THE UNIVERSITY *of* EDINBURGH

Using (meta)genomic approaches to improve the accuracy of rumen microbiome analysis

Rebecca Hannah Smith

A thesis submitted for the degree of Doctor of Philosophy (PhD)

College of Medicine and Veterinary Medicine

The University of Edinburgh

2022

Declaration

I declare that this thesis has been written and composed entirely by myself and that the work presented is entirely my own unless otherwise specified in this declaration and throughout this thesis. This work has not been submitted for any other degree or professional qualification.

Chapter 2 contains a publication entitled “Investigating the impact of database choice on the accuracy of metagenomic read classification for the rumen microbiome”, in the journal Animal Microbiome (Smith *et al.*, 2022). Co-authors Mick Watson (Roslin Institute), Laura Glendinning (Roslin Institute and Alan W. Walker (Rowett institute, U. Aberdeen) contributed to the study and manuscript. In particular, Mick Watson created some of the figures included in the manuscript and assisted with the design of the study, Laura Glendinning and Alan W. Walker both contributed feedback throughout the study and to the wording of the manuscript.

Chapter 3 includes data analysis of microbial genomes which were cultured by myself and Gillian Donachie (Rowett Institute, U. Aberdeen). In addition, Gillian conducted an ethanol shock experiment, extracted the DNA of approximately a third of the samples in this chapter, and sequenced the 16S rRNA gene for all samples.

Chapter 4 includes work conducted during an industrial placement with Fios Genomics Ltd., the industrial partner of my PhD CASE Studentship. The conception of this study was a result of joint discussion of myself with my supervisor Mick Watson and my Fios manager Paul McAdam, collaborating research skills with business needs. Paul provided technical supported throughout (e.g., arranged the computational environment) and communicated business needs, and all data analysis was conducted by myself. Signed Rebecca Smith, 31/12/2022

Abstract

The rumen is home to a rich microbiota that metabolises the lignocellulosic feed ingested by the animal, and produces short-chain fatty acids (SCFA) that can be used by the host animal for growth. The rumen microbiota is critical to agriculture and food security, directly contributing to the production of meat and milk.

Culturing has provided some insight into the microbes that live in the rumen, but the majority of the rumen microbiota have yet to be cultured. Culture-independent molecular approaches such as shotgun metagenomics and 16S rRNA gene profiling have therefore become appealing additional methods for studying the rumen microbiome. Such analyses rely on reference information available in databases, which are traditionally populated by cultured microbes. This poses an inherent bias, as taxa that have been cultured will be over-represented in reference databases.

Firstly, this thesis examines the impact of reference database choice on the results of taxonomic classification of rumen metagenomic data. To measure classification accuracy to the read-level, ground-truth data was simulated from known rumen isolate genomes (from the Hungate1000 project). In this study it was demonstrated that the choice of reference database hugely impacted classification results, and for the rumen specifically, the accuracy of classification depended on representation of microbes from this environment in the reference database. The use of custom reference databases that contained culture-derived genomes from the rumen increased classification rate and accuracy at all taxonomic levels. When uncultured metagenome-assembled genomes (rumen MAGs) were included in reference databases, there was an improvement in classification rate, but this resulted in only limited improvements in classification accuracy due to incomplete and informal taxonomy labels. Importantly, this work highlights that the use of

standard, and widely used, reference databases resulted in the classification of rumen data with poor accuracy and suggests that custom reference databases are needed to substantially improve classification accuracy.

To further explore the use of MAGs as representative genomes of uncultured species, Cultured and uncultured rumen genomes were then compared to investigate any differences that may be potential limitations of using MAGs. Bacteria from rumen samples were cultured and isolate genomes sequenced. These cultured genomes were then phylogenetically clustered with rumen MAGs, to create genome pairs consisting of a cultured genome and MAG that were thought to belong to the same microbial strain. The presence of certain functions relevant to the rumen environment were compared for all genome pairs. For all functions relating to nitrogen metabolism, and SCFA and alcohol conversions, the presence or absence of relevant gene pathways was observed to be the same (i.e. they were present or absent for both genomes in the pair) for all seven genome pairs. Carbohydrate active enzymes (CAZys) showed more variation, with only three of the genome pairs having the identical predictions of functions being present or absent. This work also suggested an association between the species and how similar the genome pair (culture-derived genome and MAG) are to one-another. In particular, it would seem that there is less variation between a culture-derived genome and a MAG for species that have a relatively closed pangenome, and more variation for species with a more open pangenome.

The microbiome field is moving towards high-throughput methods, accompanied by new approaches that need to be evaluated for their suitability and accuracy. This work concludes that MAGs may be useful as reference genomes for as-yet uncultured microbes, and highlights factors that may make a MAG more or less suitable as an accurate and representative genome.

Lastly, this thesis presents a project that evaluated the suitability of published pipelines for the functional classification of metagenomic data in an industry setting. Two pipelines, HUMAnN3 and Carnelian, were chosen based on the needs of the business. Two simulated metagenomic datasets, one generated from microbes that are members of the human gut microbiome and one from members of the rumen microbiome, were annotated with functional information from UniRef90/UniProt90 to create a ground-truth. The datasets were then classified by each pipeline, and the functional classification results were compared with the ground-truth annotations of each dataset to assess accuracy. This work found that the HUMAnN3 pipeline classified function in the most similar way to the ground truth annotations. Building on the work presented in Chapter 2, this work highlights issues in reference database bias when classifying microbial function. This project concluded that the HUMAnN3 pipeline was the most suitable for the business to incorporate into their offered services.

Overall, this thesis investigated the accuracy of popular methods to classify the taxonomy and function of metagenomics data. It revealed that the limited number of rumen microbial reference genomes is likely to be a major issue in the rumen microbiome field of research, significantly reducing the accuracy of classification. Furthermore, this work demonstrates that MAGs can resemble culture-derived genomes. Given the pressing need for ruminal reference genomes, the use of MAGs as representative reference genomes for uncultured microbes has the potential to revolutionise the rumen microbiome field. However, robust classification relies on consistent and accurate taxonomic labelling, including that of reference genomes regardless of whether they are culture-derived or metagenome-derived. Continuing improvement in reference databases is required to ensure accurate and valuable insights into the critically important, yet incompletely understood, rumen environment.

Lay summary

The rumen is an organ within the digestive tract of animals such as sheep and cows. A complex collection of microorganisms, known as the microbiota, resides within the rumen. These microbes have adapted over hundreds of thousands of years to be able to digest what the animal eats, breaking down hard-to-digest fibres, and making nutrients that the cow absorbs for energy and growth. There is therefore much interest in manipulating the rumen microbiome in order to boost meat and milk production by farmed livestock. In addition to farmers, scientists are also interested in the rumen microbiome because some rumen microbes produce methane gas, which is contributing to global warming, and understanding how this happens could lead to opportunities to reduce these emissions.

It is difficult to grow many rumen microbes in the laboratory, which makes it challenging to study many of them. This means we must rely on additional methods other than growing them in the laboratory to analyse them. Instead, we can look at the DNA sequences of rumen microbes, known as their genomes. One DNA sequencing-based method to study microbes is called “shotgun metagenomics”, which involves sequencing many fragments of the total sum of microbial genomes in the rumen. This type of sequencing produces data containing millions of short fragments of DNA. By matching the resulting sequence data with reference databases, it may be possible to determine what microbes are present in the rumen, and how they function. However, as most rumen microbes have not been grown and characterised in the laboratory, they are often absent from reference databases. The first results chapter of this thesis demonstrated how choosing different reference databases can significantly impact what microbes are reported to be in the data, with commonly used approaches often giving very inaccurate results. The third results chapter of this thesis contains work that was conducted During a placement in industry. In a similar way to the work presented in the

first Results chapter, shotgun metagenomics data containing fragments of DNA was used to determine how different computational analysis methods impact the subsequent attempts to assign functional capability to the detected microbes of interest. That chapter demonstrated that different computational methods resulted in different functional assignments. Best practice recommendations were provided to the company that hosted me.

If we want to look at the genomes of microbes that have never been grown in the laboratory, we can put the sequenced DNA fragments into groups that we think originate from the same genome in the sample. The fragments in each group are then assembled, by piecing the fragments together to form longer chains. Once assembled, these chains of assembled fragments begin to resemble complete genomes. The second results chapter of this thesis compared genomes of microbes that had been previously cultured in the laboratory, with chains of assembled fragments from shotgun metagenomics data. This work aimed to see how similar they were to one-another. How similar an assembled and culture-derived genome were to one-another varied between species, with less similarity seen for species that have more variation in their genome, and more similarity seen for species that have less variation in their genome. Therefore, this thesis overall describes challenges of metagenomics-based rumen microbiota analyses and suggests mitigations that might help to navigate these challenges.

Acknowledgements

Firstly, I would like to acknowledge my incredible supervisors Prof. Mick Watson, Dr Alan Walker, and Dr Laura Glendinning. Mick, thank you for everything you have done to support me over the years. You have encouraged me to think critically, work independently, and it is because of you that I have become the researcher I am today. Alan, thank you for your unwavering support. You have imparted so much wisdom, and I will carry your advice with me for the rest of my life. Thank you for always inspiring me, and for teaching me fit fit, fits fit fit. Laura, thank you for being such a great person, and for stepping in as my supervisor. Your encouragement has been invaluable and I am very grateful.

Thank you to members of the Watson group for all of the feedback over the years. Amanda, Deepali, Rob, Cezar, Trong, Sebastian, Jenny and Alice, it has been so lovely to be in your group. Thanks also to Gillian and Nate in the Walker group for their friendly support during my visits to Aberdeen.

I would also like to thank my thesis committee: Liz Baggs, Andy Law and Jo Stephens for all of their support and guidance over the years.

Thank you to EASTBIO DTP for funding my work and allowing me to pursue a PhD. Thanks also to my CASE industry partner Fios Genomics Ltd., particularly Paul McAdam who supported me throughout my placement.

Lastly, I would like to express my sincere thanks to my dear family and friends who have supported me throughout this PhD. To my wife, Lucy, thank you for your unwavering support over the years. Mum and Dad, I am standing on your shoulders. This thesis is dedicated to you.

Table of contents

Using (meta)genomic approaches to improve the accuracy of rumen microbiome analysis ..... *i*

Declaration ..... *i*

Abstract ..... *ii*

Lay summary ..... *v*

Acknowledgements ..... *vii*

Table of contents ..... *viii*

List of abbreviations ..... *xii*

List of figures ..... *xiii*

List of tables ..... *xvii*

Chapter 1: An introduction to the rumen environment and metagenomics *1*

1.1: Physiology of the rumen **3**

1.2: The bovine rumen microbiota **5**

1.2.1: The development of the bovine rumen microbiota 5

1.2.2: The composition and function of the bovine rumen microbiota 5

1.3: The global importance of the rumen, and intervention opportunities **10**

1.3.1: Ruminant production and food security 10

1.3.2: Feed efficiency 13

1.3.3: Environmental considerations 15

1.4: Methods to study the rumen microbiome **17**

1.4.1: Culturing the rumen microbiome 17

1.4.2: Sequencing the DNA of the rumen microbiome 19

1.4.2.1: Single genome approaches 19

1.4.2.2: Community approaches 20

1.5: Assigning taxonomy to the rumen microbiome **24**

1.5.1: Traditional understanding of taxonomy 24

1.5.2: Taxonomy in the age of DNA sequencing 25

1.6: Understanding microbial function 27

1.6.1: Describing and classifying function 27

1.6.2: High-throughput functional classification of (meta)genomic data 29

1.7: Identifying gaps in our knowledge of the rumen microbiome 30

1.8: Aims and objectives of this thesis 33

Chapter 2: Measuring the impact of reference database choice on taxonomic classification results, rate and accuracy 34

2.1: Introduction 34

2.2: Research paper 34

2.3: Conclusion 53

Chapter 3: Assessing the suitability of Metagenome-assembled genomes for microbiome analysis 54

3.1: Introduction 54

3.1.1: Metagenome-assembled genomes could be used as reference sequences for uncultured species 54

3.1.2: Considerations when using MAGs 56

3.2: Materials and methods 58

3.2.1: Cultured bacterial genomes from the rumen 58

3.2.1.1: Culturing rumen microbiota 58

3.2.1.2: Sequencing the cultured isolate bacterial genomes 66

3.2.1.3: Assessing the quality of the cultured isolate bacterial genomes and filtering 67

3.2.2: Metagenome-assembled-genomes (MAGs) from rumen metagenomic data 69

3.2.2.1: Rumen samples, metagenomic data and pre-assembly processing 69

3.2.2.2: Metagenomic assembly, binning, and assessing quality 70

3.2.3: Comparing culture-derived and metagenome-assembled genomes 71

3.2.3.1: Clustering culture-derived and metagenome-assembled genomes 71

3.2.3.2: Choosing clusters to take forward for comparison 72

3.2.3.3: Comparing seven pairs of culture-derived and metagenome-assembled genomes of the same strain 72

3.3: Results 74

3.3.1: Microbial genomes isolated from the rumen 74

3.3.1.1: Genome assembly comparison 74

3.3.1.2: Assessing quality and removing contamination 77

3.3.2: Determining the taxonomy of the culture-derived genomes 81

3.4: Metagenome-assembled-genomes (MAGs) created from rumen metagenomic data. 84

3.4.1: Assessing the quality of rumen metagenome-derived bins 84

3.4.2: Taxonomy and phylogeny of rumen genome bins 85

3.5: Clustering culture-derived and metagenome-assembled rumen genomes 87

3.5.1: Strain level clustering results 87

3.5.2: Comparing functional information from culture-derived genomes and MAGs of the same strain 105

3.6: Discussion 112

3.7: Conclusions 119

Chapter 4: Assessing the functional prediction of metagenomics data 121

4.1: Context of this project 121

4.2: Introduction 121

4.2.1: Microbiome analysis and industry 121

4.2.2: Functional classification of sequence data 122

4.2.3: Industrial considerations and aims of project 123

4.3: Materials and methods: 125

4.3.1: Pipeline selection 125

4.3.2: Data and creation of ground truth metagenomic data 134

4.3.2.1: For all pipelines: protein mapping and estimating protein abundance in the simulated data 136

4.3.2.2: For HUMAnN3: creating ground truth data that is annotated with UniRef IDs and read counts 137

4.3.2.3: For HUMAnN3: creating ground truth data with EC annotations 138

4.3.2.4: For Carnelian: creating ground truth data that is annotated by UniRef90 139

4.3.2.5: For Carnelian: creating ground truth data that is annotated by UniProtKB 139

4.3.3: Running the HUMAnN3 pipeline 140

4.3.4: Running the Carnelian pipeline 142

4.4: Results 144

4.4.1: The HUMAnN3 pipeline 144
4.4.1.1: UniRef cluster annotations 144
4.4.1.2: Regrouped EC number annotations 147
4.4.2: The Carnelian pipeline 152
4.4.2.1: Comparing the Carnelian classifications with the UniRef-derived EC annotated ground truth 152
4.4.2.2: Comparison of the Carnelian classifications with the UniProt-derived EC annotated ground truth 155
4.4.3: Overall performance of both pipelines 157
4.4.3.1: Direct comparison of Carnelian and HUMAnN3 raw outputs 157
4.4.3.2: Summarising the comparisons between the ground truth predictions and classifications 158

4.5: Discussion 160

4.5.1: Industrial outcomes 160
4.5.2: Performance of both pipelines 160
4.5.2.1: The HUMAnN3 pipeline 160
4.5.2.2: The Carnelian pipeline 164
4.5.2.3: Comparing the HUMAnN3 and Carnelian pipelines 167
4.6: Conclusions 168

Chapter 5: Discussion 169

5.1: General discussion 169

5.2: Conclusion 182

Chapter 6: Bibliography 183

Chapter 7: Appendix 215

List of abbreviations

ANI - Average Nucleotide Identity

BCFA - Branched-Chain Fatty Acid(s)

DNA - Deoxyribose Nucleic Acid

ICNP - International Code of Nomenclature of Prokaryotes

LCA - Lowest common ancestor

MAG - Metagenome-assembled-genome

Nt - Nucleotide

RNA - Ribonucleic Acid

rRNA - Ribosomal Ribonucleic Acid

SAG - Single amplified genome

SCFA - Short-Chain Fatty Acid(s)

UK - United Kingdom of Great Britain and Northern Ireland

USA - United States of America

WGS - Whole Genome Sequencing

List of figures

Figure 1.1: The growth of literature on microbiome research, as catalogued in PubMed during the past 20 years. 2

Figure 1.2: An overview of the digestive system of the domesticated cow (*Bos taurus*). 3

Figure 1.3: Ruminant production. A: The quantity of meat produced from ruminants, as well as the total human population for the years 1999, 2009 and 2019. B: The contributions from four ruminants (buffalo, cattle, goats, and sheep) to the global ruminant production of meat and milk in 2019. C: Global trade of ruminant meat, displayed as the total value of imports and exports ($USD, millions) for 1999, 2009, 2019. 12

Figure 1.4: An example of a metagenomics workflow, showing the processing of a microbiome sample through to bioinformatic analysis. 22

Figure 3.1: Methodology for culturing anaerobic microbes from rumen samples. 61

Figure 3.2: Methodology of the ethanol shock treatment to select for spore-forming bacteria. 65

Figure 3.3: DNA that was stored in unsuitable EDTA-containing buffer was re-eluted into a suitable buffer prior to sequencing. 66

Figure 3.4: Comparing the average assembly metrics for the culture-derived genomes, when assembled by myself and MicrobesNG. 74

Figure 3.5: N50 values for each culture-derived genome when assembled by myself or MicrobesNG (n=62). 76

Figure 3.6: A summary of the culture-derived genomes quality determined by CheckM before and after filtering. 78

Figure 3.7: Five examples of culture-derived genomes and how filtering impacted the quality of them. 79

Figure 3.8: Taxonomy of the culture-derived isolate genomes. 82

Figure 3.9: Summary of assembly metrics for the genome bins. 85

Figure 3.10: Taxonomy of genome bins as determined by GTDB, with the frequency shown at the phylum and family levels. 86

Figure 3.11: Cluster 3R was the culture-derived genome "40175wA6_BS64" and the MAG "RUG12079". 92

Figure 3.12: Cluster 3S was the culture-derived genome "40175wC7_BSR25_1" and the MAG "RUG14306". 94

Figure 3.13: Cluster 4G was the culture-derived genome "40175wE3_BS32" and the MAG "RUG14882". 96

Figure 3.14: Cluster 5B was the culture-derived genome "40175wG3_BS35" and the MAG "RUG13721". 98

Figure 3.15: Cluster 5J was the culture-derived genome "40175wB4_BS41R_2" and the MAG "RUG13906". 100

Figure 3.16: Cluster 5W was the culture-derived genome "40175wB4_BS41R_1" and the MAG "RUG11184". 102

Figure 3.17: Cluster 6F was the culture-derived genome "40175wG4_BS46" and the MAG "RUG10119". 104

Figure 3.18: Heatmap displaying the coverage of pathways and electric transport chains (ETCs) for each cluster. 106

Figure 3.19: Heatmap displaying the presence or absence of metabolic functions for each cluster. 109

Figure 4.1: The workflow that was followed to create the annotated ground truth data. 137

Figure 4.2: The modification made to the HUMAnN3 pipeline script such that the raw taxonomy and gene counts were printed instead of the transformed counts. 141

Figure 4.3: The frequency of UniRef cluster annotations in the HUMAnN3 classifications vs. Ground truth data. 144

Figure 4.4: The frequency of EC number annotations in the HUMAnN3 classifications vs. Ground truth data. 148

Figure 4.5: A comparison of the Carnelian output (y-axis) and the ground truth data (x-axis) as scatterplots for the human gut (A) and rumen (B) simulated metagenomes. 153

Figure 4.6: Comparison of UniProt-annotated ground truth and Carnelian EC annotations for the rumen data. 156

Figure 4.7: A comparison of the two pipelines HUMAnN3 and Carnelian for (A) the human gut metagenome and the (B) rumen metagenome. .. 157

Figure 4.8: Comparing the accuracy of each functional classification pipeline against ground truth predictions using the R^2 metric. 158

Supplementary Figure S3.1: Alignment of culture-derived genome "40175wG1_BS27" and the MAG "RUG12079". 215

List of tables

Table 3.1: Composition of growth media used to culture rumen microbes. 60

Table 3.2: Genome clusters containing a MAG and isolate genome of the same strain. 88

Table 3.3: The taxonomy of each paired culture-derived and metagenome-assembled cluster. 89

Table 3.4: The 16S, 23S and 5S ribosomal RNA genes present in each genome. 111

Table 4.1: Factors that needed to be considered when selecting a pipeline that was suitable for the aims and needs of the business. 125

Table 4.2: A summary of the functional annotation pipelines considered for this comparative study, together with the advantages and disadvantages of each. 129

Table 4.3: The sampleinfo_file.tsv 143

Table 4.4: A selection of UniRef clusters that were either unclassified or inaccurately classified by the HUMAnN3 pipeline, corresponding to the annotations on Figure 4.3. 145

Table 4.5: The top 10 most frequent EC numbers as classified by the HUMAnN3 pipeline. 150

Table 4.6: The most abundant EC numbers classified by the Carnelian pipeline and the predicted abundance in the ground truth data. 154

Supplementary Table S3.1: CheckM results assessing the quality of the culture-derived genomes before and after filtering. 215

Supplementary Table S3.2: Comparing the assembly metrics for each culture-derived genome, when assembled by myself and by MicrobesNG. 221

Supplementary Table S3.3: Taxonomy of the ruminal culture-derived isolate genomes. 226

Supplementary Table S3.4: Assembly metrics for the rumen metagenome-derived genome bins. 231

Supplementary Table S3.5: The taxonomy of each MAG according to GTDB shown at the phylum, family, genus, and species levels. 235

Supplementary Table S4. 1: Commands used to run the HUMAnN3 pipeline 239

Supplementary Table S4. 2: Commands used to run the Carnelian pipeline 241

Chapter 1: An introduction to the rumen environment and metagenomics

Alongside human socio-development, animal husbandry has evolved and resulted in the domestication of animals for agriculture, including some ruminants such as cattle and sheep (Scheu *et al.*, 2015; Larson and Burger, 2013). As a result, ruminants are found in great numbers all over the world. In 2019 FAOSTAT reported there were 4.5 billion ruminants worldwide, primarily buffalo, cattle, goats, and sheep, with cattle constituting over 35% of ruminants on Earth. These livestock animals have provided vital sources of dietary protein to humans for thousands of years and, as low to middle income countries have increased in wealth, global demand for their meat and dairy produce continues to grow (Huws *et al.*, 2018).

As will be discussed in more detail below, a key factor in animal health, meat and dairy production, and planetary impacts, is the microbiota that resides within their gastrointestinal tracts.

The term microbiota is defined by Berg *et al.* as “a characteristic microbial community occupying a reasonable well-defined habitat”. The word “microbiome” refers to the genetics and physiology of an ecosystem, including the microbes that inhabit it (Berg *et al.*, 2020). Over the past 20 years microbiome research has grown in popularity, reflected in the number of publications containing the keywords “microbiome” or “microbiota”, which increased by 200% from 2015 to 2020 (Figure 1.1).

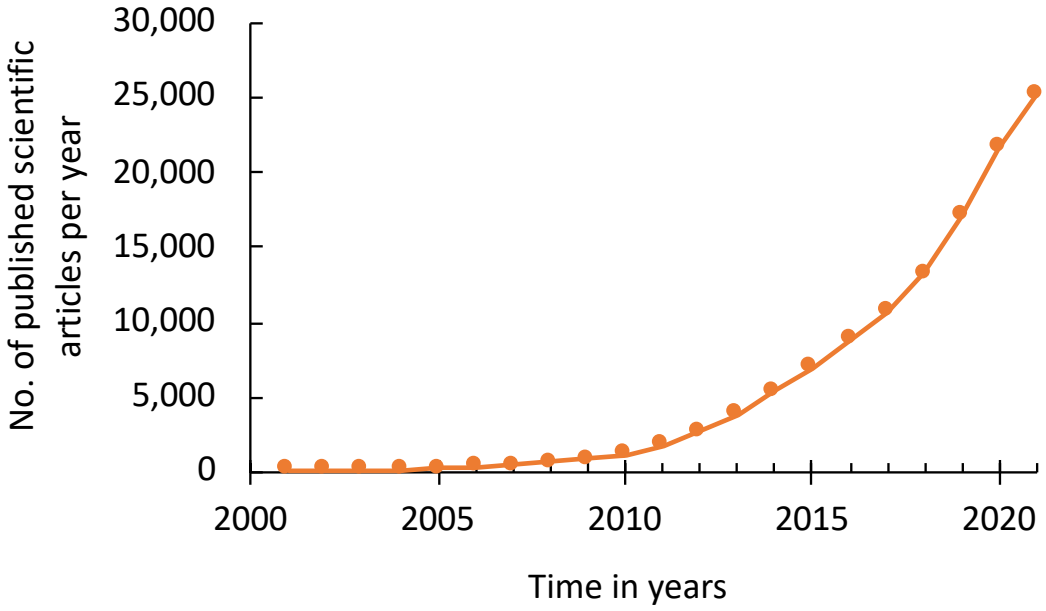


Figure 1.1: The growth of literature on microbiome research, as catalogued in PubMed during the past 20 years. Shown is the frequency in publications on PubMed containing the keywords “microbiome” or “microbiota” for the years 2000 to 2021. Source: https://pubmed.ncbi.nlm.nih.gov/

This growth is not only observed in scientific publications but is also reflected in wider media reporting, with increasing general public interest in the microbes that inhabit hosts (Prados-Bo and Casino, 2021). Rumen microbiome research is also gaining traction, as evidenced by the number of publications in PubMed that contained “rumen microbiome” increasing by over 300% from 2015 to 2020.

1.1: Physiology of the rumen

The rumen has developed as a result of millions of years of co-evolution between the microbiota and host (Mackie, 2002). Rumen physiology and anatomy shows adaptations to the host animal, particularly in different types of feeders. For example, concentrate feeders such as roe deer have a smaller rumen with simpler structures, adapted for the quick fermentation of the less fibrous feed (Hofmann, 1989). In contrast, cattle consume a high fibre diet consisting of grasses and roughage, and as a result their rumen is large and has multiple sacs allowing for the slow fermentation of the roughage (Figure 1.2). Other structures also show adaptation, for example in roughage feeders like cattle the abomasum is larger, as is the omasum, which has numerous laminae for maximum nutrient absorption (Hofmann, 1989).

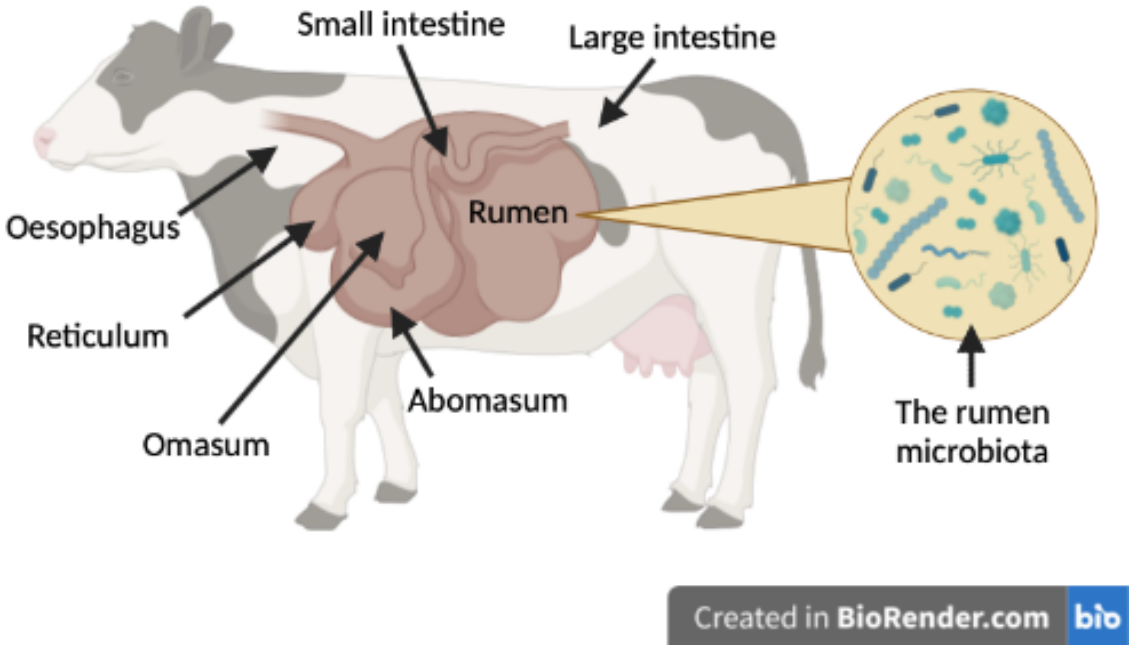


Figure 1.2: An overview of the digestive system of the domesticated cow (*Bos taurus*). Of note is the rumen, which is a largely anaerobic chamber that contains a rich microbiota that ferments the ruminant's digesta.

In cattle, food is first chewed in the mouth, where mechanistic digestion begins. Once swallowed, the muscle of the oesophageal wall propels the mixture of chewed food and saliva into the rumen, where it mixes with previously swallowed food and saliva (Maekawa *et al.*, 2002). In ruminants, previously swallowed digesta is pushed from the contraction of muscle in the reticulum, and travels back up the oesophagus to the mouth where it is further chewed and once again swallowed. This process, known as rumination, further breaks down the fibrous digesta, reducing particle size (Membrive, 2016).

Following the rumination process, the rumen contains a mixture of liquid and solid fractions, which may be floating or sunken sediment depending on their density (Nagaraja, 2016). The gas above the liquid digesta typically contains ~65% carbon dioxide and ~35% methane, making the rumen largely anaerobic (Nagaraja, 2016). Any oxygen that enters the rumen, through swallowing for example, is absorbed by facultative anaerobes (Russell, 2009). The walls of the rumen contain papillae that protrude into the organ, increasing the surface area for maximum nutrient absorption. In most ruminants, the locations in rumen with the most and largest papillae are the sides of the dorsal and ventral sacs (Harfoot, 1981).

The reticulo-omasal orifice separates the reticulum and omasum, ensuring that larger fibrous matter remains in the reticulum for further break-down, and a mostly liquid slurry moves through the omasum (Balch, 1950). In the omasum, much of this liquid is removed before the digesta moves into the abomasum and intestines. The abomasum resembles a non-ruminant “stomach”, excreting mucus, and enzymes that continue to break-down the digesta, while absorbing nutrients. Once digesta leaves the abomasum it enters the first portion of the small intestine, known as the duodenum, where it is further digested as it mixes with bile from the gallbladder and pancreatic excretions (Harfoot, 1981). Lastly, the mostly-digested matter passes through

the large intestine where remaining nutrients are absorbed into the bloodstream, before the digesta passes out of the animal as faeces.

1.2: The bovine rumen microbiota

1.2.1: The development of the bovine rumen microbiota

Shortly after birth, the bovine rumen is colonised by microbes (Stewart *et al.*, 1997). Between 3 and 12 days after birth, many of the microbes that are found in a mature rumen have already colonised the rumen (Rey *et al.*, 2014; Jami *et al.*, 2013). Once solid food is ingested, typically around 9-15 days after birth, the change in diet influences a rapid change in the microbial community, for example cellulolytic bacteria begin to flourish (Jami *et al.*, 2013). Thereafter, it has been demonstrated that the microbial composition of the cow rumen is strongly impacted by the diet of the host (Snelling *et al.*, 2019). As the metabolic function of the cow rumen is broadly similar across animals, it has been suggested that microbiomes may share functionality, although more research is required to determine the functionality of most individual members of the rumen microbiota (Jami and Mizrahi, 2012).

1.2.2: The composition and function of the bovine rumen microbiota

The bovine rumen microbiota has adapted and evolved over time to digest the highly fibrous feed that the animal consumes. Of significance is the conversion of human-indigestible lignocellulosic plant material into metabolites, which can then be used by the animal as fuel, and ultimately contribute to the production of human-edible protein such as meat and milk (Stergiadis *et al.*, 2021; Anil Kumar *et al.*, 2015). Key metabolites are short-chain fatty acids (SCFAs), which are created during anaerobic fermentation of dietary substrates by the rumen microbiota, and are then absorbed and

used by the animal as an energy source, for example, for growth or homeostasis (Seshadri *et al.*, 2018; Kamra, 2005). The main source of energy for the rumen microbiota is carbohydrates, including a range of complex fibres such as cellulose, hemicelluloses and starches (Hungate, 1966). The rumen microbiota can also metabolise other energy sources including lipids, peptides/amino acids and other nitrogenous compounds (Jenkins, 1993; Bach *et al.*, 2005), which are fermented into branched-chain fatty acids (BCFAs) or utilised for microbial growth (Allison and Bryant, 1963).

The microbial population of a mature reticulo-rumen, often grouped together as they continuously mix, is incredibly diverse, with members of the microbiota spanning the domains Archaea, Eukarya and Bacteria. It is also very dense, and it is estimated that there are between 10^{10} - 10^{11} bacteria, 10^{5} - 10^{6} protozoa and 10^{6} – 10^{8} archaea per millilitre (Harfoot, 1981; Ziemer et al., 2008; Matthews et al., 2019) In addition, there are large numbers of viruses in the rumen, including eukaryotic and archaeal viruses and bacteriophage (Gilbert and Klieve, 2015). Viruses of the order Caudovirales are abundant in the rumen (Gilbert et al., 2020). These viruses are responsible for maintaining the rumen microbial populations through cell lysis and biofilm production (Anderson et al., 2017). Additionally, bacteriophage play key roles in the evolution of rumen bacteria as they contribute to metabolic processes via horizontal gene transfer (Brüssow et al., 2004). Studying the genomes of rumen viruses has become possible with metagenomic approaches, including the bioinformatics tools VirSorter (Roux et al., 2015) and VirFinder (Ren et al., 2017). As with bacteria, viral taxonomic classification accuracy is limited by poor representation in reference databases (Roux et al., 2016).

The bacterial fraction of the rumen microbiota is the most populous. The most abundant bacterial phyla in the cow rumen are *Bacteroidetes* and *Firmicutes* (Henderson *et al.*, 2015), less abundant bacterial phyla include

Proteobacteria, *Spirochaetes* and *Actinobacteria* (Kim *et al.*, 2011). These bacteria vary in function, but many have adapted to metabolise the host animal's diet. For example, cellulose degraders include *Fibrobacter succinogenes*, *Ruminococcus flavefaciens* and *Ruminococcus albus (Jami et al.*, 2013), and hemicellulose degraders include members of the genera *Prevotella*, *Butyrivibrio* and *Pseudobutyrivibrio* (Mizrahi *et al.*, 2021). Many of these bacteria have been observed colonising plant material after only 5 minutes incubation, and reach stable numbers after 15 minutes (Kim *et al.*, 2011). Huws *et al.* describe a two-phase colonisation process where the primary phase of bacterial colonisers appear likely to digest soluble nutrients, and the secondary phase likely to digest plant structures (Huws *et al.*, 2016). The bacterial composition of the rumen is influenced by food type, which in turn influences nutrient production. For example, dietary supplements such as red clover have been shown to influence the microbiota and result in an increase of fatty acid production (Huws *et al.*, 2010), and so supplementation may provide opportunities to manipulate the rumen microbiome and positively impact ruminant production, for example.

In contrast to the highly diverse bacterial component of the bovine rumen microbiota, the Archaeal fraction is comparatively simpler in structure. The most abundant archaeal phylum in the cow rumen is *Euryarchaeota (Kim et al.*, 2011), which predominantly contains species that are methanogens (Whitford *et al.*, 2001; Tajima *et al.*, 2001). The majority of methanogenic archaea in the rumen are cross-feeders that can utilise by-products from the fermentation of fibre by other rumen microbes such as formate, carbon dioxide and hydrogen as substrates (Rother and Krzycki, 2010). The catabolic pathways for methanogenic archaea can therefore be categorised into carbon dioxide-reducing, methylotrophic, and aceticlastic (Boone *et al.*, 1993). Most species grow using H_2 and formate as a source of energy, resulting in carbon dioxide being reduced to methane (Janssen and Kirs, 2008). Methanogenic archaea utilise the H_2 that is a product of carbohydrate

metabolism, thus preventing the accumulation of H_2 (Patra et al., 2017). As the build up of H_2 directly inhibits rumen fermentation, the archaea directly contribute to improving the productivity of the rumen (McAllister and Newbold, 2008).

The Eukarya fraction in the rumen include protozoa and fungi. Ciliate protozoa are introduced to young animals by saliva, either through grooming or rumination (Dehority, 1993; Becker and Hsiung, 1929) and have been shown to not colonise the rumen if the animal was reared in isolation (Bryant and Small, 1960). As with the bacterial component described above, he protozoa of the rumen ferment carbohydrates into hydrogen, carbon dioxide and SCFA including formic, propionic, butyric, acetic, and lactic acid (Harfoot, 1981). The other roles played by ciliate protozoa in the rumen remain under speculation (Williams and Coleman, 1997). However, Newbold *et al.* suggest that the larger holotrich protozoa, of the order *Vestibuliferida,* support methane production (see Section 1.3.3) while the smaller entodiniomorph protozoa, of the order *Entodiniomorphida,* provide protein to other members of the microbiota (Newbold *et al.*, 2015) and the host animal (Wright, 2015).

In the rumen, anaerobic fungi colonise ingested plant material, digesting cellulose and lignin (Akin and Rigsby, 1987) (Bauchop, 1979). The abundance of fungi in the rumen has been shown to correlate in abundance with the amount of dietary fibre (Bauchop, 1981). The life cycle of ruminal fungi includes the release of flagellated zoospores, which attach to plant tissues. A network for nutrient transport is built in the form of an extensive rhizoid, and once established sporangium are built for the release of new spores (Akin and Borneman, 1990). The success of fungi in the rumen is due in part to the speed at which plant material is colonised, for example members of the order *Neocallimastigales* were found to colonise ryegrass as rapidly as within five minutes (Edwards *et al.*, 2008). The products of cellulose fermentation by obligate anaerobic fungi in the rumen, which

include formate, acetate, lactate, ethanol, carbon dioxide, and hydrogen, may be directly utilised by the host animal or further fermented by other microbes (Bauchop and Mountfort, 1981). Together, the rumen microbiota provides the host animal with nutrients that contribute to among other things, their growth, reproduction, and lactation.

1.3: The global importance of the rumen, and intervention opportunities

Livestock science faces many challenges, including improving feed efficiency, mitigating methane emissions, fertility and climate adaptions, all the while maintaining the nutritional value of meat and milk and protecting the health and wellbeing of the animal (Hayes *et al.*, 2013). There are therefore a number of key reasons why the study of ruminants, and of their constituent rumen microbiomes, is of considerable global and economic importance.

1.3.1: Ruminant production and food security

The majority of ruminant farming provides meat and milk, which are essential and valuable sources of protein globally. Of all the ruminants, cattle produce by far the most meat and milk, buffalo produce the second-highest amount of meat, and sheep produce the second-highest amount of milk (Figure 1.3). In 2019, 740 million tonnes of fresh cow's milk and almost 74 million tonnes of beef were produced worldwide (FAOSTAT, United Nations). In addition, many other resources originate from ruminants, for example in the year 2019, 129 million tonnes of milk-derived products such as dried milk, cheese, butters, cream, and yoghurt were produced (FAOSTAT, United Nations).

The production of meat from ruminants has been steadily increasing over the last 20 years, alongside related enteric emissions, and the world population (Figure 1.3). Furthermore, rising demand for meat and dairy is likely to continue in the near future as UN projections indicate that the human population could increase to 9.15 billion people by the year 2050 (Alexandratos and Bruinsma, 2012). Having enough food to sustain our population, and achieve and maintain global food security, is therefore an ever-growing concern (Hayes *et al.*, 2013). Indeed, if the production and consumption of food proportionally increases with population growth,

worldwide agricultural production and consumption is projected to increase by approximately 60% by 2050 in comparison to 2005/07.

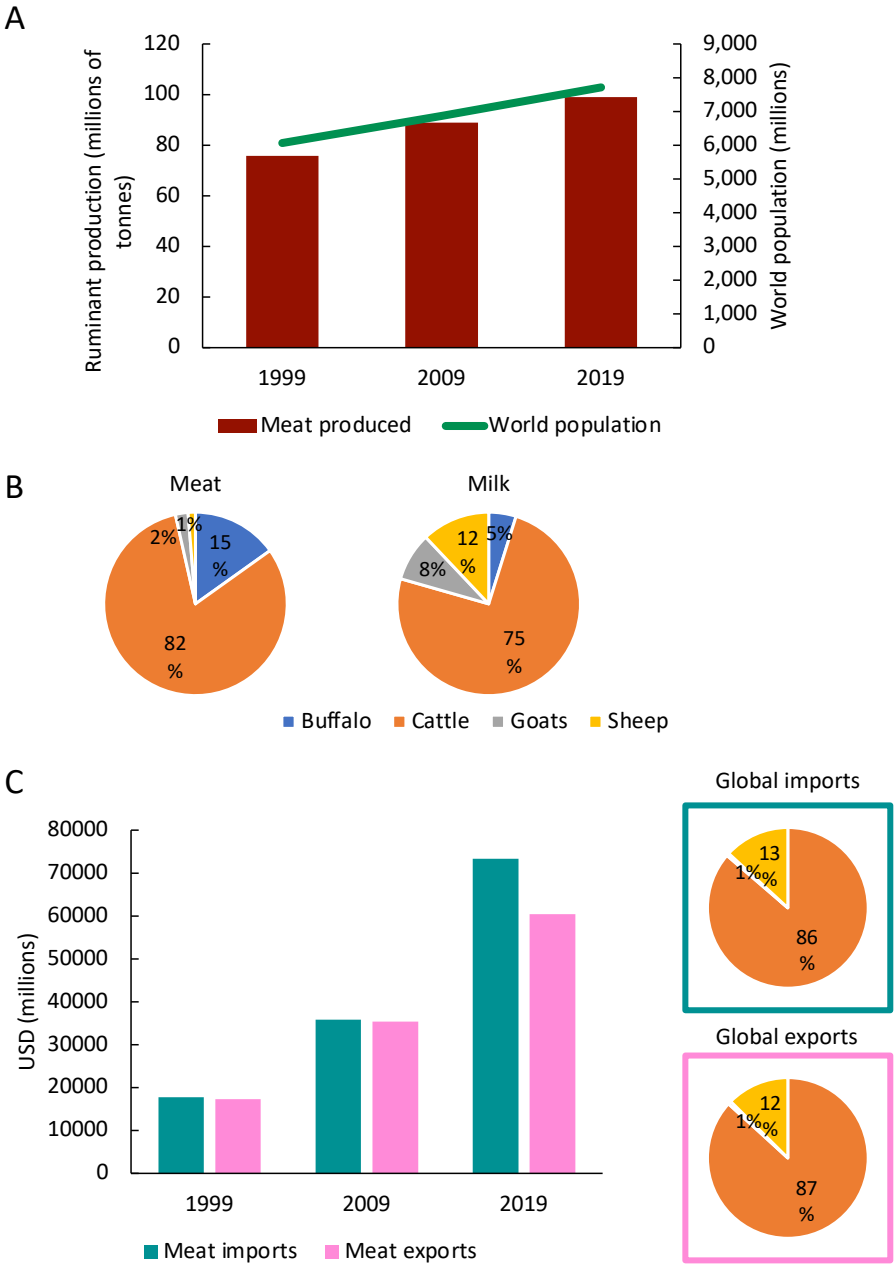


Figure 1.3: Ruminant production. A: The quantity of meat produced from ruminants, as well as the total human population for the years 1999, 2009 and 2019. B: The contributions from four ruminants (buffalo, cattle, goats, and sheep) to the global ruminant production of meat and milk in 2019. C: Global trade of ruminant meat, displayed as the total value of imports and exports ($USD, millions) for 1999, 2009, 2019. The global imports and exports for 2019 are also shown divided by the contributions from each ruminant (corresponding to the same key as B). Ruminant data taken from: Food and Agriculture Organization of the United Nations, Production and Trade data domains. World population data taken from: United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects 2019, Online Edition. Rev. 1.

As well as feeding an increasing number of people around the world, ruminant production also contributes tremendously to global economies and world trade. Shown in Figure 1.3C is the global trade of ruminant meat, showing imports and exports have continued to increase over the last three decades. In 2020, the export of ruminant products totalled more than $104.8 billion USD globally. Of this, over $51 billion USD was for beef, and $5.3 billion for fresh whole cow's milk (FAOSTAT, United Nations). Cattle therefore undeniably provide not only essential food, but also substantially contribute to the global economy.

As discussed in Section 1.2.2, ruminants are able to digest complex polysaccharides, such as cellulose and hemicellulose, that are found in grasses and other plant biomass, owing to their rumen microbiome (Mizrahi *et al.*, 2021). As this plant biomass cannot be digested without the rumen microbiome, ruminants convert the energy that would otherwise be lost from the food chain, into meat and milk. Given this significant contribution towards food security and production, it is therefore undoubtedly of utmost importance to better understand the role the rumen microbiome plays in animal growth and subsequent food production, and to investigate whether manipulating the microbiome can lead to larger yields.

1.3.2: Feed efficiency

Feed efficiency is a measure of production per unit of feed; essentially how much milk or meat is produced given how much food the animal consumes. By better understanding the interactions between host diet and the microbiome, there is therefore the potential to improve the feed efficiency performance of the host animal (Weimer, 2015). In support of this assertion, many previous studies have shown that modifying diet has been shown to change microbial abundance (e.g. (Fernando *et al.*, 2010)). Feed efficiency has also been linked to not only microbial abundance, but functionality. For

example, Xue *et al.* found that in dairy cows that had a high milk yield several *Prevotella* species were significantly more abundant. In addition, functions related to the production of SCFAs acetate, butyrate, and propionate, were significantly enriched in animals that had a high milk yield (Xue *et al.*, 2020).

Within the rumen microbiome there are multiple pathways that can influence various mechanisms, for example glycolysis and carbohydrate metabolism (Ungerfeld, 2020), which in turn impact feed efficiency (Auffret *et al.*, 2020). As modifying the microbiota appears to have the potential to improve feed efficiency, the efficacy of dietary supplementation has been explored. The rumen microbiota metabolises carbohydrates into SCFAs via the glycolytic pathway (Ungerfeld, 2020). Kim *et al.* observed that supplementing beef cows with live *Enterococcus faecium* strain SROD increased overall production of SCFAs including propionate (Mamuad *et al.*, 2019). In addition, red clover supplementation has had a positive effect on ruminant production (*Huws et al.*, 2010), and has been shown to improve rumen fermentation by improving fibre catalysis and increasing the production of SCFAs (Harlow *et al.*, 2020). This is due to Biochanin A and polyphenol oxidases present in the red clover, which inhibits ammonia-producing microbes, allowing cellulolytic bacteria to proliferate and improve fermentation (Harlow *et al.*, 2020).

1.3.3: Environmental considerations

Although feed efficiency is of utmost importance from a global food security and economic perspective, these are not the only potential reasons for rumen microbiota manipulation. Global warming is a pressing threat to the human species, contributed to by greenhouse-gas emissions, and exacerbated by a growing human population. The agricultural sector, including the farming of livestock and the transport of livestock and feed, accounts for approximately 22% of global greenhouse-gas emissions (McMichael *et al.*, 2007). Of the specific gases that contribute to global warming, ruminants are the largest source of anthropogenic methane emissions (Ripple *et al.*, 2014), with ruminant production accounting for 14% of annual global emissions of this gas (Gerber *et al.*, 2013). In addition to the greenhouse gas methane, ruminants indirectly expel ~60% of anthropogenic nitrous oxide (N_2O), a greenhouse gas 300 times more potent than CO_2 (Eckard *et al.*, 2010). As a result, in the agricultural research community, there are efforts being made to reduce the environmental impact of ruminant farming, for example, by reducing methanogenesis.

As discussed previously in Section 1.2.2, methanogens are the rumen-dwelling microbes that are associated with the production of methane, for example the archaeal genus *Methanobrevibacter* (Wallace *et al.*, 2015; Martínez-Álvaro *et al.*, 2020). The presence of competing microbes could potentially reduce the prevalence or activity of such methanogenic species. For example, Wallace *et al.* observed that members of the phylum *Proteobacteria* were 4x more abundant in cattle with lower methane emissions (Wallace *et al.*, 2015). This included members of the family *Succinivibrionaceae*, which is a capnophile and produces succinate as a product of fermentation by the reduction of oxaloacetate to succinate (Pope *et al.*, 2011). This removes exogenous H_2, which may cause a reduction in methane production as it is a substrate for methanogenesis (Conrad, 1999).

One intervention approach is to reduce the amount of H_2 in the rumen, which is a main substrate for methanogenesis. This can be done, for example, by reducing the abundance of microbes which produce it (Tapio *et al.*, 2017), or by increasing H_2 sink pathways (Ungerfeld, 2020). Newbold *et al.* found that removing ciliate protozoa reduced methanogenesis by up to 11%. This may be due to protozoa producing H_2 in mitochondria-like organelles called hydrogenosomes, and so removing these protozoa reduces the production of H_2, that subsequently becomes available for methanogenesis by archaea (Newbold *et al.*, 2015).

1.4: Methods to study the rumen microbiome

If we want to increase feed efficiency or reduce greenhouse gas emissions, improved understanding of the rumen microbiome, and how to reliably manipulate it, is required. In this section I will discuss the various methods that are currently employed to characterise the rumen microbiota. Firstly, I will discuss more traditional culturing methods to characterise microbes. Secondly, I will discuss DNA sequencing methods, including culture-free methods to study the microbial community.

1.4.1: Culturing the rumen microbiome

The field of microbiology is underpinned by the cultivation and isolation of microbes in the laboratory. Robert Hungate, a pioneer of rumen microbiology, was responsible for developing fundamental approaches for cultivation and characterisation of obligately anaerobic microbes from the rumen in the 1940's, nowadays sometimes referred to as the "Hungate method" (Hungate, 1969). This technique aims to mimic the rumen environment; firstly, with the use of rumen fluid in the culturing media to provide essential nutrients, and secondly, to ensure the culture remains anaerobic with the use of a reducing agent and butyl rubber stoppers to prevent oxygen from entering the tubes and media (Hungate, 1944). Hungate developed his approach, after attempts to isolate cellulolytic bacteria failed due to other bacteria out-competing the cellulolytic bacteria (Hungate, 1947). In subsequent developments, Bryant *et al*. suggested that inoculating diluted rumen contents directly onto solid agar plates is advantageous over inoculating directly into liquid media, as it prevents fast growing microbes from out-competing slower growing microbes, which would otherwise reduce the diversity of isolates growing in liquid media (Bryant, 1959).

There are a number of key advantages to having microbes available in culture. For one, it allows investigative experiments to be carried out in the laboratory, thereby providing novel insights into the functionality of these microbes (Decroos *et al.*, 2005). Second, having a microbe in culture allows it to be further developed as a novel biotherapeutic (Prakash *et al.*, 2011). Finally, cultured isolates can have their genomes sequenced, generating further novel insights into their potential functionality, and adding valuable knowledge to existing reference databases (Walker *et al.*, 2014). To this end, the Hungate1000 project sought to create a representative collection of rumen microbial genomes by performing genome sequencing on a global collection of previously cultured rumen isolates. This was important as, prior to the project, there existed only 14 bacterial and 1 methanogen reference genomes from the rumen microbiota (Seshadri *et al.*, 2018). In 2018, the Hungate genome collection was released, containing 501 culture-derived bacterial and archaeal reference genomes (Seshadri *et al.*, 2018). These culture-derived genomes span 9 phyla, 15 donor animal species and 18 countries. 67% of the sequences were sourced from cows, and 51% were from New Zealand. Reflecting the broad-level diversity of the rumen microbiota, the majority of the Hungate genome collection are members of the *Firmicutes* and *Bacteroidetes*, accounting for approximately 81% of sequences. A comparison with the Global Rumen Census sequences (Henderson *et al.*, 2015) revealed that the Hungate genome collection was estimated to represent approximately 75% of genus-level rumen bacterial and archaeal taxa.

Although cultivation of microbes has many advantages, it also has a number of limitations. For example, it is often highly laborious, requiring the use of many media and growth conditions to isolate a broad diversity of microbes. For this reason, it is also often not particularly high-throughput, although modern “culturomics” approaches have attempted to improve this situation (Zehavi *et al.*, 2018). It can also be expensive to isolate rumen microbes, as

many are obligate anaerobes and so require expensive equipment such as anaerobic cabinets, and specialist media components such as rumen fluid, which are not necessarily easily obtained (Seshadri *et al.*, 2018). Finally, it is often not suited to the isolation of minority components of the microbiota, as in order to increase the chances of recovering rarer members of the microbiota, many colonies would have to be picked or various methods to enrich for a minority of species would be needed (Lau *et al.*, 2016). Furthermore, some members of the microbiota may be difficult to culture as optimal conditions for effectively supporting their growth are not yet known, or they may require specific co-factors that are provided by other members of the rumen microbiome (Kingston-Smith *et al.*, 2013; Creevey *et al.*, 2014).

1.4.2: Sequencing the DNA of the rumen microbiome

1.4.2.1: Single genome approaches

Culturing microbes from the rumen can provide single-species, or clonal samples, that are suitable for whole genome sequencing (WGS). WGS produces a high quantity of DNA reads that are then assembled into contigs (Bankevich *et al.*, 2012). Once assembled, this becomes a reference genome, which allows the genome of an individual species to be analysed. However, as the ruminal microbes that have been cultured do not represent all of the species in the rumen, there are a relatively limited number of whole reference genomes (Leahy *et al.*, 2013). Furthermore, as the rumen microbiota exist as a community, examining the genomes of the community may provide a clearer picture of the microbiome (Henderson *et al.*, 2015).

1.4.2.2: Community approaches

As the majority of the rumen microbiota remain uncultured, culture-independent techniques are often used to determine the composition of the wider microbial community. These can be broadly grouped into metabarcoding and metagenomics. Metabarcoding is the process of sequencing marker genes that can be used to reliably predict the taxonomy and phylogeny of all of the detected microbes within a community. An example is sequencing the 16S rRNA gene, which remains the most commonly-used way to study the composition of microbiota (Větrovský and Baldrian, 2013). The 16S rRNA gene encodes the small subunit of the ribosome, making it essential for protein synthesis (Maguire and Zimmermann, 2001). Therefore, the 16S rRNA gene is found in all bacteria, and similar genes are found in members of the domain archaea (Clarridge and Alerts, 2004). Furthermore, changes to the DNA sequence of the 16S rRNA gene are considered an accurate measure of time and evolution for bacterial species (Janda and Abbott, 2007).

A typical metabarcoding method would involve the amplification and sequencing of the marker gene (Pollock *et al.*, 2018), before assigning taxonomy to individual amplicon sequences by comparing to a reference database, for example SILVA (Quast *et al.*, 2013). The advantages of metabarcoding are that it is relatively inexpensive and offers reasonably accurate taxonomic classification (Johnson *et al.*, 2019). A disadvantage of metabarcoding is that, because only single marker genes, and often sub-regions of these marker genes when using modern sequencing approaches, are used, the resolving power of the sequence data is often limited to the genus level, and it is not possible to obtain strain-level resolution (Snipen *et al.*, 2021). In addition, metabarcoding, like other types of sequencing can be biased, as several factors including the methods of DNA extraction (Kennedy *et al.*, 2014), stochastic effects of PCR and the choice of primers can

influence microbial composition and biological conclusions (Hamady and Knight, 2009).

Metagenomic sequencing differs from metabarcoding as it involves the direct sequencing of reads from theoretically all of the extracted DNA present in a sample, rather than focussing only on specific marker genes (Scholz *et al.*, 2012). This technique can produce vast amounts of sequencing read data, derived from the cumulative genomes of the microbial community. In addition, rather than just providing taxonomic information, as with metabarcoding, metagenomics provides information on the functional potential of the microbiota (Franzosa *et al.*, 2018). The greater sequencing depth also means that the resolving power of this technique is much higher than with metabarcoding, and recent developments have shown that it is possible to gain species-level resolution from metagenomic data (Van Rossum *et al.*, 2020).

A typical metagenomics workflow is shown in Figure 1.4. This workflow starts with a sample containing the microbiota of a particular environment, such as the rumen. The DNA is then extracted and prepared for sequencing according to the requirements for the chosen sequencing technology, (Lamble *et al.*, 2013). The DNA sample is then sequenced, producing sequencing reads that are approximately proportionate to the abundance of the original microbe in the sample (Quince *et al.*, 2017). Raw sequence reads are then processed, for example adapter sequences may be trimmed (Bolger *et al.*, 2014), host contamination may be removed, and the quality of the reads may be assessed and lower-quality reads filtered from the analysis. Depending on the analysis, single FastQ files can be treated individually or combined with multiple files to make a metagenomic dataset (Figure 1.4). In addition, metagenomic assembly can differ depending on the type of analysis, for example, after QC the reads can be assembled *de novo* into contigs and scaffolds (Peng *et al.*, 2011), and this can be with just one

sample, known as single-sample assembly or with multiple samples, known as co-assembly (Figure 1.4) (Nurk *et al.*, 2017).

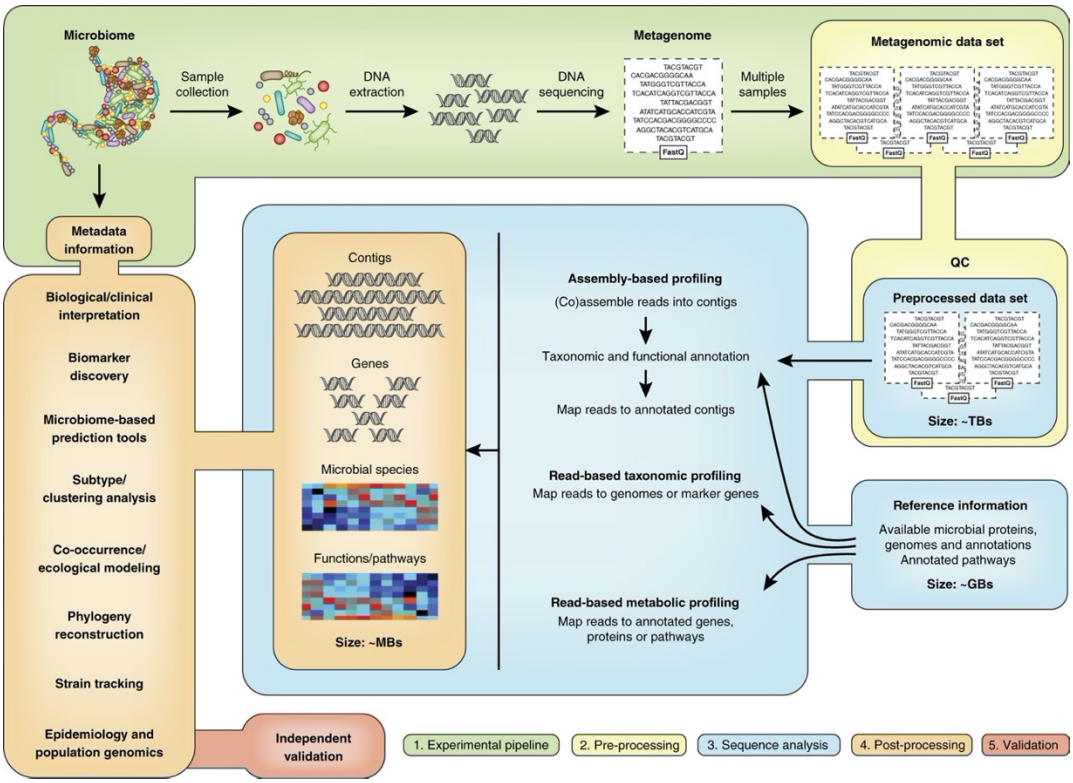


Figure 1.4: An example of a metagenomics workflow, showing the processing of a microbiome sample through to bioinformatic analysis. Reprinted with permission from Springer Nature: Nature Biotechnology. Shotgun metagenomics, from sampling to analysis. Quince, C., Walker, A., Simpson, J. *et al.* Nat Biotechnol 35, 833–844 (2017) https://doi.org/10.1038/nbt.3935. Copyright © 2017, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. License number: 5431280794788, granted on 17/11/2022.

