes, thus forming genome-scale metabolic models (GEM). Constraint based models (CBM) represent the most powerful method for large-scale metabolic network reconstruction[107]. CBMs rely on assumptions of steady-state which are justified by fast metabolism being relatively invariant to regulation[108]. This simplifies the dynamics of the model, leaving a linear system of equations that is computationally tractable. By using flux balance analysis (FBA) it is then possible to solve for the flux vector by optimizing a certain objective[109] (e.g. growth or the production of a specific metabolite). The assumption of steady-state does impose some limitations on the use cases of metabolic models, but it still enables valuable insights into the activities of metabolic networks.

The main motivation to use computational methods in multi- or holo-omics is to gain insights into the interactions across the host-microbiome boundary, but also to understand the marginal variation within either host or microbiome. Choices related to the computational workflow, statistical approaches, and data visualization have profound effects on the potential to gain biological insight. As holo-omics is an emerging field there are many other methods and approaches that could be taken as alternatives to the ones presented here.

2. Aim of the thesis

The holobiont perspective entails that the microbiomes of animals should be taken into account when considering the biology of the host[110]. While holobiont theory has existed for decades[111,112], technological advancements have created new insight into how microorganisms fulfill important roles within the confinements of their host. Especially in the gastrointestinal tract of bilaterians where microorganisms perform crucial metabolic functions that enable the host to assimilate nutrients and energy from the feed[4,113] (**Paper #3**). It is known that external factors, such as diet, can affect the composition of these microorganisms[114], but fundamentally their response cannot be predicted as a comprehensive understanding of their dynamics does not yet exist (**Paper #3**). This problem is further exacerbated by the fact that abundant clades of microbial species within these environments are still largely uncharacterized[113]. Moreover, in humans it is indicated that more subtle factors, such as living conditions and genotype of the host, can direct changes upon the composition of these microbiomes[115,116]. This suggests direct interactions between the host and microbiome, which means that the microbiome can possibly be used as a metabolic modulator of the host. Access to this knowledge could fundamentally change understanding of holobiont systems, and open a lasting potential to enhance production and health of many animals, particularly those in agri- and aquaculture. To enable insight into these abstract hypotheses, the rumen of cattle and adjacent host tissues, the rumen wall and liver, were used as an exemplar holobiont system to build a high resolution dataset. The cattle rumen is a large anaerobic chamber, where ingested plant fibers are degraded and fermented by a multitude of microorganisms that have co-evolved for millions of years with their cattle host. Since these microbial organisms are key to degradation, the host depends on them to unlock the nutrition and energy from the plant fiber based diet through the metabolic cascade that takes place[4]. The aim of this thesis is to assemble a rumen-centric view of the cattle holobiont, to put together missing links that enable a deeper characterization of both within-microbiome as well as microbiome-host interactions. To accomplish this, the thesis will develop (**Paper #1–2**) and apply multi-omic technologies on both sides of the host-microbiome boundary, and use advanced methods to achieve an integrated holo-omic view (**Paper #3**) of the study system (**Paper #4**). This will allow tracking of the metabolic cascade from CAZyme directed plant fiber degradation, through production of fermentation end products and to investigate how these are assimilated by and possibly affecting the biology of the host.

3. Main results & discussion

The papers outlined in this thesis cover the efforts taken to apply molecular omic techniques to the biological context of the ruminant, which is an exemplar holobiont system. The papers can be classified either as methodological development, or biological discovery. The methodological developments are necessary as multi- and holo-omics requires insight into many technical topics which have limitations in their ways to present the underlying biological patterns. This is both in terms of analyzing individual omics layers in the methods papers **#1** and **#2** but also to study the statistical methods that allow integration of these layers across the host-microbiome boundary in the review paper **#3**. The research paper **#4** encompasses the application of the methods that this thesis has collected and built along the way to showcase the biological interpretations of the holobiont.

Paper #1 - Long-Read Metagenomics and CAZyme Discovery

In a microbiological environment where plant fibers play the main role as contributors to the energy demands of the living organisms, CAZymes are key. **Paper #1** outlines a comprehensive workflow (**Paper #1 fig. 1**)—from complex samples to detection and classification of the CAZyme genes in the microorganisms that encode them. To understand how these microorganisms utilize the energy and carbon assimilated from the plant fibers, their overall metabolic capacity must be determined. Thus, the review describes the methodological steps to first obtain DNA from complex microbiome samples and how to bring metagenomic reads from modern long-read DNA sequencing technology into genomic representations, as well as how these can be binned and dereplicated to finally construct MAGs. Arguing that the disposition and quality of these MAGs can have profound effects on downstream analyses, the review describes how to assess their quality. Several tools are discussed that can be used to annotate the metabolic functions on their genomes and how their taxonomies can be identified. Finally, by the use of specific publicly available scripts, the paper shows how the acquired taxonomical and metabolic functional annotations can be visualized together (**Paper #1 fig. 2**).

One of the major reflections from **Paper #1** was that installing and applying all the computational tools necessary for MAG annotation and CAZyme characterization is an unnecessarily complicated process, which could be better automated. This led to the idea that many publicly available tools could be integrated into a single user friendly pipeline (**Paper #2**) that computes and presents the results in a way that enables direct biological interpretation. Problems related to installation and runtime issues mean that time that could be spent analyzing data and interpreting these to understand biology is

instead spent on troubleshooting, which is analogous to technical issues experienced in the wet lab. This led to the conceptualization and implementation of the CompareM2 software package (**Paper #2**), which includes core community adopted tools to analyze MAGs and the functional potential of these on several qualitative and biological levels, and makes it possible to gain biological insight without unnecessary overhead from the user.

Paper #2 - CompareM2 is a genomes-to-report pipeline for comparing microbial genomes

From the increasing number of MAGs being recovered from metagenomic studies, the need arises for a complete and efficient pipeline that accepts MAGs and analyzes them on many qualitative and biological levels: Quality control, functional annotation, metabolic pathway analysis, phylogenetic analysis, and core- and pan-genome partition characterization, including clustering of genes (**Paper #2 fig. 2**). This pipeline is useful for analyzing large inventories of genomes and is designed to allow direct biological interpretation. The pipeline consists of many open source tools which are integrated using Snakemake—a framework for reproducible data analysis that allows efficient parallel computation on multi-core and multi-node high-performance computers (HPC), which is crucial for scalable performance on large datasets. Many of the integrated tools use databases which the CompareM2 pipeline automatically downloads and installs. Within this pipeline, a portable report document was implemented to dynamically embed the main results of the pipeline, allowing for quick access to the biological results and their interpretation. As CompareM2 is distributed on Bioconda and uses a containerized Docker compatible Apptainer image that speeds up installation, it avoids many of the issues common to installing large sets of bioinformatic software packages.

To showcase the benefits of the CompareM2 implementation a quantitative comparison was performed (**Paper #2 fig. 1**), analyzing two predominant taxonomic groups from the metagenomic recovery of the digesta samples in the animal trial from **#4**. The dataset was selected to reflect the ability of the pipeline to process both bacterial and archaeal MAGs, namely bacteria *Prevotella* spp. of class Bacteroidia as well as archaea *Methanobrevibacter* spp. of class Methanobacteria. As there is a lack of similar tools within microbial ecology, the comparisons between CompareM2 and other tools relied on tools developed for clinical microbiology. Contrasting CompareM2 against two other popular, somewhat—but to the highest degree possible—comparable tools, it is found that our implementation is 4.1–7.2 times faster (wall time) for analysis of the *Methanobrevibacter* spp. MAGs and 3.1–7.8 times faster for the analysis of the

Prevotella spp. MAGs (**Paper #2 table 2**) on a multi-core computer with a high potential for computational parallel optimization. This supports the conclusion that CompareM2 efficiently integrates the included bioinformatic tools. As speed is not the only important factor for such pipelines the qualitative differences were also compared (**Paper #2 table 1**), the results indicating that CompareM2 implementation is both easy to install and use, and enables straightforward biological interpretations due to its dynamic report. CompareM2 follows the open source culture, paving the way for user contributions to adapt the tool for metagenomic analyses in the long term.

Holo-omic studies require thorough analysis of several molecular layers, which means that for a holo-omic study to be feasible within a short time frame and with a limited budget, streamlined methods must exist to efficiently analyze these individual layers. In this context, CompareM2 represents a contribution to streamline one such aspect of studying a holo-omic system, moving the field forwards in the larger quest of analyzing host-microbiome interactions.

Paper #3 - Integrating host and microbiome biology using holo-omics

Paper #3 presents a comprehensive review of the computational methods and bioinformatic tools that are available to perform integrative holo-omics (**Paper #3 fig. 1**). In particular it explores the biological implications of holobionts, and showcases how holo-omic methods can capture host-microbiome interactions. The review defines holo-omics as a special case of multi-omics where both the host and an adjacent microbiome are analyzed. This entails applying multi-omics methods (**Paper #3 fig. 3**) but taking into account that completely different databases of genes and pathways must be integrated across the host-microbiome boundary. Holo-omics can be considered a big data problem where the datasets are too large to be analyzed using conventional methods, contrasting biological observations directly by frequentist inference or bayesian statistics. Instead the data must be filtered and reduced, or connected in a way that allows underlying signals to percolate through and become available for statistical tests, enabling biological interpretation. This problem is related to the curse of dimensionality where a high number features explain the differences within a small number of samples (**Paper #3 fig. 2**). Another challenge with these large datasets is overfitting, where a model captures a signal that is merely spurious and does not bear any biological relevance. First, **Paper #3** presents an idealized holobiont system, gives an overview of known host-microbiome effects, and describes qualitative differences between known holobionts (**Paper #3 table 2**). It then goes on to present and contrast the qualitative differences between state of the art statistical methods that can be

broadly classified into dimensionality reduction or network analysis methods. The main quality of these methods is their ability to define clusters of features with similar presence or expression patterns across samples or over time.

Dimensionality reduction methods can be applied to individual layers which allows identification of distinct clusters that can be compared across a host-microbiome boundary. **Paper #3** presents principal component analysis (PCA) as a ubiquitous method that can be used to gain an overview of the overall variation within a single layer. It can be refined with principal coordinates analysis (PCoA) that allows transformation of the observations via any relevant distance method which can increase the sensitivity towards certain taxonomic domain specific relationships, e.g. the Bray-Curtis dissimilarity that quantifies differences in species composition between samples. The advantage of these methods is that they create a compressed view of the data by allowing to select only a few relevant orthogonal principal components. There are also non-linear alternatives like t-distributed stochastic neighbor embedding (t-SNE), non-metric multidimensional scaling (NMDS) and uniform manifold approximation and projection (UMAP) which all allow the visualization of a multidimensional manifold within a smaller number of dimensions. Other methods within dimensionality reduction include non-negative matrix factorization (NMF) and multiset correlation and factor analysis (MCFA) (**Paper #3 fig. 4**). NMF factorizes the observed data into two matrices that ideally collectively have fewer total cells. An interpretation of these resulting matrices is that one of them represents the combinations of archetypes and the other represents the linear combinations of the learned archetypes, which resemble the observed samples. This allows the method to represent a large number of samples by bringing attention to the common combined features.

Network methods are based on graphs that represent relationships between interacting entities within a system. WGCNA builds a network of co-occurring genes or expressed biochemical features. From the adjacencies of such a network it clusters these compounds based on their topological overlap. Within each of these clusters, eigengenes are defined. These capture the principal component which compresses the data into a single orthogonal combination of the contained features. Finally these cluster-representing eigengenes can be used to link metabolic functions across layers, and may present more interpretative alternatives to other clustering methods. Another network based method discussed in the review is transkingdom network analysis (TkNA) which is designed to allow causal inference of master regulators within a treatment-response experiment. It uses bipartite betweenness centrality, a common graph theory metric, to identify nodes within the correlation network with a high flow of information. **Paper #3** also outlines integration tools based upon partial

least-squares discriminant analysis (PLS-DA) and co-inertia analysis (CIA), and presents latent Dirichlet allocation (LDA) as a network based method with a potential to cluster relevant biological groups within holo-omic studies, which has yet to become adopted by the community. Conclusively **Paper #3** reveals that one of the major limitations of statistical methods for holo-omics is the trivial problem of causal inference. Inferring which side of the holo-omic boundary directs a factor, can prove challenging. Overall, this methodological review offers a springboard for integrating the holo-omic data derived from an animal experiment, which is showcased in the final chapter of this thesis (**Paper #4**).

Paper #4 - Protozoal populations drive system-wide variation in the rumen microbiome

The final paper of this thesis (**#4**) presents a biological research effort to bring together all the methods and perspectives that have been reviewed, tested and developed in papers **#1**–**#3**. To gain insight into a relevant holobiont model an experimental feedlot trial was performed in collaboration with Scotland's Rural College (SRUC). In total 80 cattle, 40 of an Aberdeen-Angus cross, and 40 of the breed Luing, were sampled for their rumen digesta via esophageal tube collection at five time points over the duration of the trial, and directly sampled from their rumen contents, rumen wall and liver tissue upon slaughter. Before slaughter the animals were subjected to rigorous measurements of various key performance traits as well as their methane yield, which had a range of 20.5–27.3 g/kg dry matter intake (**Paper #4 fig. 1**).

From the rumen digesta of all animals at slaughter, microbiome DNA was isolated and taxonomically profiled using 16S rRNA gene amplicon sequencing analysis of archaea and bacteria. Taking into account methane yield and breed, 24 animals were selected for in-depth characterisation of host and microbiome samples. The 12 highest-emitting and 12 lowest-emitting animals were chosen for multi-omic analysis of three sample locations: host-liver and -rumen wall, as well as rumen digesta which represents the microbiome. Rumen samples collected from these animals were analyzed with long read sequencing using Oxford Nanopore Technologies' R10.4 flow cell. For both host tissue and digesta samples, short read sequencing was performed using the Illumina Novaseq S4 platform. The long reads from the rumen metagenome were hybrid-assembled with short reads before binning and dereplication, which resulted in 700 MAGs of at least medium[21] quality distributed as 44 archaeal and 656 bacterial, together representing 122 unique genera. All archaeal MAGs originated from *Methanobacteriaceae* encompassing the genera *Methanobrevibacter* and

Methanosphaera. Among bacterial MAGs, the most prevalent phyla were Bacillota and Bacteroidota. Between these and other phyla the most prevalent genera were *Prevotella*, *Cryptobacteroides*, *Sodaliphilus* and UBA4372. Among all of the MAGs, most were of an uncharacterized species, and for 184 of the MAGs, species were not assigned as there was no match in the genome taxonomy database[117] (GTDB). If this is not due to the incompleteness of these MAGs, it points to a general lack of characterization of the archaeal and bacterial species of the rumen. The median completeness and contamination of the MAGs as reported by CheckM2[118] was 84.7% and 2.68%, respectively. All MAGs were annotated to predict open reading frames to assign function, and were consolidated into a database used for mapping the functional data generated from proteomics and transcriptomics. To this database, the genes from eukaryotic sourced genomes (protozoal SAGs, fungal genomes, and a cattle host reference genome)[15,48] were also added. In total, this database consisted of 4.2 M proteins with an average length of 426.8 amino acid residues sourced from a total of 777 archaeal, bacterial, protozoal, cattle host, and fungal genomes.

For functional analyses, RNA sequencing was performed on host tissues and the microbiome to track the expression patterns of the vast array of populations that constitute the cattle holobiont. The liver samples yielded 70.5 M reads per sample, on average 91% of these mapped to the host genome. The rumen wall samples yielded 78.3 M reads per sample. On average only 17% of these mapped to the cattle host genome and 57% mapped to the archaeal and bacterial genomes of the digesta. This corresponds to the fact that remains of digesta were not rinsed off the epithelial surface of the rumen wall samples before RNA extraction. The digesta samples yielded 112.8 M reads per sample, on average 2.3% mapped to archaea, 38.4% mapped to bacteria, 5.2% mapped to fungi and 53.9% mapped to protozoa. All sequencing efforts, 16S rRNA gene amplicons, shotgun DNA, and RNA, were undertaken in collaboration with DNASense ApS, Denmark.

The final translation of transcripts into protein was evaluated with metaproteomics of the digesta samples and proteomics of the liver and rumen wall samples, all of which were performed with liquid chromatography-tandem mass spectrometry (LC-MS/MS) using the Bruker timsTOF Pro platform. The digesta, rumen wall and liver sample measurements produced a range of 25k–40k scans leading to a range of 40k–81k peptide-to-spectrum matches (PSM), which were mapped to the genomic database using Fragpipe[119]. Of the protein groups recovered from the rumen digesta samples collected during slaughter, 0.9% mapped to the host, 1.1% mapped to archaea, 47.2%

mapped to bacteria, 0.4% mapped to fungi, and 50.5% mapped to protozoa. Since this database was larger than the target design for the software, a custom workstation with large quantities of RAM and swap storage was required to complete this computational step. For further analysis of the proteomic and metaproteomic samples, network analysis was undertaken upon the (meta)proteomic intensities. To circumvent limitations imposed by missing values, these were imputed using missranger[92]. The imputed data were analyzed with WGCNA to cluster detected proteins into modules[102]. A large number of these modules were not correlated with any known phenotype characteristic or sample grouping of the animals. Cross correlations of modules between host and microbiome allowed the identification of co-expressed protein groups across the host-microbiome axis.

To better understand how host and microbiome expression was contributing to the greater phenotype of the system, untargeted metabolomics was performed using a HPLC-MS/MS platform. This analysis identified the intensity of 591 unique metabolites in the digesta, 197 in the rumen wall, and 326 in the liver samples. As added layers, metabolites representative of pre- and post- digestion were measured including plant fiber abundances using microarray polymer profiling[95] (MAPP), in collaboration with Newcastle University. This led to the identification of 43 unique plant fiber and associated protein targets across the digesta samples. Finally, for fermentation end-products, six volatile fatty acid targets: acetic acid, propionic acid, iso- and butyric acid, iso- and valeric acid were collected.

Based on the extensive multilayered datasets to determine host-microbiome interactions, dimensionality reduction techniques identified a well separated and bistable clustering between two groups of animals (**Paper #4 fig. 2a-d**). Intriguingly the pattern was not correlated with any phenotypic characteristics ascertained during the trial, nor the genetic background (i.e., breed) of the animals. To reaffirm that these patterns were not related to any technical effects e.g. handling of the animals or an analytical sample batch, rigorous statistical tests were performed but did not reveal any correlations that would suggest any known technical factor to be responsible.

Curiously, the bistable clustering pattern was ubiquitous across several omic layers that were analyzed for the 24 animals across their rumen digesta, rumen wall and liver samples. For the rumen digesta metatranscriptome, rumen wall metatranscriptome, and digesta metaproteome, the first principal components unambiguously separated across the two clusters. In other words, the two unidentified clusters were the largest

contributor of variation in these layers. The congruence between the molecular layers shows that the larger part of metabolism was likely affected by this compositional difference. However, closer examination of metabolite data from the rumen digesta, the rumen wall and liver revealed less impact with PC-driven variation only being detected at PC4, PC3 and PC10, respectively. The lack of functional impact in the host was furthermore supported by host expression data with limited cluster driven variation being observed in the proteomes (PC4) and transcriptomes (PC3) of the rumen wall epithelial samples as well as the transcriptome (PC5) and metaproteome (PC15) of the liver.

Since the greatest variation for the clustering patterns was observed within the microbiome, these layers were explored further at the domain level, which identified two rumen community types (RCT) A and B that demonstrated large differences in select populations of protozoa, bacteria and archaea (**Paper #4 figures 2e and 3a**). Literature searches revealed that, for protozoa at least, the observed RCT-A and -B were already partially described over half a century ago, when J. M. Eadie (1962, Microbiology)[36] observed with a light microscope that certain groups of protozoa would never co-exist. Among these, community type A was enriched with Ophryoscolecids (now Ophryoscolecinae[15]) specifically *Polyplastron multivesiculatum* and other species under genera *Ophryoscolex and Diploplastron* (now Diplodiniinae). Conversely, community "type B" was defined as being enriched with species of genus *Eudiplodinium* and *Epidinium*, sometimes with the coexistence of ophryoscolecids. Later, Williams & Coleman (1992, Springer)[30] revisited these community types and defined type A with the presence of *Polyplastron multivesiculatum* and possibly *Diploplastron affine*, and type B with the presence of any *Epidinium* spp. In 2016, Kittelmann et al.[37] used the same scheme and made an updated definition of the types. Iterative amendments to the definition of these community types raise attention to the fact that the technology used for the quantification may impose differences in perceived abundances. Our results differ from the historic definitions by suggesting that the community types might not lead to complete exclusion of the less-abundant protozoa, even though significant differences in abundances are observed between A and B.

Considering the taxonomic profiles and the related functional implications of the community types, RCT-B is significally enriched for subfamilies *Diplodiniinae* and *Ophryoscolecinae* (Li 2022 ISME J) and is defined by species *Diplodinium dentatum, Epidinum cattanei*, and *Epidinium caudatum*. The strongest representative is *Epidinium*

cattanei (**Paper #4 figure 4b**) that has by far the highest abundance. Epidinia encode a vast array of glycoside hydrolase CAZymes[15,60] and are known to actively attach to plant material[30]. This fits well into the observation of the plant fiber multiarray polymer profiling (MAPP) that several hemicellulose fibers were less abundant in RCT-B. This is possibly a response to faster degradation (**Paper #4 fig. 5a**) by these CAZymes, suggesting that *Epidinium* spp. directs the overall glycan landscape. This may be related to the strong metatranscriptomic presence of *Prevotella* spp. and *Sodaliphilus* spp. (**Paper #4 fig. 3**) which are known fiber degraders[4]. In terms of fermentation end products, butyrate was significantly more abundant in RCT-B (**Paper #4 fig. 4c**) which may be linked with *Epidinium* spp. being observed to express several genes related to butyrate-metabolism (**Paper #4 fig. 4b)**.

In RCT-A the responsibilities for digestion are shared more broadly amongst protozoa (**Paper #4 fig. 4b**) with significant enrichment for protozoal populations affiliated to families *Entodiniinae* and *Isotrichidae* with genera *Entodinium bursa*, *Entodinium caudatum*, *Entodinium longinucleatum*, *Isotricha intestinalis*, *Isotricha YL-2021a*; and *Polyplastron multivesiculatum* (**Paper #4 fig. 3c**). MAGs taxonomically classified as *Methanobrevibacter* spp. were identified as strong contributors of proteomic and transcriptomic features that separate the RCTs (**Paper #4 fig. 2e**). These are known to hitchhike on protozoa[34], which could be linked to the conditional populations of protozoa, as phylogenetically distinct populations of *Methanobrevibacter* spp. are observed between the RCTs. Amongst bacteria, *Faecousia* spp., *Merdiplasma* spp., and *Acutalibacteraceae* are prevalent in RCT-A (**Paper #4 fig. 4a**). These clades encode parts of the Wood-Ljungdahl pathway which is a means of carbon assimilation by the use of hydrogen as electron donor. Amongst these, specifically RUG762 of *Acutalibacteraceae* is of interest as it represents the bacterial clade with the strongest PC loadings in metaproteomics. Further annotation of RUG762 predicted that it produces methionine via a cobalamin-dependent 5-methyltetrahydrofolate–homocysteine methyltransferase, rather than complete reductive acetogenesis (**Paper #4 fig. 5b**). In terms of hydrogen sinks, reductive acetogenesis is energetically unfavorable compared to methanogenesis, although it has been described to be competitive during higher partial pressure of H_2[13]. Unfortunately, this hypothesis cannot be tested within the context of this study as H_2 was not measured during sampling.

Seeing that the metabolic cascade of plant fiber degradation and production of VFAs and amino acids are affected by the RCTs (**Paper #4 fig. 4c**), it seems reasonable to

expect that the host animal is affected. However, as there are no significant correlations between the RCTs and any key performance trait, it does not seem like the RCTs impose any effects that strongly influence animal production (**Paper #4 fig. S2**). To investigate deeper if the host was in any way affected, proteomic network analysis was applied on the rumen digesta, rumen wall, and liver. Amongst the hundreds of protein clusters identified by this method in microbiome and host, only two were identified in the host, both with RCT-A (**Paper #4 fig. S5**). These protein clusters were identified in the rumen wall and one was significantly enriched for cysteine and methionine metabolism as defined by the KEGG pathway database. This suggests that the amino acid production profile of RCT-A may have an effect in the host, but must be tested more vigorously in future investigations.

Protozoa may affect the metabolic cascade of their RCT in several ways: Affecting the glycan landscape through plant fiber degradation; preying on other microorganisms; producing VFAs; and as a habitat for hitchhiking microorganisms. This means that they could potentially be the main drivers of the RCTs. Although the protozoa of class Litostomatea remain largely uncharacterized, this paper represents an advancement towards better characterizing the RCTs and the activities that may be driving their composition.

4. Concluding remarks & future perspectives

The methanogenic biology of the rumen does not exist in an isolated system, but rather plays a role in the global carbon cycle. Current practices of using ruminants, specifically cattle, as production animals for beef and dairy products present challenges for food security[4,120] and safety for humanity in the long term, as they are a major source of anthropogenic greenhouse gas emissions[4,121,122]. Similarly to the developments between rumen microorganisms and the cattle, human and cattle have a history of coevolution: The concept of cattle breeding is deeply rooted in some human cultural identities as evidenced by the genomically encoded persistence of the lactase gene, which has been shown to have entered the human lineage within the last ~2,000–20,000 years[123]. While in many countries reduction in consumption of meat and dairy is seen as a strategy to reduce anthropogenic emissions from ruminants, animal husbandry has deep cultural and socioeconomic ties in many regions of the world. This suggests that stopping the use of cattle as production animals is not a realistic means to mitigate methane emissions in a historic context. Rather, part of the solution to climate challenges may be to better understand how the production of methane in the rumen can be mitigated by directing the fermentative pathways to alternative energy-yielding metabolites of benefit to the animal.

To be able to succeed in modulating the metabolic cascade of the rumen it is necessary to build a better understanding of how the host and its microbiome work together to drive assimilation of nutrients and energy from cattle's plant fiber based diet. This thesis takes several steps ahead on this path by developing tools for scalable characterization of MAGs from complex environments, by formalizing methods that can be used for integrating multi-omic molecular features across the host-microbiome boundary, and lastly by using the acquired tools to dive further into the biology around the cattle rumen.

The first paper (**#1**) described how DNA can be extracted from complex samples and analyzed to reconstruct MAGs which can be characterized in terms of their metabolic potential. One important consideration that was not included is that when extracting such biomolecules (for example DNA, RNA, protein, and metabolites), it should be done with regard to the disposition of the cell wall of the microorganisms being investigated. For example, bacteria can have a thick layer of peptidoglycan, protozoa can be fragile, and fungi may grow into and embed themselves into the plant fibers. In order to obtain a representative amount of biomolecules from each of these taxonomic domains, ideally fractional or differential extractions should be performed to achieve efficient lysis of the cells without adverse effects, i.e., DNA fragmentation. This should be investigated deeper, and taken up as a common practice in microbiome analyses to

enable painting a more representative picture of the diverse microorganisms of the rumen, and similar environments.

One of the major outcomes from **Paper #1** was the accessibility of high throughput genome annotation that is a fundamental feature of any multi/holo-omics workflow. In this context, **Paper #2** settled the need for a generalized pipeline for quick but thorough analysis of MAGs. This is necessary for large scale characterization of genomes from complex environments and empowers the database used for mapping of proteomic and transcriptomic features in any multi-omic experiment. As this pipeline is designed with modularity in mind, it will be updated to reflect the technological evolution and incorporate new tools, to ensure its adoption by the larger scientific community within meta-omics.

As part of the multi- and holo-omic strategies that were explored throughout this thesis, several technical challenges arose that ideally require in-depth future investigations. One particular example was encountered with our use of metaproteomics. From a biological perspective, protein abundances can be seen as offering the most relevant insight into the presence of individual microorganisms and the metabolic functions they are performing in their natural habitat. However, current LC-MS/MS-based metaproteomics techniques have inherent challenges with missing values and in distinguishing similar proteins. Furthermore, severe technical problems were encountered when running the software responsible for performing the peptide-to-spectrum matches. This was largely due to a mismatch between the design of the proteomic software and the size of our proteomic database. This anecdote will not stand alone as the technological evolution of sequencing techniques leads to the recovery of even more MAGs representing species that are currently uncharacterized. This will enlarge the proteomic databases used in metaproteomics and further exacerbate the computational issues encountered in projects with a similar approach. This highlights the need for continued technological development towards metaproteomics, to avoid saturation and to enable insight into datasets with ever higher resolution.

Beyond individual omic layers, multi- and holo-omics necessitates methods that can handle multiple complex data types. **Paper #3** formalized the holobiont and its possible host-microbiome interactions, and characterized and juxtaposed powerful methods for analyzing data representing them. However, one conclusion from this review was that a methodological gold standard for holo-omic integration lies further in the future. Holo-omics is a young field, and it can be expected that an ideal standard for analysis will come together when the appropriate multi-omic technologies evolve and enable construction of higher resolution host-microbiome datasets for critical assessment of the holobiont concept. Ideally, such a methodological standard is fully

parameterized and able to handle diverse holo-omic experimental setups to enable a quick turnover of biological hypotheses.

To showcase the methodological approaches that were developed and explored in this thesis, an animal trial was designed and analyzed in depth. **Paper #4** exemplified integrating the molecular layers of host and microbiome to gain a deeper understanding of the cattle rumen holobiont. This led to the detection of two rumen community types, labeled RCT-A and RCT-B. The observations of RCT-A and -B were unexpected, and led us on a journey to characterize the organismal factors and the metabolic cascade that defines their existence. A future direction will be to identify the origin and determining factors that enabled the RCTs to establish and persist within individual cattle hosts. As the RCTs are seemingly not linked to the breed of the cattle, it can be deduced that the host genotype likely plays no role in their definition. That the RCTs have little to no effect on the host may speak to the plasticity of the rumen system overall. Maybe, because the microorganisms on the two sides of RCT-A and -B are functionally redundant and able to fulfill the same functions at large. The temporal analysis indicated that the RCTs were stable over time. A possible explanation of this could be that the cattle enlisted, in this trial, were seeded with their RCT prior to commencement, which could possibly be directed by mother-offspring contact. This hypothesis could be tested by following offspring from birth until maturity while correlating the RCT between mother and offspring. An alternative hypothesis explaining the RCT-stability over time is that certain predation patterns of the protozoa drive their apparent exclusivity. In the case that predatory protozoa target specific prey, they may be able to reinforce a distinct community profile. This hypothesis could be tested by isolating and inoculating the predatory marker species into cattle with alternate RCTs. The importance of further characterizing the RCTs can be illustrated with reference to the large-scale industrial development and investment into feed additives for cattle production, e.g. bromoform and 3-nitrooxypropanol. In the case that there is any interaction effect between a feed additive and an RCT, further characterization is key to obtaining an efficient response.

This thesis described the use of holo-omics, culminating in a study where the cattle rumen was used as an example system. The key finding, namely the RCTs, poses several distinct challenges for further analysis. First and foremost is a continued push to further characterize the RCTs beyond the limited boundaries of the trial analyzed herein—to understand their origin and to comprehensively test hypotheses for the metabolic cascades they impose. The second challenge is to look deeper into uncharacterized host-microbiome interactions, taking into account and adjusting for background effects imposed by the RCTs and other within-microbiome variation. In the greater context of holo-omics methodology development, this work brings forward

new queries arising from the within-microbiome variation that was encountered in the ruminant work. Namely, does it stand in the way of elucidating unrelated host-microbiome effects? For example, by imposing artifacts on host-microbiome signals—it should finally be considered whether the cattle rumen with its enigmatic microbiome is an appropriate model for holo-omic methodological development. One final conclusion of this PhD is that current bottlenecks for gaining better insights into holobionts is the lack of abstract annotations of pathways, as well as the bioinformatic pipelines to quickly test biological hypotheses from complex datasets.

References

1. Sagan, L. On the Origin of Mitosing Cells.
2. Vosseberg, J. *et al.* The emerging view on the origin and early evolution of eukaryotic cells. *Nature* **633**, 295–305 (2024).
3. Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. & Gordon, J. I. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.* **6**, 776–788 (2008).
4. Mizrahi, I., Wallace, R. J. & Moraïs, S. The rumen microbiome: balancing food security and environmental impacts. *Nat. Rev. Microbiol.* **19**, 553–566 (2021).
5. Kunath, B. J. *et al.* From proteins to polysaccharides: lifestyle and genetic evolution of *Coprothermobacter proteolyticus*. *ISME J.* **13**, 603–617 (2019).
6. Seshadri, R. *et al.* Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36**, 359–367 (2018).
7. Kanehisa, M. The KEGG Database. in *'In Silico' Simulation of Biological Processes* 91–103 (John Wiley & Sons, Ltd, 2002). doi:10.1002/0470857897.ch8.
8. Solden, L. M. *et al.* Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nat. Microbiol.* **3**, 1274–1284 (2018).
9. Johnson, K. A. & Johnson, D. E. Methane emissions from cattle. *J. Anim. Sci.* **73**, 2483–2492 (1995).
10. Sivalingam, V., Haugen, T., Wentzel, A. & Dinamarca, C. Effect of Elevated Hydrogen Partial Pressure on Mixed Culture Homoacetogenesis. *Chem. Eng. Sci. X* **12**, 100118 (2021).
11. Beauchemin, K. A., Ungerfeld, E. M., Eckard, R. J. & Wang, M. Review: Fifty years of research on rumen methanogenesis: lessons learned and future challenges for mitigation. *Animal* **14**, s2–s16 (2020).
12. Ragsdale, S. W. Enzymology of the Wood–Ljungdahl Pathway of Acetogenesis. *Ann. N. Y. Acad. Sci.* **1125**, 129–136 (2008).
13. Melgar, A. *et al.* Effects of 3-nitrooxypropanol on rumen fermentation, lactational performance, and resumption of ovarian cyclicity in dairy cows. *J. Dairy Sci.* **103**, 410–432 (2020).
14. Andersen, T. O. *et al.* Metabolic influence of core ciliates within the rumen microbiome. *ISME J.* **17**, 1128–1140 (2023).

15. Li, Z. *et al.* Genomic insights into the phylogeny and biomass-degrading enzymes of rumen ciliates. *ISME J.* **16**, 2775–2787 (2022).

16. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).

17. Solomon, K. V. *et al.* Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes. *Science* **351**, 1192–1195 (2016).

18. Lagier, J.-C. *et al.* Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* **16**, 540–550 (2018).

19. Staley, J. T. & Konopka, A. MEASUREMENT OF IN SITU ACTIVITIES OF NONPHOTOSYNTHETIC MICROORGANISMS IN AQUATIC AND TERRESTRIAL HABITATS. *Annu. Rev. Microbiol.* **39**, 321–346 (1985).

20. Wang, B. L. *et al.* Microfluidic high-throughput culturing of single cells for selection based on extracellular metabolite production or consumption. *Nat. Biotechnol.* **32**, 473–478 (2014).

21. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

22. Jami, E. & Mizrahi, I. Composition and Similarity of Bovine Rumen Microbiota across Individual Animals. *PLOS ONE* **7**, e33306 (2012).

23. Hart, E. H., Creevey, C. J., Hitch, T. & Kingston-Smith, A. H. Meta-proteomics of rumen microbiota indicates niche compartmentalisation and functional dominance in a limited number of metabolic pathways between abundant bacteria. *Sci. Rep.* **8**, 10504 (2018).

24. Jami, E. & Mizrahi, I. Composition and Similarity of Bovine Rumen Microbiota across Individual Animals. *PLOS ONE* **7**, e33306 (2012).

25. Moraïs, S. & Mizrahi, I. Islands in the stream: from individual to communal fiber degradation in the rumen ecosystem. *FEMS Microbiol. Rev.* **43**, 362–379 (2019).

26. Wang, D. *et al.* Multi-omics revealed the long-term effect of ruminal keystone bacteria and the microbial metabolome on lactation performance in adult dairy goats. *Microbiome* **11**, 215 (2023).

27. Stanton, T. B. & Canale-Parola, E. Treponema bryantii sp. nov., a rumen spirochete that interacts with cellulolytic bacteria. *Arch. Microbiol.* **127**, 145–156 (1980).

28. Blackburn, T. H. & Hungate, R. E. Succinic Acid Turnover and Propionate Production in the Bovine Rumen. *Appl. Microbiol.* **11**, 132–135 (1963).

29. Scheifinger, C. C. & Wolin, M. J. Propionate Formation from Cellulose and Soluble Sugars by Combined Cultures of Bacteroides succinogenes and Selenomonas ruminantium. *Appl. Microbiol.* **26**, 789–795 (1973).

30. Williams, A. G. & Coleman, G. S. *The Rumen Protozoa*. (Springer New York, New York, NY, 1992). doi:10.1007/978-1-4612-2776-2.

31. Park, T. & Yu, Z. Do Ruminal Ciliates Select Their Preys and Prokaryotic Symbionts? *Front. Microbiol.* **9**, 1710 (2018).

32. Terrapon, N., Lombard, V., Drula, E., Coutinho, P. M. & Henrissat, B. The CAZy Database/the Carbohydrate-Active Enzyme (CAZy) Database: Principles and Usage Guidelines. in *A Practical Guide to Using Glycomics Databases* (ed. Aoki-Kinoshita, K. F.) 117–131 (Springer Japan, Tokyo, 2017). doi:10.1007/978-4-431-56454-6_6.

33. Solomon, R. *et al.* Protozoa populations are ecosystem engineers that shape prokaryotic community structure and function of the rumen microbial ecosystem. *ISME J.* **16**, 1187–1197 (2022).

34. Newbold, C. J., de la Fuente, G., Belanche, A., Ramos-Morales, E. & McEwan, N. R. The Role of Ciliate Protozoa in the Rumen. *Front. Microbiol.* **6**, 1313 (2015).

35. Wallace, R. J. & McPherson, C. A. Factors affecting the rate of breakdown of bacterial protein in rumen fluid. *Br. J. Nutr.* **58**, 313–323 (1987).

36. Eadie, J. M. Inter-Relationships between Certain Rumen Ciliate Protozoa. *Microbiology* **29**, 579–588 (1962).

37. Kittelmann, S. *et al.* Natural variation in methane emission of sheep fed on a lucerne pellet diet is unrelated to rumen ciliate community type. *Microbiology* **162**, 459–465 (2016).

38. Vogels, G. D., Hoppe, W. F. & Stumm, C. K. Association of methanogenic bacteria with rumen ciliates. *Appl. Environ. Microbiol.* **40**, 608–612 (1980).

39. Paul, R. G., Williams, A. G. & Butler, R. D. Hydrogenosomes in the rumen entodiniomorphid ciliate Polyplastron multivesiculatum. *Microbiology* **136**, 1981–1989 (1990).

40. Müller, M. Review Article: The hydrogenosome. *Microbiology* **139**, 2879–2889 (1993).

41. Konings, W. N., Albers, S.-V., Koning, S. & Driessen, A. J. M. The cell membrane plays a crucial role in survival of bacteria and archaea in extreme environments. *Antonie Van Leeuwenhoek* **81**, 61–72 (2002).

42. Janssen, P. H. & Kirs, M. Structure of the Archaeal Community of the Rumen. *Appl. Environ. Microbiol.* **74**, 3619–3625 (2008).

43. Vanwonterghem, I. *et al.* Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nat. Microbiol.* **1**, 1–9 (2016).

44. Lang, K. *et al.* New Mode of Energy Metabolism in the Seventh Order of Methanogens as Revealed by Comparative Genome Analysis of "Candidatus Methanoplasma termitum". *Appl. Environ. Microbiol.* **81**, 1338–1352 (2015).

45. Henderson, G. *et al.* Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci. Rep.* **5**, 14567 (2015).

46. Kittelmann, S., Naylor, G. E., Koolaard, J. P. & Janssen, P. H. A Proposed Taxonomy of Anaerobic Fungi (Class Neocallimastigomycetes) Suitable for Large-Scale Sequence-Based Community Structure Analysis. *PLOS ONE* **7**, e36866 (2012).

47. Fliegerova, K., Kaerger, K., Kirk, P. & Voigt, K. Rumen Fungi. in *Rumen Microbiology: From Evolution to Revolution* (eds. Puniya, A. K., Singh, R. & Kamra, D. N.) 97–112 (Springer India, New Delhi, 2015). doi:10.1007/978-81-322-2401-3_7.

48. Ahrendt, S. R., Mondo, S. J., Haridas, S. & Grigoriev, I. V. MycoCosm, the JGI's Fungal Genome Portal for Comparative Genomic and Multiomics Data Analyses. in *Microbial Environmental Genomics (MEG)* (eds. Martin, F. & Uroz, S.) 271–291 (Springer US, New York, NY, 2023). doi:10.1007/978-1-0716-2871-3_14.

49. Gaborieau, B. *et al.* Prediction of strain level phage–host interactions across the Escherichia genus using only genomic information. *Nat. Microbiol.* **9**, 2847–2861 (2024).

50. Simmonds, P. *et al.* Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **15**, 161–168 (2017).

51. Castledine, M. & Buckling, A. Critically evaluating the relative importance of phage in shaping microbial community composition. *Trends Microbiol.* **32**, 957–969 (2024).

52. Scheller, H. V. & Ulvskov, P. Hemicelluloses. *Annu. Rev. Plant Biol.* **61**,

263–289 (2010).

53. Chen, J. *et al.* Alpha- and beta-mannan utilization by marine Bacteroidetes. *Environ. Microbiol.* **20**, 4127–4140 (2018).

54. Kaur, R., Sharma, M., Ji, D., Xu, M. & Agyei, D. Structural Features, Modification, and Functionalities of Beta-Glucan. *Fibers* **8**, 1 (2020).

55. FRY, S. C. The Structure and Functions of Xyloglucan. *J. Exp. Bot.* **40**, 1–11 (1989).

56. Mohnen, D. Pectin structure and biosynthesis. *Curr. Opin. Plant Biol.* **11**, 266–277 (2008).

57. Murphy, C. L. *et al.* Horizontal Gene Transfer as an Indispensable Driver for Evolution of Neocallimastigomycota into a Distinct Gut-Dwelling Fungal Lineage. *Appl. Environ. Microbiol.* **85**, e00988-19 (2019).

58. Terrapon, N., Lombard, V., Drula, E., Coutinho, P. M. & Henrissat, B. The CAZy Database/the Carbohydrate-Active Enzyme (CAZy) Database: Principles and Usage Guidelines. in (ed. Aoki-Kinoshita, K. F.) 117–131 (Springer Japan, Tokyo, 2017). doi:10.1007/978-4-431-56454-6_6.

59. Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M. & Henrissat, B. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol. Biofuels* **6**, 41 (2013).

60. Gavande, P. V., Goyal, A. & Fontes, C. M. G. A. Chapter 1 - Carbohydrates and Carbohydrate-Active enZymes (CAZyme): An overview. in *Glycoside Hydrolases* (eds. Goyal, A. & Sharma, K.) 1–23 (Academic Press, 2023). doi:10.1016/B978-0-323-91805-3.00012-5.

61. Wardman, J. F., Bains, R. K., Rahfeld, P. & Withers, S. G. Carbohydrate-active enzymes (CAZymes) in the gut microbiome. *Nat. Rev. Microbiol.* **20**, 542–556 (2022).

62. McKee, L. S. *et al.* Polysaccharide degradation by the Bacteroidetes: mechanisms and nomenclature. *Environ. Microbiol. Rep.* **13**, 559–581 (2021).

63. Bayer, E. A., Morag, E. & Lamed, R. The cellulosome — A treasure-trove for biotechnology. *Trends Biotechnol.* **12**, 379–386 (1994).

64. Van Soest, P. J. *Nutritional Ecology of the Ruminant*. (Comstock, Ithaca London, 1994).

65. Stevens, C. E. & Sellers, A. F. Pressure events in bovine esophagus and reticulorumen associated with eructation, deglutition and regurgitation. *Am. J.*

Physiol.-Leg. Content **199**, 598–602 (1960).

66. Cecava, M. J. 1 - Rumen Physiology and Energy Requirements. in *Beef Cattle Feeding and Nutrition (Second Edition)* (eds. Perry, T. W. & Cecava, M. J.) 3–24 (Academic Press, San Diego, 1995). doi:10.1016/B978-012552052-2/50004-4.

67. Dieho, K. *et al.* Morphological adaptation of rumen papillae during the dry period and early lactation as affected by rate of increase of concentrate allowance. *J. Dairy Sci.* **99**, 2339–2352 (2016).

68. Storm, A. C., Kristensen, N. B. & Hanigan, M. D. A model of ruminal volatile fatty acid absorption kinetics and rumen epithelial blood flow in lactating Holstein cows. *J. Dairy Sci.* **95**, 2919–2934 (2012).

69. Mizrahi, I. Rumen symbioses. in *The prokaryotes: Prokaryotic biology and symbiotic associations* 533–544 (Springer-Verlag Berlin Heidelberg, 2013).

70. *Quantitative Aspects of Ruminant Digestion and Metabolism*. (CABI Pub, Wallingford, Oxfordshire, UK ; Cambridge, MA, 2005).

71. Rocha, E. P. C. The Organization of the Bacterial Genome. *Annu. Rev. Genet.* **42**, 211–233 (2008).

72. Freire, B., Ladra, S. & Paramá, J. R. Memory-Efficient Assembly Using Flye. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 3564–3577 (2022).

73. Gupta, R. S. Protein Phylogenies and Signature Sequences: A Reappraisal of Evolutionary Relationships among Archaebacteria, Eubacteria, and Eukaryotes. *Microbiol. Mol. Biol. Rev.* **62**, 1435 (1998).

74. Esteban, G. F., Finlay, B. J. & Warren, A. Chapter 7 - Free-Living Protozoa. in *Thorp and Covich's Freshwater Invertebrates (Fourth Edition)* (eds. Thorp, J. H. & Rogers, D. C.) 113–132 (Academic Press, Boston, 2015). doi:10.1016/B978-0-12-385026-3.00007-3.

75. NovaSeq 6000 System | Key specifications and performance parameters. https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html.

76. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* **3**, lqab019 (2021).

77. Sereika, M. *et al.* Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat. Methods* **19**, 823–826 (2022).

78. Data analysis (DATD_5000_v1_revU_22Aug2016). *Oxford Nanopore*

Technologies https://nanoporetech.com/document/data-analysis (2017).

79. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

80. Beaulaurier, J. *et al.* Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.* **36**, 61–69 (2018).

81. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

82. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

83. Van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01773-0.

84. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).

