

*Effect of enteric parasites on the enteric
metagenome of bovine livestock*

Mumdooh J. Sabir BSc.

This thesis submitted for the degree of Masters by
research to the University of East Anglia

Faculty of Medicine and Health Sciences
University of East Anglia

May 2022

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived therefrom must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Declaration

The research described within this thesis titled ‘Effect of enteric parasites on the enteric metagenome of bovine livestock’ is, to the best of my knowledge, original and my own work, except where due reference has been made.



.....

Mumdooh J. Sabir

Acknowledgements

First of all, I would like to express my thanks and gratitude to my supervisors, Professor Neil Hall and Dr. Kevin Tyler for all the invaluable support and guidance they offered me during my study over the COVID-19 pandemic to do this research. Also, I owe a great debt to Professor Neil Hall who gave me the opportunity to benefit from the Earlham Institute at Norwich Bioscience Institutes. Without my supervisors' unlimited support and their patience, I would not have achieved this academic development.

I would also like to thank the Hall group at Earlham Institute; Neil Hall, Dr. Ross Low, Dr. Mark McMullan, Dr. Matt Brawn, Michelle Grey, Dr. Sally Warring, Helene Yvanne, Jessica Peers and Laura Bird who welcomed me in the group and supported me during these days which will not be forgotten during COVID-19 crisis in 2020. Special thanks to my fantastic mentor Dr. Ross Low who encouraged me to keep my enthusiasm and guided me from scratch. He also generously arranged for everything that I needed to complete this research. Thanks to Jessica Peers who walked me through and taught me about some laboratory basics. In addition, Thanks to Dr. Calum Raine for the controversial talks about colours and the coffee breaks.

I also would like to express my gratitude to my colleagues at King Abdulaziz University Dr. Majid Kamli, Dr. Irfan Rather, Dr. Ahmed Bahieldin and Dr. Salah Aboaba who observed my progress remotely and provided me with unlimited support.

I dedicate this thesis to my parents Prof. Jamal Sabir and Dr. Nahid Hajrah, my brother Meshaal Sabir, my sister Mernan Sabir my beloved wife Noor Khushaim and my sister-in-law Nahla Sindi. As they were the source of motivation, support and inspiration during my study and the difficult times amid the UK lockdowns due to COVID-19 pandemic emergence.

Abstract

Cryptosporidium is a protozoan parasite and it is the causative agent of cryptosporidiosis in cattle. Symptoms of the infection may include abdominal pain, vomiting and diarrhoea, however infections may also be asymptomatic. The severity of cryptosporidiosis infection depends on many factors, including host immunity which plays an important role in the outcome of infection. Agricultural losses globally due to *Cryptosporidium* infection in cattle amount to several billion dollars. Previous studies utilised genotyping techniques to identify and classify *Cryptosporidium* and relate this to virulence. Here, the impact of the gut microbiome on infection in cattle was studied. The gut microbiome is the community of bacteria resident in the digestive system. Ruminants including cattle have a stomach with four compartments specialised to pre-gastric digestion. It has previously been demonstrated that specific alterations in the bovine microbiome can facilitate growth of cryptosporidial parasites. Therefore, cryptosporidiosis can lead to long-term dysbiosis of the gut. It has been reported that infections with *C. parvum* showed changes in the microbiota as well as a shift in metabolites in the host gut. In this study, a bioinformatic pipeline based on shotgun metagenomic sequencing to detect *Cryptosporidium* infection was utilised employing computational analysis of the sequence obtained from faecal DNA to investigate the changes to the gut microbiome between infected and non-infected hosts. Crucial changes to the microbiome in cattle and humans resulting from *Cryptosporidium* infection were uncovered. The results show infection with *Cryptosporidium* decreases bacterial diversity in the gut microbiome of the host.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate 'take down' action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

Table of Contents

Acknowledgements	1
Abstract	2
Table of contents	3
Abbreviations	10
CHAPTER 1: Introduction	11
1.1 Overview	11
1.2 <i>Cryptosporidium</i> infection	13
1.3 The <i>Cryptosporidium</i> Life Cycle.	14
1.4 Genomics	16
1.5 The <i>Cryptosporidium</i> Genome	17
1.6 The Microbiome	19
1.7 Microbiome Analysis	19
1.8 Analysis of Microbiomes using NGS	24
1.9 Bovine Gut Microbiome	26
1.10 Metabarcoding and metagenomic studies of the bovine digestive system	28
1.11 The Gut Microbiome Changes in Response to Pathogen infection	34
1.12 The Role of Microbiome in Cryptosporidiosis	36
CHAPTER 2: Benchmarking methods for detecting <i>Cryptosporidium</i> using metagenomic shotgun methods	38

2.1 Introduction	38
2.1.1 Mapping methods	38
2.1.2 Using metagenomics to find pathogens	38
2.2 Material and methods	39
2.2.1 Genome sequencing	41
2.2.2 K-mer Density Analysis	42
2.2.3 Taxonomic classification and quantification	42
2.2.4 Mapping and and counting aligned <i>Cryptosporidium</i> reads	42
2.3 Results	44
2.3.1 K-mer Density Analysis	44
2.3.2 Identifying <i>Cryptosporidium</i> BWA vs. CENTRIFUGE	45
2.4 Discussion	47
CHAPTER 3: Analysis of the interaction of <i>Cryptosporidium</i> and the gut microbiome in humans and cattle	50
3.1 Introduction	50
3.1.1 The impact of infection on microbiomes	50
3.2 Methods	51
3.2.1 Samples used in this study	51
3.2.2 Taxonomic classification and quantification of humans and University of Kent cattle samples	54
3.2.3 Mapping and and counting aligned <i>Cryptosporidium</i> reads of humans and University of Kent cattle samples	54
3.2.4 analysis of pre existing dataset	55
3.3 Results	55
3.3.1 Identifying <i>Cryptosporidium</i> BWA vs. CENTRIFUGE for human and University of Kent samples	55

3.3.2 Analysis of the microbiota in <i>cryptosporidium</i> infected human samples	59
3.3.3 Analysis of the microbiota in <i>cryptosporidium</i> infected bovine samples	62
3.3.4 Analysis of the microbiota in Tanzanian pre existing dataset	65
3.3.5 The interaction between <i>Cryptosporidium</i> and <i>Fusobacterium</i>	69
3.4 Discussion	71
CHAPTER 4: General discussion	74
References	77
Appendices	92

List of Tables

Table 2.1 Spiked human faecal samples from King's College London. Number of <i>Cryptosporidium</i> parvum (IOWA) added to each 250 mg aliquot of faecal material. S1T to S6T were heat treated at 70oC for 5 min.	40
Table 2.2 Spiked rabbit samples from University of East Anglia. Number of <i>Cryptosporidium</i> parvum (IOWA) added to each 250 mg aliquot of faecal material. MMNC is the unspiked sample which serves as a negative control.	41
Table 2.3 Details of 7 <i>Cryptosporidium</i> reference genomes used to build a single database to be used in the mapping phase with BWA.	43
Table 2.4 Comparison between <i>Cryptosporidium</i> reads per million from BWA and CENTRIFUGE for both, King's College London spiked human samples with <i>Cryptosporidium</i> parvum (IOWA) added to each 250 mg aliquot of faecal material. S1T to S6T were heat treated at 70oC for 5 min from. And rabbit samples acquired from University of East Anglia, where they are spiked with <i>Cryptosporidium</i> parvum (IOWA) added to each 250 mg aliquot of faecal material. Both samples S0 and MMNC served as a negative control.	45
Table 3.1 Human samples acquired from Public Health Wales where all samples were tested positive for <i>Cryptosporidium</i> using a PCR test .	51
Table 3.2 Cattle samples acquired from University of Kent where 2 samples are negative and the rest are positive for <i>Cryptosporidium</i> by PCR test	52
Table 3.3 Tanzanian Cattle samples downloaded from MG-RAST based on their collection location with their total number of reads.	53
Table 3.4 Comparison between <i>Cryptosporidium</i> reads per million from BWA and CENTRIFUGE for human samples from Public Health Wales	55
Table 3.5 Comparison between <i>Cryptosporidium</i> reads per million from BWA and CENTRIFUGE for cattle samples from University of Kent	57