Metagenomic data can also be used to create Metagenome-assembled genomes (MAGs), which are uncultured draft genomes (Kang *et al.*, 2015). Briefly, metagenomic reads are assembled into contigs, and most assembly software use a de Bruijn graph approach (e.g. IDBA for single-sample assembly (Peng *et al.*, 2012), MEGAHIT for co-assembly (Li *et al.*, 2015). The coverage information, or sequencing depth, of each read is determined by mapping the reads to the assembly (for example with BWA-MEM (Li, 2013)). After mapping, the depth of coverage information is calculated, for

example the mapping information could be converted to a BAM file using SAMtools (Li *et al.*, 2009) and the coverage information could be calculated using the jgi_summarize_bam_contig_depths script from MetaBAT2 *(Kang et al.*, 2019). Using the coverage information as an estimation of abundance, the data is then binned into groups of reads thought to belong to the same genome (for example using MetaMAT2). Once dereplicated (for example using dRep (Olm *et al.*, 2017), the quality of the genome bin, or MAG, is usually assessed with CheckM, which is a tool that evaluates the completeness, contamination and strain heterogeneity of a genome using conserved bacterial and archaeal genes (Parks *et al.*, 2015).

Although the threshold for what is considered a “high quality” MAG has varied between research groups in the past, the Genomic Standards Consortium (GSC) released guidance for the Minimum Information about a Metagenome-Assembled Genome (MIMAG). This states that a high-quality MAG should be at least 90% complete, with less than 5% contamination, while a medium-quality MAG was defined as at least 50% completeness, but no more than 10% contamination (Bowers *et al.*, 2017).

Metagenomic sequencing is not without disadvantages, for instance, the quantity of data produced is expensive to store, and requires extensive time and computational power to analyse (Scholz *et al.*, 2012). Another issue is that the coverage of each genome will vary in a metagenomic sample. This can make assembly difficult as low abundant genomes may be lost or poorly reconstructed (Quince *et al.*, 2017). Despite these challenges, for environments like the rumen where most members of the microbiome remain uncultured, metagenomic sequencing remains an attractive option to gain novel information about some of these previously uncharacterised species.

Despite the limitations, metagenomic sequencing is therefore expanding our knowledge about “Microbial dark matter” - a term used to describe the vast

expanse of microbial genomes from microbes that have not yet been cultivated (Rinke *et al.*, 2013). As the majority of microbial diversity remains uncultured (Rappé and Giovannoni, 2003; Pace, 1997), cultivation-free methods, as previously described in this thesis, are needed to study the majority of microbes on Earth.

1.5: Assigning taxonomy to the rumen microbiome

Previously in this thesis, I have described the taxonomy of the rumen microbiota (see Section 1.2.1) and how DNA sequencing is used to study the microbiota (see Section 1.4.2). In this section I will explain the concept of taxonomy, the importance of nomenclature, and how high-throughput metagenomic data is changing the taxonomy *status quo*.

1.5.1: Traditional understanding of taxonomy

Taxonomy is the classification of living organisms on Earth and can be considered as two subjects: descriptive and phylogenetic taxonomy. Descriptive taxonomy is the process by which living organisms are named, and phylogenetics is the placement of the organism in relation to others, and within the tree of life (Godfray, 2002). Analysing microbes under the microscope was pioneered by Antonie van Leeuwenhoek in the mid-seventeenth century, and initial descriptions of microbes was appearance-based, including features such as whether they were motile, or whether they were rod- or coccus-shaped (Van Leeuwenhoek, 1996). Some order began to be applied in the 1700s, when Carl Linnaeus published "*Systema Naturae*", which presented a binomial classification system for describing taxonomic names, a system that is still largely used today. Taxonomy still follows a hierarchal lineage, starting with Domain and proceeded by Phylum,

Class, Order, Family, Genus, and ending with species, strain, or sub-strain (Sneath, 2001).

To bring further order to taxonomic descriptions, the International Code of Nomenclature of Prokaryotes (ICNP) was established, and oversees and approves such naming (Parker *et al.*, 2019). Traditionally, taxonomic classification was based on phenotypes such as metabolic activities, cell and/or colony morphology and, later, features such as cell wall fatty acid analysis (Suzuki and Komagata, 1983) and DNA-DNA hybridisation (Goris *et al.*, 2007). However, taxonomy as a discipline has faced new challenges as more microbial genomes have been sequenced, as it has become clear that traditional taxonomic classifications were incorrect and did not reflect true microbial phylogeny and evolutionary history as we now understand it (Godfray, 2002).

1.5.2: Taxonomy in the age of DNA sequencing

One of the significant benefits of sequence data, either from metabarcoding or metagenomics, is that it provides additional resolution to allow for the inference of phylogenetic relationships between microbes than traditional culture-based taxonomic classification methods (Olsen *et al.*, 1986). High-throughput metabarcoding, specifically sequencing the 16S rRNA gene, revolutionised the phylogenetic analysis of prokaryotes from the 1980s onwards, and 16S rRNA gene sequence data revealed that many traditional taxonomic assignments, particularly those of rumen dwelling genera such as *Clostridium* or *Eubacterium*, were incorrect (Konstantinidis and Tiedje, 2007). For 16S rRNA genes, a sequence similarity of 95% is an accepted cut-off for genus level (Tindall *et al.*, 2010), and 98.65% has been accepted as a threshold for species (Kim *et al.*, 2014).

The use of whole genome sequences, or MAGs, provides even greater taxonomic resolution than marker gene-based analyses. Average nucleotide identity (ANI) is the main approach used with genome data, and compares the proportion of shared genes between sequences (Konstantinidis and Tiedje, 2005). ANI builds upon the more traditional DNA-DNA hybridisation method, which was the first technique that allowed the quantitation of DNA sequence similarity (McCarthy and Bolton, 1963). In the context of taxonomy, ANI is the percentage of nucleotides that are homologous between the query and subject genome sequences (Yoon *et al.*, 2017). As ANI looks at conserved regions that are shared between the genomes that are being compared, it is not influenced by mechanisms of genome change such as horizontal transfer (de Albuquerque and Haag, 2022). A threshold of 95% ANI has been shown to accurately distinguish between microbial species (Goris *et al.*, 2007; Richter and Rosselló-Móra, 2009), and 99% for strains (Konstantinidis and Tiedje, 2005).

Using ANI to determine the taxonomy of a genome sequence relies on the comparison of a novel sequence with a genome of known taxonomy (Tindall *et al.*, 2010), usually derived from a reference database. Two of the databases in the National Centre of Biotechnology Information (NCBI) are RefSeq (O'Leary *et al.*, 2016) and GenBank (Benson *et al.*, 2013). RefSeq is a curated collection of sequencing data, including genomic, transcriptomic and protein sequences, whereas GenBank is an archive that contains data from many DNA and protein sequence databases. The NCBI taxonomy database seeks to combine taxonomy and nomenclature with sequencing data (Federhen, 2012). GTDB (Genome Taxonomy Database) is a database for a new microbial taxonomy framework that was proposed by Parks *et al.* (Parks *et al.*, 2018). This was after a meta-analysis of prokaryotic genomes (including culture-derived from NCBI RefSeq and GenBank, and MAGs from the Sequence Read Archive (SRA)) revealed more branches on the tree of life, including 20 novel phyla (Parks *et al.*, 2017). As GTDB is a genome-

based approach to taxonomy, rather than nomenclature-based, it uses sequence data to assign phylogeny. This is particularly advantageous for uncultured species, as previously the assignment of taxonomy to microbial genomes has required pure culture (Hugenholtz *et al.*, 2021).

Sequence data clearly has the potential to significantly improve taxonomy. However, current ICNP guidelines state that genomes derived from uncultured species are ineligible for a formal taxonomic name (Tindall *et al.*, 2010). In addition, although MAGs have the potential to add significantly to reference databases, by allowing data from uncultured species to be added, there is a lack of guidelines and consistency in how these MAGS are named (Murray *et al.*, 2020). Recently, efforts have been made to provide consistency in assigning names to prokaryote genomes, including MAGs. The SeqCode registry has released guidelines for nomenclature of all prokaryotes regardless of cultivation, including *Candidatus* taxa already in the literature (Hedlund *et al.*, 2022).

1.6: Understanding microbial function

1.6.1: Describing and classifying function

In addition to describing taxonomy, there is a clear need to move towards more mechanistic approaches that seek to uncover the functionality of the rumen microbiota (Martin *et al.*, 2016; Poyet *et al.*, 2019). Functional annotation is the process of assigning putative function to sequence data, for example predicting the protein or enzyme encoded by nucleotide or amino acid sequence (Carola and Rolf, 2011).

To assign function to DNA or amino acid data, often the sequence is compared against a database containing reference sequences and associated metadata. There are many reference databases that contain

functional information, and which can therefore be used to assign function to data. However, they can vary in terminology and resolution (Chen *et al.*, 2017). An example of a protein reference database is UniProt *(Apweiler et al.*, 2004), which stands for the Universal Protein Resource, and is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR) (The UniProt Consortium, 2021). Within UniProt are several databases: the UniProt Knowledgebase (UniProtKB), which contains protein sequences and functional annotations, the UniProt Archive (UniParc), which contains a non-redundant collection of all publicly available protein sequence data, and the UniProt Reference Clusters (UniRef), which contains a subset of UniProtKB with the aim of making it easy to find the proteins being searched for. An example of a metabolic pathway reference database is MetaCyc/BioCyc (Karp *et al.*, 2019). There are some reference databases that specialise in a sub-section of proteins or metabolic activity. For example, ENZYME ExPASy (Bairoch, 2000), which contains the Enzyme Commission (EC) number of a protein, any known information about nomenclature, catalytic activity and cofactors, and points to other reference databases that store relevant information such as a protein sequence (Bairoch, 2000). Another example is the CAZy (Carbohydrate-Active EnZymes) reference database, which specifically stores information about enzymes with carbohydrate catalytic activity (Cantarel *et al.*, 2009).

Consequently, the reference database used will determine the functional classification results. For example, if some data was classified using the CAZy (Carbohydrate-Active Enzyme) database, only enzymes involved in carbohydrate metabolism would be classified. Any other enzymes and non-catalytic proteins would not be classified, regardless of whether they were present in the data. It would therefore be an excellent choice for an investigation specifically interested in carbohydrate metabolism, for example rumen microbiome analysis, but might not be appropriate for other

investigations. Conversely to CAZy, UniProt will classify enzymes that are not related to carbohydrate metabolism, and non-enzyme proteins, and so may be a better choice for broader investigations into function.

1.6.2: High-throughput functional classification of (meta)genomic data

Inferring function, for example using sequence homology, is not guaranteed, and ultimately the only way to ensure a protein has the expected function, is to physically characterise the protein. It is widely known that a large number of genes responsible for encoding proteins linked to essential or core processes are conserved between species (Ashburner *et al.*, 2000). Based on this premise, homologs are often used to extrapolate the function of a protein that has not been confirmed experimentally (Altenhoff *et al.*, 2012), sometimes known as hypothetical proteins. Orthologs are genes which have evolved in separate species over time, originally from a common ancestral gene, and paralogs are genes that have evolved by duplication (Gerlt and Babbitt, 2000; Gevers *et al.*, 2004). While both genes will be very similar in sequence, orthologs are likely to retain the same function throughout evolution whereas paralogs may have diverged and are not guaranteed to share a particular function (Loewenstein *et al.*, 2009; Altenhoff *et al.*, 2012).

A challenge when classifying (meta)genomic data is that annotating function relies on accurately predicting the function of genes and proteins, which first relies on the accurate prediction of gene structures (Burge and Karlin, 1997; Guigó *et al.*, 2000). As reference databases can be abundantly populated with hypothetical genes that have been deduced by sequence similarity only and have no experimentally-verified function, only the minority of annotations in reference databases are potentially valid (Bengtsson-Palme *et al.*, 2016; Schnoes *et al.*, 2009). This means, as with other high-throughput computational techniques, that annotations cannot be trusted unconditionally.

The only way to ensure that a classified function is correct, is to confirm physiology, for example by characterising a microbe in the laboratory.

1.7: Identifying gaps in our knowledge of the rumen microbiome

Previously, this thesis has discussed the global importance of ruminants (see Section 1.3.1) and how characterising the rumen microbiota could lead to understanding and identifying ways to improve ruminant production (see Section 1.3.2).

Recent developments in DNA-sequence-based technologies mean that there are now exciting opportunities to characterise the microbiota in previously unprecedented detail. However, given that there are so many bioinformatics tools available, it is important that these are optimised, to ensure that results are interpreted correctly (Escobar-Zepeda *et al.*, 2015). Characterising the rumen microbiota requires a collaborative effort by microbiologists, bioinformaticians and agriculturalists. DNA sequencing allows for the study of the microbial genomes at nucleotide and, although not always accurate, protein level. It would perhaps be most ideal to cultivate all species in the rumen, and examine their morphology, physiology, and metabolism. However, many members of the rumen microbiota have not been cultivated. Therefore, currently the main option to investigate what roles these microbes may have, is by inferring functionality at the DNA sequencing level. This may improve cultivation efforts as information gained from genomic sequences could inform culturing, for example metabolic information can be annotated, informing growth media composition (Liu *et al.*, 2022).

In Section 1.4.1, I described culturing the rumen microbiota and the Hungate 1000 genome collection. Even though the Hungate genome collection provides useful reference genomes for some rumen taxa, it does not

represent all taxa from the rumen. Using culture-independent methods, Kim *et al.* and Creevey *et al.*, estimate the phylum *Bacteroidetes* accounts for 27% and 38% of known rumen bacteria taxa respectively. Whereas members of this phylum only account for 3% of cultured isolates, spanning just six genera. In addition, Kim *et al.* (2014) and Creevey *et al.* (2014) found multiple phyla for which no, or very few, isolates have been cultured. Some of these phyla have assigned taxonomy such as *Cyanobacteria*, *Planctomycetes*, *Acidobacteria*, as well as those who do not have a formal taxonomy, such as "*OP10*" (Dunfield *et al.*, 2012), "*SR1*" (Campbell *et al.*, 2013) and "*TM7*" (Philip *et al.*, 2001). Although *Cyanobacteria* are photosynthetic and populate aquatic environemnts, they have been widely reported in rumen studies (Neves et al., 2017). It has been proposed that these bacteria are not true *Cyanobacteria*, but are a non-photosynthetic relation called *Melainabacteria* (Soo et al., 2014), a more likely rumen dweller owing to the capablity of carbohydrate metabolism (Di Rienzi et al., 2013). The rumen microbiome is complex, and the taxonomies of these microbiota are far from well understood.

MAGs provide an opportunity to study the genomes of uncultivated, and potentially novel species (Albertsen *et al.*, 2013), including phylogenetic analysis and the assignment of taxonomic and functional information (Jo, 2004). Watson combined metagenomic data from 11 rumen studies and analysed over 33,000 rumen MAGs, which reduced to ~7,500 MAGs after dereplication (Watson, 2021). Watson used the Chao 1 index to estimate that there are over 13,600 species in the rumen. The largest collection of culture-derived rumen genomes (Seshadri *et al.*, 2018) contains only a few-hundred species, which means that if the estimation made by Watson is accurate, less than 4% of species in the rumen have been cultured to date. As most of the rumen microbiota remain uncultured, MAGs provide a potential opportunity to resolve and study these genomes.

In summary, arguably the most significant issue underpinning rumen microbiome research is that the majority of the rumen microbiota is thought to remain as-yet uncultured. This thesis will explore the applications of metagenomic data, and how such data allows for the analysis of rumen microbial genomes that have not yet been characterised with laboratory-based microbiology.

1.8: Aims and objectives of this thesis

The second chapter of this thesis presents work that investigated the importance of the choice of reference databases for the subsequent accuracy of taxonomic classification of metagenomic data. The aim of this work was to measure the impact of reference database composition on taxonomic classification results of rumen metagenomic data using a variety of reference databases and ground-truth data. In addition, this work aimed to investigate the impact of adding MAGs to the reference database.

Building upon the concepts discussed in chapter one, namely the use of MAGs as reference genomes, the third chapter of this thesis aimed to compare MAGs with culture-derived genomes of the same strain, with the aim of assessing if there are any limitations to using MAGs as a reference genome *in lieu* of a culture-derived reference genome.

The fourth chapter of this thesis describes work undertaken during an industrial placement with Fios Genomics Ltd. The aims of this work were to assess the suitability of publicly available functional metagenomic data classification pipelines for an industrial application, and estimate the accuracy of selected pipelines for the business needs of Fios Genomics Ltd.

Overall, my PhD aimed to investigate bioinformatic tools used for metagenomic analyses, with the overall aim of assessing the accuracy of rumen microbiome research, and suggesting how it might be improved.

Chapter 2: Measuring the impact of reference database choice on taxonomic classification results, rate and accuracy

2.1: Introduction

Despite the global significance of ruminants, the rumen microbiome is not fully understood, and many members of the rumen microbiota remain as-yet uncultured. Metagenomic sequencing has allowed for the study of genomes from uncultured microbes. However, commonly used classification techniques to characterise metagenomic data rely on the information stored in reference databases. Due to the limited number of culture-derived reference genomes isolated from the rumen, the rumen microbiome is poorly represented in reference databases, which has been reflected in low classification rates.

This work aimed to classify rumen data with the popular bioinformatics tool Kraken2 and investigate the impact of reference database choice on classification results and accuracy. Genomes derived from uncultured microbes are quickly becoming the “next best thing” in metagenomics-based research, and so this work also aimed to assess the impact of adding rumen MAGs to the reference database.

2.2: Research paper

This work has been published as “Investigating the impact of database choice on the accuracy of metagenomic read classification for the rumen microbiome” in Animal Microbiome (Smith *et al.*, 2022). The journal Animal Microbiome is Open-Access, and the publication is freely available for reproduction under a Creative Commons Attribution 4.0 International License

(see *https://creativecommons.org/licenses/by/4.0/*). The data from this study is available to download at *https://doi.org/10.7488/ds/3444*.

Animal Microbiome

RESEARCH

Open Access

Investigating the impact of database choice on the accuracy of metagenomic read classification for the rumen microbiome

Rebecca H. Smith[1*], Laura Glendinning[1], Alan W. Walker[2] and Mick Watson[1]

Abstract

Microbiome analysis is quickly moving towards high-throughput methods such as metagenomic sequencing. Accurate taxonomic classification of metagenomic data relies on reference sequence databases, and their associated taxonomy. However, for understudied environments such as the rumen microbiome many sequences will be derived from novel or uncultured microbes that are not present in reference databases. As a result, taxonomic classification of metagenomic data from understudied environments may be inaccurate. To assess the accuracy of taxonomic read classification, this study classified metagenomic data that had been simulated from cultured rumen microbial genomes from the Hungate collection. To assess the impact of reference databases on the accuracy of taxonomic classification, the data was classified with Kraken 2 using several reference databases. We found that the choice and composition of reference database significantly impacted on taxonomic classification results, and accuracy. In particular, NCBI RefSeq proved to be a poor choice of database. Our results indicate that inaccurate read classification is likely to be a significant problem, affecting all studies that use insufficient reference databases. We observed that adding cultured reference genomes from the rumen to the reference database greatly improved classification rate and accuracy. We also demonstrated that metagenome-assembled genomes (MAGs) have the potential to further enhance classification accuracy by representing uncultivated microbes, sequences of which would otherwise be unclassified or incorrectly classified. However, classification accuracy was strongly dependent on the taxonomic labels assigned to these MAGs. We therefore highlight the importance of accurate reference taxonomic information and suggest that, with formal taxonomic lineages, MAGs have the potential to improve classification rate and accuracy, particularly in environments such as the rumen that are understudied or contain many novel genomes.

Keywords: Metagenome-assembled genomes, Metagenome, Rumen, Microbiome, Reference databases, Read classification, Taxonomy

Background

Ruminants are vital for global food security, providing high-quality protein for the increasing food demands of an expanding human population. The rumen is home to a complex microbial ecosystem containing bacteria, archaea, fungi, protozoa and viruses. The relationship between the host and these microbes is symbiotic, as they ferment lignocellulosic feed into volatile fatty acids, which are a key energy source for the host animal [1]. Subsequently the rumen microbiome significantly contributes to global food security and world trade. Cattle alone contribute substantially to the economy; in 2018 the global production value of beef exceeded $110 billion USD, and cow's milk exceeded $280 billion USD (FAOSTAT). Understanding the rumen is paramount

*Correspondence: r.h.smith@ed.ac.uk

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, Midlothian EH25 9RG, UK
Full list of author information is available at the end of the article

to the success of many avenues of agricultural research, including feed-conversion efficiency [2], [3], methane emissions [4–7] and investigating the impact of diet on the spread of antibiotic resistance [8].

In spite of the importance of ruminants, the rumen continues to be an under-characterised environment [9] with many rumen-dwelling microbes remaining uncultured, and as such absent from public reference databases. To mitigate this issue, efforts have been made to culture rumen-dwelling microbes, such as the Hungate 1000 project. This significantly improved knowledge surrounding rumen microbiome community structure as these cultured microbes are estimated to represent up to 75% of ruminal bacterial and archaeal genera [10]. However, while culturing efforts have undoubtedly improved the availability of rumen isolated genomes, culturing is laborious, and some species may prove difficult to isolate in the laboratory. As a result, it is known that many ruminant genera remain to be cultured, and are therefore without sequence information [11], meaning reference databases still have important limitations.

Metagenomics is the simultaneous study of DNA extracted from organisms within an environment or microbiome (reviewed in [12]). Metagenome-assembled genomes (MAGs) are draft genomes that have been assembled *'de novo'*, without a reference genome, from binning metagenomic sequencing data [13]. As this process does not require culturing, MAGs can considerably expand on the number of reference genomes derived from culture collections. Additionally, MAG assembly is high-throughput, hundreds or thousands of MAGs can be assembled during a single analysis. MAGs therefore have the potential to transform microbiome analysis by shedding light on the previously poorly described "uncultured majority" [14], [15], and a recent cross-study examination of over 33,000 rumen MAGs concludes that there are still more rumen microbial species to discover [16]. As the rumen microbiome still remains predominantly uncultivated, the use of culture-independent techniques such as MAG assembly are therefore becoming increasingly valuable. Many novel MAGs have been recently published from ruminants [13, 17–25], and these allow the discovery of novel putative genes and functionality in the rumen [26–28].

Studying the microbial composition of an environment using metagenomic data, necessitates the assignment of taxonomic labels to sequence reads, referred to as taxonomic read classification. Classification can be to varying taxonomic levels or ranks. Two of the most commonly used bioinformatics tools available for metagenomic read classification are Kraken [29], and its successor, Kraken 2 [30]. Regardless of classification tool used, reference database quality and comprehensiveness fundamentally underpin the accuracy of results, and classification results can vary dramatically depending on which reference database is used. However, reference databases are known to be highly skewed towards certain well studied species. Blackwell et al. showed that 90% of genomes in the European Nucleotide Archive (ENA), a large publicly available microbial sequence archive, originate from just 20 microbial species [31]. This is important because Meric et al. demonstrated that the number of genomes used to build the index, and the taxonomic system used to classify genomes, can significantly impact classification rates [32]. Similarly, Nasko et al. demonstrated that classification accuracy is impacted by the version of the popular publicly available sequence database RefSeq [33] that is used [34], and Marcelino et al. showed that the reference database needs to represent all domains of life within the microbiome to minimise false positives [35]. Of note, some rumen metagenomics studies report very poor read classification rates when using RefSeq alone [13], [17]. The Hungate 1000 project provides excellent additional reference genomes for taxonomic classification [10] but, given that there are hundreds of currently uncultured and uncharacterised genera in the rumen, the Hungate collection alone may not be fully representative. Subsequently, although the Hungate genomes may improve the classification rate of metagenomic data [13], these may not be true hits, and therefore may not always improve the accuracy of classification. Stewart et al. have twice demonstrated that the addition of MAGs to reference databases improves metagenomic read classification rate by 50–70%, but the addition of Hungate collection genomes showed little improvement (10%) [13], [17]. However, the impact of the addition of MAGs and Hungate collection genomes to reference databases on classification accuracy, not just classification rate, is not yet known.

In this study, simulated data generated from known rumen microbial genomes, was used to test the accuracy of metagenomic read classification using a range of reference databases. This work focused on the read classification tool, Kraken2, which has been shown to be highly accurate and fast [36] and allows for the easy construction of custom reference databases. We found that classification accuracy varies significantly between reference databases, and taxonomic levels. This work emphasises the importance of reference database choice, as well as highlighting the potential low accuracy of taxonomic classification using commonly-applied present approaches. Furthermore, this study demonstrates that the addition of MAGs to reference databases substantially improves read classification accuracy at some taxonomic levels. This work proposes that this improvement has the most potential when using MAGs assembled

from the same environment as the classification data, and when using reference MAGs that have a full taxonomic lineage assigned to them.

Results

Classification rate is heavily impacted by reference database

In order to assess the impact of reference database choice on the classification of metagenomic data, a simulated metagenomic dataset was created from rumen microbial genomes. The taxonomy of the simulated metagenomic dataset was classified using Kraken2 and a variety of reference databases. Briefly, the 'Hungate' database contains rumen microbial genomes. The 'RefSeq' and 'Mini' databases contain the complete bacterial, archaeal and viral genomes in RefSeq, the human genome, as well as a collection of known vectors (UniVec_Core), with the 'Mini' database built to just 8 GB in size. The 'RUG' database contains rumen uncultured genomes (RUGs), which are MAGs that have been assembled from rumen metagenomic data. The 'RefHun' database contained the same sequences as the 'RefSeq' database, with the addition of the cultured isolate genome sequences in the 'Hungate' database. Similarly, the 'RefRUG' database contains the same sequences as the 'RefSeq' database, with the addition of the MAG sequences in the 'RUG' database. Further information on the contents of each database and how they were made can be found in the Methods section, and in Table 1.

As a first test, we looked simply at how much of the simulated metagenomic data was classified (classification rate), regardless of whether or not the classification was accurate. The overall classification rate, meaning the percentage of reads classified by Kraken2 to any taxonomic level when using that particular database, is shown in Fig. 1. Also shown in Fig. 1 is the percentage of reads that were unclassified by Kraken2, meaning they were not classified to any taxonomic level when using that particular database. As expected, since the simulated dataset was derived from the Hungate collection genomes, when the Hungate reference database was used Kraken2 classified almost all reads, with a classification rate of 99.95%. The Kraken2 Mini and RefSeq reference databases resulted in the classification of 39.85% and 50.28% of the reads respectively. Interestingly, of the 460 Hungate genomes used to create the simulated data, 119 were present in RefSeq at the time of analysis. However, as Kraken 2 chooses which genomes to include in each Standard database, not all 119 Hungate genomes in RefSeq were necessarily included in the RefSeq or Mini databases. This indicates that the RefSeq database is not fully representative of the data, which will have impacted on the classification results. The RUG reference database alone had a classification rate of 45.66%, which is a higher rate than the Mini Kraken 2 database but lower than the RefSeq database. Adding the RUG data to the RefSeq database (RefRUG) resulted in 70.09% of reads being classified, which is approximately 1.4x as many reads than were classified with the RefSeq database alone. Finally, as expected, adding the Hungate database to the RefSeq database (RefHun) resulted in near complete classification of the reads. However, there was no apparent benefit to classification rate with the addition of RefSeq (RefHun), when compared to the Hungate database alone (Fig. 1).

After observing the overall classification rates for each reference database, the next step was to examine the classification rates at various taxonomic levels for each reference database. Figure 2 separates the overall classification rate for each reference database into the classification rate at various taxonomic levels. Overall classification rates, regardless of accuracy, are also shown in Additional file 1: Table S1. In general, there was a decline in the classification rate for each database moving down the taxonomic levels from phylum, to family, to genus and finally species.

Anomalously, with some reference databases, classification rate at the genus level was lower than at the species level. This was also observed to a lesser extent in the classification rates at the family level. For example, the RUG database had a classification rate of 45.16% at phylum level, 42.36% at family level, 27.99% at genus level and 43.93% at species level. This is due to a feature of the data itself, as some of the Hungate and RUG genomes used to build the reference databases do not have complete taxonomic lineages. For example, the Hungate genome "*Bacteroidales* bacterium KHT7" (taxonomy ID: 1,855,373) has labels at the kingdom, phylum, class, order and species levels, but no labels at the family and genus levels. Of the 460 Hungate genomes, 8 do not have a label at the family level, and 73 do not have a label at the genus level. Another example is the RUG "*Ruminococcaceae* bacterium RUG10048" (taxonomy ID: 1,898,205), which has the label *Ruminococcaceae* at the family level, and the label "*Ruminococcaceae* bacterium" at the species level, but has no label at the genus level. Of the 4941 RUGs, 3849 have no labels at the genus level, and 1753 have no labels at the family level. 4293 of the RUGs had a non-specific species label, for example "uncultured *Bifidobacterium* sp.". Therefore, as these genomes do not have a taxonomic label at these levels, reads from these genomes appear as unclassified.

The addition of RefSeq to the Hungate reference database (RefHun database) did not significantly impact the classification rate at the higher taxonomic levels compared to the Hungate reference alone (Fig. 2). However,

Table 1 *The contents of each reference database and instructions on how they were built*

Database	Contents	Construction
Hungate	Custom database containing 460 rumen microbial reference genomes from the Hungate collection (see Additional file 2: Table S3)	For file in /hungate_genomes/*.fasta do kraken2-build --add-to-library $file --db hungate_only_db_k2 done kraken2-build --build --threads 16 --db hungate_only_db_k2
Mini	The complete collection of genomes in RefSeq for bacterial, viral and archaeal domains, the human genome and UniVec_Core vectors. The database was built to 8 GB in size to replicate the "MiniKraken" functionality of Kraken1	kraken2-build --download-library bacteria --db mini_standard_db_k2 --use-ftp kraken2-build --download-library archaea --db mini_standard_db_k2 --use-ftp kraken2-build --download-library viral --db mini_standard_db_k2 --use-ftp kraken2-build --download-library human --db mini_standard_db_k2 --use-ftp kraken2-build --download-library UniVec_Core --db mini_standard_db_k2 --use-ftp kraken2-build --db mini_standard_db_k2 --build --max-db-size 8,000,000,000 --threads 4
RefSeq	The complete collection of genomes in RefSeq for bacterial, viral and archaeal domains, the human genome and UniVec_Core vectors	kraken2-build --download-library bacteria --db standard_db_k2 --use-ftp kraken2-build --download-library archaea --db standard_db_k2 --use-ftp kraken2-build --download-library viral --db standard_db_k2 --use-ftp kraken2-build --download-library human --db standard_db_k2 --use-ftp kraken2-build --download-library UniVec_Core --db standard_db_k2 --use-ftp kraken2-build --build --threads 16 --db standard_db_k2
RUG	Custom database containing 4,941 rumen metagenome-assembled genomes (named "RUGs" - see Stewart et al. [17])	For file in /rug_drafts/*.fna do kraken2-build --add-to-library $file --db rug2_only_db_k2 done kraken2-build --build --threads 8 --db rug2_only_db_k2
RefRUG	The complete collection of genomes in RefSeq for bacterial, viral and archaeal domains, the human genome and UniVec_Core vectors with the addition of 4,941 rumen metagenome-assembled genomes (named "RUGs" - see Stewart et al. [17] and the RUG database)	kraken2-build --download-library bacteria --db standard_rug2_db_k2 --use-ftp kraken2-build --download-library archaea --db standard_rug2_db_k2 --use-ftp kraken2-build --download-library viral --db standard_rug2_db_k2 --use-ftp kraken2-build --download-library human --db standard_rug2_db_k2 --use-ftp kraken2-build --download-library UniVec_Core --db standard_rug2_db_k2 --use-ftp for file in /rug_drafts/*.fna do kraken2-build --add-to-library $file --db standard_rug2_db_k2 done kraken2-build --build --threads 16 --db standard_rug2_db_k2
RefHun	The complete collection of genomes in RefSeq for bacterial, viral and archaeal domains, the human genome and UniVec_Core vectors with the addition of 460 reference genomes from the Hungate collection (see Hungate database section of this table and Additional file 2: Table S3)	kraken2-build --download-library bacteria --db standard_hungate_db_k2 --use-ftp kraken2-build --download-library archaea --db standard_hungate_db_k2 --use-ftp kraken2-build --download-library viral --db standard_hungate_db_k2 --use-ftp kraken2-build --download-library human --db standard_hungate_db_k2 --use-ftp kraken2-build --download-library UniVec_Core --db standard_hungate_db_k2 --use-ftp for file in /hungate_genomes/*.fasta do kraken2-build --add-to-library $file --db standard_hungate_db_k2 done kraken2-build --build --threads 16 --db standard_hungate_db_k2

Table 1 (continued)

Database	Contents	Construction
HunRUG	The 460 reference genomes from the Hungate collection (see Hungate database section of this table and Additional file 2: Table S3), and 4,941 rumen metagenome-assembled genomes (named "RUGs" - see Stewart et al. [17] and the RUG and RefRUG databases).	For file in /hungate_genomes/*.fasta do kraken2-build --add-to-library $file --db hungate_rug2_db_k2 done For file in /rug_drafts/*.fna do kraken2-build --add-to-library $file --db hungate_rug2_db_k2 done kraken2-build --build --threads 16 –db hungate_rug2_db_k2
RefHunRUG	The complete collection of genomes in RefSeq for bacterial, viral and archaeal domains, the human genome and UniVec_Core vectors with the addition of 460 reference genomes from the Hungate collection (see Hungate database section of this table and Additional file 2: Table S3), and 4,941 rumen metagenome-assembled genomes (named "RUGs" - see Stewart et al. [17] and the RUG and RefRUG databases).	kraken2-build --download-library bacteria --db standard_hungate_rug2_db_k2 --use-ftp kraken2-build --download-library archaea --db standard_hungate_rug2_db_k2 --use-ftp kraken2-build --download-library viral --db standard_hungate_rug2_db_k2 --use-ftp kraken2-build --download-library human --db standard_hungate_rug2_db_k2 --use-ftp kraken2-build --download-library UniVec_Core --db standard_hungate_rug2_db_k2 --use-ftp For file in /hungate_genomes/*.fasta do kraken2-build --add-to-library $file --db standard_hungate_rug2_db_k2 done For file in /rug_drafts/*.fna do kraken2-build --add-to-library $file --db standard_hungate_rug2_db_k2 done kraken2-build --build --threads 16 --db standard_hungate_rug2_db_k2

The eight reference databases each contain different reference sequences, as described in the Table.

*The additional HunRUG and RefHunRUG reference databases, showed very similar results to the Hungate and RefHun reference databases, and so are only included in the Additional file 1: Fig. S2. Also shown are the commands used to download and/or add to the library for each database, and build each database using Kraken 2

at the lower taxonomic levels, the RefHun database appeared to slightly reduce the classification rate when compared to the Hungate database alone. For example, at the species level with the Hungate database 92.69% of reads were classified, whereas with the RefHun database 89.27% of reads were classified.

Classification accuracy is strongly impacted by reference database

Although classification rate is an important feature, it is clearly more important that data that are classified are-done so accurately. The next logical step was therefore to use ground truth data to investigate the read classification accuracy of each reference database on the simulated metagenomic data. Figure 3 shows the classification accuracy of reads when classified using each reference database, at various taxonomic levels. The same data in tabular form is shown in Additional file 1: Table S2. The percentage of correctly classified reads reduced when moving down the taxonomic levels from phylum to species, for all databases. At the phylum level, the majority of taxonomic labels assigned to classified reads were correct when using all reference databases, or were otherwise unclassified. Indeed, fewer than 4% of classified reads were classified incorrectly for any of the databases at the phylum level.

At the family level and above, no reads were classified incorrectly by Kraken2 with the Hungate database. The addition of Hungate genomes to the RefSeq database (RefHun) also increased the percentage of correctly classified reads substantially compared with using the RefSeq database alone, from 40.93 to 97.82%. Use of some of the reference databases resulted in reads being incorrectly classified at the family level. While classification using the RefSeq database correctly classified a higher percentage of reads than the Mini database (40.93% vs. 35.62%), it also incorrectly classified a higher percentage (7.07% vs. 2.74%), and the ratio of correct:incorrect was better when using the Mini database. Classification using the RUG database resulted in 35.76% of reads being classified correctly, which was less accurate than the RefSeq database but comparable to the Mini database. Additionally, use of the RUG database classified 5.71% of reads incorrectly, which was lower than the RefSeq database but

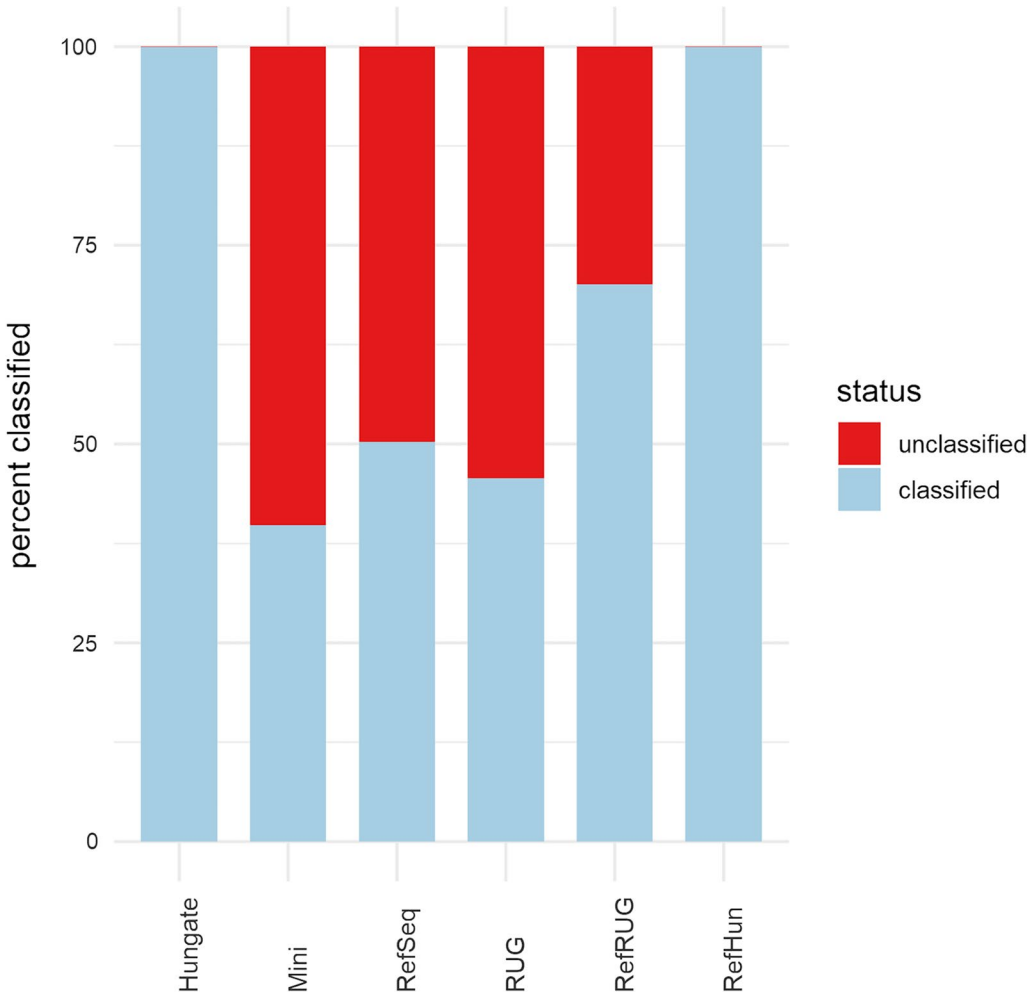


Fig. 1 Overall classification rate of reads for six reference databases. The classification rate of the data for each database are shown in the bars along the x-axis. Details about the databases can be found in Table 1. The y-axis denotes the percentage of reads from the simulated metagenomic dataset which were classified or unclassified by Kraken2 to any taxonomy level using each reference database

higher than the Mini database. Adding the RUG genomes to the RefSeq database (RefRUG) improved almost all classification metrics when compared to using RefSeq alone. However, use of the RefRUG database resulted in a higher number of reads that were classified incorrectly (Fig. 3). Use of the Hungate database correctly classified 97.99% of reads, and the remaining 2.01% were either unclassified or do not have a known truth due to missing taxonomic labels in the reference sequences. These reads are assigned the "truth_unknown" status.

At the genus level, using the RefSeq reference database both classified more reads correctly than the Mini database, which had a better ratio of correct:incorrect assignments. Using the RUG database resulted in fewer reads being classified correctly at the genus level, and resulted in a higher percentage of unclassified reads. However, use of the RUG database again resulted in fewer reads being incorrectly classified than with the RefSeq database. Similar to the family level results, adding the RUG data to RefSeq improved on most metrics when compared to using only the RefSeq database. Use of the Hungate database correctly classified 82.56% of reads, notably caused by reads categorised into the previously mentioned "truth_unknown" status, which accounted for 16.32% of the reads at genus level. Use of the Hungate database resulted in the incorrect classification of very few reads, which was echoed in the RefHun database. Compared to the RefSeq database, classification with the RefHun database classified more reads correctly (81.90% vs. 35.97%), and classified fewer reads incorrectly (0.01% vs. 7.85%).

At the species level, use of both of the RefSeq and the Mini databases classified a similar proportion of reads correctly (22.74% vs. 20.65%). However, using the RefSeq

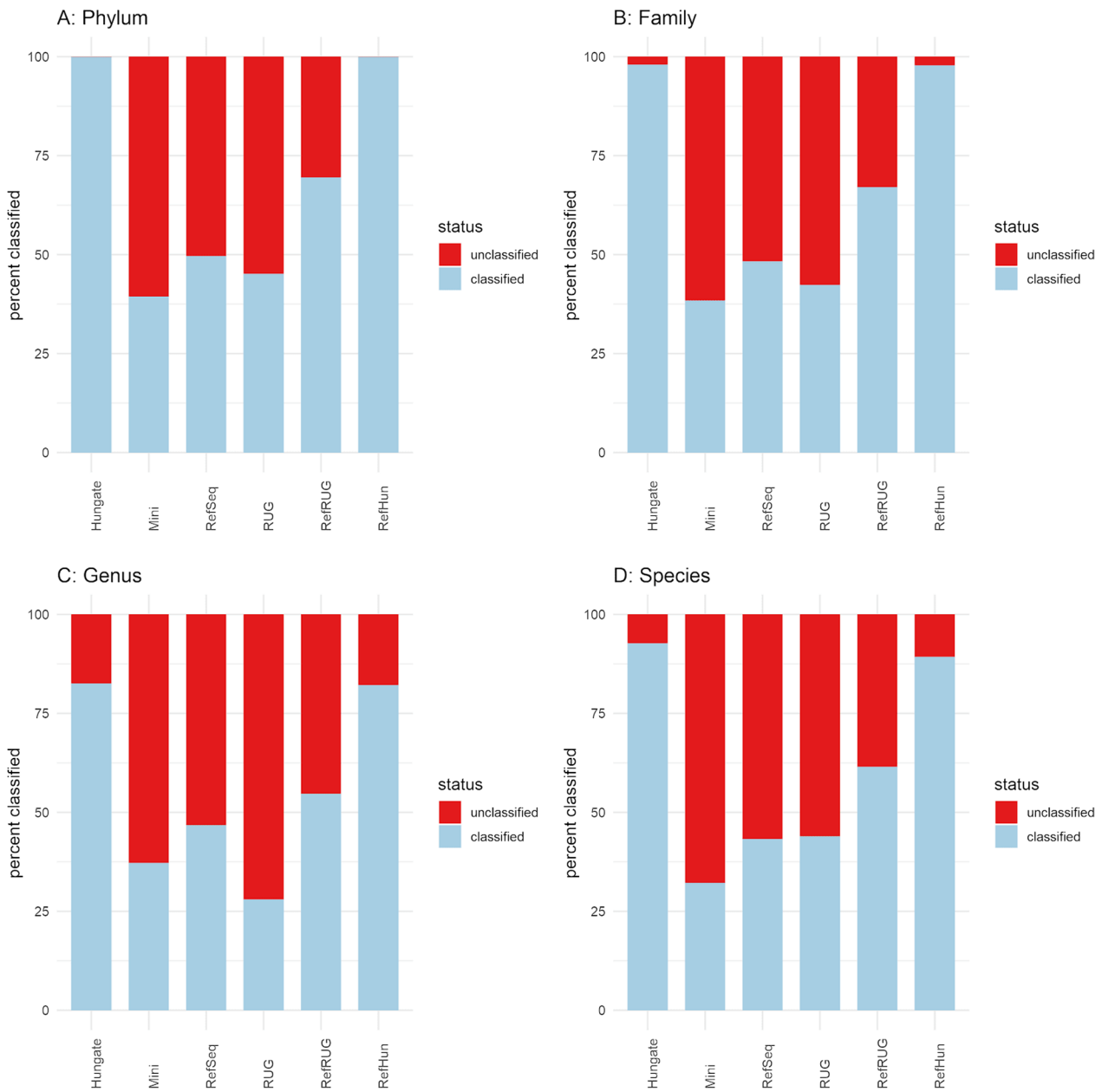


Fig. 2 Classification rate of reads, shown at various taxonomic levels for six reference databases. Classification rate refers to whether the reads were classified or unclassified, and are shown as a percentage at the **A** Phylum, **B** Family, **C** Genus and **D** Species levels. The y-axis shows the percentage of reads from the simulated dataset that were classified or unclassified when classified using Kraken2. The six reference databases used during classification are shown as bars plotted along the x-axis

database incorrectly classified almost the same proportion (20.53%), whereas using the Mini database incorrectly classified approximately half that amount (11.55%). As expected for a smaller database, classification with the Mini database had a higher proportion of reads that were unclassified at any level compared to RefSeq (60.15% vs. 49.72%). A summary of the number of genera and species in the ground truth data, and the number that were classified using each of the reference databases, is shown in Additional file 1: Fig. S1. Reference databases that include RefSeq (RefSeq, Mini, RefHun, RefRUG) classified thousands more false positives than databases that did not (Hungate, RUG). This did not improve when using the RUG database, as it failed to classify many genera and

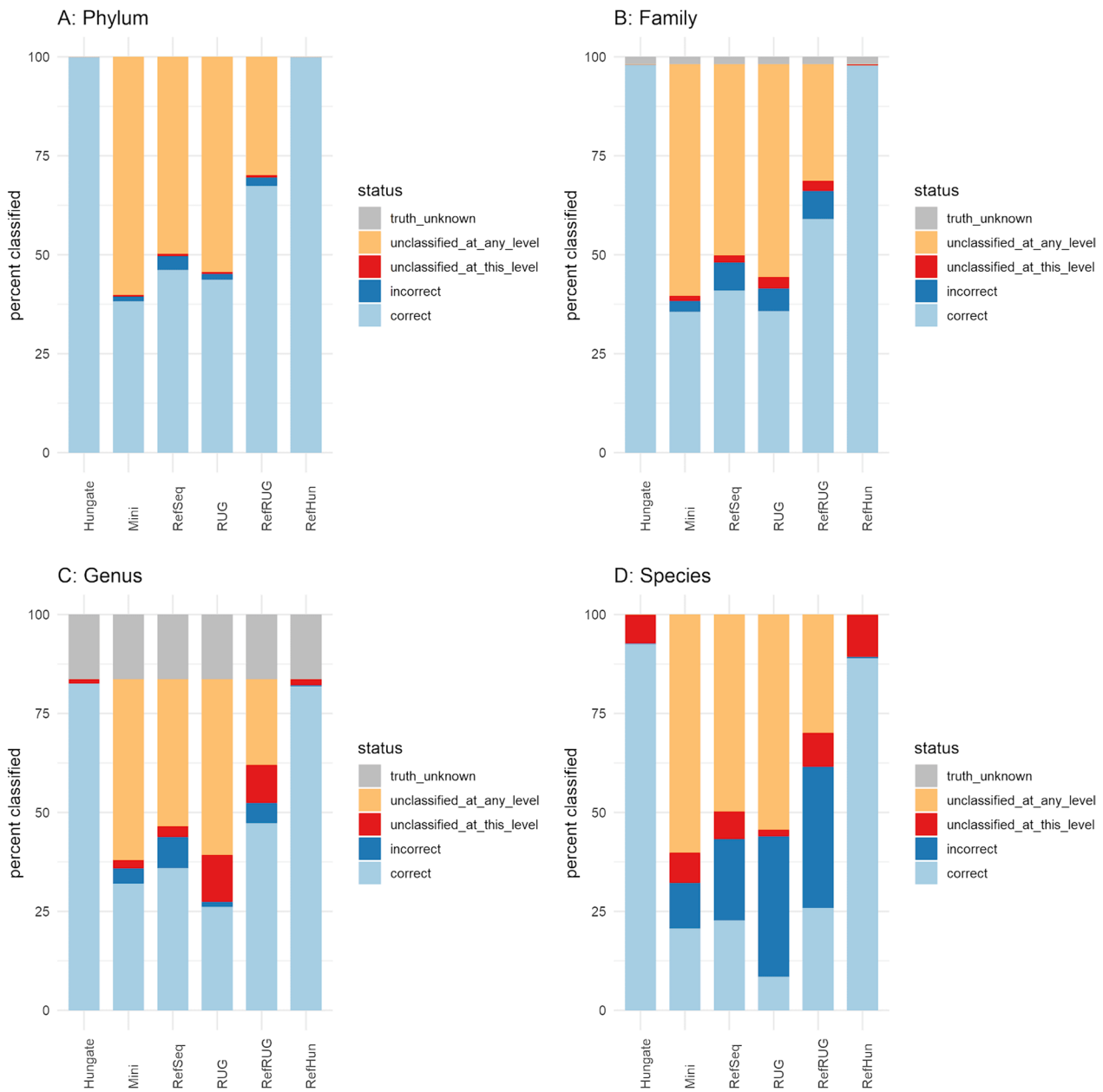


Fig. 3 The accuracy of taxonomic classification using each reference database and across the various taxonomic levels. Classification status of reads compared to the ground truth for six reference databases at various taxonomic levels. The graphs refer to the percentage of reads, shown along the y-axis, at the **A** Phylum, **B** Family, **C** Genus and **D** Species levels. Each bar represents reads classified by Kraken2, using each reference database as shown along the x-axis. The bars represent the percentage of classified reads at various classification status, as shown in the key. "Truth unknown" refers to the reads that originate from genomes that do not have an assigned family or genus. "Unclassified at any level" refers to reads that were not classified to any taxonomic level. "Unclassified at this level" refers to reads that were classified at other taxonomic levels, but not the level being examined in each graph. "Correct" and "incorrect" refer to reads that were classified correctly or incorrectly by Kraken2 using the respective database

species that were in the ground truth data. Additionally, classification of the data using the RUG database failed to classify any reads for certain abundant taxa.

After some investigation, it was discovered that there were marked differences in the annotated taxonomies present in the RUG and Hungate genomes, shown in Table 2. Several taxa were present in the Hungate data but were seemingly not present in the RUG data. As the Hungate collection contains highly abundant rumen microbial genomes, it is likely that these taxa are also

present in the assembled RUG genomes, but that their taxonomy is not accurately annotated. Further investigation revealed that this was indeed a result of some RUGs not having an assigned taxonomy at the family and/or genus levels. Examples are the family *Bacteroidaceae* and genus *Bacteroides*, which are both present in the Hungate data but not annotated as such in the RUG data, explaining why no reads were classified for these taxa at those levels.

The poor performance of RUGs at this level, as demonstrated in classification accuracy for the RUG database, also impacted the RefRUG database. Use of both reference databases including RUGs resulted in over 35% of reads being incorrectly classified. This can be explained by the use of generic species labels for the RUG dataset, which when compared to the formally named Hungate collection genomes in the ground truth were classified as incorrect. The addition of the RUG genomes to the RefSeq database (RefRUG) increased the percentage of correctly classified reads slightly, from 22.74 to 25.87%.

Once more, using the Hungate reference database resulted in the best performance, with the vast majority of reads classified correctly (92.56%), and only a small proportion of misclassifications (0.13%). There were, however, approximately 7% of reads that were not classified at the species level. The classification metrics when using the RefHun reference database were markedly closer to the results obtained when using the Hungate database than the RefSeq database. The addition of the Hungate genomes to the RefSeq database (RefHun) increased the percentage of correctly classified reads from 22.74 to 88.92%, and the decreased number of incorrectly classified reads from 20.53 to 0.35%, clearly demonstrating the huge gains in accuracy that can be obtained when closely matching sequences are present in reference databases.

Table 2 The frequency of families and genera in the Hungate and RUG datasets, and overlap between the two datasets

Status	Family	Genus
Present in Hungate but not RUG	25	48
Present in RUG but not Hungate	8	8
Present in both RUG and Hungate	23	33

Shown are the families and genera present in the Hungate and RUG datasets, including overlapping taxa. The Hungate data was used to generate the simulated data, and was included in the Hungate and RefHun reference databases. Similarly, the RUG data was included in the RefRUG and RUG reference databases.

Composition of the reference database used impacts upon the accuracy of taxonomic read classification and taxonomic read abundance

Having demonstrated that the accuracy of taxonomic read classification changes considerably depending on the reference database used, this study next examined the impact of reference database choice on the taxonomic abundance of a microbial community. This was done using the same simulated data and reference databases as before, but by examining classification results in the form of taxonomic read abundance. Figure 4 shows a selection of scatterplots that compare the taxonomic abundance of the ground truth simulated metagenomic data with that of the classified data. The closeness-of-fit of the taxonomic read abundance (Fig. 4) to the linear regression was measured using the R^2 statistic, and is shown in Fig. 5. The R^2 statistic summarises how similar the classified taxonomic abundance was to the taxonomic abundance of the ground truth simulated data, and is therefore another indication of classification accuracy using each of the reference databases at various taxonomic levels.

A cornerstone of microbiome research is community structure, which can be observed as a sample's taxonomic abundance. To investigate this, the most abundant taxa in the ground truth data were observed in the classified data. Barplots displaying the taxonomic read abundance of the ground truth data, as well as the read abundance once the data was classified using each of the reference databases, are shown in Fig. 6. Each plot shows the taxonomic distribution of the top 10 most abundant taxa for the ground truth data and the abundance of these taxa in the classified data, at that particular taxonomic level.

Overall, the Hungate and RefHun databases performed very well at classifying the data, as shown in Figs. 4, 5 and 6. There was a slight reduction in accuracy at the species level, where the R^2 value was 0.97, but this had little effect on the classification of abundant taxa (see Fig. 6). To further assess the beneficial impact of including representative genomes in the reference database, additional reference databases containing the Hungate and

(See figure on next page.)

Fig. 4 Comparing taxonomic abundance of the ground truth metagenomic data with that of the classified data. Scatterplots show the comparison between the simulated metagenomic data (ground truth, x-axis) and classified reads (y-axis). Data is plotted as a percentage of classified reads for the classified data, and a percentage of simulated reads for the ground-truth data. The data has been transformed by log10. A $y = x$ line (shown in red) has been added to demonstrate how data points would appear on the graph if the number of ground-truth and classified reads were the same. A linear regression has been added (shown in blue) and used to calculate the R^2 statistic, see Fig. 6. Comparisons are shown at the Phylum, Family, Genus and Species levels, for the Hungate, Mini, RefSeq, RUG, RefRUG and RefHun reference databases

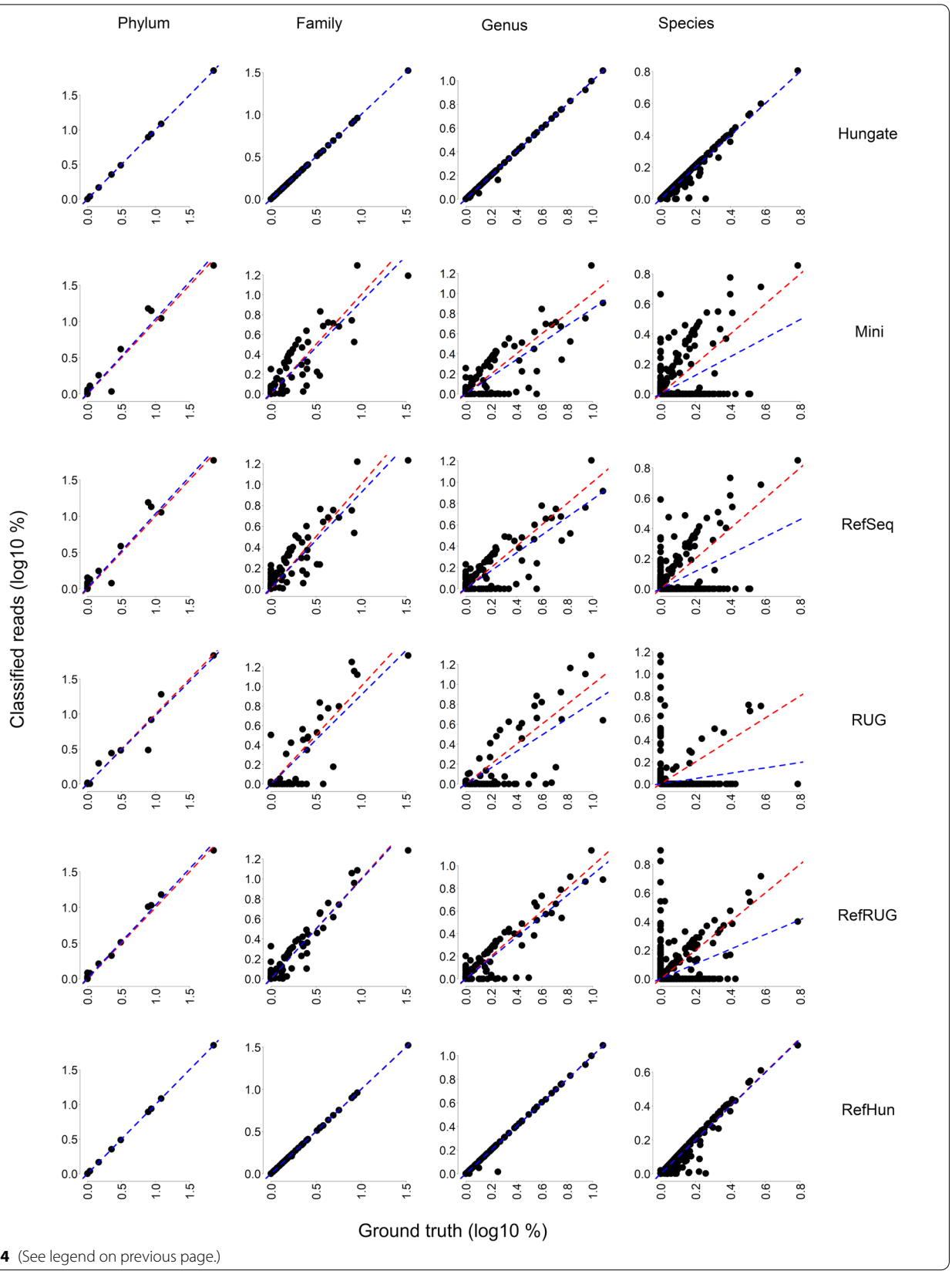


Fig. 4 (See legend on previous page.)