85. Edfors, F. *et al.* Gene‐specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* **12**, 883 (2016).

86. O'Neil, D., Glowatz, H. & Schlumpberger, M. Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity. *Curr. Protoc. Mol. Biol.* **103**, 4.19.1-4.19.8 (2013).

87. Shakya, M., Lo, C.-C. & Chain, P. S. G. Advances and Challenges in Metatranscriptomic Analysis. *Front. Genet.* **10**, (2019).

88. Wilkins, M. R. *et al.* From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis. *Bio/Technology* **14**, 61–65 (1996).

89. Brito, G. C. & Andrews, D. W. Removing bias against membrane proteins in interaction networks. *BMC Syst. Biol.* **5**, 169 (2011).

90. Hettich, R. L., Pan, C., Chourey, K. & Giannone, R. J. Metaproteomics: Harnessing the Power of High Performance Mass Spectrometry to Identify the Suite of Proteins That Control Metabolic Activities in Microbial Communities. *Anal. Chem.* **85**, 4203–4214 (2013).

91. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).

92. Mayer, M. missRanger: Fast Imputation of Missing Values. (2024).

93. Woźnica, K. & Biecek, P. Does imputation matter? Benchmark for predictive models. Preprint at http://arxiv.org/abs/2007.02837 (2020).

94. waters. UPLC/MS Monitoring of Water-Soluble Vitamin Bs in Cell Culture Media in Minutes. *www.waters.com* http://www.waters.com/waters/library.htm?lid=134636355.

95. Bakshani, C. R., Sangta, J., Sommano, S. & Willats, W. G. T. Microarray Polymer Profiling (MAPP) for High-Throughput Glycan Analysis. *J. Vis. Exp.* 65443 (2023) doi:10.3791/65443.

96. Roehe, R. *et al.* Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance. *PLOS Genet.* **12**, e1005846 (2016).

97. Anowar, F., Sadaoui, S. & Selim, B. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Comput. Sci. Rev.* **40**, 100378 (2021).

98. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, (2008).

99. Cai, T. T. & Ma, R. Theoretical foundations of t-SNE for visualizing high-dimensional clustered data. *J. Mach. Learn. Res.* **23**, 301:13581-301:13634 (2022).

100. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).

101. Zhang, B. & Horvath, S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, (2005).

102. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

103. Barabási, A.-L. Scale-Free Networks: A Decade and Beyond. *Science* **325**, 412–413 (2009).

104. Strand, M. A., Jin, Y., Sandve, S. R., Pope, P. B. & Hvidsten, T. R. Transkingdom network analysis provides insight into host-microbiome interactions in Atlantic salmon. *Comput. Struct. Biotechnol. J.* **19**, 1028–1034 (2021).

105. Pei, G., Chen, L. & Zhang, W. Chapter Nine - WGCNA Application to

Proteomic and Metabolomic Data Analysis. in *Methods in Enzymology* (ed. Shukla, A. K.) vol. 585 135–158 (Academic Press, 2017).

106. Øyås, O. & Stelling, J. Genome-scale metabolic networks in time and space. *Curr. Opin. Syst. Biol.* **8**, 51–58 (2018).

107. Terzer, M., Maynard, N. D., Covert, M. W. & Stelling, J. Genome-scale metabolic networks. *WIREs Syst. Biol. Med.* **1**, 285–297 (2009).

108. Gunawardena, J. Time-scale separation – Michaelis and Menten's old idea, still bearing fruit. *FEBS J.* **281**, 473–488 (2014).

109. Varma, A. & Palsson, B. O. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Bio/Technology* **12**, 994–998 (1994).

110. Roughgarden, J., Gilbert, S. F., Rosenberg, E., Zilber-Rosenberg, I. & Lloyd, E. A. Holobionts as Units of Selection and a Model of Their Population Dynamics and Evolution. *Biol. Theory* **13**, 44–65 (2018).

111. Weger, L. A. D. *et al.* Flagella of a plant-growth-stimulating Pseudomonas fluorescens strain are required for colonization of potato roots. *J. Bacteriol.* **169**, 2769 (1987).

112. Margulis, L. & Fester, R. *Symbiosis as a Source of Evolutionary Innovation: Speciation and Morphogenesis*. (MIT Press, 1991).

113. Huws, S. A. *et al.* Addressing Global Ruminant Agricultural Challenges Through Understanding the Rumen Microbiome: Past, Present, and Future. *Front. Microbiol.* **9**, (2018).

114. Singh, R. K. *et al.* Influence of diet on the gut microbiome and implications for human health. *J. Transl. Med.* **15**, 73 (2017).

115. Keohane, D. M. *et al.* Microbiome and health implications for ethnic minorities after enforced lifestyle changes. *Nat. Med.* **26**, 1089–1095 (2020).

116. Spor, A., Koren, O. & Ley, R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.* **9**, 279–290 (2011).

117. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

118. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).

119. Yu, F., Deng, Y. & Nesvizhskii, A. I. MSFragger-DDA+ Enhances Peptide Identification Sensitivity with Full Isolation Window Search. 2024.10.12.618041 Preprint at https://doi.org/10.1101/2024.10.12.618041 (2024).

120. Barona, E., Ramankutty, N., Hyman, G. & Coomes, O. T. The role of pasture and soybean in deforestation of the Brazilian Amazon. *Environ. Res. Lett.* **5**, 024002 (2010).

121. Vergé, X. P. C., Dyer, J. A., Desjardins, R. L. & Worth, D. Greenhouse gas emissions from the Canadian dairy industry in 2001. *Agric. Syst.* **94**, 683–693 (2007).

122. Owen, J. J. & Silver, W. L. Greenhouse gas emissions from dairy manure management: a review of field-based studies. *Glob. Change Biol.* **21**, 550–565 (2015).

123. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31–40 (2007).

Paper #1

2023, Book chapter within Methods in Molecular Biology, Springer Nature

doi: 10.1007/978-1-0716-3151-5_19

Chapter 19

Long-Read Metagenomics and CAZyme Discovery

Alessandra Ferrillo, Carl Mathias Kobel, Arturo Vera-Ponce de León, Sabina Leanti La Rosa, Benoit Josef Kunath, Phillip Byron Pope, and Live Heldal Hagen

Abstract

Microorganisms play a primary role in regulating biogeochemical cycles and are a valuable source of enzymes that have biotechnological applications, such as carbohydrate-active enzymes (CAZymes). However, the inability to culture the majority of microorganisms that exist in natural ecosystems restricts access to potentially novel bacteria and beneficial CAZymes. While commonplace molecular-based culture-independent methods such as metagenomics enable researchers to study microbial communities directly from environmental samples, recent progress in long-read sequencing technologies are advancing the field. We outline key methodological stages that are required as well as describe specific protocols that are currently used for long-read metagenomic projects dedicated to CAZyme discovery.

Key words Long-read metagenomics, Carbohydrate-active enzymes, Microbial communities, Assembly, Binning

1 Introduction

The continuing initiative to find novel carbohydrate-active enzymes (CAZymes) derives from societal and industrial interest in utilizing plant biomass as a substrate for "bio-products" such as fuels, chemicals, and plastics. Cellulose, the most abundant form of carbon on earth, is notoriously difficult to deconstruct using currently available enzyme technology, whereas a variety of digestive ecosystems, such as gastrointestinal tract of herbivores [1] or termites guts [2, 3], are able to efficiently utilize lignocellulolytic biomass. This functional capacity is controlled by microorganisms that are difficult to isolate and cultivate, which restricts direct access to their genetic and enzymatic machinery. Cultivability "bottlenecks" can be addressed by applying culture-independent methods, such as metagenomics, whereby total DNA is directly extracted,

D. Wade Abbott and Wesley F. Zandberg (eds.), *Carbohydrate-Protein Interactions: Methods and Protocols*, Methods in Molecular Biology, vol. 2657, https://doi.org/10.1007/978-1-0716-3151-5_19,

"shotgun-sequenced," and analyzed from the microbial sample without any need for prior isolation.

Shotgun metagenomics can theoretically generate sequences (reads) from all of the genomes present in the sample, collectively referred to as the metagenome. Recent progress in bioinformatics permits the complete reconstruction of a significant fraction of constituent genomes within a sample, otherwise known as metagenome-assembled genomes (MAGs). These methods have already shown their potential to find new non-cultivable polysaccharide-degrading bacteria and fungi [4–6], as well as interpretations of synergistic relationships between uncultured phylotypes in a cellulolytic community [7, 8]. Excitingly, long-read DNA sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore are creating new opportunities to improve the quality of MAGs, including their reconstruction into a single circular contig [9–11]. Despite the inherent benefits, the use of shotgun metagenomics also comes with several defined challenges. The diversity of microbial species in digestive ecosystems combined with the immense data output of the high-throughput sequencing (HTS) platforms produces terabytes of data that needs to be annotated before it can be meaningfully interpreted. Today, a wide variety of different HTS platforms and bioinformatic packages tailor-made for metagenome projects are available to scientists. Therefore, careful considerations of available resources and technical challenges that are relevant for a particular sample are still required. It is the aim of this chapter to describe the application of the biological and bioinformatic methods that are currently available for long-read metagenomics in this rapidly developing field and to provide exemplar step-by-step protocols to follow for samples originating from plant biomass-degrading ecosystems (Fig. 1).

2 Materials

2.1 DNA Extraction, Alternative A: Commercial Kits

1. DNeasy PowerSoil Pro kit (QIAGEN, Germany).
2. Short Read Eliminator (SRE) Kit (PacBio, USA).
3. *Optional* (*see* Subheading "DNA extraction Alternative A – rapid DNA extraction with short-read elimination"; "Short-Read Elimination"): ProNex® Size-Selective Purification System (Promega, USA).

2.2 DNA Extraction, Alternative B: Buffers

1. *Dissociation buffer*: 0.1% Tween 80, 1% methanol, and 1% tertiary butanol (v/v) in Milli-Q water. Adjust to pH 2 by adding HCl.
2. *Cell wash buffer*: 10 mM Tris–HC1 (pH 8.0) and 1 M NaCl, sterilized by autoclaving.

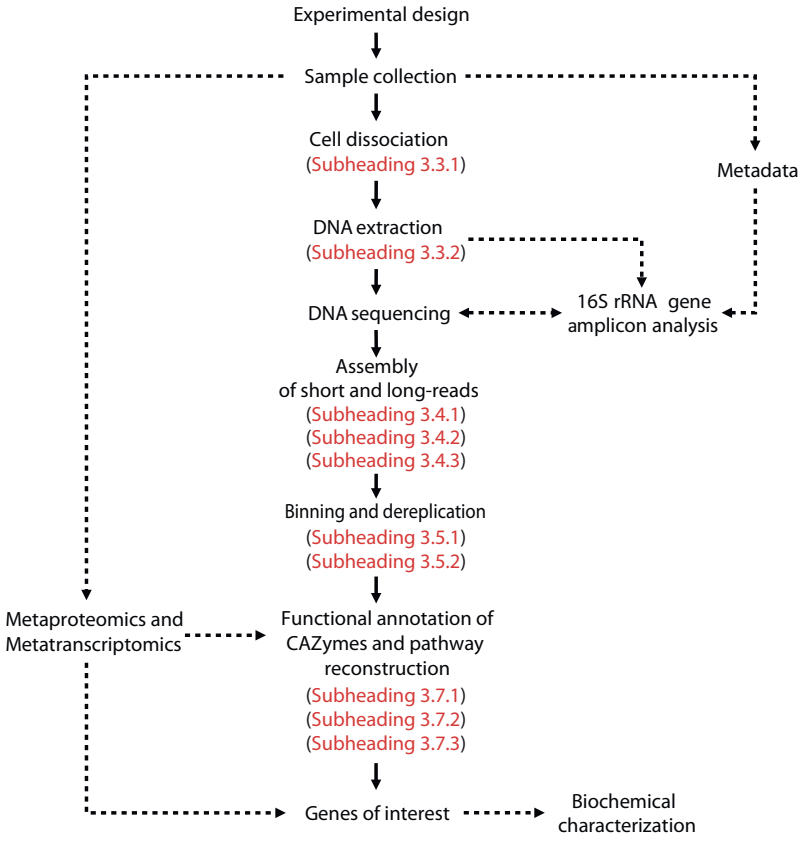


Fig. 1 Flow diagram of a typical metagenomic project dedicated to CAZyme discovery. Workflow represented by solid lines is discussed with specific protocols included (red text). Dotted lines indicate additional complementary techniques that are not presented in detail within this chapter

3. *RBB + C buffer*: 500 mM NaCl, 50 mM Tris–HCl, 50 mM EDTA, and 4% SDS.
4. *NaCl/CTAB buffer*: 4.1 g NaCl and 10 g cetyltrimethylammonium bromide (CTAB), dissolved in 80 mL of Milli-Q water. Heat to 68 °C to enhance the dissolving process. Add Milli-Q water to a final volume of 100 mL.
5. *TE buffer*: 10 mM Tris–HCl (pH 7.6) and 1 mM EDTA (pH 8.0).
6. *5M NaCl:* 29.2 g NaCl, dissolved in 100 mL Milli-Q water. Sterilize by autoclaving.
7. *1M Tris–HCl*: 121.1 g Tris-base, dissolved in 800 mL Milli-Q water. Adjust to pH 7.6 with HCl. Add Milli-Q water until to a final volume of 1.0 L.

8. *4% SDS*: 4 g SDS in 96 mL Milli-Q water. Heat to 68 °C to enhance the dissolving process. Adjust to pH 7.4. Add Milli-Q water to final volume of 100 mL.
9. *EDTA solution*: 186.12 g EDTA·Na_2·$2H_2O$ (molecular weight 372.24), dissolved in 800 mL Milli-Q water. While stirring vigorously on a magnetic stirrer, add NaOH pellet or 10 N NaOH to adjust the solution pH 8.0. Adjust the volume to 1 L with deionized/Milli-Q water. Sterilize by autoclaving.

2.3 Computational Software

1. NanoPlot v1.20.0 (https://github.com/wdecoster/NanoPlot).
2. Fastp v0.23.1 (https://github.com/OpenGene/fastp).
3. Filtlong v0.2.1 (https://github.com/rrwick/Filtlong).
4. MEGAHIT v1.2.9 (https://github.com/voutcn/megahit).
5. Flye v2.9 (https://github.com/fenderglass/Flye).
6. Minimap2 v2.17 (https://github.com/lh3/minimap2).
7. BWA-MEM2 (https://github.com/bwa-mem2/bwa-mem2).
8. Racon v1.4.20 (https://github.com/isovic/racon).
9. Medaka v1.6.1 (https://github.com/nanoporetech/medaka).
10. Polypolish v0.5.0 (https://github.com/rrwick/Polypolish).
11. MetaBAT2 v2.12 (https://bitbucket.org/berkeleylab/metabat).
12. dRep v3.2.2 (https://github.com/MrOlm/drep).
13. CheckM2 v0.1.3 (https://github.com/chklovski/CheckM2).
14. GTDB-Tk v2.1.0 (https://github.com/Ecogenomics/GTDBTk).
15. DRAM v1.2.4 (https://github.com/WrightonLabCSU/DRAM).
16. dbCAN (https://bcb.unl.edu/dbCAN2/index.php).
17. PhyloPhlAn v3.0.2 (https://github.com/biobakery/phylophlan).
18. R v4.2.1 (https://www.r-project.org/).

3 Methods

3.1 Strategy Development

For metagenomic projects dedicated to CAZyme discovery, the goal is relatively straightforward: obtain large contiguous DNA fragments that are coupled with the fewest possible misassemblies so that as many complete genes and operons are available for screening. Therefore, understanding which choice of sequencing platform and the amount of sequencing (sequencing depth) required to achieve a superior quality dataset is crucial when

designing the project [12]. Different sequencing technologies have vastly different read lengths and base pair (bp) yield and require different amounts of sample DNA and downstream bioinformatic analyses (*see* review [13]). The suitability of a specific type of HTS data to a given sample largely depends on the community structural complexity. Typically, shotgun sequencing results in deep coverage of dominant species and less reads from lower abundant species. To accommodate this discrepancy and gain greater access to the rarer community members, the sequencing effort must be increased; however, computational limitations (e.g., RAM, CPUs, storage) need to be considered. Therefore, an initial evaluation of the community structure using 16S rRNA gene amplicon analysis [14, 15] is recommended to help determine the complexity as well as the choice of HTS platform and the depth of sequencing to obtain the dataset necessary for CAZyme gene searches.

The taxonomic composition determined by 16S rRNA gene amplicon analysis can further assist in choosing metagenomes to assemble together (co-assembly), especially in circumstances where the sequencing depths need to be improved while keeping the computational demands attainable [16]. Eukaryotic DNA, such as from host, plant, or eukaryotic microorganisms (fungi, protozoa), can also affect the efficiency of sequencing and downstream analysis of the prokaryotic population. For example, ciliated protozoa can represent a large portion of the microbial biomass in the rumen ecosystem, and recent analysis in our lab indicates that the presence of their DNA potentially has a negative influence on the overall microbiome assembly. Thus, while ciliated protozoa are important contributors to the CAZy repertoire in the rumen [17], removal of eukaryotic "contamination" might enhance the recovery of prokaryote high-quality (HQ) MAGs.

3.2 Sample Collection and Metadata

The collection and processing of environmental samples is the first stage when planning a metagenomic project. Key considerations include the number of samples needed to adequately view the temporal dynamics or variability among the ecosystem as well as fulfill operating requirements of downstream bioinformatic software that require multiple samples (*see* Subheading 3.5). Moreover, other "omic" techniques, such as metaproteomics and metatranscriptomics, can be used to efficiently complement the metagenomic analysis and its outcomes (*see* Fig. 1 and Subheading 3.8). Extra subsamples can be easily stored in standardized ways in order to be analyzed later if required. In addition to the number of samples, the accompaniment of "metadata" can greatly enhance the ability to interpret the sequence data and particularly for comparative, spatial, or temporal series analysis. Metadata should appropriately describe the samples and the methods used. A suite of standard languages, called the minimum information about any (×) sequence checklists (MIxS) [18], provides format for recording

environmental and experimental data. These standards include MIGS (minimum information about a genome sequence), MIMS (minimum information about a metagenome sequence), and MIMAG (minimum information about a metagenome-assembled genome) checklists [19, 20]. What is recorded depends on where the samples come from but usually includes, among others, temperature, pH, substrate, sample handling, DNA extraction method, sequencing technology, and the bioinformatic methods used.

3.3 Cell Dissociation and DNA Extraction Methods

When looking for CAZymes, samples vary greatly from different environments, such as guts, soils, excrements, sediments, bioreactors, and other plant-associated biomass. These habitats present different characteristics, such as the presence of host cells [2, 21], microbial eukaryotes [22], enzymatic inhibitors (such as humic acids) [23], or biofilms [24], and therefore require specific protocols. Samples of plant biomass-degrading communities often require an additional processing step to remove microbial cells that are attached to plant fibers [24]. In Subheading 3.3.1, we describe a high-yield method that enables the dissociation of the microbial cells from the substrate and the recovery of the full sample's diversity. The resulting cell biomass generated from this dissociation protocol is suitable for both kit-based DNA extractions [25, 26] as well as Mamur's derived protocols [24].

DNA extraction is an important part of the experimental design since it can have a major impact on the subsequent NGS platform and downstream bioinformatic analyses that are used. DNA extraction and purification is still considered a bottleneck for metagenomic analyses, compounded by the fact that there is not one common method that fits every environmental sample. Indeed, the quality and the quantity of DNA required vary from a sequencing technology to another and may influence the choice of the DNA extraction method.

Long-read technology platforms (Oxford Nanopore Technology (ONT), PacBio) require high concentrations of non-fragmented high-molecular-weight (HMW) DNA and the efficiency of the sequencing is heavily influenced by the DNA integrity. While accommodating high-throughput sample handling, commercial kits usually result in fragmentation of the DNA molecules during the mechanical lysis of the cells (*see* Subheading 3.3.2). These instances will require gentle non-invasive methods, or to use additional size selection kits to remove short DNA fragments. In instances where a low cell biomass prohibits high nanogram or microgram quantities of DNA, whole genome amplification of starting material can also be necessary. As with any amplification method, sequence biases can occur [27, 28] and their impacts depend on the amount of starting material and the required number of amplification rounds to produce sufficient amount of DNA for sequencing.

3.3.1 Cell Dissociation for Plant-Associated Biomass Samples

The sample preparation is initially dependent on whether the sample has been stored as biomass sample at −80 °C or in 1/5 volume of phenol/ethanol (5%/95%) pH 8.0 at 4 °C (*see* **Note 1**).

1. Transfer 1.0 g of biomass or 1.5 mL sample and phenol/ethanol mix with a wide-bore pipette to a 2 mL tube.
2. Centrifuge at 14,000× *g* at room temperature for 2 min.
3. Discard supernatant and resuspend biomass in 500 µL of dissociation buffer by vortexing for 30 s.
4. Centrifuge at 100× *g* for 20 s at room temperature and transfer cell-containing supernatant to a new cell-collection tube and centrifuge at 14,000× *g* for 5 min at room temperature. Discard cell-free supernatant.
5. Repeat dissociation buffer **steps 3** and **4** two to three more times, transferring each cell-containing supernatant in the same cell-collection tube (*see* **Note 2**).
6. Resuspend the concentrated cell pellet in 1 mL of cell wash buffer.
7. Centrifuge at 100× *g* for 20 s at room temperature and transfer cell-containing supernatant to a new tube. Centrifuge at 14,000× *g* for 5 min at room temperature. Discard cell-free supernatant.
8. Resuspend cell pellet in 1 mL of cell wash buffer.
9. Centrifuge at 14,000× *g* for 2 min at room temperature and discard supernatant. Wet cell pellet should weigh ~200 mg (*see* **Note 3**).
10. Proceed to DNA extraction.

3.3.2 DNA Extraction

DNA Extraction Alternative A: Rapid DNA Extraction with Short-Read Elimination

DNA Extraction Using DNeasy® PowerSoil® Pro Kit

Protocol Modified from DNeasy PowerSoil Pro Kit Handbook (03/2021)

1. Spin a PowerBead Pro Tube briefly to ensure that the beads have settled at the bottom. Add up to 250 mg of a biomass sample (or the cell pellet retrieved in Subheading 3.3.1) and 800 µL of Solution CD1 (*see* **Note 4**). Vortex briefly to mix or invert several times.
2. Secure the PowerBead Pro Tube horizontally on a vortex adapter for 1.5–2 mL tubes and vortex at maximum speed for 10 min (*see* **Note 5**).
3. Centrifuge the PowerBead Pro Tube at 15,000× *g* for 1 min and transfer the supernatant (500–600 µL) to a clean 2 mL microcentrifuge tube. In this step, the supernatant may still contain some plant fiber particles.
4. Add 200 µL of Solution CD2 and vortex for 5 s. Centrifuge at 15,000× *g* for 1 min, and, while avoiding the pellet, transfer up to 600 µL of the supernatant to a clean 2 mL microcentrifuge tube.

5. Add 600 µL of Solution CD3 and vortex for 5 s. The solution CD3 together with the silica membrane in the MB Spin Column will precipitate contaminants, leaving only the DNA bound on the membrane.
6. Load 650 µL of the lysate onto an MB Spin Column and centrifuge at 15,000× *g* for 1 min.
7. Discard the flow-through and repeat **step 6** to ensure that all the lysate has passed through the MB Spin Column.
8. Carefully place the MB Spin Column into a clean 2 mL collection tube.

   CAUTION Avoid splashing any flow-through onto the MB Spin Column, as the solution likely contains contaminants.
9. Add 500 µL of Solution EA to the MB Spin Column. Centrifuge at 15,000× *g* for 1 min.
10. Discard the flow-through and place the MB Spin Column back into the same 2 mL collection tube.
11. Add 500 µL of Solution C5 to the MB Spin Column. Centrifuge at 15,000× *g* for 1 min.
12. Discard the flow-through and place the MB Spin Column into a new 2 mL collection tube.
13. Centrifuge at up to 16,000× *g* for 2 min. Carefully place the MB Spin Column into a new 1.5 mL elution tube.
14. Add 50–100 µL of Solution C6 to the center of the white filter membrane, make sure to not touch it, and pay attention that the entire membrane is wet. Centrifuge at 15,000× *g* for 1 min to elute the DNA. Discard the MB Spin Column and store the eluted DNA at 4 °C for days or frozen for longer term.

Short-Read Elimination

Protocol Modified from Short Read Eliminator Kit Family-Handbook v2.0 (07/19), Size Selection Protocol for SRE SX

1. Prepare fresh 70% EtOH wash buffer and store at room temperature.
2. Adjust the extracted DNA sample to a total volume of 60 µL and concentration between 25 and 150 ng μL^{-1} (*see* **Note 6**). The DNA sample can be diluted in Buffer EB.
3. Add 60 µL of Buffer SRE or Buffer SRE XL to the sample. Mix thoroughly by gently tapping the tube or by gently pipetting up and down using wide-bore tips.
4. Centrifuge at 10,000× *g* for 30 min at room temperature. Carefully remove the supernatant (*see* **Note 7**).
5. Add 200 µL of the 70% EtOH wash solution to the tube and centrifuge at 10,000× *g* for 2 min at room temperature.
6. Carefully remove the wash solution from tube without disturbing the DNA pellet (*see* **Note 7**).
7. Repeat **steps 5** and **6**.

8. Add 50–100 µL of Buffer EB to the tube and incubate at room temperature for 20 min (*see* **Note 8**).
9. After incubation, gently tap the tube to ensure that the DNA is properly resuspended and mixed, and analyze the quantity and purity of the DNA by Qubit and NanoDrop, respectively. The range of depletion can be estimated using agarose gel electrophoresis or automated electrophoresis (e.g., TapeStation system, Agilent).
10. *Optional*: During short-read elimination, loss of DNA is expected. If the final concentration falls below the required concentration for sequencing, the DNA yield can be increased by re-precipitation in a smaller elution volume. Alternatively, magnetic bead-based systems, such as the ProNex® Size-Selective Purification System can be used to concentrate the yield while also selecting for DNA of a desired length.

DNA Extraction Alternative B: HMW DNA

HMW DNA Extraction

1. Resuspend cell pellet (from Subheading 3.3.1) in 1 mL RBB + C buffer.
2. Incubate for 20 min at 70 °C, mix tube by inversion every 5 min.
3. Split into 2 × 1.5 mL tubes and add NaCl to 0.7 M and 1:10 volume of CTAB buffer.
4. Heat at 70 °C for 10 min.
5. Add an equal volume of chloroform. Mix well and centrifuge at 14,000× *g* for 15 min at room temperature. Transfer aqueous phase to new tube (*see* **Note 9**).
6. Add equal volume of phenol/chloroform/isoamylalcohol (25:24:1). Mix well, centrifuge at 14,000× *g* for 15 min at room temperature, and transfer aqueous phase to new tube. **CAUTION** Phenol, chloroform, and isoamylalcohol are harmful. Handle using appropriate safety equipment and measures.
7. Add 2 × vol of 95% ethanol and mix gently until DNA spools. Use a sterile loop to transfer the DNA to a tube containing 200 µL 70% ethanol (*see* **Note 10**). **CAUTION** Ethanol is flammable. Handle using appropriate safety equipment and measures.
8. Centrifuge at 14,000× *g* for 2 min at room temperature and carefully discard supernatant. Briefly air-dry the pellet.
9. Resuspend in 20–30 µL TE buffer (pH 8.0) and incubate at room temperature for 30–60 min to allow DNA to dissolve, before measuring the DNA concentration (*see* **Note 11**).
10. *Optional*: Some sequencing technologies, such as PacBio sequencing, are extremely sensitive to environmental contaminants. If the DNA quality does not fit the requirements, PacBio

recommends using DNeasy PowerClean Pro DNA Clean-Up Kit (QIAGEN) following the manufacturer's instructions to remove the contaminants and sequence inhibitors.

3.4 Sequencing and Assembly

Numerous HTS technologies are now available, providing cheaper, faster, and higher-throughput sequencing (*see* reviews [13, 29]). Methods that produce short reads (up to 550 base pair [bp] in length), such as Illumina, can generate high sequencing depth at comparatively low costs. Illumina's NovaSeq 6000 can generate up to 6 terabyte of bases per run (2 × 250 bp, dual flow cell run on SP flow cells), whereas the Illumina MiSeq can produce up to 15 Gbp (with 2 × 300 bp), with both platforms exhibiting a mean error rate <1% [30]. However, the high quantity of data for samples with high species complexity often leads to increased difficulties for metagenomic assembly, due to computational requirements. In theory, longer read sequencing technologies can overcome many of the known assembly problems associated with short reads because they have the potential to resolve complex repeats and span entire open reading frames (ORFs). These technologies have traditionally been accompanied with other inherent issues, such as lower sequencing depth and higher error rates. Examples of "third generation" sequencing technology include ONT and single-molecule high-fidelity (HiFi) developed by PacBio. PacBio HiFi can provide high-quality sequences greater than 99% accuracy and about 30–50 kbp in read length [13, 31]. ONT provides reads in the same length range, but is often associated with higher error rates, particularly related to insertions and deletion in homopolymers. A common strategy to overcome this is to use accurate short reads from Illumina to correct the errors ("polishing") of the assembled long reads [9, 32]. Promisingly, recent advances in ONT have demonstrated the potential to reconstruct near-complete microbial genomes from isolates with a modal read accuracy of 99%, without the need of polishing using short reads [33].

Assembly is a key stage required to generate large contiguous sequence fragments (contigs), which are required to maximize the number of ORFs and operons available for downstream CAZyme screening. Assembly algorithms that process metagenomic data are highly sensitive to the read coverage for community members, which is correlated with the species complexity in a sample and the metagenomic sequencing depth. A plethora of assembly algorithms are currently available (reviewed by [29, 34]), including several that are designed to handle large metagenomic datasets such as MEGAHIT [35], metaSPAdes [36], and metaFlye [37]. Many short-read assemblers use a de Bruijn graph approach and initially deconstruct each read into a series of oligomers of a set "word" length (commonly referred to as "*k*-mers"). The *k*-mer length is often a user-specified parameter, with longer *k*-mers overcoming repetitive/non-unique regions in the metagenome at a cost

of reduced coverage and accuracy. In contrast, short *k*-mers generate contigs with higher coverage, but are often shorter in length. Algorithms designed for longer reads have traditionally used an overlap-consensus approach, where sufficiently similar reads (based on an overlapping nucleotide region) would be merged into a contig. Enhanced sequencing depth and improved basecalling, accompanied with an increased popularity of ONT and PacBio, have led to the development of new, long-read assemblers that utilize a modified de Bruijn graph or repeat graph approach, such as Flye [38]. Subheading "Assembly of short read produced by Illumina using *MEGAHIT*" details the assembly of Illumina data using MEGAHIT [35], an iterative (iterates from a small *k* to a large *k*) de Bruijn graph de novo assembler for short-read sequencing data with highly uneven sequencing depth, which is typically characteristic of many metagenomic datasets. Subheading "Assembly of long-reads produced by Oxford Nanopore using *metaFlye*" describes the assembly of ONT data using Flye in metagenome mode ("metaFlye"). While de Bruijn graphs require exact *k-mer* matches, the repeat graphs utilized by Flye are built using approximate matches to handle noisy sequences.

Alternative approaches to reduce the computational strain of metagenomic assembly include the use of taxonomic binning or normalization methods to select subsets of reads that are then assembled separately as well as hybrid assemblies that use data from multiple sequencing platforms [39, 40]. Hybrid assemblies are still infrequent for metagenomic analysis, despite indications that combined approaches yield improvements in assembly contiguity and per-base accuracy, which are important in CAZyme discovery projects that seek to interrogate larger saccharolytic gene clusters (*see* Subheading 3.7.4). In particular, several studies have shown that high confidence reads from Illumina can be used to correct the errors inherent in ONT and PacBio sequences [9, 41]. A combination of ONT long reads and Illumina short reads provides an improvement in assembly statistics such as total assembly size and large contig size [9]. It also improves the assembly of the universal marker genes, which assists in binning and enables enhancements in genome reconstruction of uncultured microorganisms that inhabit complex communities [31, 42].

Subheadings 3.4.1, 3.4.2, and 3.4.3 outline the various stages required to assemble Nanopore and Illumina raw reads, and generate HQ MAGs (*see* **Note 12**). While Illumina data can provide high-quality metagenomic assemblies from complex microbial communities, long reads enhance the reconstruction of HQ MAGs, which are conducive to CAZyme searches. Specific examples include large datasets that have been assembled from the rumen microbiome [32, 43]. The outlined workflow is based on short-read sequencing by NovaSeq Illumina technology and long-read sequencing using Oxford Nanopore Technology. The ONT sequences were attained through the ligation sequencing kit

SQK-LSK109 (Oxford Nanopore Technologies) using flow cells from the R.9.4 generation on a MinION sequencer. While this is currently a rapidly applied approach, it should be noted that the newest ONT nanopore R10.4.1 flow cell technology significantly improves the error rate upon basecalling of the reads. This advance now provides a way to assemble high-coverage metagenomic genomes without the need for Illumina polishing [33], which greatly simplifies the complexity of the genome reconstruction pipeline and minimizes any bias that might arise from the assumptions that the short-read polishing pipelines rely on.

When assembling metagenomic datasets from complex communities, chimeric assemblies (misassemblies) can occur. Misassemblies can be prevented by producing paired-end reads, whereby one read of the pair may map to a common sequence or a repetitive element (ambiguous region with risk for chimeric assembly), whereas the other can potentially map to a non-ambiguous region and will limit misassembly. Contigs should therefore always be inspected. Abrupt change in GC% content and read coverage within the same contig can indicate chimeric assembly. Changes in the contigs' characteristics can be visualized using different tools such as Anvi'o [44] or MGAviewer [45].

3.4.1 Quality Control and Filtering

1. Filter and trim low-quality Illumina raw reads using *FastP*:

```
fastp –input1 *.R1.fq.gz –input2 *.R2.fq.gz --output1 trimmed.
R1.fq.gz ––output2 trimmed.R2.fq.gz -h trimmed_report.html -R
trimmed_report -q 30
```

   The flag "-q" represents the quality value. A complete report of filtering and trimming can be found in the "trimmed_report.html" file.

2. Inspect the quality of the long reads generated by Oxford Nanopore sequencing using *NanoPlot*:

```
Nanoplot –N50 --fastq input.fq.gz -o output.Nanoplot.dir
```

   Nanoplot requires fastq files which can be compressed (bgzip, bzip2, or gzip) and will create multiple output files, including plots for visualization of the quality metrics, such as read length histogram. The flag "--N50" in the command line indicates the N50 mark in the read length histogram, while "-o" specifies the directory in which the output is generated.

3. Filter out low-quality long reads using *Filtlong*:

```
filtlong --min_length 5000 --keep_percent 95 input.fastq.gz |
gzip > trimmed_ONT.fastq.gz
```

Filtlong requires fastq files as input files. The parameter "--min_length" will discard any read which is shorter than the set cutoff, in this case 5 kbp, while "--keep percentage" throws out the worst 5% of the read. More parameters and information can be found in the github page of Filtlong. After the filtering, perform a quality check of the trimmed reads using the NanoPlot command (above) to check the efficiency of the trimming.

3.4.2 Metagenomic Assembly

Input: Trimmed paired reads in fastq format.

Output: A directory containing the assembly fasta file with contig sequences.

Usage:

Assembly of Short Read Produced by Illumina Using *MEGAHIT*

```
megahit -t <No.CPUs> -m <RAM> -1 trimmed_R1.fq.gz -2 trimmed_R2.fq.gz -o <output directory>
```

Assembly of Long Reads Produced by Oxford Nanopore Using *metaFlye*

Input: Basecalled and trimmed reads in fasta or fastq format, the files must be compressed in gz.

Output: A directory containing the ONT draft assembly fasta file.

Usage:

```
Flye --meta --threads <No.CPUS> --nano-raw trimmed_ONT.fastq.gz –o <output directory>
```

The flag "--nano-raw" is the default mode for regular ONT data, whereas the parameter "--meta" indicates the metagenome mode (uneven coverage mode).

3.4.3 Error Correction of ONT Long Reads Using Racon, Medaka, and Polypolish

A combination of several polishing strategies and tools is often needed to maximize the accuracy of the final contigs. Thus, error correction of ONT long reads usually contains several steps, including both long-read polishing (**steps 1** and **2**) and short-read polishing (**step 3**):

1. Error correction of ONT long-reads with *Racon*:

   First, map the ONT reads to the sequences from the assembly to generate an overlap file, e.g., using Minimap2 with the following command line:

```
minimap2 -x ava-ont assembly.fasta trimmed_ONT.fastq.gz > overlap.paf
```

   Then use Racon to polish the draft ONT assembly:

   Input: The trimmed ONT reads, the assembly file in fasta format, and overlap file(s).

Output: Consensus fasta sequences.
Usage:

```
racon -t <No.CPUs> trimmed_ONT.fastq.gz overlaps.paf assembly.fasta > racon.consensus.fasta
```

2. Error correction of ONT long-reads with *Medaka*:
   Input: The trimmed ONT reads and the fasta file generated by Racon.
   Output: Consensus fasta sequences.
   Usage:

```
medaka_consensus -i input.fastq.gz -d racon.consensus.fasta -o <output directory> -m r941_min_sup_g507
```

   The flag "-m" indicates the model of the basecaller and should be changed accordingly. The consensus.fasta generated by Medaka will be saved to the output directory.
3. Error correction of ONT long-reads with *Polypolish*:
   Before running Polypolish, use an aligner, e.g., bwa-mem2, to align the accurate short reads from Illumina sequencing against the ONT consensus contigs generated by Medaka. This will create SAM files.

```
bwa-mem2 mem -t <No.CPUs> medaka.consensus.fasta trimmed_R1.fastp.fq.gz > aligned.R1.sam
```

```
bwa-mem2 mem -t <No.CPUs> medaka.consensus.fasta trimmed_R2.fastp.fq.gz > aligned.R2.sam
```

   Run Polypolish to polish the ONT assemblies with the Illumina short reads:

```
polypolish_insert_filter.py --in1 aligned.R1.sam --in2 aligned.R2.sam --out1 filtered_R1.sam --out2 filtered_R2.sam
```

```
polypolish medaka.consensus.fasta filtered_R1.sam filtered_R2.sam > polished_assembly.fasta
```

3.5 Binning

Binning is the post-assembly taxonomic assignment of contigs into genome bins/MAGs that enables the study of individual organisms (and their interactions), directly from deeply sequenced metagenomes. Therefore, the task of a binning tool is to assign an identifier to every assembled contig, with each identifier ideally representing a single population genome [46]. The most common binning tools today are based on unsupervised and reference-independent algorithms that traditionally use oligonucleotide composition to group

contigs with similar usage, thus effectively differentiating between contigs of different populations, in particular focusing on their tetranucleotide frequencies [47]. Today, binning tools increasingly leverage additional information to improve genome recovery even in the presence of multiple genomes from individual species in a sample, such as paired-end read linkage [48], mean contig coverage [49], per-sample (differential) coverage [50], or combinations thereof [51, 52]. High-quality genomes can be recovered (Subheading 3.5.1), in particular if multiple metagenomes of the same community were generated, which subsequently can be mined for new CAZymes. Dereplication or comparison of bins is often required when multiple assemblies are available (Subheading 3.5.2) or when several binning tools have been applied independently. In addition, binning results should be inspected carefully by, e.g., looking at taxonomic assignments of individual contigs, visualizing the underlying differential coverage information (as done in Albertsen et al. [53]), or using an automated method for assessing the quality of metagenome-derived microbial genomes [54] (Subheading 3.5.3). Genome bins that have been quality checked and, if necessary, refined are referred to as metagenome-assembled genomes, or MAGs. Accurate taxonomic assignment of the reconstructed MAGs is necessary to identify populations within the studied ecosystem and can be used as anchoring points for hypothesis on metabolic functions, including carbohydrate-degrading populations (Subheading 3.5.4). Overall, computational tool development for binning is a very active research area. The "Critical Assessment of Metagenomic Information" (CAMI) initiative [55] continuously benchmarks tools for binning, metagenome assembly, and profiling on various benchmark datasets reflecting common experimental setups and properties of underlying microbial communities. Up-to-date evaluation results for several use cases and commonly utilized software are available at: https://data.cami-challenge.org/.

3.5.1 Binning with MetaBAT2

Input: Metagenome assemblies from long-read and short-read sequencing (*see Protocol 3.4.1*) and read mapping file(s) in BAM format, one file per sample.

Output: One fasta file per genome bin.

Usage:

```
runMetaBat.sh assembly.fa sample1.bam [sample2.bam ...]
```

Usually, there is a trade-off between a tool's sensitivity and specificity. MetaBAT's default settings work reasonably well for most use cases. However, for very simple or very complex communities, non-default options might improve binning results. To display the complete list of options, please run: `metabat2 -h`

3.5.2 Dereplication of Bins with dRep

Input: All genome bins (generated from both short- and long-read sequencing), in fasta format.

Output: Dereplicated genome bins will be written to this folder. These bins should be used in the downstream analysis.

Usage:

```
dRep dereplicate <output directory> –g <bin directory> -p <No.
CPU> -comp <value> -con <value>
```

The dereplicate parameter will reduce the sets of genomes based on high gene similarity. It is, in addition, possible to filter the bins based on completeness ("-comp") and contamination ("-con").

3.5.3 Quality Assessment with CheckM2

Input: Dereplicated genome bins in fasta format.

Output: Contamination and completeness estimates for each genome bin.

Usage:

```
checkm2 predict --threads <No. CPUs> --input <bin directory>
-x fna --output-directory <output directory>
```

As a result, CheckM2 produce a tabular file ("quality_report.tsv") with the ID of each genome bin and their completeness and contamination. The flag "-x" refers to the extension of the input files (default: fna). Genome bins of high quality, hereby referred to as MAGs, are used in the downstream analysis.

3.5.4 Assigning Taxonomy to MAGs Using GTDB-Tk

Input: MAGs in fasta format.

Output: A folder with two main output files with the prefix "ar53" and "bac120" for Archaea and Bacteria, respectively.

Usage:

```
gtdbtk classify_wf --genome_dir <your MAGs> --out_dir <output
directory> --cpus <No.CPUs>
```

3.6 Gene Calling

Once a metagenomic dataset has been adequately assembled and taxonomically assigned, gene calling or ORF prediction is required to identify protein or RNA coding regions within the (meta)-genome. Depending on the assembly, its feasibility, and its success, gene calling can be performed on assembled contigs or raw reads (for long-read HTS data). There are two different ways for ORF prediction. The "sequence similarity-based" method and the "ab initio" gene-calling method [56]. The "sequence similarity-based" method uses homology searches to identify genes similar to those already present in databases. This method possesses high specificity and the ability to characterize functions of predicted genes. The "ab

initio" gene-calling approach relies on dependencies between codon frequencies and genome nucleotide composition to discriminate coding from noncoding regions. Frequently, metagenomic assemblies result in many genes that are partially sequenced or fragmented. In addition, metagenomic data from diverse communities can have too low similarities with sequences from databases due to evolutionary distance or short contig/read lengths, which can prevent the identification of homologs and poor detection of novel genes. Therefore, *ab initio* tools such as Prodigal [57] are essential for metagenomic analysis, especially when looking for novel enzymes [56]. Prodigal has been successfully used to predict ORFs from various metagenomes [9, 10, 32] and is often a key component of MAG annotation tools (e.g., Distilling and Refining Annotations of Metabolism, DRAM), thus requiring minimal effort to operate (DRAM will be covered in Subheading 3.7.3).

3.7 Enzymes/ Pathway Annotation

Gene calling is typically followed by functional annotation, which details comparisons of predicted ORFs to previously annotated sequences present in functional databases. The objective is to generate accurate annotations to correctly identified orthologues. There are multiple approaches to annotate ORFs and numerous tools and databases are publicly available. These include, among others, COG (Clusters of Orthologous Groups) for functional grouping [58], Pfam (the protein families database) for the identification of protein families and domains [59], TIGRfam for full-length protein families [60], and Enzyme Commission (E.C.) numbers for numerical classification scheme for enzymes, based on the chemical reactions they catalyze [61]. In addition, particular databases enable reconstruction of pathway maps for cellular and organismal functions. Key examples include BRENDA (BRaunschweig ENzyme Database [62]), KEGG (Kyoto Encyclopedia of Genes and Genomes) [63], and MetaCyc (Metabolic Pathway Database) [64].

Many functional annotation resources are collectively available via web-based platforms that provide support for visualization and comparative analysis of metagenomic datasets. For example, the US Department of Energy-Joint Genome Institute hosts the Integrated Microbial Genomes with Microbiome Samples—Expert Review (IMG/MER) system, which provides support for functional annotation and curation of metagenomic datasets of interest. A typical pipeline analysis in IMG/MER starts with the user uploading metagenomic contigs and/or unassembled reads. Protein-coding genes are identified using four *ab initio* gene-calling tools: GeneMark, MetaGeneAnnotator, Prodigal, and FragGeneScan. Predicted proteins are compared with protein families and proteomes of selected "core" genomes. Protein sequences are compared with COG using RPS-BLAST [65] and Pfam and TIGRfam using HMMER 3 [66]. Finally, protein-coding genes are

associated with KEGG Orthology terms, EC numbers, and phylogeny using USEARCH [67] against a nonredundant reference database composed of the public genomes available on IMG and KEGG database. Further information and a procedure to submit data on IMG can be found on [68] and on IMG website (https://img.jgi.doe.gov/).

3.7.1 CAZyDB, dbCAN, and Multimodular CAZymes

For metagenomic projects dedicated to CAZyme discovery, it is recommended that ORFs are annotated using a specialized database. The most comprehensive database is the CAZyme database [69] (hereafter called CAZyDB), which specializes in the display and analysis of genomic, structural, and biochemical information on carbohydrate-active enzymes. The CAZyDB contains families of catalytic and ancillary modules that are presented as glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), auxiliary activities (AAs), and the non-catalytic carbohydrate-binding modules (CBMs). The CAZyDB identifies evolutionarily related families using the classification introduced by Bernard Henrissat [70], which are based on significant amino-acid sequence similarity with at least one biochemically characterized founding member. The CAZy group actively develops tools for unambiguous high-throughput modular and functional annotation of CAZymes in sequences issued from genomic and metagenomic efforts. Annotation of unpublished datasets therefore requires collaboration with CAZy researchers; otherwise query proteins are required to be deposited as finished entries in GenBank (or EMBL and DDBJ) and will be analyzed via their operational routines.