Table of Figures

Figure 1.1 Morphology of Apicomplexan cell showing cell wall membrane structure (Alveola) with compositions and other internal cell organelles (Nucleus, Golgi body, endoplasmic reticulum, and epicoplast). Adapted from Schnittger <i>et al.</i> , (2013).	12
Figure 1.2 Outline of the <i>Cryptosporidium parvum</i> life cycle [Adapted and modified from (Bones <i>et al</i> , 2019)]. Oocysts ingested release sporozoites, which infiltrate the ileum and grow into trophozoites and type I meronts, which include type I merozoites. Type I Merozoites can evolve into trophozoites or type II meronts, which release type II merozoites, which then mature into undifferentiated gamonts. Gamonts can be classified as macrogamonts or microgamonts, with the latter producing microgametes that fertilise macrogamonts. Sporulation takes place inside the host, releasing thick-walled oocysts into the environment as well as thin-walled oocysts that self-infect the same host organism.	15
Figure 1.3 An illustration showing various routes of oocyte transmission from animal to human in a cyclic form through environmental resources (the figure was designed using Mind the Graph paid version)	16
Figure 1.4 shows the 16S rRNA gene's conserved, variable and hypervariable sections, as well as the numerous primer pairs used for metagenomic sequencing. Blue represents conserved regions, grey represents variable regions and red represents hypervariable regions. Some hypervariable sections, such as H3 and H9 are longer than H5 or H7, so the nine hypervariable regions are not evenly distributed. Primers are constructed in a conserved area to produce a PCR product that spans one or more hypervariable regions (V4) (e.g., V1-2 or V3-5). Single-end (Roche 454) or double-end (Illumina) PCR product sequencing gives data for microbiota profiling (Adapted from Shahi <i>et al</i> , 2017).	20
Figure 1.5 Meta-Analysis of the 16S-rRNA Gut Microbiome Data. Where x-axis denotes differential abundance for the classes A) <i>Bifidobacterium</i> , B) <i>Lactobacillus</i> , C) <i>Snodgrassella</i> , and finally, D) (unknown genus) the order <i>Pasteurellales</i> . While y-axis represents the percentage abundance of microbiome components (Gkantiragas and Gabrielli, 2021).	21
Figure 1.6 Microbial composition of microbiome in Rumens (cattle) and human gut system. A) Composition, B) Similarity of Bovine Rumen Microbiota across Individual Animals (Adapted from Jami E, and Mizrahi I (2012). C) different regions of the human GI tract constitute different microorganisms, C) Composition of the human gut microbiome showing dominant bacterial communities and fungal or protozoan species covering a small fraction (Adapted from Srivastava <i>et al</i> , 2022).	32

Figure 1.7 An example of colonisation of the human gut by bovine bacteria present in Parmesan cheese showing the cross transfer of microbes from cattle to humans (Adapted from Milani <i>et al</i> , 2019).	34
Figure 2.1 Kat density plot comparing read file 1(x-axis) from sequencing lane against read 1 (y-axis) from another sequencing lane for samples (a) S0, (b) S1 and (c) S1T.	44
Figure 2.2 Number of <i>Cryptosporidium</i> read per million produced by both CENTRIFUGE (Black bars) and BWA (Gray bars) for human samples with no heat treatment (A), human samples with heat treatment (B) and rabbits samples (C)	46
Figure 2.3 setting a detection cutoff at 60 reads per million (dotted Red line) for <i>Cryptosporidium</i> reads given by CENTRIFUGE. Five samples exceed the cutoff, one human sample with no heat treatment (A), one human sample with heat treatment (B) and three rabbit samples (C)	47
Figure 2.4 A comparison between the number of <i>Cryptosporidium</i> reads per million and the numbers of oocysts spiked into human and rabbit samples.	48
Figure 3.1 Number of <i>Cryptosporidium</i> reads per million produced by both CENTRIFUGE with a detection cutoff (dotted red line) where two samples exceed the cutoff.	56
Figure 3.2 Number of <i>Cryptosporidium</i> read per million produce by both CENTRIFUGE and BWA with a detection cutoff (dotted red line) where only one sample exceeds the cutoff.	58
Figure 3.3 PCA analysis for both bacterial and eukaryotic components of infected bovine where it shows that samples UKUB17, UKCU18, UKH101, UKH102, UKH105 and S0 (control sample form King's College London) are clustering together. Whereas sample UKH104 is an outlier which suggests that there has been a change in its both eukaryotic and bacterial microbiota.	59
Figure 3.4 The bacterial component of human unspoked samples	60
Figure 3.5 Eukaryotic microbiota a changes associated with <i>Cryptosporidium</i> infection in human samples	61
Figure 3.6 PCA analysis for both bacterial and eukaryotic components of infected bovine	62
Figure 3.7 The bacterial component of European cattle samples	63
Figure 3.8 The Eukaryotic component of European cattle samples	64
Figure 3.9 <i>Cryptosporidium</i> reads for Tanzanian cattle produced by CENTRIFUGE	65
Figure 3.10 PCA analysis for both bacterial and eukaryotic components of Tanzanian cattle	66

Figure 3.11 Bacterial component in Tanzanian cattle from Rungwe region	67
Figure 3.12 Bacterial component in Tanzanian cattle from Lushoto region	67
Figure 3.13 Eukaryotic component in Tanzanian cattle from Rungwe region	68
Figure 3.14 Eukaryotic component in Tanzanian cattle from Lushoto region	68
Figure 3.15 Comparison between <i>Cryptosporidium</i> and <i>Fusobacterium</i> in samples of humans from Public Health Wales and cattle from University of East Anglia	69
Figure 3.16 Comparison between <i>Cryptosporidium</i> and <i>Fusobacterium</i> in Tanzanian cattle samples	70
Figure 3.17 Correlation between <i>Cryptosporidium</i> and <i>Fusobacterium</i> Tanzanian cattle samples.	70
Figure 3.18 Correlation between <i>Cryptosporidium</i> and <i>Fusobacterium</i> in infected humans and European cattle samples. Where samples (UKH104) from Public Health Wales infected humans and (N5) from the University of Kent infected cattle seem to be severely infected.	72
Figure 3.19 Correlation between <i>Cryptosporidium</i> and <i>Fusobacterium</i> in humans without sample UKH104 and European cattle samples without sample N5 .	72

Abbreviations

Abbreviation	Description
BLAST	Basic Local Alignment Search Tool
<i>C.</i>	<i>Cryptosporidium</i>
CH ₄	Methane
CO ₂	Carbon dioxide
CRT	Cyclic Reversible Termination
DNA	Deoxyribonucleic acid
ETE	The Environmental for Tree Exploration
GI	Gastrointestinal
H ₂	Hydrogen
H ₂ S	Hydrogen Sulfide
MG-RAST	Metagenomic Rapid Annotations using Subsystems Technology
NGS	Next Generation Sequencing
OTU	operational taxonomic unit
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
rDNA	Ribosomal DNA
RDP	Ribosomal Database Project
RFLP	Restriction Fragment Length Polymorphism
rRNA	Ribosomal ribonucleic acid
SBL	Sequencing by Ligation
SBS	Sequencing by Synthesis
SNA	Single-Nucleotide Addition
T-RFLP	Terminal Restriction Fragment Length Polymorphism

CHAPTER 1

Introduction

1.1 Overview

Cryptosporidium is a protozoan that was initially identified from gastric mucosa back in 1907. It is among parasites within an incredibly vast and diversified group of phylum Apicomplexa (Fayer *et al*, 2000; Tyzzer 1907). Other parasites which belongs to the phylum *Apicomplexa* are from genus *Plasmodium*, *Toxoplasma*, *Isospora*, *Eimeria*, *Sarcocystis*, *Theileria*, *Cyclospora*, and *Babesia*, which infect both humans and animals. The majority of them have a unique sort of organelle termed an apicoplast, which consists of a non-photosynthetic plastid and an apical complex structure. The apicomplexan uses the organelle as a means of penetrating a host cell (Frénal *et al.*, 2017). The general morphology or cell structure of *Apicomplexa* cell has been shown in Figure 1.1. In 1982 *Cryptosporidium* gained recognition as a human pathogen when the first case of Cryptosporidiosis was reported in Acquired Immune Deficiency Syndrome (AIDS) patients (Hunter and Nichols, 2002). The Center for Disease Control and Prevention has categorised *Cryptosporidium* as an emerging protozoan parasite where it is estimated that 748,000 cases in humans annually in the USA alone are due to Cryptosporidiosis (Scallan *et al*, 2011). In general residents, patients, school children, and the healthy population, the highest estimated prevalence of *Cryptosporidium* infection was found in Mexico (69.6 percent), Nigeria (34.0 percent), Bangladesh (42.5 percent), and the Republic of Korea (8.3 percent) respectively (Dong *et al.*, 2020). Although the number of cases of cryptosporidiosis in the UK has been dropping for several years, there are still roughly 4000 cases reported each year in England and Wales (McKerr *et al.*, 2019)

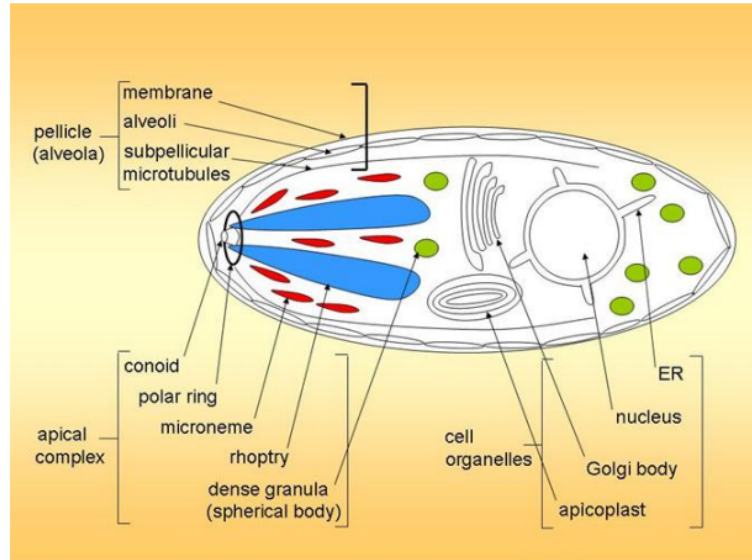


Figure 1.1 Morphology of an apicomplexan cell showing pellicular (alveola) structure, apical complex and other internal cell organelles (Nucleus, Golgi body, endoplasmic reticulum, and apicoplast). Adapted from Schnittger *et al.* (2013).