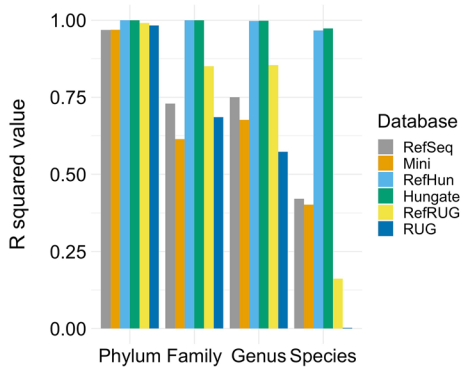


Fig. 5 R^2values of the comparisons between taxonomy of the simulated metagenomic dataset and classified taxonomy at various taxonomic levels. The key denotes each reference database used to classify the data, and these are shown as individual bars at each taxonomic rank, displayed on the x-axis. The R^2 value is the statistical measure of the correlation of data to the linear regression, measured using the scatterplots shown in Fig. 4

RUG genomes were made (see Additional file 1: Fig. S2). Specifically, we combined the Hungate and RUG databases into a new reference database ('HunRUG'), and also added RefSeq to the Hungate and RUG genomes ('RefHunRUG'). The results were overall very similar in accuracy to those observed previously with just the RefHun database (Additional file 1: Fig. S2), further emphasising the particularly beneficial impact of having well characterised reference sequences with full and accurate taxonomic labelling.

Using the RefSeq and Mini reference databases accurately classified the data at phylum level, but there was a distinct drop in accuracy at the class level, which continued further down the taxonomic levels. At the phylum level, the Mini and RefSeq databases over-estimated *Proteobacteria* and *Actinobacteria*, but under-estimated *Firmicutes*. At the family level, the Mini and RefSeq databases overestimated the *Streptococcaceae* and *Bifidobacteriaceae*, yet underestimated the *Lachnospiraceae* and *Erysipelotrichaceae*. At the genus level the Mini and RefSeq databases overestimated *Streptococcus* and *Bifidobacterium*, and underestimated *Ruminococcus* and *Prevotella*. At the species level, the RefSeq and Mini databases did not classify any reads to four of the ten most abundant species: *Clostridium clostridioforme*, *Lachnospira multipara*, *Ruminococcus flavefaciens* or *Kandleria vitulina*.

The RUG and the RefRUG databases were similarly accurate at the phylum level, but began to diverge in

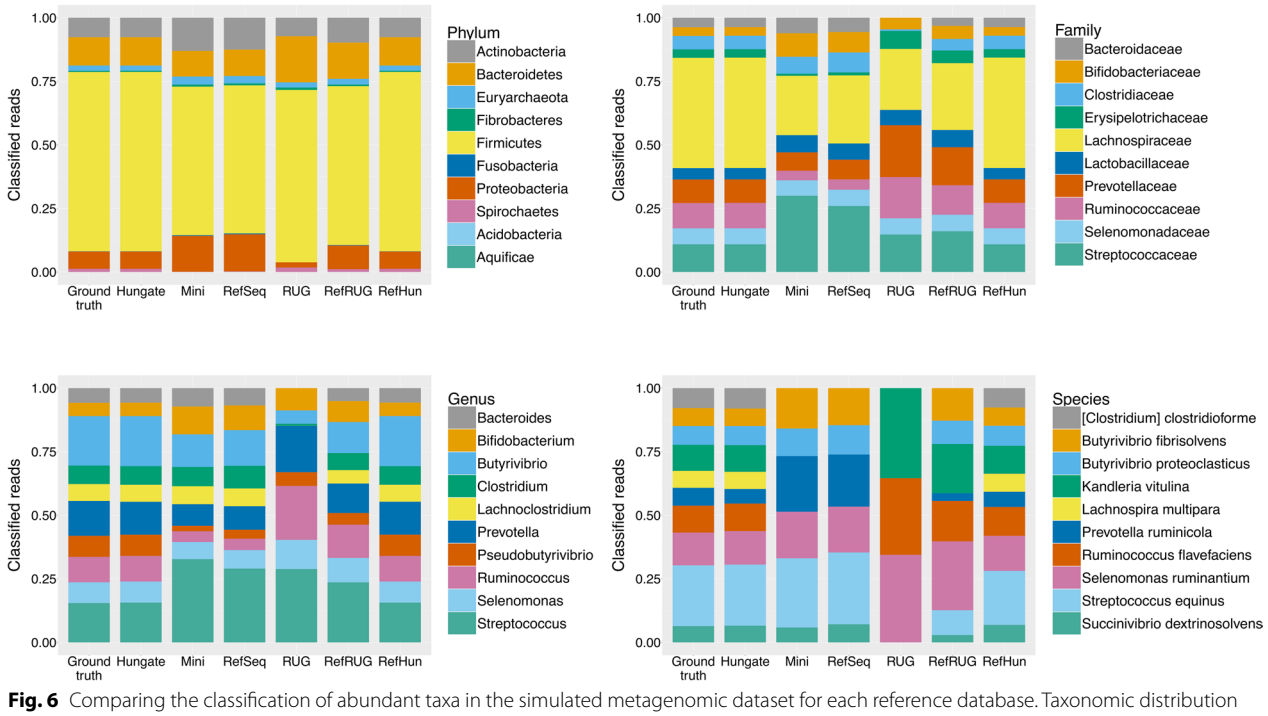


Fig. 6 Comparing the classification of abundant taxa in the simulated metagenomic dataset for each reference database. Taxonomic distribution for the top ten most abundant taxa in the simulated metagenomic dataset, classified at the Phylum, Family, Genus and Species levels with Kraken2 using six different reference databases. The y-axis denotes the percentage of reads classified at each level. The bars along the x-axis each represent the classification results for each database, split by taxonomy as shown in the keys for each level

classification accuracy at lower taxonomic levels. In general, the RefRUG database classified the data more accurately than the RUG database, and this was likely due to the issues surrounding taxonomic labelling of the RUGs, as described above. At the family level, the RUG database did not classify any reads as *Bacteroidaceae*, and at the genus level there were a lack of reads classified as *Bacteroides*. This was simply because these taxonomic labels do not appear in the RUG collection. At the species level, the RUG database classified just three of the top ten most abundant taxa in the simulated metagenome (Fig. 6). This resulted in a poor correlation in Fig. 4 and a very low R^2 value of 0.002 (Fig. 5). Interestingly, however, two out of the three species (*Ruminococcus flavefaciens* and *Kandleria vitulina*) were completely missed during classification by the RefSeq database, but were classified when the RUG data was added to the RefSeq database (RefRUG database). However, the species *Clostridium clostridioforme* and *Lachnospira multipara* were not classified when using the RefRUG reference database or indeed any databases other than Hungate or RefHun.

Discussion

Accuracy and rate of metagenomic data classification is heavily impacted by the choice of reference database

Research into microbiomes has increased substantially over the last two decades, driven by advances in DNA sequencing technologies. However, DNA-sequence based methods depend fundamentally on the quality of reference databases that are used to assign taxonomy or function to the sequence data. This study, which used a simulated metagenomic dataset, demonstrates the huge difference that choice of reference database can have on the accuracy of the results obtained. Kraken 2 was selected for this analysis as it is often reported to perform well when compared to other data classification software [36–38], has been previously used to test reference database impact [34], and allows for the creation and use of custom reference databases.

RefSeq, the open-access database from NCBI, is a popular choice of reference database when classifying metagenomic data. However, using the RefSeq database we show that less than 40% of reads at genus level, and less than 25% of reads at species level, were accurately classified (Fig. 3). Although this issue impacts all taxonomic levels, classification using these databases at the species level was particularly unreliable. When the data was classified using the RefSeq database, this study observed that nearly 50% of species taxonomy assignments were incorrect. This finding indicates that such a frequency of inaccurate classification may also be occurring in the many other studies that use the RefSeq database, compromising classification results. Use of the Mini database, which is optimised for use when there are limited computational resources available, also resulted in the classification of less than 40% of reads overall. This suggests that studies relying on the RefSeq or Mini database for classification will likely have a large proportion of inaccurate taxonomy assignments, which could impact strongly on subsequent interpretations and conclusions based on those results.

Genomes from cultured isolates derived from the environment of study hugely increase classification rate and accuracy

Current reference databases are hugely biased towards microbes that have been isolated from well-studied environments, such as the 20 microbial species contributing to 90% of the reference genomes in the ENA [31]. The rumen is an under-studied environment, which has consequently impacted the number of ruminant microbial reference genomes present in public databases such as NCBI RefSeq. At the time of writing, of the 460 Hungate genomes used to create the simulated data, only 119 are present in NCBI RefSeq. The Kraken "Standard" database contains a subset of NCBI RefSeq, and so the RefSeq database may not contain all 119 of these Hungate genomes.

The Hungate reference database used here contained all of the Hungate genomes, and so is fully representative of the data that was classified. As expected, classification with the Hungate database resulted in classification of the majority of reads, and was the most accurate out of all the databases. However, at the species level, 7.31% of reads were not classified. Interestingly, these reads were unclassified rather than incorrectly classified. This reduction in classification at the species level was likely due to the phenomenon described by Nasko et al.: the so-called "minimiser collision". This is where two distinct k-mers are minimised to identical minimisers (l-mers). In other words, if reads are highly similar, Kraken2 may be unable to distinguish between reference genomes at the species level, and so would assign taxonomy at the lowest common ancestor, therefore assigning taxonomy to a higher level [30].

In an attempt to understand the impact that including reference genomes from cultured representatives can have on classification accuracy of metagenomic data, we added the Hungate genomes to RefSeq, creating the RefHun reference database. Classification using the RefHun reference database showed significant improvements in classification rate and accuracy compared to the RefSeq database alone. This demonstrates that when classifying environmental data, classification accuracy can improve considerably by including more genomes derived from taxonomically well characterised cultured isolates in

reference databases. Continued efforts to isolate, and formally taxonomically characterise, previously uncultured microbes from the rumen microbiome, and indeed any other understudied environment, is likely to have significant benefits for the accuracy of metagenomics-based studies.

MAGs have the potential to improve metagenomic data classification even further, but are currently limited by their poorly defined taxonomy

While the addition of cultured isolate genomes clearly improves classification accuracy, it must be acknowledged that cultivation of microbes, and formally describing their taxonomy, are hugely time-consuming and labour-intensive activities [39]. Furthermore, many microbes may prove difficult to cultivate under laboratory conditions [40]. There are therefore significant bottlenecks that preclude the required widespread cultivation and characterisation of microbes. Therefore, the incorporation of MAGs, which can be generated without having to cultivate microbes in the laboratory, and can be done at far greater scale, in reference databases is an extremely promising additional or alternative avenue to improve classification of metagenomics datasets. In support of this, the addition of RUGs (MAGs) to the RefSeq database in this study (RefRUG) improved classification rate, which confirms the observations of other studies. Stewart et al. observed poor classification rates of rumen metagenomic data when using RefSeq, and reported the addition of Hungate collection genomes led to a classification rate increase of 2-fold, and the addition of RUGs led to an increase of 5-fold [13]. In a different study, Stewart et al. noted an increase of 10% in classification rate when adding Hungate collection genomes, and a 50–70% increase when adding RUGs to the reference database [17]. Xie et al. observed improvements in taxonomic classification rate with the addition of rumen MAGs to the reference database, compared with using Genbank and RMG entries alone [22].

Although addition of RUGs increased classification rate, using the RUG database resulted in the classification of reads with varying accuracy. In some respects, the effect was positive. For example, at the family and genus levels classification using the RUG database resulted in less reads being incorrectly classified than when using the RefSeq database. However, it is clear that there are likely to be significant issues with accuracy when using common current reference databases to classify metagenomic data. In this study, the ground truth information was available, which means we can say with certainty that some of the data was classified incorrectly. However, in real world scenarios, the correct taxonomy of the newly-sequenced data is of course unavailable, which means that the accuracy of classification results is difficult to quantify. We term such incorrectly classified reads as false positives, because in real world studies these incorrect classifications would be considered genuine. Marcelino et al. hypothesise that false positives occur as a result of conserved regions of reference genomes and sequence contamination in databases [35]. The use of each database classified some reads as false positives, although the highest number of false positives were classified by the reference databases containing RefSeq. In particular, classification using the RefSeq, Mini and RefRUG databases resulted in the apparent detection of thousands of species that were simply not there. The occurrence of false positives in this study indicates that false positives could be a common occurrence in metagenomic read classification.

More concerningly, addition of the RUG MAGs resulted in very poor overall classification accuracy, despite the addition of much more comprehensive reference material to the database. The likely explanation for this finding comes from the fact that, when the taxonomic labels in the Hungate and RUG data were compared at the family and genus levels, it was discovered that less than half of the total taxa were supposedly present in both datasets. As both data sets originate from the rumen, this is unlikely and is most probably a result of the incomplete and informal taxonomy labels used for the MAGs. This highlights the issue that reference sequences with incomplete or informal taxonomic labels may not be appropriate for classifying taxonomy. This issue can be resolved by ensuring all reference sequences, whether cultured isolate or MAG-derived, have complete, and accurate, labels across all taxonomic levels.

Taxonomy currently relies on consistent nomenclature to classify all organismal names across all living domains on Earth. NCBI taxonomy contained over 280,000 informal bacterial species (as of May 2017) [41, 42] and the NCBI databases contain 3760 genomes for unclassified or candidate bacteria at the time of writing. Issues arise when taxa are placed into a taxonomy database with informal names or incomplete lineages. For example, some of the Hungate collection genomes do not have an assigned rank at family or genus level. Additionally, assembled genomes (MAGs) often have an informal species name that does not follow traditional binomial nomenclature [43]. This issue was well demonstrated in this study, as classification using the RUG database failed to classify any reads from seven of the top 10 species in the ground truth data. This is surprising as these species are highly abundant in the rumen, and so you would expect to see them in the highly comprehensive RUG database. Of the 78 labels assigned at the species level by

the RUG database, 56 had informal names, for example "uncultured *Lachnospiraceae* bacterium RUG10034".

As MAGs are draft genomes, and can often be novel species or even novel clades, it can be difficult to correctly assign phylogeny and taxonomy. This is a significant problem, as metagenomics studies increasingly demonstrate that the rumen contains many genomes that cannot be easily placed into the current NCBI taxonomy. For example, Stewart et al. [17] found that of 4941 MAGs, 4303 could not be assigned a species, 3849 could not be assigned a genus, 1753 could not be assigned a family and 140 could not be assigned a phylum. However, this issue of uncertain phylogeny placement is not unique to MAGs, an example being the genus *Clostridium*, which has been demonstrated to actually consist of multiple genera [44]. While informal names may cause issues in the context of binomial nomenclature, there is still some value to providing sequences or taxa with some form of name or label. Namely, it allows for the tracing of the sequence or taxa across multiple studies. This has proved useful before, an example being the candidate TM7 phylum proposed by Rheims et al. in 1996 [45], which was identified using sequence-based approaches as being widespread in numerous environments before being renamed *Saccharibacteria* [46]. Regardless of whether genomes are derived from cultured isolates or MAGs, mistakes or gaps in taxonomic descriptors will impact the accuracy of taxonomic classification.

It has been suggested that a change in microbial taxonomy towards a genome-based approach would improve upon the current taxonomy [47, 48]. The Genome Taxonomy Database (GTDB) uses a genome-based taxonomy, assigning the taxonomy of genomes based on their phylogeny [49]. Glendinning et al. observed many discrepancies between the phylogeny of MAGs and NCBI taxonomy, which was not found when using GTDB [24].

Conclusion

In this study, we compared taxonomic classification results with ground truth simulated metagenomic data. Our results showed that classification rate, classification accuracy and taxonomic read classification were heavily impacted by the choice of reference database used. In particular, RefSeq alone is a poor choice for classifying ruminant metagenomic data. Notably, our results indicate the extent to which ruminant metagenomic data could be inaccurately classified, an issue that has the potential to affect all studies that use insufficient reference databases. We demonstrate that custom reference databases substantially improve classification accuracy, and that genomes derived from cultured representatives and MAGs improve classification rate in all cases, but only improve classification accuracy for levels in which they have assigned taxonomy. This highlights the opportunity of using MAGs to improve taxonomic classification results in under-characterised environments, but also emphasises the importance of complete taxonomic lineages for MAGs.

Methods

Simulation of known truth dataset

The composition of a given environmental microbiome sample is of course unknown, and so it is difficult to measure classification accuracy on metagenomic data. Instead, data of known composition ("ground truth data"), such as simulated datasets or mock communities [50] are typically used to assess accuracy.

Here, InSilicoSeq (version 1.4.6) was used to generate simulated metagenomic data: 50 million paired-end reads using the HiSeq model with an exponential distribution [51] from known sequences. The input genomes used to create the data were 460 publicly available bacterial and archaeal reference genomes from the Hungate collection [10]. Since some of the Hungate collection are multi-contig, they were treated as draft genomes during data generation, using the *--draft* option. Complete genomes with a single contig were treated as such, using the *--genomes* option. A list of the Hungate genome files, and which are single or multi-contig, can be found in Additional file 2: Table S3.

As the simulated reads originated from the Hungate genomes, each read had a corresponding genome and therefore corresponding taxonomy. In this study the simulated data is referred to as "ground truth", as the true taxonomy of each read is known. The number of reads simulated from each genome, and therefore for each taxonomy, were determined (using Ete3 [52]). The number of reads produced for each genome provided the number of reads produced for each taxon at the phylum, family, genus and species levels. This "ground truth" information was used to assess the classification accuracy of each read (see Figs. 3 and 4, and Additional file 1: Fig. S1 and Tables S1 and S2).

Design, choice and creation of reference databases

Six reference databases were used to classify the simulated metagenome, the details of which can be seen in Table 1. Each database was built using NCBI taxonomy downloaded on 07/03/2020. NCBI libraries for the RefSeq database were downloaded on 24/03/2020.

The Hungate reference database contains genomes from 460 rumen-dwelling microbes cultured in the Hungate 1000 project. These were the same genomes that were used to create the simulated metagenome; therefore, this database was fully representative of the data being classified. The Hungate database therefore acted as

the 'best case' scenario for database choice, and can be seen as a positive control, as each read from the simulated metagenome should be represented in the Hungate database.

The RefSeq database is the standard Kraken2 [30] reference database (see [53]) widely used for taxonomy classification. It contains the complete collection of genomes in RefSeq for bacterial, archaeal and viral domains, the human genome and a collection of vectors (UniVec_core).

The Mini reference database is also a popular database for Kraken2 users, designed for users with low-memory computing environments. Both the Standard and Mini databases contain the same RefSeq reference genomes, but the Mini database was built using a hash function to down-sample minimisers, as described in the Kraken 2 manual and shown in Table 1 (*--max-db-size function*). The hash file for the Standard Kraken 2 database is 43 GB, whereas it is only 7.5 GB for the Mini Kraken 2 database. As this database is significantly smaller than the Standard reference database, read classification requires less memory. As the Mini reference database may be the first choice for users with limited computational resources, it was included in this study.

The RUG reference database contains 4941 rumen MAGs assembled by Stewart et al. [17]. Whilst different from the cultured Hungate genomes, these assembled genomes were assembled from metagenomes also originating in the rumen. This custom database was included in the study to investigate the impact of a reference database containing assembled genomes on taxonomic classification.

The RefRUG and RefHun reference databases contain the complete collection of genomes in RefSeq (bacterial, viral and archaeal domains, the human genome and UniVec_Core vectors) in addition to the RUGs and Hungate genomes, respectively. These were included to investigate whether adding genomes or draft genomes from the same type of environmental microbiota as the data being classified improves taxonomic classification.

Read classification using Kraken2

The simulated metagenome was classified using Kraken2 (version 2.0.8_beta) with the eight reference databases described above. Default settings were used with the *--paired* option to accommodate the paired-end reads of the simulated metagenome.

Classification status was extracted from the Kraken output files and used to assign reads to one of two classes: classified or unclassified. The taxonomic ID for each read was extracted from the Kraken output files, and classified reads were compared to their known ground truth at the species, genus, family and phylum level (using Ete3). The reads were firstly grouped into "correct" or "incorrect" and then subsequently into "correct", "incorrect", "unclassified at this level", "unclassified at any level" and "truth unknown".

Finally, the Kraken 2 report files were used to compare read classification counts for each taxonomic level against the ground truth, and R^2 calculated as the sum-of-squares of absolute deviation from the ground-truth.

Abbreviations

MAG: Metagenome assembled genome; RUG: Rumen uncultured genome; NCBI: The National Centre for Biotechnology Information; ENA: European nucleotide archive.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42523-022-00207-7.

Additional file 1: Table S1. Classification rate of reads for six reference databases at various taxonomic levels. Classification rate refers to whether the read was classified, or unclassified, regardless of accuracy. Each row denotes the six databases used to classify reads with Kraken2. The "Overall" column refers to the percentage of reads that were classified or unclassified by Kraken2 regardless of taxonomic level. Subsequent columns refer to the percentage of reads that were classified or unclassified by Kraken2 at various taxonomic levels as shown in the column headers. **Table S2.** Classification status of reads compared to the ground truth for the six reference databases at various taxonomic levels. The databases and detailed classification status are shown in the first column. Subsequent columns contain the percentage of reads at that taxonomic level, which had been classified by the database and had the particular classification status outlined in the first column. "Correct" and "incorrect" refer to reads that were classified correctly or incorrectly by Kraken2 using the respective database. "Truth unknown" refers to the reads that originate from genomes that do not have an assigned family or genus. "Unclassified at any level" refers to reads that were not classified to any taxonomic level. "Unclassified at this level" refers to reads that were classified at other taxonomic levels, but not the level being examined in a given column. **Fig. S1** The frequency of genera and species in the ground truth data, and in the classification results for each reference database. The total frequency is shown in the top two graphs, the middle graphs show the frequency of false positives occurring, and the bottom two graphs show the frequency of false negatives. **Fig. S2** Scatterplots show the comparison between the simulated metagenomic data (ground truth, x-axis) and classified reads (y-axis) when classified using the HunRUG (A) and RefHunRUG (B) reference databases. Data are plotted as a percentage of classified reads for the classified data, and a percentage of simulated reads for the ground-truth data. The data were transformed by log10. A y=x line (shown in red) was added to demonstrate how data points would appear on the graph if the number of ground-truth and classified reads were the same. A linear regression was added (shown in blue) and used to calculate the R2 statistic. The R2 statistic is shown (C) for each reference database at the Phylum, Family, Genus and Species levels.

Additional file 2: Table S3 A list of the Hungate genome files used to create the simulated data. Shown in the table are the Hungate genome files used to create the simulated data. They are separated into the complete (single-contig) and draft (multi-contig) genomes, as this meant they were treated differently. The tool InSilicoSeq was used to create the simulated data, and has the capability to handle draft genomes. The draft, multi-contig genomes were used with the --draft option, and the complete, single-contig genomes were used with the --genomes option. These are the same files added to the custom databases containing Hungate genome sequences (Hungate, RefHun, RefHunRUG and HunRUG).

Acknowledgements
For the purpose of open access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We would like to thank all of those who were involved in creating and publicly sharing both the Hungate Collection data and the RUG data.

Author contributions
R.H.S. created the simulated data, conducted data analyses and bioinformatics, made figures, and contributed to writing the manuscript. M.W. conceived the study, carried out bioinformatics work and created figures. M.W., A.W.W. and L.G. supervised the project and contributed to writing the manuscript. All authors read and approved the final manuscript.

Funding
The Roslin Institute forms part of the Royal (Dick) School of Veterinary Studies, University of Edinburgh. This project was supported by the Biotechnology and Biological Sciences Research Council (BBSRC; BB/S006680/1, BB/R015023/1), including institute strategic program grant BBS/E/D/30002276. R.H.S. is supported by an EASTBIO studentship funded by BBSRC (BB/M010996/1). A.W.W. and the Rowett Institute receive core financial support from the Scottish Government Rural and Environmental Sciences and Analytical Services (SG-RESAS).

Availability of data and materials
The data used in this study was simulated using genomes from the Hungate Collection (see https://genome.jgi.doe.gov/portal/HungateCollection/HungateCollection.info.html).
The simulated metagenomic data is available at https://doi.org/10.7488/ds/3444.
The metagenomic assemblies (MAGs) used to create the RUG and RefRUG databases can be found in ENA under accession PRJEB31266 (http://www.ebi.ac.uk/ena/data/view/PRJEB31266).
Further information about the MAGs used to create the RUG database, such as genome metrics, can be found in the Stewart et al. publication [17].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no completing interests.

Author details
[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, Midlothian EH25 9RG, UK. [2]Rowett Institute, University of Aberdeen, Aberdeen AB25 2ZD, UK.

Received: 26 April 2022 Accepted: 24 September 2022
Published online: 18 November 2022

References

1. Kamra DN. Rumen microbial ecosystem. Curr Sci. 2005;89:124–35.
2. Auffret MD, Stewart RD, Dewhurst RJ, Duthie CA, Watson M, Roehe R. Identification of microbial genetic capacities and potential mechanisms within the rumen microbiome explaining differences in beef cattle feed efficiency. Front Microbiol. 2020;11:1–16.
3. Huws SA, Creevey CJ, Oyama LB, Mizrahi I, Denman SE, Popova M, et al. Addressing global ruminant agricultural challenges through understanding the rumen microbiome: past, present, and future. Front Microbiol. 2018;9:1–33.
4. Martínez-Álvaro M, Auffret MD, Stewart RD, Dewhurst RJ, Duthie CA, Rooke JA, et al. Identification of complex rumen microbiome interaction within diverse functional niches as mechanisms affecting the variation of methane emissions in bovine. Front Microbiol. 2020;11:1–13.
5. Roehe R, Dewhurst RJ, Duthie CA, Rooke JA, McKain N, Ross DW, 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. 2016;12:1–20.
6. Wallace RJ, Rooke JA, McKain N, Duthie CA, Hyslop JJ, Ross DW, et al. The rumen microbial metagenome associated with high methane production in cattle. BMC Genom. 2015;16:1–14.
7. Auffret MD, Stewart R, Dewhurst RJ, Duthie CA, Rooke JA, Wallace RJ, et al. Identification, comparison, and validation of robust rumen microbial biomarkers for methane emissions using diverse Bos Taurus breeds and basal diets. Front Microbiol. 2018;8:1–15.
8. Auffret MD, Dewhurst RJ, Duthie CA, Rooke JA, John Wallace R, Freeman TC, et al. The rumen microbiome as a reservoir of antimicrobial resistance and pathogenicity genes is directly affected by diet in beef cattle. Microbiome. 2017;5:1–11.
9. Henderson G, Cox F, Ganesh S, Jonker A, Young W, Janssen PH, et al. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Sci Rep. 2015;5:1–15.
10. Seshadri R, Leahy SC, Attwood GT, Teh KH, Lambie SC, Cookson AL, et al. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. Nat Biotechnol. 2018;36:359–67.
11. Creevey CJ, Kelly WJ, Henderson G, Leahy SC. Determining the culturability of the rumen bacterial microbiome. Microb Biotechnol. 2014;7:467–79.
12. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35:833–44.
13. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, et al. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen Robert. Nat Commun. 2018;9:1–11.
14. Rappé MS, Giovannoni SJ. The uncultured microbial majority. Annu Rev Microbiol. 2003;57:369–94.
15. Lewis WH, Tahon G, Geesink P, Sousa DZ, Ettema TJG. Innovations to culturing the uncultured microbial majority. Nat Rev Microbiol. 2021;19:225–40.
16. Watson M. New insights from 33,813 publicly available metagenome-assembled-genomes (MAGs) assembled from the rumen microbiome. Preprint at https://www.biorxiv.org/content/https://doi.org/10.1101/2021.04.02.438222v1.full (2021).
17. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37:953–61.
18. Solden LM, Naas AE, Roux S, Daly RA, Collins WB, Nicora CD, et al. Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. Nat Microbiol. 2018;3:1274–84.
19. Glendinning L, Genç B, Wallace RJ, Watson M. Metagenomic analysis of the cow, sheep, reindeer and red deer rumen. Sci Rep. 2021;11:3–12.
20. Wilkinson T, Korir D, Ogugo M, Stewart RD, Watson M, Paxton E, et al. 1200 high-quality metagenome-assembled genomes from the rumen of African cattle and their relevance in the context of sub-optimal feeding. Genome Biol. 2020;21:1–25.
21. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533–42.
22. Xie F, Jin W, Si H, Yuan Y, Tao Y, Liu J, et al. An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. Microbiome. 2021;9:1–20.
23. Svartström O, Alneberg J, Terrapon N, Lombard V, De Bruijn I, Malmsten J, et al. Ninety-nine de novo assembled genomes from the moose (Alces alces) rumen microbiome provide new insights into microbial plant biomass degradation. ISME J. 2017;11:2538–51.
24. Glendinning L, Stewart RD, Pallen MJ, Watson KA, Watson M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. Genome Biol. 2020;21:1–16.
25. Peng X, Wilken SE, Lankiewicz TS, Gilmore SP, Brown JL, Henske JK, et al. Genomic and functional analyses of fungal and bacterial consortia that enable lignocellulose breakdown in goat gut microbiomes. Nat Microbiol. 2021;6:499–511.

26. Li J, Zhong H, Ramayo-Caldas Y, Terrapon N, Lombard V, Potocki-Veronese G, et al. A catalog of microbial genes from the bovine rumen unveils a specialized and diverse biomass-degrading environment. Gigascience. 2020;9:1–15.
27. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science. 2011;331:463–7.
28. Gharechahi J, Vahidi MF, Bahram M, Han JL, Ding XZ, Salekdeh GH. Metagenomic analysis reveals a dynamic microbiome with diversified adaptive functions to utilize high lignocellulosic forages in the cattle rumen. ISME J. 2021;15:1108–20.
29. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:1–2.
30. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:1–13.
31. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. PLoS Biol. 2021;19:e3001421.
32. Méric G, Wick RR, Watts SC, Holt KE, Inouye M. Correcting index databases improves metagenomic studies. Preprint at https://www.biorxiv.org/content/10.1101/712166v1 (2019).
33. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733-45.
34. Nasko DJ, Koren S, Phillippy AM, Treangen TJ. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. Genome Biol. 2018;19:1–10.
35. Marcelino R, Holmes V, Sorrell EC. The use of taxon-specific reference databases compromises metagenomic classification. BMC Genom. 2020;21:1–5.
36. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 2017;18:1–19.
37. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep. 2016;6:1–14.
38. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking metagenomics tools for taxonomic classification. Cell. 2019;178:779–94.
39. Pallen MJ, Telatin A, Oren A. The next million names for archaea and bacteria. Trends Microbiol. 2021;29:289–98.
40. Walker AW. Microbiota of the human body. 2016;902:5–32.
41. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. Database. 2020;2020:1–21.
42. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Br Bioinform. 2018;20:1125–39.
43. Murray AE, Freudenstein J, Gribaldo S, Hatzenpichler R, Hugenholtz P, Kämpfer P, et al. Roadmap for naming uncultivated Archaea and Bacteria. Nat Microbiol. 2020;5:987–94.
44. Collins MD, Lawson PA, Willems A, Cordoba JJ, Fernandez-Garayzabal J, Garcia P, et al. The phylogeny of the genus clostridium: proposal of five new genera and eleven new species combinations. Int J Syst Bacteriol. 1994;44:812–26.
45. Rheims H, Rainey FA, Stackebrandt E. A molecular approach to search for diversity among bacteria in the environment. J Ind Microbiol Biotechnol. 1996;17:159–69.
46. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31:533–8.
47. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36:996.
48. Thompson CC, Amaral GR, Campeão M, Edwards RA, Polz MF, Dutilh BE, et al. Microbial taxonomy in the post-genomic era: Rebuilding from scratch? Arch Microbiol. 2015;197:359–70.
49. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38:1079–86.
50. Bokulich NA, Rideout JR, Mercurio WG, Shiffer A, Wolfe B, Maurice CF, et al. mockrobiota: a public resource for microbiome bioinformatics benchmarking. Msystems. 2016;1:e00062–6.
51. Gourlé H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. Bioinformatics. 2019;35:521–2.
52. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016;33:1635–8.
53. Wood DE. Kraken 2 Standard Reference Database. https://github.com/DerrickWood/kraken2/wiki/Manual#standard-kraken-2-database. Accessed 16 Mar 2020.

Publisher's Note

2.3: Conclusion

This work found that reference database choice had a large impact on classification results and accuracy. Unexpectedly, the standard Kraken2 (RefSeq) reference database classified the data with poor accuracy, and falsely “identified” thousands of species that were actually not present in the tested data. This is concerning as Kraken2 is a popular piece of software, and these results suggest that other research may unknowingly be impacted by poor classification accuracy and false positives. I hypothesise that this is likely to be true not only for the rumen, but for other environments who are not well represented in reference databases.

The addition of rumen MAGs to the reference database improved classification rate, but did not improve classification accuracy due to the MAGs’ own insufficient taxonomy. This work has demonstrated that including MAGs in reference databases could improve the classification accuracy of rumen data. However, for that to happen, MAGs need to have formal taxonomic labels and complete lineages.

Chapter 3: Assessing the suitability of Metagenome-assembled genomes for microbiome analysis

3.1: Introduction

3.1.1: Metagenome-assembled genomes could be used as reference sequences for uncultured species

The gold standard for creating microbial reference genomes is to culture the microbe in isolation, extract then sequence genomic DNA, and create a high-quality genome assembly (Seshadri *et al.*, 2018). Reference genomes, primarily derived from other genome-sequenced cultured isolates, are used for the taxonomic and functional classification of sequencing data (Almeida *et al.*, 2021). However, as many bacteria have never been cultured in the laboratory, culture-independent approaches, such as metagenomic data analysis, are the only way to examine the genomic sequences of these as-yet uncultured microbes. As an example, Parks *et al.* expanded the tree of life by over 30% with metagenome-assembled genomes (MAGs) assembled from non-human gastrointestinal samples including the rumen, with these cumulative genomes representing 20 candidate novel phyla of the bacterial and archaeal kingdoms (Parks *et al.*, 2017). Perhaps due to the difficulty of replicating the anaerobic rumen environment in the laboratory, the majority of the rumen microbiota remain as yet uncultivated (Zehavi *et al.*, 2018).

Efforts have been made to increase the representation of cultured species from the rumen, such as the Hungate Collection, which is a collection of several-hundred rumen isolated genomes (Seshadri *et al.*, 2018). However, the recent analysis of rumen metagenomic data has demonstrated how far the rumen microbiome extends beyond microbes that have been cultured to

date. Building on Chao 1 estimations (Chao, 1984) by Stewart *et al.* (Stewart *et al.*, 2019), Watson estimates that there are over 13,000 species in the rumen (Watson, 2021). By this estimation, collective culturing efforts thus far have resulted in the isolation of less than 3% of the total number of species in the rumen. The assembly of uncultivated genomes, such as MAGS, therefore provides the chance to analyse these species.

Resolving MAGs from metagenomic data involves a process of first assembling sequencing reads into contigs, before they are binned into putative genomes and dereplicated (Strous *et al.*, 2012; Sangwan *et al.*, 2016; Stewart *et al.*, 2019; Kang *et al.*, 2015). As of 2020, the Genomes OnLine Database (GOLD) contains 22,946 genomes derived from uncultured organisms, representing almost 6% of the organisms in GOLD version 8 (Mukherjee *et al.*, 2021). Techniques to evaluate the quality of a MAG include estimations of completeness and contamination, and often rely on the presence or absence of marker genes (Parks *et al.*, 2015). MIMAG, the Minimum Information about a Metagenome-Assembled Genome, is an effort to unify the definition of what metrics make a MAG (Bowers *et al.*, 2017). For a high-quality MAG, this is completeness of at least 90% and contamination of less than 5%.

3.1.2: Considerations when using MAGs

Previously in this thesis (Chapter 2), the suitability of using MAGs as reference sequences for taxonomic classification of rumen metagenomic data was evaluated. That work resulted in the conclusion that MAGs have the potential to represent novel species that would otherwise be absent from reference databases. However, using MAGs as reference genomes poses its own challenges. Due to the nature of metagenomic binning and assembly, sections of the genome that have coverage information that diverges from the rest of the genome may be mis-assembled. This may result in missing gene information in short-read assemblies. For example, highly conserved regions, repeating sequences and mobile genetic elements may be present in multiple species and so will have differing coverage information to other regions of the genome (Maguire *et al.*, 2020).

Few studies have compared the same genome when binned and assembled from metagenomic data (resulting in a MAG) to a genome derived from a cultured isolate of the same species as the MAG. Meziti *et al.* compared isolates and MAGs from the same faecal sample and found that, on average, MAGs represented approximately ¾ of core genes and ½ of variable genes in the population (Meziti *et al.*, 2021). The conclusion of their work was that, due to issues with assembly, the quality of MAGs may actually be lower than indicated by the currently used-quality metrics (for example, completeness and contamination determined by CheckM (Parks *et al.*, 2015)). Alneberg *et al.* compared single-amplified genomes and MAGs, finding genome pairs that shared 99.51% sequence identity on average. Although 3.6% of nucleotide bases were missing on average from the MAGs, due to binning error, the two methods produced accurate genomes from uncultivated bacteria (Alneberg *et al.*, 2018). In addition, Lloyd *et al.* suggested that, as the majority of microbiomes on Earth contain novel microbiota, they likely have as-yet undiscovered functionality. Therefore, metagenomic classification may not

provide accurate insights into the microbial functionality of these environments (G. *et al.*, 2018).

At the time of writing, there have been no published studies comparing metagenome-assembled and cultured genomes originating from the rumen environment. As the majority of the rumen microbiota have yet to be cultured, examining the genomes of the rumen microbiota relies on culture-independent sequencing such as metagenomic data and genome bins. In this chapter, cultured genomes were generated from isolates from the rumen, and MAGs assembled from rumen metagenomic data. Genomes that appeared to be highly similar based on nucleotide identity were compared to evaluate the differences between the cultured and metagenome-assembled genomes. In particular, differences in genome metrics, taxonomy information and gene information were compared to identify any differences. By comparing genomes isolated from the rumen with genomes assembled from rumen metagenomic data, this work sought to provide novel insights into any limitations, or indeed benefits, of using MAGs to study uncultivated genomes from the rumen.

3.2: Materials and methods

3.2.1: Cultured bacterial genomes from the rumen

3.2.1.1: Culturing rumen microbiota

In order to create a panel of reference genomes from cultured isolates, traditional microbiology cultivation methods were carried out on two rumen-derived samples. These samples were provided by Prof. Rainer Roehe from Scotland's Rural College (SRUC), and were chosen for culturing as the same samples also had metagenomic data that had been generated in previous studies at the Beef and Sheep Research Centre of SRUC *(Martínez-Álvaro et al.*, 2020; Auffret *et al.*, 2020). The metagenomic data from these samples is described in section 3.2.2.1. These samples are referred to as S1_2013 and S2_2017, and were collected from animals BCMS ID 500677 and 203344 respectively. Glycerol was added to the rumen samples at time of collection, and these were then frozen at -80^{o}C. These frozen rumen samples were transported to the Rowett Institute, U. Aberdeen on dry ice, and then stored at -80˚C until further use.

The rumen microbes were cultured using two growth media, one with glucose, soluble starch and cellobiose as carbon sources (M2GSC) and one with galactose as a carbon source (M2SGa). This difference in carbon source was to increase biodiversity in the cultured microbes, by encouraging the growth of bacteria with different metabolisms. Details of the composition of each growth medium are shown in Table 3.1. All ingredients except cysteine were weighed into a Pyrex conical glass flask. The mineral solutions, resazurin 0.1% solution, rumen fluid and water were then added to the flask before it was covered with foil and placed in a waterbath. Once boiling, the flask was kept in the waterbath for 10 minutes. As the aim was to culture anaerobic rumen microbes, all media was bubbled with CO_2 to ensure the media remained free of oxygen. During the CO_2 bubbling of the media, the reducing agent cysteine was slowly added to keep the media anaerobic. For liquid media, 7.5 mL of the media was dispensed into Hungate tubes that had been flushed with CO_2. As the CO_2 tube was removed from the Hungate tubes, they were sealed with butyl rubber stoppers and screw caps. The Hungate tubes were then autoclaved at 121˚C for 15 minutes, and were checked to ensure the media remained uncontaminated before being used for culturing. For agar plates, the media were made as described above, with the addition of 8.0 g of agar powder. The plates were poured inside an anaerobic chamber to prevent contamination with oxygen.

Table 3.1: Composition of growth media used to culture rumen microbes. Two distinct media were prepared, one without galactose (M2GSC) and one with galactose (M2SGa) and stored at 4˚C. Two mineral solutions were added that had been previously made and stored at 4˚C, the composition of these solutions is shown in Table 3.1b. * Agar was included for culturing onto plates, but was omitted when microbes were cultured in liquid media in Hungate tubes.

Table 3.1a

Component	M2GSC	M2SGa
Bacto casitone	4.0 g	4.0 g
Yeast	1.0 g	1.0 g
$NaHCO_3$	1.6 g	1.6 g
Glucose	0.8 g	None
Galactose	None	0.8 g
Starch	0.8 g	0.8 g
Cellobiose	0.8 g	None
Rumen fluid	120 mL	120 mL
Min I	60 mL	60 mL
Min II	60 mL	60 mL
Resazurin (0.1%)	0.4 mL	0.4 mL
Agar *	8.0 g	8.0 g
Cysteine HCl	0.4 g	0.4 g
d. H_2O (to 400 mL)	~160 mL	~160 mL

Table 3.1b

Component	Mineral solution 1 (Min I)	Mineral solution 2 (Min II)
K_2HPO_4	3.0 g	None
KH_2PO_4	None	3.0 g
$(NH_4)_2SO_4$	None	6.0 g
NaCl	None	6.0 g
$MgSO_4$	None	0.6 g
$CaCl_2$	None	0.6 g
d. H_2O (to 1 L)	~997 mL	~983.8 mL

The methodology for culturing microbes from the rumen samples is described in Figure 3.1. Sealed tubes containing an aliquot of each rumen sample were defrosted for four hours at 4˚C, before being placed inside an anaerobic workstation (Don Whitley Scientific Ltd, Bingley, UK).

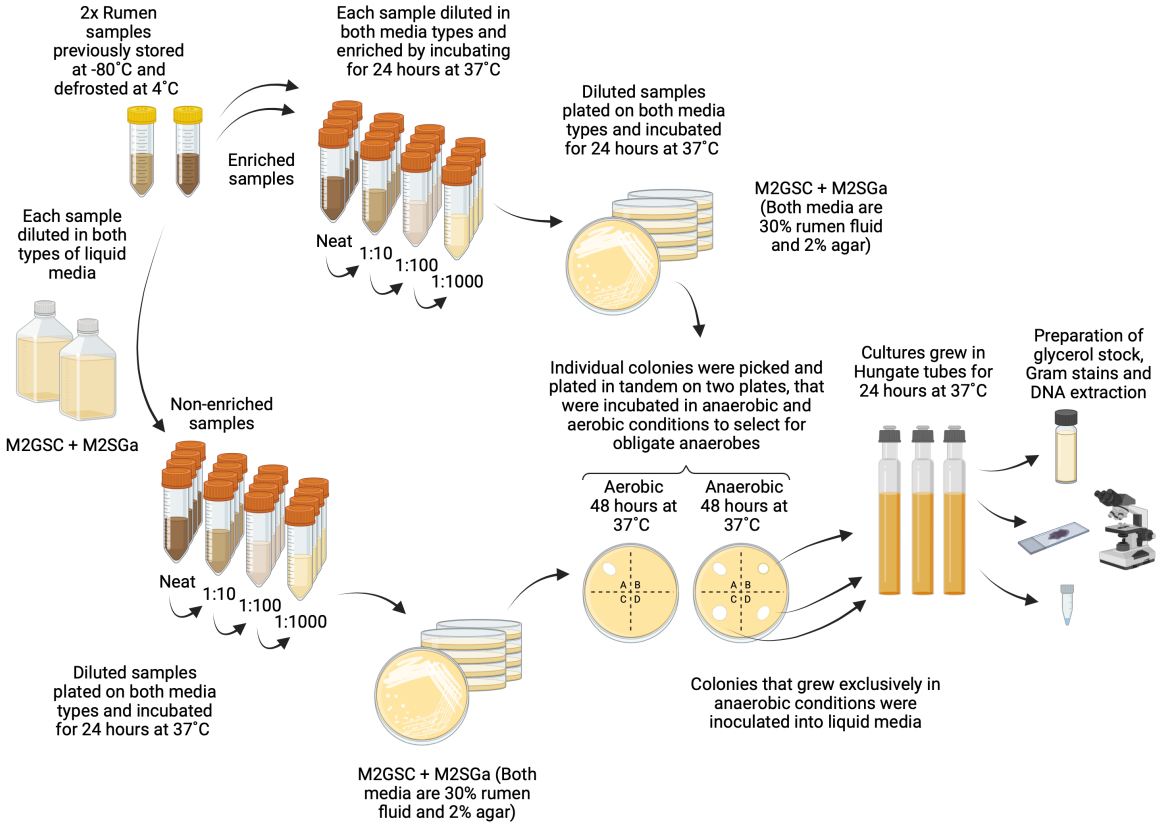


Figure 3.1: Methodology for culturing anaerobic microbes from rumen samples. With the aim of improving bacterial growth, the microbes were cultured with two slightly differing methods. One where microbes were enriched in liquid media for 24 hours at 37˚C before being plated onto agar plates, and the other which were plated onto agar plates immediately after dilution, to allow for the growth of bacteria without enrichment. Microbes were cultured anaerobically using two types of growth media, the compositions of which are described in Table 3.1. (Figure was created with BioRender.com).

The defrosted samples were then cultured on agar plates in serial dilution. Individual colonies were then picked with a sterile loop and plated in tandem onto corresponding positions on two agar plates per media type (n=4 total, 2x M2GSC and 2x M2SGa). These were then incubated at 37˚C for 48 hours, one in aerobic conditions (i.e. plates were taken out of the anaerobic cabinet and placed in a regular 37ºC incubator) and one in anaerobic conditions (i.e., the plates stayed within the anaerobic workstation). Those colonies that grew in aerobic conditions were discarded as the focus here was on cultivating anaerobes, which are the most abundant members of the rumen bacterial microbiota (Creevey *et al.*, 2014). To select for obligate anaerobes, individual colonies that only grew in anaerobic conditions were preferentially selected, although some colonies that grew both aerobically and anaerobically were also picked, and inoculated into a Hungate tube with a sterile loop of culture. These were incubated for 24 hours at 37˚C. Each colony underwent Gram staining and subsequent microscopy, and 0.85 mL of culture was added to 0.15 mL of glycerol to make a glycerol stock for long term storage at -80˚C.

The DNA of each isolate was extracted using the FastDNA™ SPIN Kit for Soil by MP Biomedicals (Irvine, California, USA) following the manufacturer's protocol with the following changes. Firstly, as this sample was not soil, instead of adding up to 500 mg of soil to a Lysing Matrix E tube as per the protocol, 1 mL of bacterial culture sample was spun down in a microcentrifuge at 14,000 x g for 10 minutes, and the pellet was resuspended in 978 µL of Sodium Phosphate Buffer and transferred to the Lysing Matrix E tube. Secondly, step 10 of the protocol calls for 500 µL of supernatant to be discarded, and for this study 1 mL was discarded. Lastly, step 11 was modified such that instead of 600 µL of the DNA and binding Matrix mixture being added to a SPIN™ filter and centrifuged at 14,000 x g for 1 minute, 750 µL of the mixture was added.

The 16S rRNA genes from the individual isolated cultures were amplified using PCR (Polymerase chain reaction) by Gillian Donachie (Rowett Institute, University of Aberdeen). Briefly, the extracted DNA was amplified using a Q5 High-fidelity Polymerase kit and dNTP (Deoxynucleotide) solution mix (both New England BioLabs, Massachusetts, USA) with 7f (AGAGTTTGATYMTGGCTCAG) and 1510r (ACGGYTACCTTGTTACGACTT) primers. The PCR reactions were in 50 µL volumes consisting of 10 µL of 5X Q5 Buffer, 1 µL of 10 mM dNTPs, 2.5 µL of 10 µM 7f Primer, 2.5 µL of 10 µM 1510r Primer, 2 µL of template DNA, 0.5 µL of Q5 Taq polymerase and 31.5 µL of nuclease-free H20. In a BioRad thermal cycler the solution was kept at 98°C for 2 minutes, before being cycled 35 times through 98°C for 30 seconds, 50°C for 30 seconds and 72°C for 2 minutes. After cycling, the solution was stored at 72°C for a further 10 minutes, before being held at 10°C until removed.

The amplification was confirmed by running 10 µL of PCR product on a 1% agarose gel alongside a 100 bp ladder (New England BioLabs, Massachusetts, USA). The PCR products were cleaned using the Promega Wizard® SV Gel and PCR Clean-up system (Wisconsin, USA) following the manufacturer's protocol for processing PCR products by centrifugation. The cleaned-up amplified DNA samples were prepared for quantification using the Qubit dsDNA HS (high sensitivity) Assay Kit (Thermo Fisher Scientific, Massachusetts, USA) and quantified in a Qubit 3.0 fluorometer (Invitrogen™, Massachusetts, USA). The samples were then prepared for sequencing by diluting with nuclease-free H_20 to ensure they were at a concentration of 25 ng/µL, and equal volumes of DNA solution and 926r (CCGTCAATTCMTTTRAGT) sequencing primer were added to a new Eppendorf tube. Samples were sent to Eurofins Genomics (Ebersberg, Germany) for LightRun Sanger barcode sequencing.

The 16S rRNA gene sequences were first visualised in Chromas (see http://technelysium.com.au/wp/chromas/) to identify and omit poor quality sequences at the start and end of the trace. The FASTA sequences were then run against the NCBI BLAST (Johnson *et al.*, 2008) and RDP databases (Maidak *et al.*, 1997) to identify the taxonomy of each isolate. Colonies that appeared mixed due to the presence of more than one peak pattern in the chromatogram traces were re-streaked and re-picked by Gillian Donachie. As the 16S rRNA gene results showed little diversity, the rumen samples were cultured again following an ethanol shock treatment. This treatment was done to select for spore-forming bacteria, which may not have been captured during the initial culturing. Please note the ethanol shock treatment, culturing work and DNA extraction of these samples was done by Gillian Donachie. The ethanol shock treatment and subsequent culturing methods are illustrated in Figure 3.2.

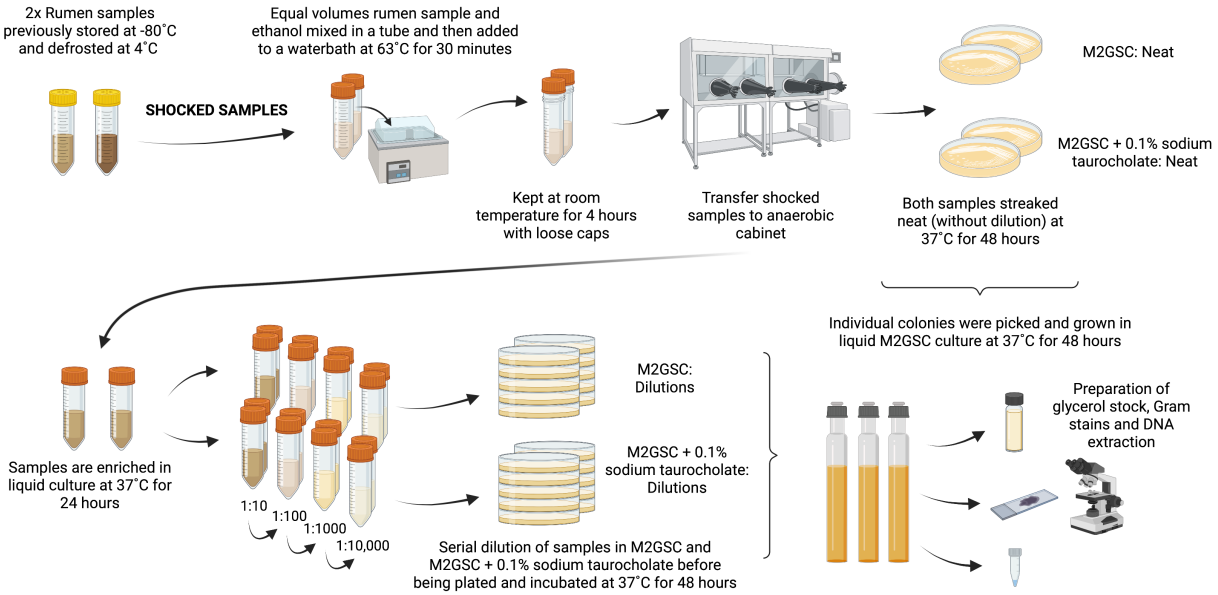


Figure 3.2: Methodology of the ethanol shock treatment to select for spore-forming bacteria. An aliquot of each defrosted rumen sample was treated with ethanol and then cultured anaerobically with two types of growth media. One was M2GSC, which is described in Table 3.1, and the other was M2GSC + 0.1% sodium taurocholate. Bacteria were enriched in liquid culture for 24 hours before being diluted from 1:10 to 1:10,000 and streaked onto agar plates of the same growth media. Individual colonies were then picked into sterile Hungate tubes with M2GSC media. Once good turbidity was observed, a glycerol stock and Gram stain of each isolated colony was prepared, and DNA was extracted. (Figure was created with BioRender.com).

The DNA of the cultured isolates that had undergone the ethanol shock treatment was extracted with the Wizard® Genomic DNA Purification Kit by Promega, following the manufacturer's instructions. The "DNA rehydration solution" in this kit was used to rehydrate and store the DNA. This reagent contains ETDA (Ethylenediaminetetraacetic acid), which is not compatible with the sequencing chemistry used by MicrobesNG. The DNA of these samples was therefore re-eluted into DNAse free water by myself using Beckman Coulter™ Agencourt AMPure XP beads, as depicted in Figure 3.3. The samples were then prepared for quantification using the Qubit dsDNA HS (high sensitivity) Assay Kit (ThermoFisher Scientific, Massachusetts, USA) following the manufacturer's instructions and quantified with a Qubit 4 fluorometer (Invitrogen™, Massachusetts, USA), to ensure they met the

criteria as requested by MicrobesNG for sequencing, before being transported on dry ice to MicrobesNG to be sequenced.

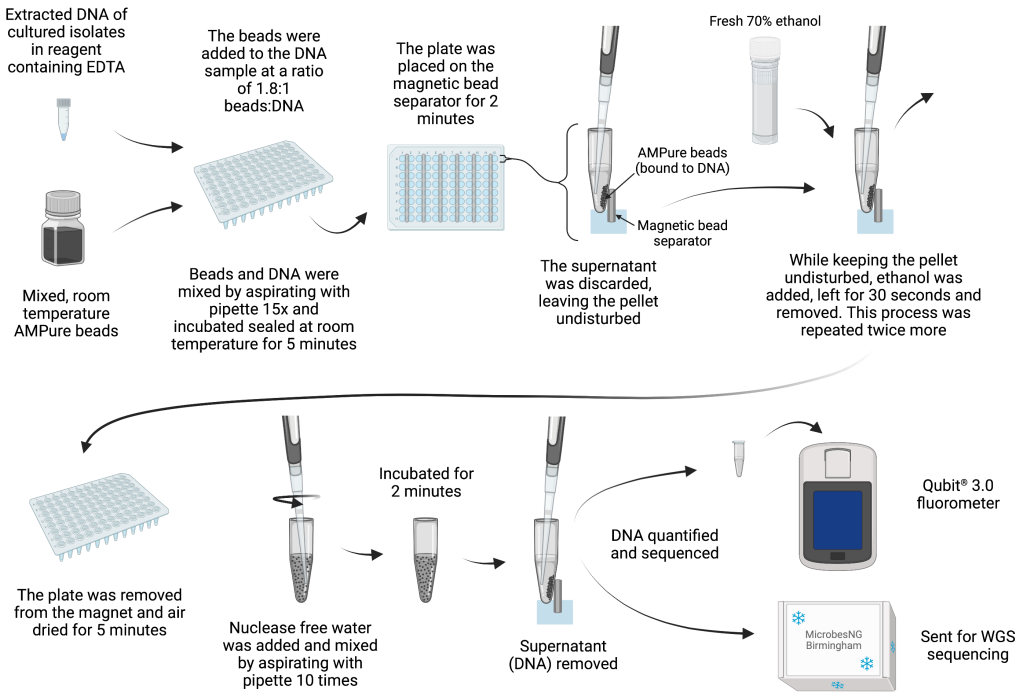


Figure 3.3: DNA that was stored in unsuitable EDTA-containing buffer was re-eluted into a suitable buffer prior to sequencing. Some of the DNA samples were eluted into a solution containing ETDA (Ethylenediaminetetraacetic acid) which needed to be re-eluted into DNAse-free water. (Figure was created with BioRender.com).