Several alternatives to CAZyDB currently exist for automated CAZyme annotation, including dbCAN [71] and CAT (obsolete) [72]; however, neither has the same levels of manual inspection by expert curators that is offered by CAZy. Importantly, dbCAN enables automated and comprehensive annotation that is based on the classification scheme of CAZyDB but relies on a defined signature domain model for each CAZyme family [71]. In addition, signature domains of each CAZyme family are represented by a hidden Markov model that is available to the public and easily amendable to local searches within unpublished metagenomic datasets. Searches against the dbCAN database can be performed either through its web platform, locally (as described below), or as part of an annotation pipeline (e.g., "DRAM"; *see* Subheading 3.7.3).

3.7.2 Search Against dbCAN3 Database

Protein sequences can be loaded on the web server: https://bcb.unl.edu/dbCAN2/index.php.

Web Server

Local Search

1. Download the dbCAN databases and tools using conda:

```
conda create -n run_dbcan python=3.8 dbcan -c conda-forge -c
bioconda

conda activate run_dbcan
```

2. Run: run_dbcan.py <input fasta file> <type of input> --out_-dir <output directory>
3. Complementary information and extra parameters can be found in the following repository: https://github.com/linnabrown/run_dbcan.

3.7.3 Functional Annotation Using DRAM

DRAM is a tool that annotates MAGs using seven different databases for functional gene annotation (dbCAN [71], VOGDB [73], KOfam [74], UniRef [75], MEROPS peptidase database [76], KEGG [63], Pfam [59]) as well as the individual software packages/algorithms needed for their application. It also includes scripts that merge and filter these results into summary tables and produces convenient heatmaps [77]. The brief guide below will show you how to run the DRAM workflow on a set of MAGs using DRAM v1.2.4. By disabling UniRef, which is the most taxing database in terms of memory usage, it is possible to cut down on the RAM usage to around 128 GB. These requirements may necessitate the use of a high-performance computing cluster (HPC) or at least a workstation computer (~128 GB RAM, 20 cores). A typical DRAM workflow can be split into two steps which we will run subsequently: (1) *annotation* and (2) *distillation*. In the (1) annotation step, each genome is run against the seven main databases using various algorithms, and the results are merged into a single table, and in step (2) distillation pathways and functional modules are highlighted based on the called genes. DRAM can annotate any fasta-formatted genomic sequence, such as scaffolds before binning, single-cell amplified genomes (SAGs), isolate–culture assemblies, or metagenome-assembled genomes (MAGs). For this example, we will consider a set of MAGs.

1. Annotate

   Input: MAGs in fasta format.

   Output: A directory named "annotation" containing several tables with summaries of called genes in each input MAG as well as fasta files containing the called genes.

   Usage:

```
DRAM.py annotate -i <your MAGs> -o annotation
```

   When specifying input MAGs, <your MAGs> can be a glob, e.g.: "path/to/mags/*.fasta". However, this glob should be contained within quotation marks ("") as the

command line argument parser will only read the first string given after the -i option key. The DRAM.py annotation command will create a sub-directory named "annotation" where all annotation results will be laid out. By default, DRAM will filter out contigs in the input MAGs with a length of less than 2.5 kilobases. The main result of the annotation step is a single file called "annotations.tsv". For each sample's contigs, it contains exact coordinates for each called gene's position and DNA strand direction, as well as each gene's identifiers for the respective database hits. Additionally, the annotations.tsv file contains a confidence rank based on cross-validating the presence of genes between databases utilizing RBH (Reciprocal Best Hits). These ranks are encoded with letters A to E, where A denotes the highest confidence. There are tables called rRNAs and tRNAs, and the called genes are available in both GFF3 and GenBank formats. The DRAM output from which the contents of the annotations.tsv table are derived is also generated in various other formats. Predicted ORFs are available in both amino-acid and nucleotide formats, with indexing that facilitates customized downstream analysis or use of the DRAM "distill" command which makes further DRAM processing possible.

2. Distill

   DRAM is bundled with an internal database consisting of 3684 genes categorized into various functional modules and pathways. The "distill" step takes the results from the first *annotation* step and uses aforementioned database to create a summary that highlights which metabolic pathway each genome is encoding. Running the "distill" command requires paths to the outputs from the first step which includes the annotations.tsv file as well as the rRNAs and tRNAs files.

   Input: A directory containing the output from the previous "DRAM.py annotate" call.

   Output: A directory named "genome_summaries" containing summaries of functional interpretations of the genes present in each input MAG, including tables and a graphical visualization.

   Usage:

```
DRAM.py distill -i annotation/annotations.tsv -o genome_summaries --trna_path annotation/trnas.tsv --rrna_path annotation/rrnas.tsv
```

   The "distill" command outputs an excel file with individual sheets for eight different functional groups of genes. Each sheet contains an enumeration of genes and the count of occurrences of each of these genes in each processed genome/MAG. DRAM uses the KEGG gene/pathway hierarchy to calculate

the completeness of individual pathways and otherwise uses a manually curated threshold for each functional group to discriminate the presence of the respective pathway. An HTML file with an interactive heatmap summarizing this completeness and presence for all these modules for each genome is also generated.

3.7.4 Identifying Plant Biomass-Degrading Loci

In addition to identifying individual CAZymes, it is valuable to observe a more global picture by identifying the gene localization and organization of CAZymes encoded in the microbial community. This enables visualization of potential plant biomass-degrading operons that are encoded within a metagenome. Several exemplary saccharolytic mechanisms are encoded by large co-regulated gene clusters and have been successfully recovered in metagenomes, including (and not limited to) Gram-negative (gn) and Gram-positive (gp) polysaccharide utilization loci (PULs) as well as cellulosomes [78, 79]. Tools such as the IMG/MER are conducive to identifying gene clusters, whereby once a metagenome has been uploaded, signature protein domains for specific gene clusters can be searched and the surrounding ORFs visualized and functionally interrogated. For gnPULs, several Pfam domains represent the archetypical outer-membrane proteins SusC and SusD [80], including the following IDs: TonB_dep_Rec (PF00593), TonB_C (PF03544) SusD (PF07980), SusD-like (PF12741), SusD-like_2 (PF12771), and SusD-like_3 (PF14322). gpPULs are typically defined by the presence of CAZymes co-localized with one of three classes of transporters [1]: ATP-binding cassette (ABC) transporters (PF00005), phosphoenol-pyruvate: carbohydrate phosphotransferase system (PTS) transporters (PF00381), or major facilitator superfamily (MFS) transporters (Pfam clan CL0015). For cellulosomes, key signatures include cohesin (PF00963) and dockerin (PF00404) domains. In addition to known PULs, other uncharacterized loci can be investigated by using gene identifiers from previously characterized CAZymes as a search query and interrogating surrounding genomic regions. However, such an approach requires manual intervention and is not amendable to large collections of CAZymes. Alternative approaches that use bioinformatics methods to identify uncharacterized loci are discussed below (Subheading 3.8).

3.8 Identifying New Gene Targets

Many CAZymes are multimodular with catalytic modules and one or more additional domains that are often substrate-targeting carbohydrate-binding modules (CBMs). “Module walking” is a method that probes the potential CAZyme activity of unknown regions or domains of ORFs that flank annotated CAZyme domains. This method has been used by Hemsworth et al. [81] to find a new LPMO family (AA11). Based on the observation that several sequences from AA9 LPMO family carry a conserved

domain of unknown function (X278), the authors searched for other multimodular proteins containing that domain, and then looked at the adjacent regions within an X278-encoding ORF and identified hypothetical domains with LPMO-like characteristics. Interestingly, those domains did not exhibit significant similarity to other LPMOs families (AA9 and AA10) and were thus considered as a new LPMO family (AA11).

Other methodologies to consider that can assist in CAZyme identification is the incorporation of additional "meta-omic" data, such as metaproteomics or metatranscriptomics, which are powerful tools when used in combination with metagenomics. In particular, mapping functional meta-omic data against reconstructed and taxonomically assigned metagenomes enables the visualization of the species identity, together with the relative quantity of key CAZymes and proteins that are expressed in response to polysaccharide cues [8]. Such techniques have the potential to detect known CAZymes that are metabolically active, as well as to identify hypothetical genes that are upregulated and/or expressed in response to growth on a particular plant biomass substrate. In such instances, these emphasized ORFs are presented as key targets for downstream biochemical characterization. Using these approaches, gene clusters coding for CAZymes and other activities related to polysaccharide utilization have been identified and functionally characterized in gut bacteria; these include, among others, mechanisms for degradation of the common food additive xanthan gum in the human gut [10] and depolymerization of feed-derived components in the rumen ecosystem [7].

Using today's molecular toolkit together with the constant improvements in sequencing technologies and bioinformatics tools, the mining for CAZymes and novel enzymes is becoming more accessible and amendable. Data can be generated from a wider range of environments, providing a direct way to interrogate uncultivable phylotypes that constitute a microbial community and enable access to untapped sources of new and interesting CAZymes [10].

3.9 Visualization and Integration of Phylogenomics and Abundance of CAZy Genes Detected by DRAM

After assessment of quality (with CheckM2, *see* Subheading 3.5.3), taxonomic (with GTDB-Tk, *see* Subheading 3.5.4) and functional annotation (with DRAM, *see* Subheading 3.7.3) of the MAGs, a visual representation of the phylogenetic clusterization (e.g., phylogenetic tree) and the annotations (e.g., heatmaps) are often desirable. Different software can be used to generate phylogenetic trees of the MAGs, including PhyloPhlAn [82], Anvi'o [83], and IQ-TREE [84]. Online tools like iTOL [85] can further provide visualization of the tree, along with annotation. In the following workflow, we present an in-house developed protocol that utilizes the PhyloPhlAn tool to produce a phylogenetic tree of the MAGs by searching and aligning 400 single-copy phylogenetic markers,

followed by the R package GGTree [86] to merge and visualize the tree with quality metrics (completeness and contamination) and taxonomy annotations of each MAG. Next, we provide an R-script that parses the information of CAZy-annotated genes by DRAM, and plot a heatmap of their abundance, clustered by their putative target glycan (e.g., starch, pectin, or chitin) and the phylogenetic group of each MAG.

3.9.1 PhyloPhlAn

The following *bash for loop* will create amino-acid fasta files from the MAGs using Prodigal [57] and subsequently generate a maximum-likelihood phylogenetic tree of all MAGs utilizing PhyloPhlAn:

```
for MAG in ./*.fasta; do; echo "starting gene prediction" ;
prodigal -a $MAG.faa -o $MAG.prodigal.out -f gff -i MAG ; done
mkdir ProteinPredictions
mv *.faa ProteinPredictions
phylophlan -i ProteinPredictions -d phylophlanDataBase -t a --
diversity high -f supermatrix_aa.cfg --verbose --nproc <No.
CPUS>
```

3.9.2 GGTree

The following R-script provides an example of how the generated MAG information can be visualized using the GGTree and GGTreeExtra Bioconductor R packages:

https://github.com/TheMEMOLab/MetaGVisualToolBox/blob/main/scripts/GenoTaxoTree.R

Input: CheckM2 tabular results, GTDB-Tk classification tables of Bacteria and Archaea, PhyloPhlAn phylogenetic tree.

Output: A figure of the phylogenetic tree produced by PhyloPhlAn annotated with taxonomy and quality information from GTDB-Tk and CheckM2 by circular heatmaps (*see* Fig. 2a).

Usage:

```
git clone https://github.com/TheMEMOLab/MetaGVisualToolBox/
Rscript MetaGVisualToolBox/scripts/GenoTaxoTree.R quality_
report.tsv gtdbtk.bac120.summary.tsv gtdbtk.ar53.summary.tsv
RAxML_result.Proteins_refined.tre OutputName
```

3.9.3 CAZy Heatmap

The following R-script provides an example of how to parse the information obtained by the DRAM.py distill command to extract all the CAZy genes encoded in the MAGs, add the taxonomic information, and then visualize the abundance of all the CAZy domains targeting particular glycans using a heatmap:

https://github.com/TheMEMOLab/MetaGVisualToolBox/blob/main/scripts/CAZYheatmap.R

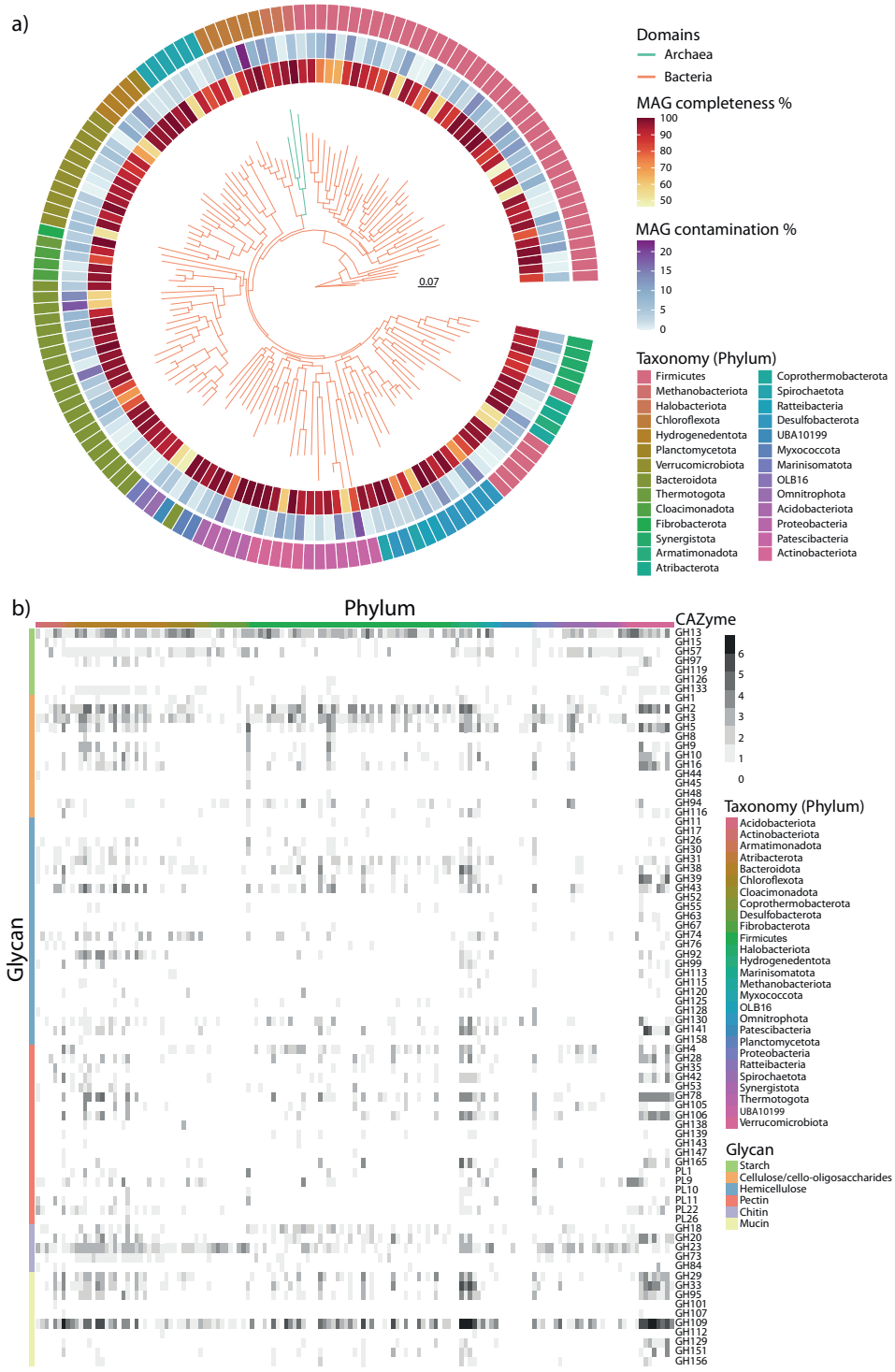


Fig. 2 Visualization of phylogeny and CAZy encoding genes in MAGs. (**a**) Phylogenetic tree, with phyla taxonomy and MAG quality (completeness and contamination). (**b**) Heatmap displaying the abundance (in log2) of CAZy

Input: metabolism_summary.xlsx file from *DRAM distill*, CheckM2 tabular results, GTDB-Tk classification tables of Bacteria and Archaea.

Output: A pdf file with a heatmap showing the abundance of CAZy encoding domains in each MAG clustered by taxonomy (columns) and glycan target (rows; *see* Fig. 2b).

Usage:

```
Rscript MetaGVisualToolBox/scripts/CAZYheatmap.R metabolism_-
summary.xlsx quality_report.tsv gtdbtk.bac120.summary.tsv
gtdbtk.ar53.summary.tsv OutputName
```

4 Notes

1. Biomass samples can be stored at 4 °C in phenol/ethanol (5%/95%) pH 8.0 for several weeks or as a biomass sample at −80 °C for longer periods. We recommend processing the samples as quickly as possible. Longer storage may result in differential lysis of microbial cells.
2. After two to three repetitions, the cell-containing supernatant in **step 4** should become much clearer as cells are removed and collected. Long exposure of samples to dissociation buffer should be avoided due to its low pH. It's recommended that no more than two to three repetitions be performed.
3. Dissociated cells may be stored at 4 °C for 1 day or − 20 °C for several weeks.
4. Some samples, such as rumen biomass, will contain undigested material. Thus, if Subheading 3.3.1 is omitted, it is important to use wide-bore pipette tips to avoid clogging.
5. Homogenizing samples at higher speed may increase yields but will likely result in more fragmented DNA. Cell lysis can alternatively be carried out using PowerLyzer 24 or TissueLyser.
6. The concentration should be determined using Qubit or PicoGreen assay. The performance of the kit depends on the input of DNA being homogeneous and not viscous (*see* page 14 in the handbook).

Fig. 2 (continued) domains encoded in different MAGs. The colors on the columns represent the different phyla classified by GTDB-Tk and on the rows the putative target glycan. Glycan target information was obtained from DRAM annotation. Letters and numbers in each row represent different catalytic CAZy families: *GH* glycoside hydrolase, *PL* polysaccharide lyase. (The MAGs shown in this figure are recovered from anaerobic digestion enrichment [87] and available from https://figshare.com/articles/dataset/MAGs/13102451)

7. The pellet may not be visible: be careful to not accidentally disturb or aspire the DNA! Always position the tubes in the centrifuge with the same orientation, and aspire the supernatant by pipetting on the opposite side (towards the thumb lid).
8. HMW DNA may take more time to resuspend. Heating to 50 °C or incubating for a longer time can help increase the recovery.
9. CTAB specifically binds to proteins at high salt concentrations. **Steps 3**–**5** should remove cell wall debris, denatured proteins, and polysaccharides complexed, while retaining the nucleic acids in solution. Aqueous phase should be clear before proceeding to **step 6**. If aqueous phase still retains an opaque-yellow color, repeat **steps 3**–**5**.
10. No DNA spool in **step 7** implies either that the DNA is of low concentration or that the DNA has sheared into relatively low-molecular weigth fragments. The DNA can still be collected by centrifugation at 14,000× *g* for 30 min at 4 °C before proceeding to **step 8**.
11. DNA can be stored at 4 °C for short periods, or at −20 °C for longer terms. It is recommended to restrict the number of freeze–thaw cycles as this can degrade HMW DNA.
12. Every metagenome is unique and requires specific consideration and analyses to adapt to its particular confines. The workflow provided here is intended as a guideline for CAZyme discovery through metagenomics and may need modifications according to the origin of the metagenome and the aim of each study.

References

1. La Rosa SL, Ostrowski MP, Vera-Ponce de León A, McKee LS, Larsbrink J, Eijsink VG, Lowe EC, Martens EC, Pope PB (2022) Glycan processing in gut microbiomes. Curr Opin Microbiol 67:102143. https://doi.org/10.1016/j.mib.2022.102143
2. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottesen EA, Zhang X, Hernandez M, Murillo C, Acosta LG, Rigoutsos I, Tamayo G, Green BD, Chang C, Rubin EM, Mathur EJ, Robertson DE, Hugenholtz P, Leadbetter JR (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450(7169):560–565. https://doi.org/10.1038/nature06269
3. Liu N, Li H, Chevrette MG, Zhang L, Cao L, Zhou H, Zhou X, Zhou Z, Pope PB, Currie CR, Huang Y, Wang Q (2019) Functional metagenomics reveals abundant polysaccharide-degrading gene clusters and cellobiose utilization pathways within gut microbiota of a wood-feeding higher termite. ISME J 13(1):104–117. https://doi.org/10.1038/s41396-018-0255-1
4. Hagen LH, Brooke CG, Shaw CA, Norbeck AD, Piao H, Arntzen M, Olson HM, Copeland A, Isern N, Shukla A, Roux S, Lombard V, Henrissat B, O'Malley MA, Grigoriev IV, Tringe SG, Mackie RI, Pasa-Tolic L, Pope PB, Hess M (2021) Proteome specialization of anaerobic fungi during ruminal degradation of recalcitrant plant fiber. ISME J 15(2):421–434. https://doi.org/10.1038/s41396-020-00769-x

5. Naas AE, Solden LM, Norbeck AD, Brewer H, Hagen LH, Heggenes IM, McHardy AC, Mackie RI, Paša-Tolić L, Arntzen M, Eijsink VGH, Koropatkin NM, Hess M, Wrighton KC, Pope PB (2018) "Candidatus Paraporphyromonas polyenzymogenes" encodes multi-modular cellulases linked to the type IX secretion system. Microbiome 6(1):44. https://doi.org/10.1186/s40168-018-0421-8
6. Peng X, Wilken SE, Lankiewicz TS, Gilmore SP, Brown JL, Henske JK, Swift CL, Salamov A, Barry K, Grigoriev IV, Theodorou MK, Valentine DL, O'Malley MA (2021) Genomic and functional analyses of fungal and bacterial consortia that enable lignocellulose breakdown in goat gut microbiomes. Nat Microbiol 6(4):499–511. https://doi.org/10.1038/s41564-020-00861-0
7. Solden LM, Naas AE, Roux S, Daly RA, Collins WB, Nicora CD, Purvine SO, Hoyt DW, Schückel J, Jørgensen B, Willats W, Spalinger DE, Firkins JL, Lipton MS, Sullivan MB, Pope PB, Wrighton KC (2018) Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. Nat Microbiol 3(11): 1274–1284. https://doi.org/10.1038/s41564-018-0225-4
8. Delogu F, Kunath BJ, Evans PN, Arntzen M, Hvidsten TR, Pope PB (2020) Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes. Nat Commun 11(1):4708. https://doi.org/10.1038/s41467-020-18543-0
9. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, Kondrotaite Z, Karst SM, Dueholm MS, Nielsen PH, Albertsen M (2021) Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. Nat Commun 12(1):2009. https://doi.org/10.1038/s41467-021-22203-2
10. Ostrowski MP, La Rosa SL, Kunath BJ, Robertson A, Pereira G, Hagen LH, Varghese NJ, Qiu L, Yao T, Flint G, Li J, McDonald SP, Buttner D, Pudlo NA, Schnizlein MK, Young VB, Brumer H, Schmidt TM, Terrapon N, Lombard V, Henrissat B, Hamaker B, Eloe-Fadrosh EA, Tripathi A, Pope PB, Martens EC (2022) Mechanistic insights into consumption of the food additive xanthan gum by the human gut microbiota. Nat Microbiol 7(4): 556–569. https://doi.org/10.1038/s41564-022-01093-0
11. Bickhart DM, Kolmogorov M, Tseng E, Portik DM, Korobeynikov A, Tolstoganov I, Uritskiy G, Liachko I, Sullivan ST, Shin SB, Zorea A, Andreu VP, Panke-Buisse K, Medema MH, Mizrahi I, Pevzner PA, Smith TPL (2022) Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. Nat Biotechnol 40(5):711–719. https://doi.org/10.1038/s41587-021-01130-z
12. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15(2):121–132. https://doi.org/10.1038/nrg3642
13. Tedersoo L, Albertsen M, Anslan S, Callahan B (2021) Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. Appl Environ Microbiol 87(17): e0062621. https://doi.org/10.1128/aem.00626-21
14. Kuczynski J, Stombaugh J, Walters WA, González A, Caporaso JG, Knight R (2011) Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Bioinformatics Chapter 10 36:Unit 10.17
15. Gilbert JA, Jansson JK, Knight R (2014) The Earth Microbiome project: successes and aspirations. BMC Biol 12:69
16. Royo-Llonch M, Sánchez P, Ruiz-González C, Salazar G, Pedrós-Alió C, Sebastián M, Labadie K, Paoli L, Ibarbalz FM, Zinger L, Churcheward B, Chaffron S, Eveillard D, Karsenti E, Sunagawa S, Wincker P, Karp-Boss L, Bowler C, Acinas SG (2021) Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean. Nat Microbiol 6(12):1561–1574. https://doi.org/10.1038/s41564-021-00979-9
17. Li Z, Wang X, Zhang Y, Yu Z, Zhang T, Dai X, Pan X, Jing R, Yan Y, Liu Y, Gao S, Li F, Huang Y, Tian J, Yao J, Xing X, Shi T, Ning J, Yao B, Huang H, Jiang Y (2022) Genomic insights into the phylogeny and biomass-degrading enzymes of rumen ciliates. ISME J 16:2775–2787. https://doi.org/10.1038/s41396-022-01306-8
18. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PS, Charlson E, Costello

EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, San Gil I, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methe BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spo A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, Yatsunenko T, Glockner FO (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29(5): 415–420. https://doi.org/10.1038/nbt.1823

19. Yilmaz P, Gilbert JA, Knight R, Amaral-Zettler L, Karsch-Mizrachi I, Cochrane G, Nakamura Y, Sansone SA, Glockner FO, Field D (2011) The genomic standards consortium: bringing standards to life for microbial ecology. ISME J 5(10):1565–1567. https://doi.org/10.1038/ismej.2011.39
20. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35(8):725–731. https://doi.org/10.1038/nbt.3893
21. Burke C, Kjelleberg S, Thomas T (2009) Selective extraction of bacterial DNA from the surfaces of macroalgae. Appl Environ Microbiol 75(1):252–256. https://doi.org/10.1128/AEM.01630-08
22. Solomon R, Wein T, Levy B, Eshed S, Dror R, Reiss V, Zehavi T, Furman O, Mizrahi I, Jami E (2022) Protozoa populations are ecosystem engineers that shape prokaryotic community structure and function of the rumen microbial ecosystem. ISME J 16(4):1187–1197. https://doi.org/10.1038/s41396-021-01170-y
23. Delmont TO, Robe P, Clark I, Simonet P, Vogel TM (2011) Metagenomic comparison of direct and indirect soil DNA extraction approaches. J Microbiol Methods 86(3): 397–400. https://doi.org/10.1016/j.mimet.2011.06.013
24. Rosewarne CP, Pope PB, Denman SE, McSweeney CS, O'Cuiv P, Morrison M (2011) High-yield and phylogenetically robust methods of DNA recovery for analysis of microbial biofilms adherent to plant biomass in the herbivore gut. Microb Ecol 61(2): 448–454. https://doi.org/10.1007/s00248-010-9745-z
25. Denman SE, Martinez Fernandez G, Shinkai T, Mitsumori M, McSweeney CS (2015) Metagenomic analysis of the rumen microbial community following inhibition of methane formation by a halogenated methane analog. Front Microbiol 6:1087
26. Cardenas E, Kranabetter JM, Hope G, Maas KR, Hallam S, Mohn WW (2015) Forest harvesting reduces the soil metagenomic potential for biomass decomposition. ISME J 9:2465–2476
27. Marine R, McCarren C, Vorrasane V, Nasko D, Crowgey E, Polson SW, Wommack KE (2014) Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. Microbiome 2:3
28. Binga EK, Lasken RS, Neufeld JD (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. ISME J 2:233–241
29. Bragg L, Tyson GW (2014) Metagenomics using next-generation sequencing. Methods Mol Biol 1096:183–201
30. Laehnemann D, Borkhardt A, McHardy AC (2016) Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Brief Bioinform 17:154–179
31. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, Knight R, Albertsen M (2021) High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat Methods 18(2):165–169. https://doi.org/10.1038/s41592-020-01041-y

32. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M (2019) Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol 37(8): 953–961. https://doi.org/10.1038/s41587-019-0202-3
33. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M (2022) Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods 19(7): 823–826. https://doi.org/10.1038/s41592-022-01539-7
34. Nagarajan N, Pop M (2013) Sequence assembly demystified. Nat Rev Genet 14:157–167
35. Li D, Liu CM, Luo R, Sadakane K, Lam TW (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676
36. Nurk S, Meleshko D, Korobeynikov A, Pevzner P (2016) metaSPAdes: a new versatile de novo metagenomics assembler. arXiv:160403071
37. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, Pevzner PA (2020) metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods 17(11):1103–1110. https://doi.org/10.1038/s41592-020-00971-x
38. Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37(5): 540–546. https://doi.org/10.1038/s41587-019-0072-8
39. Tsai YC, Conlan S, Deming C, Program NCS, Segre JA, Kong HH, Korlach J, Oh J (2016) Resolving the complexity of human skin metagenomes using single-molecule sequencing. MBio 7(1):e01948
40. Chandrakumar I, Gauthier NPG, Nelson C, Bonsall MB, Locher K, Charles M, MacDonald C, Krajden M, Manges AR, Chorlton SD (2022) BugSplit enables genome-resolved metagenomics through highly accurate taxonomic binning of metagenomic assemblies. Commun Biol 5(1):151. https://doi.org/10.1038/s42003-022-03114-4
41. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Adam MP (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30(7):693–700. https://doi.org/10.1038/nbt.2280
42. Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VG, McHardy AC, Nederbragt AJ, Pope PB (2016) Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep 6:25373
43. Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331:463–467. https://doi.org/10.1126/science.1200387
44. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319
45. Zhu Z, Niu B, Chen J, Wu S, Sun S, Li W (2013) MGAviewer: a desktop visualization tool for analysis of metagenomics alignment data. Bioinformatics 29:122–123
46. McHardy AC, Rigoutsos I (2007) What's in the mix: phylogenetic classification of metagenome sequence samples. Curr Opin Microbiol 10:499–503
47. Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner FO (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5:163
48. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335:587–590. https://doi.org/10.1126/science.1212665
49. Wu YW, Tang YH, Tringe SG, Simmons BA, Singer SW (2014) MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2:26
50. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW (2014) GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ 2: e603. https://doi.org/10.7717/peerj.603
51. Alneberg J, Bjarnason BS, Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C (2014) Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146. https://doi.org/10.1038/nmeth.3103
52. Kang DD, Froula J, Egan R, Wang Z (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex

microbial communities. PeerJ 3:e1165. https://doi.org/10.7717/peerj.1165

53. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 31: 533–538. https://doi.org/10.1038/nbt.2579
54. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114
55. Meyer F, Fritz A, Deng ZL, Koslicki D, Lesker TR, Gurevich A, Robertson G, Alser M, Antipov D, Beghini F, Bertrand D, Brito JJ, Brown CT, Buchmann J, Buluç A, Chen B, Chikhi R, Clausen P, Cristian A, Dabrowski PW, Darling AE, Egan R, Eskin E, Georganas E, Goltsman E, Gray MA, Hansen LH, Hofmeyr S, Huang P, Irber L, Jia H, Jørgensen TS, Kieser SD, Klemetsen T, Kola A, Kolmogorov M, Korobeynikov A, Kwan J, LaPierre N, Lemaitre C, Li C, Limasset A, Malcher-Miranda F, Mangul S, Marcelino VR, Marchet C, Marijon P, Meleshko D, Mende DR, Milanese A, Nagarajan N, Nissen J, Nurk S, Oliker L, Paoli L, Peterlongo P, Piro VC, Porter JS, Rasmussen S, Rees ER, Reinert K, Renard B, Robertsen EM, Rosen GL, Ruscheweyh HJ, Sarwal V, Segata N, Seiler E, Shi L, Sun F, Sunagawa S, Sørensen SJ, Thomas A, Tong C, Trajkovski M, Tremblay J, Uritskiy G, Vicedomini R, Wang Z, Wang Z, Wang Z, Warren A, Willassen NP, Yelick K, You R, Zeller G, Zhao Z, Zhu S, Zhu J, Garrido-Oter R, Gastmeier P, Hacquard S, Häußler S, Khaledi A, Maechler F, Mesny F, Radutoiu S, Schulze-Lefert P, Smit N, Strowig T, Bremges A, Sczyrba A, McHardy AC (2022) Critical Assessment of Metagenome Interpretation: the second round of challenges. Nat Methods 19(4): 429–440. https://doi.org/10.1038/s41592-022-01431-4
56. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P (2008) A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev 72(4):557–578, Table of Contents. https://doi.org/10.1128/MMBR.00009-08
57. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119
58. Galperin MY, Makarova KS, Wolf YI, Koonin EV (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43(Database issue):D261–D269. https://doi.org/10.1093/nar/gku1223
59. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49 (D1):D412–d419. https://doi.org/10.1093/nar/gkaa913
60. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E (2013) TIGRFAMs and genome properties in 2013. Nucleic Acids Res 41(Database issue):D387–D395. https://doi.org/10.1093/nar/gks1234
61. Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28:304–305
62. Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, Schomburg D (2021) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res 49 (D1):D498–d508. https://doi.org/10.1093/nar/gkaa1025
63. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2015) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 43:1–6
64. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 42(Database issue):D459–D471. https://doi.org/10.1093/nar/gkt1103
65. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A (2020) CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48(D1):D265–D268. https://doi.org/10.1093/nar/gkz991
66. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD (2018) HMMER web server: 2018 update. Nucleic Acids Res 46 (W1):W200–W204. https://doi.org/10.1093/nar/gky448

67. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460–2461. https://doi.org/10.1093/bioinformatics/btq461
68. Chen IA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter S, Varghese N, Seshadri R, Roux S, Woyke T, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC (2021) The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic Acids Res 49(D1):D751–D763. https://doi.org/10.1093/nar/gkaa939
69. Drula E, Garron ML, Dogan S, Lombard V, Henrissat B, Terrapon N (2022) The carbohydrate-active enzyme database: functions and literature. Nucleic Acids Res 50(D1):D571–D577. https://doi.org/10.1093/nar/gkab1045
70. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res 37(suppl_1):233–238
71. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y (2018) dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 46(W1):W95–W101. https://doi.org/10.1093/nar/gky418
72. Park BH, Karpinets TV, Syed MH, Leuze MR, Uberbacher EC (2010) CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. Glycobiology 20:1574–1584
73. Marz M, Beerenwinkel N, Drosten C, Fricke M, Frishman D, Hofacker IL, Hoffmann D, Middendorf M, Rattei T, Stadler PF, Töpfer A (2014) Challenges in RNA virus bioinformatics. Bioinformatics 30(13):1793–1799. https://doi.org/10.1093/bioinformatics/btu105
74. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36(7):2251–2252. https://doi.org/10.1093/bioinformatics/btz859
75. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31(6):926–932. https://doi.org/10.1093/bioinformatics/btu739
76. Rawlings ND, Barrett AJ, Bateman A (2010) MEROPS: the peptidase database. Nucleic Acids Res 38(Database issue):D227–D233. https://doi.org/10.1093/nar/gkp971
77. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, Liu P, Narrowe AB, Rodríguez-Ramos J, Bolduc B, Gazitúa MC, Daly RA, Smith GJ, Vik DR, Pope PB, Sullivan MB, Roux S, Wrighton KC (2020) DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res 48(16):8883–8900. https://doi.org/10.1093/nar/gkaa621
78. Rosewarne CP, Pope PB, Cheung JL, Morrison M (2014) Analysis of the bovine rumen microbiome reveals a diversity of Sus-like polysaccharide utilization loci from the bacterial phylum Bacteroidetes. J Ind Microbiol Biotechnol 41(3):601–606
79. Zhou Y, Pope PB, Li S, Wen B, Tan F, Cheng S, Chen J, Yang J, Liu F, Lei X, Su Q, Zhou C, Zhao J, Dong X, Jin T, Zhou X, Yang S, Zhang G, Yang H, Wang J, Yang R, Eijsink VG, Wang J (2014) Omics-based interpretation of synergism in a soil-derived cellulose-degrading microbial community. Sci Rep 4:5288
80. Martens EC, Koropatkin NM, Smith TJ, Gordon JI (2009) Complex glycan catabolism by the human gut microbiota: the bacteroidetes Sus-like paradigm. J Biol Chem 284:24673–24677. https://doi.org/10.1074/jbc.R109.022848
81. Hemsworth GR, Henrissat B, Davies GJ, Walton PH (2014) Discovery and characterization of a new family of lytic polysaccharide monooxygenases. Nat Chem Biol 10:122–126
82. Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, Zhu Q, Bolzan M, Cumbo F, May U, Sanders JG, Zolfo M, Kopylova E, Pasolli E, Knight R, Mirarab S, Huttenhower C, Segata N (2020) Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat Commun 11(1):2500. https://doi.org/10.1038/s41467-020-16366-7
83. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN, Yousef M, Fogarty EC, Trigodet F, Watson AR, Esen ÖC, Moore RM, Clayssen Q, Lee MD, Kivenson V, Graham ED, Merrill BD, Karkman A, Blankenberg D, Eppley JM, Sjödin A, Scott JJ, Vázquez-Campos X, McKay LJ, McDaniel EA, Stevens SLR, Anderson RE, Fuessel J, Fernandez-Guerra A, Maignien L, Delmont TO, Willis AD (2021) Community-led, integrated, reproducible multi-omics with anvi'o. Nat Microbiol 6(1):3–6. https://doi.org/10.1038/s41564-020-00834-3

84. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32(1): 268–274. https://doi.org/10.1093/molbev/msu300
85. Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49(W1):W293–W296. https://doi.org/10.1093/nar/gkab301
86. Yu G (2020) Using ggtree to visualize data on tree-like structures. Curr Protoc Bioinformatics 69(1):e96. https://doi.org/10.1002/cpbi.96
87. Jonassen KR, Hagen LH, Vick SHW, Arntzen M, Eijsink VGH, Frostegård Å, Lycus P, Molstad L, Pope PB, Bakken LR (2022) Nitrous oxide respiring bacteria in biogas digestates for reduced agricultural emissions. ISME J 16(2):580–590. https://doi.org/10.1038/s41396-021-01101-x

Paper #2

2024, Preprint in Biorxiv, Cold Spring Harbor Laboratory

doi: 10.1101/2024.07.12.603264

CompareM2 is a genomes-to-report pipeline for comparing microbial genomes

Carl M. Kobel*[1], Velma T. E. Aho[1], Ove Øyås[1], Niels Nørskov-Lauritsen[2], Ben J. Woodcroft[3], Phillip B. Pope[1,3,4]

1 Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway.
2 Clinical Institute, University of Southern Denmark, Odense, Denmark.
3 Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology (QUT), Translational Research Institute, Woolloongabba, Australia.
4 Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway.
* Corresponding author, carl.mathias.kobel@nmbu.no

Abstract

Here, we present CompareM2, a genomes-to-report pipeline for comparative analysis of bacterial and archaeal genomes derived from isolates and metagenomic assemblies. CompareM2 is easy to install and operate, and integrates community-adopted tools to perform genome quality control and annotation, taxonomic and functional predictions, as well as comparative analyses of core- and pan-genome partitions and phylogenetic relations. The central results generated via the CompareM2 workflow are emphasized in a portable dynamic report document. CompareM2 is free software and welcomes modifications and pull requests from the community on its Git repository at https://github.com/cmkobel/comparem2 .

Keywords:

microbiology, bacteria, archaea, genomics, metagenome assembled genomes, bioinformatics, pipeline, workflow, genomic annotation, phylogenetics, parallel computing

Background

Costs are decreasing both for sequencing of microbial genomes and complex microbiomes and for the computational resources necessary to analyze generated reads. This has led to an exponential growth in the number of available genomes and metagenome-assembled genomes (MAGs). Despite this growth, there are limits on the accessibility of software that

can analyze the evolutionary relationships and functional characteristics of microbial genomes in order to assess variation of both known and unknown species. Much of the software commonly used to analyze prokaryotic genomes has a high user entry level, requiring advanced skills for complicated installation procedures, debugging dependency issues, and circumventing operating system-specific limitations. This results in a disproportionate amount of time being spent by researchers on setup and technical preparations needed to analyze the sequenced genomic reads rather than biologically relevant analysis of scientific data. These factors define the backdrop that has motivated the conceptualization, development, and application of the CompareM2 genomes-to-report pipeline, which is designed to be an easy-to-install, easy-to-use bioinformatic pipeline that makes extensive analysis and comparison of microbial genomes straightforward.

We compared CompareM2 to several other pipelines that are designed for overlapping use cases: Nullarbor[1], Tormes[2] (stylized TORMES) and Bactopia[3] (**Table 1**). Nullarbor and Tormes do assembly and comparison and have a focus on antimicrobial resistance, spread of pathogens, and core genomes relevant for analyzing individual species. They both produce a report document that is similar to what CompareM2 produces. Bactopia does both assembly and comparative analyses, but while it does some comparative analyses in conjunction with assembly, the user must launch individual predefined workflows included in the Bactopia Tools extension to compare between the samples. Bactopia does not have a parallel scheduler for running these comparative tools. While it does not produce a report document, it does have more overlapping tools with CompareM2 when considering the Bactopia Tools extension. Neither Tormes nor Bactopia is designed for analyzing archaea, although many of the tools integrated in these pipelines are applicable for archaeal genomes when care is taken, e.g. core/pan genome reconstruction and phylogenetic analysis, etc. Furthermore, there is a lack of tools to analyze archaea which means that in many cases, researchers may opt to use non-archaeal tools for analysis of these. For this reason we have opted to compare them to CompareM2, which is designed to analyze both bacteria and archaea.

Table 1: Qualitative comparison of Nullarbor, Tormes, Bactopia and CompareM2.

	Nullarbor[1]	**Tormes[2]**	**Bactopia[3]**	**CompareM2**
Parallel workflow management (system)	yes (GNU make)	no	yes (Nextflow)	yes (Snakemake)
Built in compatibility with high performance computing (HPC) workload managers.	no	no	yes	yes

Assembly-agnostic characterization	no	no	no	yes
Officially designed for Bacteria and Archaea	no	no	no	yes
Quality control	yes	yes	yes	yes
Annotation	yes	yes	yes	yes
Core/pan genome partitioning	yes	yes	yes (using Bactopia Tools extension)	yes
Phylogenetics	yes	no	yes (using Bactopia Tools extension)	yes
Portable visual report document	yes	yes	no	yes
Automated installation	yes	yes	yes	yes
Minimal number of steps in installation instructions (after installing Conda)	NA	3	1	1
Automated database download and setup	no	yes (no checkpoints)	yes	yes
Conda environment solvable with strict channel priority	NA	no	yes	yes
Docker compatible containerization	NA	no	yes	yes
Conda recipe availability (channel)	yes (bioconda)	no	yes (bioconda)	yes (bioconda)
Age of current release	Approx. 6 years	Approx. 3 years	Approx. 1 month	Approx. 1 month
License	GPL-v2	GPL-v3	MIT	GPL-v3
Current version	2.0.20191013	1.3.0	3.0.1	2.8.1
Repository	github.com/tseemann/nullarbor	github.com/nmquijada/tormes	github.com/bactopia/bactopia	github.com/cmkobel/comparem2

Tormes has a sequential architecture, which means that it runs one sample at a time and one tool at a time. This is in contrast to CompareM2 and Bactopia, which have a parallel job scheduler where several samples and tools can be run at the same time. CompareM2 inherits this property from Snakemake, on which it is built. Bactopia on the other hand is built

on the Nextflow workflow system which in many cases is comparable to Snakemake. Central processing units (CPUs) of computers, whether in laptops, workstations, or HPCs, are seeing an increasing number of physical cores. To take advantage of this, it is necessary for software to have a parallel architecture that can utilize the full potential of the processing resources available. This is especially important on HPCs, where many independent compute nodes can run parallel jobs in a scalable manner.

Another bottleneck in bioinformatics is the interpretation of large output files and visualization of data in an informative manner. CompareM2 produces a graphical report that contains the most important curated results from each of the analyses carried out on the user-specified set of query genomes. This report contains text and figures that explain the significance of the results, which makes it easy to interpret for users with a non-bioinformatics background.

While CompareM2 can be used to compare prokaryotic isolate genomes, it also contains tools to analyze bins or MAGs from the sequencing of large microbial communities. The genome is the foundation of any multi-omics study, and such a resource of annotated genomes can be readily integrated into subsequent multi-omics analyses. For example, metaproteomic searches require a highly specific and well-annotated genome database to match MS/MS spectral data[4,5].

Results

CompareM2 congregates the most commonly used and community-tested tools to perform prokaryotic genome quality control, gene calling, functional annotation, phylogenetic analysis, and comparison of genomes across the core-pan spectrum**.** A major priority of CompareM2 is the ease of installation and use, which is achieved by containerizing all bundled software packages and automatizing the download and setup of databases. The choice of genomes to input can be any set where there is a comparable feature either within or between species. The number is limited by the computational resources but the dynamic report is designed for comparing hundreds of genomes.

Software design

CompareM2 is written as a command line program that the user calls with the input genomes that they wish to analyze. It has a text interface where the user can define optional parameters and a single executable that takes care of the overall procedure: First, it checks for presence of the Apptainer runtime, and defines reasonable defaults for database directories and configuration files, in case the user has not specified these manually as

environment variables. There also is a “passthrough arguments” feature that makes it possible to address any command line argument to any rule in the workflow. (further details in documentation https://comparem2.readthedocs.io/en/latest/). One example of a setting that can be defined via the configuration file is whether to optionally submit jobs through a workload manager like Slurm, PBS, etc., typically used on high-performance computing clusters (HPCs). Next, the executable dispatches the main Snakemake pipeline that runs all genomic analyses. This main pipeline automatically installs all necessary software environments and automatically downloads necessary databases, depending on which analyses the user has selected to run. Finally, it dispatches rendering of the dynamic report which contains the results of the main pipeline. This report is dynamic in the sense that it only includes the results which are present, which means that it can be rendered independently of which analyses the user has selected to compute.

Overall, CompareM2 is designed in such a way that the user can install the complete software in a single step. Similarly, running all analyses on a set of microbial genomes (bacterial and archaeal) can be launched in a single action, and the curated results can be studied in the dynamically rendered report. The machine requirements are a Linux-compatible OS with a Conda-compatible package manager, e.g., Miniforge, Mamba or Miniconda. There is nothing standing in the way of running CompareM2 on other operating systems, but many of the included bioinformatic tools for genomic analysis are mostly compatible with Linux-like x64-based systems. For a technical description of how CompareM2 is implemented, please see the Methods section.