Currently, about 40 species of *Cryptosporidium* are established to infect vertebrates (Norman, 2017). *Cryptosporidium* infection can lead to gastroenteritis which affects the microvillus border of gastrointestinal epithelium in many vertebrate hosts and is of both medical and veterinary importance and transmission can be either zoonotic or anthroponotic (Nime *et al.*, 1976; Bouzid *et al.*, 2013).

The diversity and distribution of *Cryptosporidium* varies worldwide. For instance, *C. parvum* is the predominant infection for cattle from European countries (Lebbad *et al.*, 2021); while in India *C. ryanae* was identified as the most prevalent infecting species (Das *et al.*, 2019). *C. parvum* and *C. hominis* are the two species primarily responsible for cryptosporidiosis in humans, *C. hominis* is an anthroponosis spreading person to person but *C. parvum* is zoonotic and is the primary agent of cryptosporidiosis in livestock. Accordingly, *C. hominis* is often more reported from metropolitan areas and *C. parvum* more reported from rural areas (Janssen and Snowden, 2020). Although recently it has been observed that the subspecies *C. parvum anthroponosum* also transmits predominantly person to person and is also more prevalent in urban areas (Nader *et al.*, 2019). The virulence of the infection varies with the host species and other factors, including the age and immune status of the host

(Janssen and Snowden, 2020). In human populations, in Europe both *C. hominis* and *C. parvum* are widespread. In the Middle East, the most common type identified was *C. parvum* while *C. hominis* was detected only in one of the studies in Kuwait (Aresshi *et al*, 2007). Bovine parasitic diarrhea (scour) caused by *C. parvum* remains one of the most severe livestock diseases, especially in young ruminants, particularly neonatal calves. Among cattle, calves aged 14 to 21 days old are most affected by the *Cryptosporidium* and cryptosporidiosis is a significant concern for animal health causing economic losses due to the high cost of medicines and mortality. It is a worldwide problem faced by the animal farmers (Santin, 2013). In many farmsteads, *C. parvum* contamination can reach a stage where it is extremely difficult to treat effectively, and it remains one of the leading causes of scouring leading to growth delays, substantial morbidity and mortality (Wells and Thomson 2014; Thomson *et al*, 2017). In livestock, young calves under 6 weeks old are at a high risk of *C. parvum* infection (Silverlas *et al*. 2009; Wells and Thomson 2014). It was reported that in 2014, the economic costs of *Cryptosporidium* in Great Britain were likely to be between £100-£200 per infected calf (Shaw 2014). These costs mainly were associated with veterinary treatments, higher nutritional requirements and lower weight gain. Prevalence of *C. parvum* in stool samples of European cattle herds has been reported to range from 13-100% (Imre and Darabus 2011). The results reinforce cattle as the main reservoir for *C. parvum*, which can spread within the species or even to humans via different routes (Hunter and Thompson 2005; Wells and Thomson 2014; Brankston *et al*. 2018; Chalmers *et al*. 2019).

1.2 *Cryptosporidium* infection

Those suffering from Cryptosporidiosis symptoms may present with abdominal pain, fever, vomiting, malabsorption and diarrhea (Bouzid *et al*, 2013). However, Cryptosporidiosis causes gastrointestinal infection of variable severity. Immunity also plays an essential role in the severity of the disease and its prognosis. It has been reported that an immunocompetent person who is affected with Cryptosporidiosis can recover without treatment in two weeks suggesting an effective immune response mechanism (Shoultz *et al*, 2016). However, *Cryptosporidium*

can lead to severe and even fatal diarrhea (Pumipuntu and Piratae, 2018) in humans with weak immunity such as AIDS as well as poorly nourished children under the age of 5. It is responsible for a significant proportion of the deaths, particularly in the developing world (Shoultz *et al*, 2016). On the other hand, in many mammals, humans and livestock, *Cryptosporidium* infection can be asymptomatic or cause only mild, transient and self-resolving symptoms (Rayan *et al*, 2014).

1.3 The *Cryptosporidium* Life Cycle.

Cryptosporidium oocysts are excreted in feces and remain infective in the environment for very long periods due to the hard wall of oocysts which protect them against physical and chemical damage in the environment. Transmission is via ingestion of infective oocysts, which can survive several months in cool, moist conditions (Robertson and Gjerde, 2007). The *Cryptosporidium* oocyst wall is very resistant to chlorine disinfectant levels making it a threat to drinking water supplies globally (Carey *et al*, 2004). In cattle and humans cryptosporidiosis can be transmitted by a faecal/oral route and so the transmission can occur from a contact with an infected host, or for instance, from a contact with contaminated grass, feed or groundwater.

Cryptosporidium has a lifecycle involving both sexual and asexual replication which is completed in a single host (Bouزيد *et al*, 2013). Oocysts become activated in the stomach and small intestine discharging four infective sporozoites during excystation. These sporozoites bind to the surface of the intestinal epithelial cells in the small intestine, attaching, wrapping themselves in membrane and developing into meronts. Within the meronts, merozoites are formed which are then released from the cells able to then invade additional epithelial cells. Some merozoites though will enter cells and produce male and female gametocytes (microgamonts and macrogamonts), fusion of male microgametes and female macrogamonts gives rise to zygotes which develop into new oocysts (Funkhouser-Jones *et al*, 2020). The detailed life cycle of *Cryptosporidium* has been shown in Figure (1.2).