3.2.1.2: Sequencing the cultured isolate bacterial genomes

DNA samples were sent to MicrobesNG (Birmingham, UK. http://www.microbesng.com), and the following methods until specified otherwise were conducted by MicrobesNG. The DNA was quantified using the Quant-iT dsDNA HS kit assay (ThermoFisher Scientific, Massachusetts, USA) in an Eppendorf AF2200 plate reader (Eppendorf UK Ltd, UK). The DNA libraries were prepared using the Illumina Nextera XT Library Prep Kit (Illumina, San Diego, California, USA) on a Hamilton Microlab STAR

automated liquid handing system (Hamilton Bonaduz AG, Switzerland). The manufacturer's protocol was followed with the input DNA being increased 2-fold and PCR elongation increased to 45 seconds. The libraries were pooled and quantified using the Kapa Biosystems Library Quantification Kit for Illumina (Roche, California, USA), before being sequenced using Illumina technology, to produce paired-end reads of 250 bp in length. MicrobesNG use Trimmomatic (version 0.30 with a sliding window quality cut-off of Q15) (Bolger *et al.*, 2014) to remove adapter sequences. In addition to the trimmed reads, assemblies and annotations were produced by MicrobesNG in-house using SPAdes (version 3.7) (Bankevich *et al.*, 2012) and Prokka (version 1.11) (Seemann, 2014).

The trimmed reads were assembled by myself using SPAdes (version 3.14.1) and the isolate option (--isolate flag) as recommended for high-coverage isolate data. This was to ensure the assemblies taken forward were of high quality. The genome assemblies that were assembled by MicrobesNG and by myself were assessed using QUAST (version 5.0.2) (Gurevich *et al.*, 2013), the results of which are shown in section 3.3.1.1. The assembly metrics for each genome, were compared for each method. Although they were highly similar, the average largest contig, total length and N50 values were higher for the assembles assembled by myself, so these were used for all subsequent analyses.

3.2.1.3: Assessing the quality of the cultured isolate bacterial genomes and filtering

The quality of the isolate bacterial genomes was assessed using CheckM-genome (version 1.0.18) (Parks *et al.*, 2015) with the lineage-specific workflow. CheckM measures how complete a genome is, named 'completeness', how contaminated a genome is, named 'contamination', and the strain heterogeneity of a genome. These are measured as a percentage,

where a genome that was complete and had no contamination or strain heterogeneity would be considered a singular genome of high quality. As determined by CheckM, most genomes had some level of contamination, which was investigated using the pipeline MAGpy (update released June 2021) (Stewart *et al*., 2019). MAGpy predicts proteins using Prodigal (Hyatt *et al*., 2010) and compares these against a UniProt (The UniProt Consortium, 2021) DIAMOND (Buchfink *et al*., 2015) database, summarising the contigs containing proteins that map to the proteins in the database at the genus level. For each genome, the CheckM and MAGpy results were considered on a case-by-case basis.

For example, if a genome had contamination of near to 100% and the MAGpy results showed contigs were mapping to two distinct genera, it was deemed that this sample could be a mixture of two genomes. To determine if these genomes were complete, the contigs mapping to each distinct genus were separated into individual FASTA files. For genomes that had between 0.01-95% contamination as determined by CheckM, the majority of the contigs mapped to the same genus, with most mapping to the same species. In these instances, the contigs which mapped to other genera were removed from the FASTA file. These now-filtered and separated genomes were then re-assessed using CheckM, to ensure they were complete and high quality (n=77). The completeness, contamination, and strain heterogeneity of the genomes prior to and after filtering/separation are shown in Supplementary Table 3.1. The taxonomy of the cultured isolates was determined using GTDB-Tk (version 1.6.0) (Chaumeil *et al.*, 2019) with the classify workflow.

3.2.2: Metagenome-assembled-genomes (MAGs) from rumen metagenomic data

3.2.2.1: Rumen samples, metagenomic data and pre-assembly processing

MAGs were assembled from two metagenomic files that correspond to the same two rumen samples that isolates were cultured from (see Section 3.2.1.1). Animal experiments were conducted at the Beef and Sheep Research Centre of Scotland's Rural College (SRUC). The first was sample 1, referred to as S1_2013 as it was collected in 2013, which is the sequencing data of the rumen of BCMS animal ID 500677. The raw sequence reads are available under European Nucleotide Archive (ENA) project PRJEB21624. The second was sample 2, referred to as S2_2017 as it was collected in 2017. It was collected from the rumen of BCMS animal ID 203344. Both samples were prepared for sequencing by Edinburgh Genomics (TruSeq libraries) and sequenced using Illumina HiSeq technology producing paired-end reads 100 bp in length.

For both samples, Trimmomatic (version 0.36) (Bolger *et al.*, 2014) was used to remove the Illumina adapters. As the rumen samples may have contained host cells, the data was mapped to the cow reference genome and any mapping reads were removed from the data. The cow reference genome (species *Bos taurus*) was downloaded from NCBI (*https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/263/795/GCF_002263795.1_ARS-UCD1.2/GCF_002263795.1_ARS-UCD1.2_genomic.fna.gz*). Additionally, the phage *PhiX* can be routinely added to samples during metagenomic sequencing, therefore the *PhiX* reference genome was also downloaded from NCBI (*ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/enterobacteria_phage_phix174_sensu_lato_uid14015/NC_001422.fna*). The cow and *PhiX* reference sequences were indexed using bwa index -p (version 0.7.17) (Li and Durbin,

2009) and the reads were mapped to the reference using bwa-mem (version 0.7.17). Any reads that mapped to either reference were removed from the data.

3.2.2.2: Metagenomic assembly, binning, and assessing quality

Each of the two rumen metagenome samples was processed as a single sample, i.e. no co-assembly. Sequencing reads were assembled into contigs using MetaSPAdes (version 3.15.3, python version 3.6.6) (Nurk *et al.*, 2017). The contigs were then indexed using bwa index -p (version 0.7.17), and reads were mapped to the assemblies using bwa-mem (version 0.7.17). Coverage information was calculated firstly by making a BAM file using SAMtools (version 1.14) (Li *et al.*, 2009) and then the MetaBAT command jgi_summarize_bam_contig_depths. Contigs were then placed into bins using MetaBAT 2 (version 2.15) (Kang *et al.*, 2019). The quality of the bins was assessed by CheckM (CheckM-genome version 1.1.3) (Parks *et al.*, 2015) (with dependencies hmmer version 3.3.2 (*hmmer.org*), pplacer version 1.1alpha19 (Matsen *et al.*, 2010), prodigal version 2.6.3 (Hyatt *et al.*, 2010), numpy version 1.21.2 (Harris *et al.*, 2020), matplotlib=3.5.0 (Hunter, 2007), pysam version 0.17.0 (*https://github.com/pysam-developers/pysam*), and dereplicated with dRep (version 3.2.2) (Olm *et al.*, 2017). In addition, dRep uses software internally and these dependencies were Prodigal (version 2.6.3) (Hyatt *et al.*, 2010), ANIcalculator (version 1) (Varghese *et al.*, 2015), MUMmer (version 3.23) (Kurtz *et al.*, 2004), Centrifuge (version 1.0.3-beta) (Kim *et al.*, 2016), MASH (version 1.1) (Ondov *et al.*, 2016) and the environment used Python 3.6.6 (*python.org*). The taxonomy of each MAG was determined using the MAGpy pipeline (update released June 2021) (Stewart *et al.*, 2019) and GTDB-Tk (version 1.6.0, release 202) *(Chaumeil et al.*, 2019).

3.2.3: Comparing culture-derived and metagenome-assembled genomes

3.2.3.1: Clustering culture-derived and metagenome-assembled genomes

Genomes were clustered using dRep (version 3.2.2, and the same dependencies as above) (Olm *et al.*, 2017). The culture-derived genomes described in section 3.2.1 (n=77) were dereplicated and clustered alongside the Hungate genome collection (n=460), which is a collection of culture-derived reference genomes from the rumen (Seshadri *et al.*, 2018). The Hungate genome collection can be found on the Joint Genome Institute genome portal (see *https://genome.jgi.doe.gov/portal/HungateCollection/HungateCollection.info.html*). These genomes were dereplicated and clustered with an ANI (average nucleotide identity) of 95%, which was done to see whether any of the bacteria isolated were different to those in the Hungate collection.

To determine pairs of genomes, one culture-derived and one assembled from metagenomic data, the culture-derived genomes described in section 3.2.1 (n=77), the MAGs described in section 3.2.2 (n=69 that had >80% completeness and <10% contamination) and the rumen MAG superset (n=4,941, referred to as the 'RUG superset' or 'RUGs' and described in Stewart *et al*. 2019 (Stewart *et al.*, 2019), were dereplicated and clustered with an ANI of 99% to cluster genomes of the same strain.

3.2.3.2: Choosing clusters to take forward for comparison

Once the cultured genomes described in section 3.2.1, the MAGs described in section 3.2.2, and the RUG superset were dereplicated and clustered using an ANI of 99%, the resulting clusters were considered to contain genomes of the same strain. Clusters containing a culture-derived genome and a rumen MAG from either the MAGs described in section 3.2.2 or the RUG superset, were considered for comparison. Within these clusters, one culture-derived genome and one MAG were chosen as representative genomes for that cluster and strain. The representative genomes were chosen by selecting the genomes with the highest completeness and lowest contamination.

3.2.3.3: Comparing seven pairs of culture-derived and metagenome-assembled genomes of the same strain

Each pair of clustered genomes was compared to see what the differences were between a culture-derived and a metagenome-assembled genome. The genomes were aligned using progressive Mauve (version 2.4.0) (Darling *et al.*, 2004), and the contigs of the MAG were re-ordered using the culture-derived genome as a reference. These alignments were then exported as figures and are shown in section 3.3.3.1. The contigs of each genome were aligned using Minimap (version 2.24) (Li, 2018) and plotted as a dot plot (see section 3.3.3.1) using D-GENIES (version 1.4) (Cabanettes and Klopp, 2018). Metabolic and gene information for each genome was determined using DRAM (version 1.2.4) (Shaffer *et al.*, 2020) with databases KEGG (release 96) (Kanehisa and Goto, 2000), PFAM (Mistry *et al.*, 2021), dbCAN (Yin *et al.*, 2012), MEROPS (Rawlings *et al.*, 2018), RefSeq Viral *(Brister et al.*, 2015), VOG DB (*http://vogdb.org/*), and UniRef90 (release 2020_06). The RNAs are detected with barrnap (*https://github.com/tseemann/barrnap*) and

tRNAscan -SE (Lowe and Eddy, 1997) which is shown in section 3.3.3.2.

3.3: Results

3.3.1: Microbial genomes isolated from the rumen

3.3.1.1: Genome assembly comparison

The genomic sequencing was done by the company MicrobesNG, who return both raw sequencing reads and assembled genomes. To ensure the assemblies used were of high quality, the sequencing reads were assembled into contigs and scaffolds, before comparing my assemblies with those assembled by MicrobesNG. The quality of the assemblies was evaluated using QUAST (Gurevich *et al.*, 2013) and the average assembly metrics are shown in Figure 3.4.

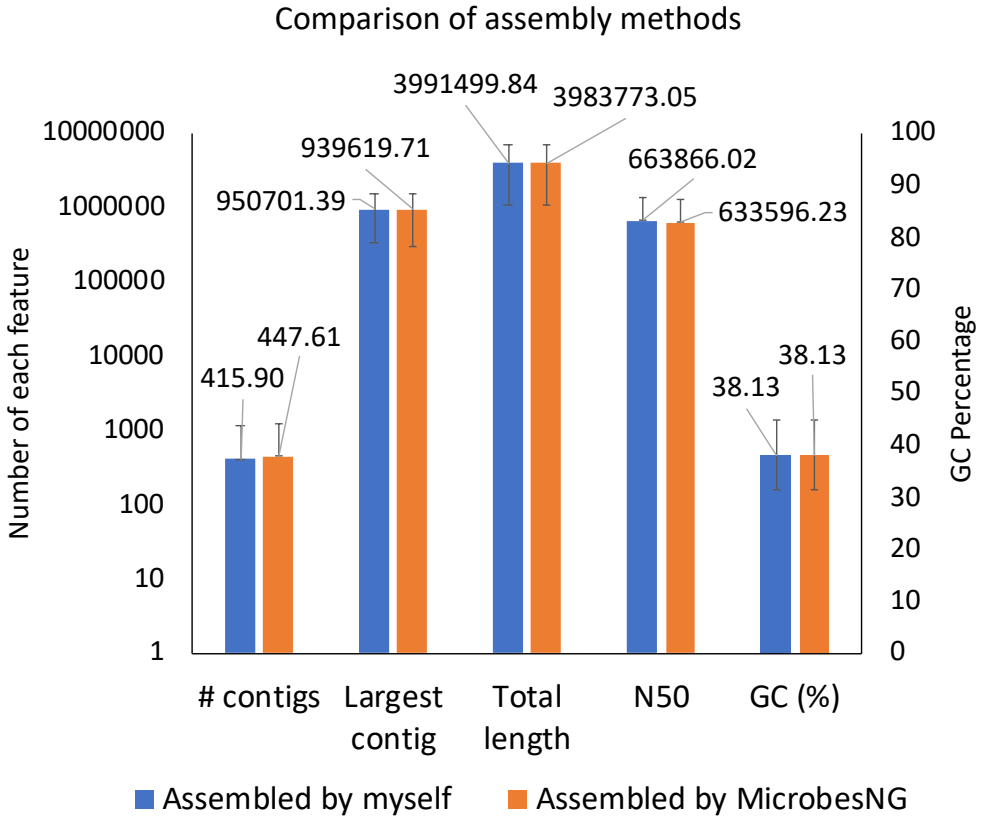


Figure 3.4: Comparing the average assembly metrics for the culture-derived genomes, when assembled by myself and MicrobesNG. All metrics are plotted on the left Y axis except for GC%.

For both assembly methods, the assembly metrics were very similar. However, the average length of the longest contig and the average total length of the assemblies were slightly higher in my assemblies than those assembled by MicrobesNG. As longer contigs indicate a higher quality assembly, this indicates my assembly method produced higher quality assemblies. The N50 metric is often used to assess assembly quality, as it directly correlates to the contiguity of an assembly. The N50 value refers to the length of the shortest contig that covers at least 50% of the assembly (Yandell and Ence, 2012). Therefore, a higher N50 number generally indicates a higher quality assembly with longer contigs compared to an assembly with a lower N50 that might have more contigs that are shorter in length. The N50 values for each genome when assembled by MicrobesNG or myself are shown in Figure 3.5. For most genomes the N50 values were similar for both assembly methods, but some genomes had a better N50 value when assembled by myself. These were 40175wA2_BS4, 40175A4_BSR38, 40175wD1_BS10, 40175wF1_BS7 and 40175wG3_BS35. The average N50 value for the genomes was 620,267 when assembled by MicrobesNG and was 650,987 when assembled by myself. As they were of slightly higher quality, the genomes assemblies that were assembled by myself were used in all further analyses.

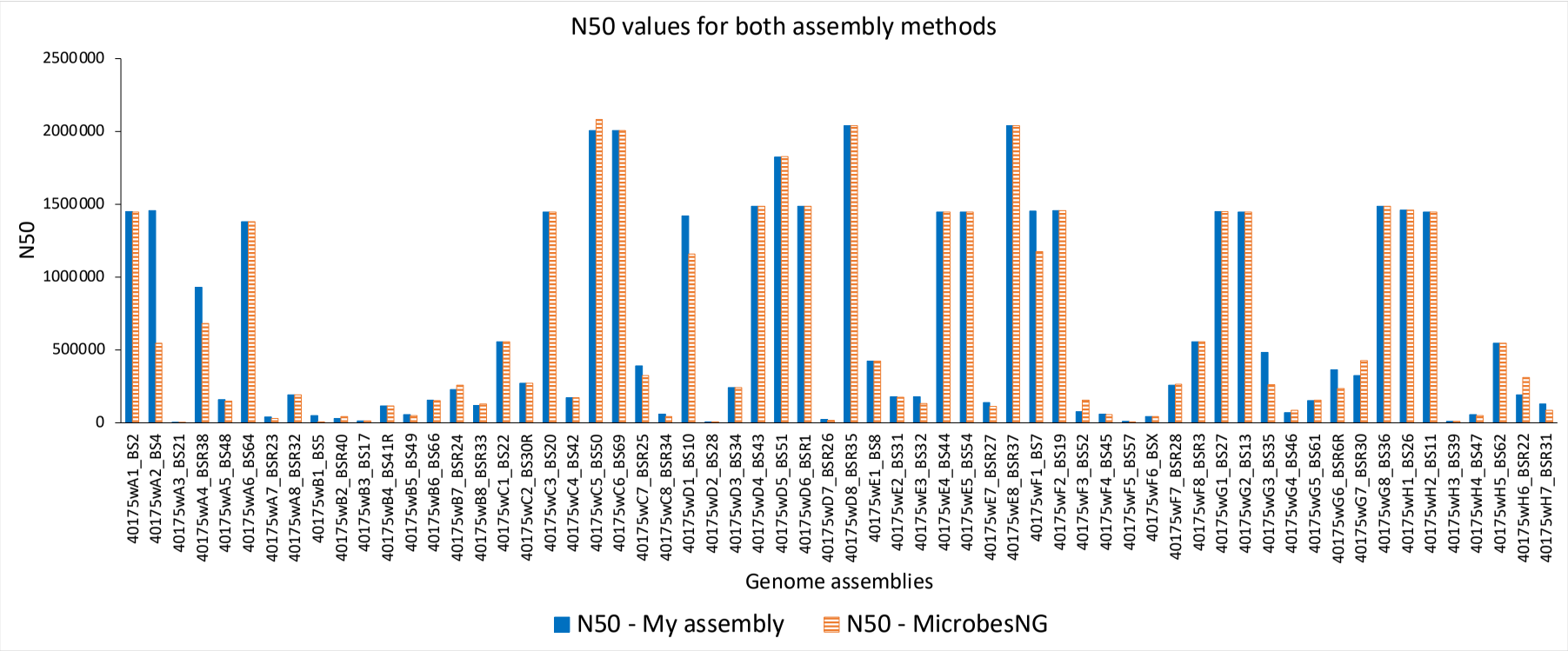


Figure 3.5: N50 values for each culture-derived genome when assembled by myself or MicrobesNG (n=62). MicrobesNG assembled the reads in-house using SPAdes version 3.7, and I assembled the reads using SPAdes version 3.14.1 with the --isolate option. The N50 value is an indicator of the lengths of contigs in the assembly, where a higher N50 indicates a better-quality assembly with longer contigs. In general, the N50 values for each genome were similar for both assembly methods, however there were a few genomes that had a higher N50 when assembled by myself, for example 40175wA2_BS4, 40175A4_BSR38, 40175wD1_BS10, 40175wF1_BS7 and 40175wG3_BS35.

3.3.1.2: Assessing quality and removing contamination

The cultured isolates underwent Gram staining and microscopy. The cellularity and morphology of the bacteria in each sample was reviewed. Some samples appeared to contain a mixture of bacteria, and as a result, the culture-derived genomes were assessed for quality using CheckM. CheckM uses marker genes to assess the completeness and contamination of a genome. For example, if a sample was completely pure and free from contamination, CheckM would deem that sample as having 0% contamination. Whereas if a sample contained two complete genomes, CheckM would deem that sample as having 100% contamination, as it contains an entire other genome. Alongside CheckM, the genomes were assessed with MAGpy, which assigns taxonomy by predicting proteins from contigs and mapping these to a DIAMOND database. For some genomes that were determined as having a high amount of contamination according to CheckM, a high number of proteins corresponded to more than one genus, which suggested the sample contained more than one genome at the time of sequencing.

Sample 40175wB4_BS41R had 100% completeness and 100% contamination, suggesting it contained more than one genome. The MAGpy results showed ~2500 proteins mapping with high percentage identity to *Lachnobacterium bovis* strains DSM 14045 OX=1122142 and OX=140626, and ~2100 proteins mapping with high percentage identity to *Acidaminococcus fermentans* strains ATCC 25085 / DSM 20731 / VR4 OX=591001 and OX=905. The contigs mapping to these distinct species were extracted from the original genome FASTA file, and written to new FASTA files, before being re-assessed with CheckM. A summary of the effect of filtering on the quality of the genomes is shown in Figure 3.6, and Figure 3.7 shows some examples of genomes that were filtered, and the impact of filtering on their completeness, contamination, and strain heterogeneity. The

CheckM results for all of both the unfiltered and filtered culture-derived genomes are shown in Supplementary Table S3.1. For this example, the new filtered genomes were 40175wB4_BS41R_1 and 40175wB4_BS41R_2, and the new CheckM results showed near-complete genomes with minimal contamination.

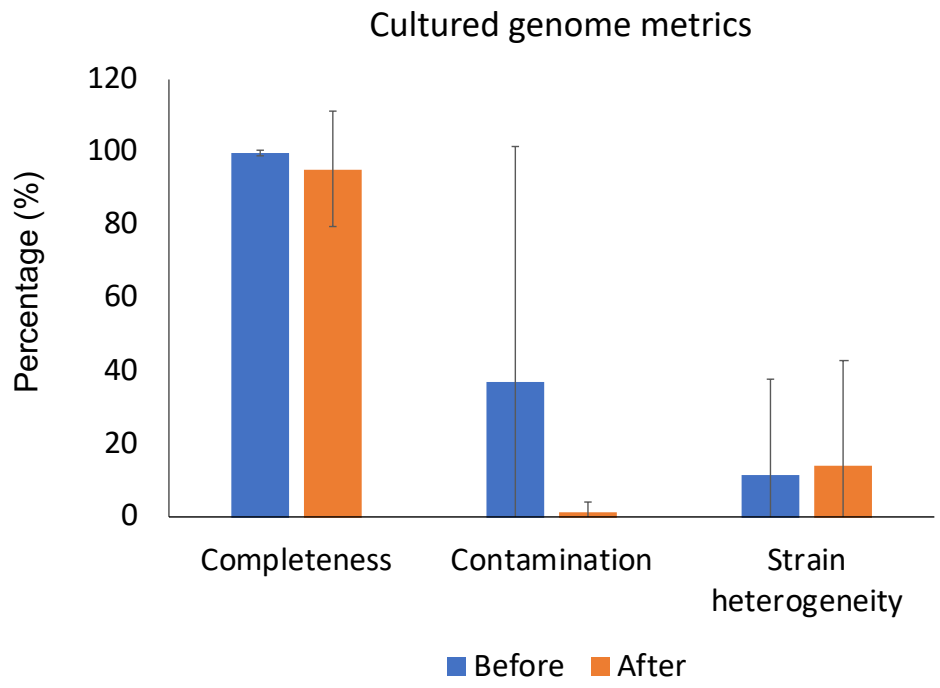


Figure 3.6: A summary of the culture-derived genomes quality determined by CheckM before and after filtering. Shown is the average completeness, contamination and strain heterogeneity, and the error bars are standard deviation.

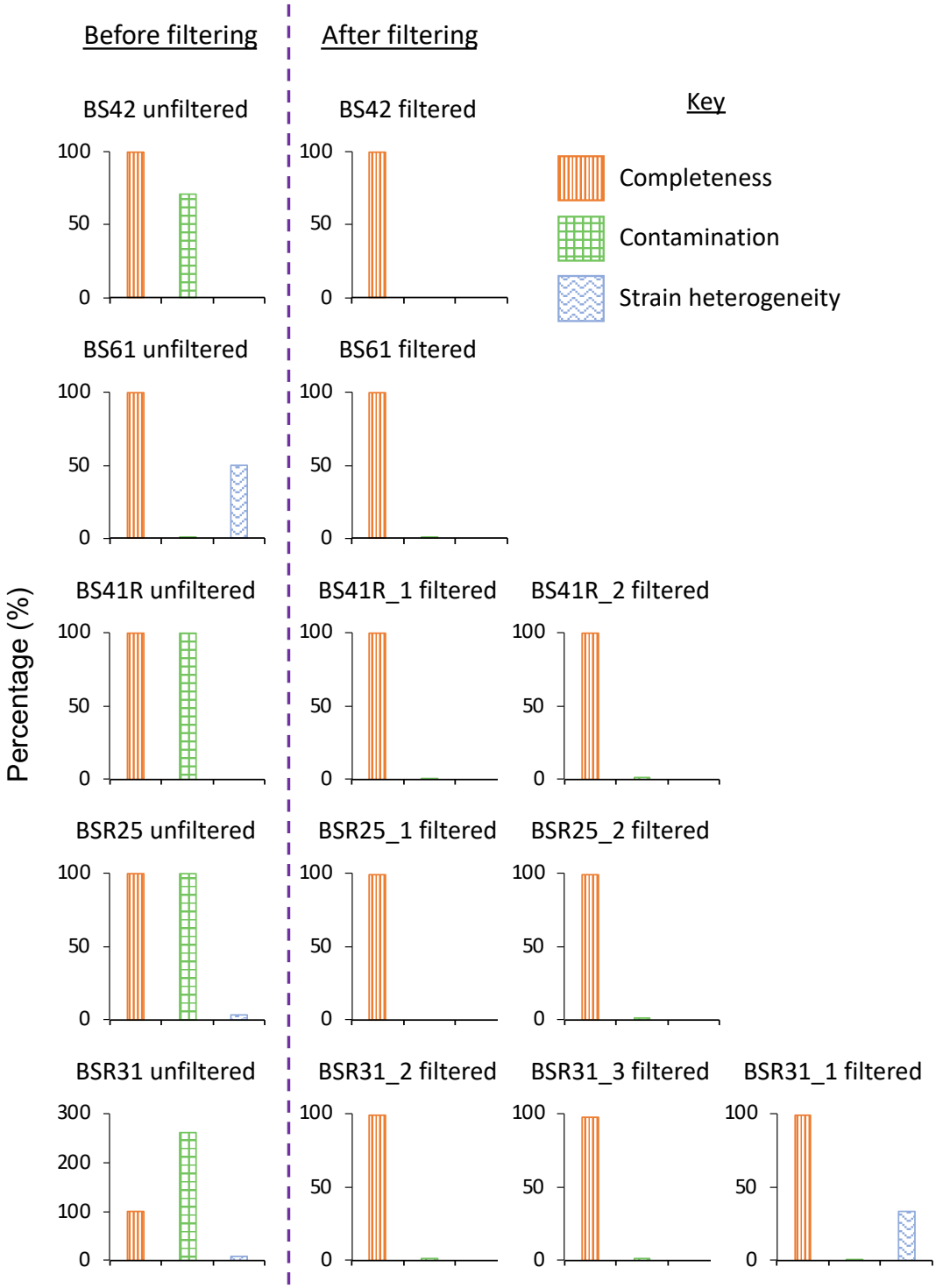


Figure 3.7: Five examples of culture-derived genomes and how filtering impacted the quality of them. Quality was assessed using CheckM, and contigs that looked like they were derived from reagent contamination were removed. For genomes that appeared mixed, the contigs were saved to separate files and re-assessed as individual genomes with CheckM, for example BS41R, which was split into the now-filtered BS41R_1 and BSR41_2. The filtered genomes were used in all further analyses.

For other genomes that had some contamination, but not enough to construct a whole other genome, the DIAMOND results were considered on a case-by-case basis. Almost all genomes had some degree of low-level bacterial contamination. Reagent contamination has been identified in other studies (Walker, 2019). It is also possible that some of these are contamination from the rumen fluid that was used in the culturing media, as while this rumen fluid is filtered and autoclaved it is likely that not all DNA is removed as dead microbial cells will remain within the liquid (Dr Alan Walker, University of Aberdeen, personal communication). This background contamination was characterised by having many contigs with a single or very few proteins mapping to a genera or species, which may have mapped with poor identity. For example, sample 40175wD1_BS10 had 6 contigs that corresponded to ~1700 proteins that mapped with high identity to the species *Streptococcus equinus* OX=1335 and were kept in the filtered version of this genome. However, there were a few contigs that mapped to other species, for example *Anoxybacillus* sp. BCO1 OX=1548750, which is a thermophilic bacterium isolated from the water of a Great Artesian Basin bore well (see https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=taxonDetail&taxon_oid=2642422507). As this is a thermophilic bacterium, it may have been introduced as reagent contamination, possibly able to survive harsh environments that reagents are exposed to in production and use (Salter *et al.*, 2014). For that reason, this particular contig was not included in the filtered genome.

Another example is 40175wB8_BSR33 which had ~2600 proteins mapping with high identity to *Clostridium isatidis* OX=182773 and was deemed as 100% complete by CheckM. This genome had some contigs that had proteins mapping to reagent contamination, such as *Bradyrhizobium betae* OX=244734, which is a likely nitrogen-fixing reagent contaminant *(Salter et al.*, 2014). It has been suggested that nitrogen-fixing laboratory contaminants originate in the production of ultra-pure water, as nitrogen is used to prevent

CO_2 and O_2 from dissolving in the water (Kulakov *et al.*, 2002). Other contaminants include “bioreactor metagenome OX=1076179”, which is an example of how informal labels in lieu of formal taxonomy can still be informative. Removing these contigs did not change the amount of contamination this genome had as deemed by CheckM, which remained at 1.08% after filtering. This may be because this genome still had contamination that was a close relative, or, as this particular genome had not been cultivated previously, it may not have been recognised as one complete genome by CheckM yet. Sample 40175wB6_BS66 had ~2400 proteins mapping to *Lachnobacterium boxis* OX=140626, but also several proteins that mapped to other rumen bacteria. These included *Pseudobutyrivibrio xylanivorans* strain DSM 14809 OX=1123012 (Grilli *et al.*, 2013), *Agathobacter ruminis* strain OX=1712665 (Rosero *et al.*, 2016), and *Pseudobutyrivibrio ruminis* strain OX=46206 (Grilli *et al.*, 2013).

3.3.2: Determining the taxonomy of the culture-derived genomes

The taxonomy of each cultured isolate genome assembly was determined using GTDB-Tk. The frequencies of each taxonomy at the phylum, family and species level are shown in Figure 3.8. Of the 77 cultured isolate genomes, all were assigned to the kingdom Bacteria, and 76 were assigned the phylum *Firmicutes*. The other genome was assigned the phylum *Actinobacteriota*. The isolates spanned 10 families, 12 genera and 19 species.

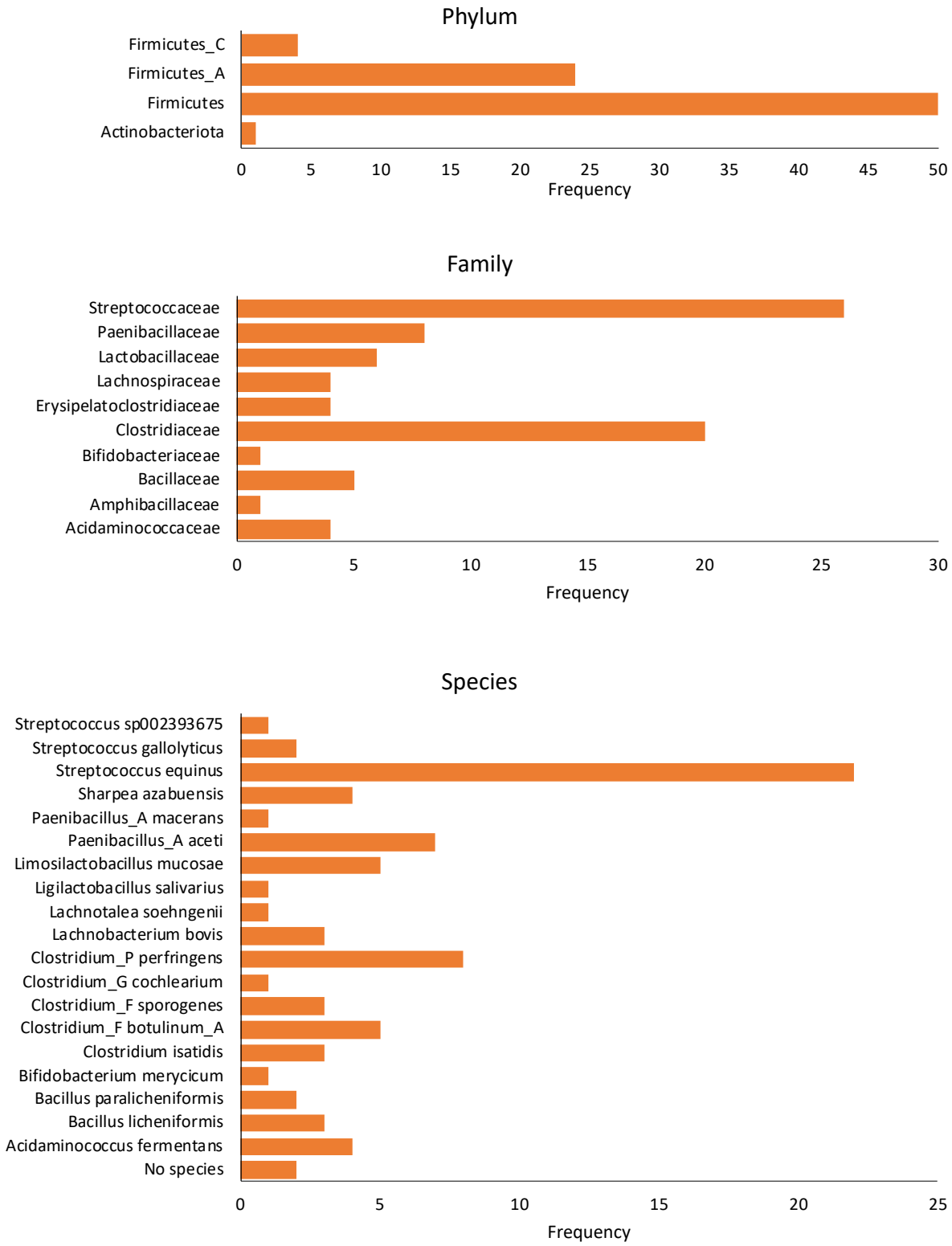


Figure 3.8: Taxonomy of the culture-derived isolate genomes. The taxonomy of each culture-derived genome was assigned using GTDB-Tk. The frequency of each phylum, family and species is shown.

The Hungate collection is an extensive collection of rumen-isolated microbial genomes. The cultured isolates were clustered with culture-derived genomes from the Hungate collection using dRep at 95% ANI (Average nucleotide identity). Of the 77 isolates, 29 did not cluster with a Hungate genome. As 95% ANI is considered a cut-off for defining a species, any isolated genomes that did not cluster with a Hungate genome may be a species present in the rumen but that was not isolated during the Hungate1000 project. Although these genomes appeared to be absent from the Hungate genomes, 27 of the 29 were assigned a species by GTDB-Tk, which suggests they are not novel species. However, two isolates were not assigned a species by GTDB-Tk, which suggests they may be completely novel and not previously cultivated. These are BSR6R ("40175wG6_BSR6R_filtered"), which was assigned the lineage d__Bacteria > p__Firmicutes > c__Bacilli > o__Bacillales_D > f__Amphibacillaceae > g__Virgibacillus and BS49_2 ("40175wB5_BS49_2_filtered") which was assigned the lineage d__Bacteria > p__Firmicutes > c__Bacilli > o__Lactobacillales > f__Streptococcaceae > g__Streptococcus. The phylum, family, genus, and species level taxonomy for the culture-derived genomes are available in Supplementary Table S3.3.

3.4: Metagenome-assembled-genomes (MAGs) created from rumen metagenomic data

3.4.1: Assessing the quality of rumen metagenome-derived bins

From the two metagenomic samples, a total of 206 genome bins were assembled, and the quality of the bins was assessed using CheckM. The acceptable cut-off for the quality of a bin can vary depending on the study, but a medium-quality draft according to the Genome Standard is one that has at least 50% completeness and less than 10% contamination (Bowers *et al.*, 2017). In this study drafts with at least 80% completeness and less than 10% contamination were selected. Of the 206 genome bins assembled, 69 met this criterion. The average assembly metrics for the 69 medium-quality bins are shown in Figure 3.9. The full metrics for each bin are shown in Supplementary Table S3.4. In addition, the genome bins were checked for contamination and quality by running the MAGpy pipeline, as was done for the cultured isolate genomes. For all 69 medium-quality bins, a high number of the predicted proteins mapped with high percentage identity to the same genus, supporting the CheckM findings of low contamination.

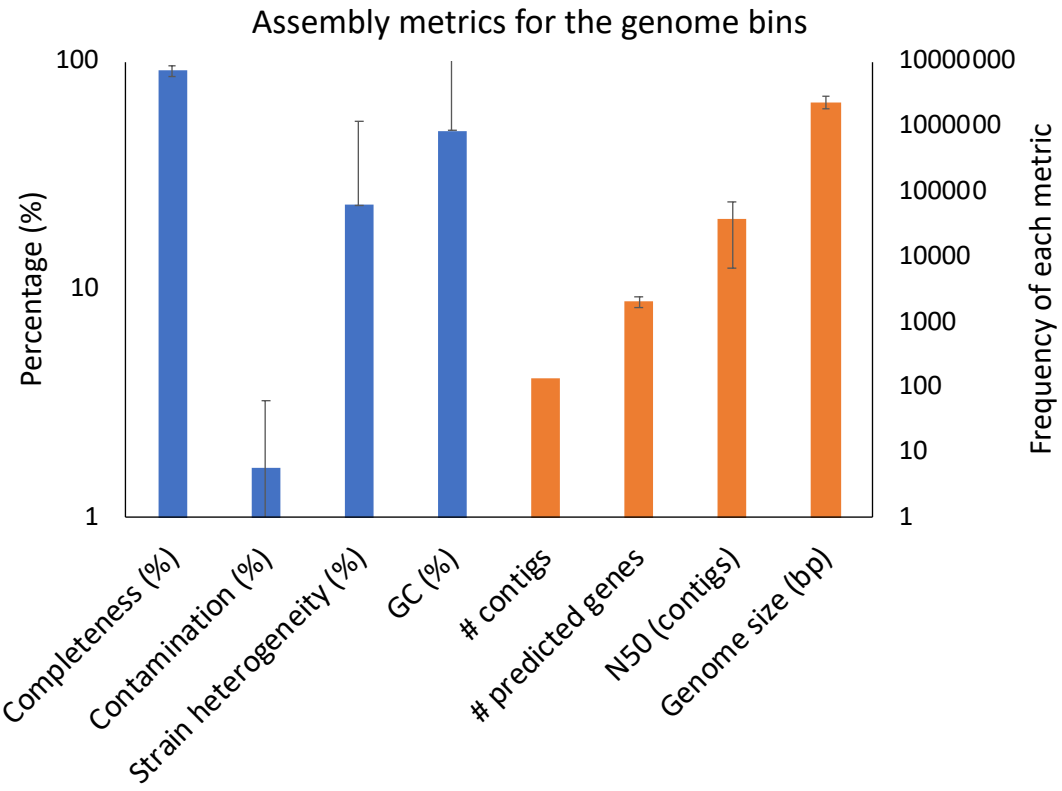


Figure 3.9: Summary of assembly metrics for the genome bins. Assembly metrics shown are the average for the 69 medium-quality bins that had >80% completeness and <1% contamination. For each bar the y-axis is the frequency of that metric.

3.4.2: Taxonomy and phylogeny of rumen genome bins

GTDB-Tk was used to determine the taxonomy of each MAG, and the taxonomy at the phylum and family levels are shown in Figure 3.10. The full taxonomic lineage is shown in Supplementary Table S3.5. The 69 medium-quality (>80% complete and <10% contaminated) MAGs were all Bacteria except for “metabat_S1_2013.144.fa”, which was Archaea. The MAGs spanned 8 phyla, with 29 belonging to *Bacteroidota* and 31 belonging to *Firmicutes*. Three MAGs did not have taxonomy assigned at the species level, which suggests they are novel to the GTDB database. These were “metabat_S2_2017.17.fa”, “metabat_S1_2013.16.fa” and

“metabat_S2_2017.10.fa”, which were classified at the genus level as *Eubacterium*, *Saccharofermentans* and *Prevotella* respectively.

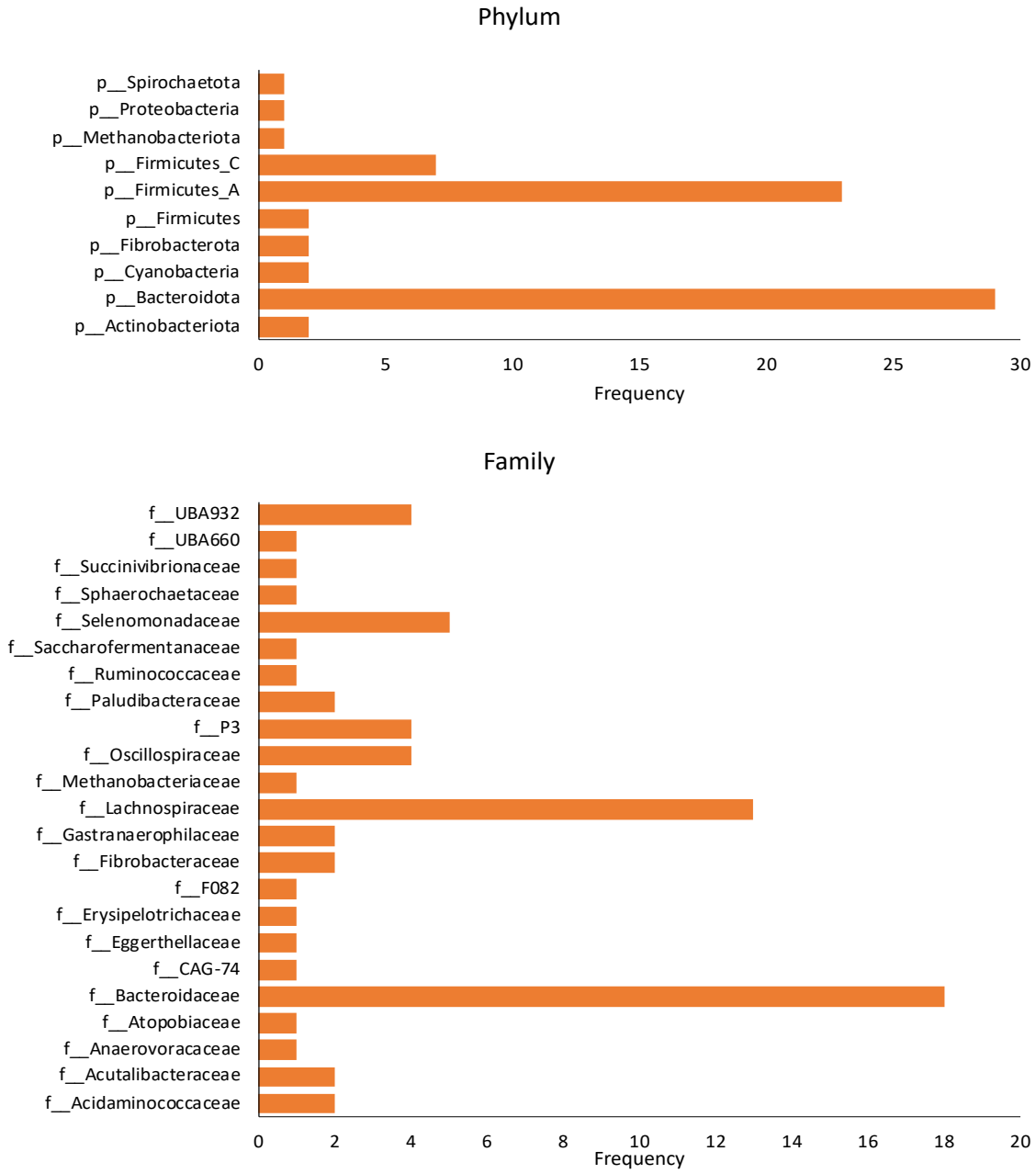


Figure 3.10: Taxonomy of genome bins as determined by GTDB-Tk, with the frequency shown at the phylum and family levels.

In order to see whether there were any genomes in the MAGs or isolates that belonged to taxa that were not present in the Hungate1000 genome collection or the previously released rumen MAG collection referred to as the "RUG superset" that was published in 2019 (Stewart *et al.*, 2019), the MAGs, culture-derived genomes, the Hungate collection and the RUG superset were dereplicated using dRep at 95% ANI. As 95% ANI is an accepted cut-off for distinguishing between species (Jain *et al.*, 2018), any genomes that did not cluster may be a species that is not in the Hungate genome collection and/or RUG superset. Two of the MAGs, "metabat_S2_2017.17.fa" and "metabat_S2_2017.10.fa", were assembled from rumen sample 2 which was frozen in 2017. These did not cluster with any other genomes. The MAG "metabat_S1_2013.16.fa" clustered with assembled genomes in the RUG superset, none of which had an assigned species by GTDB-Tk. This is interesting as it suggests a species that has been observed previously but has not yet been assigned a taxonomy in GTDB.

3.5: Clustering culture-derived and metagenome-assembled rumen genomes

3.5.1: Strain level clustering results

In order to identify genomes that could be considered members of the same strain, MAGs and culture-derived genomes were dereplicated and clustered at 99% ANI (Konstantinidis and Tiedje, 2005). Unexpectedly, the culture-derived genomes that had been cultured from the two rumen samples (see Section 3.2.1), and the MAGs that had been assembled from the two rumen samples (see Section 3.2.2), did not cluster at the species (95% ANI) or strain level (99% ANI). Consequently, the culture-derived genomes were then compared with the MAGs in the RUG superset (described in Section 3.4.2 and in (Stewart *et al.*, 2019)). This produced 7 clusters that contained a pair of genomes, one MAG and one culture-derived, of the same strain. For clusters containing more than one genome, a representative culture-derived

or metagenome-assembled genome was chosen such that one pair of genomes was selected from each cluster for comparison. The metrics of each cluster pair are shown in Table 3.2.

Table 3.2: Genome clusters containing a MAG and culture-derived genome of the same strain. Each genome pair are clustered at 99% ANI, and the genome metrics are shown.

Cluster ID	Bin Id	Completeness	Contamination	Strain heterogeneity	Genome size (bp)	# contigs	N50 (contigs)	GC%	# predicted genes
3R	40175wA6_BS64_filtered.fa	100	0	0	1930717	8	1378230	36.8	1880
	RUG12079.fa	96.07	0.75	0	1646589	107	26456	37.5	1665
3S	40175wC7_BSR25_1_filtered.fa	98.86	0	0	4244894	21	494543	45.8	4359
	RUG14306.fa	96.95	1.21	7.14	4019681	239	31316	46.2	4190
4G	40175wE3_BS32_filtered.fa	99.18	0	0	1958667	35	179478	46.7	1885
	RUG14882.fa	90.24	1.09	100	1539703	193	10272	47.2	1542
5B	40175wG3_BS35_filtered.fa	99.77	0.95	0	2203519	24	481331	60.2	1801
	RUG13721.fa	99.77	0.86	0	2221233	25	156122	60.5	1810
5J	40175wB4_BS41R_2_filtered.fa	99.98	1.2	0	2287647	118	75765	56.1	2100
	RUG13906.fa	98.58	0.6	0	2058902	93	37232	57.2	1911
5W	40175wB4_BS41R_1_filtered.fa	99.52	0.16	0	2711068	85	155494	31.4	2426
	RUG11184.fa	97.1	0.81	0	2429297	204	24589	31	2225
6F	40175wG4_BS46_filtered.fa	99.06	0.94	0	2355253	93	70474	37.4	2346
	RUG10119.fa	99.06	0.94	0	2292713	142	25743	37.1	2281

The taxonomy of each clustered pair was determined using GTDB and is shown in Table 3.3. For each MAG and isolate pair, the taxonomy was the same at all taxonomic levels, with the exception of cluster 3R. The genomes in cluster 3R are both members of the *Streptococcus* genus, however they were assigned different species.

Table 3.3: The taxonomy of each paired culture-derived and metagenome-assembled cluster. The assigned taxonomy at the phylum, family, genus, and species levels for each pair of metagenome-assembled and culture-derived genomes.

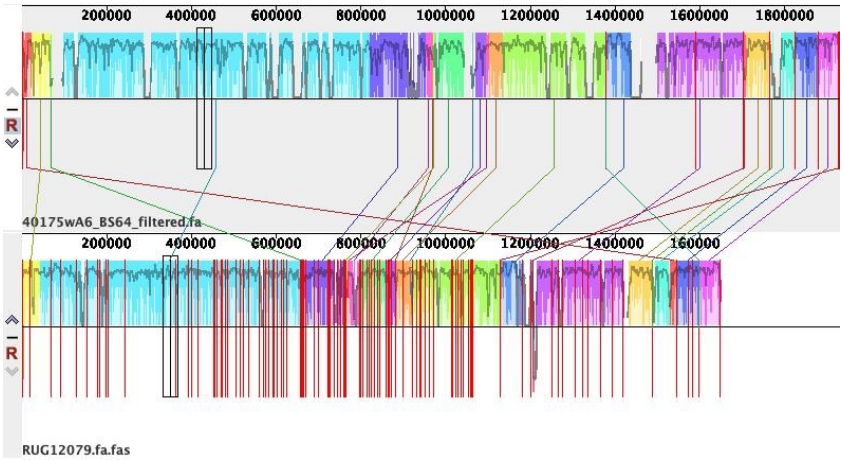
Cluster	Genome	Phylum	Family	Genus	Species
3R	RUG12079	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
	40175wA6_BS64	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus sp002393675
3S	RUG14306	p__Firmicutes	f__Bacillaceae	g__Bacillus	s__Bacillus licheniformis
	40175wC7_BSR25_1	p__Firmicutes	f__Bacillaceae	g__Bacillus	s__Bacillus licheniformis
4G	RUG14882	p__Firmicutes	f__Lactobacillaceae	g__Limosilactobacillus	s__Limosilactobacillus mucosae
	40175wE3_BS32	p__Firmicutes	f__Lactobacillaceae	g__Limosilactobacillus	s__Limosilactobacillus mucosae
5B	RUG13721	p__Actinobacteriota	f__Bifidobacteriaceae	g__Bifidobacterium	s__Bifidobacterium merycicum
	40175wG3_BS35	p__Actinobacteriota	f__Bifidobacteriaceae	g__Bifidobacterium	s__Bifidobacterium merycicum
5J	RUG13906	p__Firmicutes_C	f__Acidaminococcaceae	g__Acidaminococcus	s__Acidaminococcus fermentans
	40175wB4_BS41R_2	p__Firmicutes_C	f__Acidaminococcaceae	g__Acidaminococcus	s__Acidaminococcus fermentans
5W	RUG11184	p__Firmicutes_A	f__Lachnospiraceae	g__Lachnobacterium	s__Lachnobacterium bovis
	40175wB4_BS41R_1	p__Firmicutes_A	f__Lachnospiraceae	g__Lachnobacterium	s__Lachnobacterium bovis
6F	RUG10119	p__Firmicutes	f__Erysipelatoclostridiaceae	g__Sharpea	s__Sharpea azabuensis
	40175wG4_BS46	p__Firmicutes	f__Erysipelatoclostridiaceae	g__Sharpea	s__Sharpea azabuensis

Alignments of each genome pair, with the culture-derived genome as the reference and the MAG as the query, are shown in Figures 3.11-3.17. Overall, the culture-derived genomes and their MAG counterparts aligned well, showing a high level of sequence similarity.

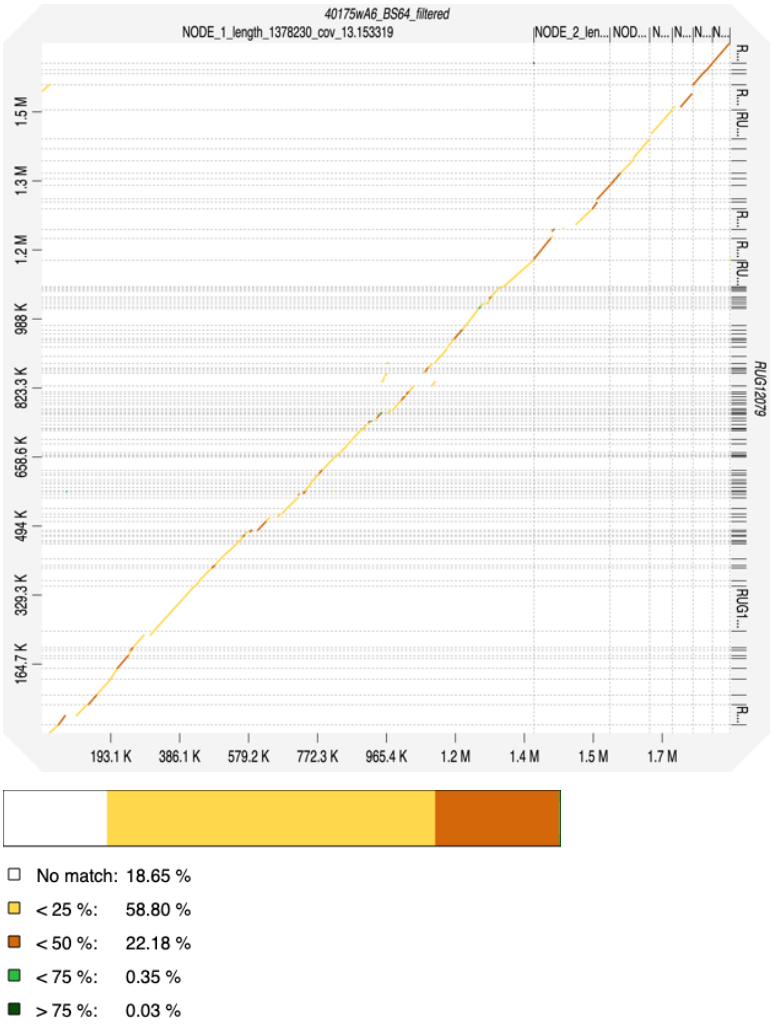
Figure 3.11 shows the alignment of Cluster 3R, which contained the cultured genome “40175wA6_BS64” and the MAG “RUG12079”. The assemblies of these genomes aligned well, and the dot plot was contiguous, which suggests that the genome sequences are similar to one-another. However, the percentage identity was the lowest of all the pair comparisons, with the majority of contigs aligning at <25% identity. This is unexpected as the alignments suggest that the genome sequences are similar, but a low identity suggests variation in the sequences. To investigate whether this was a result of the MAG assembly, the MAG “RUG12079” was aligned with another culture-derived genome “40175wG1_BS27” that that clustered with > 95% ANI (see Supplementary Figure S3.1). The alignment with “40175wG1_BS27” was contiguous, but the contigs aligned with a much higher identity that the alignment with “40175wA6_BS64”. Both the culture-derived genomes and the MAG clustered at > 95% ANI, which suggests they are members of the same species. This is also supported by the contiguous alignments. However, as the majority of contigs in the assemblies for “40175wA6_BS64” and “RUG12079” aligned with poor identity, but those for “40175wG1_BS27” and “RUG12079” aligned with high identity, there must be differences in sequence in the culture-derived genome “40175wA6_BS64”. It may be that although the orthologous protein coding genes share > 95% ANI, there are differences in the non-coding regions that cause the contigs to have poor identity. The taxonomy of these culture-derived species was assigned using GTDB, which stated that “40175wG1_BS27” belongs to the species *Streptococcus equinis*, whereas “40175wA6_BS64” was assigned a species label of *Streptococcus* “sp002393675”. Although it may be that “40175wA6_BS64” has a poor quality assembly, it is also possible that this

culture-derived genome is a relative of *Streptococcus equinis*, and has acquired some of these protein coding genes from horizontal gene transfer.

A



B

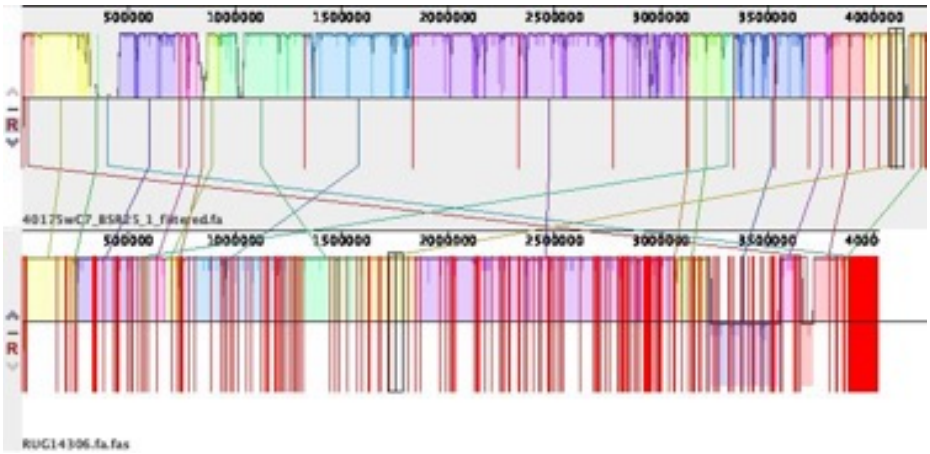


No match: 18.65 %
< 25 %: 58.80 %
< 50 %: 22.18 %
< 75 %: 0.35 %
> 75 %: 0.03 %

Figure 3.11: Cluster 3R was the culture-derived genome "40175wA6_BS64" and the MAG "RUG12079". The two genomes are shown as (A) an alignment plot and (B) a dot plot. The colours in the alignment plot (A) denote regions of the genome. For the dot plot (B), the colours of the contigs denote the percentage of similarity between the contigs of the two genomes.

The comparison of assemblies in cluster 3S, which contained the culture-derived genome “40175wC7_BSR25_1” and the MAG “RUG14306”, are shown in Figure 3.12. The dot plot shows the two assemblies are very contiguous and had a high identity. The mauve alignment shows regions of the genome that appear to be present in both genome assemblies. Together, these results suggest that these genomes are very similar in sequence to each other.