For demo reports, please see https://comparem2.readthedocs.io/en/latest/30%20what%20analyses%20does%20it%20do/#rendered-report .

Benchmarking

Initially, we wanted to compare CompareM2 to Nullarbor[1], Tormes[2] and Bactopia[3]. As none of these tools support the external long-reads based assembly, binning and dereplication pipeline where our MAGs were sourced from, we inputted the finished MAGs as is into these tools. Unfortunately this was not possible for Nullarbor, as it is not able to run without reads[6]. Nonetheless, we have included Nullarbor in **Table 1** for the purpose of a qualitative comparison.

We compared the running times of CompareM2, Tormes, and Bactopia when scaling up the number of input MAGs to analyze on a single workstation. We considered two different genera: *Methanobrevibacter*, which are archaea from the class Methanobacteria, and

Prevotella, which are Gram-negative bacteria from the class Bacteroidia. Our MAGs have an average genome size of 2.19 Mb for *Methanobrevibacter* and 3.07 Mb for *Prevotella*. Species prediction and genome sizes are measured on the analyzed MAGs with GTDB-Tk[7] and assembly-stats[8] using CompareM2 itself.

Although Bactopia, Tormes, and CompareM2 are designed for overlapping use cases, they are still very different, because they implement different kinds of analyses. In order to make them as comparable as possible, we ran only the analyses with pairwise overlap between CompareM2 and each of the two other tools. This was done using CompareM2's "until" parameter to specify exactly which rules to run. CompareM2 in "Bactopia mode" includes rules sequence_lengths, assembly_stats, prokka, abricate, and mlst, whereas CompareM2 in "Tormes mode" includes rules prokka, abricate, assembly_stats, mlst, panaroo, and gtdbtk.

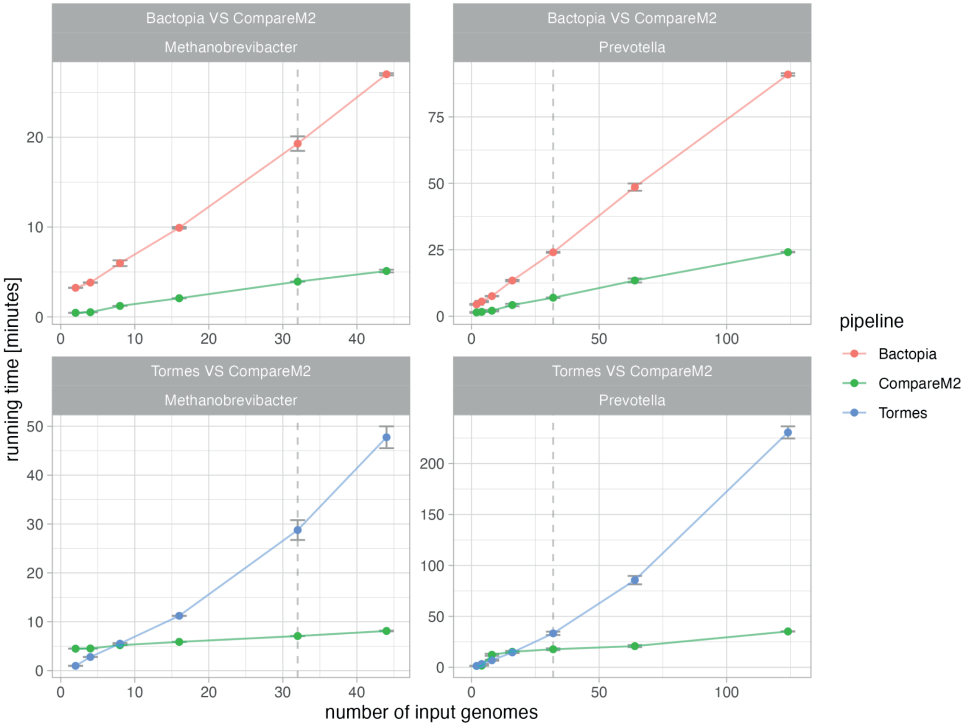


***Fig. 1**: Wall running time analysis comparing CompareM2 to Bactopia and Tormes. In each comparison, CompareM2 was run in a mode where it creates a comparable set of results to the pipeline it is compared to. All analyses ran with 3 replicates, the error bars show means ± the standard deviation of these replicates. A vertical dashed line highlights input size = 32 which is equal to the number of cores used in each benchmark.*

We analyzed the running time of Bactopia, Tormes and CompareM2 when increasing the input size (number of input genomes) (**Fig. 1**). The running times of all tools were approximately linear functions of the input size (time = input size × slope + constant) but with big differences between pipelines in the slope. There are hints of an exponential component in the scaling of running time for Tormes and CompareM2 in Tormes mode, likely because these pipelines construct core genomes, which is a computationally expensive problem where all genes, in the case of Panaroo and Roary, are compared in a pairwise manner[9,10]. The running time per genome was generally higher for *Prevotella* than *Methanobrevibacter,* which is expected since *Prevotella* has a slightly larger average genome size, meaning that the total number of genes to be processed is larger.

For running time per number of input genomes, CompareM2 outperformed both Tormes and Bactopia significantly. When analyzing 44 *Methanobrevibacter* MAGs, CompareM2 is 4.1 times faster than Bactopia and 7.2 times faster than Tormes. For 124 *Prevotella MAGs*, CompareM is 3.1 and 7.8 times faster than Bactopia and Tormes, respectively (**Table 2**).

Table 2*: Wall running time in seconds for analyzing 44 Methanobrevibacter MAGs or 124 Prevotella MAGs. "Factor" denotes how many times slower each tool is compared to the fastest. The fastest tool is marked with an underline (in both cases CompareM2). All numbers are means of three replicates.*

	44 Methanobrevibacter genomes		124 Prevotella genomes	
	minutes	factor	minutes	factor
CompareM2	6.6	1.0	29.7	1.0
Bactopia	27.0	4.1	91.0	3.1
Tormes	47.7	7.2	230.5	7.8

Discussion

CompareM2 is significantly faster than both Tormes and Bactopia as its running time scales much better with increasing input size. Notably, running time scaled approximately linearly with a small slope even when increasing the number of input genomes well beyond the number of available cores on the machine. The running time of each pipeline comes down to the time it takes to run each included tool on each sample, so differences between pipelines in terms of running time are determined by how they allocate resources and schedule jobs efficiently in parallel.

The speed of Bactopia is strongly affected by its reads-based approach: If reads are not input by the user – which was not possible in this case because we compared genomes that were assembled using a different pipeline – Bactopia creates artificial reads with ART[11]. This is done in order for Bactopia to be able to compare genomes without reads to genomes with reads. CompareM2 on the other hand is designed to compare genomes without reads and thus does not have to spend computing resources on producing these artificial reads. It should be noted that if the user runs more comparative analyses using the Bactopia Tools extensions, the scalability will be worse since the Bactopia platform does not offer to schedule running several tools in parallel. While Tormes does not suffer from producing artificial reads, it does fall short on not having a parallel workflow management system. As it runs all samples sequentially, running each tool at a time, it is not competitive on HPCs or multi-core CPUs. Generally, the running time standard deviations are negligible because the relative time differences are large. The running time was computed on a 64-core workstation (see Methods - Benchmarking). We ran the analysis by allocating 32 cores on this machine. By running with a lower number of cores than the machine has, we lower the chances that any other component than the CPU is the main bottleneck for computation.

Since both Tormes and Bactopia are designed for different use cases, they might not represent the perfect contenders for a comparison with CompareM2. Nonetheless, to our knowledge, they are the most comparable pipelines that exist today. In the case of Tormes, the comparison highlights the benefit of having a parallel rather than sequential job scheduling setup. In the case of Bactopia, it shows that other pipelines can approach the scalability of CompareM2 but also that having a reads-based approach is not competitive and that comparative analyses can be more integrated into the main pipeline. Also, we want to highlight that Bactopia and Tormes are not the only tools relevant for comparison. As CompareM2 sports many tools for advanced annotation, it also overlaps in use case with more annotation-focused pipelines like DRAM[12].

What is characteristic about CompareM2, is that it is assembly-agnostic: It works strictly downstream of assembling and binning. It is a general-purpose pipeline that doesn't discriminate between genomes based on how they were assembled. Many other tools also include all the steps necessary to turn raw reads into genome representatives and then do varying degrees of biological characterization of these, but raw read-dependent tools were deliberately left out of CompareM2. This is because read mapping, assembling, or even binning are highly dependent on the sequencing technology used and require a highly specialized pipeline for each technological use case. Next-generation sequencing has matured, and many competitive sequencing platforms exist (sequencing-by-synthesis, single molecule sequencing, etc.). Thus, designing a toolbox that can compare genomes is a very

different discipline from designing a toolbox that can assemble these genomes in the first place. Hard-linking two such pipelines together raises the concern that one part will not fit a specific use case. CompareM2 takes a different approach which is to offer a platform where you can compare your genomes regardless of how they were assembled.

Conclusions

CompareM2 offers an easy-to-install, user-friendly, and efficient genome annotation pipeline. It can be launched using a single command and is scalable to a range of projects, from the annotation of single genomes to comparisons across complex inventories. By using widely adopted genome tools, CompareM2 performs key annotation steps including genome quality control, predicted biological gene function, and taxonomic assignment. In addition, comparative analyses like computation of core- and pan-genomes or phylogenetic relations can be executed. We expect that CompareM2 will support the productivity of genome researchers by simplifying and expediting the annotation and comparison of genome-centric data. Further development of CompareM2 will continue with its ongoing adaptation to the community consensus of microbial ecologists. Through benchmarking, we have shown that CompareM2 is highly scalable, allowing analysis of large numbers of input genomes thanks to its underlying parallel job scheduling provided by Snakemake. Via CompareM2 we seek to accelerate and democratize the analysis of genomic assemblies for anyone who has computational resources available—be that on HPCs, a workstation, or even a laptop.

Methods

Implementation of CompareM2

Snakemake

This section sketches the technical implementation of CompareM2 using Snakemake. For more details on usage and modification of running parameters, please consult the documentation at: https://comparem2.readthedocs.io .

CompareM2 is built on top of Snakemake[13]. This means that much of the functionality in it is inherited by solutions within the Snakemake framework. A major upside of this is that CompareM2 can make use of Snakemake's extensive support for parallel job scheduling and support for running on workload managers used on HPCs. CompareM2 has been tested on Slurm[14] and PBS[15] workload managers. The CompareM2 executable, which is available

on Bioconda, is run by the user and does some basic housekeeping: It sets up the necessary environment variables like the base directory of the code, the Snakemake profile that defines whether the jobs are being submitted to an HPC workload manager and whether to use Conda or Docker/Apptainer as well as checking the database paths which are used by some of the tools. Then, the main Snakemake workflow is launched on the input genomes. By default CompareM2 accepts all fasta-formatted files available in the current working directory, but a specific path of “input_genomes” can also be set. In the case of analyzing large numbers of genomes, a “fofn” (file of file names) can also be used to define the input genomes. Snakemake allows for running only specific wanted parts of the rule graph (**Fig. 2**) and common parameters can be customized by the user. Because command line arguments are passed on from the CompareM2 executable to the main Snakemake workflow, the user has full access to control the main Snakemake workflow underneath. Regardless of whether the user uses the precompiled Docker image that contains environments for the included tools, the snakefile rules that govern how the specific tools are called can still be fully modified.

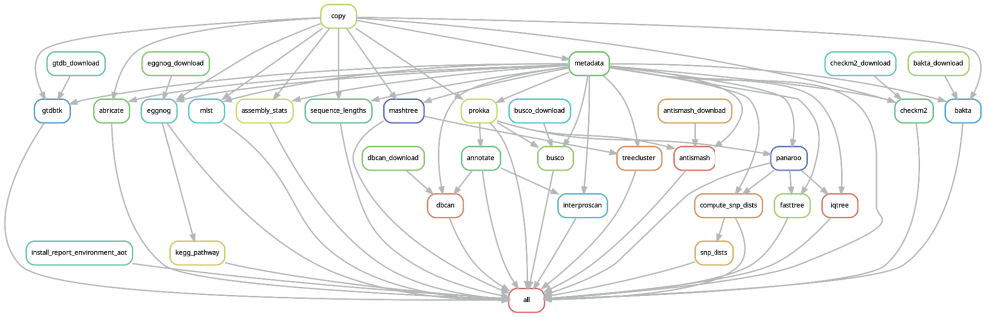


***Fig. 2**: **The underlying Snakemake workflow of CompareM2.** A directed acyclic graph (DAG) that shows the order in which to run rules that are dependent on each other. Each box represents a Snakemake rule, which is a code template that can be used to process a sample or a batch of samples. For instance, rule “prokka” is run on each input assembly, and then rule “panaroo” can analyze all samples using this output. The start and end points for the complete pipeline are represented by “copy” and “all”, respectively.*

When the main workflow is completed, the CompareM2 executable checks that relevant outputs exist, and in that case it calls the “dynamic report” sub pipeline. This pipeline is only in charge of producing the portable graphical dynamic report document that contains the main results of the main pipeline. When this report is rendered, the return code of the main pipeline is returned by the CompareM2 executable. One of the main features of this dynamic report is that it can produce a report from possibly incomplete results from the main pipeline. In many cases, a job will fail; for instance an advanced annotator like Antismash[16] will return

a non-zero exit code and not produce any output files if no genes are annotated in a genome. When this is the case, it is preferable that the rest of the main pipeline can continue and that the dynamic report pipeline can be run on the remaining results, showing the missing results. In a pipeline with this many moving parts, it is crucial that unaffected tools can continue running if something breaks down. This graphical report is based on the R Markdown[17] document rendering framework. Statistics and plots are generated with Tidyverse[18].

Launching the pipeline is a single command and the user is only required to have minimal experience with the command line interface for moving fasta genome files in and out of directories. The minimal system requirements are Linux with a Conda-compatible package manager. It is recommended that the system has Apptainer installed such that a Docker image containing all necessary binaries can be automatically downloaded and installed.

Docker and downloads

As CompareM2 uses many independent tools for analysis, it is not feasible to have all binaries in the same environment. The requirements for many different versions of the same dependencies would quickly lead to dependency hell[19]. Instead, each tool is installed into its own isolated environment. For a developmental installation of CompareM2, these isolated Conda environments are automatically installed using the “use-conda” functionality from Snakemake. In any other case when a user installs CompareM2 and has Apptainer (a high-performance Docker-compatible runtime[20]), the “use-apptainer” system is activated and a pre-compiled Docker[21] image is automatically downloaded and used instead. It contains a precompiled distribution of all Conda environments. Using this image is optional but highly recommended as it avoids potential dependency issues for the user to deal with during installation. Six of the bundled tools (Bakta, Busco, CheckM2, Dbcan, Eggnog, GTDB-Tk) need databases to run. These databases are automatically downloaded by individual rules in the Snakemake rule graph. CompareM2 only downloads these databases the first time a tool is used.

Tools included

Many tools are included in CompareM2. Here we list which they are, what they do and define the conditions in which they run.

The first step of the pipeline is to run all genomes through any2fasta[22] which acts as input validation and converts the input genome queries into a homogenized fasta format with a uniform character set.

Quality control is performed by assembly-stats[8] and seqkit[23] which both compute various basic genome statistics like genome length, count and lengths of contigs, N50, GC, etc. Busco[24] and CheckM2[25] are run to compute the completeness and contamination parameters of the input genomes.

The input genomes can be annotated with Prokka[26] or Bakta[27]. As both of these annotators produce results with a similar output structure, it is up to the user to decide which to use for downstream analysis. The default is Prokka, which is also displayed in **Fig. 2**.

Advanced annotation is carried out with the following tools with their briefly stated functions: Interproscan[28] scans protein signature databases like PFAM, TIGRFAM, and HAMAP. Dbcan[29] scans carbohydrate active enzymes (cazymes). Eggnog-mapper[30] provides orthology-based functional annotations. Gapseq[31] builds gapfilled genome scale metabolic models (GEMs). Antismash[16] finds biosynthetic gene clusters. Clusterprofiler[32] computes a pathway enrichment analysis. GTDB-Tk[7] uses an alignment of ubiquitous proteins to predict species names.

In a clinical setting, the following tools might be useful: Abricate[33] scans the NCBI[34], Card[35], Plasmidfinder[36], and VFDB[37] databases for antimicrobial resistance genes and virulence factors. MLST[38,39] calls multi-locus sequence types, is relevant for an initial grouping when tracking transmission and spread of bacteria.

In terms of phylogenetic analysis: Mashtree[40], which computes a neighbor-joined tree on the basis of mash distances. Treecluster[41] which based on customizable presets clusters the mashtree tree. Panaroo[9] produces a core genome suitable for phylogenetic analysis and defines a pangenome. This core genome is used by the following tools: Fasttree[42] computes a neighbor joined tree. IQ-TREE[43] computes a maximum-likelihood tree. Snp-dists[44] computes the pairwise snp-distances.

The CompareM2 code base with its 1k source lines of code (SLOC), installation instructions, and documentation are available in a Git repository currently hosted at GitHub: https://github.com/cmkobel/comparem2. The code is published under the GNU Public Licence version 3 which means that anyone who wishes to modify the software can do so if attributing the original authors. If users come up with concrete modifications or extensions, they are welcome to make a pull request on this repository.

Benchmarking

All running time analyses were run sequentially in random order on an AMD x86-64 “Threadripper Pro” 5995WX 64 cores, 8 memory channels, 512GiB DDR4 3200MHz ECC

(8x 64 GiB) and 4 2TB SSDs in raid0. All tests were run with three replicates. Electrical power used consisted of 89% Hydroelectric, 11% wind[45]. The running time was reported by the Snakemake benchmark function. We allowed each pipeline to use a maximum of 32 cores. Statistics and plots were generated using R[46] v4.3.1 and Tidyverse[18] v2.0.0. Scripts used to compute the benchmarking results are provided at https://github.com/cmkobel/cm2_benchmark .

Statistics of analyzed MAGs

Genome sizes were measured with assembly-stats v1.0.1. The MAGs were classified using GTDB v2.3 using database release 214. MAGs from both compared genera (*Methanobrevibacter* and *Prevotella*) were sourced from a project based on ONT R10.4 long read sequencing of the rumen content of 24 male *Bos taurus.* These genomes are accessible here: https://doi.org/10.6084/m9.figshare.26203361

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Bioconda package: https://anaconda.org/bioconda/comparem2

Complete code base repository: https://github.com/cmkobel/comparem2

Documentation: https://comparem2.readthedocs.io

Scripts for computing the benchmark results in this paper: https://github.com/cmkobel/ac2_benchmark

Competing interests

The authors declare that they have no competing interests.

Funding

This project is funded by The Novo Nordisk Foundation (Project no. 0054575-SuPAcow).

Authors' contributions

CMK conceptualized, designed, and developed the software. CMK wrote the initial version of the manuscript. VTEA, OØ, NNL, BJW, and PBP assisted in conceptualizing the software and writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We want to thank the initial testers and users of the software for feedback: Judith Guitart-Matas, Shashank Gupta, Wanxin Lai, and the 2023 and 2024 student batches of the “BIO326: Genome Sequencing; Tools and Analysis” course at Norwegian University of Life Sciences.

Author's information

Carl M. Kobel is a bioinformatician and a PhD candidate in the MEMO group at NMBU, Norway. Carl’s perspective is that microbiomes are largely undervalued and that we should better understand the minute interactions within them. Carl adopts a big data inspired approach, enjoys tinkering with hardware, and building parallelizable bioinformatics pipelines to gain insights into large microbiome datasets.

References

1. tseemann/nullarbor: :floppy_disk: ‘Reads to report’ for public health and clinical microbiology. https://github.com/tseemann/nullarbor?tab=readme-ov-file.
2. Quijada, N. M., Rodríguez-Lázaro, D., Eiros, J. M. & Hernández, M. TORMES: an automated pipeline for whole bacterial genome analysis. *Bioinformatics* **35**, 4207–4212 (2019).

3. Petit, R. A. & Read, T. D. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. *mSystems* **5**, 10.1128/msystems.00190-20 (2020).

4. Andersen, T. O. *et al.* Metabolic influence of core ciliates within the rumen microbiome. *ISME J.* **17**, 1128–1140 (2023).

5. Lazear, M. R. Sage: An Open-Source Tool for Fast Proteomics Searching and Quantification at Scale. *J. Proteome Res.* **22**, 3652–3659 (2023).

6. Nullarbor report without reads · Issue #249 · tseemann/nullarbor. https://github.com/tseemann/nullarbor/issues/249.

7. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

8. sanger-pathogens/assembly-stats. Pathogen Informatics, Wellcome Sanger Institute (2024).

9. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).

10. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

11. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).

12. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).

13. Mölder, F. *et al.* Sustainable data analysis with Snakemake. Preprint at https://doi.org/10.12688/f1000research.29032.2 (2021).

14. Jette, M. A. & Wickberg, T. Architecture of the Slurm Workload Manager. in *Job Scheduling Strategies for Parallel Processing* (eds. Klusáček, D., Corbalán, J. & Rodrigo, G. P.) 3–23 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/978-3-031-43943-8_1.

15. Henderson, R. L. Job scheduling under the Portable Batch System. in *Job Scheduling Strategies for Parallel Processing* (eds. Feitelson, D. G. & Rudolph, L.) 279–294 (Springer, Berlin, Heidelberg, 1995). doi:10.1007/3-540-60153-8_34.

16. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **51**, W46–W50 (2023).

17. Baumer, B. & Udwin, D. R Markdown. *WIREs Comput. Stat.* **7**, 167–177 (2015).

18. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

19. Fan, G. *et al.* Escaping dependency hell: finding build dependency errors with the unified dependency graph. in *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis* 463–474 (Association for Computing Machinery, New York, NY, USA, 2020). doi:10.1145/3395363.3397388.

20. Dykstra, D. Apptainer Without Setuid. *EPJ Web Conf.* **295**, 07005 (2024).

21. Miell, I. & Sayers, A. *Docker in Practice, Second Edition*. (Simon and Schuster, 2019).

22. Seemann, T. tseemann/any2fasta. (2024).

23. SeqKit2: A Swiss army knife for sequence and alignment processing - Shen - iMeta - Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/imt2.191.

24. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).

25. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).

26. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

27. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification | Microbiology Society. https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000685.

28. Zdobnov, E. M. & Apweiler, R. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).

29. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, W445–W451 (2012).

30. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

31. Zimmermann, J., Kaleta, C. & Waschina, S. gapseq: informed prediction of bacterial

metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol.* **22**, 81 (2021).

32. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

33. Seemann, T. tseemann/abricate. (2024).

34. Bacterial Antimicrobial Resistance Reference Gene ... (ID 313047) - BioProject - NCBI. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047.

35. Smith, K. W. *et al.* A standardized nomenclature for resistance-modifying agents in the Comprehensive Antibiotic Resistance Database. *Microbiol. Spectr.* **11**, e0274423 (2023).

36. Carattoli, A. & Hasman, H. PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). in *Horizontal Gene Transfer: Methods and Protocols* (ed. de la Cruz, F.) 285–294 (Springer US, New York, NY, 2020). doi:10.1007/978-1-4939-9877-7_20.

37. Liu, B., Zheng, D., Zhou, S., Chen, L. & Yang, J. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* **50**, D912–D917 (2022).

38. Jolley, K. A. & Maiden, M. C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595 (2010).

39. Seemann, T. tseemann/mlst. (2024).

40. Katz, L. S. *et al.* Mashtree: a rapid comparison of whole genome sequence files. *J. Open Source Softw.* **4**, 1762 (2019).

41. Balaban, M., Moshiri, N., Mai, U., Jia, X. & Mirarab, S. TreeCluster: Clustering biological sequences using phylogenetic trees. *PLOS ONE* **14**, e0221068 (2019).

42. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

43. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

44. Seemann, T. tseemann/snp-dists. (2024).

45. Kraftproduksjon. *ENERGIFAKTANORGE* https://energifaktanorge.no/norsk-energiforsyning/kraftforsyningen/.

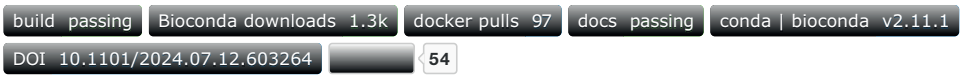
46. Wickham, H. & Grolemund, G. R for Data Science.

Supplementary information for paper #2

- Documentation hosted at https://comparem2.readthedocs.io
  Retrieved 2024-09-06

»Home

CompareM2



Note

If you're looking for the original version of CompareM, a tool to calculate AAI and codon usage, please follow this link: github.com/donovan-h-parks/CompareM

🧬CompareM2 is a genomes-to-report pipeline. It accepts prokaryotic (bacterial and archaeal) genomic assemblies and compares them in many different ways.

🦠Being designed to analyze assemblies of both isolates and metagenomes (MAGs), it is useful for anyone working with microbial genomics.

💾Installing CompareM2 on your system gives you access to many powerful state-of-the-art tools for analysis of prokaryotic genomes which will accelerate your research. It is easy to use and can be used by non-bioinformaticians.

👩‍🔬CompareM2 integrates several analyses that yield scientific results about genomic assemblies on several levels: Quality control, annotation, function and species calling as well as comparative analyses like computation of core/pan genomes and phylogenetics.

🐍CompareM2 works by calling a Snakemake workflow that can be easily modified to use different parameters for the underlying tools.

📄Central results are dynamically integrated in a compact portable report .html-document. It can be browsed in any web browser and can be easily shared as a single file. This report is generated even if some jobs in the pipeline fail. See examples.

👷CompareM2 can be run either on a local workstation (recommended >= 64GiB RAM), or a HPC (high performance computing) cluster. Both Apptainer/Singularity/Docker images and conda environment definitions are available for all dependent software to run.

If you have any questions, issues or ideas about using CompareM2, please raise an issue here.

The comprehensive documentation is available at CompareM2.readthedocs.io. And the code base is available at github.com/cmkobel/CompareM2.

Installation

Install the bioconda package

It is recommended that you have Apptainer on your system as it makes CompareM2 able to use a compressed Docker-image that speeds up installation significantly.

First, you need to install a Conda or Mamba package manager. The recommended choice is Miniforge which not only provides the required Python and Conda commands, but also includes Mamba - an extremely fast and robust replacement for the Conda package manager which is highly recommended.

Note

In case you don't use Miniforge you can always install Mamba into any other Conda-based Python distribution with: `conda install -n base -c conda-forge mamba`

Finally, CompareM2 can be installed into its own environment with the correct channels like so:

```
mamba create -c conda-forge -c bioconda -n comparem2 comparem2
```

Installing into isolated environments is best practice in order to avoid side effects with other packages.

Note

If you want to develop new rules in the CompareM2 pipeline, you should consider following the development version installation instructions. The development version contains the full git repository and is purely conda-based so you can affect the next version of the Apptainer-compatible Docker image.

Optionally: Testing the installation

Now you will be able to run CompareM2. You can use the example data in path "tests/MAGs" to check that everything works. The first time you run CompareM2 it will show the message "Pulling singularity image docker://cmkobel/comparem2." This might take some time depending on your network bandwidth as it downloads a +4GB Docker image that contains all the conda environments needed for each analysis.

```
# Activate the newly created conda environment containing the comparem2 launcher.
mamba activate comparem2

# First, create an empty directory and enter.
mkdir test_comparem2_install
cd test_comparem2_install

# Copy some test metagenomic assemblies from the test directory.
# cp $CONDA_PREFIX/share/comparem2-*/tests/E._faecium/*.fna . # Until v2.7.1 you must
unzip $CONDA_PREFIX/share/comparem2-*/tests/E._faecium/fna.zip # From 2.8.1

# Should take about a minute to complete the "fast" pseudo-rule.
comparem2 --until fast

# You can then investigate the report document that has been generated.
# open results_comparem2/report_test_comparem2_install.html

# Downloads all databases (~ 200 GB).
comparem2 --until downloads

# Run the full pipeline (~ 1 cpu-hour per genome).
comparem2
```

Advanced configuration

Shared database

If you are working on a shared computational resource like a laboratory workstation or a HPC you might want to share a database directory so that each user will not have to redundantly download each database. To set this up, the first user must decide on a directory and set reading and writing permissions for the group of users that should be able to use the database. Writing permissions are necessary for the "database representative" flags that snakemake uses to keep track of the presence of the databases. Setting this custom path is a matter of defining the "COMPAREM2_DATABASES" environment variable. You can put this into your ~/.bashrc or execute the command before using CompareM2.

```
export COMPAREM2_DATABASES="/absolute/path/to/shared_databases/comparem2_v2.5.8+"
```

HPC profiles for Snakemake

If you have experience with snakemake and are working on a high performance computing cluster (HPC), you can modify and use the cluster configuration profiles in the "profiles/" directory. You can define the use of one of these profiles by setting the "COMPAREM2_PROFILE" environment variable. You can put this into your ~/.bashrc or execute the command before using CompareM2. You can read more about snakemake profiles here or browse more default profiles here.

```
export COMPAREM2_PROFILE=${COMPAREM2_BASE}/profiles/apptainer/slurm-sigma2-saga
```

Usage

Overall, CompareM2 follows standard command line practices. CompareM2 is built on top of Snakemake. Hence, when tweaking your run, you must pass the parameters through the `--config` key. Although all Snakemake options are available to use, here we bring the ones that are necessary and useful for daily-driving CompareM2.

```
comparem2 [ --config KEY=VALUE [KEY2=VALUE]... ]
  [ --until RULE [RULE2]... ]
  [ --forcerun RULE [RULE2]... ]
  [ --printshellcmds ]
  [ --dry-run ]
  [ --version ] [ --help ] [ --cite ]
```

Usage examples

- Run *all* analyses across all fasta files in the current working directory.

  `comparem2`

- Run only jobs *until* prokka

  `comparem2 --until prokka`

- Run *all* analyses with specified input and output.

  `comparem2 --config input_genomes="path/to/genomes_*.fna" output_directory="my_analysis"`

- Use a *fofn* - a file of file names.

  `ls path/to/*.fna > my_fofn.txt; comparem2 --config fofn="my_fofn.txt"`

- Run a *dry run*.

  `comparem2 --config input_genomes="path/to/genomes_*.fna" --dry-run`

- Specify annotator. (default is "prokka")

  `comparem2 --config input_genomes="path/to/genomes_*.fna" annotator="bakta"`

- Run only the *fast* rules. (read more about pseudo rules)

```
comparem2 --config input_genomes="path/to/genomes_*.fna" annotator="bakta" --
until fast
```

- Run panaroo as well.

```
comparem2 --config input_genomes="path/to/genomes_*.fna" annotator="bakta" --
until fast panaroo
```

- And pass a command line argument directly to panaroo.

```
comparem2 --config input_genomes="path/to/genomes_*.fna" set_panaroo--
threshold=0.95 annotator="bakta" --until fast panaroo
```

Options

`--config KEY=VALUE [KEY2=VALUE]...`

Pass a parameter to the snakemake pipeline, where the following keys are available, defaults are stated as standard.

- `input_genomes="*.fna *.fa *.fasta *.fas"` Path to input genomes. As the default value indicates, all fasta type files in the present directory will be analyzed.
- `fofn="fofn.txt"` Deactivated by default. When set to a path it overrides key input_genomes. A fofn can be created with `ls *.fna > fofn.txt`
- `output_directory="results_comparem2"` All results are written here.
- `annotator="prokka"` Choice of annotation tool. Alternatively "bakta".

Passthrough arguments

From v2.8.2, CompareM2 has the ability to pass any command line argument (option-parameter pair) through to any rule in the workflow. This is done by using a generalized "passthrough argument" feature that recognizes config argument options starting with string "set_" and passes them to the correct rule upon generating the shell scripts for each rule in the workflow. The general syntax for these passthrough arguments is `set_<rule><option>=<parameter>` where rule is the rule name, option is the option key, and parameter is the parameter value.

Note

This feature requires modification of Snakemake such that it can accept special

characters through the config strings given at the command line. This modification can easily be done using the following command that ships with the bioconda package: `enable_passthrough_parameters_comparem2`

Otherwise you might receive the Snakemake error: "Invalid config definition: Config entry must start with a valid identifier."

An example can be used to explain how this feature can be used in practice: Consider using the Prokka annotator, which is capable of annotating both bacterial and archaeal genomes. By default, Prokka is set to bacterial annotation, so in case we want to annotate an archaea, we can set the "--kingdom" argument to "archaea". In this case the rule name is `prokka`, the option key is `--kingdom` and the parameter value is `archaea`. When using CompareM2, this setting can be set following the passthrough argument syntax like so:

```
# comparem2 --until set_<rule><key>=<value> # Syntax template.
comparem2 --config set_prokka--kingdom=archaea
```

Notice how the double dash prefix in "--kingdom" is part of the the set_ string. This is because many different styles of command line argument options need to be supported (e.g.: "--command_key", "--command-key", "-command_key" etc).

In some cases, command line options are flags, meaning that they need no parameter value. In this case, an empty string can be given as parameter value:

```
comparem2 --config set_prokka--rfam="" # --rfam enables searching for ncRNAs with Inf
```

In case of non-empty parameter values, use of apostrophes is optional.

Using a space separator, several command line arguments can be given at once for several different tools. In the following example we're also loosening the Panaroo core genome identity "--threshold" option down to 95% to increase the apparent number of genes in the core genome.

```
comparem2 --config set_prokka--kingdom=archaea set_panaroo--threshold=0.95 --until pa
```

CompareM2 comes with a number of sane default arguments which can be observed here. Any passthrough argument that the user gives on the command line overwrites these defaults.

Validating command line arguments

There are no limitations on which command line arguments can be passed to the passthrough argument feature. Thus, the user should follow the documentation of each individual tool to make sure that the command line arguments given are valid. In order to validate that the arguments given to rules are as expected, the full generated shell command of each rule can be printed with `-p`. It is especially useful to do this in conjunction with the `--dry-run` argument. Example below:

```
comparem2 --config set_panaroo--threshold=0.99 --until panaroo -p --dry-run
#> [...]
#>    panaroo \
#>        -o results_comparem2/panaroo \
#>        -t 16 \
#>        --clean-mode sensitive \
#>        --core_threshold 0.95 \
#>        --threshold 0.99 \
#> [...]
```

`--until RULE [RULE2]...`

Select to run up until and including a specific rule in the rule graph. Available rules: abricate annotate antismash assembly_stats bakta busco checkm2 copy dbcan eggnog fasttree gapseq gapseq_find gtdbtk interproscan iqtree kegg_pathway mashtree mlst prokka sequence_lengths snp_dists treecluster antismash_download bakta_download busco_download checkm2_download dbcan_download eggnog_download gtdb_download panaroo

There are also a number of pseudo rules, effectively "shortcuts" to a list of rules. - downloads (Run rules that download and setup up necessary databases.) - fast (Only rules that complete within a few seconds. Useful for testing.) - isolate (Only rules that are relevant for genomes of isolate origin.) - meta (Only rules that are relevant for genomes "MAGs" of metagenomic origin.) - report (Re-renders the report.)

`--forcerun RULE [RULE2]...`

Force rerunning of one or more rules that already have been completed. This is generally necessary when changing running parameters in the config (see "--config" above).

`--printshellcmds`, **`-p`**

Print the full generated shell commands of each rule in the workflow.

`--dry-run`

Run a "dry run": Shows what will run without doing it.

`--version`, `-v`

Show current version.

`--help`, `-h`

Show this help and exit.

Environment variables

No environment variables are strictly necessary to set, but the following might be useful:

- `COMPAREM2_PROFILE` (default "profile/apptainer/local") specifies which of the Snakemake profiles to use. This can be useful for running CompareM2 on a HPC or using specific settings on a large workstation. Check out the bundled profiles in path profile/* (possibly in $CONDA_PREFIX/comparem2/profile/*).
- `COMPAREM2_DATABASES` (default "databases/") specifies a database location. Useful when sharing a database installation between various users on the same workstation or HPC.

Output

CompareM2 creates a directory named "results_comparem2/" (or what the output_directory parameter is set to) that contains all of the analysis results that are computed.

Results from input genomes are in dir "samples/" and results across all samples are in the root.

The report is named "report_<title>.html" after the title of the run which defaults to the name of the current working directory.

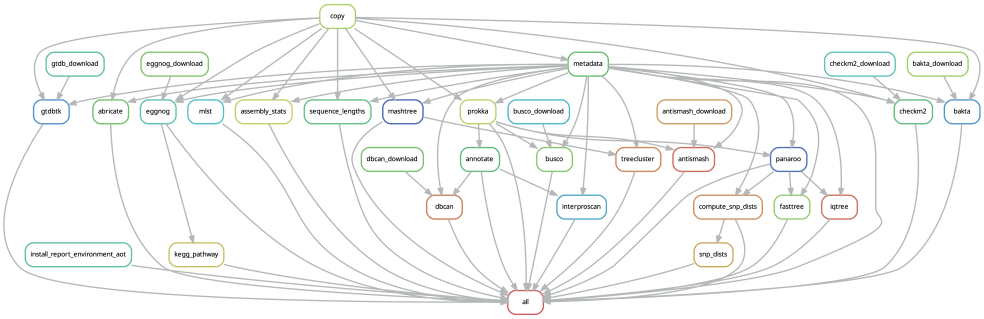
```
results_comparem2/
├── abricate/
├── assembly-stats/
├── benchmarks/
├── checkm2/
├── fasttree/
├── gtdbtk/
├── iqtree/
├── kegg_pathway/
├── mashtree/
├── metadata.tsv
├── mlst/
├── panaroo/
├── report_<title>.html
├── samples/
│   └── <sample>/
│       ├── antismash/
│       ├── bakta/
│       ├── busco/
│       ├── dbcan/
│       ├── eggnog/
│       ├── <sample>.fna
│       ├── interproscan/
│       ├── prokka/
│       └── sequence_lengths/
├── snp-dists/
├── tables/
├── treecluster/
└── version_info.txt
```

For the file tree of each of the analysis tools, please consult the respective documentation.

CompareM2 genomes-to-report pipeline. Copyright (C) 2024 contributors GNU GPL v3.

What analyses does it do?

Below is the graph the shows the order of execution of all possible analyses "rules" in CompareM2:



This figure does not show the pseudo rules such as `meta`, `isolate`, `fast`, etc.

Hint

Use `comparem2 --until <rule> [<another rule>...]` to run one or several specific analyses only. The rule names for each analysis to pick is listed below:

For each sample

First, independent analyses are run on each of the input genomic assembly files.

- `sequence_lengths` seqkit Lengths and GC-content of individual contigs.
- `assembly_stats` assembly-stats Generic assembly statistics.
- `busco` BUSCO Estimate assembly completeness and contamination.
- `checkm2` CheckM2 Estimate assembly completeness and contamination.
- `prokka` prokka Genomic annotation of Archaea and Bacteria.
- `bakta` bakta Genomic annotation of Bacteria (lacking in report, but used downstream by other tools).
- `kegg_pathway` clusterProfiler KEGG ortholog-based pathway enrichment analysis.
- `dbcan` dbCAN4 Annotation of carbohydrate-active "CAZyme" enzymes (lacking in

report).
- `antismash` antismash Detection of biosynthesis genes (lacking in report).
- `eggnog` eggnog-mapper Functional annotation.
- `interproscan` InterProScan Protein function using Tigrfam, Hamap and Pfam (lacking in report).
- `abricate` abricate Virulence and resistance gene identification.
- `mlst` mlst Multi locus sequence typing.
- `gtdbtk` GTDB-tk Species recognition.

Across samples

Then on the basis of the analysis of each input genomic assembly, these analyses are run across all samples.

- `panaroo` panaroo Pan and core genome.
- `snp_dists` snp-dists Core genome pairwise snp-distances.
- `fasttree` FastTree Phylogenetic tree of the core genome.
- `iqtree` IQ-tree Phylogenetic Tree of core genome with bootstrapping (lacking in report).
- `mashtree` Mashtree Super fast distance measurement
- `treecluster` TreeCluster Clustering of phylogenetic trees (lacking in report).
- **A nice report easy to share with your friends (See demos below)**

Pseudo-rules

There are also a few pseudo targets defined. For instance `fast` which runs sequence_lengths, assembly-stats and mashtree. There is also one named `isolate` which runs all the analyses that are relevant for clinical isolates (sequence_lengths, prokka, mlst, abricate, assembly-stats, gtdbtk, busco, checkm2, roary, snp-dists, fasttree, mashtree) as well as one named `meta` which runs the analyses that are relevant to metagenomes (aka. MAGs), these are sequence_lengths, prokka, gtdbtk, busco, checkm2, mashtree.

Hint

You can run one of these pseudorules just like any other rulename with `comparem2 --until meta` or `comparem2 --until isolate`

Rendered report

These demo reports are available for inspiration while you wait for your own report to complete.

- report_strachan_campylo.html

  32 Campylobacter genomes, Metagenome and genome sequencing from the rumen epithelial wall of dairy cattle. From Nature 2022 - Strachan et al. (doi.org/10.1038/s41564-022-01300-y).

- report_Methanoflorens.html

  6 Methanoflorens (archaeal) genomes. Representants of Bog-38 which are part of GTDB.

Future functionality

In the future we might add some of the following software packages into CompareM2. This document serves as a backlog of tools that we want to integrate when time allows.

Assembly basis (within each sample)

- AlphaFold Neural network protein folding prediction genome annotation.
- Integration of the DRAM databases for easier metabolic interpretation.
- Oriloc Identification of possible replication origins of chromids.
- RFplasmid Identification of plasmids using the pentamer-random-forest method.
- Kaptive Identification of surface polysaccharide loci for Klebsiella and Acinetobacter baumannii.
- AMRFinderPlus Identification of AMR genes and their point mutations.
- gapseq GEMs, pathway completeness and much more.

Batch basis (across all samples)

- GC3-profiling "fingerprinting" of the distribution of GC-content.
- Recombination in core genome using the Bruen's PHI statistic or ClonalFrameML.
- Identification of horizontally transferred genes?

Please add an issue on the repository if you have any ideas or requests.

Contributors

Contributors of the CompareM2 code base

- Carl Mathias Kobel (@cmkobel)
- Oliver Kjærlund Hansen (@OliverKjHansen)
- Ben J. Woodcroft (@wwood)

Citing and alternatives

Citing CompareM2

If you use CompareM2, you can support further funding by bumping the citation count on this one:

- Kobel, C. M. et al. CompareM2 is a genomes-to-report pipeline for comparing microbial genomes. 2024.07.12.603264 Preprint at https://doi.org/10.1101/2024.07.12.603264 (2024).

References for the included tools

CompareM2 would not have existed, if it hadn't been for the integrated software packages and their databases. Please reach out to carl.mathias.kobel near nmbu.no if you think something is missing.

- Bacterial Antimicrobial Resistance Reference Gene ... (ID 313047) - BioProject - NCBI. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047.
- Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification | Microbiology Society. https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000685.
- Balaban, M., Moshiri, N., Mai, U., Jia, X. & Mirarab, S. TreeCluster: Clustering biological sequences using phylogenetic trees. PLOS ONE 14, e0221068 (2019).
- Baumer, B. & Udwin, D. R Markdown. WIREs Comput. Stat. 7, 167–177 (2015).
- Blin, K. et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Res. 51, W46–W50 (2023).
- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol. Biol. Evol. 38, 5825–5829 (2021).
- Carattoli, A. & Hasman, H. PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). in Horizontal Gene Transfer: Methods and Protocols (ed. de la Cruz, F.) 285–294 (Springer US, New York, NY, 2020). doi:10.1007/978-1-4939-9877-7_20.

- Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics 38, 5315–5316 (2022).
- Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat. Methods 20, 1203–1212 (2023).
- Dykstra, D. Apptainer Without Setuid. EPJ Web Conf. 295, 07005 (2024).
- Jette, M. A. & Wickberg, T. Architecture of the Slurm Workload Manager. in Job Scheduling Strategies for Parallel Processing (eds. Klusáček, D., Corbalán, J. & Rodrigo, G. P.) 3–23 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/978-3-031-43943-8_1.
- Jolley, K. A. & Maiden, M. C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11, 595 (2010).
- Katz, L. S. et al. Mashtree: a rapid comparison of whole genome sequence files. J. Open Source Softw. 4, 1762 (2019).
- Liu, B., Zheng, D., Zhou, S., Chen, L. & Yang, J. VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res. 50, D912–D917 (2022).
- Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol. 38, 4647–4654 (2021).
- Miell, I. & Sayers, A. Docker in Practice, Second Edition. (Simon and Schuster, 2019).
- Minh, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534 (2020).
- Mölder, F. et al. Sustainable data analysis with Snakemake. Preprint at https://doi.org/10.12688/f1000research.29032.2 (2021).
- Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693 (2015).
- Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 5, e9490 (2010).
- sanger-pathogens/assembly-stats. Pathogen Informatics, Wellcome Sanger Institute (2024).
- Seemann, T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069 (2014).
- Seemann, T. tseemann/abricate. (2024).
- Seemann, T. tseemann/any2fasta. (2024).
- Seemann, T. tseemann/mlst. (2024).
- Seemann, T. tseemann/snp-dists. (2024).
- SeqKit2: A Swiss army knife for sequence and alignment processing - Shen - iMeta -

Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/imt2.191.
- Smith, K. W. et al. A standardized nomenclature for resistance-modifying agents in the Comprehensive Antibiotic Resistance Database. Microbiol. Spectr. 11, e0274423 (2023).
- Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 21, 180 (2020).
- Wickham, H. et al. Welcome to the Tidyverse. J. Open Source Softw. 4, 1686 (2019).
- Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2, 100141 (2021).
- Yin, Y. et al. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 40, W445–W451 (2012).
- Zdobnov, E. M. & Apweiler, R. InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847–848 (2001).
- Zimmermann, J., Kaleta, C. & Waschina, S. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. Genome Biol. 22, 81 (2021).

Alternative tools

What is unique about CompareM2 is that it works strictly downstream of assembling and binning. Many other tools also include all the steps necessary to turn raw reads into genome representatives, and then does varying degrees of biological characterization of these freshly created assemblies/bins/genomes. It is a conscious decision to exclude the raw read-dependent tools out of the equation for CompareM2. This is because read-mapping, assembling or even binning is highly dependent on the sequencing technology used and requires a highly specialized pipeline for each technological use case. CompareM2 takes a different approach which is to offer a portable and flexible platform where you can easily compare your genomes, no matter where they came from, regardless of the sequencing technology used to create them in the first place. Genome quality is only increasing and in the future we will not have to be worried when comparing pyrosequencing and single-molecule sequencing or hybrid approach based genomes in a single batch of CompareM2.