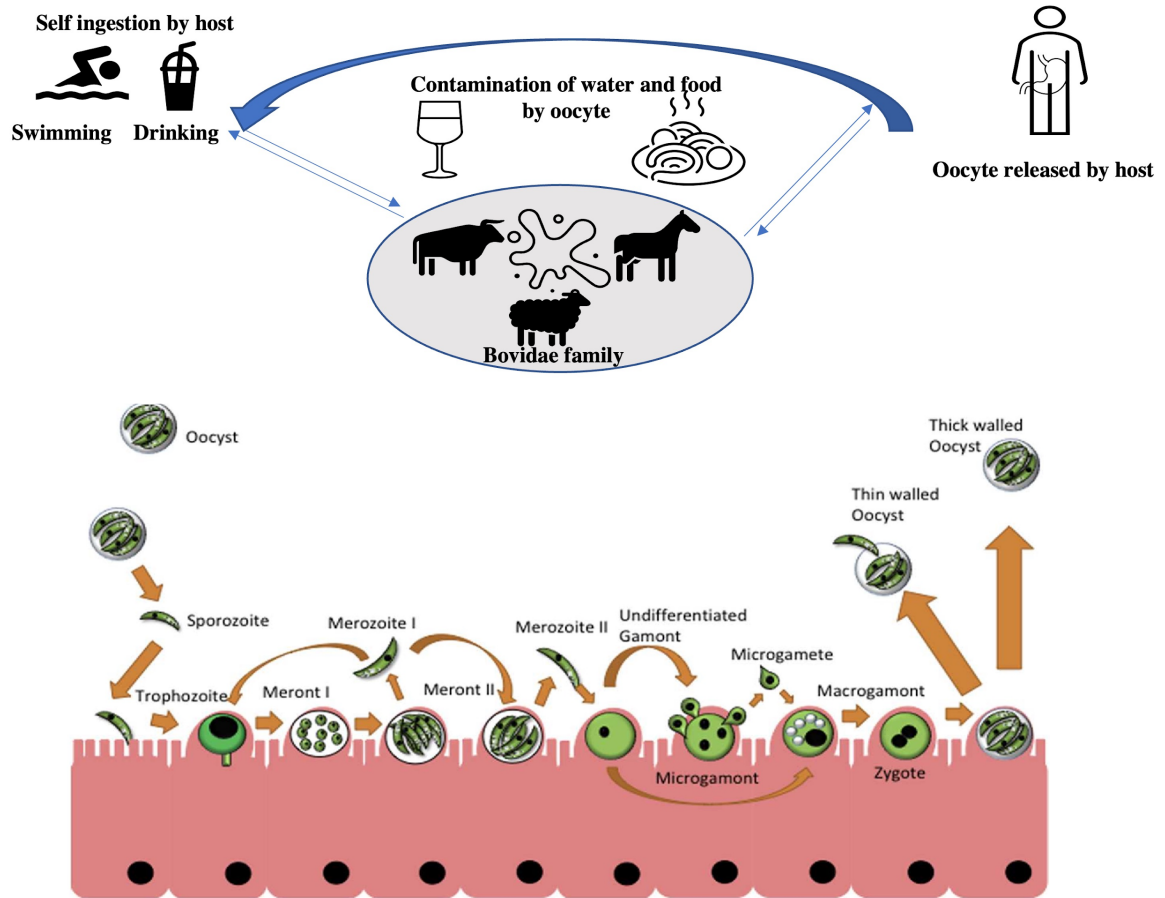


Figure 1.2 Outline of the *Cryptosporidium parvum* life cycle [Adapted and modified from (Bones *et al*, 2019)]. Oocysts ingested release sporozoites, which infiltrate the ileum and grow into trophozoites and type I meronts, which include type I merozoites. Type I Merozoites can evolve into trophozoites or type II meronts, which release type II merozoites, which then mature into undifferentiated gamonts. Gamonts can be classified as macrogamonts or microgamonts, with the latter producing microgametes that fertilise macrogamonts. Sporulation takes place inside the host, releasing thick-walled oocysts into the environment as well as thin-walled oocysts that self-infect the same host organism.

At this stage of the life cycle, the oocysts will be of two forms. The first, is the thick-walled, which will be excreted in the environment and can be found in dung. The second are thin-walled oocysts that are believed to release sporozoites into the small intestine which initiate a new round of infection (Pumipuntu and Piratae, 2018). The thick-walled oocysts which are released by the infected hosts will be ingested by new hosts eating contaminated food or drinking contaminated water and the life cycle will repeat (O'Donoghue, 1995). The

routes of transmission of oocyte through various environmental resources from animals to human and then back to animals is shown in Figure (1.3).

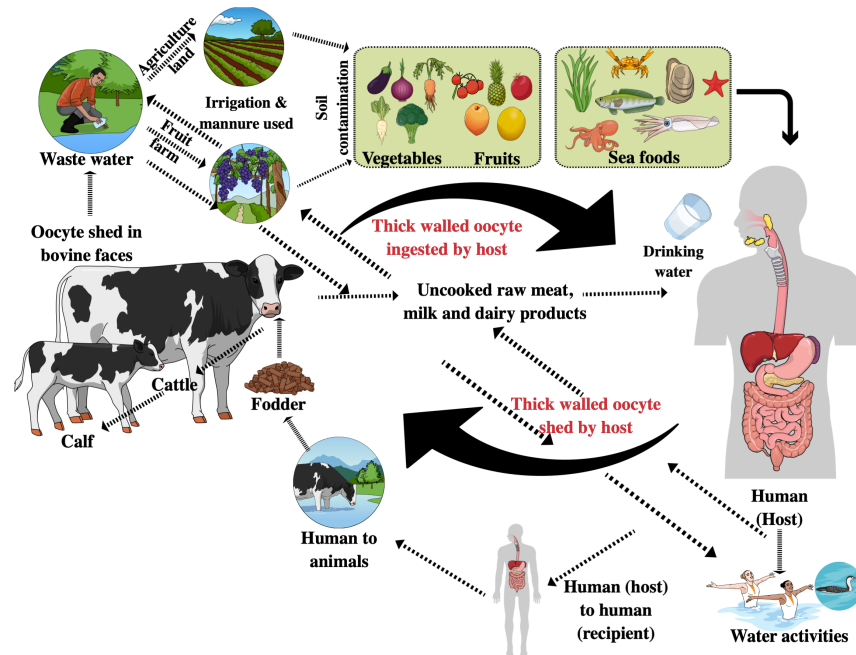


Figure 1.3 An illustration showing various routes of oocyte transmission from animal to human in a cyclic form through environmental resources (the figure was designed using Mind the Graph, paid version)

1.4 Genomics

DNA sequencing is defined as a process of extracting the sequence of nucleotides in a DNA molecule. Back in 1977, DNA sequencing technology came into existence with the development of Sanger's 'chain-termination' or dideoxy technique for sequencing technique (Sanger *et al.* 1977). And later, this technique was improved with fluorometric based detection and fragment separation through capillary-based electrophoresis. This technique was known as first generation DNA sequencing which was able to sequence only a short fragment of DNA up to around 1 kb. Later, this was improved by the shotgun sequencing technique in which overlapping DNA fragments were cloned and sequenced separately. Presently to get information about the whole genome, biologists use whole genome and next generation sequencing. This technology of Next Generation Sequencing (NGS) raises the possibility of

producing an enormous volume of data, about one billion short reads per run, dramatically increased throughput and substantially lowered expenditure. This feature of Next Generation Sequencing expands the realm of experimentation beyond determining the order of bases. Besides, sequencing the whole genome allows for studying precise large-scale comparisons of genomes and predicting pathways of evolution at a genomic level of detail and with certainty impossible to reach via earlier technologies. Further, resequencing the human genome via NGS enhanced the understanding of the link between genetic differences and human health disease (Metzker, 2010).

The most commonly used technology for Sequencing today is Illumina sequence technology, namely sequencing by synthesis (SBS) using cyclic reversible termination (CRT). This sequencing technology supports both single-read and paired-end libraries and offers a short-insert paired-end capability for high-resolution genome sequencing. The technology also combines short insert and longer reads to increase the ability to fully characterise large genomes. The termination of DNA synthesis after the incorporation of single nucleotides represents the most important feature of CRT and the use of a dideoxynucleotide, which is an adaption of the of Sanger method which used no reversible chain termination but provided the basis for the initial development of reversible blocking groups 3'-O-allyl-dNTPs and 3'-O-azidomethyl-dNTPs (Metzker *et al*, 1994) and (Canard and Sarfati, 1994). The Illumina method has an innovative DNA replication step called 'bridge amplification' that enables the DNA fragments to be amplified *in situ*, creating individual colonies of clonal DNA. This means that the sequencing progresses across thousands of fragments of DNA at the same time so that the fluorescent signal can be more easily read by the sequencing machine.

1.5 The *Cryptosporidium* Genome

Genomics is the field of biology defined as the study of genomes and their functions. A genome describes the complete DNA content of a cell or an organism. Genome sequencing of parasites (Warrenfeltz *et al*, 2020) began more than 20 years ago when a number of genome sequencing projects were initiated to understand the diseases and identify therapeutic targets.

Today many pathogenic parasites that are medically important have been subjected to whole genome sequencing (Forrester and Hall, 2014). These sequencing projects allowed the identification and discovery of genes involved in many protozoan diseases with a focus on drug resistance, drug targets and antigens that could be vaccine targets (Mu *et al*, 2007). Then next generation sequencing technologies opened new ways for parasite biologists to better understand the mechanism and functions of parasitic disease through the application of comparative genomics and functional genomics such as RNAseq (Talavera-Lopez and Andersson, 2017) . Thus, understanding the molecular mechanism by which pathology occurs by using genomics to reveal the parasite (and host) components is important (Khan *et al*, 2018). Currently, 40 *Cryptosporidium* species have been recognized in humans and animals (Feng *et al*, 2018). However, the genetic basis and molecular mechanism by which these species cause disease remains unclear.