A



B

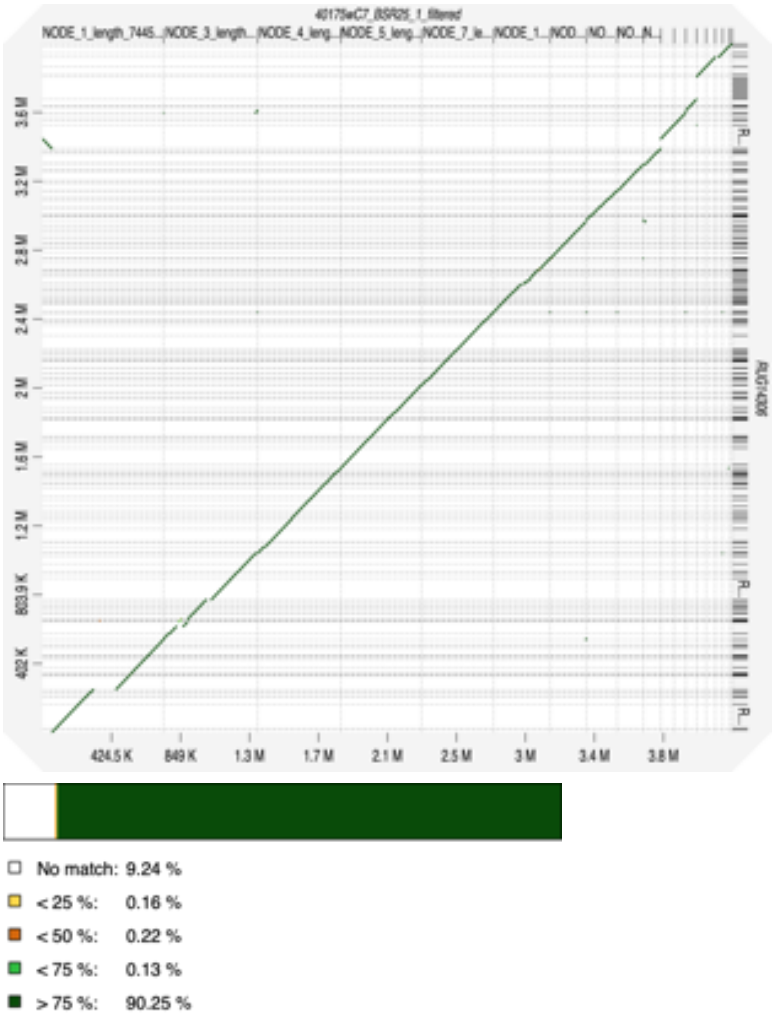
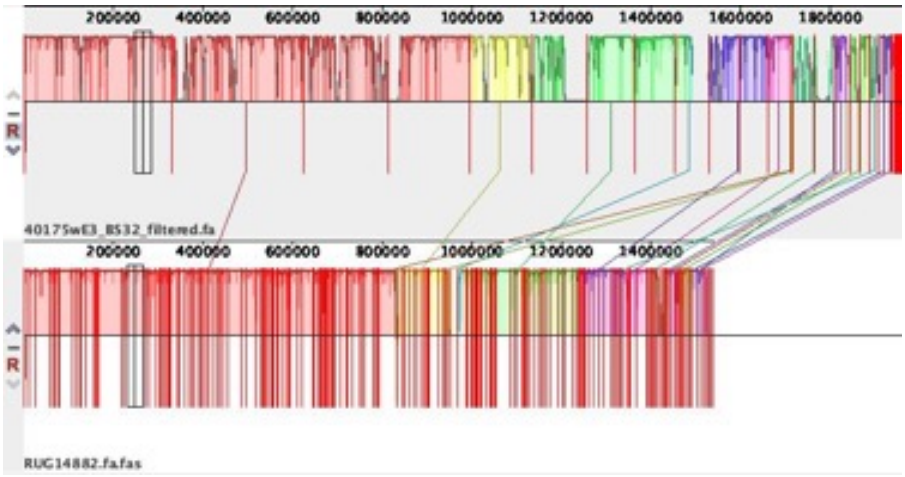


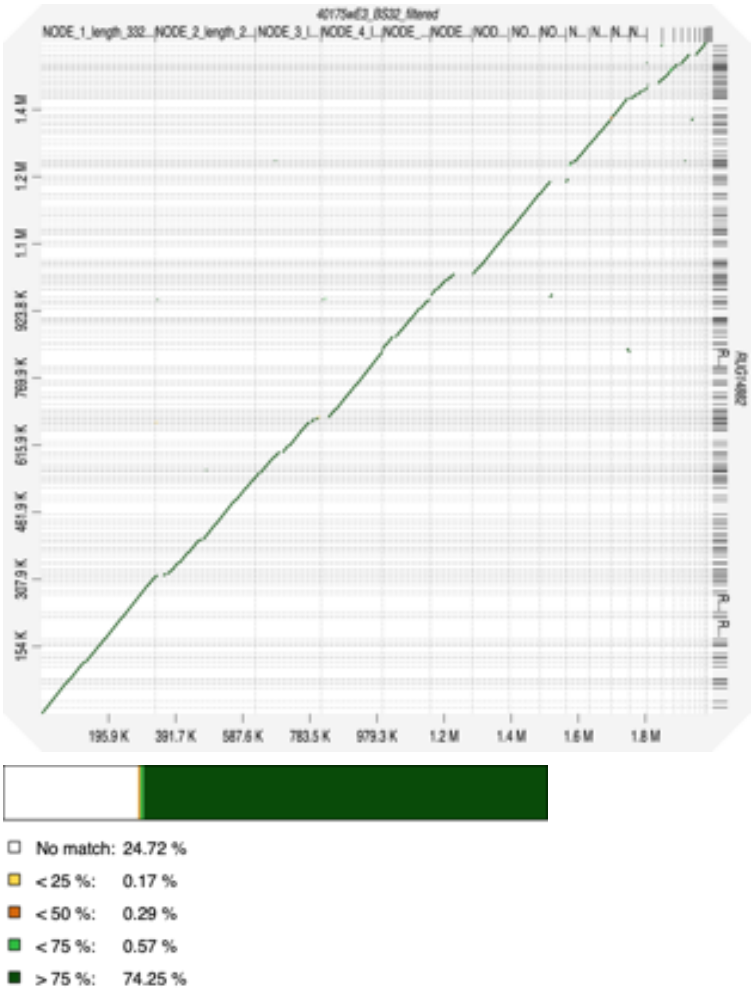
Figure 3.12: Cluster 3S was the culture-derived genome "40175wC7_BSR25_1" and the MAG "RUG14306". The two genomes are shown as (A) an alignment plot and (B) a dot plot. The colours in the alignment plot (A) denote regions of the genome. For the dot plot (B), the colours of the contigs denote the percentage of similarity between the contigs of the two genomes.

Cluster 4G, which contained the culture-derived genome “40175wE3_BS32” and the MAG “RUG14882”, had almost 25% of contigs from each assembly that did not align. This was evident in the alignment plot, shown in Figure 3.13, as there were many segments of the culture-derived genome that were absent in the assembled genome. The dot plot however shows that the contigs that did match between the two genomes did so with high percentage identity. These results suggest that the MAG and culture-derived genome in cluster 4G were not as similar as some other clusters, for example cluster 3S. There appeared to be a few gaps in the culture-derived genome assembly compared to the MAG. This could be a result of erroneous MAG binning, or it is possible the culture-derived genome was not complete. The culture-derived genome “40175wE3_BS32” was deemed as 99.18% complete by CheckM, therefore it may be more likely that the MAG contained contigs that belong to another genome.

A



B

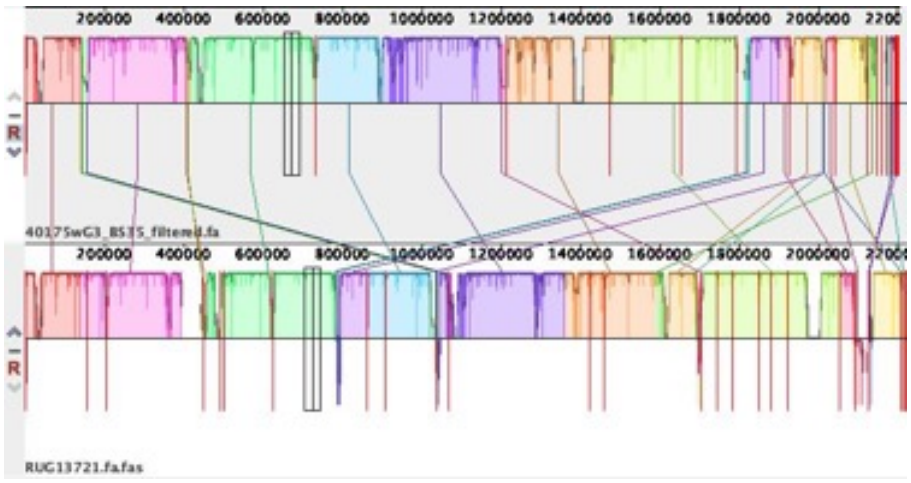


- No match: 24.72 %
- < 25 %: 0.17 %
- < 50 %: 0.29 %
- < 75 %: 0.57 %
- > 75 %: 74.25 %

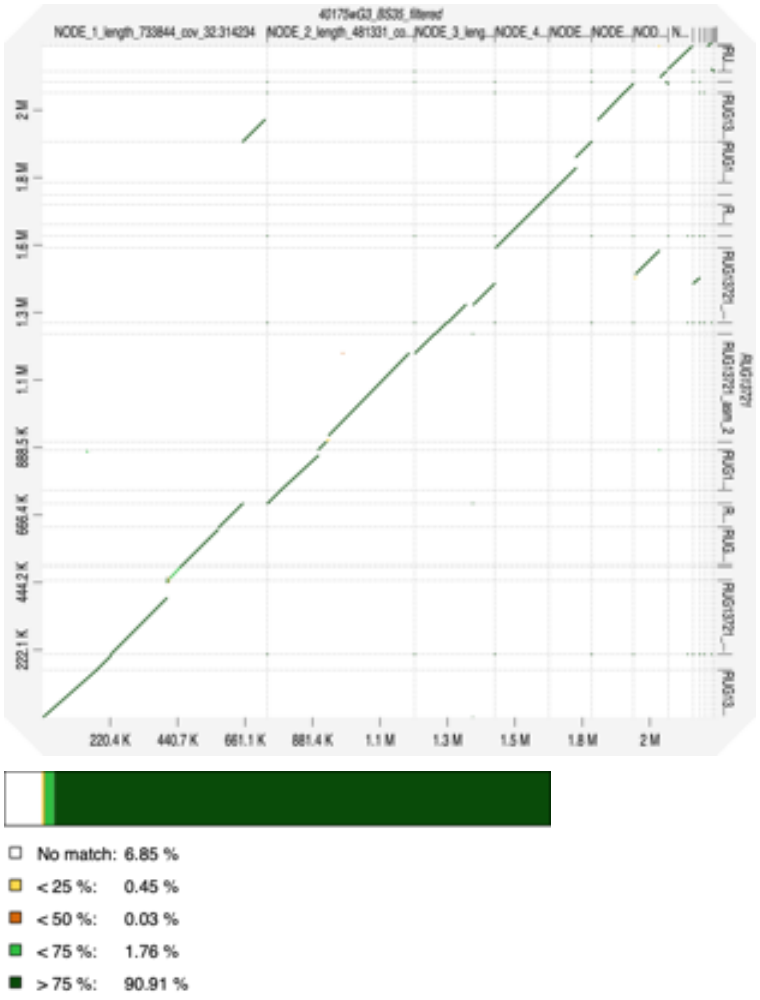
Figure 3.13: Cluster 4G was the culture-derived genome "40175wE3_BS32" and the MAG "RUG14882". The two genomes are shown as (A) an alignment plot and (B) a dot plot. The colours in the alignment plot (A) denote regions of the genome. For the dot plot (B), the colours of the contigs denote the percentage of similarity between the contigs of the two genomes.

The alignments of cluster 5B, which contained the culture-derived genome “40175wG3_BS35” and the MAG “RUG13721”, also aligned well and with high identity, as shown in Figure 3.14. The alignment plot shows the two genomes had many regions in common that were of a similar size. The dot plot comparing the genomes in cluster 5B showed some contigs that did not align well across the assemblies, which may be mobile elements or issues with assembly. Overall, the culture-derived genome and MAG in cluster 5B seemed to be very similar in sequence. Clusters 3S and 5B had genomes pairs that were the most similar to each other out of all of the clusters; with less than 10% of reads not aligning across the assemblies.

A



B

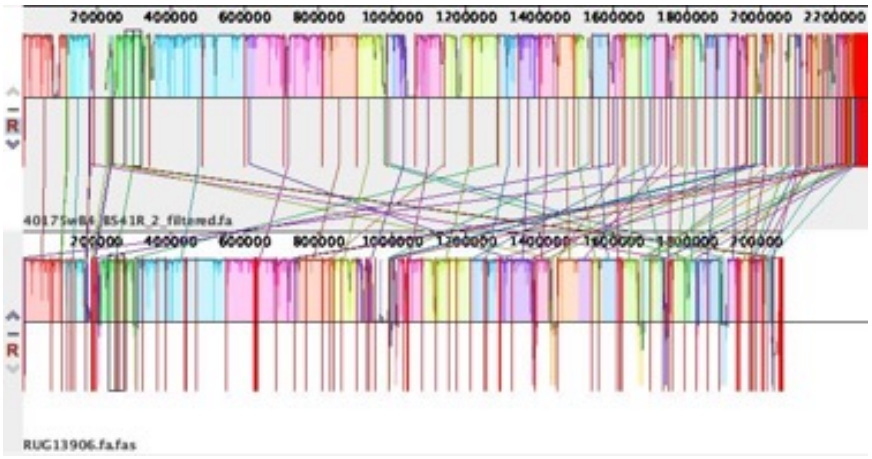


No match: 6.85 %
< 25 %: 0.45 %
< 50 %: 0.03 %
< 75 %: 1.76 %
> 75 %: 90.91 %

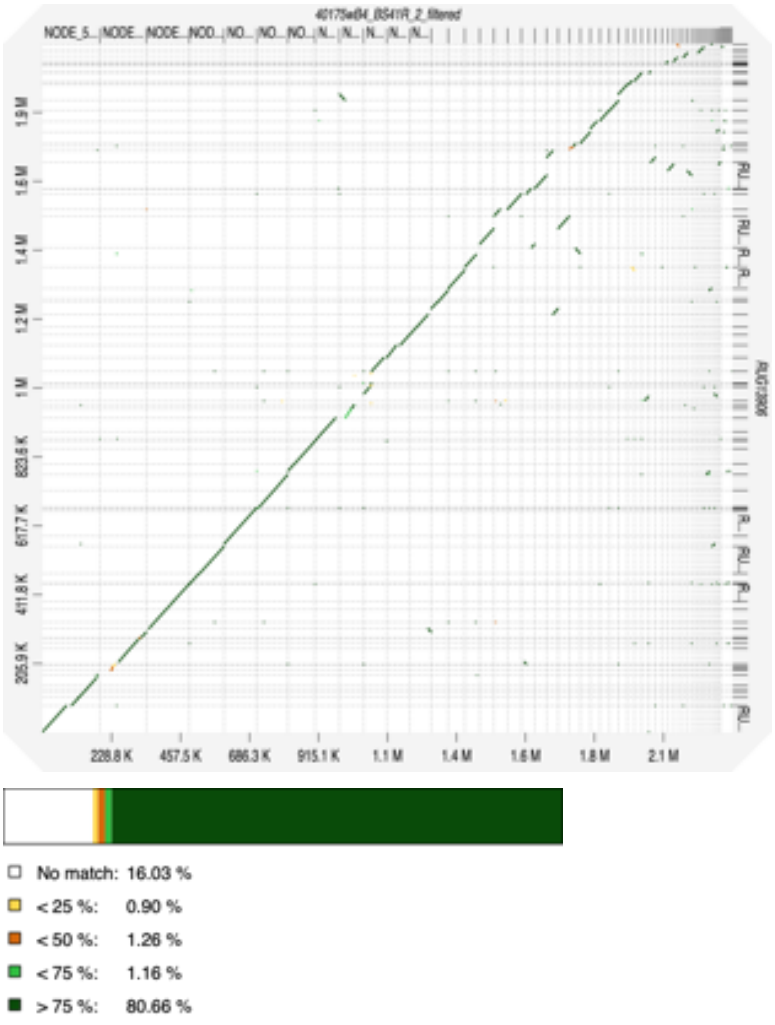
Figure 3.14: Cluster 5B was the culture-derived genome "40175wG3_BS35" and the MAG "RUG13721". The two genomes are shown as (A) an alignment plot and (B) a dot plot. The colours in the alignment plot (A) denote regions of the genome. For the dot plot (B), the colours of the contigs denote the percentage of similarity between the contigs of the two genomes.

Cluster 5J contained the culture-derived genome “40175wB4_BS41R_2” and the MAG “RUG13906”, and in the mauve alignment plot there were many segments of the genome assemblies that were in different locations across the two genomes, see Figure 3.15. In the dot plot, most contigs aligned with high identity but over 16% of the genomes did not align. These results suggest that the two genomes were not as similar to each other as some other clusters, and that there is more variation in the sequences.

A



B

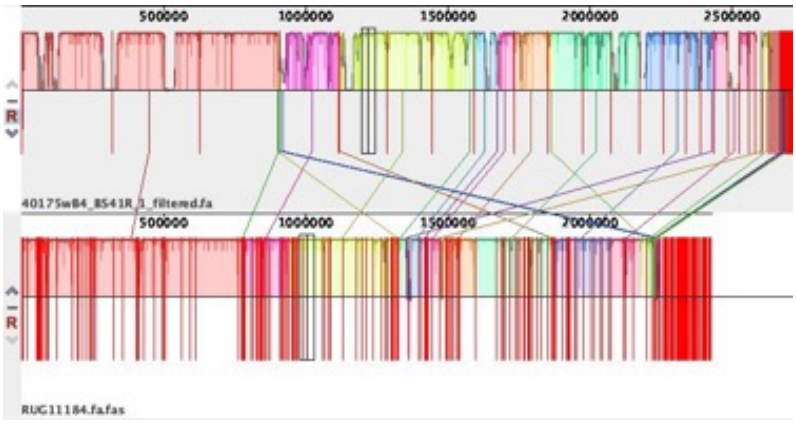


No match: 16.03 %
< 25 %: 0.90 %
< 50 %: 1.26 %
< 75 %: 1.16 %
> 75 %: 80.66 %

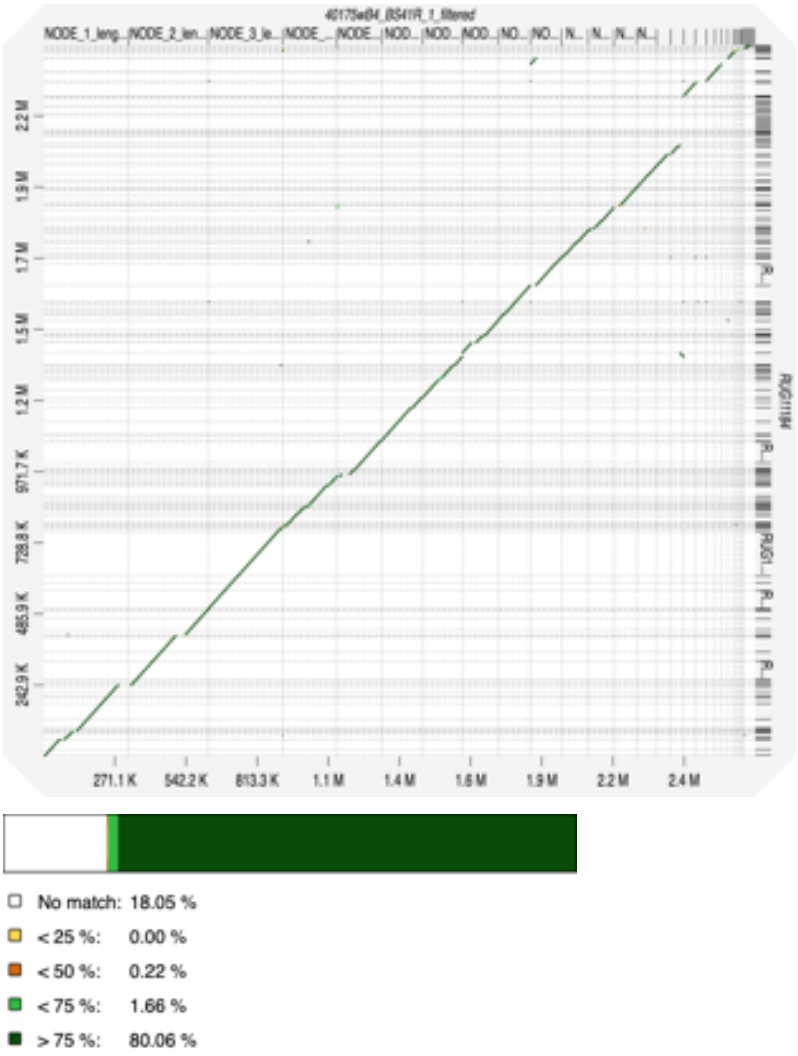
Figure 3.15: Cluster 5J was the culture-derived genome "40175wB4_BS41R_2" and the MAG "RUG13906". The two genomes are shown as (A) an alignment plot and (B) a dot plot. The colours in the alignment plot (A) denote regions of the genome. For the dot plot (B), the colours of the contigs denote the percentage of similarity between the contigs of the two genomes.

The alignments for cluster 5W are shown in Figure 3.16, and while the assemblies generally aligned with high identity, there was a gap in the dot plot. This suggested that the culture-derived genome “40175wB4_BS41R_1” contained a region that was not present in the MAG “RUG11184”. This could be due to misassembly, however there were no other contigs of a suitable length in the assembly that could be the missing contig. It is more likely, therefore, that this segment of genome was present in the culture-derived genome but absent in the RUG assembly. Segments belonging to regions of the genome that are highly repetitive, or present in multiple species in the metagenome sample, are more likely to be lost during binning and assembly.

A



B

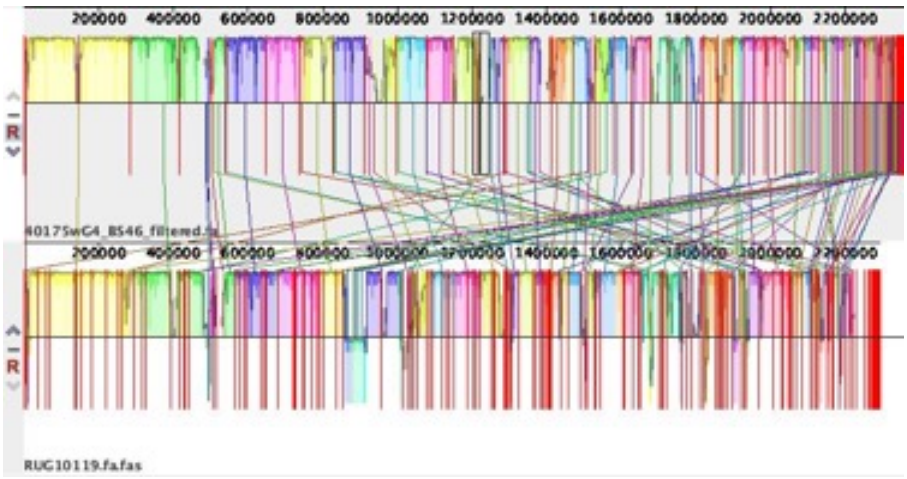


No match: 18.05 %
< 25 %: 0.00 %
< 50 %: 0.22 %
< 75 %: 1.66 %
> 75 %: 80.06 %

Figure 3.16: Cluster 5W was the culture-derived genome "40175wB4_BS41R_1" and the MAG "RUG11184". The two genomes are shown as (A) an alignment plot and (B) a dot plot. The colours in the alignment plot (A) denote regions of the genome. For the dot plot (B), the colours of the contigs denote the percentage of similarity between the contigs of the two genomes.

The alignment of cluster 6F, which contained the culture-derived genome “40175wG4_BS46” and the MAG “RUG10119”, shown in Figure 3.17, appeared to be more fragmented than some of the other alignments. This was possibly due to the N50 of “40175wG4_BS46”, which was the lowest of all the culture-derived genome assemblies. A shorter N50 indicates the assembly contained shorter length contigs, which may cause the assemblies to be less contiguous. While cluster 6F generally aligned with high identity, 8.45% of the alignment matched at 50-75% identity. This suggests that although the genomes are similar, there was some variation between the two genome sequences.

A



B

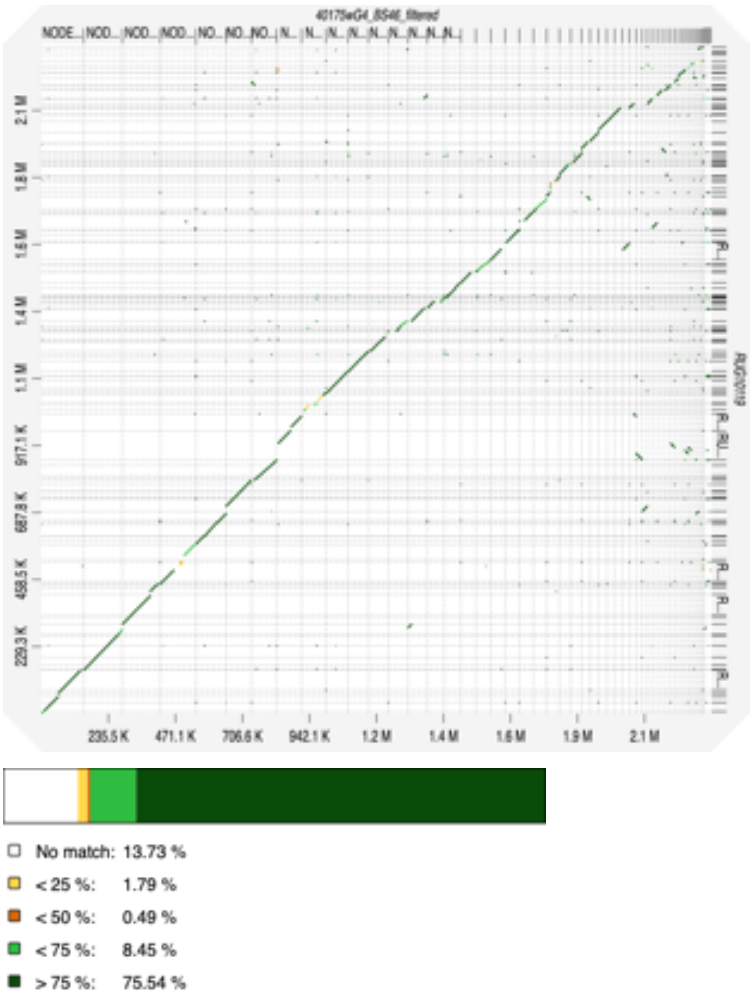


Figure 3.17: Cluster 6F was the culture-derived genome "40175wG4_BS46" and the MAG "RUG10119". The two genomes are shown as (A) an alignment plot and (B) a dot plot. The colours in the alignment plot (A) denote regions of the genome. For the dot plot (B), the colours of the contigs denote the percentage of similarity between the contigs of the two genomes.

3.5.2: Comparing functional information from culture-derived genomes and MAGs of the same strain

Previously in this work, the genome assembly metrics and taxonomy information of rumen MAGs and culture-derived genomes of the same strain were compared. As many species of the rumen microbiota do not a have a representative genome that has been derived from culture, many rumen species lack a reference genome. This lack of representation in reference databases has the potential to limit the accurate functional classification of rumen metagenomic data.

However, if rumen MAGs can provide accurate functional level information, they may be suitable as reference genomes where a culture-derived reference genome is not available. A selection of functional pathways and electron transport chain (ETC) complexes to gain a useful overview of a microbe's physiology (Kracke *et al.*, 2015; Hackmann and Firkins, 2015), were used to investigate the suitability of using rumen MAGs as functional reference genomes. For each pair of clustered genomes, the predicted completeness of functional pathways and ETC complexes were compared for the MAG and culture-derived genome (Figure 3.18).

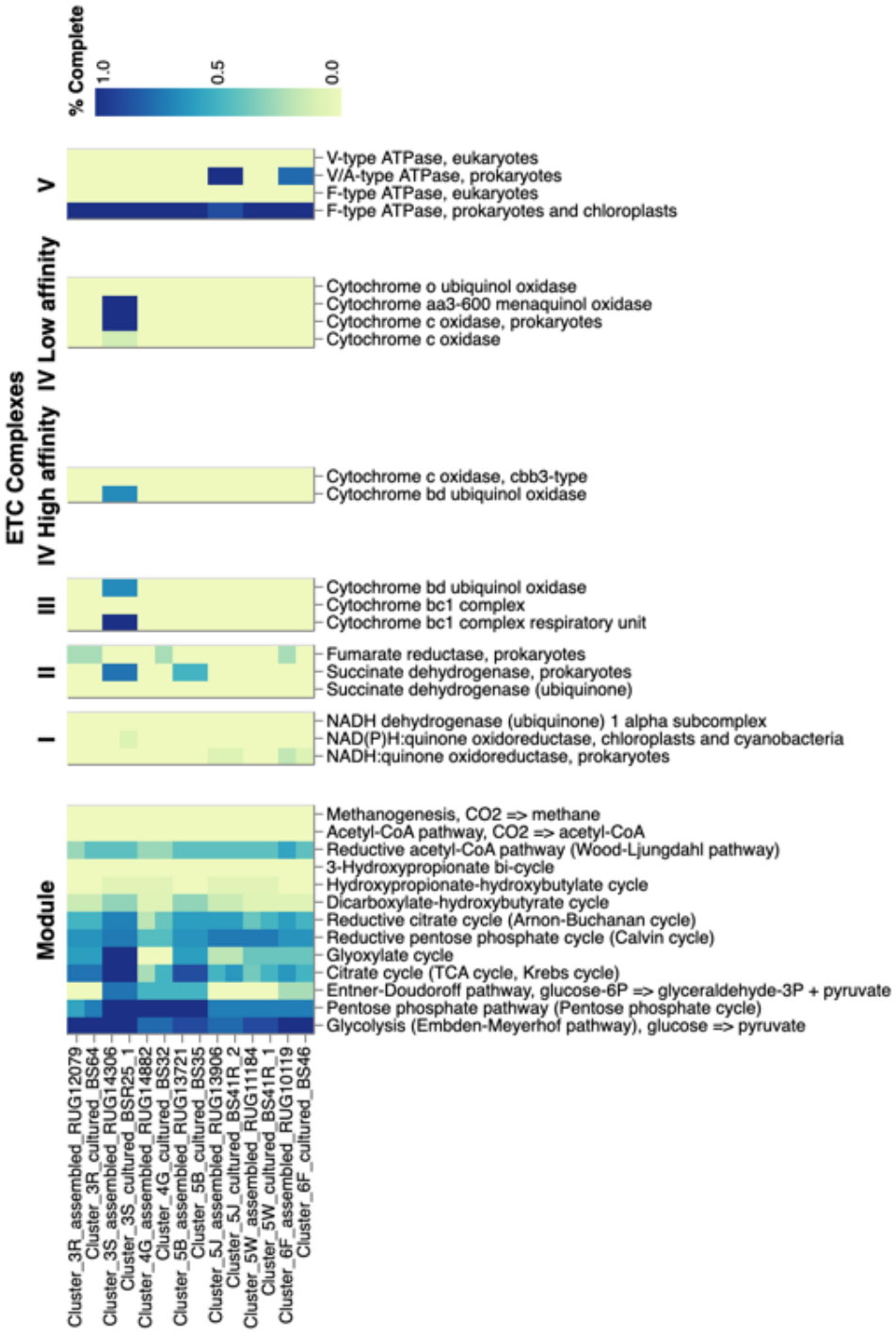


Figure 3.18: Heatmap displaying the coverage of pathways and electric transport chains (ETCs) for each cluster. The coverage of each pathway or ETC is calculated using KEGG modules, and is displayed as a percentage as shown in the key. For example, if the genome contained the genes for all subunits of an enzyme complex, it was inferred that the microbe produces this enzyme and therefore has the capability of that process. The clusters containing one metagenome-assembled and culture-derived genome pair are shown as pairs along the bottom of the heatmap.

Overall, there was a high similarity of completeness between the MAG and culture-derived genome of each paired cluster. This was especially true for ATPase ETC's (prokaryotic), several cytochrome oxidases, and succinate dehydrogenase (prokaryotic). Of the 10 pathways that were present to some degree of coverage in any cluster pair, 6 were shown to have the same coverage in the MAG and cultured isolate genome of all 7 clustered genome pairs. With the exception of cluster 6F, more often than not a culture-derived genome had higher coverage of a pathway than the MAG counterpart. These results demonstrate that MAGs have the potential to accurately predict the presence, and in some cases coverage, of functional pathways and ETC complexes, even if their overall quality may not be quite as high as those derived from cultured isolate genomes. This indicates that MAGs could be used as reference genomes to accurately classify function, which could be particularly useful when classifying data from the rumen, or other environments that are not well represented in public reference databases.

For some clustered genome pairs there were differences between the MAG and cultured genome. For example, the cultured genome in cluster 4G (BS32) saw a higher completeness for some pathway modules such as the citrate cycle, compared to the MAG (RUG14882). Conversely, the MAG (RUG10119) of cluster 6F had a higher completeness for some complexes, such as the reductive acetyl-CoA pathway, than the cultured genome (BS46). This indicates that MAGs may lose some gene-level resolution during assembly, suggesting that MAGs may not always be suitable for accurate high-resolution functional pathway prediction. Furthermore, these preliminary results could be built upon, by exploring how this accuracy and resolution could be improved. For example, the methods by which MAGs are assembled could impact the accuracy and completes of functional predictions.

Next, the prediction of metabolic functions that are particularly relevant for rumen microbiome research, for example, carbohydrate metabolism and SCFA production, were compared for the culture-derived genome and MAG pairs (Figure 3.19). Overall, if a function was predicted as being present, it was present in both the MAG and culture-derived genome of each paired cluster. This was true for all functions in the nitrogen metabolism, and the SCFA and alcohol conversions groups.

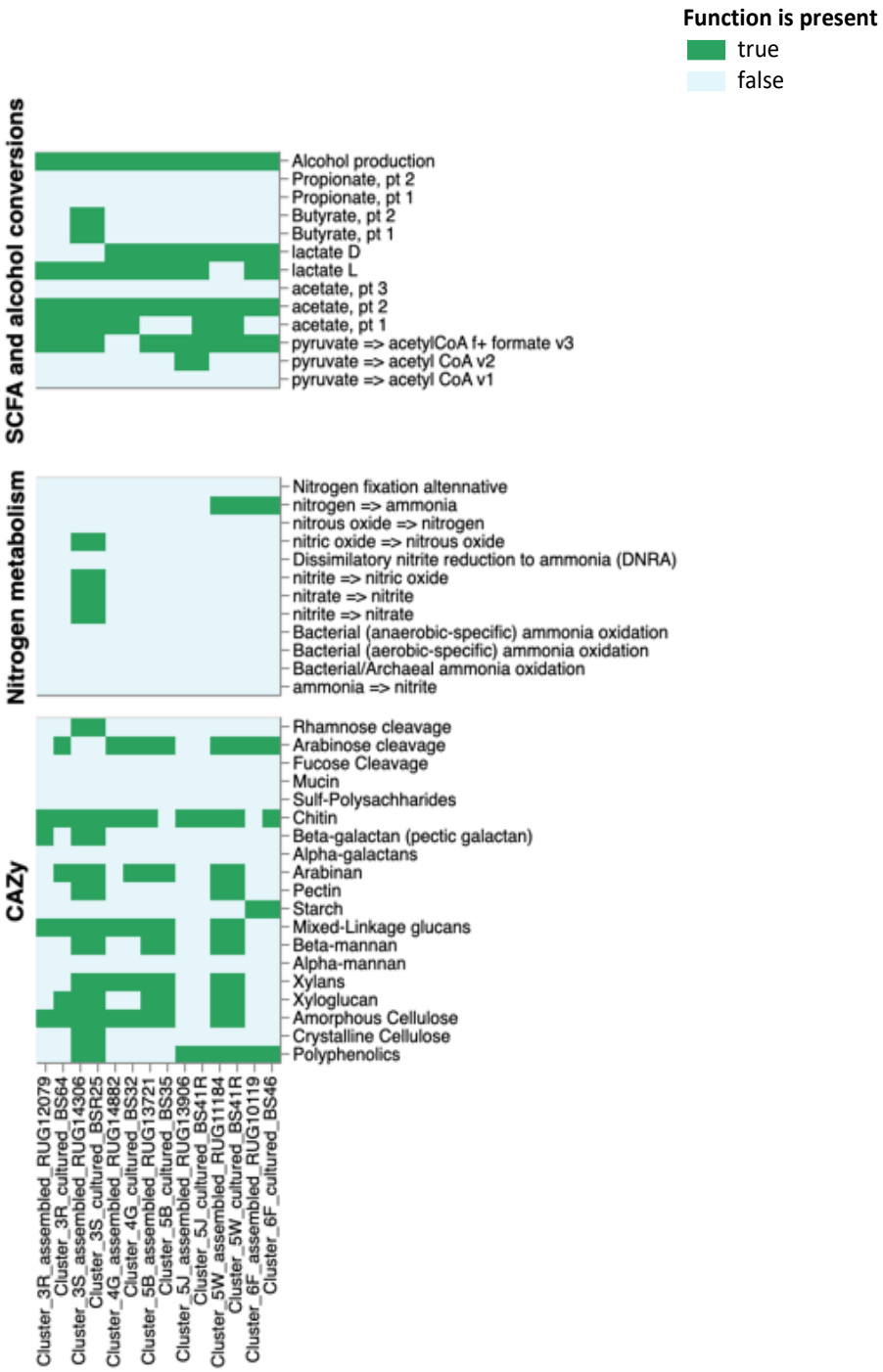


Figure 3.19: Heatmap displaying the presence or absence of metabolic functions for each cluster. The presence of a function was determined by the presence of a gene or genes required for that function or process. The clusters containing one metagenome-assembled and culture-derived genome pair are shown as pairs along the bottom of the heatmap. The functions are grouped into CAZy (Carbohydrate active enzymes), nitrogen metabolism, and SCFA (short chain fatty acid) and alcohol conversions.

However, there was more variation with carbohydrate active enzymes. The most variation was seen in cluster 3R, with four CAZymes that differed in presence between the culture-derived genome (BS64) and MAG (RUG12079). The culture-derived genome in cluster 4G (BS32) was predicted as having the function to metabolise the carbohydrate arabinan, but the function was absent in the MAG (RUG14882). In cluster 5B, the MAG (RUG13721) was predicted as having the function to metabolise the carbohydrate chitin, but the culture-derived genome (BS35) was not. The opposite was true for genomes in cluster 6F, chitin metabolism was present in the culture-derived genome (BS46) but absent in the MAG (RUG10119).

Interestingly, three of the clusters had identical functional predictions. These were clusters 3S, 5J and 5w, and the functions predicted as being present or absent by DRAM were the same for the cultured genomes and MAGs. This is promising, as it suggests that these MAGs could be used to accurately predict these functions for rumen metagenomic data.

Due to their conserved inheritance, ribosomal RNA genes are often used as marker genes for taxonomic assignment purposes. The presence or absence of 16S, 23S and 5S rRNA genes in each MAG and culture-derived genome pair, were compared (Table 3.4). In general, the rRNA genes were more present in the culture-derived genomes than the MAGs. Only 3 of the MAGs had any rRNA genes annotated, compared to all 7 of the culture-derived isolates. These results therefore suggest that MAGs may not be suitable to resolve genomes inclusive of rRNA genes. This is unsurprising as short-read MAG assembly struggles to resolve repeating or conserved regions of the genome, such as rRNA genes (Nelson *et al.*, 2020). However, MAGs assembled from long-read data may be more suitable, as long-reads are more likely to overlap genes and reduce the occurrence of mis-assembly (Liu *et al.*, 2022).

Table 3.4: The 16S, 23S and 5S ribosomal RNA genes present in each genome.

		Gene		
Cluster	Genome	5S rRNA	16S rRNA	23S rRNA
Cluster 3R	Cluster_3R_assembled_RUG12079	1	0	0
	Cluster_3R_cultured_BS64	2	0	0
Cluster 3S	Cluster_3S_assembled_RUG14306	2	0	0
	Cluster_3S_cultured_BSR25_1	7	1	0
Cluster 4G	Cluster_4G_assembled_RUG14882	0	0	0
	Cluster_4G_cultured_BS32	6	1	1
Cluster 5B	Cluster_5B_assembled_RUG13721	0	0	0
	Cluster_5B_cultured_BS35	3	0	0
Cluster 5J	Cluster_5J_assembled_RUG13906	1	0	0
	Cluster_5J_cultured_BS41R_2	2	0	0
Cluster 5W	Cluster_5W_assembled_RUG11184	0	0	0
	Cluster_5W_cultured_BS41R_1	0	1	0
Cluster 6F	Cluster_6F_assembled_RUG10119	0	0	0
	Cluster_6F_cultured_BS46	0	0	1

3.6: Discussion

The aim of this study was to identify and compare culture-derived and metagenome-assembled genomes of the same species or strain from the rumen microbiome, in order to highlight any differences or similarities between MAGs and culture-derived genomes. Identifying differences may give insights into the suitability of using a MAG as a reference genome, as MAGs have the potential to represent uncultured species in reference databases.

During this project, microbial genomes were generated from cultured bacteria isolated from two rumen samples, and MAGs were assembled from metagenomic data derived from the same samples. A limitation of this work was the culturing methodology. For instance, the temperature of the rumen is 39˚C and the incubation temperature used here was 37˚C. This temperature difference likely added a selection pressure which would have affected the type of microbes that were able to grow. Additionally, the cultures were enriched for just 24 hours before being incubated for 48 hours. Many rumen microbes require more time to grow, which means that this methodology selected for fast growers in the rumen. These factors will have impacted the diversity and taxa of microbes that were isolated. Sixteen of the culture-derived genomes were mixed, such that they were able to be separated into multiple genomes. As different methodologies were used on the samples (for example, some underwent ethanol shock treatment, some were purified and re-eluted into a different buffer etc.), these were examined to see at what step the contamination might have been introduced. These samples were processed several times using multi-well plates, at the DNA extraction, library preparation and, for those that needed it, DNA purification. The use of multi-well plates has been reported to increase the occurrence of well-to-well contamination (Minich *et al.*, 2019; Walker, 2019). Additionally, the DNA samples were transported for sequencing in a 96 well plate that was sealed

using strip lids. All but one sample in row 7 of the plate were mixed, suggesting that the cause of this contamination may have been an insecure strip lid. However, it is likely that some of these were simply mixed cultures which failed to be isolated during culturing. This was a limitation of this study, and if this work was to be continued, isolates could undergo several rounds of subculturing to ensure they are clonal isolates and not mixtures.

Of the high-quality, filtered, cultured isolates, 29 did not cluster with the Hungate genome collection, and 2 of these genomes were not assigned a species level taxonomy label by GTDBtk. This suggests that genomes BSR6R and BS49_2, which were assigned the genera *Virgibacillus* and *Streptococcus* respectively, may be novel species that have never been cultured before. *Streptococcus bovis* is a facultative anaerobe (Herrera *et al.*, 2009), *Virgibacillus* is a facultative anaerobe genus (Heyndrickx *et al.*, 1998), and has previously been observed in the rumen via 16S rRNA gene sequencing (Lima *et al.*, 2015). For all samples, glycerol stocks were prepared for long-term storage at the Rowett Institute, and it would be interesting to further characterise these two microbes and their genomes, to determine if they are indeed novel.

Of the MAGs that were assembled and binned from the two rumen samples, those that had greater than 80% completeness and less than 10% contamination, were selected for comparison with the cultured isolates, the Hungate collection (see Seshadri *et al.*, 2018) and the RUG2 superset (see Stewart *et al.*, 2019). Two MAGs did not cluster to another genome at the species level were not assigned a species level taxonomy label, which suggests they may be novel species. At the genus level, they were assigned the labels *Eubacterium*, *Saccharofermentans* and *Prevotella*. Members of the genus *Eubacterium* are present in the cow rumen, and are associated with hemicellulose degradation (Taguchi *et al.*, 2004). Chen *et al.* isolated a bacterium from the phylum *Firmicutes*, believed to belong to the novel genus

Saccharofermentans and novel species *Saccharofermentans acetigenes* (Chen *et al.*, 2010). This bacterium was unable to digest cellulose, but did ferment polysaccharides, alcohols and several hexoses, which supports the bacterium existing in the rumen. *Prevotella* is an abundant genus in the rumen, with various roles in protein and peptide metabolism (Wallace *et al.*, 1997).

To identify MAGs and cultured isolates of the same species or strain, the MAGs were clustered with the cultured isolate genomes. None of the assembled MAGs and cultured isolate genomes clustered at the species or strain level, meaning there was no direct match between these two genome types derived from the same samples. This may be due to the rumen samples, as they had been previously frozen for some years. This may have greatly skewed the type of species that remained able to be recovered. Although the ethanol shock treatment did improve the diversity of isolates, many species appeared to have died prior to cultivation attempts. It is also possible that the taxa isolated from the samples happen to be difficult to recover from metagenomic data. It has been demonstrated previously that certain microbes are repeatedly detected by culturing, and rarely by metagenomic sequencing and vice versa, and that some cultured species are observed at low abundance in sequencing data (Rajilić-Stojanović *et al.*, 2007; Shade *et al.*, 2012).

Given the lack of paired genomes from the cultured isolates and MAGs, the cultured isolate genomes were then clustered with the RUG2 superset, and seven pairs containing a MAG and culture-derived genome clustered at 99% ANI. ANI refers to the shared regions of orthologous protein coding genes, and 95% ANI is considered the threshold for defining a prokaryotic species (Richter and Rosselló-Móra, 2009; Jain *et al.*, 2018), and 99% ANI is considered a threshold for the definition of a prokaryotic strain (Stewart *et al.*, 2019). The completeness and contamination for each pair were similar, with

one pair having identical levels of completeness and contamination (cluster 6F). For some pairs the genome metrics were very similar, while others varied. For example, the genome sizes of cluster 5B varied by less than 1%, and the GC% and number of predicted genes were very similar. In contrast, the genome length of cluster 4B was less similar, varying by 27%, and the culture-derived genome 40175wE3_BS32 had over 300 more predicted proteins than the MAG RUG14882.

A pangenome is the collective genes of a microbe, and a closed-pangenome describes a microbe that has genes in common between species, with limited variation and a finite number of genes found in the population. In contrast, an open-pangenome describes a microbe with more variation between species, and comparatively a higher diversity of genes found in the population. Members of the *Streptococcus* genus are known to be incredibly diverse, with an open-pangenome (Gao et al., 2014; Zhemin Zhou et al., 2020), which is particularly true for the *Streptococcus equinis/bovis* complex (Papadimitriou *et al.*, 2014). This diversity may explain why the genomes in cluster 3R had the lowest sequencing identity, with the majority of contigs mapping at less than 25% identity across the assemblies. In addition, the variation of streptococci genomes may explain why the MAG of cluster 3R was assigned the species *Streptococcus equinus*, while the culture-derived isolate was assigned *Streptococcus* sp002393675, despite clustering at >99% ANI. This was unusual, as all other clusters were assigned the same taxonomy by GTDB-Tk.

The genome assemblies of the MAG and culture-derived genome for each strain clustered pair were aligned to one another, and some contigs did not align well between the MAG and cultured isolate genome assemblies. Cluster 4G containing culture-derived genome 40175wE3_BS32 and MAG RUG14882 saw almost 25% of contigs not matching. This is likely due to the MAG appearing to have fewer contigs than the culture-derived genome.

However, the dot plot alignment showed the overlapping sequences of the genomes to be highly similar. These genomes were assigned the species *Limosilactobacillus mucosae* (formally *Lactobacillus*), the accessory genome of which is less variable than other species of the genus (Ksiezarek *et al.*, 2022) which likely explains why the two genomes aligned with high identity. However, there were also some gaps in the alignments, suggesting that the MAG contained some sequences that were absent in the cultured isolate genome. The cultured isolate genomes were sequenced from extracted DNA of high concentration, and while some contigs were identified as contamination and removed, the filtered genomes were considered complete by CheckM. In contrast, the MAGs consisted of contigs that had been binned and assembled based on coverage information, which could have originated from multiple strains. Therefore, it is less likely that the cultured isolate genomes are incomplete, and more likely that the MAGs contain erroneous contigs that belong to another genome.

The culture-derived genome and MAG in cluster 5B were one of the most similar pairs, with over 90% of the assemblies mapping with greater than 75% identity. The alignment of cluster 5B showed some contigs that did not align well across both genomes. However, they appeared to be present in both genomes, but in different locations. It may be the case that these contigs contained genes that are shared between the genomes, but have not been placed in the correct location during assembly. For example, mobile genetic elements may be shared between microbes of the same strain, but be present in different locations of the genome (Xie, 2021).

Functional and gene information was assigned to each clustered genome pair. As coverage of a pathway or ETC was calculated as a proportion, there was more variation across genomes compared to the function analysis, which stated whether a function was present or absent in a genome. Most ETC complexes appeared to be present with the same coverage in both the

MAG and cultured isolate genomes of a given cluster. Fumarate reductase (prokaryotes) was present at similar coverage for both genomes in cluster 3R, but only for the assembled genome of cluster 6F and the culture-derived genome of cluster 4G. Interestingly, cluster 4G had differences in coverage between the culture-derived and the assembled genome for some pathways. For cluster 4G it appeared as though the MAG had missing genes, and was therefore unable to provide all information that had been learned from the cultured isolate genome. However, the opposite was true for cluster 6F, where the culture-derived genome had less or no coverage for an ETC or pathway compared to the paired MAG. This is surprising, as the cultured isolate genomes were produced from high-quantified DNA extracts and were assessed as highly complete and minimally contaminated by CheckM after filtering. Therefore, it is unlikely that the cultured isolate genome was missing multiple sections. Instead, it is more likely that the MAG of cluster 6F, despite being assessed by CheckM as over 99% complete, contained sections of other genomes from multiple strains. Another possibility is that the two clustered genomes were not from the same strain, despite clustering with extremely high ANI. In addition, the two genomes in cluster 6F were classified as the species *Sharpea azabuensis* by GTDB-tk. This species was first isolated in 2008 (Morita *et al.*, 2008) from horse faeces, and at the time of writing the genus *Sharpea* has just four reference genomes in NCBI (two with assigned taxonomy). As there are limited reference genomes, it is not yet known whether these species have an open or closed pangenome, but knowing this may give insight as to whether a high amount of gene-level variation between genomes would be expected.

Ribosomal RNA (rRNA) genes were annotated for each clustered genome, and they were present in far more of the culture-derived genomes than the MAGs. This was expected as conserved or repetitive regions are often lost during MAG binning and assembly (Meziti *et al.*, 2021). Despite the use of rRNA genes as a criterion of a high-quality MAG in addition to high

completeness and low contamination (MIMAG, (Bowers *et al.*, 2017)), MAGs often lack rRNA genes (Parks *et al.*, 2017). However in this study some of the MAGs did contain 5S rRNA genes, likely due to the 5S gene being shorter in length (approx. 120 nt (Szymanski *et al.*, 2002)) than the 16S and 23S rRNA genes, and may be captured by short-read assembly. Analysis of the 16S rRNA gene is still the most popular way to study microbiota (Pollock *et al.*, 2018), and with current methods, MAGs may not be useful to improve 16S rRNA gene-based taxonomy studies.

As MAGs are draft genomes, it might be expected that they do not contain enough complete genes to predict function to the same extent as a cultured reference genome. However, for some of the genome clusters metabolic functions were predicted as being present in both the culture-derived and assembled genomes within a pair. Some clusters had more variation between the culture-derived and assembled genome. For example, cluster 3R had differences in the predicted metabolic function, ETC complexes and pathways as well as a lower identity when aligned. As previously discussed, the taxa of cluster 3R, *Streptococcus*, is likely to have an open pangenome with variation between strains (Zhemin Zhou et al., 2020). Conversely, the genomes in cluster 3S had no differences in predicted function or pathway/complex coverage. In addition, the majority of the contigs of genomes in cluster 3S aligned with >90% identity. A contribution to this similarity may be the taxonomy of cluster 3S, which is *Bacillus licheniformis*. Dindhoria *et al.* conducted a pan-genome analysis of 13 strains of *Bacillus licheniformis* MCC 2514, which revealed a pangenome of 6008 genes, 3775 of which were core (Dindhoria *et al.*, 2022). Dindhoria *et al.* suggest that *Bacillus licheniformis* MCC 2514 has an open pangenome. However, the rarefaction curve showing the frequency of unique genes looks as though it is starting to plateau after only 13 genomes. This would suggest that *Bacillus licheniformis* has a more closed pangenome than *Streptococcus*, which has a rarefaction curve that does not appear to be reaching a plateau after over

3000 genomes were included (Zhemin Zhou et al., 2020). Although preliminary, these findings suggest that there may be more variation between MAGs and culture-derived genomes of taxa that have open pangenomes, and less variation for species that have less open pangenomes. Therefore, when thinking of using MAGs as reference genomes, one may want to first consider the taxonomy and how much gene-level variation there may be between members of that species. In addition, the success of MAG assembly must be considered when attempting to represent actual microbial species. Therefore, it is important to consider that for certain environments this may not be a useful approach. For instance, metagenomic binning and assembly methods are often optimised with bacterial genomes in mind, and so *de novo* assemblies may not accurately represent fungal genomes as well as bacterial assemblies. Methodological approaches such as long-read DNA sequencing, hybrid assembly, and single-amplified genomes (SAGs) may be more appropriate.

3.7: Conclusions

While the functional classifications of MAGs are informative, assigning functional information to genomic sequences has its limitations. It is important to remember that the only way of definitively knowing whether a microbe has the ability to carry out a given function, such as metabolising a particular substrate, is to characterise it in the laboratory. Ultimately, MAGs may not yet be suitable to provide accurate and complete genome information, and this may be impacted by differences in taxonomy. However, in this study some MAGs showed similarity to cultured isolate genomes, providing accurate gene and functional information. This was especially true for certain taxa, such as *Bacillus licheniformis*, suggesting that for microbes with a relatively closed pangenome or limited strain to strain variation, MAGs could closely resemble a culture-derived genome. This work has demonstrated the potential of using MAGs as reference genomes. The use of MAGs could

provide representative genomes where a culture-derived genome is not available. This would be particularly impactful when characterising environments such as the rumen or the soil microbiome, which have relatively few cultured isolates. In conclusion, this study has demonstrated that MAGs can provide useful information about species that are not yet cultured, but highlights the potential limitations that need to be considered.

Chapter 4: Assessing the functional prediction of metagenomics data

Industrial supervisor: Paul McAdam, Head of Next Generation Sequencing, Fios Genomics

4.1: Context of this project

This project was a collaboration between Fios Genomics and The Roslin Institute, University of Edinburgh as part of my CASE studentship requirements. This work was conducted during an industrial placement with Fios Genomics. The industrial supervisor for this project was the head of Fios Genomics' Next Generation Sequencing department, Dr Paul McAdam.

4.2: Introduction

4.2.1: Microbiome analysis and industry

Microbiome research has undergone an explosion of growth in the past 20 years (Jones, 2013). Advancements in the characterisation of the human microbiome are translating into healthcare and clinical application (Wilkinson *et al.*, 2021). The growth of human microbiome research specifically (Stulberg *et al.*, 2016) has also led to an increase in human microbiome data, with only 3 samples available from the BioSample database (NCBI) in 2010 increasing to over 123,000 in 2020 (Abdill *et al.*, 2022). The human microbiome project was a research initiative launched in the USA which sought to improve understanding of the microbiota involved in human health and disease (Proctor *et al.*, 2019). Much human microbiome research is focussed on the gut, including study of links between the microbiome and gastrointestinal disorders such as irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD) (Bouskra *et al.*, 2008; Collins and Bercik,

2009; Rhee *et al.*, 2009). Additionally, the human gut microbiota is thought to influence the immune response (Cash *et al.*, 2006; Mazmanian and Kasper, 2006), which is relevant to other research areas such as discovering drug targets (Zaneveld *et al.*, 2008), the development of vaccines (de Steenhuijsen Piters *et al.*, 2019) and cancer research (Garrett, 2015). These inter-disciplinary fields with a microbiota aspect are at the forefront of scientific research, and are receiving extensive interest, funding and investment (Eisenstein, 2020). According to the BCC Research report "Microbiome therapeutics: Global markets" dated November 2020, the global market for microbiome therapeutics is set to rise from $141.7 million (2021) to $1.3 billion by 2026 (Jhamb, 2020), reflecting the fact that microbiome research is becoming more and more profitable, and as such is a growing area of interest for industry. This also applies to metagenomics sequencing, and there are now a number of companies that either incorporate this methodology into their own R&D pipelines, or offer commercial services providing metagenomics sequencing and analyses for consumers (Gullapalli, 2020).

4.2.2: Functional classification of sequence data

For companies that use or provide metagenomics services, optimal methodology choices are clearly important for their business model. In Chapter 2 I highlighted the impact of methodology choices on taxonomic classification of metagenomics data, but it is also important to consider the accuracy of functional classification of that data. Functional annotation is the process of assigning a functional label to a nucleotide or protein sequence, which has been previously described in this thesis, see Section 1.6. Classifying functional information to high-throughput metagenomic data poses unique challenges, as the data contains fragments of potentially thousands of different genomes. As there are such vast amounts of data to

annotate, manual curation is infeasible. Instead, equally high-throughput annotation is needed.

There are two approaches to assigning function to metagenomic data: assigning function to assemblies or assembly-free methods. For methods that use assembly, metagenomic data would first be assembled into contigs before function is annotated using sequence homology, for example using BLAST (Basic local alignment search tool), which can search DNA and protein sequences, motifs and gene information (Altschul *et al.*, 1990). Other examples include Profile Hidden Markov Models (HMMs) (Eddy, 2011) or position-specific weight matrices (PWMs) as used in MEGAN4 (Huson *et al.*, 2011). A disadvantage of assembling reads into contigs is that it can require a lot of computational resources and may be slow. In contrast, assembly-free methods annotate raw reads directly using sequence homology or read mapping. Example software packages that carry out this task include MG-RAST (Meyer *et al.*, 2008) and HUMAnN3 (Beghini *et al.*, 2021). HUMAnN3 utilises DIAMOND (Buchfink *et al.*, 2015) for protein alignments, which reduces run time. Metagenomic data is large in size, and can be several gigabytes or terabytes. This means that, in addition to high storage demands, the processing of metagenomic data requires high amounts of computational resources, and the ability to spread the 'processing load' over multiple cores, known as multi-threading. These requirements translate into expense, and the cost of metagenomic data analysis is an important factor to consider when selecting what tools to use.

4.2.3: Industrial considerations and aims of project

Fios Genomics are a company that primarily specialise in offering bioinformatics services. They have an international client base, including academics, healthcare providers and biotechnology companies. The business sought to add functional metagenomic analysis to the range of

services they currently offer. The aim of this placement was therefore to select functional metagenomic classification pipelines and compare their outputs. Liaising with the company allowed me to understand the operational and business needs. This highlighted factors that needed to be considered when choosing a pipeline, which are discussed in Section 4.3.1. Regular meetings ensured the business needs were prioritised, and ensured the outcomes of the project were relevant and applicable to the company moving forward.

4.3: Materials and methods:

4.3.1: Pipeline selection

The aim of this project was to select functional annotation pipelines that seemed most suitable to be incorporated into the working practices at Fios Genomics, and compare their performance. First, discussions with the company highlighted factors that needed to be considered when choosing a pipeline, which are summarised in Table 4.1. Functional annotation pipelines were then researched and assessed based on their suitability.

Table 4.1: Factors that needed to be considered when selecting a pipeline that was suitable for the aims and needs of the business.