Below we are listing some competing pipelines that partly overlap with the use cases of CompareM2. Sorted alphabetically.

- Anvi'o
- ASA^{3}P
- Aviary

- Bactopia
- DRAM
- Nullarbor
- Tormes
- VEBA

Paper #3

2024, Molecular Omics, Royal Society of Chemistry

doi: 10.1039/D4MO00017J

Molecular Omics

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue

Cite this: *Mol. Omics*, 2024, **20**, 438

Integrating host and microbiome biology using holo-omics

Carl M. Kobel, [a] Jenny Merkesvik, [b] Idun Maria Tokvam Burgos, [c] Wanxin Lai, [b] Ove Øyås, [a] Phillip B. Pope, [abd] Torgeir R. Hvidsten [b] and Velma T. E. Aho *[a]

Holo-omics is the use of omics data to study a host and its inherent microbiomes – a biological system known as a "holobiont". A microbiome that exists in such a space often encounters habitat stability and in return provides metabolic capacities that can benefit their host. Here we present an overview of beneficial host–microbiome systems and propose and discuss several methodological frameworks that can be used to investigate the intricacies of the many as yet undefined host–microbiome interactions that influence holobiont homeostasis. While this is an emerging field, we anticipate that ongoing methodological advancements will enhance the biological resolution that is necessary to improve our understanding of host–microbiome interplay to make meaningful interpretations and biotechnological applications.

Received 1st February 2024,
Accepted 10th June 2024

DOI: 10.1039/d4mo00017j

rsc.li/molomics

Introduction

Overview and potential of holo-omics

In many biological systems and environments, both the host and its resident microbiomes are considered as important contributors to the total function of the overall system.[1] To study the biology of a living system, scientists regularly perform "omics" analyses of the various biomacromolecules that constitute a living cell, such as DNA, RNA, proteins, and metabolites (Table 1, Fig. 1).[2] Analysis of these different "layers" can be further empowered when performed in combination, an approach referred to as "multi-omics". The concept of "holo-omics" represents an additional thematic shift whereby instead of focusing on molecular details of an individual organism or an isolated reaction in a specific environment, we can consider the biology of a host–microbiome system as a single unit of action. This makes it possible to understand overarching phenomena in the holobiont.[3] In this context, the objective of

[a] *Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway. E-mail: velma.tea.essi.aho@nmbu.no*
[b] *Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway*
[c] *Faculty of Natural Sciences, Norwegian University of Science and Technology, Trondheim, Norway*
[d] *Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology (QUT), Translational Research Institute, Woolloongabba, Queensland, Australia*

Carl M. Kobel

Carl is a bioinformatician and a PhD candidate in the MEMO group at NMBU, Norway. Carl's perspective is that microbiomes are largely undervalued and that we should better understand the minute interactions within them. Carl adopts a big data inspired approach, enjoys tinkering with hardware, and building parallelizable bioinformatics pipelines to gain insights into large microbiome datasets.

Jenny Merkesvik

Jenny is a PhD candidate in the Bioinformatics and Applied Statistics group at NMBU. She is part of 3D'omics, a European Union Horizon 2020 project in which she contributes to increase our understanding of host-microbiome interplay through holo-omics. Her work is motivated by the aim of improving animal and feed production, benefitting both the animals and the growing human population, in a sustainable way.

Table 1 Glossary

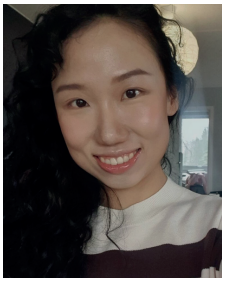
Term	Definition
Habitat	A defined ecological niche that provides environmental parameters that supports a set of organisms.
Holo-	From Ancient Greek ὅλος: hólos, "whole".
Holo-omics	Research that analyses one or more functional layers of omics data from both host and microbiome. The terms holo-omics and hologenomics might be used interchangeably because most omics layers arise from genomic DNA.
Holobiont	An ecological unit consisting of a host and its resident, interacting micro-organisms.
Host–microbiome interface	Any surface where biological features from either host or microbiome can interact.
Integrative analysis	Overlapping or relating the biological factors between two molecular layers or host–microbiome sources.
Metagenomics	Techniques used to study the collective genomic reads from all organisms in an ecological niche.
Multi-omics	Research covering more than one omics layer representing one or multiple interacting organisms. Examples of the former include human multi-omics with measurements that only reflect human biology; and microbial multi-omics without taking the host into account.
Omics	The study of all biomolecules of a specific type. This review focuses on functional omics data, which can be defined as omics data that change over time and across conditions.
Proteomics	Using a bespoke database which is based on *in silico* translation of the genomic sequences, to match mass spectrometric spectra to measure the abundance of proteins in a sample.
Transcriptomics	Techniques used to study an organism's transcriptome, *i.e.* the sum of all of its RNA transcripts.
Untargeted metabolomics	Using methods such as mass spectrometry (MS) or nuclear magnetic resonance (NMR) to measure the abundance of all the metabolites in a sample.

holo-omics is to study biomacromolecules that constitute biological interactions between a host and its microbiome (Fig. 1).

Acquiring a dataset to study host–microbiome interactions is a matter of applying various omics technologies to measure

Idun Maria Tokvam Burgos

Idun Burgos is a PhD candidate in the Systems Biology group at the Norwegian University of Science and Technology. She holds a master's degree in chemical engineering with a specialization in systems biology from the same university. Her PhD project entails studying bacterial communities through genome-scale metabolic networks, applying methods from bioinformatics, biochemical engineering, and systems biology.



Wanxin Lai

Wanxin Lai is a PhD candidate affiliated with both the Bioinformatics and Applied Statistics group and the Faculty of Chemistry, Biotechnology and Food Science at the Norwegian University of Life Sciences (NMBU). Her research focuses on integrating multi omics data through network-based approaches to uncover underlying molecular mechanisms and improve the prediction of phenotypes of interest.

Ove Øyås

Ove Øyås is a researcher in the Microbial Ecology and Meta-Omics group at the Norwegian University of Life Sciences. In 2019, he obtained a PhD in computational systems biology from ETH Zurich, where he developed a passion for understanding biology through model-based integration of omics data. Currently, Dr Øyås is working on multi-omics data analysis and modeling of the rumen gut microbiome as part of the EU-funded HoloRuminant project. Most of his research involves development of scalable computational methods that make it possible to answer new biological questions.

Velma T. E. Aho

Velma Aho, PhD, has been involved in microbiome research since 2013, starting with amplicon sequencing studies and progressing towards increasingly complex multi- and holo-omic projects, constantly striving for a deeper understanding of the roles of microbiomes in mammalian hosts. After ten years of focus on gut microbiota in Parkinson's disease at the universities of Helsinki and Luxembourg, Dr Aho is currently exploring the microbial community of the cattle rumen as part of the Microbial Ecology and Meta-Omics group at the Norwegian University of Life Sciences.

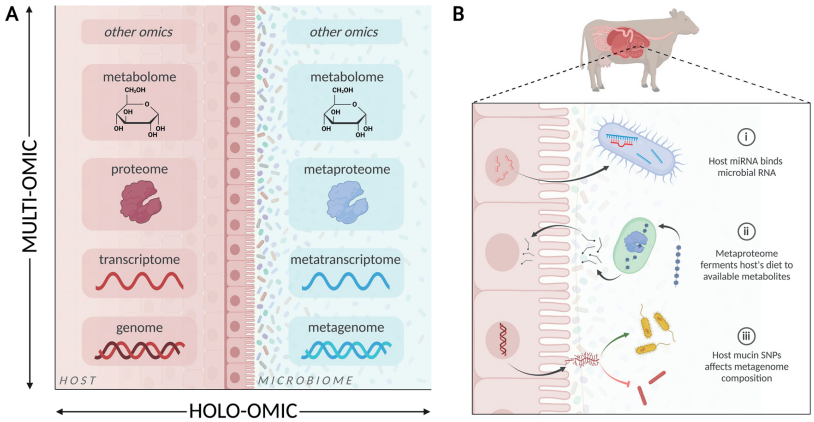


Fig. 1 Holo-omics is a specialised case of multi-omics where biological features are linked across a host–microbiome interface. (A) This interface is idealised along the horizontal axis labelled "holo-omic" as an epithelium with a large surface area where biochemical compounds can be exchanged in both directions. The vertical axis labelled "multi-omic" highlights that interactions can occur on multiple levels in terms of coding sequences and biochemical compounds. (B) Examples of molecular interactions across a host–microbiome interface.[4–6] Created with biorender.com.

the molecular features of both sides of the holobiont. While this data acquisition used to be the limiting step in such analyses, modern molecular biology tools are making this process more efficient and economical. Today's primary technical bottlenecks are (1) overcoming microbial community complexity, which can contain thousands of different genomes compared to their singular defined host, and (2) the computational analysis of holo-omic data so that the biological processes of both the host and its microbiome can be integrated computationally, interpreted, and visualised.[7] For example, performing data integration across the host–microbiome interface requires correlating individual biological features across various omics layers, which often cannot be scaled to the typical size of holo-omic datasets and can also suffer due to insufficient statistical power. To meet this challenge a new family of computational tools is needed: they must be able to cluster biological features into modules and cross-correlate features across the host–microbiome boundary, capturing the signals that represent the hypothesised cooperation between the host and its microbiome.

Symbiotic interactions in host-associated microbiomes are generally defined by the mutualistic, commensalistic or neutral effects shared between each organism, which depend on whether the benefit involved is one-way, two-way or lacking, respectively.[8] Additionally, there is a spectrum of harmful-neutral interactions within the microbiome and between certain microorganisms, and opportunistic pathogens, viruses, and phages might play a role in defining the dynamics of the microbiome.[9] Additional layers of intra-microbiome complexity should also be considered, particularly for the existence of networks of symbioses within a given microbiome which can be characterised in isolation, as with any other microbial environment. What distinguishes holo-omics is that the host variation is integrated together with any intra-microbiome relationships. Subsequently, holo-omics makes it possible to understand the intra-microbiome dynamics where a host-directed interaction is imposed on the microbiome.

For this review, we discuss in detail how actual holo-omic analyses can be performed computationally and present several frameworks to take the typically massive and complex holo-omic datasets and integrate the signal between the host and its microbiome. We consider host–microbiome studies where the host is a multicellular organism like an animal, fungus, or plant that forms a large surface or boundary from which it can interact with the microbiome that typically consists of a community of single-celled microorganisms (bacteria, archaea, eukaryotes) and possibly viruses, with varying degrees of diversity (Table 2). For simplicity, we do not consider parasite interactions in this review but focus on the beneficial interactions in holobionts.

Many known hosts are obligate symbionts, meaning the host is non-viable when the microbiome is absent. One example of an obligate holobiont is lichen, where a fungus and a community of cyanobacteria represent a complete holobiont. The fungus provides physical anchoring and nutrient assimilation whereas the cyanobacteria provide carbohydrates assimilated through photosynthesis. Additionally, these holobionts may house Alphaproteobacteria which work in conjunction to fix nitrogen for the lichen, which may otherwise be nutrient-limited.[30] On the other end of the spectrum of dependency are several types of insects, such as ants and caterpillars, which harbour few or no resident microorganisms that are unlikely to have a large impact on fitness.[31] Mammalian hosts tend to fall between these two extreme examples: they are viable when

Table 2 Selected examples of host–microbiome systems and their characteristics in terms of symbiotic benefit, dependency, species richness, and services exchanged between host and microbiome. These definitions depend on the ecological circumstances in which each host–microbiome system was considered

Holobiont system	Symbiosis	Microbiome richness	Host → microbiome services	Microbiome → host services
Cattle rumen	Mutualistic[8,10]	8500–16 994 prokaryotic species,[11,12] 52 alveolata,[13] 12 fungi[14]	Habitat, substrates[15]	Catabolism of complex plant fibres,[15] anabolism of essential chemicals
Mouse gut	Mutualistic[16]	828–1573 species[17,18]	Habitat, substrates	Catabolism of feed matter, anabolism of essential chemicals[16,19]
Salmon gut	Commensalistic	30–40 species (prokaryotes)[20]	Habitat, substrates	Unknown
Plant root-soil	Mutualistic, commensalistic[21]	2799–271 940 species[22,23]	Energy (sugars, fibres)[24]	Nutrients, nitrogen,[25] stress resistance[21]
Bee gut	Mutualistic	<10 species[26–28]	Habitat, substrates	Modulate social behaviour,[29] catabolism of carbohydrates[28]

raised in a germ-free setting, but experimental results suggest various abnormalities in such animals, ranging from changes in the immune system to altered neurodevelopment and behavior.[32,33]

Host–microbiome orchestration

The holobiont represents an evolutionary shortcut where the host and microbiome partners together orchestrate a metabolic capacity[34,35] that otherwise would have had to develop using horizontal gene transfer and recombination *via* sexual reproduction within the genome of the host organism itself.[36] In the holobiont perspective, the host provides a habitat for its associated microbiomes with defined and stable ecological factors, such as the presence or gradients of substrates and environmental factors like oxygen, temperature, and H^+ concentration.[37] In return, the microbiome provides complex biochemicals that the host otherwise would not have been able to synthesise or assimilate. In this context, host-directed internal environmental factors provide the selective pressure that defines which microorganisms are ultimately present.[38] However, many microorganisms, mainly prokaryotes, utilise promiscuous mechanisms for horizontal gene transfer. This gives them the ability to collect mobile and novel genetic elements from diverse sources such as viruses, and alternative genealogies across domains of life.[39] Mechanisms that enable the rapid evolution of microorganisms facilitate their competitive metabolic potential to assimilate both energy and nutrients from a spectrum of ecological niches. A host that has co-evolved with its microbiome can leverage its microbiome-based metabolic potential flexibility to adapt and thrive in niches that the host would have been unlikely to enter on its own.

The microbiomes of holobionts are per definition not mediated through the somatic genome of the host which means that the microbiota must have its own way of transmitting genetic material to offspring or between individuals in a population. This means that the composition of species present in a microbiome is subject to change over time as new species colonise and take over functions of others.[40] Host–microbiome co-evolution and adaptation is possible when new microbiota become part of the holobiont in a population of hosts, and are inherited vertically to offspring or between individuals in a host population. This can give rise to endemic microbiota species which are exclusively found as part of a holobiont. The microorganisms can adapt to their host and thus diverge from their ancestral population. Hosts and microbiota are able to co-adapt evolutionarily which means that they can each specialise and optimise their function in the holobiont system over generations.[41,42]

Idealised biological frameworks of holo-omic models

Studying holobiont systems using holo-omics generally requires a statistical or mechanistic framework that can capture signals or patterns in the data to infer interacting biomacromolecules or biological features across the host and its microbiome.[43] When analysing holo-omic data, its size and complexity usually means it must first be constrained by dimensionality reduction or compression, or by clustering into modular groups of co-abundant biological features. This is to make the computational analysis tractable and to simplify the interpretation of its function. Therefore, it is necessary that the methodological framework chosen to perform this data constraining is able to capture the hypothesised interaction between the host and microbiome.

Most frameworks are statistical in the sense that they test whether there are significant differences between treatments or co-appearing groups, but suitable mechanistic models are increasingly available and used for data integration as well.[43] To integrate omics data, these mechanistic models should ideally account for the dynamics of all relevant genome-scale networks in the holobiont system, but scaling to systems of this size entails major computational challenges for dynamic models in particular.[44] Because of this, mechanistic omics integration studies have mainly used genome-scale metabolic models (GEMs), which capture the steady-state flows of metabolites through an organism's network of biochemical reactions[45] and are available for a range of hosts and microorganisms.[46] By linking metabolic flows to interactions between host and microbiome, GEMs integrated with holo-omics can allow mechanistic investigation of holobiont systems. Dynamic modelling

of genome-scale interaction networks is also becoming feasible thanks to algorithmic and computational advances,[47] but most of the methods that we will discuss here take a statistical approach where they compare and compute significance between groups.

Examples of recent publications with a holo-omics approach

Since the rise of modern molecular biology tools that have facilitated holo-omic analyses, the number of publications focusing on host–microbiome interactions has been growing. For the purposes of this review, we are particularly interested in studies that include an integrative analysis of two or more omic datasets and discuss both the host and its associated microbiome.

Recent holo-omic research articles provide examples of the different types of questions that can be approached from a holo-omics point of view, ranging from experiments with model organisms to comparative evolutionary studies. In the classic experimental end of the spectrum, two studies used a mouse model to address two ''epidemics'' faced by human medicine: opioid overuse[48] and obesity.[49] Both studies included host transcriptomics, microbial shotgun metagenomics, and untargeted metabolomics, the latter capturing a mix of molecules produced by the host and the microbiome. Their results suggested that the tested medications – morphine in the opioid study, the antidiabetic drug empagliflozin in the obesity study – had effects on the host and microbial layers.[48,49] Both studies further confirmed that there are correlations between different omic layers, offering the simplest kind of evidence for host–microbiome interactions. The opioid study also tested this experimentally by showing that morphine-induced changes in host gene expression vary depending on the presence of a microbiome.[48]

In an example closer to traditional ecology, a study focusing on the gut of the termite *Labiotermes labralis* used metagenomics, metatranscriptomics, and host transcriptomics data to demonstrate that the host and the microbiome provide complementary sets of carbohydrate-active enzymes, enabling the holobiont to degrade a wide range of soil polysaccharides.[50] Finally, a study taking a holo-omics approach to evolution compared several ant- and termite-eating mammals, with findings that supported convergent evolution not only in host genomes, but also in microbiomes.[51] Specifically, the gut metagenomes of these mammals were enriched in enzymes that are necessary for subsisting on an insectivorous diet, such as chitinases and trehalases, compared to mammals with other types of diets.

While the existing publications showcase the exciting opportunities offered by holo-omic research, many of them include only one omics layer for each side of the holobiont. Comprehensive, multi-layered integrative studies remain rare, partly due to financial limitations, but also to the challenges presented by bioinformatic and statistical analyses.

State of the art in integrative models

Considerations for holo-omics tools

Although the cost of generating omics data has come down considerably in recent years, it is still a major undertaking to run controlled animal experiments to obtain matching samples from hosts and microbiomes. As a consequence, holo-omic studies typically tend to have small sample sizes. At the same time, the number of measured biological features (genes, proteins, metabolites) may reach millions, considering that complex microbial communities contain hundreds of species.

Let us consider a hypothetical holo-omic study, where we have measured the host transcriptome of the liver in 100 cows (n = 100) and the meta-transcriptome of the rumen content in those same individuals (p = 20 000 host genes + average 3000 microbial genes × 200 microbial species = 620 000 features). Let us further assume that the experiment is set up to measure methane emission, and that half of the cows were given a methane-inhibiting feed additive (treatment) that indeed reduced emissions. This dataset would pose a massive challenge for data analysis, and not primarily because it would require considerable computational resources to assemble and annotate Metagenome Assembled Genomes (MAGs) and estimate expression (read mapping). The main challenge is related to the large number of features compared to samples. Naively one would think that this dataset could be analysed using multivariate- or machine learning-based prediction methods, where the predictive model could be queried for features or combination of features that contributed significantly to the prediction; ''IF gene G on MAG5 is up AND host gene H is down THEN low methane''. However, with this many features there will be an enormous number of feature combinations that could separate low and high emitting cows, and with only 100 examples (cows) to constrain them, we would never be able to discern real biological feature-combinations from spurious ones (Fig. 2). This phenomenon is referred to as overfitting and is a consequence of the curse of dimensionality: the number of

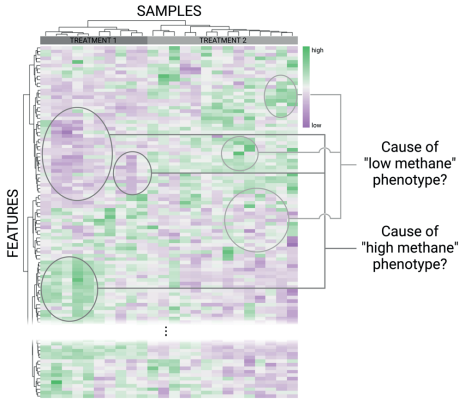


Fig. 2 Illustrating a common problem in multi-omics and holo-omics where a low number of samples with a high number of features are linked into a low number of traits (methane). The underlying data is arbitrary and represents a single omics layer. Created with biorender.com.

examples (cows) needed to identify the biologically meaningful features grows exponentially with the number of features.

Methods that divide the aforementioned examples into training and test sets, such as cross validation, would be able to tell us that we are overfitting, but will not be able to solve the problem. Even testing one feature at a time is problematic, since multiple hypothesis testing would severely limit the statistical power and thus only identify features with very large and consistent differences (*i.e.* large effect sizes) between the two treatments. Luckily, omics features are by no means independent and can be grouped into modules of co-abundant genes, proteins, or metabolites, for instance by correlation. This and other so-called dimensionality reduction approaches typically result in a few dozen distinct modules that can be used as our new features to reveal connections to methane emission and also to hypothesise putative interactions between host and microbiome. A note of caution here is that methods for module finding that rely on computing a distance matrix would require extreme amounts of memory. An approach used for instance by weighted gene co-expression network analysis (WGCNA, a method discussed later in this review) is to first group the data into ''blocks'' using k-means clustering, find modules in each group, and then combine similar modules at the end.

Integrating several omics datasets for a multi-omics approach can help us hone in on biologically meaningful patterns, if done carefully. Assuming that we added metabolomics data to the aforementioned cow example; simply concatenating the transcriptomics and metabolomics table would leave us with even more features (number of genes + number of metabolites). Instead, one could first identify genes and metabolites that are differentially abundant between ''low'' and ''high'' methane-emitting cows, and then select pathways that are enriched in both differential genes and differential metabolites. Such consensus integration methods use information about multiple types of molecules to constrain the number of possible biological interpretations.

Although there are strong functional interdependencies between rumen microbes converting feed into fatty acids and the host animal metabolizing fatty acids in the liver to produce energy, there are also clear physical boundaries separating these features, meaning that we should consider omic data origins in our holo-omic analysis design. In the case of pathway analysis, for example, one needs to consider that a pathway operates within the confines of a cell of a single organism. More generally, most integration methods are designed for a single species, and thus cannot be applied directly in a holo-omics setting. Any pattern discovered in omics data with the aim of describing host-microbiota interactions must include biomacromolecules originating from both sides of the holobiont boundary. This might be accomplished by first applying a standard (multi-)omics analysis method and then filtering the results afterwards, *e.g.* selecting modules containing genes from both the host and the microbiota. However, integrating the host-microbiota constraint as an integral part of the data analysis method could drastically reduce the search space, help deal with the curse of dimensionality and force results to include features from the host that might otherwise drown in the sea of microbial features. The methods described below are selected because we find them especially promising for solving challenges related specifically to holo-omics data sets.

Existing methodological frameworks and tools

Dimensionality reduction. The genetic repertoire of the host and its microbiomes captured by holo-omic data introduces complexities such as data sparsity, sampling variation, ecological differences, and host-specific genetic makeup. Furthermore, distinguishing between free-living and host-associated entities adds another layer of complexity. Since the number of biological features always surpasses the number of observations in holo-omic studies, dimensionality reduction is crucial to create human-interpretable visualisations to explore hidden structures and patterns, and prevent model overfitting.[52] Supervised dimensionality reduction – such as partial least squares discriminant analysis – relies on class labels or response variables to guide the dimensionality reduction process. However, such methods struggle when sample sizes are much smaller than the number of features. On the other hand, unsupervised dimensionality reduction like including matrix factorization and neighbour graph methods, allow discovery of structures in the data without relying on class labels or response variables.[52] Methods that find a few dimensions that are likely to be intrinsic come in two flavours; methods that identify a subset of relevant original features (feature selection), and methods that create new features by combining the original features (feature extraction). Feature extraction methods such as principal component analysis (PCA) (Fig. 3) and single value decomposition utilise variation preservation techniques to extract new features – so-called principal components – that are linear combinations of the original features. Principal components are commonly used for visualising clustering patterns and interpreting sample separation.[53]

Canonical correlation analysis (CCA) is a statistical technique akin to PCA in terms of finding a linear transformation of the original variables that consists of orthogonal vectors.[54] The objective of CCA is to summarise the linear relationship between two sets of variables by identifying linear combinations – called canonical variables – that maximise correlations based on pairs of loading vectors. Although CCA is not primarily designed for dimensionality reduction, it plays a crucial role in comprehending multivariate relationships by revealing the directions in which two sets of variables are most interdependent. Several extensions of CCA further enhance its applicability: (i) multiset CCAs analyse maximal correlations across multiple sets of omics data; (ii) sparse CCAs identify a subset of variables most relevant to the canonical variables by introducing sparsity constraints; (iii) regularised CCAs incorporate regularisation which is particularly beneficial when dealing with high-dimensional data or when variables are not well-captured by linear transformations; and (iv) partial least squares CCAs which focus on predicting one set of variables using another, thus combining aspects of partial least squares

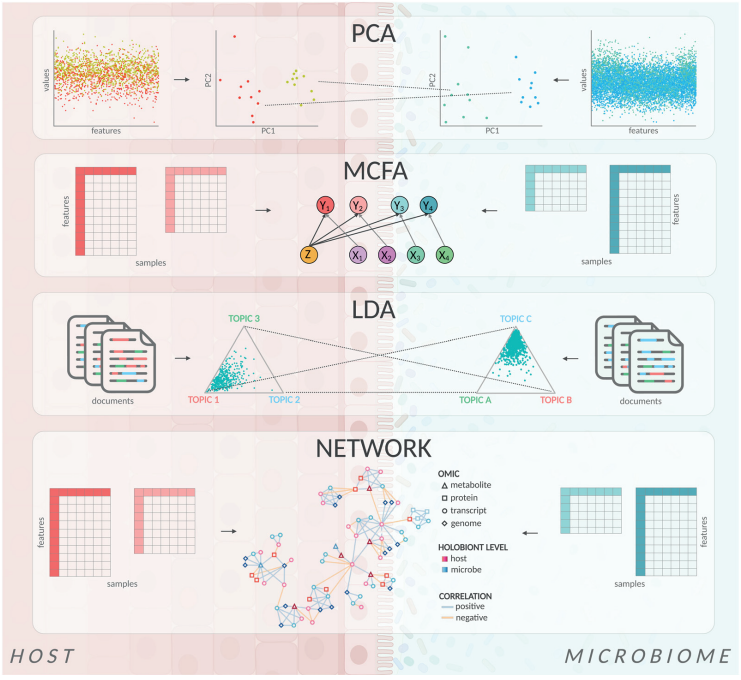


Fig. 3 Figurative summary of the methods discussed in this review. All can reduce inputs with many features to a smaller number of components in order to simplify interpretation of the underlying biological phenomena. PCA: Principal component analysis; MCFA: Multiset correlation and factor analysis; LDA: Latent dirichlet allocation. Created with biorender.com.

regression with CCA.[55] These extensions cater to diverse scenarios, offering flexibility to address specific challenges in multivariate analysis and canonical correlation.

Principal coordinates analysis (PCoA) is a linear transformation method similar to PCA which incorporates multidimensional scaling, creating dissimilarity matrices to visualise sample relationships.[56] Unlike PCA, PCoA is not limited to Euclidean measures and has been shown to be useful for comparing beta-diversity in microbial contexts. Non-metric multidimensional scaling (nMDS) is popular for amplicon/shotgun sequencing data, offering a rank-based approach that handles non-linear relationships and outliers effectively, albeit with potential distortions in global structures.[57–60] Non-linear methods like *t*-distributed stochastic neighbour embedding (*t*-SNE) and uniform manifold approximation and projection (UMAP) belong to the second type of dimensionality reduction, known as neighbour graph algorithms.[59–62] These methods emphasise preserving local structures, relying on graph layout algorithms to create probabilistic weighted graphs representing relationships between high-dimensional data points. UMAP and *t*-SNE differ primarily in their theoretical foundation for balancing the local and global structures.[53] While *t*-SNE results can vary between runs due to its stochastic nature and sensitivity to initialisation, UMAP, although also stochastic, tends to demonstrate more stability across runs. UMAP excels in preserving the global structure of the final projection while still capturing local relationships, it is hence a better choice for prediction tasks.[59,60] Nonetheless, it may struggle to distinguish closely nested clusters. It is crucial to note that all three non-linear methods are sensitive to initialisation, and it is recommended to employ the first two principal components from the linear approach as seeds for initiation. Users should implement these exploratory methods with caution, exploring various hyperparameters, running multiple projections for stability. When choosing a non-linear dimensionality reduction method, careful consideration of data scale, characteristics, and specific research goals is essential.[63]

Matrix factorisation (NMF and MCFA). Aforementioned methods for dimensionality reduction by matrix factorisation – such as PCA – enable compression of large datasets into a smaller feature space, and may thus facilitate identification of important biological factors for the variation in the observed

data. This is particularly relevant for holo-omic studies utilising a matrix factorisation approach, in which we consider complex systems through assembling a variety of data types from both sides of the holobiont, adding to the already prevalent imbalance of few biological samples and high feature counts. Challenges arise when size and heterogeneity of the dataset increases, which calls for adaptations of these matrix factorisation methods when applied in holo-omics.

Non-negative matrix factorisation (NMF)[64] is a method for dimensionality reduction that has been used both in several multi-omic studies and as a basis for additional tools for multi-omic data integration and analysis.[65–69] NMF has the same foundation as PCA, essentially decomposing a large data matrix (D) consisting of feature values (p) across biological replicates (n) into a reduced set of (r) linear expressions. These expressions are represented by two matrices smaller than the original data; one with weights (W, $p \times r$) and one with the reduced feature components (F, $r \times n$) (Fig. 4A).

In contrast to PCA, NMF requires the decomposition matrices to contain non-negative values only. This constraint causes the NMF-derived linear expressions to only consist of addends, thereby preventing cancellations between biological factors with opposing signs. NMF thus reflects the idea of assembling parts – analogous to the omic data layers – into a larger image representing the whole system. Simultaneously, the non-negativity constraint of NMF necessitates the compressed data to be seen as an approximation ($\approx$) of the real data rather than as an equality ($=$).[70] Our objective function for determining the decomposition matrices then becomes to minimise the difference between the real data (D) and the approximation (WF). This iterative approach may yield different solutions based on the initial weight and reduced component matrices, potentially affecting the outcome of the analysis.[71] Hence dimensionality reduction by NMF may be more in line with the analogy of assembling omic datasets to uncover interactions between layers of the complex system, although resulting in an approximated model with a potentially large residual difference caused by the lossy factorisation.

Another approach to holo-omic dataset integration based on matrix factorisation is multiset correlation and factor analysis (MCFA)[72] (Fig. 3 and 4C). While also seeking to compress observed data (D) into matrices for weights (W) and reduced components (F), MCFA effectively divides the model into two parts. One set of decomposition matrices fit the so-called shared space (S), consisting of reduced features with implied importance across all the included omics layers. This shared space is determined through an extension of CCA called probabilistic CCA (pCCA), and it serves the same purpose as the general decomposition seen in NMF. Additional sets of decomposition matrices are then fitted for each individual omics layer through factor analysis, based on the residual between the read data (D) and the modelled shared space. These ''private'' aspects of the model reflect contributions from factors that are only perceived as important for observations in specific omics layers. The full model then combines the shared and private spaces to approximate the real data, determining the weight and feature matrices through an expectation maximisation algorithm, with the remainder (ψ) being quantified a third addend to complete the expression.

By fitting the observed data to shared and private reduced features separately, the MCFA method may help distinguish between components with implied importance across all levels of the holobiont and those that only appear relevant for a particular omics layer. Additionally, introducing a private model layer for each omic may leave a smaller residual than had the model only covered components relevant for all included data layers. At the time of writing, MCFA has not been applied in a peer-reviewed study since its publication in August 2023, thus its versatility for holo-omic data integration has yet to be demonstrated.

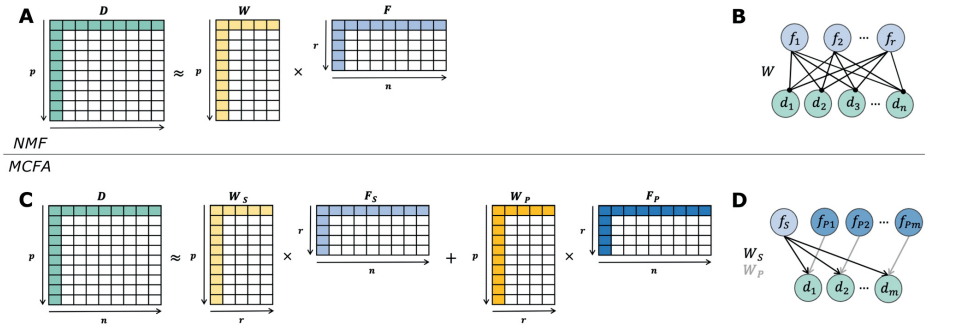


Fig. 4 Comparison of two methods for matrix factorisation; (A) and (B) non-negative matrix factorisation (NMF) and (C) and (D) multiset correlation and factor analysis (MCFA). Both methods reduce a full set of observed data d (columns of D) into linear expressions of reduced features f (columns of F) transformed by multiplication with weights (W). (A) and (C) In contrast to NMF, the MCFA method reduces the dataset into two spaces, either shared between all omics layers (S) or private to each one (P). (B) and (D) All features contribute to approximate the observed data for each shared omics layer, visualised in the same style as Fig. 4 in ref. 66.

Network analysis

Networks are graphs that represent complex relationships between interacting entities within a system.[73] The network is a ubiquitous concept in informatics that can represent many analogous systems like social interactions, flow of information, internet connections, and biological systems like genome-scale metabolic networks (GEMs), genomic co-occurrences, RNA regulation, protein–protein interactions, and metabolics-driven networks.[74] We use correlation networks as an example in this review, as they are suitable for holo-omic studies (Fig. 3). Multi-omics is a more mature concept than holo-omics, hence network methods for the former study type are more developed.[4] We suggest extending these multi-omic tools by integrating data crossing the holo-omic boundary as if it were another omics layer. In general, network analyses handle high-dimensional data well and can provide more interpretable results – compared to other approaches – in the form of node and edge statistics.

WGCNA is a popular framework for investigating associations between biological features within a single omics layer.[75] It calculates an adjacency matrix containing transformed, pairwise correlations between biological features such as genes, proteins, and metabolites. The adjacencies are transformed in order to obtain a scale-free network, in which features can be related to continuous and categorical external data like phenotypic traits or treatment groups. On the basis of these adjacencies, the topological overlap measure can be used together with hierarchical clustering to obtain a set of clusters where each biological feature becomes part of only one of the formulated clusters. In WGCNA terminology, these clusters are referred to as ''modules'' and are represented by their first principal component. This linear combination of biological features is referred to as an eigengene and is idealised to capture the most important variation of the module with limited noise. Since these modules are called without utilising information about treatments or traits, the method can be characterised as unsupervised.

WGCNA can be extended to holo-omic data[76] by relating the modules across the host–microbiome boundary. WGCNA has been applied for both clustering and dimensionality reduction in several multi and holo-omic studies related to both plant[77] and animal biology.[76,78–80] One study concerning the gut microbiome in patients with insulin sensitivity or resistance[79] applied a range of node selection and dimensionality reduction methods on their data, and used WGCNA to find clusters of hydrophilic and lipid metabolites. These were later connected to other omics layers to identify clusters associated with metabolism of the gut microbiome between the groups of patients.

Alternative clustering methods can also be employed for dimensionality reduction. A state-of-the-art example is the Leiden algorithm,[81] which is an optimisation-based form of clustering. The algorithm was used in a study of HIV patients in which they investigated health in relation to the microbiome of the patients. Specifically, they used the Leiden algorithm to detect clusters of microbiome-derived metabolites before integrating these features with other omics layers.[82] Similarly, a study of the SARS-CoV-2 used the Leiden algorithm to detect clusters of metabolites.[80]

Transkingdom network analysis (TkNA)[83] for holo- and multi-omics is a network-based method that detects biological features that differentiate treatment groups. TkNA is designed to handle a binary testing condition, such as ''disease'' and ''control''. The method consists of a comprehensive pipeline containing all the functions needed to transform normalised data into a network that can be readily visualised. TkNA creates a co-variation network and calculates node statistics like node degrees and bipartite betweenness centrality (BiBC). This approach emphasises that hub nodes with high BiBC and degree represent potential modules of the biological network. Additionally, TkNA interfaces with the Infomap[84] and Louvain[85] network clustering algorithms, which can aid in the interpretation of a biological network further.

The size and complexity of networks created from holo-omics datasets make them hard to interpret, hence it is necessary to find ways to categorise and structure the represented data. Clustering nodes and thus reducing the number of visual features to consider can help organise the network. This is exemplified in the aforementioned SARS-CoV-2 study where WGCNA was used to recognise clusters across omics layers. The cross-omic clusters correlating with disease severity revealed a relationship between host serum metabolites and microorganisms.[80]

In gene set enrichment analysis (GSEA), a gene set usually represents a metabolic pathway that performs a specific biological function. By testing whether there is an enrichment of genes from a specific pathway in a network cluster or module, we can argue that this pathway is captured by the module, thereby drawing further conclusions about its activity by interpreting the module's omics profile and association to other phenotypic metadata. GSEA can be applied on clusters that are defined using any clustering algorithm. An example is a study on the Atlantic salmon[76] where gene enrichment analysis was used to show that certain host RNA genes responded to long chain fatty acids in the feed. A similar method[86] for improving interpretability is network enrichment, in which functional information and network connectivity is integrated. Instead of testing for a significant difference between treatment groups like GSEA, network enrichment quantifies the differential representation among neighbours in the gene network.[87]

A network can be interpreted by statistical concepts that describe crucial properties of the nodes and how they are connected. Degree is simply the number of neighbours of any node. The degree can be expressed relative to the node with the highest number of neighbours, hence degree centrality. Node betweenness describes how many of the pairwise node connections in the network pass through a specific node. If this betweenness measurement is high, the node represents a bottleneck and is indicated to have a potential regulatory effect.[83] The cluster coefficient of a node describes the number of edges between its neighbours in relation to the possible number of edges between these. Coreness considers the neighbourhood of a node as it describes whether a node is part of a

''core'' of nodes that are all interconnected with a certain degree (k). Hence, a network can be characterised by the maximum coreness of all nodes. Eigenvector centrality is another network statistic computed for each node in a network. The maximum eigenvalue of the adjacency matrix is computed and is used to normalise the eigenvector, which becomes the eigenvector centralities. Generally, nodes with high eigenvector centrality are essential and interact closely with their respective neighbours.[74] In a study of periodontal disease and response to different treatment, eigenvector centrality was used specifically to find nodes in the network that were connected to other highly connected nodes.[88] This revealed microbial taxa that could be more closely associated with the patients' health status. The same study also looked at the network transitivity – describing the ratio of connected triplets to the number of possible connected triplets – for the networks over different patients and disease states. This statistic is high in the presence of clusters, and the more severe disease cases in the study were associated with lower transitivity. A higher interdependence (*i.e.* transitivity) between microbes was therefore shown to be beneficial for the patient. The severe cases were also more often associated with networks with a high diameter – meaning the shortest path between the most distant nodes – which is expected with low transitivity.

Other tools and frameworks

In addition to the methods introduced above, there are various other multi-omic integration tools that could be useful for holo-omic data analysis. A comprehensive and constantly growing community-maintained list of such tools can be found online in a dedicated Git repository.[89] Aside from a handful of methods aimed at microbiome analyses, this list mainly represents a host perspective. Nevertheless, many of the tools could be used in a host–microbiome context, including the examples highlighted below.

MixOmics[90] is a toolkit that offers both unsupervised and supervised statistical approaches for multi-table datasets, ranging from single omic analysis to complex multi-omics. The supervised method for multi-omics, titled Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO),[91] is based on partial least squares regression/projection to latent structures[92] discriminant analysis (PLS-DA)[93] and sparse generalised canonical correlation analysis (sGCCA),[54] an extension of the CCA method. The sparse version of DIABLO involves using lasso[94] to select those features from each layer that best discriminate between groups of interest. Since DIABLO does not assume any particular distributions from the input data,[91] it is applicable for holo-omic datasets, as long as each layer is normalised in a way that is appropriate to that data type. The limiting factor of this approach is that DIABLO is a supervised method aimed at classification of data into pre-established groups of interest, which makes it less useful for basic, explorative holo-omic studies. Examples where this tool has already been used include a study of the relationships of gut microbiota, dietary fatty acids, and liver gene expression in mice;[95] and the effects of cyanobacterial blooms on the microbiome and metabolome of the medaka fish species.[96]

For studies that do not involve a predefined grouping variable, mixOmics is compatible with mixKernel[97] for multi-omics integration. This explorative, unsupervised approach is based on forming a kernel – a symmetric and positive function that provides pairwise similarities between samples – to represent each layer of data.[97] These can be combined into a meta-kernel by creating one of two alternatives: (i) a consensus kernel, or (ii) a sparse kernel that preserves the topology of the original data. The meta-kernel can then be used in downstream analyses, for example kernel PCA (KPCA)[98] for visualisation of the different layers. Since mixKernel is suited for heterogeneous data, it is also applicable for holo-omics. So far, this method has not been commonly utilised in a host–microbiome context, but it successfully complemented simpler, single-table statistics when selecting plant-beneficial bacterial strains for rice cultivation based on plant growth related measurements.[99]

Another explorative method is mCIA[100] – a multi-table version of co-inertia analysis (CIA or COIA)[101] – which has been tested for selecting rice growth promoting bacteria.[99] CIA resembles sPLS in that it also searches to maximise the covariance between eigenvectors.[100] mCIA has been extended to create sparse mCIA (smCIA) which adds feature selection, improving the interpretability of the results.[102] There is also a further extension, structured sparse mCIA (ssmCIA), which enables incorporating structural information about variables, such as regulatory networks for genes.[102] However, this is less relevant for holo-omic analyses as such pre-existing information is seldomly available.

Compositional omics model-based integration or COMBI[103] is another explorative, unsupervised multi-table method. It is particularly appropriate for host–microbiome analyses since it has been designed to account for compositionality, a feature common to many microbiome measurements such as 16S rRNA gene amplicon data and shotgun metagenomic data.[104] Specifically, compositional data is handled through using the centred log-ratio transform as a link function in the models, while the integrative part of the approach is based on inferring latent variables.[103] This method also offers visualisation of the results as a multiplot showing the features with the largest loadings.

Finally, latent dirichlet allocation (LDA) is a form of unsupervised dimensionality reduction[105] (Fig. 3). It uses a specific terminology as it was originally invented for use in text mining. In a corpus – a set of text documents that represent a spectrum of topics – it allocates each word to a predetermined number of topics so that each word in the total vocabulary belongs to one topic. Each topic is a set of words that, as a whole, revolve around a semantic context. Although the topics are coherent and represent an underlying theme, the title of each topic must be defined manually by interpretation of the listed words in each topic. As a text mining tool, LDA doesn't immediately lend itself useful for biological data inquiries. But, consider substituting a corpus for an omics layer: documents become

biological samples, and genes or compounds become the words. By doing so, the model will be able to capture latent topics defined by biological features that tend to occur together in the same documents (co-abundance), forming topics that represent metabolic functions in the samples. This text-biology analogy means that LDA can be applied for use in biological studies.[106]

Conclusion

As the biological insights of holo-omics are limited by the computational model that picks up host–microbiome interactions, there is a need for better modelling tools. Typically, holo-omic analysis is performed with complex models that use clustering or network analyses coupled with functional enrichment analyses to assign biological functions to interacting groups of biochemical compounds across the host–microbiome boundary. As holo-omics is a specialised case of multi-omics, it is possible to apply multi-omic tools in a holo-omics context. In multi-omics, the omics layers are integrated by correlating clusters of biochemical compounds between layers across the samples. Carried forward, it is possible to integrate the two sides of the holobiont by correlating clusters of biochemical compounds between the host and microbiome sides across the samples.

As this is a new, fast-moving field, there still is no consensus of what is the best way to do science using holo-omics. We hope that this review can generate discussion and new ideas on how to approach the further development of holo-omic methodologies, and we are positive that gold standard methodologies will soon be established.

Author contributions

Conceptualization: CMK, OØ, PBP, TRH, VTEA; funding acquisition: PBP, TRH; supervision: OØ, PBP, TRH, VTEA; visualisation: CMK, JM; writing – original draft: CMK, JM, IMTB, WL, OØ, PBP, TRH, VTEA; writing – review & editing: CMK, JM, IMTB, WL, OØ, PBP, TRH, VTEA.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We gratefully acknowledge the financial support of the European Union's Horizon 2020 research and innovation programme under the grant agreements 101000213-HoloRuminant and 101000309-3D-omics, as well as the Novo Nordisk Foundation under 0054575-SuPAcow.