Genotyping techniques are used for the identification of *Cryptosporidium* species such as Restriction Fragment Length Polymorphism (RFLP), Polymerase Chain Reaction (PCR) and Sanger sequencing of the 18S rRNA gene (Xiao *et al*, 2017). In *Cryptosporidium parvum* (*C. parvum*) and *Cryptosporidium hominis* (*C. hominis*), the gp60 gene is used as a typing locus, and alleles are discriminated either by Sanger sequencing (Alves *et al*, 2003) or by qPCR (Xiao *et al*, 2001). Whole genome sequence and comparative genomics have been widely used for the characterization of *Cryptosporidium* spp for understanding the molecular basis and epidemiology of cryptosporidiosis. This has been enabled by technology which has provided whole genome determination from clinical samples (Hadfield *et al.*, 2015) and enabled profound insight into population structures and evolutionary relationships between populations of these parasites (Nader *et al* 2019; Tikhule *et al* 2021). *Cryptosporidium* genomic data are available in CryptoDB and NCBI database (Warrenfeltz *et al*, 2020), these genomic databases provide important information about the genetic basis of *Cryptosporidium* spp which are of great interest for public health veterinary research (Heiges *et al*, 2006). These studies have revealed that the *C. hominis* genome is 9.2 Mb in size with a GC content 31.7% and the *C. Parvum* genome is 9.1 Mb in length with a GC content of 30.3 %. They have 3,994 genes and 3,952 genes respectively (Xu *et al*, 2004)

Next generation sequencing and whole genome sequencing has been reported for the important hosts and disease models - *Mus musculus domesticus*, *Homo sapiens* and cattle (Fan et al, 2019). Further functional genomics approaches and comparative genomic analyses of *Cryptosporidium* spp have improved biologist's knowledge about their molecular genetics and evolution. Thus, next generation sequencing technology is a new opportunity for computational biologists to provide a new approach to utilize genomics which includes comparative genome sequencing, transcriptomics, proteomics, metabolomics and epigenetics for the study of parasitic diseases (Preidis *et al*, 2015).

1.6 The Microbiome

The term microbiome describes the microbial communities in a particular environment (including the body or a part of the body). The microbiomes of the gut are of particular interest as they are known to be so diverse and they can influence so many aspects of health in humans and animals (Young, 2017). These microbial communities comprise bacteria, flagellate protozoans, ciliates, anaerobic fungi and viruses, present in all animals. The gastrointestinal tract plays an important role in the digestion and abstraction of nutrients from food and hosts an abundance of microbes. The gut microbiome plays a significant role in maintaining various nutritional, physiological, immunological, and protective functions (deMenezes *et al*, 2011). Vertebrate hosts and their microbiota are thought to have co-evolved and in doing so have each gained advantageous capabilities resulting in a symbiotic relationship between the enteric microbiota and the gut and its host.

1.7 Microbiome Analysis

Within a complex ecosystem like that of the microbiome, the comparative analysis of composition can be performed through rapid identification of microbial taxa based on 16S or 18S rRNA gene sequences. Molecular identification has supplanted traditional culture with the development of NGS technology, providing a window into the variety of microbial communities. The majority of these investigations are based on 16S rRNA gene sequencing,

which contains multiple highly conserved regions interspersed with varied areas (Figure 1.4). The conserved areas serve as "anchors" for building PCR primers, while the variable regions are sequenced to identify the bacteria (Fuks *et al.*, 2018). In the approximately 1.5kbp of the 16S rRNA gene, there are 9 (V1-V9) V sections, each of which carries a distinct sequence depending on the bacterial species. PCR primers designed to be specific to the V1-V2 or V3-V4 regions can yield around 500bp amplified products. In this method genes are amplified using 16S rRNA primers that are each coupled to an adapter sequence for next-generation sequencing analysis, and the products are then attached to random recognition sequences and sequence nucleotides by Index PCR. Mi-seq can now be used to sequence the data. The sequence data is then examined using specialised software, and analyses (phylogenetic analysis, bacterial flora identification study) of the bacterial flora present can then be generated (eg. Figure. 1.5).

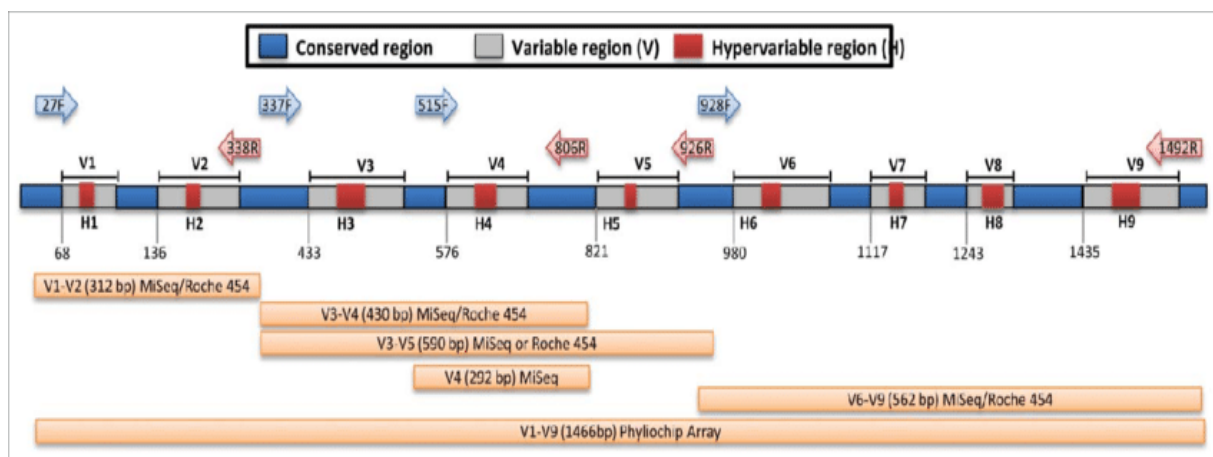


Figure 1.4 shows the 16S rRNA gene's conserved, variable and hypervariable sections, as well as the numerous primer pairs used for metagenomic sequencing. Blue represents conserved regions, grey represents variable regions and red represents hypervariable regions. Some hypervariable sections, such as H3 and H9 are longer than H5 or H7, so the nine hypervariable regions are not evenly distributed. Primers are constructed in a conserved area to produce a PCR product that spans one or more hypervariable regions (V4) (e.g., V1-2 or V3-5). Single-end (Roche 454) or double-end (Illumina) PCR product sequencing gives data for microbiota profiling (Adapted from Shahi *et al.*, 2017).

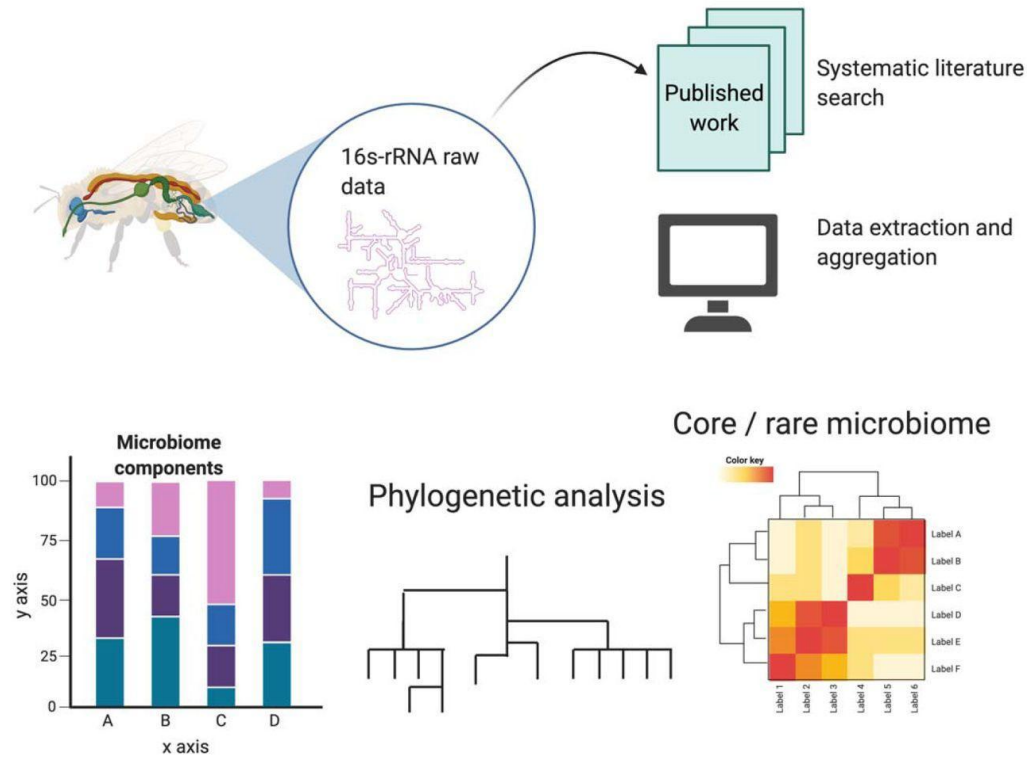


Figure 1.5 pipeline for visualization of honey bee 16S-rRNA microbiome data. Meta-Analysis provides the tabulated data, visualized in a bar chart where x-axis denotes differential abundance for the classes A) *Bifidobacterium*, B) *Lactobacillus*, C) *Snodgrassella*, and finally, D) (unknown genus) the order *Pasteurellales* and the y-axis represents the percentage abundance of microbiome components - their interrelatedness can be described by phylogenetic comparisons visualized as trees, which can then be compared with each other to determine the core components using heatmaps (Gkantiragas and Gabrielli, 2021).