Ease of use	The business operates within Docker containers, so it was vital that the pipeline could be pulled as a Docker Image.
	For ongoing technical support, it was important to the business that the chosen pipeline was well maintained and kept up-to-date.
	To ensure the pipeline could be smoothly integrated into the business' services, it was important that the pipeline was relatively straightforward to use and that supporting documentation was detailed and clear.
	The client receives a full report of the analysis in PDF format, and so the chosen pipeline would ideally produce

	results either in a visual format or that which could be simply made into a visual output.
Financial cost	The financial cost of adding the pipeline was an important factor. It was crucial to consider the licensing of the pipeline and any tools it uses therein.
	Additionally, the chosen pipeline would ideally not require excessive computational resources or time to run.
Attractiveness to clientele	The pipeline would be designed to be used for the functional profiling of a metagenomics dataset.
	As it is common for clientele to be interested in the functional profiling of human-associated microbiota, tools that were specifically designed for human-associated microbiota were desirable.
	To ensure services offered are competitive, the pipeline would ideally be popular, well-cited, and published in a high impact journal.
	Several clients prefer analysis results to be available as normalised results, meaning they have been adjusted in the context of the sample. Ideally the pipeline would

	include a normalisation step, or would produce outputs that would be simple to normalise in-house.

The business was particularly interested in specific pipelines as they have been well described and utilised previously in the context of the human microbiome, such as Carnelian (Nazeen *et al.*, 2020). Carnelian has been demonstrated to find functional trends among different populations, and so was attractive to Fios Genomics as a pre-built comparative pipeline. As Carnelian assigns function using EC numbers, it can be directly compared to metabolism and pathway information, which the business viewed as practical for downstream analysis. As some of Fios Genomics' clientele are particularly interested in enzyme activity and metabolism, this was considered an advantageous feature of the software. In addition, Carnelian uses a gapped k-mer binning approach, which as a method is less computationally resource heavy than other approaches such as read alignment (Clausen *et al.*, 2018). Another key advantage is that Carnelian can be trained to directly classify reads with functional information, even if the read corresponds to a protein absent from reference databases (Nazeen *et al.*, 2020). Therefore, the Carnelian pipeline is theoretically able to accurately classify the function of proteins in novel species.

The HUMAnN3 pipeline, the successor of HUMAnN (Abubucker *et al.*, 2012) and HUMAnN2 (Franzosa *et al.*, 2018), was chosen to be included in this analysis as it is open-source, and boasts impressive precision (Beghini *et al.*, 2021). In addition, HUMAnN3 has the ability to regroup functional classification to represent other labelling methods, including EC, which as previously mentioned, is an advantage from a business perspective. HUMAnN3 assigns function using UniProt Knowledgebase (UniProtKB), assigning microbial function by annotating sequence data with UniRef clusters. Briefly, HUMAnN3 uses the database ChocoPhlAn 3 (also Biobakery 3 (Beghini *et al.*, 2021)), which contains pangenomes annotated

from UniRef90/UniProtKB (Suzek *et al.*, 2007) (and NCBI (Sayers *et al.*, 2022). By annotating pangenomes with cluster information, HUManN3 essentially creates pan-proteomes. These annotations are refined by selecting a protein for that cluster that represents the desired species using taxonomic information provided by MetaPhlAn (also Biobakery 3 (Beghini *et al.*, 2021)). The pipeline then classifies function using these representative proteins which are then used to combine annotations from various sources that are associated with the UniProtKB entry– e.g., GO terms, EC numbers, KEGG modules, KO identifiers, Pfam accessions and eggnog accessions. The variety in functional terms was considered an advantage from Fios Genomics' perspective, as they could adapt the final report to show specific functional information at the request of the client.

Table 4.2 summarises some of the functional annotation pipelines considered for this comparative study, and the advantages and disadvantages of each pipeline. Crucially, any software had to be free of licensing restrictions for commercial and academic purposes, including any tools used within pipelines. Indeed, because of how important financial overheads are to the business, licensing was the most common reason that a pipeline was not considered for this project. For example, several pipelines use the KEGG pathway for annotation, a resource which is not freely available in the public domain, as all commercial use of KEGG and associated resources require a license. Ultimately, the Carnelian and HUMAnN3 pipelines were chosen as they satisfied the most factors that were important to the business. As HUMAnN3 annotates reads using sequence homology whereas Carnelian uses a gapped k-mer binning method, the company felt that they would be an interesting comparison.

Table 4.2: A summary of the functional annotation pipelines considered for this comparative study, together with the advantages and disadvantages of each.

Pipeline	Pros	Cons	Chosen?
Carnelian	Carnelian has been designed with comparative functional classification in mind, particularly between a test and control group within a population Carnelian is freely available with an MIT (Massachusetts Institute of Technology) license. The EC (Enzyme Commission) "gold standard" database provides direct mapping to KEGG metabolic pathways	Annotates function as metabolic pathways in EC terms, which is only useful for clients if they are looking at enzymes specifically. Uses an internal "gold standard" database, which only includes prokaryotic and metabolomic annotations.	Yes – As Carnelian has MIT licencing it is freely available for commercial use, however as the EC annotation is suitable for direct mapping with KEGG, this is advantageous as it is a commonly used database and so may be favoured by clientele. Additionally, Carnelian was designed with comparative functional metagenomics in mind, which is advantageous as several clients have projects involving a comparison

	Available as a Docker container.		between two datasets.
HUMAnN3	Annotates reads directly via sequence homology and read mapping Users have a choice of translated search database: Full UniRef90, EC-filtered UnniRef90, full UniRef50, EC-filtered UniRef50 Quantifies gene families and pathways Extensive documentation and accessible online help (e.g., bioBakery forum)	Required >16 GB memory and 10 GB storage for pipeline alone Full UniRef90 database is an additional 20.7 GB	Yes – This pipeline is extensively documented, well-cited, uses publicly available reference databases and allows for UniRef-annotated classification and regrouped (e.g., EC-annotated) classification.

	Popular choice in publications and well-cited.		
Eggnog-mapper v2	Considers gene homology and evolution. Claims that using orthology predictions for functional annotation is more precise than homology searches, because with homology-based strategies paralogs may be falsely considered as having the same function.	Does not estimate abundance. Requires ~ 50 GB storage. Uses KEGG to annotate function.	No – this pipeline is relatively resource-heavy, and requires KEGG, which is not suitable as it requires a licence for commercial use.
MetaErg	Multiple functional annotation searches: functional categories, protein domains, KEGG	Documentation is less clear than other pipelines. HTML result pages would be	No – As it outputs classification results into HTML pages, and it cannot be run within a single Docker image, the

	Orthology terms, Gene Ontology terms, EC numbers, so could be useful to adapt to client's analyses. Fully open source (Academic free license).	difficult to translate into Fios client reports. Docker image doesn't include Signal, which needs to be installed separately.	business felt it was not suitable.
MetaLAFFA	Estimates abundance at the gene and community levels. Gene abundance is based on functional orthologs. Results can be in terms of relative abundance, or converted to copy number per genome (using MUSiCC).	Aggregates KO abundances into KEGG pathway abundances. Implemented in Snakemake, which is not a pipeline workflow manager that the business use.	No – HUMAnN3 also uses a UniRef90 via DIAMOND approach, so this would likely give similar results, and the use of KEGG made this pipeline unsuitable for the business.

	GNU license v3.0 (No restrictions to use by non-academics).		
YAMP	GLU general public license. Written in Nextflow, useful for Fios, but also available as a Docker container. Output includes taxonomy composition, relative abundance of microbes, genes and pathways, and pathway coverage.	Uses HUMAnN pipeline for functional annotation.	No – As it uses HUMAnN pipeline internally it posed no advantage over using the HUMAnN3 pipeline itself
FMAP	Performs sequence alignment, gene family abundance calculations and differential abundance of	KEGG subscription must be obtained.	No – KEGG subscription made this an unsuitable option for the business

	genes, operons and pathways.		

4.3.2: Data and creation of ground truth metagenomic data

This project used ground truth data, which in this context refers to data which had known functional annotations. As discussed in Section 1.6.1, functional information is stored in databases, that are then used to classify the function of metagenomic data. UniProt is a widely-used reference database containing protein and functional information. As it is popular, and used by several functional classification pipelines, UniProt/UniRef was used to annotate the ground truth data used in this analysis. Two simulated metagenomes were chosen for this project, which represent two different microbiomes: the human gut and the cow rumen.

The simulated human gut metagenome was created by Huttenhower *et al.*, and is evaluated in the HUMAnN2.0 article (Franzosa *et al.*, 2018). In brief, the simulated human gut metagenome consists of 10 million reads, 100 bp in length, taken from the genomes of the 20 most abundant species in Human Microbiome Project (HMP). These species are *Alistipes onderdonkii*, *Alistipes putredinis*, *Alistipes shahii*, *Bacteroides caccae*, *Bacteroides cellulosilyticus*, *Bacteroides dorei*, *Bacteroides massiliensis*, *Bacteroides ovatus*, *Bacteroides stercoris*, *Bacteroides thetaiotaomicron*, *Bacteroides uniformis*, *Bacteroides vulgatus*, *Barnesiella intestinihominis*, *Dialister invisus*, *Eubacterium rectale*, *Faecalibacterium prausnitzii*, *Parabacteroides distasonis*, *Parabacteroides merdae*, *Prevotella copri*, and *Ruminococcus bromii*.

The simulated cattle rumen metagenome was simulated using 460 cultured rumen-dwelling microbial genomes from the Hungate 1000 project (Seshadri *et al.*, 2018), the genomes of which are available at *https://genome.jgi.doe.gov/portal/HungateCollection/HungateCollection.info.html*. The metagenome consists of 50 million paired end reads, 126 bp in length, and was created using the tool Insilicoseq (version 1.4.6) with the HiSeq exponential model (MIT licence, (Gourlé *et al.*, 2019)). This dataset is described in more detail in the manuscript included in Chapter 2 of this thesis and is available for download at *https://doi.org/10.7488/ds/3444*.

4.3.2.1: For all pipelines: protein mapping and estimating protein abundance in the simulated data

The process for creating the ground truth data is outlined in Figure 4.1. The reference genomes which were used to create each of the simulated metagenomes were downloaded as FASTA files. Proteins encoded by those reference genomes were predicted using Prodigal (version 2.6.3) (Hyatt *et al.*, 2010). A BWA index (BWA version 0.7.17) (Li and Durbin, 2009) was created of all reference genomes for each sample, and the simulated reads for that sample were then mapped back to the combined index using BWA-MEM and SAMtools (version 1.9) (Li *et al.*, 2009). The files containing predicted proteins were concatenated and filtered to include only protein coding regions. Bedtools (version 2.30.1) (Quinlan, 2014) was then used to count how many reads overlapped a gene by >30%, and a script parsed the final counts to an output file. Each comparison required a different ground truth, and so the next steps varied for the creation of each ground truth as explained below.

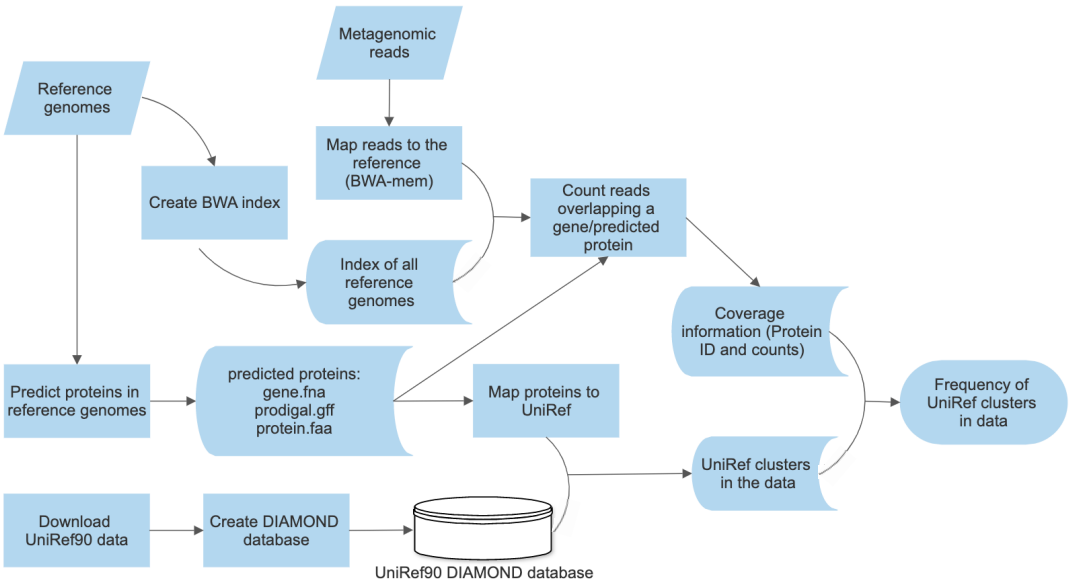


Figure 4.1: The workflow that was followed to create the annotated ground truth data. The same process was followed for both the human gut and rumen data. For the ground truth data to be compared with Carnelian, the current version of UniRef90 was downloaded, and for comparison with HUMAnN3, version 2019_01 was downloaded.

4.3.2.2: For HUMAnN3: creating ground truth data that is annotated with UniRef IDs and read counts

HUMAnN3 annotates UniRef90 labels from ChocoPhlAn pangenomes. ChocoPhlAn3 is a database containing reference genes and gene catalogues, based on the contents of the UniProt Proteomes portal as of January 2019 (Beghini *et al.*, 2021). UniProt release 2019_01 was announced on 16/01/2019, and it was this version of UniProt that was used to annotate the ground truth for comparison with HUMAnN3 classification results.

A DIAMOND (Buchfink *et al.*, 2015) database was built using the downloaded UniRef2019_01 FASTA sequences. For each metagenome, the proteins were concatenated into one file, and then mapped against the DIAMOND database. The resulting tab-separated files were then filtered such that alignment length of proteins had to be >90% of the query length and

percentage identity had to be >90% to the database. The results were then parsed to produce a table containing the UniRef IDs and the frequency they occurred within each sample. These counts formed the ground truth data, which was then directly compared with the HUMAnN3 output, the results of which are shown in section 4.4.1.1.

4.3.2.3: For HUMAnN3: creating ground truth data with EC annotations

As described in section 4.3.2.2, a DIAMOND database was built from UniRef version 2019_01 to create the ground truth data to compare with the classifications made by HUMAnN3. The simulated metagenomic reads for each sample were mapped to the database. The mapped reads were filtered such that alignment length was >99% of the query length and percentage identity was >90%. These reads were then parsed to produce a file containing UniRef cluster IDs and the frequency they occurred within each sample. UniRef cluster IDs were formatted such that they contained a UniProt ID within them. For example, “UniRef90_A0A009ES05” is a cluster ID, and “A0A009ES05” is a UniProt ID. The UniRef cluster IDs were converted to UniProt IDs, accompanied by frequency of these proteins in each sample.

UniProt version 2019_01 Sprot and Trembl files were downloaded from the UniProt server. From these files, the UniProt ID numbers and their corresponding complete EC numbers were extracted. This information was combined with the count information to create a table containing EC numbers and abundance counts for each simulated metagenome, which was then used to compare the EC annotation results of HUMAnN3 (results in Section 4.4.1.2).

4.3.2.4: For Carnelian: creating ground truth data that is annotated by UniRef90

The current version of UniRef (2021_01) was downloaded as Carnelian uses the current version of UniProt to annotate their proteins (Nazeen *et al.*, 2020). A DIAMOND database was built and for each simulated metagenome the proteins were mapped to the database. As with the ground truth data created to compare against HUMAnN3 results, the mapped reads were filtered such that alignment length was >99% of the query length and percentage identity was >90%. The frequency of UniRef IDs for each sample were extracted, and the UniRef IDs were then converted to UniProt IDs.

The current versions of UniProt (2021_01) Sprot and Trembl files were downloaded (on 10/02/21), and the UniProt ID and associated complete EC numbers were extracted. These were combined with the frequency information to produce EC numbers and their frequency for each sample. These ground truth counts were then compared with the Carnelian output, the results of which are in section 4.4.2.1.

4.3.2.5: For Carnelian: creating ground truth data that is annotated by UniProtKB

To test whether the method of ground truth creation caused a bias in favour of HUMAnN3, a different method was used to create a new ground truth that would be less biased. The original method used to create the ground truth to assess the classification results of Carnelian was made using UniRef90 (see section 4.3.2.4.). However, this version of the ground truth was made using a step that involved converting UniRef90 IDs to UniProtKB IDs. The UniProt IDs were then converted to EC numbers. However, as UniRef90 IDs contain clusters of proteins, this effectively converted a group of proteins into one single protein.

The new ground truth was created using a method similar to the method used to create the ground truth for comparison with HUMAnN3. UniProtKB was downloaded in FASTA format and used to build a DIAMOND database. DIAMOND (v0.9.22) was then used to annotate the simulated gut and rumen metagenomic data. The DIAMOND results were filtered such that alignment length was >99% of the query length and percentage identity was >90%. The UniProtKB IDs and ground truth counts were then parsed to an output file, to be compared with classified data (see Section 4.4.2.2).

4.3.3: Running the HUMAnN3 pipeline

The commands used to install and run the HUMAnN3 pipeline are shown in Supplementary Table S4.1. Briefly, HUMAnN3 v3.0.0.alpha.4 was run within a Docker container. The BioBakery3 databases ChocoPhlAn, UniRef90, and the utility mapping database were then downloaded, and the locations of each database were updated in the config file. HUMAnN3 offers various output options, but the standard is “gene family RPK”, which stands for reads per kilobase. In the documentation from the software developer, it details how this number is created:

“This is computed as the sum of the scores for all alignments for a gene family. An alignment score is based on the number of matches to the reference gene for a specific sequence. It is divided by the length of the reference gene in kilobases to normalize for gene length. Each alignment score is also normalized to account for alignments for a single sequence to multiple reference genes. Alignments are not considered if they do not pass the e-value, identity, and coverage thresholds.”

This normalisation is well suited for quantifying functional annotation; however, the business was interested in the comparison of the classification results from each pipeline. Replicating this transformation in the ground truth

data was not feasible as we were unable to replicate HUMAnN3's RPK transformation. As the transformation could not be replicated in the ground truth data, a valid comparison could therefore not be performed. Instead, a script within HUMAnN3 was edited such that it would print out the un-transformed, here-on referred to as raw counts, data for each gene family before they were transformed (Figure 4.2). Additional print commands were added to the HUMAnN3 python script store.py at line 470, which meant that when the pipeline was run the raw gene and microbial counts were printed. HUMAnN3 was then run on the concatenated FASTQ files for both datasets. This command saved the standard output of the pipeline to a file named humann_rumen.out, which was then parsed with an in-house R script to extract UniRef90 clusters and raw counts and write them to a new output file.

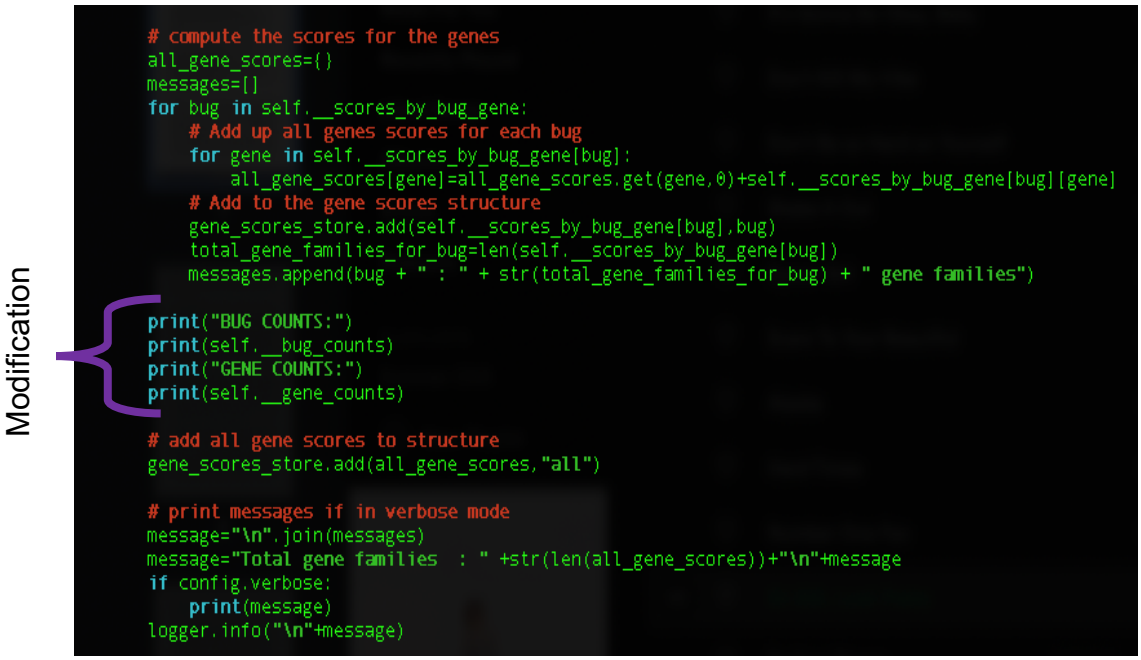


Figure 4.2: The modification made to the HUMAnN3 pipeline script such that the raw taxonomy and gene counts were printed instead of the transformed counts. The added text is referred to by the brace and is labelled "Modification".

The added text is referred to by the brace in Figure 4.2 and is labelled "Modification". This changed the store.py script such that it printed out un-

transformed counts for taxonomy and genes. The modified script store.py was located within the HUMAnN3 pipeline (at /usr/local/lib/python3.6/dist-packages/humann/store.py). The command used to run HUMAnN3 is shown in Supplementary Table S4.1.

As Fios Genomics were interested in classification to enzyme resolution, the raw UniRef counts were also converted to raw EC counts. This was carried out using the documentation for HUMAnN3, which describes built-in data normalisation or transformation options. The script human_regroup_table was used to regroup the raw UniRef cluster counts to EC counts using the option --groups uniref90_level4ec. This resulted in the HUMAnN3 functional classification results being resolved to the EC level, which were then compared with the EC-annotated UniRef version 2019_01 (see Section 4.4.1.2) and the Carnelian pipeline classification results (see Section 4.4.3.1).

4.3.4: Running the Carnelian pipeline

The Carnelian pipeline was installed, trained and run using the information given on the developer's webpage (see *http://cb.csail.mit.edu/cb/carnelian/*). Carnelian was run within a Docker container, and the command for mounting the container is in Supplementary Table S4.2.

The pipeline and dependencies were installed, and the FragGeneScan and EC-2010_DB were downloaded and extracted using the SETUP.sh script. The installation was then tested to ensure it was working correctly. The Carnelian workflow has three main components: training, functional binning, and annotating samples, also referred to as functional profiling and abundance. Firstly, the carnelian model was trained with the gold standard dataset. The commands for training and running the pipeline can be found in Supplementary Table S4.2. Within the functional binning step, the samples have to be processed. As the samples contained simulated metagenomic

reads, the adapter sequence trimming step was skipped. For each sample the forward and reverse reads were converted from FASTQ to FASTA, before being concatenated into one file. The samples were then annotated (see Supplementary Table S4.2).

The functional profiling and abundance analysis for both datasets required the label files for both samples to be moved to the same directory. The documentation stated that a sample information file had to be created, but there was no further information on how to create the file. Following discussions with the software developers it was confirmed that the sampleinfo_file is a tab-separated file with three columns: sample_id, group and fraglen (Table 4.3). The sample_id contains the unique sample identifiers as used in the file names, the group refers to the category of the samples (in this case, rumen and human gut), and fraglen contains the average read length of sequences in amino acids for each sample. The fraglen value was therefore calculated by averaging the read length for each sample and dividing by 3.

Table 4.3: The sampleinfo_file.tsv

sample_id	group	fraglen
hungate	rumen	41
human	gut	33

The functional profiling and abundance analysis was run, which produced two types of output: effective and raw counts. Raw counts were used in all subsequent comparisons with the ground truth data and HUMAnN3 classifications.

4.4: Results

4.4.1: The HUMAnN3 pipeline

4.4.1.1: UniRef cluster annotations

The main objective of the industrial placement was to assess the suitability and accuracy of functional classification pipelines for metagenomics data derived from the bovine rumen and the human gut. Firstly, the accuracy of the HUMAnN3 pipeline was assessed by comparing the classification results with ground truth predictions (the methods to create the ground truth data are described in Section 4.3.2.2). The modified HUMAnN3 pipeline was run on both the simulated human gut and rumen metagenomic data, which produced raw counts of UniRef clusters. To assess the accuracy of HUMAnN3, these classified UniRef counts were compared directly with the ground truth UniRef counts (see Figure 4.3).

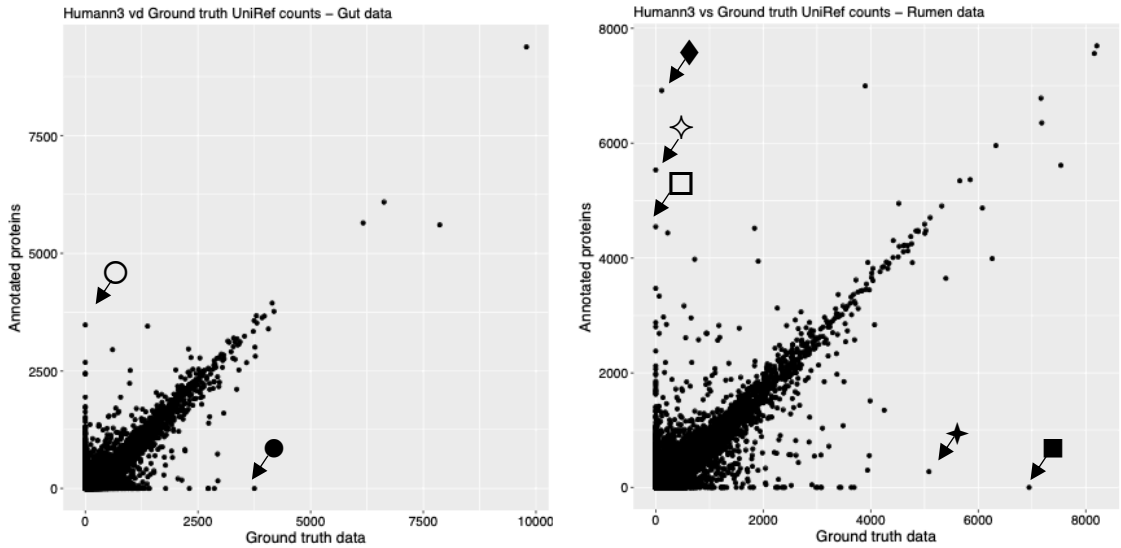


Figure 4.3: The frequency of UniRef cluster annotations in the HUMAnN3 classifications vs. Ground truth data. Comparing the frequency of UniRef annotations in the ground truth data (x-axis) and the HUMAnN3 pipeline (y-axis) for the simulated human gut metagenome (A) and simulated rumen metagenome (B). Clusters of interest are annotated with a symbol to their right and are described in Table 4.4.

Table 4.4: A selection of UniRef clusters that were either unclassified or inaccurately classified by the HUMAnN3 pipeline, corresponding to the annotations on Figure 4.3.

Annotation on Figure 4.3	Metagenome	Ground truth abundance	HUMAnN3 abundance	UniRef90 cluster
♦	Rumen	110	6919	UniRef90_Q04IK6
✧	Rumen	0	5533	UniRef90_ K9ID55
✦	Rumen	5082	276	UniRef90_Q3JYR6
■	Rumen	6946	0	UniRef90_A0A0U0K9F7
□	Rumen	0	4548	UniRef90_A4W4F3
●	Human gut	3751	0	UniRef90_D4IMY6
○	Human gut	0	3482	UniRef90_B0NLV0

In general, HUMAnN3 accurately classified some of the UniRef clusters predicted in the ground truth data, and the scatterplots in Figure 4.3 show a clear trend line for both datasets. However, of the 745,457 UniRef clusters predicted in the rumen data ground truth, 12.6% were classified by HUMAnN3 within 10% above and below the ground truth abundance. In the human gut data, the ground truth predicted 50,523 UniRef clusters, and HUMAnN3 classified 13.9% of them within 10% above and below the ground truth abundances. This means that HUMAnN3 generally classified the human gut data with more accuracy than the bovine rumen data.

As this project sought to assess whether the HUMAnN3 pipeline was accurately classifying UniRef clusters, I felt it was important to identify any clusters that appeared to be classified inaccurately. Some clusters that were classified at a considerably different abundance than what was predicted in the ground truth, were annotated in Figure 4.3, and are detailed in Table 4.4. For the rumen data, these clusters were selected as the difference between the ground truth and classified abundance that was greater than 4500 counts, and for the human gut data, the difference was greater than 3000 counts. As these clusters were the most erroneously classified by HUMAnN3, I wanted to investigate what proteins they were, as it might explain why they were not classified well.

In the rumen data, cluster UniRef90_Q04IK6, was estimated to appear in the ground truth data 110 times, and was classified by HUMAnN3 6919 times. Cluster Q04IK6 is a DNA-directed RNA polymerase subunit beta, EC number 2.7.7.6 and gene rpoB. Cluster UniRef90_K9ID55 was classified 5533 times in the data by HUMAnN3, but not at all in the ground truth estimations. Cluster K9ID55 refers to a “Conserved domain protein” and the gene PLO_2151. BLAST results showed a >90% identity with three proteins named Cell wall-associated hydrolase. Cluster UniRef90_Q3JYR6 appeared 5082 times in the ground truth estimations but only 276 times by HUMAnN3. Cluster Q3JYR6 is an enzyme called leucine-tRNA ligase, enzyme EC 6.1.1.4, and gene leuS. Cluster UniRef90_A0A0U0K9F7 was estimated to appear 6946 times in the ground truth data but not at all by HUMAnN3. This protein cluster no longer appears in UniRef, and the UniProtKB entry A0A0U0K9F7 was made redundant on December 11th 2019, which may be why it was not classified by HUMAnN3. Before it was made redundant, this entry was named DNA-directed RNA polymerase subunit beta, and the EC number of this reaction was 2.7.7.6. Running the amino acid sequence for UniProtKB entry A0A0U0K9F7 on BlastP – the top hit gave a 100% identity match to UniProtKB/Swiss-Prot accession C1CA06. Protein entry C1CA06

(RPOB_STRP7) is named DNA-directed RNA polymerase subunit beta, correlating to the gene rpoB, and as A0A0U0K9F7, the EC number is 2.7.6.6. UniRef90_A4W4F3 is no longer a cluster in UniRef90, but UniProtKB A4W4F3 refers to an enzyme called Leucine--tRNA ligase. Just as protein Q3JYR6 discussed previously, A4W4F3 has the EC number 6.1.1.4 and corresponds to the gene leuS.

In the human gut data, cluster UniRef90_B0NLV0 was classified 3482 times by HUMAnN3, but was not predicted in the ground truth data. This cluster belongs to an enzyme with EC 2.1.1.72 named “Site-specific DNA-methyltransferase (adenine-specific), correlating to the gene BACSTE_00431. Cluster UniRef90 D4IMY6 was predicted 3751 times in the ground truth data, but not at all by HUMAnN3. This cluster refers to an enzyme called “Site-specific DNA-methyltransferase (adenine-specific)” with EC 2.1.1.72. Several of these proteins appear to be the same, for example having the same function, name or EC number, but have been labelled differently by either the processes used to make the ground truth data annotations, or the HUMAnN3 pipeline. As these UniRef clusters correspond to the same enzyme, the classifications are accurate but appeared inaccurate as this comparison was to the protein cluster level.

4.4.1.2: Regrouped EC number annotations

In addition to the UniRef cluster annotations described above, the HUMAnN3 pipeline allows the UniRef cluster counts to be regrouped to other functional annotations, including EC (Enzyme commission) number. The business was interested in analysing enzymes in metagenomic data, as it is the focus of several clients. Therefore, I regrouped the raw UniRef cluster counts to provide EC number frequencies for both the human gut and rumen metagenomic data. These frequencies were compared with the EC-annotated ground truth (for methods for the creation of the EC-annotated

ground truth see Section 4.3.2.3). The HUMAnN3 EC classifications showed a clear correlation with the EC-annotated ground truth predictions for the human gut and rumen data (Figure 4.4). In comparison to the UniRef cluster-annotated ground truth and HUMAnN3 classifications, the EC-annotated data showed an obvious correlation with far fewer erroneous classifications. This may be because when groups of protein IDs were converted to one EC there was less variation between the ground truth and classified data frequencies. As the EC number classifications were far more reliable than the UniRef clusters, classifying to the EC number resolution might be attractive to the company.

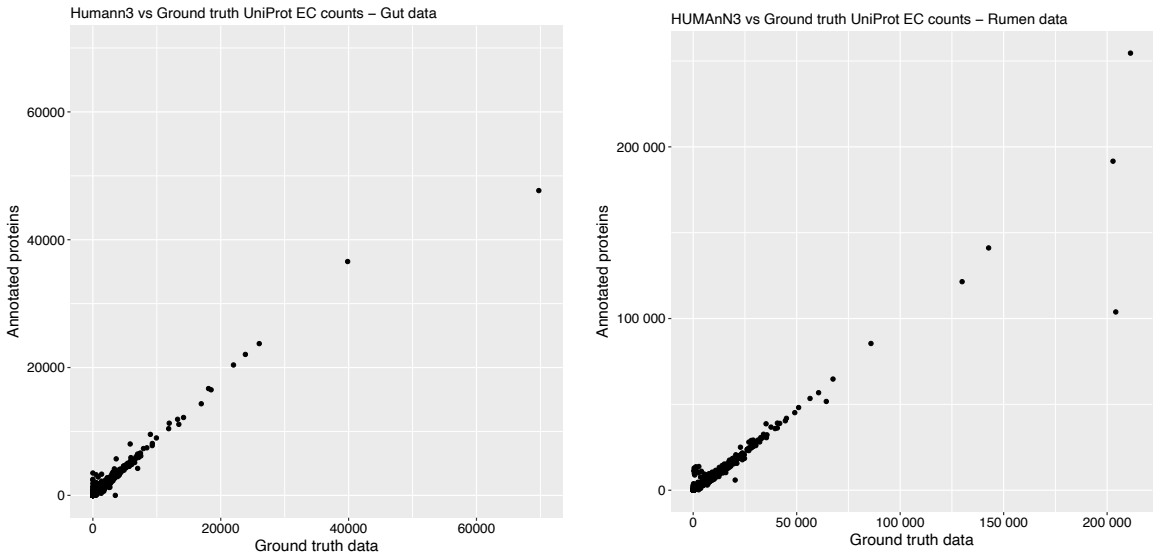


Figure 4.4: The frequency of EC number annotations in the HUMAnN3 classifications vs. Ground truth data. Comparing the frequency of EC number annotations in the ground truth data (x-axis) and the HUMAnN3 pipeline (y-axis) for the simulated human gut metagenome (A) and simulated rumen metagenome (B).

The 10 EC numbers that were classified the most by HUMAnN3 are shown in Table 4.5. These highly abundant enzymes are involved in essential cellular processes, such as DNA replication and translation. EC 2.7.13.3 refers to the catalytic activity of histidine protein kinase, which are multifunctional proteins

that typically play a role in signal transduction across the cell membrane (Engin, 2021).

EC 2.7.13.3 was present in the human gut ground truth and classified data, as the most abundant enzyme class. However, in the rumen data, EC 2.7.13.3 was annotated as occurring in the ground truth data at a frequency of 204,178 but was classified at a frequency of 103,841 by HUMAnN3. This difference in abundance was not seen in the human gut data, where the enzyme was classified at a similar abundance to the ground truth estimation. It may be that this is the result of an error during the creation of the ground truth data.

Table 4.5: The top 10 most frequent EC numbers as classified by the HUMAnN3 pipeline. The frequency of the 10 most abundant enzymes classified by the HUMAN3 pipeline and in the ground truth for the rumen and human gut data is shown in Table 4.5a. The function of these enzymes is shown in Table 4.5b (information from UniProt (The UniProt Consortium, 2021) and Expasy- (Bairoch, 2000)).

Table 4.5a:

Rumen data			Human Gut data		
EC	Ground truth	HUMAnN3 pipeline	EC	Ground truth	HUMAnN3 pipeline
3.6.4.12*	211292	254600	2.7.13.3*	69750	47690
2.7.7.7*	202847	191646	3.2.1.23*	39850	36590
2.7.7.6*	142744	141129	3.6.4.12*	26020	23744
5.99.1.3*	129935	121472	2.7.7.7*	23853	22054
2.7.13.3*	204178	103841	2.7.7.6*	21996	20401
6.3.5.5*	85902	85477	3.2.1.22	18094	16710
7.1.2.2	67553	64762	5.99.1.3*	18494	16519
3.2.1.23*	60594	56800	5.2.1.8*	16955	14332
6.1.1.20	56456	53461	7.2.1.1	14184	12186
5.2.1.8*	64304	51720	6.3.5.5*	13250	11905

Table 4.5.b:

EC (Enzyme commission) number	Function
3.6.4.12*	DNA helicase
2.7.7.7*	DNA-directed DNA polymerase
2.7.7.6*	DNA-directed RNA polymerase
5.99.1.3* -> transferred to 5.6.2.2	DNA topoisomerase (ATP=hydrolysing)
2.7.13.3*	Histidine kinase
6.3.5.5*	Carbamoyl-phosphate synthase (glutamine-hydrolysing)
5.2.1.8*	Peptidylprolyl isomerase
3.2.1.23*	Beta-galactosidase
7.1.2.2	H^+-transporting two-sector ATPase
3.2.1.22	Alpha-galactosidase
7.2.1.1	NADH:ubiquinone reductase (Na^+-transporting)
6.1.1.20	Phenylalanine--tRNA ligase

Enzymes marked with a (*) were shared in the top 10 classified enzymes for both the human gut and rumen data

4.4.2:The Carnelian pipeline

4.4.2.1: Comparing the Carnelian classifications with the UniRef-derived EC annotated ground truth

Having assessed the accuracy of the HUMAnN3 pipeline with ground truth data, I next carried out a similar process using Carnelian. The Carnelian pipeline was included in this analysis as it classifies function to the EC level, which can be directly mapped to KEGG pathways (Kanehisa and Goto, 2000). Additionally, it is freely available with an MIT license, and so is suitable for commercial use. It was also chosen as the ability to compare multiple samples within one analysis was considered to be attractive to Fios. It is also available as a Docker container, and the "gold standard" database that Carnelian uses is only 5.5 GB in size. As Fios were interested in the performance of Carnelian, ground truth protein predictions were again used to assess the functional classification results of the Carnelian pipeline. As Carnelian classifies data using EC numbers, the ground truth data was annotated with EC numbers for comparison. Briefly, the ground truth data was annotated with UniRef clusters, before these were converted to UniProt accessions which were used to look up EC numbers in UniProt Trembl and Sprot (information about the methodology to create the ground truth data for this comparison is in Section 4.3.2.4).

To estimate the accuracy of the Carnelian pipelines, the classifications made by the Carnelian were compared with the ground truth predictions (see Figure 4.5). The most abundant EC classifications for the rumen and human gut datasets are shown in Table 4.6.

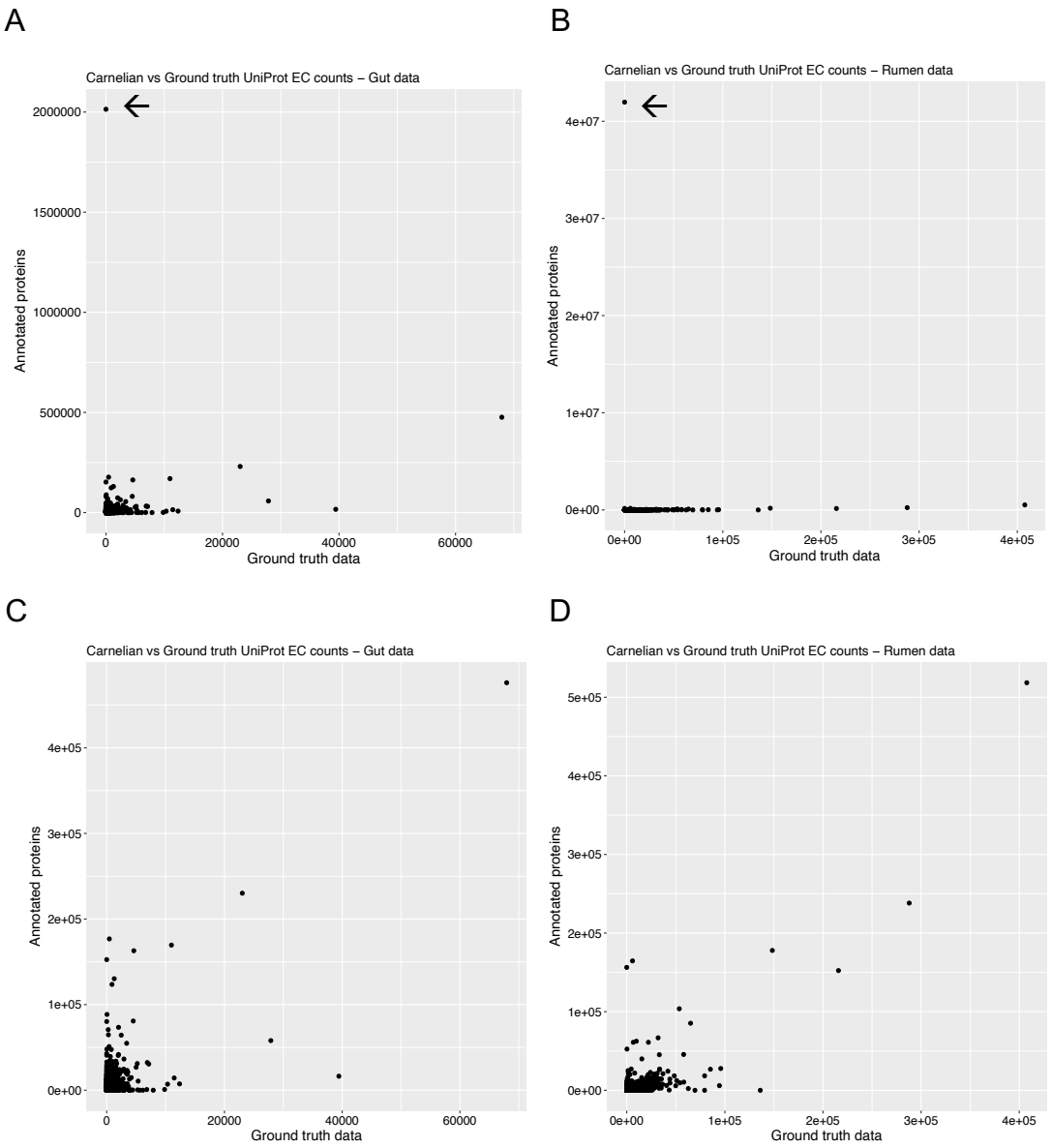


Figure 4.5: A comparison of the Carnelian output (y-axis) and the ground truth data (x-axis) as scatterplots for the human gut (A) and rumen (B) simulated metagenomes. An outlier (pointed to in A and B) was removed as it was distorting the graphs. The scatterplots without the single outlier are shown for the human gut (C) and rumen (D) data. The data plotted is the counts of EC (Enzyme Commission), representing the frequency of these enzymes in the data. The frequency of EC numbers in the data as annotated by the Carnelian pipeline is plotted on the y-axis, and the frequency of EC numbers according to the ground truth data is plotted on the x-axis.

Table 4.6: The most abundant EC numbers classified by the Carnelian pipeline and the predicted abundance in the ground truth data.

Human Gut			Rumen		
EC	Ground truth	Carnelian	EC	Ground truth	Carnelian
2.4.1.333 ←	0	2013761	2.4.1.333 ←	0	41966663
2.7.13.3	67944	476107	2.7.13.3	407672	518486
3.6.4.12	23032	230139	3.6.4.12	287861	238149
1.6.5.11	450	176772	2.7.7.6	148369	177921
2.7.7.6	10993	169542	1.6.5.11	5894	164635
2.7.7.7	4605	162944	5.99.1.3	0	156352
5.99.1.3	0	152596	2.7.7.7	215759	152282
2.7.11.1	1280	130258	2.7.11.1	53554	103716
3.2.1.4	897	123578	5.2.1.8	65055	85338
3.1.4.52	49	88542	3.2.1.4	31980	66726
5.2.1.8	4512	80815	3.1.4.52	9822	62404

The most abundant classification by Carnelian was the enzyme class EC 2.4.1.333, which was classified as occurring over 41 million times in the rumen data and over 2 million times in the human gut data, and is annotated by an arrow in Figure 4.6 and Table 4.6. This EC number refers to 1,2-β-D-glucan:phosphate α-D-glucosyltransferase (UniProt link: *https://www.uniprot.org/uniprot/Q92AT0*). This enzyme catalyses the reversible phosphorolysis of beta-(1->2)-D-glucans. The reaction is ((1->2)-beta-D-glucosyl)(n) + phosphate <=> ((1->2)-beta-D-glucosyl)(n-1) + alpha-D-glucose 1-phosphate). To see whether this was a reasonable classification, I examined the datasets. The human gut and rumen data contained 10 million and 50 million paired end reads respectively. This means that if this classification were true, then 82% of the reads in the rumen dataset would contain gene sequences belonging to this enzyme class. This is very surprising, as I would not expect this enzyme to be more abundant

than essential housekeeping proteins. Therefore, it seems likely that Carnelian was falsely reporting the frequency of EC 2.4.1.333 in the data. This could be a result of an issue with the training step of Carnelien, as the model is trained on a pre-determined set of enzymes. Another possibility is that there was an error at the classification step of Carnelian. However, I think it is more likely that this is an issue with wider enzyme reference databases, and that incorrectly labelled reference data has caused this over-classification. The datapoints for this enzyme were treated as outliers and removed from the data (see Figure 4.5 C and D). EC 5.99.1.3 has been transferred to EC 5.6.2.2, DNA topoisomerase (ATP-hydrolysing), which was not classified by Carnelian but was predicted in the ground truth data 136079 times. As shown in Table 4.6, EC 5.99.1.3 was reported by Carnelian in both the human gut and rumen data, but was absent from the ground truth. After some investigation I realised this was because EC 5.99.1.3 became obsolete in 2018, and as the ground truth used in this comparison was made with the current version of UniProt it would not be present. However, this was surprising as Carnelian also claimed to use the current version of UniProt, and so should not contain EC 5.99.1.3 either.

4.4.2.2: Comparison of the Carnelian classifications with the UniProt-derived EC annotated ground truth

When the Carnelian classification results were compared with the ground truth estimations in Section 4.4.2.1, the results did not correlate well. The ground truth data for comparisons with the results of the Carnelian pipeline were annotated with UniRef clusters (Section 4.4.2.1). This ground truth was made by annotating the predicted proteins with UniRef clusters, which were then converted to UniProt accessions and then used to extract EC numbers from the UniProt Trembl and Sprot files. As the UniRef clusters are groups containing multiple proteins, the method for converting UniRef clusters to UniProt accessions by removing the underscore character may have added a

bias to the ground truth data. To test this, a new ground truth data was created by annotating the data with UniProt accessions to start with, instead of UniRef cluster IDs (further information about this process can be found in Section 4.3.2.5). This new UniProt-annotated ground truth was then compared with the Carnelian classification results, as shown in Fig. 4.6. The results were very similar, implying that the original methodology of converting UniRef cluster IDs to UniProt accessions produced a very similar ground truth to when the data was annotated without the conversion step. As EC numbers often contain multiple proteins related to a particular function or enzyme activity, it may be the case that UniRef clusters and EC groups often share the same proteins.

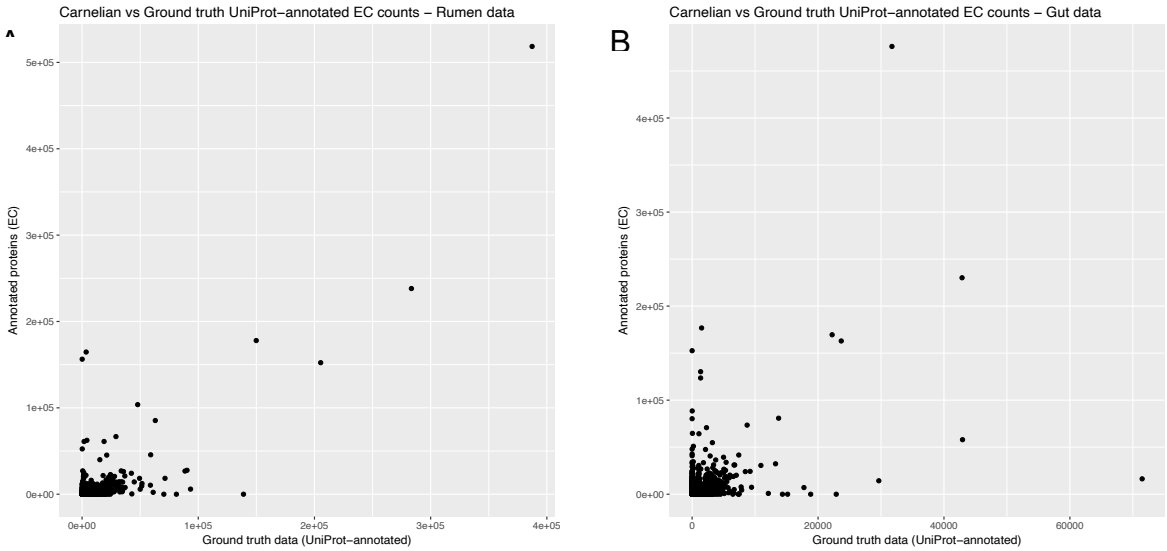


Figure 4.6: Comparison of UniProt-annotated ground truth and Carnelian EC annotations for the rumen data. As with Figure 4.5, the erroneously annotated EC 2.4.1.333 (n=4196663) was removed from the data.

4.4.3: Overall performance of both pipelines

4.4.3.1: Direct comparison of Carnelian and HUMAnN3 raw outputs

As Fios were interested in whether the Carnelian pipeline classified metagenomic data in a similar pattern to HUMAnN3, the raw classification results from the Carnelian pipeline and the regrouped EC number classification results from the HUMAnN3 pipeline, were compared (see Figure 4.7). While this was not a perfect comparison as each pipeline uses a different version of UniRef, it was interesting to see that each pipeline classified proteins with differing abundance.

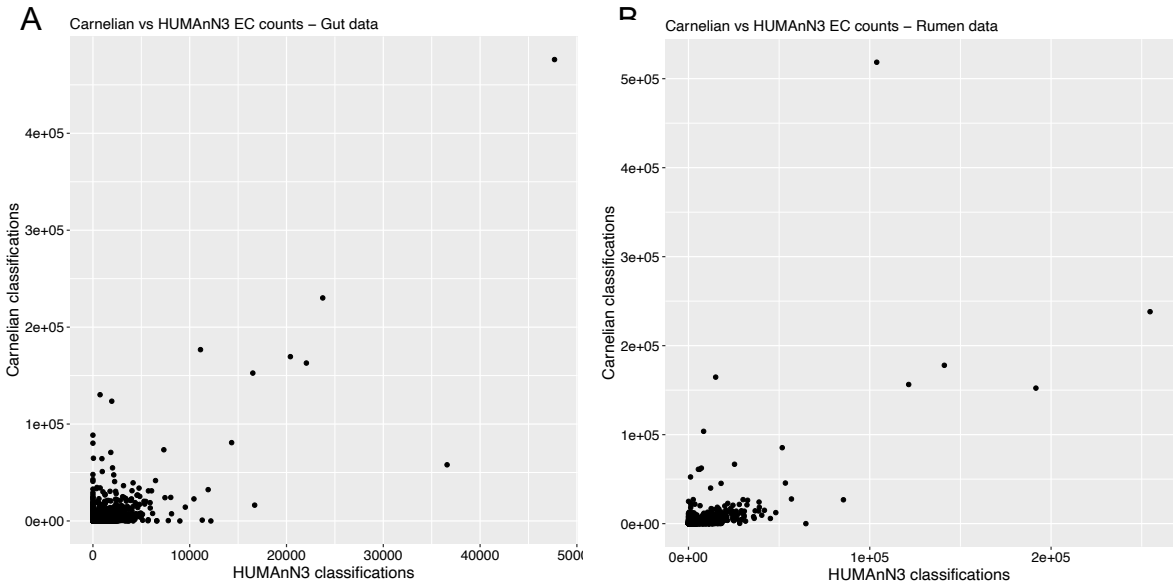


Figure 4.7: A comparison of the two pipelines HUMAnN3 and Carnelian for (A) the human gut metagenome and the (B) rumen metagenome. The HUMAnN3 classifications were regrouped to EC number, and are plotted on the x-axis, and the raw Carnelian EC number counts are plotted on the y-axis. As with Figure 4.5 and 4.6, EC 2.4.1.333 (n=4196663) was removed from the data.

4.4.3.2: Summarising the comparisons between the ground truth predictions and classifications

In order to quantify the difference between the ground truth and classification results for each pipeline, a linear regression was added to the comparative scatterplots. The R^2 value was calculated for each comparison (see Figure 4.8). The R^2 value is the statistical measure of goodness-of-fit of the data points to the linear regression, which quantifies how similar the two abundances compare to one-another. In a situation where a pipeline estimated the abundance of proteins or clusters completely accurately, all data points would perfectly fit the linear regression and the R^2 value would equal 1. For the purposes of this analysis, the R^2 value can be considered an indication of the accuracy of each pipeline.

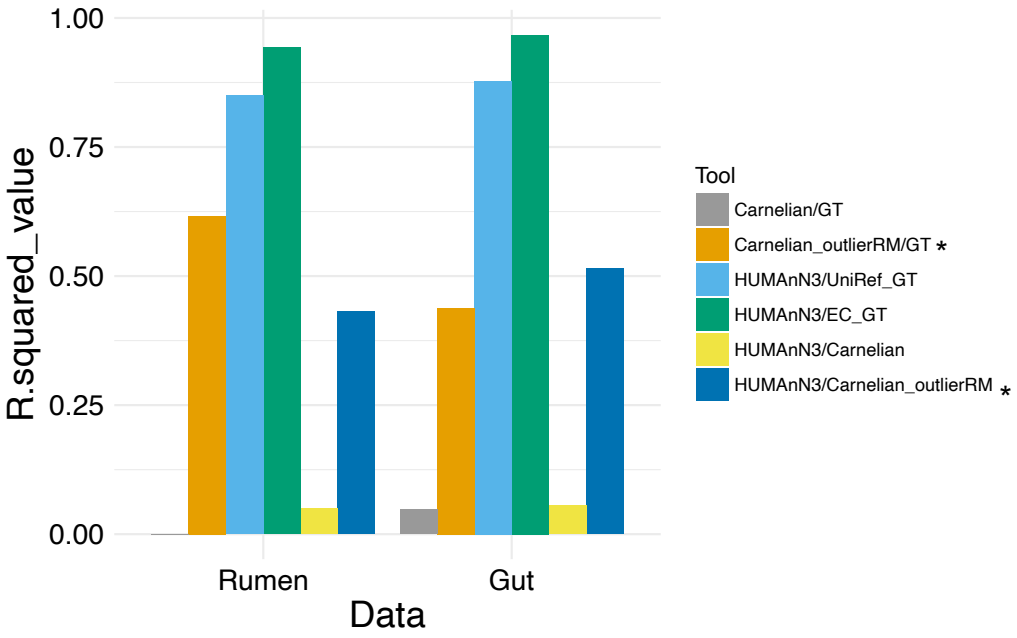


Figure 4.8: Comparing the accuracy of each functional classification pipeline against ground truth predictions using the R^2 metric. The R^2 value (y-axis) for each comparison is represented as a bar. The key shows each comparison, and the * denotes the comparisons with the Carnelian results where the EC 2.4.1.333 has been removed. Carnelian/GT refers to a comparison of the Carnelian classification results and the ground truth, HUMAnN3/UniRef_GT refers to a comparison of the HUMAnN3 classification results and the UniRef-annotated ground truth, HUMAnN3/EC_GT refers to a comparison of the regrouped (EC counts) HUMAnN3 classification results and the EC-annotated ground truth.

The highest R^2 value was observed when comparing the EC-annotated ground truth with the regrouped HUMAnN3 EC classification results. If we consider the R^2 metric as an estimation of accuracy, the HUMAnN3 pipeline classified EC groups with 97% and 94% accuracy for the human gut and rumen data respectively. The HUMAnN3 pipeline also performed well when classifying UniRef cluster abundance, with accuracy of 88% for the human gut data and 85% for the rumen data. The Carnelian pipeline classification results were not as similar to the ground truth predictions as those obtained with HUMAnN3. The EC 2.4.1.333 was erroneously reported as being extremely abundant, and so this was removed from the comparisons. Removing this outlier improved the similarity between the ground truth predictions and the Carnelian pipeline, increasing the accuracy of Carnelian by ~40% for the human gut data and ~60% for the rumen data. Overall, even with the outlier removed, the Carnelian pipeline did not perform as well as HUMAnN3, classifying data with 44% accuracy for the human gut data and 62% accuracy for the rumen data.

4.5: Discussion

4.5.1: Industrial outcomes

This placement with Fios Genomics Ltd was designed to choose and compare functional classification pipelines for metagenomics data. There were several factors to consider when choosing the pipelines, and both HUMAnN3 and Carnelian were chosen as they met almost all of these factors (see Table 4.2). The performance of each pipeline was assessed by comparing each with ground truth annotations.

4.5.2: Performance of both pipelines

4.5.2.1: The HUMAnN3 pipeline

The HUMAnN3 pipeline performed well, classifying most of the data to a similar abundance as the ground truth estimations. If we use the R^2 value to estimate accuracy, the HUMAnN3 pipeline classified the human gut data with 88% accuracy and the rumen data with 85% accuracy. Of course, this assumes that the ground truth estimations are correct, which will be discussed further in Section 4.5.2.2. Most clusters were classified to a similar abundance by HUManN3 as was predicted in the ground truth. However, some clusters were notably abundant in either the HUMAnN3 or ground truth estimations, and not the other. These protein clusters include UniRef90_Q04IK6 and UniRef_K9ID55, which were classified at high abundance in the rumen data by HUMAnN3, but were either absent or at low abundance in the ground truth data. Conversely, clusters UniRef90_Q3JYR6 and UniRef90_A0A0U0K9F7 were predicted at high abundance in the ground truth data, but were classified at either low abundance or were absent from the HUMAnN3 classifications. Interestingly, these protein clusters are functionality similar (Consortium, 2019).