References

1 J. Roughgarden, S. F. Gilbert, E. Rosenberg, I. Zilber-Rosenberg and E. A. Lloyd, Holobionts as Units of Selection and a Model of Their Population Dynamics and Evolution, *Biol. Theory*, 2018, **13**, 44–65.
2 G. T. Jung, K.-P. Kim and K. Kim, How to interpret and integrate multi-omics data at systems level, *Anim. Cells Syst.*, 2020, **24**, 1–7.
3 L. Xu, G. Pierroz, H. M.-L. Wipf, C. Gao, J. W. Taylor, P. G. Lemaux and D. Coleman-Derr, Holo-omics for deciphering plant-microbiome interactions, *Microbiome*, 2021, **9**, 69.
4 N. Malmuthuge and L. L. Guan, Noncoding RNAs: Regulatory Molecules of Host–Microbiome Crosstalk, *Trends Microbiol.*, 2021, **29**, 713–724.
5 S. L. La Rosa, M. L. Leth, L. Michalak, M. E. Hansen, N. A. Pudlo, R. Glowacki, G. Pereira, C. T. Workman, M. Ø. Arntzen, P. B. Pope, E. C. Martens, M. A. Hachem and B. Westereng, The human gut Firmicute Roseburia intestinalis is a primary degrader of dietary β-mannans, *Nat. Commun.*, 2019, **10**, 905.
6 P. Fan, B. Bian, L. Teng, C. D. Nelson, J. Driver, M. A. Elzo and K. C. Jeong, Host genetic effects upon the early gut microbiota in a bovine model with graduated spectrum of genetic variation, *ISME J.*, 2020, **14**, 302–317.
7 L. Nyholm, A. Koziol, S. Marcos, A. B. Botnen, O. Aizpurua, S. Gopalakrishnan, M. T. Limborg, M. T. P. Gilbert and A. Alberdi, Holo-Omics: Integrated Host-Microbiota Multi-omics for Basic and Applied Biological Research, *iScience*, 2020, **23**, 101414.
8 P. R. Myer, Bovine Genome-Microbiome Interactions: Metagenomic Frontier for the Selection of Efficient Productivity in Cattle Systems, *mSystems*, 2019, **4**(3), DOI: **10.1128/msystems.00103-19**.
9 V. Aggarwala, G. Liang and F. D. Bushman, Viral communities of the human gut: metagenomic analysis of composition and dynamics, *Mobile DNA*, 2017, **8**, 12.
10 I. Mizrahi, in *The Prokaryotes: Prokaryotic Biology and Symbiotic Associations*, ed. E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt and F. Thompson, Springer, Berlin, Heidelberg, 2013, pp. 533–544.
11 M. Kim, M. Morrison and Z. Yu, Status of the phylogenetic diversity census of ruminal microbiomes, *FEMS Microbiol. Ecol.*, 2011, **76**, 49–63.
12 L. Yuan, C. Hensley, H. M. Mahsoub, A. K. Ramesh and P. Zhou, in *Progress in Molecular Biology and Translational Science*, ed. J. Sun, Academic Press, 2020, vol. 171, pp. 15–60.
13 Z. Li, X. Wang, Y. Zhang, Z. Yu, T. Zhang, X. Dai, X. Pan, R. Jing, Y. Yan, Y. Liu, S. Gao, F. Li, Y. Huang, J. Tian, J. Yao, X. Xing, T. Shi, J. Ning, B. Yao, H. Huang and

Y. Jiang, Genomic insights into the phylogeny and biomass-degrading enzymes of rumen ciliates, *ISME J.*, 2022, **16**, 2775–2787.
14 I. V. Grigoriev, R. Nikitin, S. Haridas, A. Kuo, R. Ohm, R. Otillar, R. Riley, A. Salamov, X. Zhao, F. Korzeniewski, T. Smirnova, H. Nordberg, I. Dubchak and I. Shabalov, MycoCosm portal: gearing up for 1000 fungal genomes, *Nucleic Acids Res.*, 2014, **42**, D699–D704.
15 E. Jami, A. Israel, A. Kotser and I. Mizrahi, Exploring the bovine rumen bacterial community from birth to adulthood, *ISME J.*, 2013, **7**, 1069–1079.
16 D. Laukens, B. M. Brinkman, J. Raes, M. De Vos and P. Vandenabeele, Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design, *FEMS Microbiol. Rev.*, 2016, **40**, 117–132.
17 S. Kieser, E. M. Zdobnov and M. Trajkovski, Comprehensive mouse microbiota genome catalog reveals major difference to its human counterpart, *PLoS Comput. Biol.*, 2022, **18**, e1009947.
18 M. Chu and X. Zhang, Bacterial Atlas of Mouse Gut Microbiota, *Cell. Microbiol.*, 2022, **2022**, e5968814.
19 S. P. Rosshart, B. G. Vassallo, D. Angeletti, D. S. Hutchinson, A. P. Morgan, K. Takeda, H. D. Hickman, J. A. McCulloch, J. H. Badger, N. J. Ajami, G. Trinchieri, F. P.-M. de Villena, J. W. Yewdell and B. Rehermann, Wild Mouse Gut Microbiota Promotes Host Fitness and Improves Disease Resistance, *Cell*, 2017, **171**, 1015–1028.e13.
20 A. V.-P. De León, M. Hoetzinger, T. Hensen, S. Gupta, B. Weston, S. M. Johnsen, J. A. Rasmussen, C. G. Clausen, L. Pless, A. R. A. Veríssimo, K. Rudi, L. Snipen, C. R. Karlsen, M. T. Limborg, S. Bertilsson, I. Thiele, T. R. Hvidsten, S. R. Sandve, P. B. Pope and S. L. La Rosa, The Salmon Microbial Genome Atlas enables novel insights into bacteria-host interactions via functional mapping, *BioRxiv*, 2023, DOI: **10.1101/2023.12.10.570985**.
21 B. Bai, W. Liu, X. Qiu, J. Zhang, J. Zhang and Y. Bai, The root microbiome: Community assembly and its contributions to plant fitness, *J. Integr. Plant Biol.*, 2022, **64**, 230–243.
22 H. R. Barajas, S. Martínez-Sánchez, M. F. Romero, C. H. Álvarez, L. Servín-González, M. Peimbert, R. Cruz-Ortega, F. García-Oliva and L. D. Alcaraz, Testing the Two-Step Model of Plant Root Microbiome Acquisition Under Multiple Plant Species and Soil Sources, *Front. Microbiol*, 2020, **11**, DOI: **10.3389/fmicb.2020.542742**.
23 C. R. Fitzpatrick, J. Copeland, P. W. Wang, D. S. Guttman, P. M. Kotanen and M. T. J. Johnson, Assembly and ecological function of the root microbiome across angiosperm plant species, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **115**, E1157–E1165.
24 A. Pascale, S. Proietti, I. S. Pantelides and I. A. Stringlis, Modulation of the Root Microbiome by Plant Molecules: The Basis for Targeted Disease Suppression and Plant Growth Promotion, *Front. Plant Sci*, 2019, **10**, DOI: **10.3389/fpls.2019.01741**.
25 M. I. A. Cavassim, S. Moeskjær, C. Moslemi, B. Fields, A. Bachmann, B. J. Vilhjálmsson, M. H. Schierup, J. P. W. Young and S. U. Andersen, Symbiosis genes show a unique pattern of introgression and selection within a Rhizobium leguminosarum species complex, *Microb. Genomics*, 2020, **6**(4), DOI: **10.1099/mgen.0.000351**.
26 K. Raymann and N. A. Moran, The role of the gut microbiome in health and disease of adult honey bee workers, *Curr. Opin. Insect. Sci.*, 2018, **26**, 97–104.
27 G. Bonilla-Rosso and P. Engel, Functional roles and metabolic niches in the honey bee gut microbiota, *Curr. Opin. Microbiol.*, 2018, **43**, 69–76.
28 W. K. Kwong and N. A. Moran, Gut microbial communities of social bees, *Nat. Rev. Microbiol.*, 2016, **14**, 374–384.
29 J. Liberti, T. Kay, A. Quinn, L. Kesner, E. T. Frank, A. Cabirol, T. O. Richardson, P. Engel and L. Keller, *The gut microbiota affects the social network of honeybees*, *Nat. Ecol. Evol.*, 2022, **6**, 1471–1479.
30 S. T. Bates, G. W. G. Cropsey, J. G. Caporaso, R. Knight and N. Fierer, Bacterial Communities Associated with the Lichen Symbiosis, *Appl. Environ. Microbiol.*, 2011, **77**, 1309–1314.
31 T. J. Hammer, J. G. Sanders and N. Fierer, Not all animals need a microbiome, *FEMS Microbiol. Lett.*, 2019, **366**, fnz117.
32 P. Luczynski, K.-A. McVey Neufeld, C. S. Oriach, G. Clarke, T. G. Dinan and J. F. Cryan, Growing up in a Bubble: Using Germ-Free Animals to Assess the Influence of the Gut Microbiota on Brain and Behavior, *Int. J. Neuropsychopharmacol.*, 2016, **19**, pyw020.
33 M. Jans and L. Vereecke, A guide to germ-free and gnotobiotic mouse technology to study health and disease, *FEBS J.*, 2024, DOI: **10.1111/febs.17124**.
34 N. A. Moran, H. Ochman and T. J. Hammer, Evolutionary and Ecological Consequences of Gut Microbial Communities, *Annu. Rev. Ecol. Evol. Syst.*, 2019, **50**, 451–475.
35 E. B. V. Arnam, C. R. Currie and J. Clardy, Defense contracts: molecular protection in insect-microbe symbioses, *Chem. Soc. Rev.*, 2018, **47**, 1638–1651.
36 L. Margulis, Serial endosymbiotic theory (SET) and composite individuality, *Microbiol.: Today*, 2004, **31**, 173–174.
37 W. H. Hoover and T. K. Miller, Rumen Digestive Physiology and Microbial Ecology, *Vet. Clin. North Am. Food Anim. Pract.*, 1991, **7**, 311–325.
38 L. G. M. Baas-Becking, *Geobiologie; of inleiding tot de milieukunde*, WP Van Stockum & Zoon NV, 1934.
39 Y. Peng, J. Cai, W. Wang and B. Su, Multiple Inter-Kingdom Horizontal Gene Transfers in the Evolution of the Phosphoenolpyruvate Carboxylase Gene Family, *PLoS One*, 2012, **7**, e51159.
40 I. Mizrahi and E. Jami, A method to the madness, *EMBO Rep.*, 2021, **22**, e52269.
41 E. Rosenberg and I. Zilber-Rosenberg, The hologenome concept of evolution after 10 years, *Microbiome*, 2018, **6**, 78.
42 J. B. Russell and J. L. Rychlik, Factors That Alter Rumen Microbial Ecology, *Science*, 2001, **292**, 1119–1122.
43 E. Noor, S. Cherkaoui and U. Sauer, Biological insights through omics data integration, *Curr. Opin. Syst. Biol.*, 2019, **15**, 39–47.

44 A. J. Lopatkin and J. J. Collins, Predictive biology: modelling, understanding and harnessing microbial complexity, *Nat. Rev. Microbiol.*, 2020, **18**, 507–520.

45 C. Ramon, M. G. Gollub and J. Stelling, Integrating –omics data into genome-scale metabolic network models: principles and challenges, *Essays Biochem.*, 2018, **62**, 563–574.

46 C. Gu, G. B. Kim, W. J. Kim, H. U. Kim and S. Y. Lee, Current status and applications of genome-scale metabolic models, *Genome Biol.*, 2019, **20**, 121.

47 P. Borzou, J. Ghaisari, I. Izadi, Y. Eshraghi and Y. Gheisari, A novel strategy for dynamic modeling of genome-scale interaction networks, *Bioinformatics*, 2023, **39**, btad079.

48 U. Kolli, R. Jalodia, S. Moidunny, P. K. Singh, Y. Ban, J. Tao, G. N. Cantu, E. Valdes, S. Ramakrishnan and S. Roy, Multi-omics analysis revealing the interplay between gut microbiome and the host following opioid use, *Gut Microbes*, 2023, **15**, 2246184.

49 J. Shi, H. Qiu, Q. Xu, Y. Ma, T. Ye, Z. Kuang, N. Qu, C. Kan, N. Hou, F. Han and X. Sun, Integrated multi-omics analyses reveal effects of empagliflozin on intestinal homeostasis in high-fat-diet mice, *iScience*, 2023, **26**, 105816.

50 M. Marynowska, D. Sillam-Dussès, B. Untereiner, D. Klimek, X. Goux, P. Gawron, Y. Roisin, P. Delfosse and M. Calusinska, A holobiont approach towards polysaccharide degradation by the highly compartmentalised gut system of the soil-feeding higher termite Labiotermes labralis, *BMC Genomics*, 2023, **24**, 115.

51 S.-C. Cheng, C.-B. Liu, X.-Q. Yao, J.-Y. Hu, T.-T. Yin, B. K. Lim, W. Chen, G.-D. Wang, C.-L. Zhang, D. M. Irwin, Z.-G. Zhang, Y.-P. Zhang and L. Yu, Hologenomic insights into mammalian adaptations to myrmecophagy, *Natl. Sci. Rev.*, 2023, **10**, nwac174.

52 L. Van Der Maaten, E. O. Postma and H. J. van den Herik, *et al.*, Dimensionality reduction: A comparative review, *J. Mach. Learn. Res.*, 2009, **10**, 13.

53 F. Anowar, S. Sadaoui and B. Selim, Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE), *Comput. Sci. Rev.*, 2021, **40**, 100378.

54 A. Tenenhaus, C. Philippe, V. Guillemot, K.-A. Le Cao, J. Grill and V. Frouin, Variable selection for generalized canonical correlation analysis, *Biostatistics*, 2014, **15**, 569–583.

55 X. Zhuang, Z. Yang and D. Cordes, A technical review of canonical correlation analysis for neuroscience applications, *Hum. Brain Mapp.*, 2020, **41**, 3807–3833.

56 J. C. Gower, *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, 2015, pp. 1–7.

57 F. M. Ibarbalz, M. V. Pérez, E. L. M. Figuerola and L. Erijman, The Bias Associated with Amplicon Sequencing Does Not Affect the Quantitative Assessment of Bacterial Community Dynamics, *PLoS One*, 2014, **9**, e99722.

58 J. B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, 1964, **29**, 1–27.

59 T. T. Cai and R. Ma, Theoretical foundations of t-SNE for visualizing high-dimensional clustered data, *J. Mach. Learn. Res.*, 2022, **23**, 301:13581–301:13634.

60 J. Gauß, Topological and Practical Aspects of Data Separability in Complex High-Dimensional Data.

61 L. McInnes, J. Healy and J. Melville, *arXiv*, 2020, preprint, arXiv:1802.03426, DOI: **10.48550/arXiv.1802.03426**.

62 L. Van der Maaten and G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res*, 2008, **9**(86), 2579–2605.

63 M. Rahmatbakhsh, A. Gagarinova and M. Babu, Bioinformatic Analysis of Temporal and Spatial Proteome Alternations During Infections, *Front. Genet*, 2021, **12**, DOI: **10.3389/fgene.2021.667936**.

64 D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*, 1999, **401**, 788–791.

65 R. Tappu, J. Haas, D. H. Lehmann, F. Sedaghat-Hamedani, E. Kayvanpour, A. Keller, H. A. Katus, N. Frey and B. Meder, Multi-omics assessment of dilated cardiomyopathy using non-negative matrix factorization, *PLoS One*, 2022, **17**, e0272093.

66 A. R. Kriebel and J. D. Welch, UINMF performs mosaic integration of single-cell multi-omic datasets using non-negative matrix factorization, *Nat. Commun.*, 2022, **13**, 780.

67 Z. Yang and G. Michailidis, A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data, *Bioinformatics*, 2016, **32**, 1–8.

68 S. Mallik, A. Sarkar, S. Nath, U. Maulik, S. Das, S. K. Pati, S. Ghosh and Z. Zhao, 3PNMF-MKL: A non-negative matrix factorization-based multiple kernel learning method for multi-modal data integration and its application to gene signature detection, *Front. Genet*, 2023, DOI: **10.3389/fgene.2023.1095330**.

69 P. Chalise and B. L. Fridley, Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm, *PLoS One*, 2017, **12**, e0176278.

70 A. Akalin, *11.3 Matrix factorization methods for unsupervised multi-omics data integration|Computational Genomics with R.*

71 F. Esposito, A Review on Initialization Methods for Non-negative Matrix Factorization: Towards Omics Data Experiments, *Mathematics*, 2021, **9**, 1006.

72 B. C. Brown, C. Wang, S. Kasela, F. Aguet, D. C. Nachun, K. D. Taylor, R. P. Tracy, P. Durda, Y. Liu, W. C. Johnson, D. Van Den Berg, N. Gupta, S. Gabriel, J. D. Smith, R. Gerzsten, C. Clish, Q. Wong, G. Papanicolau, T. W. Blackwell, J. I. Rotter, S. S. Rich, R. G. Barr, K. G. Ardlie, D. A. Knowles and T. Lappalainen, Multiset correlation and factor analysis enables exploration of multi-omics data, *Cell Genomics*, 2023, **3**, 100359.

73 D. Jiang, C. R. Armour, C. Hu, M. Mei, C. Tian, T. J. Sharpton and Y. Jiang, Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities, *Front. Genet*, 2019, **10**, DOI: **10.3389/fgene.2019.00995**.

74 Z. Liu, A. Ma, E. Mathé, M. Merling, Q. Ma and B. Liu, Network analyses in microbiome based on high-throughput multi-omics data, *Briefings Bioinf.*, 2021, **22**, 1639–1655.

75 P. Langfelder and S. Horvath, WGCNA: an R package for weighted correlation network analysis, *BMC Bioinf.*, 2008, **9**, 559.

76 M. A. Strand, Y. Jin, S. R. Sandve, P. B. Pope and T. R. Hvidsten, Transkingdom network analysis provides insight into host–microbiome interactions in Atlantic salmon, *Comput. Struct. Biotechnol. J.*, 2021, **19**, 1028–1034.

77 J. Xie, Y. Ma, X. Li, J. Wu, F. Martin and D. Zhang, Multi-feature analysis of age-related microbiome structures reveals defense mechanisms of Populus tomentosa trees, *New Phytol.*, 2023, **238**, 1636–1650.

78 B. Czech, Y. Wang, K. Wang, H. Luo, L. Hu and J. Szyda, Host transcriptome and microbiome interactions in Holstein cattle under heat stress condition, *Front. Microbiol*, 2022, **13**, DOI: **10.3389/fmicb.2022.998093**.

79 T. Takeuchi, T. Kubota, Y. Nakanishi, H. Tsugawa, W. Suda, A. T.-J. Kwon, J. Yazaki, K. Ikeda, S. Nemoto, Y. Mochizuki, T. Kitami, K. Yugi, Y. Mizuno, N. Yamamichi, T. Yamazaki, I. Takamoto, N. Kubota, T. Kadowaki, E. Arner, P. Carninci, O. Ohara, M. Arita, M. Hattori, S. Koyasu and H. Ohno, Gut microbial carbohydrate metabolism contributes to insulin resistance, *Nature*, 2023, **621**, 389–395.

80 W. C. Albrich, T. S. Ghosh, S. Ahearn-Ford, F. Mikaeloff, N. Lunjani, B. Forde, N. Suh, G.-R. Kleger, U. Pietsch, M. Frischknecht, C. Garzoni, R. Forlenza, M. Horgan, C. Sadlier, T. R. Negro, J. Pugin, H. Wozniak, A. Cerny, U. Neogi, P. W. O'Toole and L. O'Mahony, A high-risk gut microbiota configuration associates with fatal hyperinflammatory immune and metabolic responses to SARS-CoV-2, *Gut Microbes*, 2022, **14**, 2073131.

81 V. A. Traag, L. Waltman and N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities, *Sci. Rep.*, 2019, **9**, 5233.

82 F. Mikaeloff, M. Gelpi, R. Benfeitas, A. D. Knudsen, B. Vestad, J. Høgh, J. R. Hov, T. Benfield, D. Murray, C. G. Giske, A. Mardinoglu, M. Trøseid, S. D. Nielsen and U. Neogi, Network-based multi-omics integration reveals metabolic at-risk profile within treated HIV-infection, *eLife*, 2023, **12**, e82785.

83 N. K. Newman, M. Macovsky, R. R. Rodrigues, A. M. Bruce, J. W. Pederson, S. S. Patil, J. Padiadpu, A. K. Dzutsev, N. Shulzhenko, G. Trinchieri, K. Brown and A. Morgun, *Nat. Protoc.*, 2024, **19**, DOI: **10.1038/s41596-024-00960-w**.

84 M. Rosvall, D. Axelsson and C. T. Bergstrom, The map equation, *Eur. Phys. J.-Spec. Top.*, 2009, **178**, 13–23.

85 V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks, *J. Stat. Mech.: Theory Exp.*, 2008, **2008**, P10008.

86 J.-H. Hung, T.-H. Yang, Z. Hu, Z. Weng and C. DeLisi, Gene set enrichment analysis: performance evaluation and usage guidelines, *Briefings Bioinf.*, 2012, **13**, 281–291.

87 A. Alexeyenko, W. Lee, M. Pernemalm, J. Guegan, P. Dessen, V. Lazar, J. Lehtiö and Y. Pawitan, Network enrichment analysis: extension of gene-set enrichment analysis to gene networks, *BMC Bioinf.*, 2012, **13**, 226.

88 L. Sisk-Hackworth, A. Ortiz-Velez, M. B. Reed and S. T. Kelley, Compositional Data Analysis of Periodontal Disease Microbial Communities, *Front. Microbiol*, 2021, **12**, DOI: **10.3389/fmicb.2021.617949**.

89 Github: mikelove/awesome-multi-omics, **https://github.com/mikelove/awesome-multi-omics**, (accessed February 2024).

90 F. Rohart, B. Gautier, A. Singh and K.-A. L. Cao, mixOmics: An R package for 'omics feature selection and multiple data integration, *PLoS Comput. Biol.*, 2017, **13**, e1005752.

91 A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt and K.-A. Lê Cao, DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays, *Bioinformatics*, 2019, **35**, 3055–3062.

92 H. Abdi, Partial least squares regression and projection on latent structure regression (PLS Regression), *WIREs Comput. Stat.*, 2010, **2**, 97–106.

93 K.-A. Lê Cao, S. Boitard and P. Besse, Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems, *BMC Bioinf.*, 2011, **12**, 253.

94 R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, *J. R. Stat. Soc., B: Stat. Methodol.*, 1996, **58**, 267–288.

95 M. Schoeler, S. Ellero-Simatos, T. Birkner, J. Mayneris-Perxachs, L. Olsson, H. Brolin, U. Loeber, J. D. Kraft, A. Polizzi, M. Martí-Navas, J. Puig, A. Moschetta, A. Montagner, P. Gourdy, C. Heymes, H. Guillou, V. Tremaroli, J. M. Fernández-Real, S. K. Forslund, R. Burcelin and R. Caesar, The interplay between dietary fatty acids and gut microbiota influences host metabolism and hepatic steatosis, *Nat. Commun.*, 2023, **14**, 5329.

96 A. Gallet, S. Halary, C. Duval, H. Huet, S. Duperron and B. Marie, Disruption of fish gut microbiota composition and holobiont's metabolome during a simulated Microcystis aeruginosa (Cyanobacteria) bloom, *Microbiome*, 2023, **11**, 108.

97 J. Mariette and N. Villa-Vialaneix, Unsupervised multiple kernel learning for heterogeneous data integration, *Bioinformatics*, 2018, **34**, 1009–1015.

98 B. Schölkopf, A. Smola and K.-R. Müller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Comput.*, 1998, **10**, 1299–1319.

99 M. Truu, S. K. Gopalasubramaniam, G. Muthukrishanan and J. Truu, Application of data integration for rice bacterial strain selection by combining their osmotic stress response and plant growth-promoting traits, *Front. Microbiol*, 2022, **13**, DOI: **10.3389/fmicb.2022.1058772**.

100 C. Meng, B. Kuster, A. C. Culhane and A. M. Gholami, A multivariate approach to the integration of multi-omics datasets, *BMC Bioinf.*, 2014, **15**, 162.

101 S. Dray, D. Chessel and J. Thioulouse, Co-Inertia Analysis and the Linking of Ecological Data Tables, *Ecology*, 2003, **84**, 3078–3089.

102 E. J. Min and Q. Long, Sparse multiple co-Inertia analysis with application to integrative analysis of multi -Omics data, *BMC Bioinf.*, 2020, **21**, 141.

103 S. Hawinkel, L. Bijnens, K.-A. L. Cao and O. Thas, Model-based joint visualization of multiple compositional omics datasets, *NAR: Genomics Bioinf.*, 2020, **2**, lqaa050.

104 G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn and J. J. Egozcue, Microbiome Datasets Are Compositional: And This Is Not Optional, *Front. Microbiol*, 2017, **8**, DOI: 10.3389/fmicb.2017.02224.
105 D. M. Blei, Latent Dirichlet Allocation, *J. Mach. Learn. Res.*, 2003, **3**, 993–1022.
106 C. Tataru, M. Peras, E. Rutherford, K. Dunlap, X. Yin, B. S. Chrisman, T. Z. DeSantis, D. P. Wall, S. Iwai and M. M. David, Topic modeling for multi-omic integration in the human gut microbiome and implications for Autism, *Sci. Rep.*, 2023, **13**, 11353.

Paper #4

Manuscript, November 2024

Protozoal populations drive system-wide variation in the rumen microbiome

Carl M. Kobel[1], Andy Leu[2], Arturo Vera-Ponce de León[1], Ove Øyås[1], Wanxin Lai[3], Ianina Altshuler[1,4], Live H. Hagen[3], Rasmus D. Wollenberg[5], Cassie R. Bakshani[6,7], William G. T. Willats[6], Laura Nicoll[8], Simon J. McIlroy[2], Torgeir R. Hvidsten[3], Chris Greening[9], Gene W. Tyson[2], Rainer Roehe[8], Velma T. E. Aho[1]*, Phillip B. Pope[1,2,3]*

1 Faculty of Biosciences, Norwegian University of Life Sciences, 1432 Ås, Norway.
2 Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology (QUT), Translational Research Institute, Woolloongabba, QLD, Australia.
3 Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1432 Ås, Norway.
4 Microbiome Adaptation to the Changing Environment laboratory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
5 DNASense ApS, Aalborg, Denmark.
6 Department of Biology, School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom.
7 Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom.
8 Scotland's Rural College, Edinburgh, United Kingdom.
9 Department of Microbiology, Biomedicine Discovery Institute, Monash University, Melbourne, Australia

* These authors contributed equally to this work.

Abstract

While rapid progress has been made to characterize bacterial and archaeal populations residing in the rumen microbiome, insight into how they interact with keystone protozoal species remains elusive. Here, we reveal two distinct rumen community types (RCT-A and RCT-B) that are not strongly associated with host phenotype nor genotype but instead linked to protozoal community patterns. We leveraged a series of multi-omic datasets to show that the dominant Epidinium spp. in animals with RCT-B employ a plethora of fiber-degrading enzymes that present enriched Prevotella spp. a favorable carbon landscape to forage upon. Conversely, animals with RCT-A, dominated by genera Isotricha and Entodinium, harbor a more even distribution of fiber, protein, and amino acid fermenters, reflected by higher detection of metabolites from both protozoal and bacterial metabolism. We reveal that microbiome variation transcends key protozoal and bacterial populations, which should act as an important consideration for future development of microbiome-based technologies.

Introduction

The herbivore rumen is a highly specialized organ that has co-evolved in symbiosis with a complex microbiome, made up of thousands of microbial populations whose interactions collectively convert plant material into energy-yielding metabolites for the host's sustenance.

The rumen microbiome acts as an interface between the nutrient potential of the feed and the metabolism of the host animal, and is represented by a wide radiation of the tree of life covering all domains: Archaea, Bacteria, and Eukarya (ciliate protozoa and fungi)[1,2]. From ingested plant material, cellulose, pectin, xylans, xyloglucans, and other polysaccharides are degraded by microbially encoded carbohydrate-active enzymes (CAZymes) down to their component monosaccharide units, which are subsequently fermented into several intermediates. Most importantly, pyruvate is converted to volatile fatty acids (VFAs) such as acetate, propionate, and butyrate[3]. Along this fermentation pathway, hydrogen (H_2) is produced, which predominantly flows into methanogenesis but can also be incorporated into VFAs through alternative hydrogen sinks such as the reduction of fumarate[4]. The rumen epithelial wall is able to transport most of the VFAs directly into the blood, whereas more complex metabolites take a longer path, being assimilated by the posterior gastrointestinal tract[5].

Rumen microbiome structure and function is shaped by many dynamic variables, such as diet, age, health status, animal husbandry, behavior, and breed. Efforts to monitor and predict overall rumen microbiome function for the purpose of improved animal production have up to date mainly focused on recovering isolate and genome representation of the various populations in the rumen. However, the superior amenability of bacteria and archaea to current molecular microbiology techniques has created significant domain-specific information bias, with recovery of greater than 50,000 bacterial and archaeal genomes compared to ~50 for eukaryotic species[1,2]. The ciliate protozoa, specifically the class Litostomatea, subclass Trichostomatia, have a relatively large biomass in the rumen (up to 50%[1]), and are ubiquitous among ruminants. Although single-celled, they have complex organelles and physiological features such as mouthlike adoral openings that lead to a tongue-like extrusible peristome, which ingests feed particles into an esophagus-like structure[6]. This, combined with their outside being covered with undulating cilia for propulsion, makes many of them voracious predators[6]. To add to their versatility, they express carbohydrate-active enzymes (CAZymes) and are able to degrade plant fibers[2]. Decades of in vitro work have shown that rumen ciliates often act as a microhabitat for archaea and bacteria[7], especially *Methanobrevibacter* spp., which form metabolic mutualisms with several ciliate species by recycling the H_2 produced by the ciliates as a main metabolic end product[8,9]. Providing *in vivo* context to the wider ecological impacts of rumen protozoal populations has proven immensely challenging but is necessary to advance microbiome-based solutions to animal productivity and sustainability, for example in the context of methane mitigation.

Rapid advancement of biotechnological tools has improved accessibility of representative data for resident rumen microbiota, yet information on how species interact within these

multidomain ecosystems is still limited. In this study, we applied long-read metagenomics, existing single-cell amplified eukaryotic genomes, and genome-centric multi-omics of both host and its microbiome to improve resolution of inter-domain relationships and the influence they exert at a systems level. Two breeds of cattle from a highly controlled experiment were phenotyped for key performance traits, and rumen, epithelial, and liver samples were analyzed across all molecular layers—genes, transcripts, proteins, and metabolites—by the application of their respective -omics (Fig. 1). Taxonomic analysis identified two clear rumen microbiome structural patterns across the entire animal cohort that were not strongly correlated to breed, any of the recorded animal performance metrics, or methane emissions. Looking deeper across the microbial domains, we identified two distinct protozoal population types that we hypothesize to drive systems-wide microbiome differences, ultimately affecting the interlinked metabolisms that channel the flow of nutrients across the feed-microbiome-host axis.

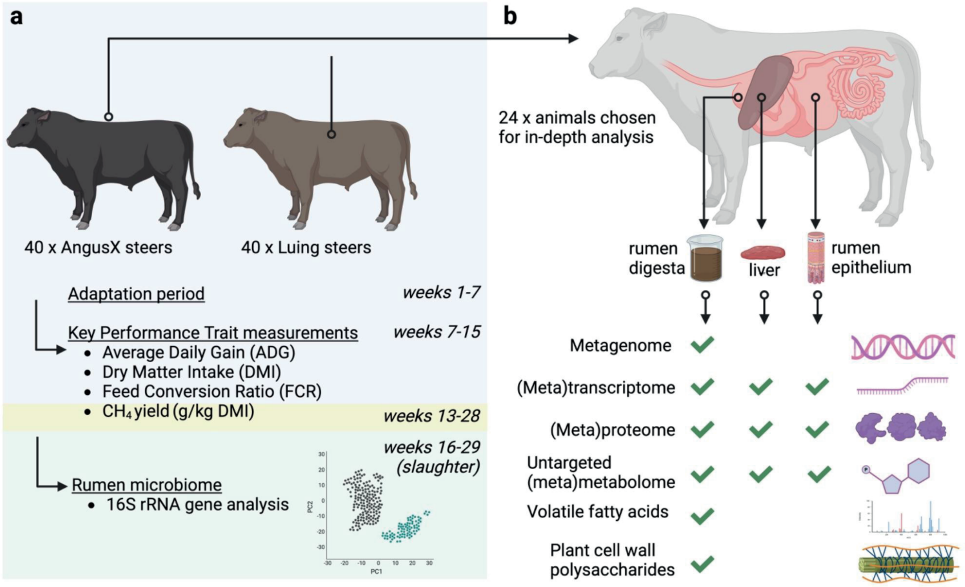
Results

One controlled animal experiment reveals two distinct rumen microbiome patterns

To create greater resolution and depth of understanding within the rumen microbiome, we analyzed samples from a controlled feedlot trial of adult beef cattle fed a total mixed ration of forage and concentrate (ratio: 51:49). From an initial 80 animals representing two breeds that commenced the trial, 36 Aberdeen-Angus cross (AAX) and 35 Luing animals completed the experimental period with all planned measurements, including key performance traits (KPTs) such as dry matter intake (DMI), live weight gain (LWG), feed conversion ratio (FCR) and methane yield (g/kg DMI). For microbiome analysis, rumen samples were taken for all 71 animals at five timepoints across the experimental period and subjected to 16S rRNA gene amplicon sequencing, with an additional final timepoint sampled at slaughter **(Fig. 1a)**. A subset of 24 animals (12 AAX, 12 Luing), representing the highest and lowest natural levels of methane yield, were sampled across both the host and its microbiome at slaughter. The datasets generated from these 24 animals included long-read metagenomics for metagenome-assembled genome (MAG) reconstruction as well as RNA, protein and metabolite analysis of rumen digesta, rumen epithelia and liver tissue **(Fig. 1b)**. As expected, the recorded KPTs showed breed-dependent differences in animal metrics, such as a higher liveweight and dry matter intake (DMI) in AAX animals, and a trend for higher methane emissions (g/kg DMI) in Luing animals **(Fig. 1c)**.

For microbiome characterization, long-read metagenomic sequencing of rumen samples from the 24-animal subset produced a total of 700 high- and medium-quality metagenome-assembled genomes (MAGs, 656 classified as bacterial, 44 as archaeal;

Supplementary Table 1a). These MAGs, together with previously published fungal genomes (n=9)[10] and protozoal single amplified genomes (SAGs) (n=53)[2,11], formed the reference database for metatranscriptomic and metaproteomic analyses. Rumen metatranscriptomics identified 1,669,849 expressed genes (of which 1,299,827 from bacteria, 80,325 from archaea, 252,768 from protozoa, and 9,529 from fungi), whereas metaproteomics identified 35,655 protein groups (16,823 from bacteria, 380 from archaea, 18,000 from protozoa, 137 from fungi, and 315 from the cattle host) (**Supplementary Table 1b**). To further assist our interpretations of host and microbial metabolic activity we generated untargeted metabolomic data from the three different sample types available (numbers of identified metabolites: rumen: 496; rumen epithelium: 517; liver: 859; **Supplementary Table 1b**). Finally, we performed VFA measurements from rumen fluid, as well as Microarray Polymer Profiling (MAPP) of rumen digesta, determining the composition and relative abundance of glycans available to the rumen microbiome.



c

	Animals with amplicon data			Animals with multi-omic data		
	Aberdeen Angus X, N = 36[1]	Luing N = 35[1]	p-value[2]	Aberdeen Angus X, N = 12[1]	Luing, N = 12[1]	p-value[2]
Liveweight (kg)	715 (688, 735)	686 (666, 712)	***0.008***	720 (707, 742)	695 (679, 709)	***0.012***
Age (days) at start of performance test	449 (436, 469)	466 (445, 470)	0.252	455 (435, 470)	466 (440, 469)	0.628
LWG	1.64 (1.51, 2.01)	1.73 (1.43, 2.00)	0.798	1.72 (1.59, 1.99)	1.70 (1.56, 2.11)	0.835
DMI	12.05 (11.59, 12.86)	11.71 (11.01, 12.41)	***0.040***	12.39 (11.82, 13.59)	11.91 (11.05, 12.41)	***0.034***
FCR	7.24 (6.52, 7.91)	6.88 (5.91, 7.87)	0.494	7.26 (6.23, 7.78)	6.81 (5.78, 7.56)	0.383
CH4 (g/kg/DMI)	22.5 (20.5, 25.3)	24.8 (21.8, 27.3)	0.068	19.7 (18.9, 26.1)	26.1 (21.4, 30.6)	0.129

[1]Median (IQR)

[2]Welch Two Sample t-test

Figure 1. Experimental, sampling and data generation design of a controlled beef cattle animal trial. a. Animal experimental setup. A total of 80 animals across two breeds (Aberdeen-Angus cross and Luing) were enrolled of which 71 completed the 4-5 month study period that culminated in their slaughter. Key performance traits such as live weight gain (LWG), dry matter intake (DMI), feed conversion ratio (FCR) and methane yield (g/kg DMI) were measured for all animals and rumen samples periodically collected across the duration of the trial. **b.** Sampling design for a subset of 24 animals, selected on the widest recorded level of natural methane yield variation. At slaughter, three sample locations were collected: Rumen digesta, rumen wall tissue, liver. Samples were characterized on several molecular layers: Genomes, transcripts, proteins, untargeted metabolomics. **c.** Key performance traits and other animal production metrics that were determined for all enrolled animals. IQR: interquartile range. Significant p-values are marked with bold italics.

Against expectations, microbiome analysis of the 71 animals using the 16S rRNA gene sequence data did not reveal clear associations for any alpha or beta diversity metrics with breed, methane yield, or any of the other measured animal KPT (**Extended Data Fig. 1**). However, beta diversity plots illustrated two groups of animals whose microbiome structure distinctly clustered together, which could also be captured using probabilistic modeling (Dirichlet Multinomial Mixtures[12]) (**Fig. 2a**). Surprisingly, these two clusters, hereafter referred to as Rumen Community Type-A and -B (RCT-A and RCT-B), did not correspond to any measured animal KPT nor to any technical grouping that arose from the experimental workflow (**Extended Data Fig. 2a**). Furthermore, these community types were stable across time: the animals consistently stayed in the same cluster over the six timepoints sampled during the experiment (**Extended Data Fig. 2b**).

Curiously, RCT-A and -B were detectable across several omic layers in the 24-animal subset that we analyzed in more detail. Principal coordinates analysis (PCoA) of MAG abundances reflected the same pattern that was detected in the 16S rRNA gene sequence data **(Fig. 2b)**. In Principal Component Analyses (PCA) of digesta and rumen wall epithelium metatranscriptomics as well as digesta metaproteomics, the first principal components (PCs) clearly differentiated RCT-A and -B , thus mirroring the 16S rRNA gene sequence analysis results of the entire animal cohort **(Fig. 2c-d)**. In other words, the microbiome clustering into either RCT-A or -B was the largest contributor of variation across these layers. The congruence between the molecular layers affirmed that elements of metabolism are affected by this compositional difference. Untargeted metabolomics of digesta and the rumen wall epithelium reflected this pattern on PCs 4 and 3, respectively **(Fig. 2e)**. Finally, host proteomics and transcriptomics from wall and liver data also showed a trend towards the two community types, although not always statistically significant ($p < 0.05$ only for PC15 from liver proteomics; $0.1 > p > 0.05$ for other host data; **Fig. 2f)**.

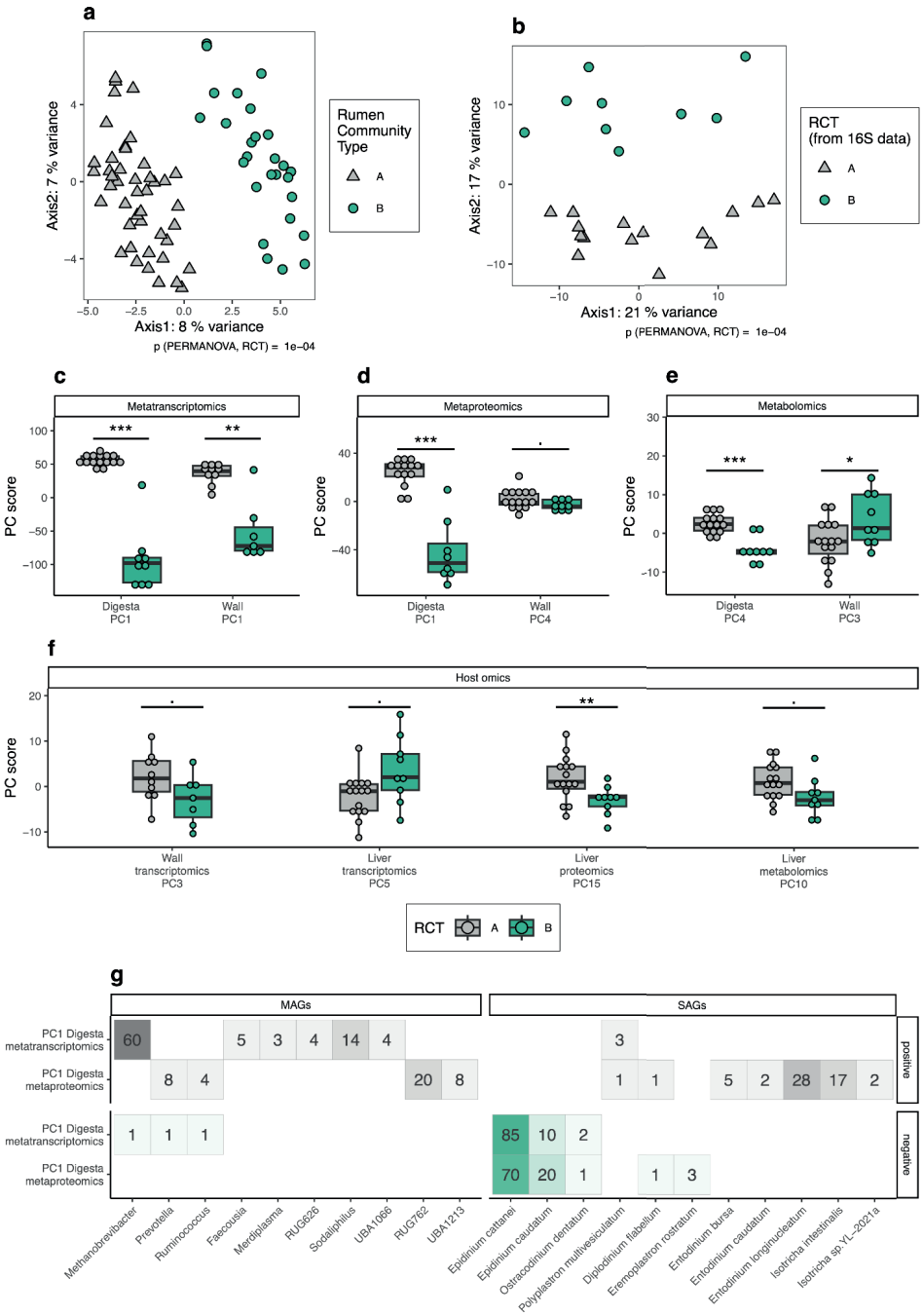


Figure 2. Microbiome analyses revealing two distinct groups of animals, labeled as Rumen Community Type-A and -B. a. Principal Coordinate Analysis (PCoA) plot of robust Aitchison distances of 16S rRNA gene amplicon sequence analysis of 71 animals across two breeds (AAX = 36, Luing = 35), showing the two optimal clusters based on Dirichlet Multinomial Mixtures. **b.** PCoA plot of robust Aitchison distances of shotgun metagenomic data from a subset of 24 animals with the highest and lowest methane yield (g/kg DMI). **c-f.** Dot and box plots of microbiome-derived **c.**

metatranscriptomic, **d.** metaproteomic, and **e.** metabolomic data as well as **f.** host-omic data, showing the first Principal Component (PC) with a significant difference between the two clusters (t-test) for each sample material and measurement type. Box hinges represent 1st and 3rd quartiles, and whiskers range from hinge to the highest and lowest values within 1.5*IQR of the hinge. **e.** Taxonomic summaries of the 100 features with the highest positive and negative loadings for PC1 in digesta metatranscriptomics and -proteomics.

Protozoal patterns associate with rumen community types

To explain the biological drivers causing the system-wide microbiome variation observed as RCT-A or -B, we examined its pattern across the different microbial domains present within the rumen samples of this study, incorporating the archaeal and bacterial MAGs as well as the single amplified genomes (SAGs) for protozoal populations. The taxonomic classifications of the transcripts and proteins with the strongest contributions to the significant principal components from rumen content **(Fig. 2g)** clearly indicated that the RCT-A and -B clustering extended to the abundance profiles of detected protozoal species. Based on these and the differential abundance comparisons of metatranscriptomic and metaproteomic data, animals that exhibited the RCT-A pattern were enriched for families Entodiniinae and Isotrichidae and were defined by the higher abundance of *Entodinium bursa*, *Entodinium caudatum* and *Entodinium longinucleatum*; *Isotricha intestinalis* and *Isotricha YL-2021a*; as well as *Ostracodinium gracile* and *Polyplastron multivesiculatum* (**Fig. 3)**. Conversely, animals with the RCT-B pattern were enriched for subfamilies Diplodiniinae and Ophryoscolecinae[2] and were defined by species *Diplodinium dentatum, Epidinum cattanei*, *Epidinium caudatum,* and *Ophryoscolex caudatus* **(Fig. 3)**.

Coexistence or exclusion patterns of protozoal species have been repeatedly observed over half a century[13]. J. Margaret Eadie first microscopically determined in 1962 that certain genera of protozoa, or “community types”, were detected across select cohorts of animals in both sheep and cattle herds[13]. Protozoal community type A was defined with *Polyplastron multivesiculatum* and other species under genera *Ophryoscolex and Diploplastron* (now Diplodiniinae), while type B was defined by high abundance of *Eudiplodinium* and *Epidinium* spp., either together or alone[13]. Within our microbiome data, we observed protozoal expression patterns (both RNA and protein) that bore remarkable similarities with the definitions of Eadie’s original A/B types. RCT-A animals resembled protozoal community type A with *Polyplastron* detected at significantly higher expression levels, while *Epidinium* spp. were enriched in RCT-B animals considering both metatranscriptomic and metaproteomic comparisons.