The evolutionary conservation of the 16S rRNA gene in prokaryotes or the 18S rRNA gene in eukaryotes can be used as a marker gene to reduce complexity. This ‘metabarcoding’ approach can resolve evolutionary relationships between individuals without cultivating the organisms (Deng *et al*, 2008). These approaches have elucidated the vast diversity of bacterial genera and species (Larue *et al*, 2005). Preferentially, inferences can be made by sequencing Polymerase Chain Reaction (PCR) amplicons from the ribosomal 16S rRNA gene having restricted domains for archaea and bacteria (Janda and Abbott, 2007). In 16S rRNA libraries the primers used can produce low quality regions. Thus, for short read sequencing, it is recommended to use the trimming software like cutadapt, sickle, fastqMcf. In 16S rRNA it is advisable to trim

along all the sequences using a quality threshold to generate high quality sequences for analyses. In taxonomic classification regions of the bacterial 16S rRNA gene are clustered by two widely accepted approaches. Primarily the sequences are clustered into phylotypes based on their previous similarities with annotated sequences in referential databases (Liu *et al*, 2008). Operational taxonomic units (OTUs) are essentially clusters of sequences reaching a defined similarity threshold that can be *de novo* from 16S sequences (Sun *et al*, 2012). Reference databases developed for the annotation of the sequenced bacterial metagenomes, including the Green-genes database, the Ribosomal Database Project (RDP) and the SILVA database (Balvočiūtė and Huson, 2017). OTUs can be compared to these datasets to provide a phylogenetic position for the clusters and therefore allow comparison of taxonomic composition across different samples. Highly dissimilar sequences are challenging as they form less defined clusters. Also, the read-counts must be normalised before comparison using the total sum and upper quartile normalisation procedures. The normalisation produces sample specific dispersion as well as minimises the variability in samples (Dillies *et al*, 2013). The complete linkage method represents hierarchical clustering of similar bacterial patterns. Quantification of beta diversity can be done either via phylogenetic relationships or non-phylogenetic (taxon based). For example, UniFrac (unique fraction), which can measure the amount of 'unique evolution' of a community in comparison to others. The non-phylogenetic approach includes Bray-Curtis dissimilarity; it is robust to analyze microbiome data. Thus, while using the different primers for different regions 16S rRNA gene led to disagreement due to distinct binding affinities of it to flanking regions and the varied resolution capacity of each variable region in taxa (Seorgel *et al*, 2015). To put it concisely, performing 16S rRNA sequencing is simple methodologically. To begin with, DNA needs to be extracted from the sample. Then, the targeted hypervariable regions are amplified by PCR, samples are then combined together in equal proportions prior to sequencing. On the other hand, shot-gun metagenomics accounts for a high-quality sequence which enables the specific taxonomic and functional sequencing and also led to the discovery of novel bacteria or their genes, preferably in archaea (Norman *et al*, 2015). Therefore, it is important to analyze meta-transcriptomics, proteomics, metabolomics and metadata, like clinical and dietary

information to derive mechanistic models that explain the structure and function of the microbiome.

Metagenomics, unlike metabarcoding, involves sequencing of all microorganisms in the sample. Amplicon sequencing based on 16S rRNA gene results in deciphering microbial biodiversity in a great range of environments including human gut (Yatsunenko *et al*, 2012) and (Lozupone *et al*, 2007), plant roots (Lundberg *et al*, 2012), ocean thermal vents (McCliment *et al*, 2006) and hot springs (De León *et al*, 2013). However, amplicon sequencing results in the production of varying estimates of microbial diversity (Logares *et al*, 2014). In addition, sequencing error and incorrectly assembled amplicons result in the occurrence of artificial unidentified sequences (Wylie *et al*, 2012). Amplicon sequencing cannot directly resolve biological functions of the taxa identified. As microbiomes mostly comprise unculturable bacteria, often a large number of the detected species are novel and difficult to classify (Acinas *et al*, 2004). On the other hand, metagenomic shotgun DNA sequencing represents an alternative powerful environmental sequencing approach that refers to both community biodiversity and metabolic function as well as the possible detection of uncultured microbiota. Instead of targeting a specific genomic locus for amplification in 16S amplicon sequencing, DNA is firstly sheared, then, independently sequenced. Resulted DNA sequences are aligned to various genomic locations for the myriad genomes present in the sample. Most reads will be resulted from protein coding sequences that provide insight into the biological functions of the metagenome (Sharpton, 2014). MG-RAST is also available for annotation purposes. For comparing microbiomes an approach called beta diversity can be used which considers the community difference of varied environments (Tuomisto, 2010).

Owing to the ease of use, reproducibility, sensitivity, and speed of dealing with nucleic acids, traditional microbiological methods have been quickly replaced by newer molecular approaches. These methods enable high-throughput investigation of complex and diverse microbial communities such as those found in soil, freshwater, and saltwater, as well as the microbiota that coexists with a host organism (plant, mouse, human, etc). These approaches,

for example, have been successfully utilised to characterise the plant (rhizosphere), animal, and human microbiome, particularly the complex gut microbiota (Boughner & Singh, 2016). Therefore, it is important to analyse omics methods including metagenomics, meta-transcriptomics, meta-proteomics, metabolomics and metadata, having their own set of benefits when it comes to studying microbial diversity, adaptive evolution, metabolic alterations, and resistance mechanisms to deepen our understanding of single strains or communities under harsh environments (Li and Wen, 2021). These omics techniques have the potential to discover candidate genes, enzymes, metabolites regulating metabolic pathways that may build comprehensive metabolic models to describe how community members interact and how the external environment affects their actions in their respective species. Furthermore, integrated multi-omics analysis can link complete information at various biomolecular levels, resulting in more precise and full global adaptability processes in ever changing activities of microorganisms from enteric bovine livestock. Therefore, the future development path is to continue investigating resistance mechanisms and to apply omics data to the creation of highly efficient engineering strains.

1.8 Analysis of Microbiomes using NGS

The Next Generation Sequencing technology has also allowed the study of human microbiomes at a resolution that was not previously possible and enabled the creation of ‘field guides’ to the taxa found particularly in the human gut (Eisenstein, 2020). Field guides in human microbiomes refer to the taxa specific to a particular organ, to a particular health condition, or to the link between differential abundance of a given taxon and being commensal, opportunistic or pathogenetic. However, many recent studies refer to the difficulties pertaining to this area of research, as microbial communities tend to change dynamically, and some changes are so far unexplainable (Dethlefsen, 2007). For instance, the proposed speculated field guides are of limited use as they cannot clearly distinguish healthy from unhealthy microbiomes due to the vast differences between the microbiomes of apparently healthy individuals (Eisenstein, 2020). This complication is due to the involvement of complex combinations of environmental conditions, genetic makeup and lifestyle (diet) that

globally influenced human microbiomes (Vangay *et al*, 2018) this explains the difficulties in determining whether an individual is healthy or at risk of a given metabolic disease, ex., Type-2 Diabetes, based on microbiome composition. Understanding the clinical implications of those differences is also a challenge, given the extensive interactions among microbes and with their host, as well as the conditions in which that individual survives. Interestingly, preterm birth can profoundly affect the signature of microbiomes of children, for whom normal signatures require a comparatively long time to be established (Callahan *et al*, 2017). Researchers have speculated that a healthy microbiome might not be healthy in another context (Eisenstein, 2020).

It is generally held that the description of a ‘commensal’ microbe is temporally and spatially influenced as this microbe can be converted to advantageous microbe due to the change in relative abundance or if this microbe exists in a different environmental condition (Rothschild *et al*, 2017) also a commensal microbe could become pathogenic in a different environment. Thus, researchers are required to think beyond the term field guides and begin to dissect interactive forces between microbes, on one hand, and between microbes and their hosts, on the other hand. Therefore, recent applications of gut microbiome engineering in medicine and the diet industry ought to handle the problem of imprecise definition of either healthy microbiome or dysbiosis due to the large compositional complexity of both in either cases. For example, the importance of microbiome engineering, particularly in humans (targeting pre/probiotic, microbiome transfer, signalling molecule, drugs, and synthetic biology), animals (targeting feeding enzymes, pre/probiotic, and), plants (targeting microbiome transfer, signalling molecules), and soil (targeting agricultural management) has been reported to promote gut health, disease prevention, resist environmental stresses, and improve soil fertility respectively (Foo *et al*, 2017). Accordingly, researchers are required to conduct more experiments in order to refute the role of the commensal versus opportunistic gut microbiota in human health and disease. Moreover, individual microbes can vary in colonization density of a given niche possibly due to the physicochemical differences, of which some are highly abundant in one niche while less abundant in another niche (Biesbroek, 2012).