The protein A0A0U0K9F7 was made redundant in UniProtKB on the 11th December 2019, but was named DNA-directed RNA polymerase subunit beta. In UniProtKB, a protein is decided to be redundant if it is identical to another protein found in the same species, in this case *Streptococcus pneumoniae*. A BlastP (Johnson *et al.*, 2008) search on the amino acid sequence of A0A0U0K9F7 revealed a 100% identity match to the UniProtKB/Swiss-Prot accession C1CA06, which has the same name of DNA-directed RNA polymerase subunit beta. As A0A0U0K9F7/C1CA06 and Q04IK6 both have the same name, correspond to the EC number 2.7.7.6, and the gene rpoB, this is strong evidence that they are the same protein. However, this raises the question – why were these proteins not predicted at a similar abundance during the creation of the ground truth data, and when classified by HUMAnN3? The HUMAnN3 pipeline annotates UniRef90 labels from ChocoPhlAn pangenomes, which is based on the contents of the UniProt proteomes portal as of January 2019 (Beghini *et al.*, 2021). Because of this, UniProt release 2019_01 (released on 16/01/2019) was the version of UniProt used to annotate the ground truth data to compare with the HUMAnN3 results. This was done with the intention of using the same version of UniProtKB, which would reduce variation caused by database contents. However, as HUMAnN3 annotates with UniProt proteomes, it may have been that this accession was already redundant in the proteomes portal, or had been replaced by other accessions such as Q04IK6. Accession A0A0U0K9F7 belongs to proteome UP000040760, whereas Q04IK6 belongs to proteome UP000001452. Proteome UP000040760 has since been excluded for reasons stating the genome was too large and was likely contaminated (see *https://www.uniprot.org/proteomes/UP000040760*).

The cluster UniRef90_K9ID55 was not predicted in the ground truth data but was classified over 5000 times by the HUMAnN3 pipeline in the rumen data. K9ID55 is a conserved domain protein, corresponding to the gene PLO_2151, and inferred in the genome of *Pediococcus acidilactici* NGRI

0510Q. The proteome associated with this protein is UP000009880. The UniProtKB entry for this protein states that the sequence was derived from an EMBL/GenBank/DDBJ WGS entry, and the information has been imported from ENA (GAC46679.1) and is therefore preliminary. There is no detail included about the function of this protein. Running the amino acid sequence on BlastP showed that this protein sequence has a higher percentage identity to multiple proteins, most of which are hypothetical, but one is a cell wall-associated hydrolase (accession: CDW60789.1, query cover 79%, percentage identity 96.77%).

Unlike K9ID55, Q3JYR6 is a reviewed protein, and is a leucine--tRNA ligase. This protein is encoded by the gene leuS, corresponds to the EC number 6.1.1.4, and was isolated in *Streptococcus agalactiae* serotype Ia *(*strain ATCC/A909/CDC SS700*).* This catalyses the reaction [ATP + L-leucine + $tRNA^{Leu}$ = AMP + diphosphate + L-leucyl-$tRNA^{Leu}$] and thus would play a vital role in the cell during the translation stage of protein synthesis (Rosario *et al.*, 2004). It is therefore not surprising that it was highly abundant.

As K9ID55 and Q3YR6 were predicted in the rumen data ground truth and classified by HUMAnN3 to a similar abundance, it could have been a similar situation as with A0A0U0K9F7 and Q04IK6 where the two proteins are highly likely to be the same protein. However, as the genes are not the same and a BlastP comparison showed no significant similarity between the two amino acid sequences, it seems unlikely that these proteins are the same. As K9ID55 is only 39 amino acids in length, I would suggest that this domain forms part of a bigger protein that was abundant in the data, which is why HUMAnN3 classified it to a high abundance.

In the gut data, several proteins were classified at high abundance by HUMAnN3 but were not present in the ground truth data. The two proteins

that were classified the most by HUMAnN3 are clusters UniRef90_B0NLV0 and UniRef90_E4MAS8. UniRef90_B0NLV0 is no longer a cluster in UniRef, but protein B0NLV0 is now a member of UniRef90_D6CWH4, which is a group of Site-specific DNA-methyltransferases (adenine-specific). UniProtKB entry B0NLV0 corresponds to the gene BXY_13910 and EC number 2.1.1.72. This protein has DNA binding activity, endonuclease activity, as well as the site-specific DNA-methyltransferase activity, which is likely an essential protein involved in DNA replication and/or transcription (Hornby, 1993).

Interestingly, protein cluster UniRef90_D4IMY6, which was present in high abundance in the ground truth predictions but was not classified by HUMAnN3, is also a Site-specific DNA-methyltransferase (adenine-specific), and protein D4IMY6 corresponds to the gene AL1_1960 and EC number 2.1.1.72. As proteins B0NLV0 and D4IMY6 both correspond to the same EC number, they catalyse the same reaction. It is therefore possible that HUMAnN3 failed to classify the same UniProt ID as the ground truth, despite the same version of UniProtKB being used, as there are differences in the proteomes that were used to annotate ChocoPhlan in bioBakery3.

When comparing the EC-annotated ground truth with the regrouped EC classification by HUMAnN3, there was a strong correlation with both the rumen and human gut data. Using R^2 as an estimation of accuracy, we observed that HUMAnN3 classified the rumen data with EC numbers with 94% accuracy, and the human gut data with 97% accuracy. This comparison between classification with HUMAnN3 (EC numbers) and the ground truth data was therefore the most accurate of all the comparisons. The human gut is a well-studied environment, and as such proteins found in the human gut microbiome are relatively well characterised (Almeida *et al.*, 2019). By contrast, the rumen is not such a well-studied environment (Stewart *et al.*,

2018), and this may be the reason why there was a slight reduction in accuracy between the two datasets.

There was more of a correlation between the EC-annotated ground truth and the classification results, than with the UniRef-annotated data. This was most likely due to the nature of EC groups and UniRef clusters. Although both contained multiple proteins grouped by function, enzymatic activity is a specific molecular function, and is more likely to be confirmed experimentally. It may also be that as most proteins will not be enzymes, the number of enzymes in the data was fewer, and the proteins in databases are most likely curated, there is a smaller chance of mis-labelling data with EC numbers. Additionally, it may be that multiple proteins that could be members of different UniRef clusters, belonged to the same EC number. This was observed with proteins B0NLV0 and D4IMY6, which both corresponded to EC number 2.1.1.72, despite being in different UniRef clusters.

4.5.2.2: The Carnelian pipeline

The Carnelian pipeline classifies data to EC number using a gold standard database that has been annotated with UniProtKB. This used the current version of UniProtKB at the time of analysis, which was 2021_01, and so the ground truth used to compare with Carnelian classifications was also annotated with this version of UniProtKB. During the creation of the ground truth data, UniRef clusters were converted to UniProtKB IDs by taking the latter part of the UniRef ID as a UniProtKB ID (more info in Section 4.3.2.4). The comparison of the ground truth data that was made using this method and the Carnelian pipeline classifications produced very poor R^2 values of $5.78e^{-5}$ for the rumen data, and 0.048 for the human gut data. These values were exacerbated by the EC number 2.4.1.333, which was absent in the ground truth predictions but was classified by Carnelian at very high abundance. In the rumen data, Carnelian classified EC 2.4.1.333 over 40

million times and in the human gut data over 2 million times. This EC number refers to the enzyme 1,2-beta-D-glucan:phosphate alpha-D-glucosyltransferase, which catalyses the reversible phosphorolysis of beta-(1->2)-D-glucans.

The next most abundant protein Carnelian classified were the same in both datasets, EC number 2.7.13.3 which is Histidine protein kinase. This catalyses the reaction [ATP + protein L-histidine <=> ADP + protein N-phospho-L-histidine], and includes multifunctional proteins found in non-animal kingdoms that typically play a role in signal transduction across the cell membrane (Engin, 2021). As EC number 2.4.1.333 only has one UniProtKB protein that corresponds to the enzyme (Q92AT0) and EC number 2.7.13.3 has 184 UniProtKB proteins that correspond to this enzyme, it seems unlikely that Carnelian was reporting an accurate abundance. Indeed, other enzymes that were predicted by the ground truth data and classified by the Carnelian pipeline in high abundance are essential to the cell, so even if EC 2.4.1.333 did appear in high abundance, it most likely would not be more abundant than essential housekeeping proteins such as DNA helicase. It may be, therefore, that there was an issue with the database entry for this enzyme, or that it contained common or repeating sequences that would be found on many reads. Removing this enzyme from the classification results improved the R^2 value to 0.62 for the rumen data and 0.44 for the human gut data.

However, as this version of the ground truth was made by converting UniRef cluster IDs to UniProtKB IDs, it was also possible that this resulted in converting a group of proteins into one single protein. To test whether or not this method added a bias to the ground truth, a different ground truth was made without the conversion step by annotating the data with UniProtKB IDs. Removing this conversion step made little difference to the overall results, so it is does not appear that it added a bias to the results. This may be because

Carnelian classifies to EC number, and by doing so will classify multiple proteins from multiple UniRef clusters to the same protein if appropriate. This may remove any intrinsic bias as any change at the protein level might still be classified as the same enzyme, and would therefore not be impacted by the different methods used to create the ground truth data.

4.5.2.3: Comparing the HUMAnN3 and Carnelian pipelines

Some of the classifications by the HUMAnN3 and Carnelian pipelines were the same EC number for each dataset, for example 2.7.13.3, 3.6.4.12, and 2.7.7.6. This is to be expected as both the human gut and the rumen contain genomes derived from microbes that are adapted to the mammalian gastrointestinal tract and there will therefore be conserved enzymes in both datasets (Ben David *et al.*, 2015). As both pipelines classified these enzymes, it increases confidence that the classifications were correct. However, the EC number 5.99.1.3 was not predicted in the ground truth data but was classified by the Carnelian pipeline. This EC number was predicted in the ground truth data that was made for comparison with HUMAnN3. Additionally, this EC number was classified by the HUMAnN3 pipeline itself. EC 5.99.1.3 has been transferred to 5.6.2.2, which is why it was not present in the ground truth as this was annotated with the current (2021_01) version of UniProtKB. It is surprising that this EC number was classified by the Carnelian pipeline, as this prediction was made using the same version of UniProtKB.

Overall, the classification results and the accuracy of the results differed significantly between the pipelines. HUMAnN3 classified data with good accuracy, and as it classified both datasets to the UniRef cluster and EC number resolutions, in my opinion it is a good choice for classifying function of metagenomic data. While Carnelian was able to efficiently classify the enzyme-encoding genes in both datasets, it does appear to classify data with less accuracy than HUMAnN3. Based on these results, I would personally choose HUMAnN3 to classify function, and this was my recommendation to Fios Genomics Ltd. As this was a preliminary study, additional work would need to be done to measure the accuracy of these tools before any firm conclusions were made. However, this work does highlight the importance of

choosing the most appropriate bioinformatics tool for the analysis, and the importance of testing the tools you are using with simulated data. In a laboratory, a control sample might be used to ensure the method is working as expected, and it should be the same in computational biology.

4.6: Conclusions

In general, the HUMAnN3 pipeline classified data that was more similar to the ground truth predictions than Carnelian. As this is a preliminary study, further analysis would be needed to determine the accuracy of each pipeline robustly. As the ground truth predictions were more similar to the methods by which HUMAnN3 classify data, this analysis does have some bias. However, it was not possible to replicate the methods by which Carnelian classifies data, and therefore the ground truth for comparison with Carnelian was made using varying methods. Fios Genomics Ltd viewed this project as a success, and they were satisfied with the project results. The company asked for a report detailing the analysis and results which was circulated within the bioinformatics and data analysis teams. As this preliminary data shows HUMAnN3 performed well at classifying both the human gut and rumen data, as well as being suitable for EC annotation, Fios intend to include HUMAnN3 as a functional annotation pipeline in the services they offer to clients for analysing metagenomic data.

Chapter 5: Discussion

5.1: General discussion

The rumen microbiota plays essential roles in the breakdown and fermentation of dietary fibres consumed by the host animal, and synthesises short-chain fatty acids, that ultimately contribute to ruminant production (Auffret *et al.*, 2020). Improving our understanding of the roles that the rumen microbiome plays is important for animal health and nutrition, for ensuring food security, and for mitigating issues such as methane emissions. Despite laudable previous efforts, only a subset of the rumen microbiota has been cultured to date. This leaves most ruminal microbes uncultured, and therefore uncharacterised (Creevey *et al.*, 2014). Without prior cultivation in the laboratory, these species do not have high-quality isolate-derived reference genomes, and will likely be absent from reference databases. This is of critical importance, as rumen microbiome research is now driven in large part by high-throughput methodologies such as metagenomic sequencing, which provide the opportunity to study the genomes of uncultured species. As such, common techniques for analysing the microbial community of the rumen include taxonomic and functional classification of metagenomic data. However, work prior to this thesis demonstrated that rumen metagenomic data often showed poor classification rates when using publicly available reference databases, likely as a result of underpopulated reference databases (Stewart *et al.*, 2018). Therefore, the reliability and accuracy of classification results should be questioned.

The second chapter of this thesis aimed to investigate the impact of reference database choice on the taxonomic classification results of rumen metagenomic data. Several reference databases were built, comprising different contents, including custom reference databases that contained rumen culture-derived genomes and MAGs, as well as standard reference

databases that contained reference genomes from RefSeq. A ground truth simulated metagenomic dataset was created from culture-derived ruminal genomes with known taxonomy, and classified with Kraken2 using the various reference databases. The impact of reference database choice on classification rate was assessed, and the accuracy of the classifications was measured to the read level, for all taxonomic levels.

This investigation found that the standard Kraken2 reference database (which included the complete reference genomes of the bacterial, archaeal, and viral domains in RefSeq) proved a poor choice for classifying the taxonomy of rumen metagenomic data, both in terms of classification rate and accuracy. Of particular concern was the false reporting of thousands of species that were absent from the data. RefSeq is a widely used resource that continues to grow in size, and is the standard reference database when classifying taxonomy with the popular tools Kraken or Kraken2 (Segerman, 2020). The implications of these findings are significant, as they suggest that the many other studies using this methodology, may unknowingly have largely inaccurate classification results. This likely applies not just to the rumen, but to any environment that is poorly characterised and has a lack of relevant reference genomes in RefSeq. Metagenomics is increasingly applied to various types of environmental samples, and it is clear that results will continue to be limited in their accuracy until reference databases can be improved.

In the case of the rumen, the value of having well-characterised reference genomes was beautifully demonstrated in my project. Indeed, the use of custom reference databases containing culture-derived ruminal microbes from the Hungate1000 collection improved the classification rate and accuracy to almost 100% at the phylum and family levels. To some extent, this was unsurprising as the data was simulated from the same genomes that were used to build the reference databases containing culture-derived

genomes. This is a limitation of the study and, ideally, it would have been good to test different input genomes to truly assess whether adding well-characterised, culture-derived genomes from the same environment would improve classification results. In the future, as more well-characterised ruminal reference genomes become available, it would be interesting to compare how well the Hungate1000 collection-containing reference databases improved their classification rate and accuracy versus the standard RefSeq choice. Based on the results presented in this thesis, it is reasonable to anticipate that classification rate and accuracy would indeed be improved, and that continually adding additional reference genomes would improve the situation even further.

Interestingly, in contrast to the results at the phylum and family levels, when culture-derived genomes from the Hungate1000 collection were included in reference databases, there was a noticeable reduction in classification rate at the species level, with over 7% of reads unclassified. This was unexpected as the data was fully represented in the reference databases that contained the Hungate1000 collection. A similar phenomenon was observed by Nasko *et al.* when classifying taxonomy using different versions of RefSeq (Nasko *et al.*, 2018). Nasko *et al.* hypothesised that as the number of closely related species increases with each version of RefSeq, the ability to create unique minimisers (l-mers, minimised k-mers) reduces. This resulted in what I have called minimiser collisions, as the minimisers can no longer be used to distinguish between different very closely related species in the reference database. As Kraken2 classifies to the lowest common ancestor (LCA), this results in the read being classified to the genus level, instead of the species level. Therefore, there is a reduction in classification at the species level when using the later versions of RefSeq, where there is a greater number of species. I believe the same phenomenon was happening when the data was classified using reference genomes that contained the ruminal culture-derived genomes from the Hungate1000 collection. To further explore this in

the future, it would be interesting to measure the sequence similarity of the genomes and build reference databases to contain genomes with differing amounts of similarity. Additionally, it would be interesting to see whether there was a threshold that results in the greatest number of accurate classifications to the species level, but with the fewest minimiser collisions. This information could be used to inform other studies that classify taxonomy using k-mers and improve classification at the species level specifically. In addition, this work highlights the importance of evaluating reference database choice when classifying the taxonomy of metagenomic data. Other approaches may classify rumen metagenomic data with more accuracy. For example, GTDB-Tk uses sequence-based phylogeny to classify taxonomy to bacterial and archaeal genomes (Chaumeil *et al.*, 2019). As GTDB-Tk does not require a representative reference genome in RefSeq like Kraken2, it may classify the taxonomy of rumen metagenomic data with more accuracy than Kraken2. It would be interesting to create multiple ground truth datasets and quantify the accuracy of classification using GTDB for not only the rumen environment, but for others that lack culture-derived representation in RefSeq (Choi *et al.*, 2017).

The need for accurate classification of metagenomics data has been a common thread running through this thesis. In chapter 2, the accuracy of taxonomic classification was limited due to informal and/or incomplete taxonomy labels of genomes in the reference databases. Although this was true for a minority of culture-derived ruminal genomes (from the Hungate1000 collection), it was particularly impactful for rumen MAGs (from the RUG2 superset in Stewart *et al.* (Stewart *et al.*, 2019)). Despite informal labels and/or incomplete taxonomic lineages, including rumen MAGs in reference databases improved classification rate and reduced the misclassification of reads at the family and genus levels. However, including rumen MAGs in reference databases did not improve classification accuracy, most likely due to issues with their taxonomy. I hypothesise that rumen

MAGs with complete and accurate formal taxonomic labels would improve the classification accuracy substantially. If this work was continued, it would be interesting to make a reference database to test this, and classify not only simulated reads from isolates as in Chapter 2, but also simulated reads from isolates that had not yet been added to RefSeq to test the impact of adding MAGs to a reference database on classifying (for all intents and purposes) novel species.

Traditionally, the giving of taxonomic names to novel prokaryotic species is limited to cultured isolates, with prokaryotic nomenclature regulated by the ICNP (International Code of Nomenclature of Prokaryotes) (Parker *et al.*, 2019). However, the growing number of uncultured genomes, including MAGs, require formal and consistent names for the same reasons that cultured isolates do. Specifically, this would bring order to the field, and avoid filling reference databases with duplicate genomes with different names and/or incomplete taxonomic lineages, which, as my results in Chapter 2 showed, severely impacts the accuracy of classification steps *(Hugenholtz et al.*, 2021). Arguably, the need for more systematic characterisation of MAGs is even more pressing than for cultured isolates because approximately 75% of microbial phylogenetic diversity is exclusively represented by genomes derived from uncultured species (Nayfach *et al.*, 2021). The current *status quo* of only providing names to the cultured microbes severely limits the organisation of the majority of the tree of life (Murray *et al.*, 2020). There have been recent attempts to bring a more formal naming/taxonomic structure to non-cultured-derived genomes. For example, SeqCode is a framework for registering prokaryotic genome sequences (Hedlund *et al.*, 2022). The aim of SeqCode is to provide nomenclature to all prokaryotic-derived genome sequences, including MAGs and *Candidatus* names. By registering a genome in SeqCode, the sequence of the genome and nomenclature of the species can be shared, which may reduce repetitions in the reference database. Given that my work has demonstrated the critical

importance of MAGs having consistent taxonomic names if they are to be used for reference database purposes, it will be interesting to see if microbiome researchers adopt this approach. If so, this is likely to bring additional consistency and accuracy to taxonomic annotation of metagenome sequence data.

Having established in Chapter 2 that MAGs have the potential to improve classification accuracy if given formal taxonomic labels, Chapter 3 further explored the concept of using MAGs as reference genomes, aiming to assess the representativeness and accuracy of MAGs as reference genomes for uncultivated taxa. Culture-derived genomes and MAGs thought to belong to the same strain (having clustered at >99% ANI) were compared to evaluate any differences, and potential limitations, of using a MAG to characterise the taxonomy and function of a microbe compared to a culture-derived genome. Overall, the extent of how similar, or different, the MAGs and culture-derived genomes were to one another varied and seemed to be linked to the taxa of that species. This work observed that for genomes that had originated from a species with a relatively open pangenome, the functional predictions varied between the culture-derived genome and the MAG considerably. In my opinion this is somewhat unsurprising as a large accessory genome would translate into a higher level of variation between genomes derived from the same bacterial species (Rouli *et al.*, 2015). In support of this assertion, Meziti *et al.* observed that highly complete *Escherichia coli* (>95% completeness) MAGs on average captured 77% of core genes and 50% of variable genes of an *E. coli* population isolated from stool samples (Meziti *et al.*, 2021). *E. coli* is known to have an open pangenome and a large accessory genome (Decano and Downing, 2019; Rasko *et al.*, 2008), so it would be interesting to see a comparison at the gene level with a species that is known to have a relatively closed pangenome. In this study, a culture-derived genome and a MAG that were both assigned the species *Bacillus licheniformis* (see Section 3.3.3) had

identical predicted function. Although this shows promise that, for some species, there may not be any disadvantages to using MAGs *in lieu* of a culture-derived genome, it is important to further test this on a variety of species with differing sized accessory genomes. If I were to continue this investigation I would predict function more extensively, for example annotating genes and then comparing gene prediction profiles between MAGs and culture-derived genomes, to gain a more detailed view of any information that is absent in MAGs.

One of the limitations of the work described in Chapter 3 was that there was no direct matches of cultured-isolate derived genomes with MAGs that were assembled from the same original rumen samples that the cultures were isolated from. This could be explained by the methodological limitations of the work. For instance, the rumen samples had been frozen for several years which likely would have killed some microbes. Additionally, the isolates were cultured at 37˚C whereas the rumen has an average temperature of 39˚C. The cultures were enriched for 24 hours before being cultured for 48 hours, a longer enrichment and incubation time would have allowed for a greater diversity of microbes to grow. Furthermore, increasing the number of subculturing steps would have increased the likelihood of isolating a pure colony. There are also limitations to the computational methodologies, for instance the possible absence of cultured species in the metagenome data may have been due to low abundance microbes in the sample not assembling well (Ayling *et al.*, 2020). Regardless, this meant it was not possible with 100% certainty to disentangle strain-level variations in pan-genome content from possible methodological artefacts when comparing the two types of genomes. One way to address this problem in future experiments might be to take a cultured microbe with a high-quality reference genome and spike it into a rumen sample at high dosage, and then generate a metagenomic sequence library from the spiked sample. Additionally, mock communities could provide suitable known truth metagenomic data. These

options would ensure that the identical strain of interest would definitely be present and assembled in the metagenomic dataset. This would mean it would be possible to do a direct comparison between the traditionally-assembled genome from the cultured isolate and the metagenome-assembled genome from the same strain, and make empirical observations about how directly comparable each type of genome was to each other. In this way the quality and accuracy of MAGs could be definitively assessed.

Currently, the quality of MAGs is usually assessed using marker genes, for example CheckM (Parks *et al.*, 2015). Although, Meziti *et al.* suggest that the quality of MAGs may be over-estimated, due to the gene-level variability that can occur between species not being considered during MAG assembly (Meziti *et al.*, 2021). To ensure the criteria for what is considered 'high-quality' for a MAG is consistent, the Genomic Standards Consortium (GSC) has therefore developed two standards for genomes derived from uncultured species: the Minimum Information about a Single Amplified Genome (MISAG) and the Minimum Information about a Metagenome-Assembled Genome (MIMAG) (Bowers *et al.*, 2017). These standards ensure consistency with language, as previously MAGs could be referred to as "high quality" when they were only 80% complete. With these new standards, to be considered high-quality a MAG has to have at least 90% completeness and less than 5% contamination. Additionally, rRNA genes and tRNA genes should also be present.

Advancements in methodologies and technologies have the potential to improve MAG assembly. Long-read sequencing can be advantageous for metagenomics, capturing longer sections of the genomes in an environmental sample. Compared to short-read sequencing, which may need several contigs or scaffolds to cover a region of a genome, long-read sequencing can cover longer sections of a genome in a single read. Consequently, there are less opportunities for misassembled contigs,

resulting in MAGs assembled from long-read data generally being higher-quality than those assembled from short-read data (L. Liu et al., 2022; Tao et al., 2022). This is especially relevant in microbial genomes as horizontal gene transfer events, which can cause issues with mis-assembly in short-read data, can be identified with long-read metagenomic data (Douglas and Langille, 2019). In addition, 16S rRNA gene sequences, are a challenge to derive from short-read metagenomic data (Yuan et al., 2015; Hiseni et al., 2022). However, as previously discussed they are required for a "high-quality" MAG according to the guidelines set forth by MIMAG. Long-read sequencing data can capture single-contig bacterial genomes (Cuscó et al., 2021). Not only do these single-contig genomes contain rRNA gene sequences, but they provide a complete genome of an uncultured species (Singleton *et al.*, 2021).

MAG quality is not only impacted by sequencing methods, but the bioinformatic methods that combine that data into draft genomes. For example, SemiBin is a bioinformatics tool that uses a siamese neural network that utilises contig information derived from GTDB to improve metagenomic binning and MAG quality (Pan et al., 2022). The combination of long- and short-read data during assembly, known as hybrid assembly, has its advantages (Bertrand et al., 2019). The long-reads capture any repetitive or conserved sequences which improves the quality of the assembly, while the short-reads increase the depth of coverage which can increase the confidence of structural variant calling (Chen et al., 2022). Assembling reads from one metagenomic sample (single-sample assembly) or multiple metagenomic samples (co-assembly) can impact MAG quality and biodiversity (Delgado and Andersson, 2022). Bickhart *et al*. demonstrated that a hybrid approach of long-read sequencing and binned reads from Hi-C single-cell sequencing produced very high-quality MAGs (Bickhart et al., 2022).

Single-cell sequencing provides an excellent opportunity to assemble reads that are confidently from the same genome (Kogawa *et al.*, 2018). Single-amplified genomes (SAGs) have revealed detailed insights into the taxonomy and function of uncultured microbial genomes (Chijiiwa et al., 2020). As SAGs are assembled from clonal cells of the same microbe, there is no possibility of strain-level variation causing misassembly (Singleton *et al.*, 2021). Alneberg *et al.* recently compared SAGs and MAGs of very high sequence similarity (>89% ANI), and found that the MAGs were missing ~3.6% of the sequence compared with the SAG, demonstrating the clear potential for improvements if uncultured genomes are assembled with a SAG approach (Alneberg *et al.*, 2018). While this is promising, in my opinion, further exploration of MAGs and SAGs together with culture-derived genomes is needed before they can be reliably used as reference genomes. It is important to note though that, even if the quality of SAGs is higher than that of MAGs, they still would not improve the accuracy of classification after adding to reference databases without first assigning full and accurate taxonomy to them, as was demonstrated here in Chapter 2.

High-quality uncultured genomes have implications far beyond the rumen. MAGs and SAGs have shed light on the taxonomy and functionality of novel uncultured species in freshwater environments(Shaffer et al., 2020), seawater (Ranran et al., 2023), and the human gut (Hosokawa et al., 2022). Under-studied environments such as deep ocean hydrothermal plumes are extremely harsh environments that contain many uncultured species that have greatly benefited from information provided by MAGs (Molari et al., 2023; Zhichao Zhou et al., 2020) where in some cases almost 70% of draft genomes are unknown at the Phylum level (Speth et al., 2022). MAGs are being used to detangle genes of previously unknown function at an unprecedented rate. Vanni *et al.* recently identified over 283,000 genes of unknown function belonging specifically to the lineage of a novel Phyla "*Candidata Patescibacteria*" from metagenomic data and MAGs

(https://elifesciences.org/articles/67667). Hypersaline mats are incredibly diverse ecosystems, and despite being thought to contain information about some of the earliest microbiomes on the planet, remain highly understudied. MAGs have allowed for the first ever exploration of the so-called “microbial dark matter” metabolism in hypersaline mats (Wong et al., 2020). By analysing the data of uncultured genomes, more information about the taxonomy and function of host and non-host microbiomes may be brought to light.

While much of my thesis explored the impact of methodological approaches and reference databases on the accuracy of taxonomic classifications, one of the other key features of metagenomics is that it allows functional predictions to be made as well. Indeed, one of the findings in Chapter 3 was that functional predictions did sometimes differ between culture-derived genomes and MAGs. It is clear, therefore, that chosen methodological approaches will also be as important for functional classifications as they are for taxonomic ones.

One of the problems for anyone entering the field of metagenomics research is that there is a continually expanding and vast array of software available for the analysis of data, involving constituent steps in the process from assembly to classification, and it is not always clear which tool or pipeline is best for the chosen analysis (Prakash and Taylor, 2012; Segata *et al.*, 2013). Accuracy is clearly a key goal, but it is important to acknowledge the importance of recognising the differences between bioinformatics tools and pipelines, and that results may vary depending on what is used. In Chapter 4, I therefore compared two functional classification pipelines on behalf of the company Fios Genomics Ltd. The aims of this work were to assess the suitability of the pipelines for including in the company’s business infrastructure and estimate the accuracy of the two chosen pipelines HUMAnN3 and Carnelian. Classifying function can be challenging from a

business perspective as some of the commonly used computational analysis tools for metagenomics data, such as KEGG, require an expensive license for commercial use. The HUMAnN3 and Carnelian pipelines were therefore chosen because they best met the needs of the company. More specifically, and for example, it was essential that any pipelines to be integrated into their services were available to run within a Docker container for reproducibility and version control.

In my experiments while on industrial placement, I predicted the function of simulated metagenomic data from the bovine rumen and human gut to create ground truths, which were then compared with the classification results of each pipeline. The results suggested that HUMAnN3 was more accurate than Carnelian for classifying function, and predicted function at a similar abundance to the ground truth both at protein (UniProtKB) and enzyme (EC) resolution. Fios Genomics Ltd were very pleased with the outcome of the project and have since used my results to inform their business practices moving forward. A challenge of this work was choosing a method to annotate the ground truth data. The ground truth data was annotated with UniProt/UniRef90, which was also used by HUMAnN3 to classify the data. As HUMAnN3 classifies the data with the same database that was used to annotate the ground truth data, the labels are far more likely to be the same. Thus, despite HUMAnN3 classifying the data with high accuracy, the method of ground truth annotation potentially adds a bias that needs to be acknowledged as a limitation of the study. If I were to continue this work in an academic setting, I would compare more pipelines, for example eggNOG-mapper v2 (Cantalapiedra *et al.*, 2021). In order to assess the accuracy of eggNOG-mapper2, I would create a ground truth dataset annotated with eggNOG 5.0 (Huerta-Cepas *et al.*, 2019), with the aim of ensuring that the functional annotations were consistent and comparable between the ground truth estimations and the classifications. Importantly, many pipelines other than the ones I looked at in Chapter 4 use the KEGG database, which is

widely used in metagenomics research (the primary manuscript for the KEGG database has been cited over 25,000 times) (Kanehisa and Goto, 2000). However, during this placement KEGG could not be incorporated into the analysis due to the costs of licensing for industrial use.

5.2: Conclusion

In conclusion, this thesis has highlighted that the use of standard reference databases can result in highly inaccurate taxonomic classification of metagenomics data. The use of MAGs as reference genomes would likely improve classification accuracy as it would vastly increase representation of genomes derived from uncultured species. However, issues such as incomplete and informal taxonomic labels need to be addressed before MAGs could be used effectively as reference genomes. Additionally, this thesis evaluated genomic and functional differences between MAGs and culture-derived genomes, and these preliminary findings suggest that in some cases MAGs could provide reasonably accurate information in lieu of a culture-derived genome. Lastly, this thesis measured the accuracy of two functional classification pipelines for an industrial application of metagenomics and based on these results, the company are aiming to include the most appropriate pipeline in their bioinformatics services moving forwards. Overall, the findings in this thesis emphasise the importance of continued efforts to improve reference databases, and metagenomics analysis pipelines, but suggest that such activities have the potential to significantly enhance our ability to expand our knowledge of the rumen microbiome.

Chapter 6: Bibliography

Abdill, R. J. et al. (2022) Public human microbiome data are dominated by highly developed countries. *PLOS Biology*. 20 (2), e3001536.

Abubucker, S. et al. (2012) Metabolic reconstruction for metagenomic data and Its application to the human microbiome. *PLOS Computational Biology*. 8 (6), e1002358.

Akin, D. E. & Borneman, W. S. (1990) Role of rumen fungi in fiber degradation. *Journal of Dairy Science*. 73 (10), 3023–3032.

Akin, D. E. & Rigsby, L. L. (1987) Mixed fungal populations and lignocellulosic tissue degradation in the bovine rumen. *Applied and Environmental Microbiology*. 53 (9), 1987–1995.

Albertsen, M. et al. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*. 31 (6), 533–538.

de Albuquerque, N. R. M. & Haag, K. L. (2022) Using average nucleotide identity (ANI) to evaluate microsporidia species boundaries based on their genetic relatedness. *Journal of Eukaryotic Microbiology*. n/a (n/a), e12944.

Alexandratos, N. & Bruinsma, J. (2012) World agriculture towards 2030/2050: The 2012 revision. An FAO perspective. World Agriculture

Allison, M. J. & Bryant, M. P. (1963) Biosynthesis of branched-chain amino acids from branched-chain fatty acids by rumen bacteria. *Archives of Biochemistry and Biophysics*. 101 (2), 269–277.

Almeida, A. et al. (2019) A new genomic blueprint of the human gut microbiota. *Nature*. 568 (7753), 499–504.

Almeida, A. et al. (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*. 39 (1), 105–114.

Alneberg, J. et al. (2018) Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome*. 6 (1), 173.

Altenhoff, A. M. et al. (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLOS Computational Biology*. 8 (5), e1002514.

Altschul, S. F. et al. (1990) Basic local alignment search tool. *Journal of Molecular Biology*. 215 (3), 403–410.

Anderson, C. L. et al. (2017) Dietary energy drives the dynamic response of bovine rumen viral communities. *Microbiome*. 5 (1), 155.

Anil Kumar, P. et al. (2015) *Rumen microbiology: from evolution to revolution*. 1st edition. Springer New Delhi.

Apweiler, R. et al. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*. 32 (suppl_1), D115–D119.

Ashburner, M. et al. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*. 25 (1), 25–29.

Auffret, M. D. et al. (2020) Identification of microbial genetic capacities and potential mechanisms within the rumen microbiome explaining differences in beef cattle feed efficiency. *Frontiers in Microbiology*. 11 (June), 1–16.

Ayling, M. et al. (2020) New approaches for metagenome assembly with short reads. *Briefings in Bioinformatics*. 21 (2), 584–594.

Bach, A. et al. (2005) Nitrogen metabolism in the rumen. *Journal of Dairy Science*. 889–21.

Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Research*. 28 (1), 304–305.

Balch, C. C. (1950) Factors affecting the utilization of food by dairy cows: the rate of passage of food through the digestive tract. *British Journal of Nutrition*. 4 (4), 361–388.

Bankevich, A. et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19 (5) p.455–477.

Bauchop, T. (1979) Rumen anaerobic fungi of cattle and sheep. *Applied and Environmental Microbiology*. 38 (1), 148–158.

Bauchop, T. (1981) The anaerobic fungi in rumen fibre digestion. *Agriculture and Environment*. 6 (2), 339–348.

Bauchop, T. & Mountfort, D. O. (1981) Cellulose fermentation by a rumen anaerobic fungus in both the absence and the presence of rumen methanogens. *Applied and Environmental Microbiology*. 42 (6), 1103–1110.

Becker, E. R. & Hsiung, T. S. (1929) The method by which ruminants acquire their fauna of infusoria, and remarks concerning experiments on the host-specifitiy of these protozoa. *Proceedings of the National Academy of Sciences*. 15 (8), 684–690.

Beghini, F. et al. (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3 Peter Turnbaugh et al. (eds.). *eLife*. 10e65088.

Bengtsson-Palme, J. et al. (2016) Strategies to improve usability and preserve accuracy in biological sequence databases. *Proteomics*. 16 (18), 2454–2460.

Benson, D. A. et al. (2013) GenBank. *Nucleic Acids Research*. 41 (D1), D36–D42.

Berg, G. et al. (2020) Microbiome definition re-visited: old concepts and new challenges. *Microbiome*. 8 (1), 103.

Bertrand, D. et al. (2019) Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature Biotechnology*. 37 (8), 937–944.

Bickhart, D. M. et al. (2022) Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nature Biotechnology*. 40 (5), 711–719.

Bolger, A. M. et al. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 30 (15), 2114–2120.

Boone, D. R. et al. (1993) 'Diversity and taxonomy of methanogens', in James G Ferry (ed.) *Methanogenesis*. Boston, MA: Springer US. pp. 35–80.

Bouskra, D. et al. (2008) Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis. *Nature*. 456 (7221), 507–510.

Bowers, R. M. et al. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*. 35 (8), 725–731.

Brister, J. R. et al. (2015) NCBI viral genomes resource. *Nucleic Acids Research*. 43 (D1), D571–D577.

Brüssow, H. et al. (2004) Phages and the Evolution of Bacterial Pathogens : from Genomic Rearrangements to Lysogenic Conversion Phages and the Evolution of Bacterial Pathogens : from Genomic Rearrangements to

Lysogenic Conversion. *Microbiology and molecular biology reviews*. 68 (3), 560–602.

Bryant, M. P. (1959) Bacterial species of the rumen. *Bacteriological Reviews*. 23 (3), 125–153.

Bryant, M. P. & Small, N. (1960) Observations on the ruminal microorganisms of isolated and inoculated calves. *Journal of Dairy Science*. 43 (5), 654–667.

Buchfink, B. et al. (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods*. 12 (1), 59–60.

Burge, C. & Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*. 268 (1), 78–94.

Cabanettes, F. & Klopp, C. (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way Thomas Tullius (ed.). *PeerJ*. 6e4958.

Campbell, J. H. et al. (2013) UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proceedings of the National Academy of Sciences*. 110 (14), 5540–5545.

Cantalapiedra, C. P. et al. (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*. 38 (12), 5825–5829.

Cantarel, B. L. et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Research*. 37 (suppl_1), D233–D238.

Carola, S. & Rolf, D. (2011) Metagenomic analyses: past and future trends. *Applied and Environmental Microbiology*. 77 (4), 1153–1161.

Cash, H. L. et al. (2006) Symbiotic bacteria direct expression of an intestinal

bactericidal lectin. *Science*. 313 (5790), 1126–1130.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*. 11 (4), 265–270.

Chaumeil, P.-A. et al. (2019) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*. 36 (November 2019), 1925–1927.

Chen, C. et al. (2017) 'Protein bioinformatics databases and resources', in Cathy H Wu et al. (eds.) *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*. New York, NY: Springer New York. pp. 3–39.

Chen, L. et al. (2022) Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nature Communications*. 13 (1), 3175.

Chen, S. et al. (2010) Saccharofermentans acetigenes gen. nov., sp. nov., an anaerobic bacterium isolated from sludge treating brewery wastewater. *International Journal of Systematic and Evolutionary Microbiology*. 60 (12), 2735–2738.

Chijiiwa, R. et al. (2020) Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. *Microbiome*. 8 (1), 5.

Choi, J. et al. (2017) Strategies to improve reference databases for soil microbiomes. *The ISME Journal*. 11 (4), 829–834.

Clarridge, J. E. & Alerts, C. (2004) Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* 17 (4), 840–862.

Clausen, P. T. L. C. et al. (2018) Rapid and precise alignment of raw reads

against redundant databases with KMA. *BMC Bioinformatics*. 19 (1), 307.

Collins, S. M. & Bercik, P. (2009) The relationship between Intestinal microbiota and the central nervous system in normal gastrointestinal function and disease. *Gastroenterology*. 136 (6), 2003–2014.

Conrad, R. (1999) Contribution of hydrogen to methane production and control of hydrogen concentrations in methanogenic soils and sediments. *FEMS Microbiology Ecology*. 28 (3), 193–202.

Consortium, T. U. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*. 47 (D1), D506–D515.

Creevey, C. J. et al. (2014) Determining the culturability of the rumen bacterial microbiome. *Microbial Biotechnology*. 7 (5), 467–479.

Cuscó, A. et al. (2021) Long-read metagenomics retrieves complete single-contig bacterial genomes from canine feces. *BMC Genomics*. 22 (1), 330.

Darling, A. C. E. et al. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*. 14 (7), 1394–1403.

Ben David, Y. et al. (2015) Ruminococcal cellulosome systems from rumen to human. *Environmental Microbiology*. 17 (9), 3407–3426.

Decano, A. G. & Downing, T. (2019) An Escherichia coli ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Scientific Reports*. 9 (1), 17394.

Decroos, K. et al. (2005) Isolation and characterisation of an equol-producing mixed microbial culture from a human faecal sample and its activity under gastrointestinal conditions. *Archives of Microbiology*. 183 (1), 45–

55.

Dehority, B. (1993) ‘The rumen protozoa’, in *Parasitic protozoa*. Second pp. 1–35.

Delgado, L. F. & Andersson, A. F. (2022) Evaluating metagenomic assembly approaches for biome-specific gene catalogues. *Microbiome*. 10 (1), 72.

Dindhoria, K. et al. (2022) Bacillus licheniformis MCC 2514 genome sequencing and functional annotation for providing genetic evidence for probiotic gut adhesion properties and its applicability as a bio-preservative agent. *Gene*. 840146744.

Douglas, G. M. & Langille, M. G. I. (2019) Current and Promising Approaches to Identify Horizontal Gene Transfer Events in Metagenomes. *Genome Biology and Evolution*. 11 (10), 2750–2766.

Dunfield, P. F. et al. (2012) Electing a candidate: a speculative history of the bacterial phylum OP10. *Environmental Microbiology*. 14 (12), 3069–3080.

Eckard, R. J. et al. (2010) Options for the abatement of methane and nitrous oxide from ruminant production: A review. *Livestock Science*. 130 (1), 47–56.

Eddy, S. R. (2011) Accelerated Profile HMM Searches. *PLOS Computational Biology*. 7 (10), e1002195.

Edwards, J. E. et al. (2008) Dynamics of initial colonization of nonconserved perennial ryegrass by anaerobic fungi in the bovine rumen. *FEMS Microbiology Ecology*. 66 (3), 537–545.

Eisenstein, M. (2020) Early investments powering the ascent of microbiome therapeutics. *biopharma dealmakers*.

Engin, E. D. (2021) ‘Bacterial protein kinases’, in Ayse Basak Engin & Atilla

Engin (eds.) *Protein Kinase-mediated Decisions Between Life and Death*. Cham: Springer International Publishing. pp. 323–338.

Escobar-Zepeda, A. et al. (2015) The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*. 61–15.

Federhen, S. (2012) The NCBI taxonomy database. *Nucleic Acids Research*. 40 (D1), D136–D143.

Fernando, S. C. et al. (2010) Rumen microbial population dynamics during adaptation to a high-grain diet. *Applied and Environmental Microbiology*. 76 (22), 7482–7490.

Franzosa, E. A. et al. (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*. 15 (11), 962–968.

G., L. K. et al. (2018) Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *mSystems*. 3 (5), e00055-18.

Gao, X.-Y. et al. (2014) Comparative genomics of the bacterial genus Streptococcus illuminates evolutionary implications of species groups. *PLOS ONE*. 9 (6), e101229.

Garrett, W. S. (2015) Cancer and the microbiota. *Science*. 348 (6230), 80–86.

Gerber, P. J. et al. (2013) *Tackling climate change through livestock – A global assessment of emissions and mitigation opportunities.* Food and Agriculture Organization of the United Nations (FAO), Rome.

Gerlt, J. A. & Babbitt, P. C. (2000) Can sequence determine function? *Genome Biology*. 1 (5), reviews0005.1.

Gevers, D. et al. (2004) Gene duplication and biased functional retention of

paralogs in bacterial genomes. *Trends in Microbiology*. 12 (4), 148–154.

Gilbert, R. A. et al. (2020) Rumen Virus Populations: Technological Advances Enhancing Current Understanding . Frontiers in Microbiology 11.

Gilbert, R. A. & Klieve, A. V (2015) 'Ruminal viruses (bacteriophages, archaeaphages)', in Anil Kumar Puniya et al. (eds.) *Rumen Microbiology: From Evolution to Revolution*. New Delhi: Springer India. pp. 121–141.

Godfray, H. C. J. (2002) Challenges for taxonomy. *Nature*. 417 (6884), 17–19.

Goris, J. et al. (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*. 57 (1), 81–91.

Gourlé, H. et al. (2019) Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*. 35 (3), 521–522.

Grilli, D. J. et al. (2013) Isolation of Pseudobutyrivibrio ruminis and Pseudobutyrivibrio xylanivorans from rumen of Creole goats fed native forage diet. *Folia Microbiologica*. 58 (5), 367–373.

Guigó, R. et al. (2000) An assessment of gene prediction accuracy in large DNA sequences. *Genome Research*. 10 (10), 1631–1642.

Gullapalli, R. R. (2020) Evaluation of commercial next-generation sequencing bioinformatics software solutions. *The Journal of Molecular Diagnostics*. 22 (2), 147–158.

Gurevich, A. et al. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 29 (8), 1072–1075.

Hackmann, T. J. & Firkins, J. L. (2015) Electron transport phosphorylation in rumen butyrivibrios: unprecedented ATP yield for glucose fermentation

to butyrate . Frontiers in Microbiology 6.

Hamady, M. & Knight, R. (2009) Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Research*. 19 (7), 1141–1152.

Harfoot, C. G. (1981) 'Anatomy, physiology and microbiology of the ruminant digestive tract', in WILLIAM W B T - Lipid Metabolism in Ruminant Animals CHRISTIE (ed.) *Lipid Metabolism in Ruminant Animals*. Pergamon. pp. 1–19.

Harlow, B. E. et al. (2020) Isoflavone supplementation, via red clover hay, alters the rumen microbial community and promotes weight gain of steers grazing mixed grass pastures. *PLOS ONE*. 15 (3), e0229200.

Harris, C. R. et al. (2020) Array programming with NumPy. *Nature*. 585 (7825), 357–362.

Hayes, B. J. et al. (2013) The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends in Genetics*. 29 (4), 206–214.

Hedlund, B. P. et al. (2022) SeqCode: a nomenclatural code for prokaryotes described from sequence data. *Nature Microbiology*. 7 (10), 1702–1708.

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

Herrera, P. et al. (2009) Ecology and pathogenicity of gastrointestinal Streptococcus bovis. *Anaerobe*. 15 (1), 44–54.

Heyndrickx, M. et al. (1998) Virgibacillus: a new genus to accommodate Bacillus pantothenticus (Proom and Knight 1950). Emended description of Virgibacillus pantothenticus. *International Journal of Systematic*

Bacteriology. 48 (1), 99–106.

Hiseni, P. et al. (2022) Questioning the Quality of 16S rRNA Gene Sequences Derived From Human Gut Metagenome-Assembled Genomes . Frontiers in Microbiology 12.

Hofmann, R. R. (1989) Evolutionary steps of ecophysiological adaptation and diversification of ruminants: a comparative view of their digestive system. *Oecologia*. 78 (4), 443–457.

Hornby, D. P. (1993) 'DNA methyltransferases (EC 2.1.1.72 and EC 2.1.1.73)', in Michael M Burrell (ed.) *Enzymes of Molecular Biology*. Totowa, NJ: Humana Press. pp. 201–211.

Hosokawa, M. et al. (2022) Strain-level profiling of viable microbial community by selective single-cell genome sequencing. *Scientific Reports*. 12 (1), 4443.

Huerta-Cepas, J. et al. (2019) EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*. 47 (D1), D309–D314.

Hugenholtz, P. et al. (2021) Prokaryotic taxonomy and nomenclature in the age of big sequence data. *The ISME Journal*. 15 (7), 1879–1892.

Hungate, R. E. (1969) '*A roll tube method for cultivation of strict anaerobes*', in J R Norris & D W B T - Methods in Microbiology Ribbons (eds.) Academic Press. pp. 117–132.

Hungate, R. E. (1947) Studies on cellulose fermentation: the culture and Isolation for cellulose-decomposing bacteria from the rumen of cattle. *Journal of Bacteriology*. 53 (5), 631–645.

Hungate, R. E. (1944) Studies on cellulose fermentation: the culture and Physiology of an anaerobic cellulose-digesting bacterium. *Journal of*

Bacteriology. 48 (5), 499–513.

Hungate, R. E. (1966) *The rumen and its microbes*. Elsevier.

Hunter, J. D. (2007) Matplotlib: a 2D graphics environment. *Computing in Science & Engineering*. 9 (3), 90–95.

Huson, D. H. et al. (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Research* . 21 (9), 1552–1560.

Huws, S. A. et al. (2018) Addressing global ruminant agricultural challenges through understanding the rumen microbiome: past, present, and future. *Frontiers in Microbiology*. 91–33.

Huws, S. A. et al. (2010) Forage type and fish oil cause shifts in rumen bacterial diversity. *FEMS Microbiology Ecology*. 73 (2), 396–702.

Huws, S. A. et al. (2016) Temporal dynamics of the metabolically active rumen bacteria colonizing fresh perennial ryegrass. *FEMS Microbiology Ecology*. 92 (1), fiv137.

Hyatt, D. et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 11 (1), 119.

Jain, C. et al. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*. 9 (1), 1–8.

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

Jami, E. & Mizrahi, I. (2012) Composition and similarity of bovine rumen microbiota across individual animals. *PLoS ONE*. 7 (3), 1–8.

Janda, J. M. & Abbott, S. L. (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls.

Journal of Clinical Microbiology 45 (9) p.2761–2764.

Janssen, P. H. & Kirs, M. (2008) Structure of the archaeal community of the rumen. *Applied and Environmental Microbiology*. 74 (12), 3619–3625.

Jenkins, T. C. (1993) Lipid metabolism in the rumen. *Journal of Dairy Science*. 76 (12), 3851–3863.

Jhamb, K. (2020) *Microbiome therapeutics: global markets*.

Jo, H. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*. 68 (4), 669–685.

Johnson, J. S. et al. (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*. 10 (1), 5029.

Johnson, M. et al. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Research*. 36 (suppl_2), W5–W9.

Jones, S. (2013) Trends in microbiome research. *Nature Biotechnology*. 31 (4), 277.

Kamra, D. N. (2005) Rumen microbial ecosystem. *Current Science*. 89 (1), 124–135.

Kanehisa, M. & Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 28 (1), 27–30.

Kang, D. D. et al. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*. 2015 (8), 1–15.

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

Joseph Gillespie (ed.). *PeerJ*. 7e7359.

Karp, P. D. et al. (2019) The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*. 20 (4), 1085–1093.

Kennedy, N. A. et al. (2014) The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLOS ONE*. 9 (2), e88982.

Kim, D. et al. (2016) Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*. 26 (12), 1721–1729.

Kim, M. et al. (2011) Status of the phylogenetic diversity census of ruminal microbiomes. *FEMS Microbiology Ecology*. 76 (1), 49–63.

Kim, M. et al. (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*. 64 (Pt_2), 346–351.

Kingston-Smith, A. H. et al. (2013) Comparative metabolite fingerprinting of the rumen system during colonisation of three forage grass (Lolium perenne L.) varieties. *PLOS ONE*. 8 (11), e82801.

Kogawa, M. et al. (2018) Obtaining high-quality draft genomes from uncultured microbes by cleaning and co-assembly of single-cell amplified genomes. *Scientific Reports*. 8 (1), 2059.

Konstantinidis, K. T. & Tiedje, J. M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences*. 102 (7), 2567–2572.

Konstantinidis, K. T. & Tiedje, J. M. (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current Opinion in Microbiology*. 10 (5), 504–509.

Kracke, F. et al. (2015) Microbial electron transport and energy conservation – the foundation for optimizing bioelectrochemical systems . Frontiers in Microbiology 6.

Ksiezarek, M. et al. (2022) Genomic diversity of genus Limosilactobacillus. *Microbial Genomics*. 8 (7), 000847.

Kulakov, L. A. et al. (2002) Analysis of bacteria contaminating ultrapure water in industrial systems. *Applied and Environmental Microbiology*. 68 (4), 1548–1555.

Kurtz, S. et al. (2004) Versatile and open software for comparing large genomes. *Genome Biology*. 5 (2), R12.

Lamble, S. et al. (2013) Improved workflows for high throughput library preparation using the transposome-based nextera system. *BMC Biotechnology*. 13 (1), 104.

Larson, G. & Burger, J. (2013) A population genetics view of animal domestication. *Trends in Genetics*. 29 (4), 197–205.

Lau, J. T. et al. (2016) Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome Medicine*. 8 (1), 72.

Leahy, S. C. et al. (2013) Genome sequencing of rumen bacteria and archaea and its application to methane mitigation strategies. *Animal*. 7235–243.

Van Leeuwenhoek, A. (1996) *The collected letters of Antoni Van Leeuwenhoek - volume 14*. 1st edition. A Committee of Dutch Scientists.

Li, D. et al. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 31 (10), 1674–1676.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*. 00 (00), 3.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34 (18), 3094–3100.

Li, H. et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25 (16), 2078–2079.

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25 (14), 1754–1760.

Lima, F. S. et al. (2015) Prepartum and postpartum rumen fluid microbiomes: characterization and correlation with production traits in dairy cows. *Applied and Environmental Microbiology*. 81 (4), 1327–1337.

Liu, L. et al. (2022) Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes. *Microbiome*. 10 (1), 209.

Liu, S. et al. (2022) Opportunities and challenges of using metagenomic data to bring uncultured microbes into cultivation. *Microbiome*. 10 (1), 76.

Loewenstein, Y. et al. (2009) Protein function annotation by homology-based inference. *Genome Biology*. 10 (2), 207.

Lowe, T. M. & Eddy, S. R. (1997) tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research*. 25 (5), 955–964.

Mackie, R. I. (2002) Mutualistic fermentative digestion in the gastrointestinal tract: diversity and evolution. *Integrative and Comparative Biology*. 42 (2), 319–326.

Maekawa, M. et al. (2002) Effect of concentrate level and feeding management on chewing activities, saliva production, and ruminal pH of

lactating dairy cows. *Journal of Dairy Science*. 85 (5), 1165–1175.

Maguire, B. A. & Zimmermann, R. A. (2001) The ribosome in focus. *Cell*. 104 (6), 813–816.

Maguire, F. et al. (2020) Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microbial Genomics*. 6 (10), 1–12.

Maidak, B. L. et al. (1997) The RDP (Ribosomal Database Project). *Nucleic Acids Research*. 25 (1), 109–110.

Mamuad, L. L. et al. (2019) Rumen fermentation and microbial community composition influenced by live Enterococcus faecium supplementation. *AMB Express*. 9 (1), 123.

Martin, V. J. et al. (2016) Transitioning from descriptive to mechanistic understanding of the microbiome: the need for a prospective longitudinal approach to predicting disease. *The Journal of Pediatrics*. 179240–248.

Martínez-Álvaro, M. et al. (2020) Identification of complex rumen microbiome interaction within diverse functional niches as mechanisms affecting the variation of methane emissions in bovine. *Frontiers in Microbiology*. 111–13.

Matsen, F. A. et al. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*. 11 (1), 538.

Matthews, C. et al. (2019) The rumen microbiome: a crucial consideration when optimising milk and meat production and nitrogen utilisation efficiency. *Gut Microbes*. 10 (2), 115–132.

Mazmanian, S. K. & Kasper, D. L. (2006) The love–hate relationship between bacterial polysaccharides and the host immune system. *Nature Reviews*

Immunology. 6 (11), 849–858.

McAllister, T. A. & Newbold, C. J. (2008) Redirecting rumen fermentation to reduce methanogenesis. *Australian Journal of Experimental Agriculture*. 48 (2), 7–13.

McCarthy, B. J. & Bolton, E. T. (1963) An approach to the measurement of genetic relatedness among organisms. *Proceedings of the National Academy of Sciences*. 50 (1), 156–164.

McMichael, A. J. et al. (2007) Food, livestock production, energy, climate change, and health. *The Lancet*. 370 (9594), 1253–1263.

Membrive, C. M. B. (2016) 'Anatomy and physiology of the rumen', in Danilo Domingues Millen et al. (eds.) *Rumenology*. Cham: Springer International Publishing. pp. 1–38.

Meyer, F. et al. (2008) The metagenomics RAST server - A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. 91–8.

Meziti, A. et al. (2021) The reliability of Metagenome-Assembled Genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same fecal sample. *Applied and Environmental Microbiology*. 87 (6), e02593-20.

Minich, J. J. et al. (2019) Quantifying and understanding well-to-well contamination in microbiome research. *mSystems*. 4 (4), e00186-19.

Mistry, J. et al. (2021) Pfam: The protein families database in 2021. *Nucleic Acids Research*. 49 (D1), D412–D419.

Mizrahi, I. et al. (2021) The rumen microbiome: balancing food security and environmental impacts. *Nature Reviews Microbiology*. 19 (9), 553–566.

Molari, M. et al. (2023) A hydrogenotrophic Sulfurimonas is globally abundant

in deep-sea oxygen-saturated hydrothermal plumes. *Nature Microbiology*. 8 (4), 651–665.

Morita, H. et al. (2008) Sharpea azabuensis gen. nov., sp. nov., a Gram-positive, strictly anaerobic bacterium isolated from the faeces of thoroughbred horses. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*. 58 (12), 2682–2686.

Mukherjee, S. et al. (2021) Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Research*. 49 (D1), D723–D733.

Murray, A. E. et al. (2020) Roadmap for naming uncultivated Archaea and Bacteria. *Nature Microbiology*.

Nagaraja, T. G. (2016) 'Microbiology of the rumen', in Danilo Domingues Millen et al. (eds.) *Rumenology*. Cham: Springer International Publishing. pp. 39–61.

Nasko, D. J. et al. (2018) RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biology*. 19 (1), 1–10.

Nayfach, S. et al. (2021) A genomic catalog of Earth's microbiomes. *Nature Biotechnology*. 39 (4), 499–509.

Nazeen, S. et al. (2020) Carnelian uncovers hidden functional patterns across diverse study populations from whole metagenome sequencing reads. *Genome Biology*. 21 (1), 1–18.

Nelson, W. C. et al. (2020) Biases in genome reconstruction from metagenomic data Gerard Lazo (ed.). *PeerJ*. 8e10119.

Neves, A. L. A. et al. (2017) Enhancing the Resolution of Rumen Microbial Classification from Metatranscriptomic Data Using Kraken and Mothur . Frontiers in Microbiology 8.

Newbold, C. J. et al. (2015) The role of ciliate protozoa in the rumen. Frontiers in Microbiology 6.

Nurk, S. et al. (2017) MetaSPAdes: A new versatile metagenomic assembler. *Genome Research*. 27 (5), 824–834.

O'Leary, N. A. et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. 44 (D1), D733–D745.

Olm, M. R. et al. (2017) dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal*. 11 (12), 2864–2868.

Olsen, G. J. et al. (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annual Review of Microbiology*. 40 (1), 337–365.