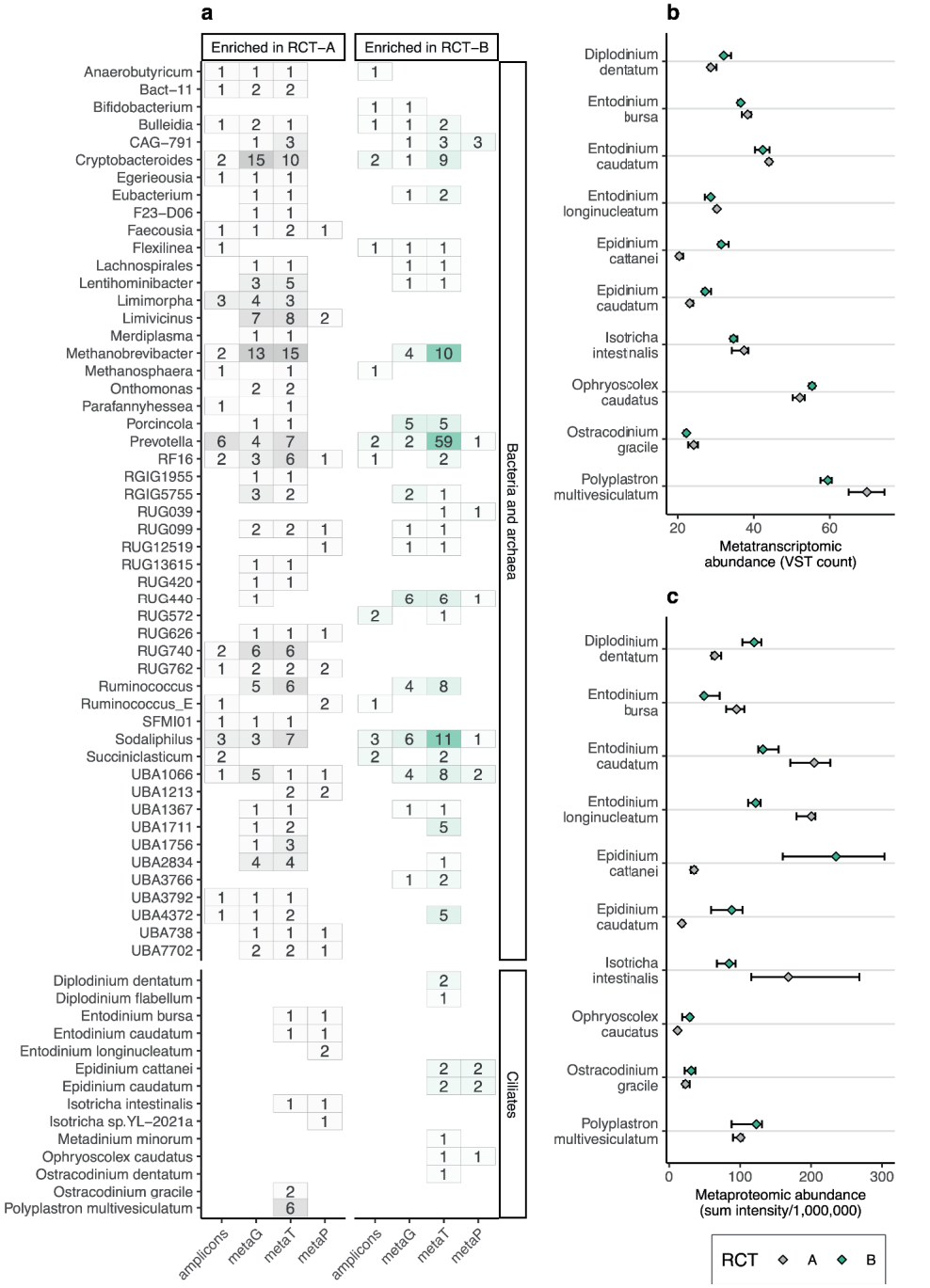


Figure 3. Differential abundances of taxa across omics. a. Summary of differential abundance results. Numbers indicate significantly different taxa (adjusted $p < 0.05$; ASVs or MAGs per genus for bacteria and archaea, SAGs per species for ciliates). For bacteria and archaea, only genera with differentially abundant taxa supported by at least two omics in the same direction are shown. Amplicons: 16S rRNA gene sequence data ASV counts compared with DESeq2; metaG: MAG

relative abundance compared with Wilcoxon Rank Sum tests; metaT: transcript counts summarized per MAG or SAG, compared with DESeq2; metaP: LFQ intensities summarized per MAG or SAG compared with Wilcoxon Rank Sum tests. Multiple comparison correction is included in the testing procedure for DESeq2 (Benjamini-Hochberg FDR); additional FDR correction was implemented for Wilcoxon Rank Sum tests. **b.-c.** Ciliate abundances for the species seen more than once in panel a. **b.** Metatranscriptomics, **c.** Metaproteomics. Diamonds indicate medians, whiskers IQR.

One noticeable difference between previously defined protozoal community types and the microbiome patterns we observed herein was the coexistence of *P. multivesiculatum* and *Epidinium* spp. in RCT-B animals, a scenario previously suggested by others to constitute a type AB protozoal community type[9,14,15]. According to Williams & Coleman, the introduction of *P. multivesiculatum*-containing (A-type) rumen fluid through a rumen cannula into a B-type rumen results in the "complete elimination of *Epidinium* spp. suggesting consequential predatory dynamics of the protozoal populations"[16]. The *P. multivesiculatum* > *Epidinium* spp. dominance was later supported by Kittelmann et al. who postulated that animals observed with both species likely are undergoing a rumen transition from type B to type AB and finally type A (over a 2 week period)[9]. To explore the interrelationships of these species, we examined the metatranscriptomes and metaproteomes of rumen samples collected from six animals over five months. In line with the stability of the bacterial and archaeal community structure (**Extended Data Fig. 2b**), the coexistence patterns of *P. multivesiculatum* and *Epidinium* spp. were consistent over time, indicating a constant low but detectable presence of *P. multivesiculatum* in type B and of *Epidinium* spp. in type A animals (**Extended Data Fig. 3**). This brings to question whether an antagonistic relationship indeed exists between these two protozoal genera.

Protozoal community types affect bacterial and archaeal structure and function

Examining further the metatranscriptome and metaproteome data, we sought to identify the concerted microbial populations that were driving the system-wide variation we observed, and if there exists biological relationships to explain these patterns. We looked at differential expression analysis as well as the features with the strongest loadings in our abovementioned PCA analysis, which both highlighted that specific bacterial, archaeal and protozoal populations were indeed more prevalent in either RCT-A or -B (**Fig. 2e, Fig. 3**). Collectively, for animals categorized as RCT-A, the metaproteomes from their rumen were largely dominated by *Isotricha* spp, *Entodinium* spp, and the clostridial lineage *Acutalibacteraceae* (RUG762) while transcriptomes for various *Methanobrevibacter* spp., *Sodaliphilus* spp., *Faecousia* spp. and *Lachnospiraceae* (UBA1066) were also prevalent. In contrast, both the metatranscriptome and metaproteome for rumen samples from RCT-B animals showed far higher detection of *Epidinium* spp., while high transcript levels were also observed for many *Prevotella* populations (**Fig. 3**).

To link biology to these observed structural patterns, we explored the annotated functions of the differentially detected populations more deeply with specific attention to the key functional stages of rumen digestion, namely fiber hydrolysis, fermentation of organic material and production of energy-yielding volatile fatty acids (**Fig. 4**). By far the most active fibrolytic population observed in RCT-B animals was *Epidinium* spp. which contained a plethora of CAZymes predicted to act upon cellulose, arabinoxylans, beta-mannans and arabinogalactan protein glycans commonly found in grasses and grains (**Fig. 4b**). Epidinia are the most reputable among the rumen ciliates to actively attach and degrade plant material, as visually confirmed across a series of prior studies[16]. In a scenario where epidinia are more proliferant in animals and engaging in plant material deconstruction, it is reasonable to expect their activity and size will impact the glycan landscape that is available for neighboring microbial populations. Indeed, our MAPP analysis of rumen digesta particles was suggestive of differences in various beta-glucan, xylans, xyloglucans, and arabinogalactan proteins between the epidinia-dominated RCT-B animals and the RCT-A animals (**Fig. 5a**). In turn, many fiber-degrading bacterial lineages, such as *Sodaliphilus* and *Prevotella* spp., were detected at higher levels in metatranscriptomic data arising from RCT-B animals (**Fig. 3**), supporting our hypothesis that system-wide effects are driven by protozoal activity. Within RCT-B animals, a higher proportion of butyrate was detected (**Fig. 4c**). This was corroborated by elevated metaproteomic detection of central butyrate-producing enzymes in epidinia species (**Fig. 4b**) as well as a prior meta-analysis of protozoa which calculated that defaunation will substantially decrease ruminal butyrate levels[17].

In the absence of elevated epidinia metabolism within RCT-A animals, both PCA and differential abundance analyses indicated the primary responsibilities for digestion was shared more broadly across the protozoal species *Entodinium* spp. and *P. multivesiculatum* as well as bacteria affiliated to family *Acutalibacteraceae* or genera *Faecousia* and *Merdiplasma* (**Fig. 4a-b**). The *Isotricha* species that dominated RCT-A animals were, as expected[17], not primarily degraders of plant material, though we suspect their influence still impacted heavily upon other bacterial populations. For example, populations affiliated to RUG762 (*Acutalibacteraceae*), had some of the strongest loadings for RCT-A animals within the metaproteomic PCA analysis (**Fig. 2e**) and were differentially abundant across all molecular datasets (16S rRNA, DNA, RNA, and protein) analyzed in this study (**Fig. 3a**). Closer annotation of their metabolic features suggested RUG762 populations were engaged largely in protein and amino acid fermentation, and this was supported by metaproteomic and metabolomic enrichment of the enzymes and metabolites for aspartate, glutamine and branched chain amino acid metabolism in RCT-A animals (**Fig. 4c, Fig. 5b**). Fermentation end products were predicted to be propionate and branched-chain volatile fatty acids, which were also detected at higher proportions in RCT-A animals (**Fig. 4c**). The protein and amino acids for ruminal fermentation could plausibly arise from the grain fraction of the animal's diet

(355 g/kg DM in the concentrate component). However, *Isotricha* spp. have been shown to excrete cellular nitrogen in the form of amino acids, principally alanine, proline, glutamic acid, and aspartic acid[16,18]. If such excretion of amino acids indeed occurs in RCT-A animals dominated by *Isotricha* spp. our observations of elevated RUG762 metabolism are plausibly interlinked, though we acknowledge this hypothesis must be tested in future experiments that examine cellular proximity and nutrient transfer between these populations.

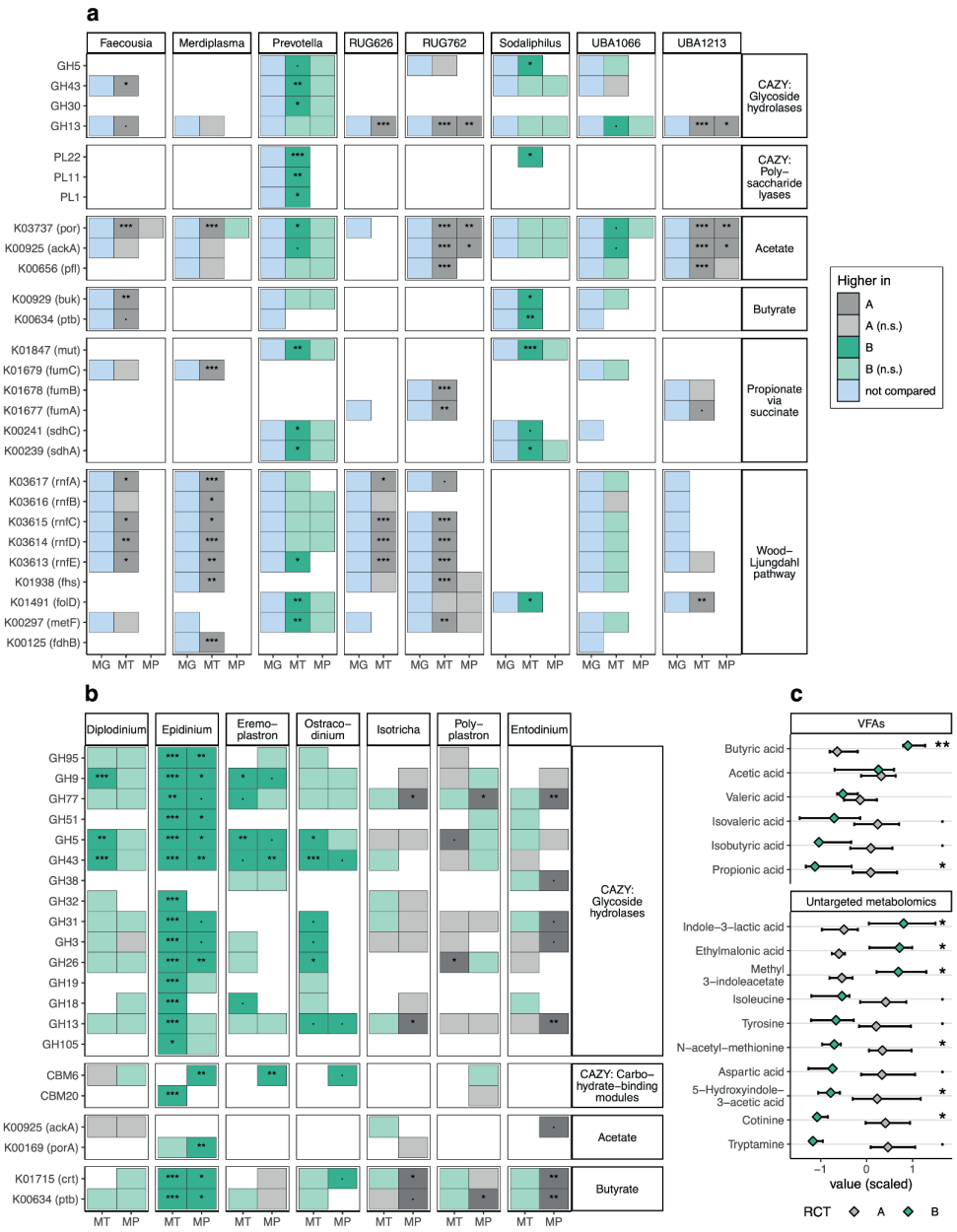


Figure 4. Functional differences between rumen community types. a.-b. Summaries of Kegg Orthologs in pathways of interest and CAZymes in **a.** bacterial and **b.** protozoal genera of interest, selected based on Fig 2e. MG = metagenomic data, blue cells indicating that the function was present in a MAG classified to the genus in question. MT = metatranscriptomic data, stars reflecting adjusted *p*-values from DESeq2. MP = metaproteomic data, stars reflecting fdr-adjusted *p*-values from Wilcoxon Rank Sum tests. Empty spaces indicate that the function was not detected at all in the genus in question. **c.** Summary of metabolite measurements. Volatile Fatty Acid (VFAs) panel shows all measured VFAs, untargeted metabolomics panel only metabolites with annotation level 1 or 2a and a multiple comparison corrected *p*-value < 0.1. Diamonds indicate medians, whiskers IQR. In all panels, *** : $p < 0.001$, ** $p < 0.01$, * : $p < 0.05$, · : $0.1 > p > 0.05$.

It was interesting to note that for RCT-A animals a grouping of *Methanobrevibacter*-affiliated populations were detected at higher abundance and/or with loadings clearly associating them with RCT-A in our PCA analysis, despite there not being significant differences in measured methane yield across the two groups of animals (**Extended Data Fig. 2a**). The holotrich *Isotricha* species have been repeatedly shown[17] to associate with different methanogenic populations than entodiniomorphids (e.g. epidinia), and our data also followed this trend with *Methanobrevibacter* populations in epidinia-dominated RCT-B animals seemingly of distinct strains compared to RCT-A (**Fig. 3a, Extended Data Fig. 3**). Functional examination of bacterial populations enriched in RCT-A animals (**Fig. 3a**) identified several taxa, including *Faecousia* and *Merdiplasma* species, that were predicted to encode uptake hydrogenases and/or the Wood-Ljungdahl pathway (WLP) (**Fig. 4a**). This pathway potentially facilitates reductive acetogenesis and acts as an alternative hydrogen sink to methanogenesis[19]. Indeed, aforementioned RUG762 populations were also suspected to encode a partial WLP as well as the associated energy conservation machinery such as the electron-bifurcating hydrogenase, ferredoxin:NAD-oxidoreductase (Rnf) complex, and F_OF_1 ATP synthase (**Fig. 5b**). However, rather than complete reduction of CO_2 to acetate via the acetyl-CoA synthase/carbon monoxide dehydrogenase complex, we suspect RUG762 populations are instead producing methionine via a cobalamin-dependent 5-methyltetrahydrofolate–homocysteine methyltransferase. Under normal rumen conditions, reductive acetogens are believed to be outcompeted energetically by methanogens, except for instances of higher hydrogen partial pressure[19]. In this context, and in the absence of measured hydrogen levels from our samples, we speculate that non-differential methane yield levels across the RCT-A and -B animals are the result of increased competition for hydrogen across methanogens and other hydrogenotrophs in the rumen. Additional support for this hypothesis comes from our observations that *Methanosphaera* spp. were also differentially observed at higher levels in RCT-A animals (**Fig. 3a**), which follows previous studies that show this group of methanol-utilizing methanogens is detected in animals with elevated hydrogen levels[20].

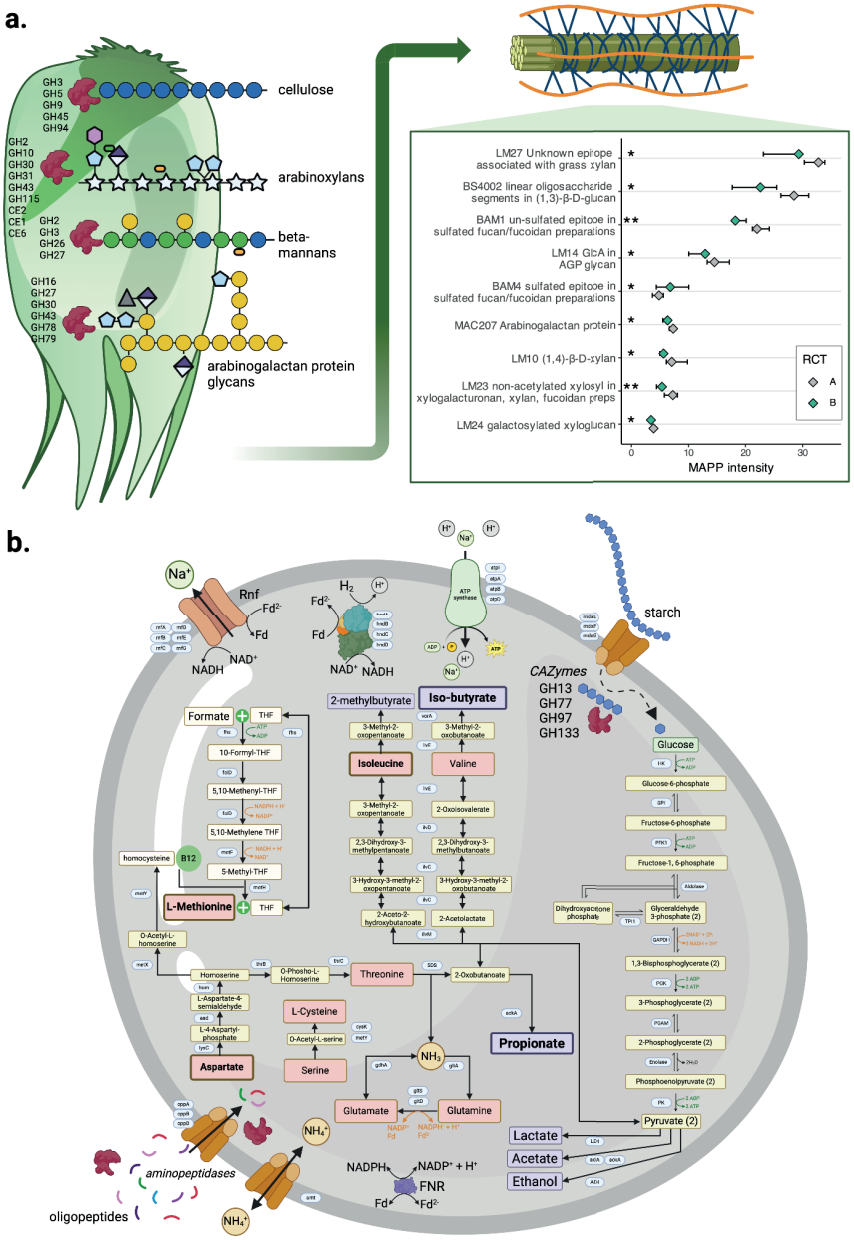


Figure 5: Metabolic predictions from major populations strongly featured in RCT-A and -B animals that are predicted to influence rumen function. a. Illustration of *Epidinium cattanei*, the protozoal species most strongly associated with RCT-B, predicted to engage a broad range of CAZymes to degrade dietary plant material. Given the size and activity of *E. cattanei* their fiber degrading metabolism is predicted to impact the rumen structure and function of RCT-B animals, which was supported via Microarray Polymer Profiling (MAPP) of various hemicellulosic plant fibers, which highlighted differential abundances. In the MAPP inset, diamonds indicate medians, whiskers interquartile ranges, and stars represent uncorrected *p*-values from Wilcoxon rank sum tests. **b.** Pathway reconstruction for the *Acutalibacteraceae*-affiliated RUG762 population, strongly associated

with RCT-A, highlighting amino acid (red boxes) metabolism via fermentative and a partial Wood-Ljungdahl Pathway, which was supported by the associated energy conservation machinery such as the electron-bifurcating hydrogenase, ferredoxin:NAD-oxidoreductase (Rnf) complex, and FoF1 ATP synthase. Bold text indicates differentially abundant metabolites from Fig 4.

Implications for the host animal

Despite the distinct systems-wide microbiome shifts that were reconstructed in RCT-A and -B animals, we were surprised to observe limited data that would suggest these structural and functional effects are being passed onward to the host animals. This was apparent in animal performance measurements (**Extended Data Fig. 2a**), microbial and host metabolomic data as well as host expression data in gut epithelial and liver tissues, which showed only minor changes to a limited number of features (**Fig. 3**). The clearest difference was the relative composition of several amino acids and VFAs, with propionate and branched chain volatile fatty acids higher in RCT-A animals, while butyrate levels were higher in RCT-B (**Fig. 4c**). Since VFAs are the major energy source for the host animal and are taken up directly through the rumen wall epithelium[3], we further applied a series of network analyses using rumen and epithelial proteomic data to ascertain if underlying expression patterns were indeed evident between metabolically linked microbial and host pathways. From rumen metaproteomes, weighted gene correlation network analysis (WGCNA)[21] identified a wide variety of co-expression modules (ME) that contained mixtures of protozoal, bacterial and archaeal proteins; many of these modules were, unsurprisingly, strongly correlated with the RCT variable (**Extended Data Fig. 5**). In the epithelial proteomics data, WGCNA identified only two co-expression modules, comprised largely of host proteins, that were correlated with the RCT groupings, none of which were enriched with proteins functionally inferred in VFA metabolism (**Extended Data Fig. 6**). Of note, interlinked patterns of rumen digesta (ME9 and ME13) and epithelial (ME1) modules were enriched in proteins annotated in cysteine and methionine metabolism and RUG762 populations suggesting possible metabolic interplay of amino acids, though this needs future testing for validation. The lack of striking host effects arising from microbiome differences in RCT-A and -B animals highlights the extraordinary plasticity of the rumen microbiome and its ability to absorb structural variation that on the surface would appear to inflict real functional impact.

Discussion

Rumen protozoa are large and complex compared to their bacterial and archaeal neighbors and their presence and distribution within the livestock rumen has been heavily documented for well over 130 years[16]. Despite their long-standing history their impact across the total rumen ecosystem remains poorly understood at a molecular level due to technical restrictions that have impeded their study, and which have only recently been overcome with

omics methodologies. Herein we were excited to link the molecular patterns and functional interpretations in our data to community types first postulated over 60 years ago via light microscopy[13]. When first describing protozoal community types in 1962 J. Margaret Eadie explicitly stated: *"It is concluded that inter-relationships of the type described may play an important role in determining the components of a particular rumen microfauna."*[13]. We show that for the animals in this study, the system-wide rumen microbiome structure indeed extended beyond the protozoal components originally proposed in community types A and B to encompass bacterial and archaeal populations.

Advancing the original Eadie hypothesis, our multi-layered omic datasets offered plausible interpretations on how two independent modes of metabolic interactions are interlinked across the rumen microbiome of RCT-A and -B animals. Of particular note was the seemingly direct influence certain protozoal species (e.g. *Epidinium* spp) play at higher trophic levels such as fiber hydrolysis, which likely impacts fiber structural configuration and availability for bacterial fibrolytic populations. On the other hand, protozoal metabolism of *Isotricha* spp. was predicted to indirectly affect how nutrients enter the food chain via excretion of metabolites such as amino acids and hydrogen, which impacted the structure and function of intermediate fermenters. While this study goes some way into explaining the microbiome-wide effects that particular protozoa can exert, major questions regarding the origin of their structural configuration still remain. We speculate the original seeding took place via animal-animal contact likely during early life transition that started with mother-calf contact and gradually extended to other animals across the greater herd. Unfortunately, behavioral data prior to animal enrolment and pen groupings used in this animal trial were not recorded, though it was clear that grouping of RCT-A and -B animals together in randomized pens had no immediate nor long term influence upon microbiome structure.

It is without question that our knowledge of the rumen microbiome has rapidly improved with advancements in (meta)genome technologies, increases in mass spectrometry sensitivity and evolution of computational methods that can accurately reconstruct high quality genomes from complex microbial assemblages. Moreover, we show that the acceleration in genome recovery of protozoal populations and their supplementation into rumen microbiome databases has massively impacted our ability to estimate the transdomain microbial trophic cascades that convert complex plant material into energy-yielding nutrients for the host animal. Moving forward, several outstanding knowledge gaps need to be prioritized so that greater microbiome resolution can be routinely gained. Laboratory-based experiments that validate both proximity and metabolic interactions between protozoa, bacteria and archaea will lead to improved interpretations of how protozoa modulate rumen biology and formulate tools to potentially intervene where desired. Finally, more extensive surveys of increased animal numbers, varying diets, breeds and management practices will need to be analyzed at a depth comparative to the present study to ascertain the wider implications of

protozoal-bacterial-archaeal interactions, and how that knowledge can be applied to improve microbiome modulation strategies that make meaningful impact.

Material and Methods

Ethics statement

The animal experiment was conducted at the Beef and Sheep Research Centre of Scotland's Rural College (6 miles south of Edinburgh, UK). The experiment was approved by the Animal Experiment Committee of SRUC and was conducted in accordance with the requirements of the UK Animals (Scientific Procedures) Act 1986.

Experimental design and measurement of key performance traits

An initial group of 80 animals representing two breeds of beef cattle (Aberdeen-Angus cross (AAX, n = 40), and Luing (n = 40) was selected for the experiment. Of these, 71 (AAX: n = 36; Luing: n = 35) successfully completed the designed sampling scheme. All animals were provided a typical basal diet consisting of whole crop barley (300 g/kg DM), grass silage (200 g/kg DM), barley (355 g/kg DM), maize dark grains (120 g/kg DM), molasses (15 g/kg DM) and minerals (10 g/kg DM). For half of the animals, the experimental design originally involved supplementation with *Asparagopsis taxiformis* red algae vegetative tissue (thallus) at 0.3% of the organic mass (OM). *A. taxiformis* is a feed additive which has been shown to reduce methane emissions in past studies[22–25]. However, due to adverse effects observed in animals during the planned three-week seaweed adaptation phase, supplementation was terminated after just 15 days. All animals were given a further 5 weeks to adapt to basal feed before performance testing was carried out. Due to this delay, the heaviest 32 animals, balanced for breed, underwent a shorter performance test period of 4 weeks instead of the normal 8 weeks. During performance testing, daily feed intake was recorded using electronic feeders (HOKO, Insentec, Marknesse, The Netherlands). Twice weekly, duplicated samples of each diet component were collected to determine dry matter content and to calculated dry matter intake (DMI). Body weight of each animal was measured weekly to estimate average daily gain (ADG) using a linear regression model including time on test. Feed conversion ratio (FCR) was calculated for each animal as average daily DMI divided by ADG.

At the end of the experimental period, the animals' methane emissions were measured in respiration chambers. One week prior to entering the respiration chambers, the cattle were single-housed in training pens, identical in size and shape to the pens inside the chamber, to adapt to individual housing. The cattle were allocated to six respiration chambers based on

the criterion of minimisation of the variation in body weight. They remained in the respiration chambers for 3 days, which included one day for adaptation and a 48-hours measurement period for methane emissions.

Of the 71 animals that completed the trial, 24 were selected for multi-omic analysis, including equal numbers of the two breeds, and representing the full range of methane emissions. For a further subset of six animals (out of 24), samples were also analyzed for a time series collected during the experimental period using orogastric tubing, as described below.

Rumen content and tissue sample collection

On live animals, longitudinal rumen fluid samples (50 ml) were collected using a stomach tube (16×2700 mm Equivet Stomach Tube; Jørgen Kruuse A/S, Langeskov, Denmark) nasally and aspirating manually. Samples were collected prior to the adaptation phase to seaweed, before and after the performance test as well as immediately after leaving the respiration chambers. Additionally, rumen fluid samples (50 ml) were obtained after the animals were slaughtered in a commercial abattoir, immediately after the rumen was opened to be drained. Immediately after sampling, the rumen digesta was filtered through two layers of muslin and a 5 ml sample of the filtered liquid was transferred into a 30ml universal containing 10 ml of PBS-Glycerol, then stored in a freezer at -80°C.

Rumen cell wall samples were collected from the central region of the ventral sac before the rumen had been washed. The ruminal tissue was dipped into a 125 ml beaker containing a PBS solution to remove the ruminal digesta. The tissue was sliced using a sterile scalpel and transferred to a 30 ml universal tube containing 5 ml RNALater. Additionally, liver samples were taken by the meat inspector, with a section cut out using a sterile scalpel and then stored in a 30 ml universal tube with 5 ml RNALater. All tissue samples were stored in a freezer at -80 °C before being analyzed. Further details regarding the sampling and experimental procedures carried out at SRUC can be found in previously published studies[26,27] which followed a similar protocol.

16S rRNA gene amplicon sequence data

Rumen digesta sample DNA extraction, PCR amplification and sequencing of 16S rRNA gene amplicons was performed at DNASense ApS (Aalborg, Denmark).

Sample DNA extraction

Rumen digesta DNA was extracted using the FastDNA Spin kit for Soil (MP Biomedicals, USA) with the following exceptions to the standard protocol: 500 µL of sample, 480 µL Sodium Phosphate Buffer and 120 µL MT Buffer were added to a Lysing Matrix E tube. Bead

beating was performed at 6 m/s for 4x40s. Gel electrophoresis using Tapestation 2200 and Genomic DNA screentape (Agilent, USA) was used to validate product size and purity of a subset of DNA extracts. DNA concentration was measured using Qubit dsDNA HS/BR Assay kit (Thermo Fisher Scientific, USA).

Sequencing library preparation

Amplicon libraries for the 16S rRNA gene variable region 4 (abV4-C) were prepared using a custom protocol based on an Illumina protocol[28]. Up to 10 ng of extracted DNA was used for PCR amplification. Each reaction (25 µL) contained (12.5 µL) PCRBIO Ultra mix and 400 nM of each forward and reverse tailed primer mix. The PCR program was as follows: initial denaturation at 95 °C for 2 min, 30 cycles of amplification (95 °C for 15 s, 55 °C for 15 s, 72 °C for 50 s) and a final elongation at 72 °C for 5 min. Duplicate reactions were performed for each sample and the duplicates pooled afterwards. The primers targeting the abV4-C region were the following, designed according to[28] : [515FB] GTGYCAGCMGCCGCGGTAA and [806RB] GGACTACNVGGGTWTCTAAT[29], with tails that enable attachment of Illumina Nextera adaptors necessary for sequencing in a subsequent round of PCR. The amplicon libraries were purified using the standard protocol for CleanNGS SPRI beads (CleanNA, NL) with a bead to sample ratio of 4:5. DNA was eluted in 25 µL of nuclease free water (Qiagen, Germany). DNA concentration was measured using Qubit dsDNA HS Assay kit (Thermo Fisher Scientific, USA). Gel electrophoresis using Tapestation 2200 and D1000/High sensitivity D1000 screentape (Agilent, USA) was used to validate product size and purity of a subset of libraries.

Sequencing libraries were prepared from purified amplicon libraries using a second PCR. Each reaction (25 µL) contained PCRBIO HiFi buffer (1x), PCRBIO HiFi Polymerase (1 U/reaction) (PCRBiosystems, UK), adaptor mix (400 nM of each forward and reverse) and up to 10 ng of amplicon library template. PCR was done with the following program: initial denaturation at 95 °C for 2 min, 8 cycles of amplification (95 °C for 20 s, 55 °C for 30 s, 72 °C for 60 s) and a final elongation at 72 °C for 5 min. The resulting libraries were purified following the same protocol as above for the first PCR.

DNA sequencing

The purified sequencing libraries were pooled in equimolar concentrations and diluted to 2 nM. The samples were paired-end sequenced (2x300 bp) on a MiSeq (Illumina, USA) using a MiSeq Reagent kit v3 (Illumina, USA) following the standard guidelines for preparing and loading samples on the MiSeq. > 10 % PhiX control library was spiked in to overcome low complexity issues often observed with amplicon samples.

Sequence data analysis

Quality trimming and amplicon sequence variant (ASV) inference for the 16S rRNA gene amplicon sequence data was performed with dada2[30] following the recommended Big Data Paired-end workflow[31] using default parameters, except for the following choices for the filterAndTrim step: truncLen = 240 for forward, 200 for reverse reads; trimLeft = 20 for forward, 30 for reverse reads; maxEE = 2, and truncQ = 6. The reference database for taxonomic classification was the dada2 formatted version of release 214 of the Genome Taxonomy Database (GTDB)[32].

Metagenomics

DNA extraction and sequencing as well as initial metagenomic sequence data analysis for rumen digesta samples was performed at DNASense ApS (Aalborg, Denmark).

DNA extraction

DNA intended for sequencing on the Illumina platform was extracted during the workflow for 16S rRNA gene amplicon data, as described above. DNA intended for ONT sequencing was extracted with the DNeasy PowerSoil Kit (Qiagen, Germany) and further cleaned with the DNeasy PowerClean Pro Cleanup Kit (Qiagen, Germany). A custom SPRI (Solid Phase Reversible Immobilization) short fragment removal step was implemented to remove fragments shorter than approximately 1500-2000 bp. DNA concentration and purity was assessed using the Qubit dsDNA HS Assay kit (Thermo Fisher Scientific, USA) and the NanoDrop One device (Thermo Fisher Scientific, USA). DNA size distribution was evaluated using the Genomic DNA ScreenTape on the Agilent Tapestation 2200 (Agilent, USA).

Illumina sequencing

Extracted DNA was fragmented to approximately 550 bp using a Covaris M220 with microTUBE AFA Fiber screw tubes and the settings: Duty Factor 10 %, Peak/Displayed Power 75W, cycles/burst 200, duration 40s and temperature 20 °C. The fragmented DNA was used for metagenome preparation using the NEB Next Ultra II DNA library preparation kit. The DNA library was paired-end sequenced (2 x 150 bp) on a NovaSeq S4 system (Illumina, USA).

Oxford Nanopore sequencing

SQK-LSK114 sequencing libraries were prepared according to manufacturer recommendations with a minor custom modification to allow for native barcoding using kits EXP-NBD104 and EXP-NBD114 (Oxford Nanopore Technologies, Oxford, UK). Briefly; before initiating the SQK-LSK114 protocol, native barcodes were ligated onto end-prepped

sample DNA (100-200 fmol) using NEB Blunt/TA ligase mastermix (New England Biolabs, USA). Approximately 10-20 fmol barcoded DNA library were loaded onto primed PromethION FLO-PRO114M (R10.4.1) flow cells and sequenced on the PromethION P2 Solo device running MinKNOW Release 22.07.3 (MinKNOW Core 5.3.0-rc3-p2solo).

Data preprocessing

Raw Illumina reads were filtered for PhiX using Usearch11[33] and trimmed for adapters using cutadapt[34] (v. 3.5). Forward and reverse read files were concatenated using a custom python script. Raw Oxford Nanopore fast5 files were basecalled and demultiplexed in Guppy v. 6.1.15 using the dna_r10.4.1_E8.2_400bps_sup algorithm. Adapters were removed in Porechop v. 0.2.4 using default settings. NanoStat v.1.4.0[35] was used to assess quality parameters of the basecalled data. The adapter-trimmed data was then filtered in Filtlong v. 0.2.1 with –min_length set to 1500 bp and –min_mean_q set to 96 (q-score of 14).

Metagenome de novo assembly

Draft *de novo* co-assembly for metagenomes was performed in six groups of samples/animals (combinations of control and treatment, corresponding to the seaweed supplementation, and a three-category methane variable representing low, medium and high emission levels) using Flye (v.2.9.1-b1780[36]) by setting the metagenome parameter (–meta). Draft metagenomes were first polished with Medaka (v.1.7.1) using quality-filtered Oxford Nanopore R10.4.1 data, following further polishing with minimap2 (v. 2.24-r1122[37]) and racon[38] (v.1.5.0) using Illumina data covering the relevant metagenome sample trajectory.

MAG binning

Each metagenome assembly was subjected to independent and automated genome binning using Metabat2 v. 2.15[38] and Vamb[39] (v. 4.1.1). MAGs from each metagenome were subsequently dereplicated using dRep[40] (v. 3.2.2) setting minimum MAG length to 250000 bp (-l). All dereplicated MAGs from each metagenome assembly were finally pooled and dereplicated again (cross-dereplicated) with dRep.

Hybrid Metagenomic Assembly and Binning

Samples containing paired short-read and nanopore metagenomes were processed using a hybrid assembly approach, followed by MAG recovery through the Aviary[41] v0.5.7 pipeline (https://github.com/rhysnewell/aviary) using the recover workflow with default settings. The resulting assemblies were manually inspected using Bandage to identify and verify closed genomes. A total of 4,469 redundant MAGs were recovered. Completeness and contamination rates were assessed with CheckM2[42] v1.0.1 using the lineage wf command.

Only MAGs with >70% completeness and <10% contamination were retained for further analysis. To address potential multi-mapping issues during meta-omic relative abundance calculations, the genomes were dereplicated using a custom script. Pairwise Average Nucleotide Identity (ANI) values were calculated for all MAGs using Skani[43]. Genomes with >97% ANI and >50% alignment were clustered using complete linkage clustering. The highest-quality MAG within each cluster was selected as the representative genome. The quality score was calculated using the following metric: completeness - 5*contamination - 5*num_contigs/100 - 5*num_ambiguous_bases/100000, as described by Parks et al. (2020)[44]. This clustering process was iteratively repeated until no further clustering of representative MAGs was possible. This resulted in a nonredundant set of 700 MAGs.

Rumen microbial genome database for metatranscriptomics and metaproteomics

For metatranscriptomic and metaproteomic data analyses, we built databases consisting of 6 parts representing different sources and taxonomic domains:

A. 700 MAGs assembled from our digesta samples, both archaea and bacteria.
B. Bos taurus host genome ARS-UCD1.3[45] GCF_002263795.2 (NCBI BioprojectPRJNA391427).
C. Entodinium caudatum genome[46] (NCBI Bioproject PRJNA380643).
D. 52 protozoal SAGs genomes[2] (NCBI Bioproject PRJNA777442).
E. 9 fungal genomes from phylum neocallimastigomycota[10].
F. 14 bacterial genomes of genus Campylobacter genomes[47].

This total rumen microbial genome database consisted of 4.2 M proteins with an average length of 426.8 amino acids totalling 1.8 G amino acid letters.

Annotation of genomes and characterization of proteins

For the 700 recovered MAGs (A) and 14 *Campylobacter* spp. genomes (F), Prokka[48] v1.14.6 was used for annotation and to translate the coding sequences. CheckM2[49] v1.0.2 was used for assessment of completeness and contamination. The remaining database parts (B–E) were downloaded as amino acid sequences. Translated genes of the complete rumen genome database (A–F) were characterized functionally using eggnog-mapper[50] v2.1.12, resulting in the identification of PFAM[51], CAZy[52], and KEGG[53] orthologs. Pathway enrichment analysis was calculated using the KEGG orthologs and KEGG pathway database[54] (downloaded on 2023-08-28) via clusterProfiler[55] v4.10.0. Taxonomic identification of MAGs were done with GTDB-tk[56] v2.4.0 using database r214. The genomic characterization tools mentioned above were run via CompareM2[57] v2.11.1. For screening of metabolic capacities, DRAM v1.4 was used using default settings.

Meta- and host transcriptomics

RNA extraction and sequencing for rumen digesta, wall and liver samples, as well as bioinformatic analyses for rumen wall and liver sequence data, were performed at DNASense ApS (Aalborg, Denmark).

RNA extraction

RNA extraction for rumen digesta, rumen wall and liver samples was performed with the standard protocol for RNeasy PowerMicrobiome Kit (Qiagen, Germany) with minor modifications: custom reagent volumes were used, PM4 buffer was replaced with 70 % ethanol in initial extraction mix, and bead beating was performed at 6 m/s for 4x40s. Gel electrophoresis using Tapestation 2200 and RNA screentape (Agilent, USA) was used to validate product integrity and purity of RNA extracts. RNA concentrations were measured using Qubit RNA HS/BR Assay kit (Thermo Fisher Scientific, USA). The extracted RNA was treated with the TURBO DNAfree (Thermo Fisher Scientific, USA) to ensure removal of all DNA in the samples. Afterwards the RNA was quality controlled using RNA screentape (Agilent, USA) and Qubit RNA HS/BR Assay kit (Thermo Fisher Scientific, USA).

Sequencing library preparation

RNA extracts were rRNA depleted using the Ribo-Zero Plus rRNA Depletion Kit (Illumina, USA), and residual DNA from RNA extraction was removed using the DNase MAX kit (MoBio Laboratories Inc.). The samples were purified using the standard protocol for CleanPCR SPRI beads (CleanNA, NL) and further prepared for sequencing using the NEBNext Ultra II Directional RNA library preparation kit (New England Biolabs). Library concentrations were measured using Qubit HS DNA assay and library DNA size estimated using TapeStation with D1000 ScreenTape. The samples were pooled in equimolar concentrations and sequenced (2 x 150 bp, PE) on a Novaseq platform (Illumina, USA). All kits were used as per the manufacturer's instructions with minor modifications.

Host transcriptome mapping

Forward and reverse sequencing cDNA reads were quality-filtered and trimmed for Illumina adapters using Cutadapt v. 3.7[58] used in paired-end mode. For liver and rumen wall data, the reads were subsequently mapped against the Bos Taurus Genome Reference ARS-UCD1.3 (Genbank assembly accession GCA_002263795.3). The genome and its associated gene transfer format file (GTF) were downloaded and indexed using STAR v2.7.10a[59], setting a sjdbOverhang of 149 bp. Adapter-trimmed sample reads were mapped against the indexed genome of ARS-UCD1.3 using STAR v2.7.10a in paired-end mode, with the option -outReadsUnmapped Fastx enabled. Alignments were ported to coordinate-sorted BAM

files, and FeatureCounts v2.0.1 of the SubRead package[60] was used to quantify CDS mappings as counts. Where nothing else is stated, the default settings were used for all bioinformatic tools.

Rumen wall metatranscriptome mapping

For rumen wall samples, the forward and reverse cDNA reads that did not map against the *Bos taurus* genome were bioinformatically depleted for rRNA using Ribodetector v. 0.2.7[61] and then mapped against the rumen MAGs. Prior to mapping, the concatenated MAGs were indexed using STAR[62] v2.7.10a. The rRNA-depleted and quality filtered DNA reads were mapped against the MAGs with STAR, setting alignIntronMax to 1. All alignments were ported to coordinate-sorted BAM files.

Rumen content metatranscriptomics

Rumen content data were mapped against the Bos taurus genome (Genome Reference ARS-UCD1.3) using minimap2 v 2.2. All non-paired mapped reads were retrieved using samtools v 1.17[63] with the following parameters samtools fastq -f 12 -F 256 -c 7 -1 read1.fq.gz -2 read2.fq.gz. rRNA reads present in the samples were bioinformatically removed using SortMeRNA v 4.3.6[64] with the following SILVA databases: silva-bac-16s-id90, silva-arc-16s-id95, silva-bac-23s-id98, silva-arc-23s-id98, silva-euk-18s-id95 and silva-euk-28s-id98, and the parameters –out2–paired_out –fastx–thread 64. These reads were used to quantify the expression of coding sequences (CDS) encoded in the rumen microbial genome database using Kallisto[65] v0.50.0. The resulting 'raw-counts' tables were gathered into a single table using the Bioconductor tximport[66] v1.26.1 library in R 4.2.2.

Meta- and host proteomics

Proteomic and metaproteomic measurements and all bioinformatic analyses for rumen digesta, wall and liver samples were performed at the Norwegian University of Life Sciences (NMBU; Ås, Norway).

Protein extraction and digestion

Protein extraction was performed following a previously published protocol[1]. Briefly, 300 µL of rumen fluid sample or a representative amount of wall or liver tissue sample was combined with 150 µL lysis buffer (30 mM DTT, 150 mM Tris-HCl (pH = 8), 0.3% Triton X-100, 12% SDS) and 4 mm glass beads (≤160 µm), then vortexed and rested on ice for 30 mins. Sample lysis was performed with a FastPrep-24 Classic Grinder (MP Biomedical, Ohio, USA) for 3 × 60 s at 4.0 m/s[67], followed by centrifugation at 16,000 × g for 15 min at 4 °C. Lysate was removed and its absorbance measured at A750 on BioTek Synergy H4

Hybrid Microplate Reader (Thermo Fisher Scientific Inc., Massachusetts, USA). 40–50 µg of protein was prepared in SDS-buffer, heated in a water bath for 5 min at 99 °C, and analyzed by SDS-PAGE with Any-kD Mini-PROTEAN TGX Stain-Free gels (Bio-Rad, California, USA) in a 2 minute run for sample clean-up, before staining with Coomassie Blue R-250. Visible bands were excised and divided into 1 mm^2 pieces before reduction, alkylation and trypsin digestion. Peptides were concentrated and eluted using C18 ZipTips (Merck Millipore, Darmstadt, Germany) following manufacturer's instructions.

Mass spectrometry

The peptide samples were analyzed by coupling a nano UPLC (nanoElute, Bruker) to a trapped ion mobility spectrometry/quadrupole time of flight mass spectrometer (timsTOF Pro, Bruker). The peptides were separated by a PepSep Reprosil C18 reverse-phase (1.5 µm, 100Å) 25 cm X 75 µm analytical column coupled to a ZDV Sprayer (Bruker Daltonics, Bremen, Germany). The temperature of the column was kept at 50°C using the integrated oven. Equilibration of the column was performed before the samples were loaded (equilibration pressure 800 bar). The flow rate was set to 300 nl/min and the samples were separated using a solvent gradient from 5 % to 25 % solvent B over 70 minutes, and to 37 % over 9 minutes. The solvent composition was then increased to 95 % solvent B over 10 min and maintained at that level for an additional 10 min. In total, a run time of 99 min was used for the separation of the peptides. Solvent A consisted of 0.1 % (v/v) formic acid in milliQ water, while solvent B consisted of 0.1 % (v/v) formic acid in acetonitrile.