Gene catalogues might be a good alternative to 16S for amplicon sequencing when trying to detect and study targeting pathogens at a low density. This method uses mapping of raw microbiome data against published reference genomes of the target pathogene with low density. Therefore pathogen's abundance, diversity and genetic background in terms of resistome and other genes that can cause functional shifting of the microbiome might be detected. However there are technical hurdles to this as low bacterial density and susceptibility to bias caused by environmental contamination are two major problems when studying pathogenic bacteria. Mitigating the pernicious effect of contamination in low density artificial communities is also a major hurdle to analysis of microbiomes was recently discussed (Eisenhofer, 2019). Saladié and others indicated that the choice of DNA extraction protocol has dramatic effects in the study of microbes pathogens with low density (Saladié, 2020).

1.9 Bovine Gut Microbiome

Ruminants including cattle have a stomach with four compartments specialised to pre-gastric digestion. These four compartments are called the rumen, reticulum, omasum and abomasum of which the rumen is the largest and most studied. In new-born calves, rumen remains non-functional for around three weeks, and later starts functioning, and after 6 months, it becomes fully functional. However, the rumen is the home of diverse bacteria, protozoa, and fungi whose function is to break the feed. The diversity of the microbial communities is highly influenced by the age and the type of feed they consume. Ruminants, in a symbiotic relationship with the microorganisms in the rumen, degrade and use fibrous feed as a source of energy and nutrients, thus act synergistically to ferment plant structural and non-structural carbohydrates and proteins. Around eighty-eight difference bacterial genera cultured from various rumens were identified with meta-analysis approach (Kim *et al.* 2011). The calf rumen remains sterile in the womb and gets exposed to microflora for the first time by ingestion of microbiota while in the birth canal. The subsequent establishment of anaerobic microbes comes from the mother and the environment. The rumen can be thought of as an anaerobic and methanogenic fermentation compartment containing microorganisms capable of using cellulolytic food and increasing their output (i.e. hay, straw, silage and grass). The rumen

microbiome, or the community of microbes that live in the rumen, is distinguished by its high population density, wide diversity (which includes bacteria, archaea, protozoa, and fungi), and their complicated interactions. These microorganisms are found in three intersecting micro-environments in the rumen: the solid phase, which makes up 70% of the microbiological bulk, the liquid phase, which makes up 25% of the microbial mass, and the rumen epithelial cells and protozoa, which make up 5% of the microbial mass. Interestingly, ruminants can utilise the components that the human body is unable to break down, such as lignocellulose, thanks to the action of their microbiota. In the rumen, the interaction between the host and the bacteria is synergistic as the host contributes heat, moisture, and food, while the bacteria create protein and by-products of digestion, such as the aforementioned VFAs, for the host.

The predominant phyla present in the rumen of 2-week-old calves are Firmicutes and Bacteroidetes, with Proteobacteria accounting for 20% of the sequenced bacteria. In contrast, analysing the microbiome of rumen from grain-fed animals using pyrosequencing approach (Callaway *et al*, 2010) *Succinivibrio*, *Prevotella*, *Megasphaera*, *Bacteroides*, *Butyrivibrio*, *Clostridium* and *Ruminobacter* were observed at 3% or greater. *Prevotella* and *Butyrivibrio* are two genera that predominate the forage-fed cattle (Dehority and Orpin, 1988). By the use of Terminal Restriction Fragment Length Polymorphism (T-RFLP) technique which is used to determine microbial communities features, it was found that high levels of *Prevotella* were observed in the ruminal bacteria from cattle grazing pasture. *Prevotella* is one of major ruminal genres across most bovine diets, and first bacteria isolated from rumen belonged to same genus (de Menezes *et al*, 2011). Apart from the bacterial microbiome, there is enough enumeration of protozoans in the rumen which have been studied through microscopy, within which *Entodinium* accounts about 95% and is dominant genus in the rumen, especially in animals fed high grain diets. With molecular approaches like DNA fingerprinting, *Epidinium caudatum*, *Entodinium caudatum* and *Isotricha prostoma* were found to be harboured in the rumen (Sylvester *et al*, 2004). Neocallimastigomycota, anaerobic fungi have been isolated from rumen and feces of numerous animals that feed on plants including ruminant and mono-gastric animals (Griffith *et al*, 2010; Liggenstoffer *et al*, 2010). The abundance of the

rumen fungal community has been estimated to be around 10% of the total microbial biomass, with variations according to diet and individual (Krause *et al*, 2013). The heterogeneity of rumen microbial environment and its potential role in animal physiology raise numerous interesting questions concerning gene transfer and the evolution of common roles among its microbial members within niche-like nature, thus produces an understanding of lateral gene transfer (Kav *et al*, 2012; Mizrahi, 2012). However, with Next Generation Sequencing technology it can be suggested that the rumen plasmids play an important role in rumen functions like respiration and metabolism (Kav *et al*, 2012). Thus, it is clear that for ruminant livestock efficiency of food digestion and metabolism and hence important agricultural traits such as optimal growth and milk production require the presence of an effective and efficient microbiome (Jami *et al.*, 2014).

1.10 Metabarcoding and metagenomic studies of the bovine digestive system

The term metabarcoding refers to the identification of the community composition via analysis of the sequences of marker genes, while metagenomics refers to the analysis of the whole genomes in an environmental niche. Ruminants harbour a diverse symbiotic or commensal microbiota in their rumen. This microbiota involves bacteria, archaea, ciliated protozoa, fungi and viruses (Mizrahi *et al*, 2021). Main action of this microbiota is the degradation of complex plant fibres and polysaccharides, thereby providing the host with nutrients such as volatile fatty acids, proteins and vitamins. In return, the host provides the substrates and environment suitable for the survival of the commensal microbiota (Choudhury, *et al*, 2012). It is known that host genetics and diet participate in shaping its rumen microbiota (Li, 2019). Other factors influencing variation in rumen microbiota involve lactation cycles and age (Jami *et al*, 2013) and (Jewell, *et al*, 2015). Ruminants do not manufacture the enzymes required to digest most complex plant polysaccharides, thus the rumen provides a rich and dense habitat for anaerobic microorganisms to perform this metabolic function. Cellulolytics, amylolytics, proteolytics; and other functional categories of rumen microbes; breakdown a wide range of feed components or further metabolise some of the compounds generated by other microbes (Hungate, 1966). For example, Methanogens or methane-forming archaea that metabolize

hydrogen produced by some fermentative microbes to produce methane. The methane produced during this fermentation adds to global human greenhouse gas emissions and results in a 2–12% reduction in the animal's feed energy (Ripple *et al.*, 2014; Johnson & Ward, 1996). Variations in methane formation and feed conversion to animal products are caused by differences in rumen microbial populations (Kittelmann *et al.*, 2014; Jami *et al.*, 2014; Carberry *et al.*, 2012). As a result, understanding these communities is critical to comprehending ruminal conversions of plant material into both unwanted and beneficial ruminant products. In a survey of rumen microbiota of 742 rumen from 32 species/sub-species of ruminants across continents, influences of diet, geographical regions, and genetic background of the host are highly linked (Henderson, 2015). Among high-throughput omics technologies, metagenomics approaches, for example, revealed that phylum Bacteroidetes is energetically less favourable in the host rumen compared to phylum Firmicutes as the latter increases nutrient availability to the host (Hart *et al.*, 2018). Phyla Firmicutes and Bacteroidetes were reported to be the most abundant members in rumen and gut microbiomes at both the DNA and protein levels and comprise key fibre degraders (Hart *et al.*, 2018). Other taxa that are efficient in hemicellulose degradation include genera *Prevotella*, *Butyrivibrio* and *Pseudobutyrvibrio* (Moraïs and Mizrahi, 2019). The latter genera are considered as dominant taxa in the colonisation and degradation of hemicellulose as they comprise a broader enzymatic repertoire that enables them to digest fibre efficiently (Hart *et al.*, 2018). *Selenomonas ruminantium* and members of the Succinic Vibrionaceae family are also considered as dominant members of the rumen microbiome as they utilise secondary fermentation products of other microbes (Scheifinger and Wolin, 1973).

The 16S rRNA gene sequencing in rumen and cecum of ruminants showed *Prevotella* as an important gene in their community (Pitta *et al.*, 2010). About 41 sequences of 16S rRNA genes from the bovine rumen concluded that *M. ruminantium* and *M. stadtnanae* are the predominant species (Whitford *et al.*, 2001). In another study, *Methanobrevibacter* phylotypes have also been observed in high frequency (Wright *et al.*, 2007). Studies have initially focussed on the composition, structure and functions of the microbial communities - including bacteria, fungi, protozoans and viruses (Wang *et al.*, 2015), and their impact on metabolites and gaseous

products produced, mainly CO₂, H₂, H₂S, CH₄ obtained from them due to their environmental impact (Pimentel *et al*, 2013).