Ondov, B. D. et al. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*. 17 (1), 132.

Pace, N. R. (1997) A molecular view of microbial diversity and the biosphere. *Science*. 276 (5313), 734–740.

Pan, S. et al. (2022) A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nature Communications*. 13 (1), 2326.

Papadimitriou, K. et al. (2014) Comparative genomics of the dairy isolate Streptococcus macedonicus ACA-DC 198 against related members of the Streptococcus bovis/Streptococcus equinus complex. *BMC Genomics*. 15 (1), 272.

Parker, C. T. et al. (2019) International code of nomenclature of Prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*. 69

(1), S1.

Parks, D. H. et al. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*. 36 (10), 996.

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

Parks, D. H. et al. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*. 2 (11), 1533–1542.

Patra, A. et al. (2017) Rumen methanogens and mitigation of methane emission by anti-methanogenic compounds and substances. *Journal of Animal Science and Biotechnology*. 8 (1), 13.

Peng, Y. et al. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. 28 (11), 1420–1428.

Peng, Y. et al. (2011) Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics*. 27 (13), i94–i101.

Philip, H. et al. (2001) Investigation of candidate division TM7, a recently recognized major lineage of the domain bacteria with no known pure-culture representatives. *Applied and Environmental Microbiology*. 67 (1), 411–419.

Pollock, J. et al. (2018) The madness of microbiome: attempting to find consensus “best practice” for 16S microbiome studies. *Applied and Environmental Microbiology*. 84 (7), e02627-17.

Pope, P. B. et al. (2011) Isolation of Succinivibrionaceae implicated in low

methane emissions from tammar wallabies. *Science*. 333 (6042), 646–648.

Poyet, M. et al. (2019) A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nature Medicine*. 25 (9), 1442–1452.

Prados-Bo, A. & Casino, G. (2021) Microbiome research in general and business newspapers: How many microbiome articles are published and which study designs make the news the most? *PLoS ONE*. 16 (4 April), 1–14.

Prakash, S. et al. (2011) Gut microbiota: next frontier in understanding human health and development of biotherapeutics. *Biologics: Targets and Therapy*. 571–86.

Prakash, T. & Taylor, T. D. (2012) Functional assignment of metagenomic data: Challenges and applications. *Briefings in Bioinformatics*. 13 (6), 711–727.

Proctor, L. M. et al. (2019) The integrative human microbiome project. *Nature*. 569 (7758), 641–648.

Quast, C. et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*. 41 (D1), D590–D596.

Quince, C. et al. (2017) Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*. 35 (9), 833–844.

Quinlan, A. R. (2014) BEDTools: the Swiss-Army tool for genome feature analysis. *Current Protocols in Bioinformatics*. 47 (1), 11.12.1-11.12.34.

Rajilić-Stojanović, M. et al. (2007) Diversity of the human gastrointestinal tract microbiota revisited. *Environmental Microbiology*. 9 (9), 2125–2136.

Ranran, H. et al. (2023) Long-Read Metagenomics of Marine Microbes Reveals Diversely Expressed Secondary Metabolites. *Microbiology Spectrum*. 0 (0), e01501-23.

Rappé, M. S. & Giovannoni, S. J. (2003) The uncultured microbial majority. *Annual review of microbiology*. 57369–394.

Rasko, D. A. et al. (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of Bacteriology*. 190 (20), 6881–6893.

Rawlings, N. D. et al. (2018) The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research*. 46 (D1), D624–D632.

Ren, J. et al. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*. 5 (1), 69.

Rey, M. et al. (2014) Establishment of ruminal bacterial community in dairy calves from birth to weaning is sequential. *Journal of Applied Microbiology*. 116 (2), 245–257.

Rhee, S. H. et al. (2009) Principles and clinical implications of the brain–gut–enteric microbiota axis. *Nature Reviews Gastroenterology & Hepatology*. 6 (5), 306–314.

Richter, M. & Rosselló-Móra, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences*. 106 (45), 19126–19131.

Di Rienzi, S. C. et al. (2013) The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria Roberto Kolter (ed.). *eLife*. 2e01102.

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

Ripple, W. J. et al. (2014) Ruminants, climate change and climate policy. *Nature Climate Change*. 4 (1), 2–5.

Rosario, G. et al. (2004) Determination of the core of a minimal bacterial gene set. *Microbiology and Molecular Biology Reviews*. 68 (3), 518–537.

Rosero, J. A. et al. (2016) Reclassification of Eubacterium rectale (Hauduroy et al. 1937) prévot 1938 in a new genus agathobacter gen. nov. as Agathobacter rectalis comb. nov., and description of Agathobacter ruminis sp. nov., isolated from the rumen contents of sheep and cows. *International Journal of Systematic and Evolutionary Microbiology*. 66 (2), 768–773.

Van Rossum, T. et al. (2020) Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology*. 18 (9), 491–506.

Rother, M. & Krzycki, J. A. (2010) Selenocysteine, pyrrolysine, and the unique energy metabolism of methanogenic archaea Jerry Eichler (ed.). *Archaea*. 2010453642.

Rouli, L. et al. (2015) The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*. 772–85.

Roux, S. et al. (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*. 537 (7622), 689–693.

Roux, S. et al. (2015) VirSorter: mining viral signal from microbial genomic data Kimberly Bishop-Lilly (ed.). *PeerJ*. 3e985.

Russell, J. B. (2009) ‘Rumen’, in Moselio B T - Encyclopedia of Microbiology (Third Edition) Schaechter (ed.) *Encyclopedia of Microbiology (Third Edition)*. Oxford: Academic Press. pp. 163–174.

Salter, S. J. et al. (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*. 12 (1), 1–12.

Sangwan, N. et al. (2016) Recovering complete and draft population genomes from metagenome datasets. *Microbiome*. 4 (1), 8.

Sayers, E. W. et al. (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Research*. 50 (D1), D20–D26.

Scheu, A. et al. (2015) The genetic prehistory of domesticated cattle from their origin to the spread across Europe. *BMC Genetics*. 16 (1), 54.

Schnoes, A. M. et al. (2009) Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*. 5 (12), .

Scholz, M. B. et al. (2012) Next generation sequencing and bioinformatic bottlenecks: The current state of metagenomic data analysis. *Current Opinion in Biotechnology*. 23 (1), 9–15.

Seemann, T. (2014) Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. 30 (14), 2068–2069.

Segata, N. et al. (2013) Computational meta'omics for microbial community studies. *Molecular Systems Biology*. 9 (1), 666.

Segerman, B. (2020) The most frequently used sequencing technologies and assembly methods in different time segments of the bacterial surveillance and RefSeq genome databases. Frontiers in Cellular and Infection Microbiology 10.

Seshadri, R. et al. (2018) Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nature Biotechnology*. 36 (4), 359–367.

Shade, A. et al. (2012) Culturing captures members of the soil rare biosphere. *Environmental Microbiology*. 14 (9), 2247–2252.

Shaffer, M. et al. (2020) DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*. 48 (16), 8883–8900.

Singleton, C. M. et al. (2021) Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nature Communications*. 12 (1), 2009.

Smith, R. H. et al. (2022) Investigating the impact of database choice on the accuracy of metagenomic read classification for the rumen microbiome. *Animal Microbiome*. 4 (1), 57.

Sneath, P. H. A. (2001) 'Bacterial nomenclature', in David R Boone et al. (eds.) *Bergey's Manual® of Systematic Bacteriology: Volume One : The Archaea and the Deeply Branching and Phototrophic Bacteria*. New York, NY: Springer New York. pp. 83–88.

Snelling, T. J. et al. (2019) Temporal stability of the rumen microbiota in beef cattle, and response to diet and supplements. *Animal Microbiome*. 1 (1), 16.

Snipen, L. et al. (2021) Reduced metagenome sequencing for strain-resolution taxonomic profiles. *Microbiome*. 9 (1), 79.

Soo, R. M. et al. (2014) An Expanded Genomic Representation of the Phylum Cyanobacteria. *Genome Biology and Evolution*. 6 (5), 1031–1045.

Speth, D. R. et al. (2022) Microbial communities of Auka hydrothermal sediments shed light on vent biogeography and the evolutionary history of thermophily. *The ISME Journal*. 16 (7), 1750–1764.

de Steenhuijsen Piters, W. A. A. et al. (2019) Interaction between the nasal microbiota and S. pneumoniae in the context of live-attenuated influenza vaccine. *Nature Communications*. 10 (1), 2981.

Stergiadis, S. et al. (2021) Unravelling the role of rumen microbial communities, genes, and activities on milk fatty acid profile using a combination of omics approaches. Frontiers in Microbiology 11.

Stewart, C. S. et al. (1997) 'The rumen bacteria', in P N Hobson & C S Stewart (eds.) *The Rumen Microbial Ecosystem*. Dordrecht: Springer Netherlands. pp. 10–72.

Stewart, R. D. et al. (2018) Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen Robert. *Nature Communications*. 91–11.

Stewart, Robert D. et al. (2019) Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnology*. 37 (8), 953–961.

Stewart, Robert D et al. (2019) MAGpy: a reproducible pipeline for the downstream analysis of metagenome-assembled genomes (MAGs). *Bioinformatics*. 35 (12), 2150–2152.

Strous, M. et al. (2012) The binning of metagenomic contigs for microbial physiology of mixed cultures. Frontiers in Microbiology 3.

Stulberg, E. et al. (2016) An assessment of US microbiome research. *Nature Microbiology*. 1 (1), 15015.

Suzek, B. E. et al. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*. 23 (10), 1282–1288.

Suzuki, E. I. & Komagata, K. (1983) Taxonomic significance of cellular fatty acid composition in some coryneform bacteria. *International Journal of*

Systematic Bacteriology. 33 (2), 188–200.

Szymanski, M. et al. (2002) 5S Ribosomal RNA database. *Nucleic Acids Research*. 30 (1), 176–178.

Taguchi, H. et al. (2004) Partial characterization of structure and function of a xylanase gene from the rumen hemicellulolytic bacterium Eubacterium ruminantium. *Animal Science Journal*. 75 (4), 325–332.

Tajima, K. et al. (2001) Phylogenetic analysis of archaeal 16S rRNA libraries from the rumen suggests the existence of a novel group of archaea not associated with known methanogens. *FEMS Microbiology Letters*. 200 (1), 67–72.

Tao, Y. et al. (2022) Improved Assembly of Metagenome-Assembled Genomes and Viruses in Tibetan Saline Lake Sediment by HiFi Metagenomic Sequencing. *Microbiology Spectrum*. 11 (1), e03328-22.

Tapio, I. et al. (2017) The ruminal microbiome associated with methane emissions from ruminant livestock. *Journal of Animal Science and Biotechnology*. 8 (1), 7.

The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*. 49 (D1), D480–D489.

Tindall, B. J. et al. (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. *International Journal of Systematic and Evolutionary Microbiology*. 60 (1), 249–266.

Ungerfeld, E. M. (2020) Metabolic hydrogen flows in rumen fermentation: principles and possibilities of interventions. Frontiers in Microbiology 11.

Varghese, N. J. et al. (2015) Microbial species delineation using whole genome sequences. *Nucleic Acids Research*. 43 (14), 6761–6771.

Větrovský, T. & Baldrian, P. (2013) The Variability of the 16S rRNA Gene in

Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE*. 8 (2), 1–10.

Walker, A. W. (2019) A lot on your plate? Well-to-well contamination as an additional confounder in microbiome sequence analyses. *mSystems*. 4 (4), e00362-19.

Walker, A. W. et al. (2014) Phylogeny, culturing, and metagenomics of the human gut microbiota. *Trends in Microbiology*. 22 (5), 267–274.

Wallace, R. J. et al. (1997) Peptidases of the rumen bacterium, Prevotella ruminicola. *Anaerobe*. 3 (1), 35–42.

Wallace, R. J. et al. (2015) The rumen microbial metagenome associated with high methane production in cattle. *BMC Genomics*. 16 (1), 1–14.

Watson, M. (2021) New insights from 33,813 publicly available metagenome-assembled-genomes (MAGs) assembled from the rumen microbiome. *bioRxiv*. 2021.04.02.438222.

Weimer, P. J. (2015) Redundancy, resilience, and host specificity of the ruminal microbiota: implications for engineering improved ruminal fermentations . Frontiers in Microbiology 6.

Whitford, M. F. et al. (2001) Phylogenetic analysis of methanogens from the bovine rumen. *BMC Microbiology*. 1 (1), 5.

Wilkinson, J. E. et al. (2021) A framework for microbiome science in public health. *Nature Medicine*. 27 (5), 766–774.

Williams, A. G. & Coleman, G. S. (1997) ‘The rumen protozoa’, in P N Hobson & C S Stewart (eds.) *The Rumen Microbial Ecosystem*. Dordrecht: Springer Netherlands. pp. 73–139.

Wong, H. L. et al. (2020) Microbial dark matter filling the niche in hypersaline microbial mats. *Microbiome*. 8 (1), 135.

Wright, A.-D. G. (2015) ‘Rumen protozoa’, in Anil Kumar Puniya et al. (eds.) *Rumen Microbiology: From Evolution to Revolution*. New Delhi: Springer India. pp. 113–120.

Xie, Z. (2021) ‘*The methods and tools for mobile genetic element detection and their application to systems medicine*’, in Olaf B T - Systems Medicine Wolkenhauer (ed.) Oxford: Academic Press. pp. 203–207.

Xue, M.-Y. et al. (2020) Multi-omics reveals that the rumen microbiome and its metabolome together with the host metabolome contribute to individualized dairy cow performance. *Microbiome*. 8 (1), 64.

Yandell, M. & Ence, D. (2012) A beginner’s guide to eukaryotic genome annotation. *Nature Reviews Genetics*. 13 (5), 329–342.

Yin, Y. et al. (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*. 40 (W1), W445–W451.

Yoon, S.-H. et al. (2017) A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek*. 110 (10), 1281–1286.

Yuan, C. et al. (2015) Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics*. 31 (12), i35–i43.

Zaneveld, J. et al. (2008) Host-bacterial coevolution and the search for new drug targets. *Current Opinion in Chemical Biology*. 12 (1), 109–114.

Zehavi, T. et al. (2018) Insights into culturomics of the rumen microbiome. *Frontiers in Microbiology*. 91–10.

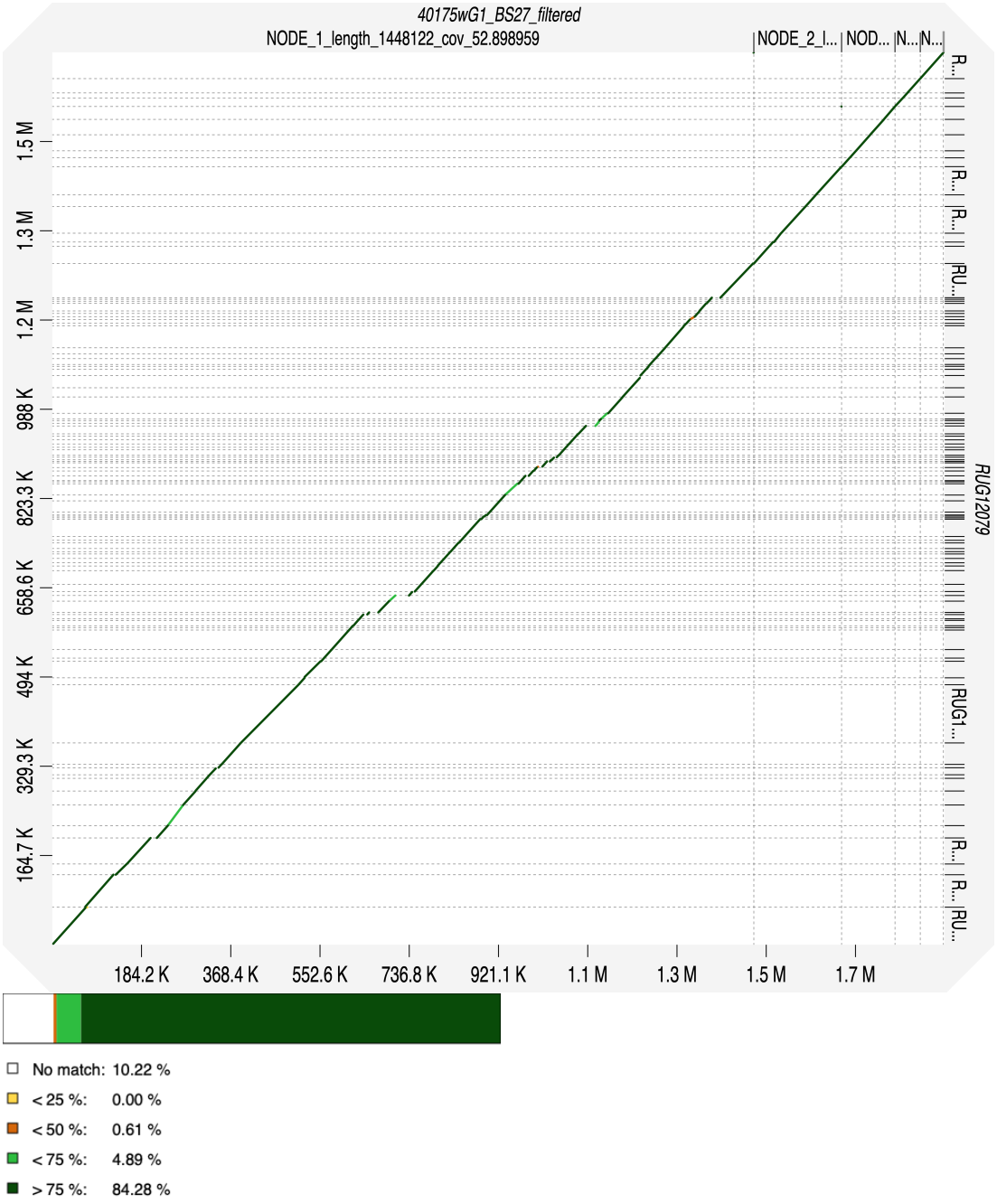
Zhou, Zhemin et al. (2020) Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. *Genome Research* . 30 (11), 1667–1679.

Zhou, Zhichao et al. (2020) Gammaproteobacteria mediating utilization of

methyl-, sulfur- and petroleum organic compounds in deep ocean hydrothermal plumes. *The ISME Journal*. 14 (12), 3136–3148.

Ziemer, C. J. et al. (2008) Comparison of microbial populations in model and natural rumens using 16S ribosomal RNA-targeted probes. *Environmental Microbiology*. 2 (6), 632–643.

Chapter 7: Appendix



Supplementary Figure S3.1: Alignment of culture-derived genome "40175wG1_BS27" and the MAG "RUG12079".

Supplementary Table S3.1: CheckM results assessing the quality of the culture-derived genomes before and after filtering. The cultured isolate genomes (n=62) were filtered at the contig level to reduce contamination. Genomes that were contaminated >=90% (n=15) were saved to multiple files to assess whether they were mixed complete genomes. All separated and filtered genomes (n=77) were re-assessed for completeness, contamination, and strain heterogeneity with CheckM. Those that were of extremely poor quality were subsequently removed from the study and are denoted by *(R)*.

Genome ID before filtering	Genome ID after filtering	Completeness	Contamination	Strain heterogeneity
40175wA1_BS2		100	0.01	100
	40175wA1_BS2_filtered	99.63	0	0
40175wA2_BS4		100	0	0
	40175wA2_BS4_filtered	100	0	0
40175wA3_BS21 *(R)*	N/A - removed from study	95.94	44.59	95.2
40175wA4_BSR38		99.19	0.27	0
	40175wA4_BSR38_filtered	99.19	0.27	0
40175wA5_BS48		100	0	0
	40175wA5_BS48_filtered	100	0	0
40175wA6_BS64		100	0	0
	40175wA6_BS64_filtered	100	0	0
40175wA7_BSR23		98.25	90.35	3.06
	40175wA7_BSR23_1_filtered	99.31	0	0
	40175wA7_BSR23_2_filtered	91.18	2.2	11.76
40175wA8_BSR32		99.26	0.44	25
	40175wA8_BSR32_filtered	99.26	0.41	33.33
40175wB1_BS5		100	5.06	100
	40175wB1_BS5_filtered	100	4	100
40175wB2_BSR40		100	176.07	41.92
	40175wB2_BSR40_1_filtered	97.14	3.57	90.62
	40175wB2_BSR40_2_filtered	82.56	1.52	85.71
	40175wB2_BSR40_3_filtered *(R)*	2.74	0	0
	40175wB2_BSR40_4_filtered	80.71	3.33	5.88
40175wB3_BS17		100	140.08	3.16
	40175wB3_BS17_1_filtered	91.8	2.25	80
	40175wB3_BS17_2_filtered	100	5.55	95.65
40175wB4_BS41R		100	100	0
	40175wB4_BS41R_1_filtered	99.52	0.16	0

	40175wB4_BS41R_2_filtered	99.98	1.2	0
40175wB5_BS49		100	100	71.43
	40175wB5_BS49_1_filtered	100	0	0
	40175wB5_BS49_2_filtered	88.37	0.86	0
40175wB6_BS66		99.52	0.16	0
	40175wB6_BS66_filtered	99.52	0.16	0
40175wB7_BSR24		98.68	0.41	33.33
	40175wB7_BSR24_filtered	98.68	0.41	33.33
40175wB8_BSR33		97.98	1.08	0
	40175wB8_BSR33_filtered	97.98	1.08	0
40175wC1_BS22		100	0	0
	40175wC1_BS22_filtered	100	0	0
40175wC2_BS30R		98.42	0.63	0
	40175wC2_BS30R_filtered	98.42	0.63	0
40175wC3_BS20		99.62	0	0
	40175wC3_BS20_filtered	99.62	0	0
40175wC4_BS42		100	71.29	0
	40175wC4_BS42_filtered	100	0	0
40175wC5_BS50		100	0	0
	40175wC5_BS50_filtered	100	0	0
40175wC6_BS69		100	0.29	0
	40175wC6_BS69_filtered	100	0	0
40175wC7_BSR25		100	100	3.57
	40175wC7_BSR25_1_filtered	98.86	0	0
	40175wC7_BSR25_2_filtered	99.31	1.38	0
40175wC8_BSR34		100	100	2.8
	40175wC8_BSR34_1_filtered	99.43	0.77	0
	40175wC8_BSR34_2_filtered	97.98	1.08	0
40175wD1_BS10		100	0	0
	40175wD1_BS10_filtered	100	0	0
40175wD2_BS28		100	217.29	0.91
	40175wD2_BS28_1_filtered	98.8	2.04	57.14
	40175wD2_BS28_2_filtered	100	5.39	68.97
40175wD3_BS34		99.48	0	0
	40175wD3_BS34_filtered	99.48	0	0

40175wD4_BS43		100	0	0
	40175wD4_BS43_filtered	100	0	0
40175wD5_BS51		100	0	0
	40175wD5_BS51_filtered	100	0	0
40175wD6_BSR1		100	0	0
	40175wD6_BSR1_filtered	100	0	0
40175wD7_BSR26		98.25	97.93	3.85
	40175wD7_BSR26_1_filtered	89.37	2.02	20
	40175wD7_BSR26_2_filtered	99.31	0	0
40175wD8_BSR35		99.19	0.36	0
	40175wD8_BSR35_filtered	99.19	0.27	0
40175wE1_BS8		100	0	0
	40175wE1_BS8_filtered	100	0	0
40175wE2_BS31		99.18	0	0
	40175wE2_BS31_filtered	99.18	0	0
40175wE3_BS32		99.18	0	0
	40175wE3_BS32_filtered	99.18	0	0
40175wE4_BS44		100	0	0
	40175wE4_BS44_filtered	100	0	0
40175wE5_BS54		100	0.12	100
	40175wE5_BS54_filtered	100	0.12	100
40175wE7_BSR27		100	100.38	5.17
	40175wE7_BSR27_1_filtered	98.62	0	0
	40175wE7_BSR27_2_filtered	98.68	0.41	33.33
40175wE8_BSR37		99.19	0.27	0
	40175wE8_BSR37_filtered	99.19	0.27	0
40175wF1_BS7		99.62	0	0
	40175wF1_BS7_filtered	99.62	0	0
40175wF2_BS19		100	0	0
	40175wF2_BS19_filtered	100	0	0
40175wF3_BS52		100	148.51	0
	40175wF3_BS52_1_filtered	100	0	0
	40175wF3_BS52_2_filtered	84.86	1.92	0
40175wF4_BS45		99.06	4.01	0
	40175wF4_BS45_filtered	99.06	0.94	0

40175wF5_BS57		96.21	26.49	87.9
	40175wF5_BS57_filtered	96.21	26.49	87.9
40175wF6_BSX		99.06	1.08	0
	40175wF6_BSX_filtered	98.11	1.08	0
40175wF7_BSR28		99.26	0.7	25
	40175wF7_BSR28_filtered	99.26	0.41	33.33
40175wF8_BSR3		100	0	0
	40175wF8_BSR3_filtered	100	0	0
40175wG1_BS27		100	0	0
	40175wG1_BS27_filtered	100	0	0
40175wG2_BS13		100	0	0
	40175wG2_BS13_filtered	100	0	0
40175wG3_BS35		99.77	0.95	0
	40175wG3_BS35_filtered	99.77	0.95	0
40175wG4_BS46		99.06	1.13	0
	40175wG4_BS46_filtered	99.06	0.94	0
40175wG5_BS61		99.52	0.4	50
	40175wG5_BS61_filtered	99.52	0.16	0
40175wG6_BSR6R		100	1.33	0
	40175wG6_BSR6R_filtered	100	1.33	0
40175wG7_BSR30		100	100	12.5
	40175wG7_BSR30_1_filtered	97.93	2.42	83.33
	40175wG7_BSR30_2_filtered	99.31	1.38	0
40175wG8_BS36		100	0.5	0
	40175wG8_BS36_filtered	100	0.5	0
40175wH1_BS26		100	0	0
	40175wH1_BS26_filtered	100	0	0
40175wH2_BS11		100	1.8	9.09
	40175wH2_BS11_filtered	100	0	0
40175wH3_BS39		100	144.47	0
	40175wH3_BS39_filtered	99.06	2.36	0
40175wH4_BS47		100	173.54	0.68
	40175wH4_BS47_1_filtered	99.38	1.4	25
	40175wH4_BS47_2_filtered	99.18	0	0
40175wH5_BS62		100	0	0

	40175wH5_BS62_filtered	100	0	0
40175wH6_BSR22		100	100.38	5.17
	40175wH6_BSR22_1_filtered *(R)*	44.83	0	0
	40175wH6_BSR22_2_filtered	98.68	0.41	33.33
	40175wH6_BSR22_3_filtered *(R)*	12.5	0	0
40175wH7_BSR31		100	261.53	7.77
	40175wH7_BSR31_1_filtered	98.97	0.41	33.33
	40175wH7_BSR31_2_filtered	99.31	1.38	0
	40175wH7_BSR31_3_filtered	97.98	1.08	0

Supplementary Table S3.2: Comparing the assembly metrics for each culture-derived genome, when assembled by myself and by MicrobesNG.

	My assembly					MicrobesNG assembly				
Genome	# contigs	Largest contig	Total length	GC (%)	N50	# contigs	Largest contig	Total length	GC (%)	N50
40175wA1_BS2	13	1449505	1846557	37.16	1449505	9	1446676	1845865	37.16	1446676
40175wA2_BS4	9	1454967	1925183	36.96	1454967	10	908424	1925083	36.96	546443
40175wA3_BS21	1586	36507	2774682	37.16	2916	1648	36507	2689247	37.14	2337
40175wA4_BSR38	16	1108587	3330676	28.35	929980	21	1107445	3330979	28.35	681197
40175wA5_BS48	24	481135	2326003	37.61	158593	28	332964	2326822	37.61	148864
40175wA6_BS64	14	1378230	1954322	36.81	1378230	14	1378230	1954277	36.81	1378230
40175wA7_BSR23	1252	1822053	8195526	36.88	40928	1523	1821094	8148852	36.83	28787
40175wA8_BSR32	81	409185	5173271	48.31	192066	83	409185	5174830	48.31	192066
40175wB1_BS5	192	124415	2090401	36.95	49067	995	37430	2339449	37.05	4852
40175wB2_BSR40	3247	505865	14803545	48.51	28553	3207	389277	14688447	48.5	41313
40175wB3_BS17	1563	125769	4895146	43.88	12078	1482	186974	4784250	43.96	12391
40175wB4_BS41R	143	323762	5061219	42.76	115103	141	324999	5065011	42.78	115685

	My assembly					MicrobesNG assembly				
Genome	# contigs	Largest contig	Total length	GC (%)	N50	# contigs	Largest contig	Total length	GC (%)	N50
40175wB5_BS49	197	481172	4274193	38.73	56734	253	333001	4252295	38.73	49941
40175wB6_BS66	53	327289	2729837	31.5	155652	59	324999	2729994	31.5	151162
40175wB7_BSR24	58	505224	5253074	48.3	227926	60	740152	5255398	48.29	257822
40175wB8_BSR33	55	310626	2720826	28.61	119967	54	310626	2720850	28.61	127764
40175wC1_BS22	8	878350	1825431	37.2	554753	10	728345	1825144	37.2	554753
40175wC2_BS30R	36	606459	3929577	34.15	269412	36	606459	3929626	34.15	269412
40175wC3_BS20	8	1446982	1844556	37.26	1446982	7	1446941	1844414	37.26	1446941
40175wC4_BS42	1024	1447448	3172653	38.08	172142	1079	1447905	3138835	38.06	172142
40175wC5_BS50	19	2005516	3364841	28.08	2005516	16	2083128	3364372	28.07	2083128
40175wC6_BS69	254	2005516	3537223	29.69	2005516	249	2005516	3530528	29.64	2005516
40175wC7_BSR25	69	744534	8239190	37.07	389154	86	744534	8238486	37.08	324788
40175wC8_BSR34	271	310626	7102741	39.12	58304	376	310626	7098621	39.12	42663

	My assembly					MicrobesNG assembly				
Genome	# contigs	Largest contig	Total length	GC (%)	N50	# contigs	Largest contig	Total length	GC (%)	N50
40175wD1_BS10	7	1418833	1814869	37.2	1418833	8	1159413	1814909	37.2	1159413
40175wD2_BS28	3366	917198	8038389	41.34	5764	3389	917198	7961456	41.32	5404
40175wD3_BS34	13	879936	1805883	32.81	242687	16	527991	1805732	32.81	242687
40175wD4_BS43	7	1485475	1880458	37.1	1485475	8	1485475	1880358	37.1	1485475
40175wD5_BS51	23	1824546	3363813	28.07	1824546	17	1825648	3363180	28.07	1825648
40175wD6_BSR1	7	1485197	1880212	37.1	1485197	8	1485197	1880112	37.1	1485197
40175wD7_BSR26	1420	1822193	8826811	38.84	22303	1661	1296608	8786433	38.81	16563
40175wD8_BSR35	16	2037547	3331233	28.36	2037547	18	2037611	3332074	28.36	2037611
40175wE1_BS8	54	1643783	3853438	27.84	423938	57	1725058	3859182	27.83	424038
40175wE2_BS31	36	331378	1965929	46.74	179076	36	331378	1964649	46.74	175635
40175wE3_BS32	36	332387	1963586	46.71	179478	37	332387	1961976	46.71	133225
40175wE4_BS44	7	1446945	1841928	37.16	1446945	8	1447133	1841839	37.16	1447133

	My assembly					MicrobesNG assembly				
Genome	# contigs	Largest contig	Total length	GC (%)	N50	# contigs	Largest contig	Total length	GC (%)	N50
40175wE5_BS54	7	1447135	1841941	37.16	1447135	8	1446678	1841593	37.16	1446678
40175wE7_BSR27	194	511352	9304035	39.48	137302	209	747154	9295364	39.5	113287
40175wE8_BSR37	16	2037609	3331390	28.36	2037609	14	2037609	3330986	28.35	2037609
40175wF1_BS7	6	1452267	1844287	37.26	1452267	9	1174430	1844642	37.26	1174430
40175wF2_BS19	9	1454531	1848605	37.14	1454531	8	1454740	1848230	37.14	1454740
40175wF3_BS52	1770	1746424	6134698	34.67	77034	1968	2055496	6111188	34.63	155292
40175wF4_BS45	344	146492	2619180	39.07	59560	349	134688	2616013	39.04	55550
40175wF5_BS57	971	77545	3129330	55.06	8860	975	70895	3087584	55.02	6720
40175wF6_BSX	121	153677	2406885	37.28	43488	124	153541	2405879	37.26	42379
40175wF7_BSR28	76	505458	5176139	48.31	256994	72	659371	5176186	48.31	266023
40175wF8_BSR3	8	877492	1824899	37.2	555076	9	878482	1825830	37.2	554462
40175wG1_BS27	8	1448122	1846472	37.14	1448122	8	1448122	1846472	37.15	1448122
40175wG2_BS13	7	1447133	1841939	37.16	1447133	8	1447133	1841839	37.16	1447133

	My assembly					MicrobesNG assembly				
Genome	# contigs	Largest contig	Total length	GC (%)	N50	# contigs	Largest contig	Total length	GC (%)	N50
40175wG3_BS35	22	733844	2208936	60.19	481331	26	630609	2208841	60.19	261398
40175wG4_BS46	90	148740	2414307	37.37	70474	87	148740	2418326	37.38	85566
40175wG5_BS61	58	327549	2722524	31.48	151162	57	323762	2720952	31.47	155488
40175wG6_BSR6R	42	585251	4772632	35.94	362820	49	571495	4771798	35.94	235986
40175wG7_BSR30	148	1395278	6657023	28.17	322510	69	2249310	6650527	28.17	425826
40175wG8_BS36	7	1485198	1880215	37.1	1485198	8	1485467	1880352	37.1	1485467
40175wH1_BS26	8	1458583	1840356	37.13	1458583	8	1458961	1840356	37.13	1458961
40175wH2_BS11	733	1447133	2483965	44.14	1447133	744	1446676	2474478	44.08	1446676
40175wH3_BS39	2040	134688	4791220	42.51	8405	2107	134688	4755260	42.47	9499
40175wH4_BS47	1666	313220	6471179	51.66	56274	1811	313220	6442647	51.65	48641
40175wH5_BS62	13	705484	1838555	37.19	544820	15	590558	1838420	37.19	544820
40175wH6_BSR22	152	442533	9340021	39.46	192064	95	616455	9330779	39.45	312127
40175wH7_BSR31	2086	545608	13965035	38.23	127975	2206	545608	13935812	38.22	87184

Supplementary Table S3.3: Taxonomy of the ruminal culture-derived isolate genomes. The taxonomy of each isolate as classified by GTDB-tk is shown at the phylum, family, genus and species level. Those marked with (*) are genome assemblies that did not cluster with any genomes from the Hungate collection.

Genome	Phylum	Family	Genus	Species
40175wA1_BS2_filtered	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
40175wA2_BS4_filtered	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
40175wA4_BSR38_filtered (*)	p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_P	s__Clostridium_P perfringens
40175wA5_BS48_filtered	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus gallolyticus
40175wA6_BS64_filtered	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus sp002393675
40175wA7_BSR23_1_filtered (*)	p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_F	s__Clostridium_F botulinum_A
40175wA7_BSR23_2_filtered	p__Firmicutes	f__Bacillaceae	g__Bacillus	s__Bacillus paralicheniformis
40175wA7_BSR26_1_filtered (*)	p__Firmicutes	f__Paenibacillaceae	g__Paenibacillus_A	s__Paenibacillus_A aceti
40175wA7_BSR26_2_filtered (*)	p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_F	s__Clostridium_F botulinum_A
40175wA8_BSR32_filtered (*)	p__Firmicutes	f__Paenibacillaceae	g__Paenibacillus_A	s__Paenibacillus_A aceti
40175wB1_BS5_filtered	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
40175wB2_BSR40_1_filtered	p__Firmicutes	f__Bacillaceae	g__Bacillus	s__Bacillus licheniformis
40175wB2_BSR40_2_filtered	p__Firmicutes	f__Bacillaceae	g__Bacillus	s__Bacillus paralicheniformis
40175wB2_BSR40_4_filtered (*)	p__Firmicutes	f__Paenibacillaceae	g__Paenibacillus_A	s__Paenibacillus_A macerans
40175wB3_BS17_1_filtered	p__Firmicutes	f__Lactobacillaceae	g__Limosilactobacillus	s__Limosilactobacillus mucosae

Genome	Phylum	Family	Genus	Species
40175wB3_BS17_2_filtered	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
40175wB4_BS41R_1_filtered	p__Firmicutes_A	f__Lachnospiraceae	g__Lachnobacterium	s__Lachnobacterium bovis
40175wB4_BS41R_2_filtered	p__Firmicutes_C	f__Acidaminococcaceae	g__Acidaminococcus	s__Acidaminococcus fermentans
40175wB5_BS49_1_filtered	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus gallolyticus
40175wB5_BS49_2_filtered (*)	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__
40175wB6_BS66_filtered	p__Firmicutes_A	f__Lachnospiraceae	g__Lachnobacterium	s__Lachnobacterium bovis
40175wB7_BSR24_filtered (*)	p__Firmicutes	f__Paenibacillaceae	g__Paenibacillus_A	s__Paenibacillus_A aceti
40175wB8_BSR33_filtered (*)	p__Firmicutes_A	f__Clostridiaceae	g__Clostridium	s__Clostridium isatidis
40175wC1_BS22_filtered	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
40175wC2_BS30R_filtered (*)	p__Firmicutes_A	f__Lachnospiraceae	g__Lachnotalea	s__Lachnotalea soehngenii
40175wC3_BS20_filtered	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
40175wC4_BS42_filtered	p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
40175wC5_BS50_filtered (*)	p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_P	s__Clostridium_P perfringens
40175wC6_BS69_filtered (*)	p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_P	s__Clostridium_P perfringens
40175wC7_BSR25_1_filtered	p__Firmicutes	f__Bacillaceae	g__Bacillus	s__Bacillus licheniformis
40175wC7_BSR25_2_filtered (*)	p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_F	s__Clostridium_F sporogenes

Phylum	Family	Genus	Species
p__Firmicutes	f__Bacillaceae	g__Bacillus	s__Bacillus licheniformis
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium	s__Clostridium isatidis
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes_C	f__Acidaminococcaceae	g__Acidaminococcus	s__Acidaminococcus fermentans
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes	f__Lactobacillaceae	g__Ligilactobacillus	s__Ligilactobacillus salivarius
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_P	s__Clostridium_P perfringens
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_P	s__Clostridium_P perfringens
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_P	s__Clostridium_P perfringens
p__Firmicutes	f__Lactobacillaceae	g__Limosilactobacillus	s__Limosilactobacillus mucosae
p__Firmicutes	f__Lactobacillaceae	g__Limosilactobacillus	s__Limosilactobacillus mucosae
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_F	s__Clostridium_F botulinum_A

Phylum	Family	Genus	Species
p__Firmicutes	f__Paenibacillaceae	g__Paenibacillus_A	s__Paenibacillus_A aceti
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_P	s__Clostridium_P perfringens
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_P	s__Clostridium_P perfringens
p__Firmicutes	f__Lactobacillaceae	g__Limosilactobacillus	s__Limosilactobacillus mucosae
p__Firmicutes	f__Erysipelatoclostridiaceae	g__Sharpea	s__Sharpea azabuensis
p__Firmicutes_C	f__Acidaminococcaceae	g__Acidaminococcus	s__Acidaminococcus fermentans
p__Firmicutes	f__Erysipelatoclostridiaceae	g__Sharpea	s__Sharpea azabuensis
p__Firmicutes	f__Paenibacillaceae	g__Paenibacillus_A	s__Paenibacillus_A aceti
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Actinobacteriota	f__Bifidobacteriaceae	g__Bifidobacterium	s__Bifidobacterium merycicum
p__Firmicutes	f__Erysipelatoclostridiaceae	g__Sharpea	s__Sharpea azabuensis
p__Firmicutes_A	f__Lachnospiraceae	g__Lachnobacterium	s__Lachnobacterium bovis

Phylum	Family	Genus	Species
p__Firmicutes	f__Amphibacillaceae	g__Virgibacillus	s__
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_G	s__Clostridium_G cochlearium
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_F	s__Clostridium_F sporogenes
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes	f__Erysipelatoclostridiaceae	g__Sharpea	s__Sharpea azabuensis
p__Firmicutes_C	f__Acidaminococcaceae	g__Acidaminococcus	s__Acidaminococcus fermentans
p__Firmicutes	f__Lactobacillaceae	g__Limosilactobacillus	s__Limosilactobacillus mucosae
p__Firmicutes	f__Streptococcaceae	g__Streptococcus	s__Streptococcus equinus
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_F	s__Clostridium_F botulinum_A
p__Firmicutes	f__Paenibacillaceae	g__Paenibacillus_A	s__Paenibacillus_A aceti
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_F	s__Clostridium_F botulinum_A
p__Firmicutes	f__Paenibacillaceae	g__Paenibacillus_A	s__Paenibacillus_A aceti
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium_F	s__Clostridium_F sporogenes
p__Firmicutes_A	f__Clostridiaceae	g__Clostridium	s__Clostridium isatidis

Supplementary Table S3.4: Assembly metrics for the rumen metagenome-derived genome bins. Genome bins that had >80% completeness and <1% contamination as determined by CheckM. Shown too are the genome metrics for each assembly, such as genome size and contig information.

Bin Id	Completeness	Contamination	Strain heterogeneity	Genome size (bp)	# contigs	N50 (contigs)	GC	# predicted genes
metabat_S1_2013.130	98.39	0	0	2058880	24	148521	33.8	1728
metabat_S1_2013.77	98.39	0	0	2095163	46	78225	36.7	1829
metabat_S1_2013.82	98.39	0	0	1714914	132	20228	64.2	1508
metabat_S1_2013.44	98.04	0	0	2566669	35	92197	44	2378
metabat_S1_2013.76	96.93	0	0	2366212	89	43432	55.7	2133
metabat_S1_2013.128	91.91	0	0	2322543	89	44954	54.9	1976
metabat_S2_2017.103	93.45	0.06	100	1557390	138	16325	58.3	1508
metabat_S1_2013.67	95.57	0.16	0	2196974	130	24168	43.9	2038
metabat_S2_2017.60	94.06	0.22	100	2197358	73	41073	47.3	1865
metabat_S1_2013.19	90.1	0.32	0	2058151	225	12170	59.4	1933
metabat_S1_2013.37	89.53	0.34	100	2586516	113	34527	47	2258
metabat_S1_2013.5	95.02	0.35	50	2803020	120	37205	58.6	2558
metabat_S2_2017.10	87.18	0.37	0	2800555	66	58367	51.4	2259
metabat_S1_2013.114	87.24	0.4	50	2171665	96	29709	48.2	1877
metabat_S1_2013.36	95.02	0.43	0	3516882	60	93855	50	2899

metabat_S1_2013.96	94.11	0.48	100	1786462	81	33611	44	1609
metabat_S1_2013.29	91.88	0.51	100	2170677	94	39573	57.4	1994
metabat_S2_2017.6	88.01	0.54	50	2265737	78	50035	49.5	1930
metabat_S1_2013.66	89.4	0.56	0	2593606	55	78686	48.4	2120
metabat_S1_2013.100	83.92	0.63	0	2107974	240	11803	53.9	1970
metabat_S1_2013.120	96.14	0.72	100	2206264	83	43275	38.3	1985
metabat_S2_2017.57	87.18	0.79	25	2060753	214	12545	53.9	1827
metabat_S1_2013.144	85.75	0.8	0	1892859	209	11465	34.5	1680
metabat_S2_2017.96	85.19	0.89	0	2741906	35	15218	42.1	2345
metabat_S1_2013.56	89.22	0.9	40	2528610	223	16250	64.3	2323
metabat_S1_2013.102	96.43	0.95	0	2755728	97	44799	48.8	2190
metabat_S1_2013.69	95.48	0.95	50	2556529	22	150757	55.5	2113
metabat_S1_2013.131	99.52	0.98	50	2854373	83	58009	41.4	2611
metabat_S2_2017.46	95.08	1.01	33.33	1837907	212	10614	33.9	1781
metabat_S1_2013.139	81.87	1.1	0	2620071	261	13463	51.1	2270
metabat_S2_2017.102	92.24	1.12	50	3181361	79	66482	47.1	2708
metabat_S1_2013.107	85.39	1.12	0	1107668	101	14182	26.6	1184
metabat_S1_2013.79	98.78	1.15	0	2615577	108	39856	57	2435
metabat_S2_2017.17	96.58	1.23	0	2498797	141	26884	33.3	2114
metabat_S2_2017.38	93.48	1.34	0	2413127	225	16454	44.6	2258
metabat_S2_2017.116	88.59	1.34	0	1671152	164	14484	58.3	1582
metabat_S2_2017.48	92.17	1.39	0	3557610	117	47521	45.9	2727
metabat_S1_2013.41	95.01	1.43	14.29	2754498	73	60122	47.4	2151

metabat_S1_2013.53	94	1.46	50	2006244	126	22895	52.8	1788
metabat_S2_2017.106	96.47	1.47	0	1895508	142	18481	39.4	1650
metabat_S1_2013.8	95.75	1.48	0	3020714	67	109604	48.6	2685
metabat_S1_2013.21	95.58	1.48	66.67	2292139	84	51959	60.4	2053
metabat_S2_2017.47	85.48	1.52	0	2858373	167	30629	50.1	2390
metabat_S1_2013.108	92.39	1.73	33.33	2788008	89	53924	54.4	2146
metabat_S2_2017.119	81.89	1.73	28.57	2173312	133	21585	49	1901
metabat_S1_2013.97	80.15	1.88	50	2118875	251	12403	59.2	2026
metabat_S1_2013.104	92.04	1.9	16.67	1605993	106	25581	62.3	1584
metabat_S2_2017.77	83.88	1.98	37.5	2515413	167	23442	50.5	2058
metabat_S1_2013.142	91.51	1.99	0	2291918	242	14120	53	1979
metabat_S1_2013.10	84.93	2.12	27.78	2720914	197	20006	46.9	2343
metabat_S1_2013.135	98.23	2.13	33.33	1788339	126	20207	56.1	1558
metabat_S2_2017.86	94.29	2.14	28.57	2422298	104	39725	58.2	2082
metabat_S1_2013.93	97.28	2.22	0	2454474	87	42976	50.2	2309
metabat_S2_2017.120	87.83	2.26	66.67	2741402	179	20725	52.3	2447
metabat_S1_2013.140	81.43	2.56	0	1725353	112	23327	33.9	1743
metabat_S1_2013.51	80.7	2.63	0	1901166	78	34234	59.2	1729
metabat_S1_2013.4	95.83	2.65	0	3160309	150	36737	51	2550
metabat_S2_2017.18	83.36	2.69	20	2008926	105	31453	54.6	1850
metabat_S2_2017.58	85.74	2.86	0	1919724	121	21167	58.5	1641
metabat_S1_2013.88	91.68	2.94	0	3525858	273	20695	46.4	3057
metabat_S2_2017.12	86.98	3.02	11.11	3176661	304	16312	57.7	2920

metabat_S1_2013.65	84.33	3.04	13.33	2048177	185	14728	59.1	1970
metabat_S1_2013.87	88.68	3.24	13.64	3438288	347	13720	51.3	2880
metabat_S1_2013.52	96.1	3.52	33.33	3161823	154	27933	62.2	2712
metabat_S2_2017.45	92.77	3.97	9.09	2522546	131	25193	51.3	2137
metabat_S1_2013.2	81.39	6.23	0	1927090	290	7910	47.4	1990
metabat_S1_2013.31	90.24	6.27	23.26	3220033	215	23402	50.1	2833
metabat_S1_2013.48	88.89	6.84	0	2092895	57	75483	34.5	1935
metabat_S1_2013.16	92.17	7.47	46.67	2548694	262	12962	47	2447

Supplementary Table S3.5: The taxonomy of each MAG according to GTDB shown at the phylum, family, genus, and species levels.

Genome	Phylum	Family	Genus	Species
metabat_S1_2013.144	p__Methanobacteriota	f__Methanobacteriaceae	g__Methanobrevibacter	s__Methanobrevibacter sp900314635
metabat_S1_2013.10	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__Prevotella sp900315765
metabat_S1_2013.100	p__Firmicutes_A	f__Lachnospiraceae	g__UBA2856	s__UBA2856 sp900319065
metabat_S1_2013.102	p__Bacteroidota	f__UBA932	g__RC9	s__RC9 sp900316045
metabat_S1_2013.104	p__Firmicutes_A	f__Oscillospiraceae	g__ER4	s__ER4 sp900317525
metabat_S1_2013.105	p__Firmicutes_C	f__Selenomonadaceae	g__Selenomonas_C	s__Selenomonas_C bovis
metabat_S1_2013.107	p__Firmicutes	f__UBA660	g__RUG705	s__RUG705 sp902788975
metabat_S1_2013.108	p__Bacteroidota	f__Bacteroidaceae	g__UBA3839	s__UBA3839 sp900319385
metabat_S1_2013.114	p__Bacteroidota	f__Paludibacteraceae	g__RF16	s__RF16 sp900321605
metabat_S1_2013.120	p__Firmicutes_A	f__Lachnospiraceae	g__RUG191	s__RUG191 sp900316395
metabat_S1_2013.128	p__Firmicutes_A	f__Lachnospiraceae	g__RUG306	s__RUG306 sp900316075
metabat_S1_2013.130	p__Bacteroidota	f__P3	g__Phil12	s__Phil12 sp900314725
metabat_S1_2013.131	p__Firmicutes_A	f__Lachnospiraceae	g__Agathobacter	s__Agathobacter sp900317585
metabat_S1_2013.135	p__Firmicutes_A	f__Anaerovoracaceae	g__Eubacterium_T	s__Eubacterium_T pyruvativorans
metabat_S1_2013.139	p__Fibrobacterota	f__Fibrobacteraceae	g__Fibrobacter	s__Fibrobacter sp900142495
metabat_S1_2013.140	p__Cyanobacteria	f__Gastranaerophilaceae	g__UBA2813	s__UBA2813 sp902774905
metabat_S1_2013.142	p__Firmicutes_A	f__Lachnospiraceae	g__CAG-791	s__CAG-791 sp900317475

Genome	Phylum	Family	Genus	Species
metabat_S1_2013.16	p__Firmicutes_A	f__Saccharofermentanaceae	g__Saccharofermentans	s__
metabat_S1_2013.19	p__Firmicutes_A	f__Lachnospiraceae	g__UBA1066	s__UBA1066 sp900317045
metabat_S1_2013.2	p__Firmicutes_A	f__Lachnospiraceae	g__Agathobacter	s__Agathobacter sp900316805
metabat_S1_2013.21	p__Firmicutes_A	f__Oscillospiraceae	g__CAG-110	s__CAG-110 sp900315595
metabat_S1_2013.29	p__Firmicutes_C	f__Selenomonadaceae	g__UBA2897	s__UBA2897 sp900315155
metabat_S1_2013.31	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__Prevotella sp900314995
metabat_S1_2013.36	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__Prevotella sp900314655
metabat_S1_2013.37	p__Firmicutes_A	f__Ruminococcaceae	g__Ruminococcus	s__Ruminococcus flavefaciens_E
metabat_S1_2013.4	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__Prevotella sp900315955
metabat_S1_2013.41	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__Prevotella sp900315635
metabat_S1_2013.44	p__Firmicutes_A	f__Lachnospiraceae	g__Catonella	s__Catonella sp900314885
metabat_S1_2013.48	p__Cyanobacteria	f__Gastranaerophilaceae	g__UBA1221	s__UBA1221 sp900320675
metabat_S1_2013.5	p__Firmicutes_C	f__Selenomonadaceae	g__Selenomonas_C	s__Selenomonas_C sp900314825
metabat_S1_2013.51	p__Firmicutes_C	f__Acidaminococcaceae	g__Succiniclasticum	s__Succiniclasticum sp900316935
metabat_S1_2013.52	p__Firmicutes_A	f__CAG-74	g__UBA2862	s__UBA2862 sp900315585
metabat_S1_2013.53	p__Firmicutes_A	f__Lachnospiraceae	g__CAG-791	s__CAG-791 sp900317555
metabat_S1_2013.56	p__Actinobacteriota	f__Atopobiaceae	g__UBA7741	s__UBA7741 sp900314495

Genome	Phylum	Family	Genus	Species
metabat_S1_2013.65	p__Firmicutes_C	f__Selenomonadaceae	g__UBA2897	s__UBA2897 sp900316275
metabat_S1_2013.66	p__Bacteroidota	f__Bacteroidaceae	g__UBA4334	s__UBA4334 sp900318775
metabat_S1_2013.67	p__Firmicutes_A	f__Lachnospiraceae	g__Faecalimonas	s__Faecalimonas sp900316755
metabat_S1_2013.69	p__Bacteroidota	f__UBA932	g__RC9	s__RC9 sp900321245
metabat_S1_2013.76	p__Firmicutes_C	f__Acidaminococcaceae	g__Acidaminococcus	s__Acidaminococcus sp900315205
metabat_S1_2013.77	p__Bacteroidota	f__P3	g__Phil12	s__Phil12 sp900315335
metabat_S1_2013.79	p__Spirochaetota	f__Sphaerochaetaceae	g__RUG023	s__RUG023 sp900315435
metabat_S1_2013.8	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__Prevotella sp900319305
metabat_S1_2013.82	p__Actinobacteriota	f__Eggerthellaceae	g__RUG013	s__RUG013 sp001486445
metabat_S1_2013.87	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__Prevotella sp900319905
metabat_S1_2013.88	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__Prevotella sp900316015
metabat_S1_2013.93	p__Firmicutes	f__Erysipelotrichaceae	g__Bulleidia	s__Bulleidia massiliensis_B
metabat_S1_2013.96	p__Firmicutes_A	f__Lachnospiraceae	g__UBA629	s__UBA629 sp900316665
metabat_S1_2013.97	p__Firmicutes_C	f__Selenomonadaceae	g__Selenomonas_C	s__Selenomonas_C sp900315575
metabat_S2_2017.10	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__
metabat_S2_2017.102	p__Bacteroidota	f__Bacteroidaceae	g__UBA4334	s__UBA4334 sp900316505
metabat_S2_2017.103	p__Firmicutes_A	f__Oscillospiraceae	g__NK3B98	s__NK3B98 sp900314485
metabat_S2_2017.106	p__Proteobacteria	f__Succinivibrionaceae	g__Succinivibrio	s__Succinivibrio sp900316175

Genome	Phylum	Family	Genus	Species
metabat_S2_2017.116	p__Firmicutes_A	f__Acutalibacteraceae	g__RUG420	s__RUG420 sp900317085
metabat_S2_2017.119	p__Bacteroidota	f__Bacteroidaceae	g__UBA1179	s__UBA1179 sp900318995
metabat_S2_2017.12	p__Firmicutes_A	f__Oscillospiraceae	g__UBA1777	s__UBA1777 sp900317375
metabat_S2_2017.120	p__Fibrobacterota	f__Fibrobacteraceae	g__Fibrobacter	s__Fibrobacter sp900313675
metabat_S2_2017.17	p__Firmicutes_A	f__Lachnospiraceae	g__Eubacterium_S	s__
metabat_S2_2017.18	p__Bacteroidota	f__P3	g__UBA1711	s__UBA1711 sp902776365
metabat_S2_2017.38	p__Firmicutes_A	f__Lachnospiraceae	g__UBA2942	s__UBA2942 sp900321525
metabat_S2_2017.45	p__Bacteroidota	f__UBA932	g__RC9	s__RC9 sp902785575
metabat_S2_2017.46	p__Firmicutes_A	f__Acutalibacteraceae	g__Ruminococcus_E	s__Ruminococcus_E sp900314705
metabat_S2_2017.47	p__Bacteroidota	f__F082	g__F082	s__F082 sp900318085
metabat_S2_2017.48	p__Bacteroidota	f__Bacteroidaceae	g__UBA3839	s__UBA3839 sp900313845
metabat_S2_2017.57	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__Prevotella sp900318625
metabat_S2_2017.58	p__Bacteroidota	f__P3	g__UBA1711	s__UBA1711 sp900314925
metabat_S2_2017.6	p__Bacteroidota	f__Paludibacteraceae	g__RF16	s__RF16 sp902765485
metabat_S2_2017.60	p__Bacteroidota	f__Bacteroidaceae	g__Prevotella	s__Prevotella sp900315775
metabat_S2_2017.77	p__Bacteroidota	f__Bacteroidaceae	g__UBA4372	s__UBA4372 sp900320565
metabat_S2_2017.86	p__Bacteroidota	f__UBA932	g__Bact-11	s__Bact-11 sp902789445
metabat_S2_2017.96	p__Bacteroidota	f__Bacteroidaceae	g__UBA4293	s__UBA4293 sp900316585

Supplementary Table S4. 1: Commands used to run the HUMAnN3 pipeline

The HUMAnN3 v3.0.0.alpha.4 Docker container was pulled using the command:
`docker run -ti -v /data/:/data/ biobakery/human`
The HUMAnN3 databases were downloaded, and the config updated, using the commands:

`humann_databases --download chocophlan full /data/humann3_repeated/databases/`
 	
`humann_databases --download uniref uniref90_diamond /data/humann3_repeated/databases/`
 	
`humann_databases --download utility_mapping full /data/humann3_repeated/databases/ --update-config yes`
 	
`humann_config --update database_folders nucleotide /data/humann3_repeated/databases/chocophlan/`
 	
`humann_config --update database_folders protein /data/humann3_repeated/databases/uniref/`

The HUMAnN3 pipeline was ran using the command:
`humann --input /data/simulated_data_hungate/hiseq_exponential_cat.fast`

```
q --nucleotide-database
/data/humann3_repeated/databases/chocophlan/ --protein-
database /data/humann3_repeated/databases/uniref/ --
output /data/humann3_repeated/paired_end_rumen/ >
/data/humann3_repeated/paired_end_rumen/humann_rumen.ou
t
```

Supplementary Table S4. 2: Commands used to run the Carnelian pipeline

The Carnelian Docker container was pulled using the command:
`docker run -ti -v /data/carnelian/:/data/ snazeen/carnelian`
The Carnelian model was trained using the command:
`carnelian train -k 8 --num_hash 4 -l 30 -c 5 ./EC-2010-DB/ model_directory`
The data was annotated by the Carnelian pipeline using the command:

`carnelian annotate -k 8 -n 4 /data/simulated_data_hungate/fasta/ /usr/local/src/carnelian/data/model_directory`
 	
`/data/carnelian/human/ /usr/local/src/carnelian/util/ext/FragGeneScan`

The functional profiling and abundance analysis was performed by Carnelian using the command:
`carnelian abundance /data/labels_dir/ /data/abundance_matrix_output_dir/ /data/sampleinfo_file.tab /data/EC-2010-DB/ec_lengths.tsv`