The timsTOF Pro was run in positive ion data dependent acquisition PASEF mode with the control software Compass Hystar version 5.1.8.1 and timsControl version 1.1.19 68. The acquisition mass range was set to 100 – 1700 m/z. The TIMS settings were: 1/K0 Start 0.85 V·s/cm2 and 1/K0 End 1.4 V·s/cm2, Ramp time 100 ms, Ramp rate 9.42 Hz, and Duty cycle 100 %. The Capillary Voltage was set at 1400 V, Dry Gas at 3.0 l/min, and Dry Temp at 180 °C. The MS/MS settings were the following: number of PASEF ramps 10, total cycle time 0.53 sec, charge range 0-5, Scheduling Target Intensity 20000, Intensity Threshold 2500, active exclusion release after 0.4 min, and CID collision energy ranging from 27-45 eV.

Data analysis

The raw spectra were analyzed using mspipeline1[68] v2.0.0 based on FragPipe[69] v19.1. Using Philosopher[70] v4.8.1, MSFragger[71] v3.7 and IonQuant v1.8.10. Spectra were analyzed slicing the rumen microbial genome database into 16 parts using msfragger.misc.slice-db=12. Mass calibration was disabled with msfragger.calibrate_mass=0. The maximum length of peptides to be generated during in-silico digestion was 35 with msfragger.digest_max_length=35. Allowed number of missed cleavages 1 and 2 was set to 1 with msfragger.allowed_missed_cleavage_{1,2}=1.

Otherwise, default settings were used. The processing was performed on an AMD x86-64 “Threadripper Pro” 5995WX 64 cores, 8 memory channels, 512GiB DDR4 3200MHz ECC (8x 64 GiB) and 4 2TB SSDs in raid0.

Proteomic intensities were log2-transformed prior to any statistical analysis. Genes in the proteomic database were annotated using eggnog e-mapper v2.1.12 using CompareM2 v2.11.1. Missing values were imputed using missRanger[72] v2.6.0.

Untargeted metabolomics

Untargeted metabolomic analyses for rumen digesta, rumen wall, and liver samples were carried out by MS-Omics Aps (Vedbæk, Denmark). Compound identification was performed at four levels: Level 1: identification by retention times (compared against in-house authentic standards), accurate mass, and MS/MS spectra; Level 2a: identification by retention times (compared against in-house authentic standards), and accurate mass; Level 2b: identification by accurate mass, and MS/MS spectra; Level 3: identification by accurate mass alone. A deviation of 3 ppm was accepted for accurate mass identification.

Sample extraction

Rumen digesta samples were vortexed and an aliquot (100 µl) transferred to a spin filter (0.22µm). The aliquot was diluted with water (100 µl) and filtered by centrifugation (7000 rpm, 2 x 5 min, 4°C). Filtered extracts were diluted 10 times in mobile phase eluent A and fortified with stable isotope labeled standards before analysis.

Rumen wall and liver samples were mixed with ceramic beads and precooled methanol/water (1:2) fortified with stable isotope labeled standards. The samples were then placed in a pre-cooled (-20°C) bead beater and homogenized (4 x 30 sec., 30 Hz) followed by ultrasonication (5 min). After centrifugation (18000 RCF, 5 min, 4°C), the supernatant of each tube was collected. The sample pellets were re-extracted as described above. The two extract supernatants were pooled and passed through a phosphor removal cartridge (Phree, Phenomenex). A precise aliquot of the extract was evaporated to dryness under a gentle stream of nitrogen, before reconstitution with 10% Eluent B in Eluent A.

LC-MS method

All samples were analyzed using a Thermo Scientific Vanquish LC coupled to an Orbitrap Exploris 240 MS instrument (Thermo Fisher Scientific). An electrospray ionization interface was used as the ionization source. Analysis was performed in positive and negative ionization mode under polarity switching. Ultra-performance liquid chromatography was performed using a slightly modified version of a published protocol[73]. Peak areas were extracted using Compound Discoverer (Thermo Scientific) version 3.2 (digesta) or 3.3 (liver

and wall). For the wall and liver samples, probabilistic quotient normalization was applied prior to further analyses, to decrease concentration effects.

Volatile fatty acid quantification

Rumen digesta samples were thawed on ice and centrifuged when still cold. 450 µL of each sample was transferred to a new tube and 50 µL of a 50% formic acid solution was added to reach a 5 % concentration of formic acid. Samples were then centrifuged again and 400 µL of the supernatant was transferred to a GC-vial, with 1000 µL of an internal standard solution added. Volatile fatty acids were separated using gas chromatography (Trace 1300 GC with autosampler, Thermo Scientific) with a Stabilwax-DA column (30m, 0.52 mm ID, 0.25 µm, Restek).

Microarray polymer profiling

Microarray polymer profiling (MAPP) entails the printing of extracted glycans as high-density microarrays which are then probed with monoclonal antibodies with specificities for different glycan epitopes. The output from MAPP provides insight into the relative abundance of epitopes across the sample set.

Alcohol insoluble residues (AIR) were prepared from each rumen digesta sample (n=24) as follows: samples were homogenized to a fine powder using a tissue lyser (Qiagen). Approximately five volumes of 70% ethanol were added, the samples vortexed for 10 minutes then centrifuged at 2,700 g for 10 minutes and the supernatants discarded, This step was repeated. Approximately five volumes of 1:1 methanol and chloroform were added to the pellet and the samples were again vortexed and centrifuged as previously. Finally, approximately five volumes of acetone were added and the same vortexing and centrifugation steps performed. The resulting pellets were AIR.

To extract glycans, 300 µL of 50 mM diamino-cyclo-hexane-tetra acetic acid, pH 7.5, were added to 10 mg AIR. After agitation in a tissue lyser (27 s-1 for 2 minutes and 10 s-1 2 hours), samples were centrifuged at 2,700 g for 10 minutes. The supernatant was removed, 300 µL 4M NaOH with 1% v/v $NaBH_4$ added to the pellet and the agitation and centrifugation steps repeated. The resultant NaOH extraction supernatants were diluted sequentially (1/2,1/5,1/5,1/5) in microarray printing buffer (55.2% glycerol, 44% water and 0.8% Triton X-100), and the four dilutions were printed in quadruplet onto nitrocellulose membranes using a non-contact microarray robot (Arrayjet, Roslin). Thus, every replicate was represented by a 16-spot subarray (four concentrations and four printing replicates). Arrays were probed with monoclonal antibodies, scanned, uploaded into microarray analysis software (Array Pro Analyzer 6.3, Media Cybernetics) and mean spot signals from each sub array calculated.

Statistics and data visualization

Unless otherwise specified, statistical analyses and visualizations were performed in the R statistical programming language[74] (v. 4.3.2). The knitr[75] package (v. 1.45) was used for reporting, renv[76] (v. 1.0.7) for package management, ggplot2[77] (v. 3.5.1) for visualizations, cowplot[78] (v. 1.1.3) for composing multipanel figure layouts, and ComplexHeatmap[79] (v. 2.15.4) for heatmaps.

16S rRNA gene ASV data was managed with phyloseq[80] (v. 1.46.0), which was also used to calculate alpha diversity indices. Rumen community types (RCTs) were defined using the ASVs data and Dirichlet Multinomial Mixtures[81] clustering implemented with mia[82] (v. 1.10.0), selecting the optimal number of clusters based on the Laplace method. Only ASVs that were present in at least half of all slaughter timepoint samples (n = 35) were used for this analysis.

All beta diversity comparisons for ASV counts and MAG relative abundances were performed using vegan[83] (v. 2.6-6), with robust Aitchison distances statistically compared with PERMANOVA (adonis2; 9999 permutations), and visualized with PCoA (package ape[84], v. 5.8). Principal Component Analysis (PCA) for all other omic data types was run with the “prcomp” function. For (meta)transcriptomic data (variance stabilizing transformed (VST) counts) and (meta)proteomic data (log2 transformed LFQ intensities with imputed missing values), the 1000 features with the highest variance were selected for PCA. For untargeted metabolomic data, where the number of features was orders of magnitude lower, all features were used, except for rumen digesta, where features with annotation level 3 were excluded.

Where statistical testing was done between two categorical variables, Fisher’s exact tests were used. Continuous variables were compared either with t-tests (KPTs and other animal-related metrics, Principal Component scores), or Wilcoxon rank sum tests (alpha diversity indices, metagenomic relative abundances, proteomic LFQ intensities, MAPP intensities, normalized intensities of metabolites from untargeted measurements, and molar percentages of volatile fatty acids) with multiple comparison correction using the “fdr” option of the “p.adjust” function. Differential abundance comparisons for count data (ASVs and meta- and host transcriptomics) were run with DESeq2[85] (v. 1.42.1), with default parameters for transcriptomic data, and the “sfType” parameter changed to "poscounts" for ASV data.

Network analysis (WGCNA)

Correlation-network based analysis was applied on the proteomic and metaproteomic samples to group co-expressed proteins into clusters. Weighted gene co-expression network analysis[21] (WGCNA) v1.73 was applied on data that included imputed missing values to construct clusters independently in the digesta, rumen wall epithelium, and liver samples. These clusters were then correlated via their eigengenes across samples to obtain host-microbiome boundary-links.

Code availability

The R code for the figures and related results tables is available at https://github.com/TheMEMOLab/supacow-share

The code to perform proteomic network analysis using WGCNA is available at https://github.com/cmkobel/holodoublevu

Figures & Tables

Figure 1. Experimental, sampling and data generation design of a controlled beef cattle animal trial.

Figure 2. Microbiome analyses revealing two distinct groups of animals, labeled as Rumen Community Type-A and -B.

Figure 3. Differential abundances of taxa across omics.

Figure 4. Functional differences between rumen community types.

Figure 5: Metabolic predictions from major populations strongly featured in RCT-A and -B animals that are predicted to influence rumen function.

Extended Data Figures & Tables

Extended Data Fig 1. 16S rRNA gene amplicon sequence data analysis results for 71 animals

Extended Data Fig 2. Comparisons of rumen microbial community types and technical and animal-related variables.

Extended Data Fig 3. Heatmap of Spearman correlations between archaea and bacteria of interest and ciliates in metatranscriptomic and metaproteomic data.

Extended Data Fig 4. Ciliate abundances in rumen digesta samples over time for select species.

Extended Data Fig 5. Heatmap of WGCNA results of rumen digesta metaproteomics.

Extended Data Fig 6. Heatmap of WGCNA results of rumen wall metaproteomics.

Supplementary Information

Supplementary Table 1. A. MAG statistics, B. Data overview/summaries.

Acknowledgements

We gratefully acknowledge the financial support of the Novo Nordisk Foundation under 0054575-SuPAcow. The authors acknowledge the Orion High Performance Computing Center at the Norwegian University of Life Sciences and Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway for providing computational resources that have contributed to computations reported in this paper. We also acknowledge Elixir Norway, supported by the Research Council of Norway's (NFR) grant 322392, for the bioinformatics and data management support received for this paper.

References

1. Andersen, T. O. *et al.* Metabolic influence of core ciliates within the rumen microbiome. *ISME J.* **17**, 1128–1140 (2023).
2. Li, Z. *et al.* Genomic insights into the phylogeny and biomass-degrading enzymes of rumen ciliates. *ISME J.* **16**, 2775–2787 (2022).
3. Seshadri, R. *et al.* Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36**, 359–367 (2018).
4. Beauchemin, K. A., Ungerfeld, E. M., Eckard, R. J. & Wang, M. Review: Fifty years of research on rumen methanogenesis: lessons learned and future challenges for mitigation. *Animal* **14**, s2–s16 (2020).
5. Storm, A. C., Kristensen, N. B. & Hanigan, M. D. A model of ruminal volatile fatty acid absorption kinetics and rumen epithelial blood flow in lactating Holstein cows. *J. Dairy Sci.* **95**, 2919–2934 (2012).
6. Furness, D. N. & Butler, R. D. The Cytology of Sheep Rumen Ciliates. I. Ultrastructure of Epidinium caudatum Crawley. *J. Protozool.* **30**, 676–687 (1983).
7. Vogels, G. D., Hoppe, W. F. & Stumm, C. K. Association of methanogenic bacteria with rumen ciliates. *Appl. Environ. Microbiol.* **40**, 608–612 (1980).
8. Ranilla, M. j., Jouany, J.-P. & Morgavi, D. p. Methane production and substrate degradation by rumen microbial communities containing single protozoal species in vitro. *Lett. Appl. Microbiol.* **45**, 675–680 (2007).
9. Kittelmann, S. *et al.* Natural variation in methane emission of sheep fed on a lucerne

pellet diet is unrelated to rumen ciliate community type. *Microbiology* **162**, 459–465 (2016).

10. Ahrendt, S. R., Mondo, S. J., Haridas, S. & Grigoriev, I. V. MycoCosm, the JGI's Fungal Genome Portal for Comparative Genomic and Multiomics Data Analyses. in *Microbial Environmental Genomics (MEG)* (eds. Martin, F. & Uroz, S.) 271–291 (Springer US, New York, NY, 2023). doi:10.1007/978-1-0716-2871-3_14.

11. Park, T., Wijeratne, S., Meulia, T., Firkins, J. L. & Yu, Z. The macronuclear genome of anaerobic ciliate *Entodinium caudatum* reveals its biological features adapted to the distinct rumen environment. *Genomics* **113**, 1416–1427 (2021).

12. Holmes, I., Harris, K. & Quince, C. Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLOS ONE* **7**, e30126 (2012).

13. Eadie, J. M. Inter-Relationships between Certain Rumen Ciliate Protozoa. *Microbiology* **29**, 579–588 (1962).

14. G, T., Tg, N. & Kk, K. Ruminal ciliated protozoa in bison. *Appl. Environ. Microbiol.* **54**, (1988).

15. Göçmen, B., Dehority, B. A. & Rastgeldi, S. Ciliated protozoa in the rumen of Turkish domestic cattle (Bos taurus L.). *J. Eukaryot. Microbiol.* **50**, 104–108 (2003).

16. Williams, A. G. & Coleman, G. S. *The Rumen Protozoa*. (Springer New York, New York, NY, 1992). doi:10.1007/978-1-4612-2776-2.

17. Newbold, C. J., de la Fuente, G., Belanche, A., Ramos-Morales, E. & McEwan, N. R. The Role of Ciliate Protozoa in the Rumen. *Front. Microbiol.* **6**, 1313 (2015).

18. Harmeyer, J. Der Aminosäurenstoffwechsel isolierter Pansenprotozoenarten (Isotricha prostoma und I. intestinalis). *Z. Für Tierphysiol. Tierernähr. Futtermittelkunde* **28**, 75–85 (1971).

19. Melgar, A. *et al.* Effects of 3-nitrooxypropanol on rumen fermentation, lactational performance, and resumption of ovarian cyclicity in dairy cows. *J. Dairy Sci.* **103**, 410–432 (2020).

20. Pitta, D. W. *et al.* The effect of 3-nitrooxypropanol, a potent methane inhibitor, on ruminal microbial gene expression profiles in dairy cows. *Microbiome* **10**, 146 (2022).

21. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

22. Brooke, C. G. *et al.* Methane Reduction Potential of Two Pacific Coast Macroalgae During in vitro Ruminant Fermentation. *Front. Mar. Sci.* **7**, (2020).

23. Machado, L. *et al.* In Vitro Response of Rumen Microbiota to the Antimethanogenic

Red Macroalga Asparagopsis taxiformis. *Microb. Ecol.* **75**, 811–818 (2018).

24. Kinley, R. D., Nys, R. de, Vucko, M. J., Machado, L. & Tomkins, N. W. The red macroalgae Asparagopsis taxiformis is a potent natural antimethanogenic that reduces methane production during in vitro fermentation with rumen fluid. *Anim. Prod. Sci.* **56**, 282–289 (2016).

25. Henderson, G. *et al.* Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci. Rep.* **5**, 14567 (2015).

26. Roehe, R. *et al.* Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance. *PLOS Genet.* **12**, e1005846 (2016).

27. Duthie, C.-A. *et al.* The effect of dietary addition of nitrate or increase in lipid concentrations, alone or in combination, on performance and methane emissions of beef cattle. *Animal* **12**, 280–287 (2018).

28. Illumina, I. 16S metagenomic sequencing library preparation, part# 15044223 Rev. *B* **1213**, 1214 (2015).

29. Apprill, A., McNally, S., Parsons, R. & Weber, L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.* **75**, 129–137 (2015).

30. Callahan, B. J. *et al.* DADA2: High resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

31. Callahan, B. A DADA2 workflow for Big Data: Paired-end (1.4 or later).

32. Ali, A. DADA2 formatted 16S rRNA gene sequences for both bacteria & archaea. Zenodo https://doi.org/10.5281/zenodo.10403693 (2023).

33. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

34. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).

35. De Coster, W., D'hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

36. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

37. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

38. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

39. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).

40. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

41. Newell, R. J. P. *et al.* Aviary.

42. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).

43. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nat. Methods* **20**, 1661–1665 (2023).

44. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).

45. Waters, S. I. & White, J. M. Biochemical and Molecular Characterization of Bovine Fertilin α and β (ADAM 1 and ADAM 2): A Candidate Sperm-Egg Binding/Fusion Complex1. *Biol. Reprod.* **56**, 1245–1254 (1997).

46. Park, T., Wijeratne, S., Meulia, T., Firkins, J. L. & Yu, Z. The macronuclear genome of anaerobic ciliate *Entodinium caudatum* reveals its biological features adapted to the distinct rumen environment. *Genomics* **113**, 1416–1427 (2021).

47. Strachan, C. R. *et al.* Differential carbon utilization enables co-existence of recently speciated Campylobacteraceae in the cow rumen epithelial microbiome. *Nat. Microbiol.* **8**, 309–320 (2023).

48. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

49. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).

50. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

51. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

52. Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* **50**, D571–D577 (2022).

53. Kanehisa, M. The KEGG Database. in *'In Silico' Simulation of Biological Processes* 91–103 (John Wiley & Sons, Ltd, 2002). doi:10.1002/0470857897.ch8.

54. Ogata, H., Goto, S., Fujibuchi, W. & Kanehisa, M. Computation with the KEGG pathway database. *Biosystems* **47**, 119–128 (1998).

55. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, (2021).

56. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

57. Kobel, C. M. *et al.* CompareM2 is a genomes-to-report pipeline for comparing microbial genomes. 2024.07.12.603264 Preprint at https://doi.org/10.1101/2024.07.12.603264 (2024).

58. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).

59. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

60. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

61. Deng, Z.-L., Münch, P. C., Mreches, R. & McHardy, A. C. Rapid and accurate identification of ribosomal RNA sequences via deep learning. *Nucleic Acids Res.* **50**, e60–e60 (2022).

62. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

63. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

64. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).

65. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

66. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2016).

67. Mićić, M., Whyte, J. & Karsten, V. *Sample Preparation Techniques for Soil, Plant, and Animal Samples*. (Springer, 2016).

68. Kobel, C. M. cmkobel/mspipeline1. (2024).

69. Yu, F., Haynes, S. E. & Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Mol. Cell. Proteomics* **20**, (2021).

70. da Veiga Leprevost, F. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **17**, 869–870 (2020).

71. Yu, F. *et al.* Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Mol. Cell. Proteomics* **19**, 1575–1585 (2020).

72. Mayer, M. missRanger: Fast Imputation of Missing Values. (2024).

73. waters. UPLC/MS Monitoring of Water-Soluble Vitamin Bs in Cell Culture Media in Minutes. *www.waters.com* http://www.waters.com/waters/library.htm?lid=134636355.

74. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2021).

75. Xie, Y. knitr: a comprehensive tool for reproducible research in R. in *Implementing reproducible research* 3–31 (Chapman and Hall/CRC, 2018).

76. Ushey [aut, K., cre, Wickham, H., Software, P. & PBC. renv: Project Environments. (2024).

77. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

78. Wilke, C. O. cowplot: Streamlined plot theme and plot annotations for'ggplot2'. *CRAN Contrib. Packag.* (2015).

79. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

80. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS One* **8**, e61217 (2013).

81. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS One* **7**, e30126 (2012).

82. Ernst, F., Shetty, S., Borman, T. & Lahti, L. mia. *Bioconductor* http://bioconductor.org/packages/mia/.

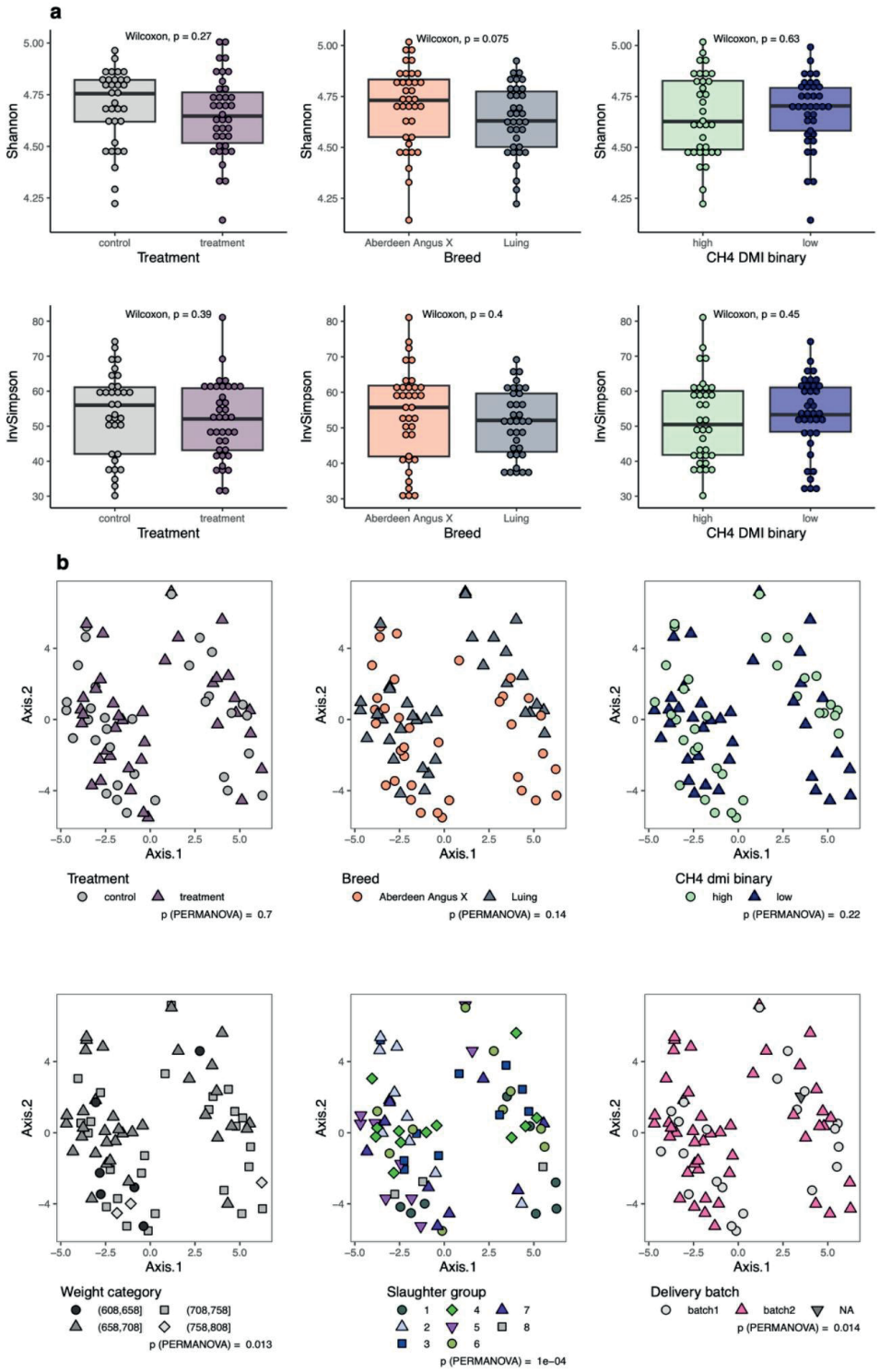
83. Oksanen, J. *et al.* vegan: Community Ecology Package. (2024).

84. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

85. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

Supplementary information for paper #4

- Extended data (figures)

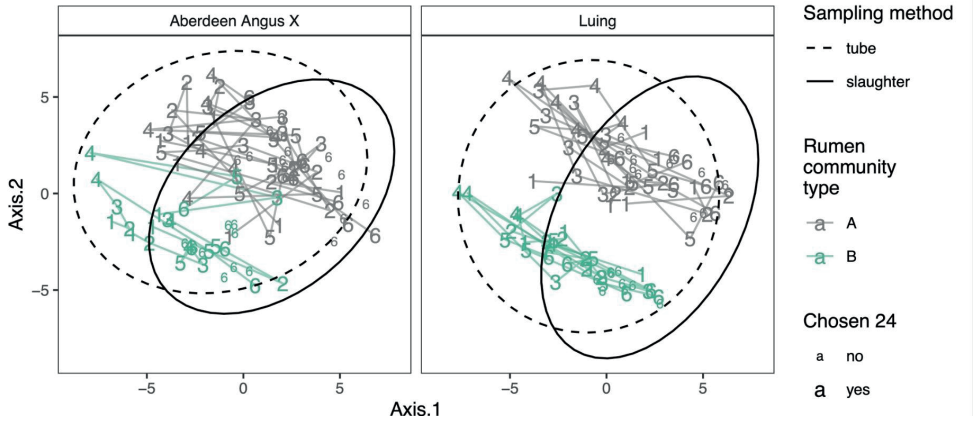

a
Wilcoxon, p = 0.27
Shannon
Treatment
control
treatment
Wilcoxon, p = 0.075
Breed
Aberdeen Angus X
Luing
Wilcoxon, p = 0.63
CH4 DMI binary
high
low
Wilcoxon, p = 0.39
InvSimpson
Wilcoxon, p = 0.4
Wilcoxon, p = 0.45
b
Axis.1
Axis.2
Treatment
control
treatment
p (PERMANOVA) = 0.7
Breed
Aberdeen Angus X
Luing
p (PERMANOVA) = 0.14
CH4 dmi binary
high
low
p (PERMANOVA) = 0.22
Weight category
(608,658]
(708,758]
(658,708]
(758,808]
p (PERMANOVA) = 0.013
Slaughter group
p (PERMANOVA) = 1e−04
Delivery batch
batch1
batch2
NA
p (PERMANOVA) = 0.014

←**Extended Data Fig 1. 16S rRNA gene amplicon sequence data analysis results for 71 animals**. **a**. Alpha diversity, showing Shannon and inverse Simpson indices for treatment (short-term seaweed supplementation), breed (Aberdeen Angus X vs Luing) and methane emission level (binary categorical variable based on median CH_4 g/kg DMI). Box hinges represent the 1st and 3rd quartiles; whiskers range from hinge to the highest and lowest values that are within 1.5*IQR of the hinge. **b**. Beta diversity visualized with Principal Coordinates Analysis of robust Aitchison distances and statistically compared with PERMANOVA (adonis), showing the three main grouping variables (treatment, breed and methane emission category) as well as the three tested variables with the lowest p-value (liveweight, slaughter group and delivery batch). All p-values shown without multiple comparison correction.

a

Variable	p	Statistic	RCT–A mean	RCT–B mean	Effect size	Test
Treatment	1.000					Fisher
Breed	0.809					Fisher
Sire	0.616					Fisher
Performance test group	1.000					Fisher
Age start performance test	0.473	–0.722	451.93	455.52	–0.175	t–test
LWG	0.078	–1.802	1.69	1.87	–0.449	t–test
DMI	0.189	–1.330	11.87	12.20	–0.328	t–test
FCR	0.105	1.645	7.24	6.76	0.400	t–test
Liveweight	0.047	–2.028	693.91	710.59	–0.485	t–test
CH4 g/day	0.191	–1.323	240.63	252.42	–0.317	t–test
CH4 g/kg DMI	0.309	–1.028	23.77	24.82	–0.258	t–test
CH4 DMI binary	1.000					Fisher
Birth month	0.671					Fisher
Pen	0.254					Fisher
Slaughter group	0.228					Fisher
Sample storage bag	1.000					Fisher
Delivery batch	0.426					Fisher

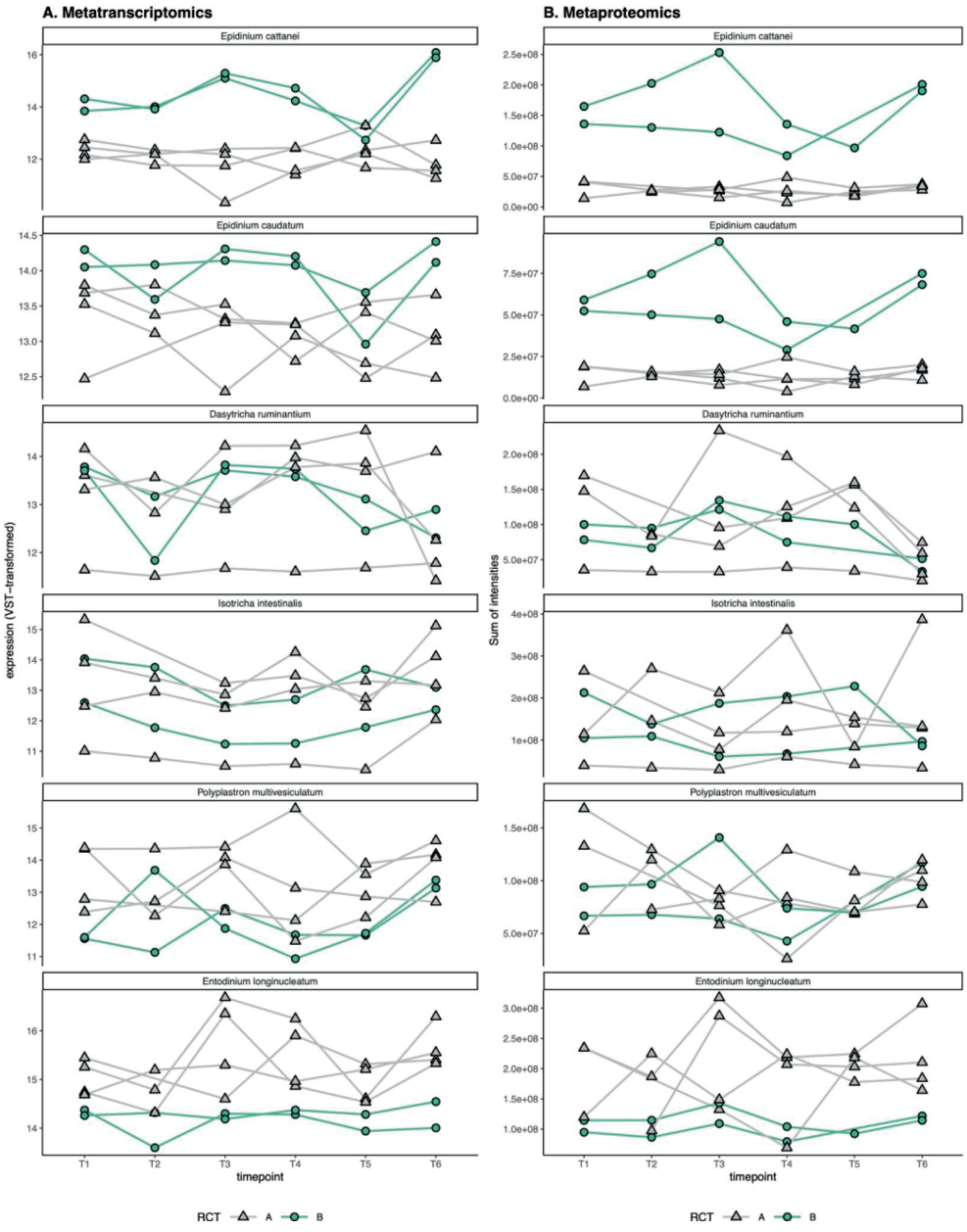
b



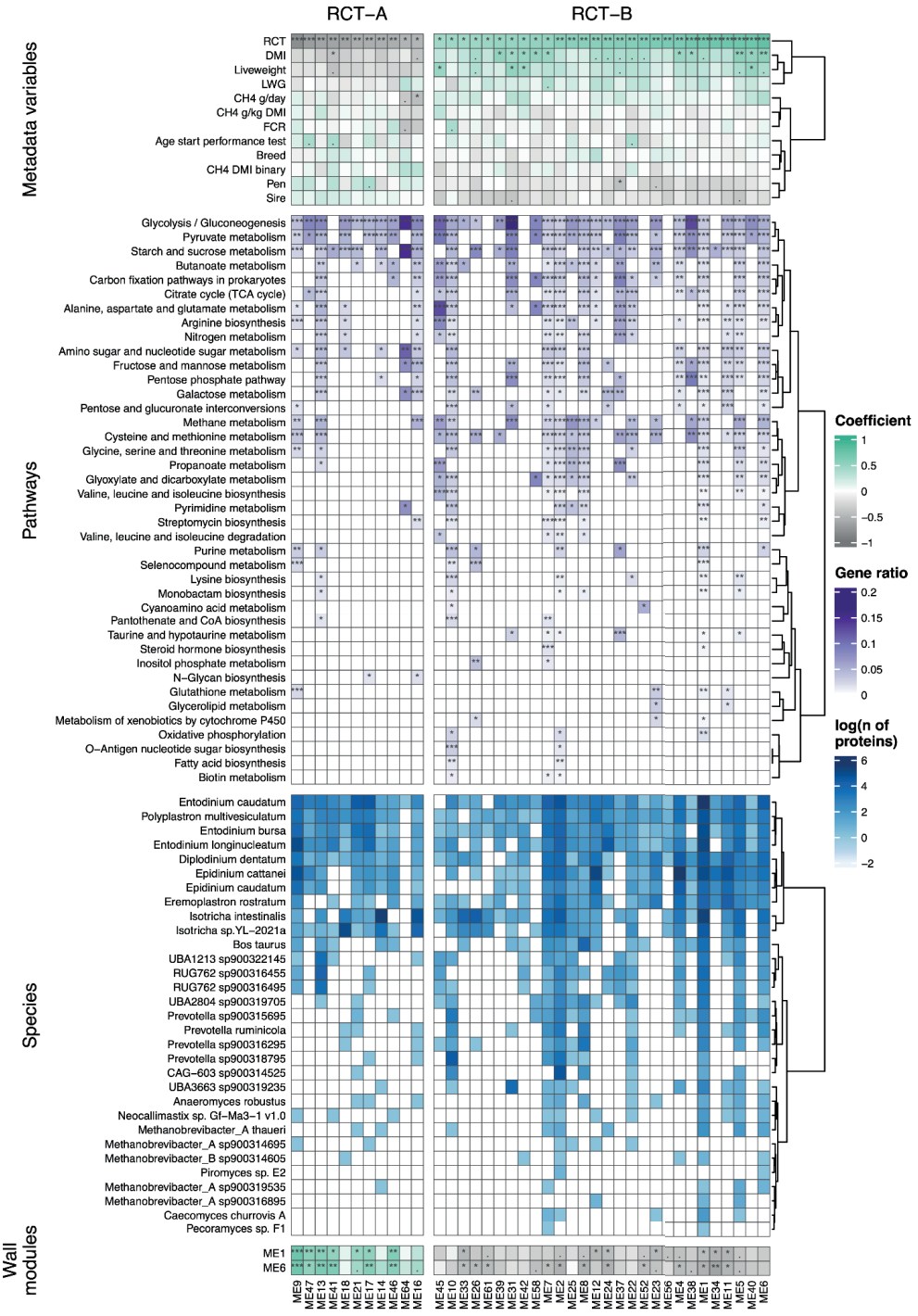
Extended Data Fig 2. Comparisons of rumen community types and technical and animal-related variables. a. Statistical comparisons of variables against the two community types, with t-tests for numerical variables and Fisher's exact tests for categorical variables. *p*-values have not been corrected for multiple comparisons. **b.** Principal Coordinates Analysis with robust Aitchison distances of rumen digesta 16S rRNA gene amplicon sequence data, showing samples from all six time points. Numbers correspond to timepoint, colors to rumen microbial community type, and the 24 animals chosen for deeper analysis are indicated with larger symbols. Samples from the same animal are connected with lines in sequential order. Ellipses indicate 95% confidence levels for sample types: tube sampling (timepoints 1-5; dashed line) or post-slaughter sampling (timepoint 6, continuous line).


A
n.s.
B
metaT
Methanobrevibacter
Prevotella
Ruminococcus
Sodaliphilus
UBA1066
RUG626
RUG762
UBA1213
Faecousia
Merdiplasma
r (Spearman)
1
0.5
0
−0.5
−1
A
n.s.
B
metaP
Methanobrevibacter
UBA1066
Sodaliphilus
Prevotella
Ruminococcus
Merdiplasma
RUG626
RUG762
UBA1213
Faecousia
Entodinium.caudatum.SAG4
Entodinium.longinucleatum.SAG1
Entodinium.longinucleatum.SAG4
Entodinium.bursa.SAG3
Polyplastron.multivesiculatum.SAG12
Entodinium.caudatum.SAG2
Polyplastron.multivesiculatum.SAG8
Polyplastron.multivesiculatum.SAG5
Polyplastron.multivesiculatum.SAG14
Polyplastron.multivesiculatum.SAG13
Polyplastron.multivesiculatum.SAG11
Isotricha.intestinalis.SAG3
Isotricha.intestinalis.SAG2
Isotricha.intestinalis.SAG12
Dasytricha.ruminantium.SAG2
Ostracodinium.gracile.SAG1
Ostracodinium.gracile.SAG3
Isotricha.sp.YL.2021a.SAG1
Isotricha.sp.YL.2021a.SAG2
Diplodinium.dentatum.SAG2
Isotricha.prostoma.SAG2
Isotricha.sp.YL.2021b.SAG5
Entodinium.caudatum.SAG11
Polyplastron.multivesiculatum.SAG7
Isotricha.sp.YL.2021b.SAG3
Ophryoscolex.caudatus.SAG10
Ophryoscolex.caudatus.SAG5
Diplodinium.dentatum.SAG3
Diplodinium.flabellum.SAG2
Ophryoscolex.caudatus.SAG7
Ent_caudatum
Isotricha.prostoma.SAG3
Isotricha.sp.YL.2021b.SAG4
Isotricha.sp.YL.2021b.SAG3
Isotricha.prostoma.SAG1
Dasytricha.ruminantium.SAG3
Isotricha.sp.YL.2021b.SAG2
Isotricha.sp.YL.2021b.SAG1
Enoploplastron.trilioricatum.SAG1
Epidinium.cattanei.SAG3
Epidinium.cattanei.SAG2
Epidinium.caudatum.SAG4
Epidinium.caudatum.SAG1
Ostracodinium.dentatum.SAG1
Ophryoscolex.caudatus.SAG13
Ophryoscolex.caudatus.SAG11
Enoploplastron.trilioricatum.SAG3
Metadinium.minomm.SAG1
Ophryoscolex.caudatus.SAG9
Eremoplastron.rostratum.SAG1
Eremoplastron.rostratum.SAG2
Diplodinium.flabellum.SAG1
Diplodinium.dentatum.SAG4

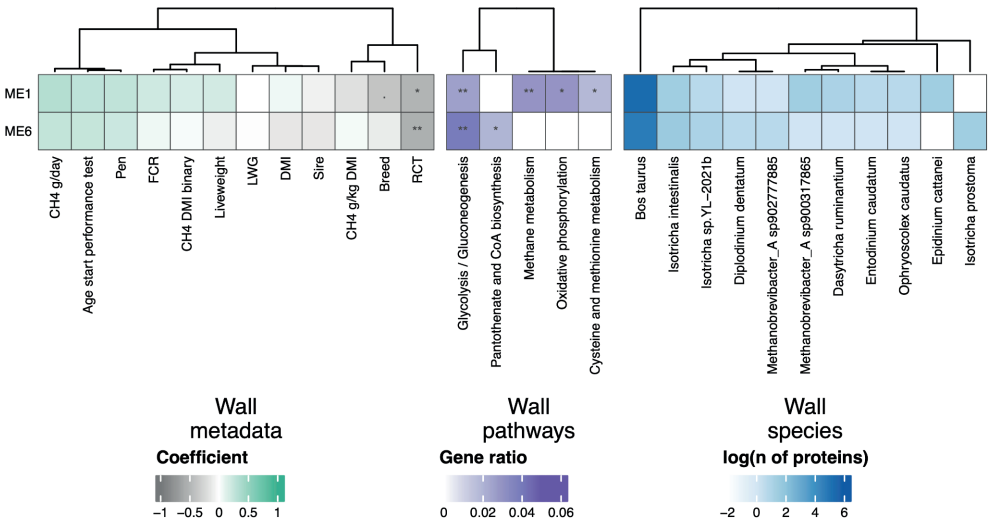
←**Extended Data Fig 3. Heatmap of Spearman correlations between archaea and bacteria of interest and ciliates in metatranscriptomic and metaproteomic data.** Rows correspond to individual archaeal and bacterial MAGs, and are labeled according to the genus classification of the MAGs. Columns correspond to individual ciliate SAGs (and "Ent_caudatum", representing the first published genome available for *Entodinium caudatum*), ordered according to their differential abundance between the two rumen community types (RCT-A and -B) in either metatranscriptomic (metaT) or metaproteomic (metaP) data. The plot shows those MAGs that were among the genera with highest loadings in Fig 2e, and are significantly correlated ($p < 0.05$) with at least 10 SAGs in both metatranscriptomic and metaproteomic data; the same MAGs are shown in both the metaT and the metaP parts of the heatmap. Stars reflect uncorrected *p*-values as follows: *** : $p < 0.001$, ** $p < 0.01$, * : $p < 0.05$, . : $0.1 > p > 0.05$.


A. Metatranscriptomics
B. Metaproteomics
Epidinium cattanei
Epidinium caudatum
Dasytricha ruminantium
Isotricha intestinalis
Polyplastron multivesiculatum
Entodinium longinucleatum
expression (VST−transformed)
Sum of intensities
timepoint
T1
T2
T3
T4
T5
T6
RCT
A
B

←**Extended Data Fig 4. Ciliate abundances in rumen digesta samples over time for select species**. Sampling for timepoints T1-T5 was performed through a nasogastric tube, while T6 was collected post-slaughter. Lines connect samples from the same animal (with a total of 6 animals with time series data). Shape and color correspond to rumen community type (RCT-A or -B). **a**. VST-normalized sums of counts per ciliate species for metatranscriptomic data; missing one point (animal 28, T2) for which sampling was not successful. **b**. Sums of LFQ intensities per ciliate species for metaproteomic data, excluding samples with < 1000 ciliate protein groups detected.


RCT–A
RCT–B
Metadata variables
RCT
DMI
Liveweight
LWG
CH4 g/day
CH4 g/kg DMI
FCR
Age start performance test
Breed
CH4 DMI binary
Pen
Sire
Pathways
Glycolysis / Gluconeogenesis
Pyruvate metabolism
Starch and sucrose metabolism
Butanoate metabolism
Carbon fixation pathways in prokaryotes
Citrate cycle (TCA cycle)
Alanine, aspartate and glutamate metabolism
Arginine biosynthesis
Nitrogen metabolism
Amino sugar and nucleotide sugar metabolism
Fructose and mannose metabolism
Pentose phosphate pathway
Galactose metabolism
Pentose and glucuronate interconversions
Methane metabolism
Cysteine and methionine metabolism
Glycine, serine and threonine metabolism
Propanoate metabolism
Glyoxylate and dicarboxylate metabolism
Valine, leucine and isoleucine biosynthesis
Pyrimidine metabolism
Streptomycin biosynthesis
Valine, leucine and isoleucine degradation
Purine metabolism
Selenocompound metabolism
Lysine biosynthesis
Monobactam biosynthesis
Cyanoamino acid metabolism
Pantothenate and CoA biosynthesis
Taurine and hypotaurine metabolism
Steroid hormone biosynthesis
Inositol phosphate metabolism
N–Glycan biosynthesis
Glutathione metabolism
Glycerolipid metabolism
Metabolism of xenobiotics by cytochrome P450
Oxidative phosphorylation
O–Antigen nucleotide sugar biosynthesis
Fatty acid biosynthesis
Biotin metabolism
Species
Entodinium caudatum
Polyplastron multivesiculatum
Entodinium bursa
Entodinium longinucleatum
Diplodinium dentatum
Epidinium cattanei
Epidinium caudatum
Eremoplastron rostratum
Isotricha intestinalis
Isotricha sp.YL–2021a
Bos taurus
UBA1213 sp900322145
RUG762 sp900316455
RUG762 sp900316495
UBA2804 sp900319705
Prevotella sp900315695
Prevotella ruminicola
Prevotella sp900316295
Prevotella sp900318795
CAG–603 sp900314525
UBA3663 sp900319235
Anaeromyces robustus
Neocallimastix sp. Gf–Ma3–1 v1.0
Methanobrevibacter_A thaueri
Methanobrevibacter_A sp900314695
Methanobrevibacter_B sp900314605
Piromyces sp. E2
Methanobrevibacter_A sp900319535
Methanobrevibacter_A sp900316895
Caecomyces churrovis A
Pecoramyces sp. F1
Wall modules
ME1
ME6
ME9 ME47 ME13 ME41 ME18 ME21 ME17 ME14 ME46 ME64 ME16
ME45 ME10 ME33 ME26 ME61 ME39 ME31 ME42 ME58 ME7 ME25 ME8 ME12 ME24 ME37 ME22 ME52 ME23 ME56 ME4 ME38 ME1 ME34 ME11 ME5 ME40 ME6
Coefficient
1
0.5
0
−0.5
−1
Gene ratio
0.2
0.15
0.1
0.05
0
log(n of proteins)
6
4
2
0
−2

←**Extended Data Fig 5. Heatmap of WGCNA results of rumen digesta metaproteomics**. Showing the 39 modules out of a total of 65 that had p < 0.05 for Pearson correlations with rumen community types (RCT). “Metadata variables” rows show Pearson correlations between modules and animal metrics. “Pathways” rows show those KEGG pathways of class "09100 Metabolism" that were significantly enriched using the hypergeometric test in more than one RCT-correlated module. “Species” rows show the numbers of protein groups per species assigned to the modules, including the top ten taxa with the most protein groups in the RCT-correlated modules for bacteria and protozoa each, and the top five for archaea and fungi. The “Wall modules” rows indicate biweight midcorrelation between digesta and wall modules.



Extended Data Fig 6. Heatmap of WGCNA results of rumen wall metaproteomics. Showing the 2 modules out of a total of 19 that had p < 0.05 for Pearson correlations with rumen community types (RCT). “Metadata variables” rows show Pearson correlations between modules and animal metrics. “Pathways” rows show those KEGG pathways of class "09100 Metabolism" that were significantly enriched using the hypergeometric test in more than one RCT-correlated module. “Species” rows show the numbers of protein groups per species assigned to the modules, including all taxa with more than 2 protein groups detected.

Errata