To fulfil the demands of an ever-growing population, it is necessary to continue to actively seek for and create new technologies and resources for producing animal proteins in an economically, environmentally, and socially acceptable manner (Myer, 2019). Many factors, including diet, animal management, host genetics, and the complex symbiotic microbiota inhabiting the gastrointestinal system, influence the nutritional condition of beef cattle and other ruminants. The focus on microbial implications on beef nutrition and feed efficiency has increased appropriately throughout the last decade. Further characterization of the complex matrix between the host and its ruminal and lower gastrointestinal microbiota holds promise for optimising bovine production because the gut microbial communities are critical for the degradation of low-quality feedstuffs into glucogenic, lipogenic, and aminogenic precursors involved in energy metabolism and sustenance via other vital nutrients.

Ruminants' gut microbiota likely outnumber their monogastric counterparts in terms of compositional and functional diversity, turning forage energy into high-quality protein and influencing production characteristics like feed efficiency. The USDA-NIFA National Program for Genetic Improvement of Feed Efficiency in Beef Cattle funded several of the earliest studies looking at the relationship between the gut microbiota and beef cattle feed efficiency. This study found a link between rumen, jejunum, cecum, and distal colon/rectum microbial populations and feed efficiency in steers (Myer *et al*, 2017). These findings were based on a feed efficiency metric that took the highest variation of average daily gain (ADG) and average daily feed intake (ADFI) within each Cartesian quadrant into consideration (Myer *et al*, 2015). In the rumen, the group of steers with higher ADG and lower ADFI (i.e., the feed-efficient group) had considerably higher relative abundances of Firmicutes taxa, demonstrating that higher feed efficiency is linked to more abundant Firmicutes populations. This link was fascinating because the Firmicutes-to-Bacteroidetes ratio has been linked to fat deposition increases and energy harvesting in dairy cattle (Ley *et al.*, 2005; Jami *et al.*, 2014). *Butyrivibrio* was found to be the genus with the greatest relative abundance that differed among the feed efficiency groups in jejunal bacterial investigations (Myer *et al*, 2016). Many

taxa of the family Lachnospiraceae were found in the cecum of feed-efficient steers, including *Butyrivibrio*, *Pseudobutyrvibrio*, and *Blautia*, indicating that energy availability to host intestinal tissues is important (Myer *et al*, 2017). Colonic bacteria like *Faecalibacterium* were found to be suggestive of feed-efficient steers (Myer *et al*, 2017), and this has been linked to increased weight gain and lower diarrhoea rates in calves (Oikonomou *et al*, 2013). As a result, *Faecalibacterium* abundances in growing animals may be beneficial for maintaining weight and preventing gastrointestinal infections. These research could be important in defining the microbial impact on beef cattle feed efficiency across the rumen and lower gastrointestinal tract, as well as showing the potential influence of particular microbial species on phenotypic rather than global microbial alterations. Similar studies are also being conducted in the field of rumen microbiology to identify the bovine genome-microbiome link, with the ultimate goal of genetically modifying the rumen microbiome for improved production efficiency (Clemmons *et al*, 2019). Interdisciplinary systems-based concepts will surely be crucial for advancements in the field of rumen microbiology. Researchers have already begun to find heritable microorganisms and microbial characteristics in human and animal systems when considering gut microbiomes as heritable phenotypes (Sasson *et al*, 2017).

Some of the examples of microbiota present in ruminants, humans, and their cross transferring provided in the literature has been shown in Figure (1.6).

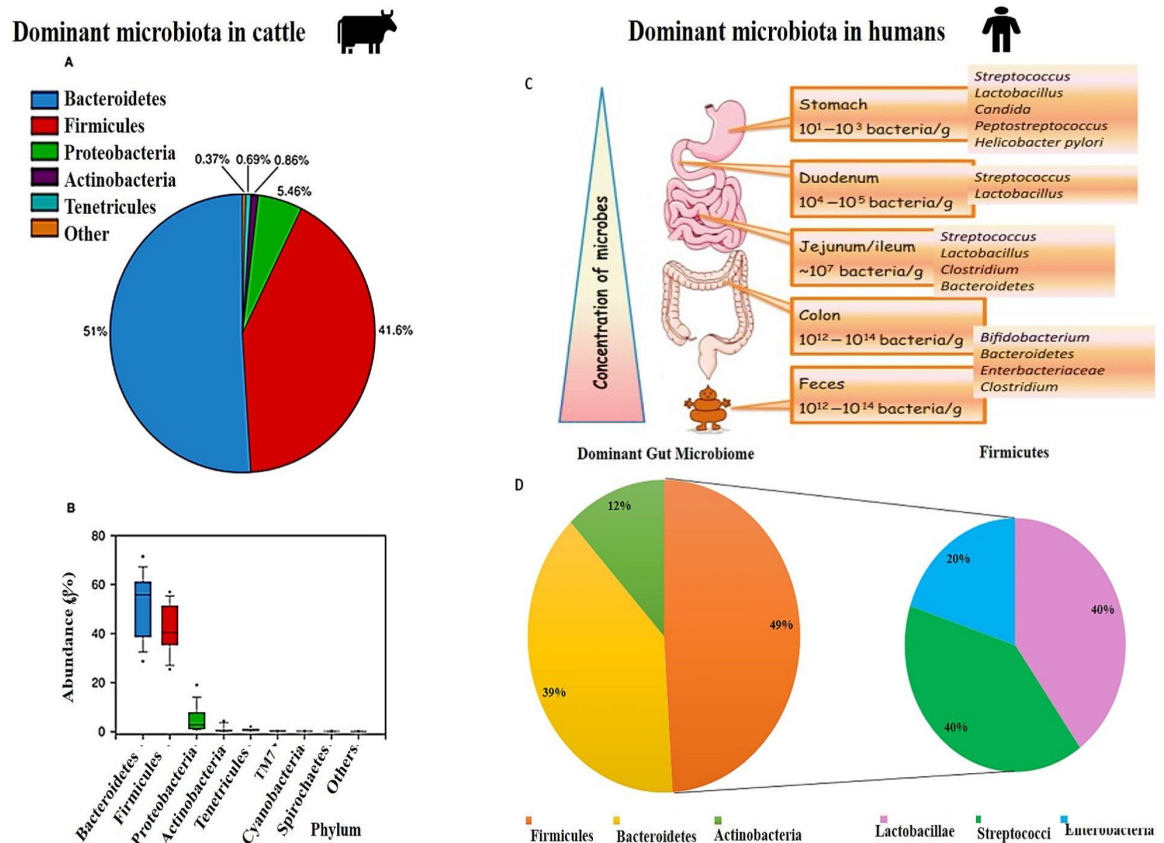


Figure 1.6 Microbial composition of microbiome in Ruminants (cattle) and human gut systems. A) Composition and abundance of bacterial taxa from rumen microbes, as determined by pyrosequencing of the 16 S rDNA gene B) similarity of Bovine Rumen Microbiota across Individual Animals (Adapted from Jami E, and Mizrahi I (2012). C) different regions of the human GI tract constitute different concentrations and groups of microorganisms, D) composition of the human gut microbiome showing dominant bacterial communities and fungal or protozoan species covering a small fraction (Adapted from Srivastava *et al*, 2022).

This figure highlights there are some common core microbial signatures to human gut microbiota and the gut microbiota of animals (principally consisting of *Bacteroidetes*, *Firmicutes*, *Actinobacteria*) forming a complex ecological community, although some are not

completely explored yet. It also describes that the variations of microbial composition of bovine rumen microbiota across individual animals (Jami E, and Mizrahi I (2012). Although it is apparent that some human gut bacterial diseases originate from animals (demonstrating that human-to-animal transmission happens) the extent to which normally non-pathogenic commensal taxa are transferred between people and animals has not been fully investigated (Ellis *et al*, 2013). However, below an example has been highlighted of colonization of the human gut by bovine bacteria present in parmesan cheese showing the cross transfer of microbes from cattle to humans (Milani *et al*, 2019). This figure depicts how dairy cattle-associated bacteria can be transferred to the human gut via consumption of parmesan cheese across the food production chain and employed a multiomics approach (shotgun metagenomics and bifidobacterial ITS profiling) (Figure 1.7). Milani *et al*. (2019) stated that the engraftment of bacteria carried across the Parmesan cheese production chain to the human gut may be aided by regular consumption of cow's milk, suggesting putative co-operative behavior amongst some food-vectored bacteria. Additionally, it could also be possible to identify persistent microbial taxa and functions that are indicative of efficient beef cattle once the influence of host genetic variation on rumen microbiome composition and heritable microbiota is established. These activities will help to accelerate microbial ecology research and production agriculture so that high-quality protein can be produced efficiently and sustainably for human consumption, resulting in a food-secure global population.

Cross-transfer of microbes from cattle to humans

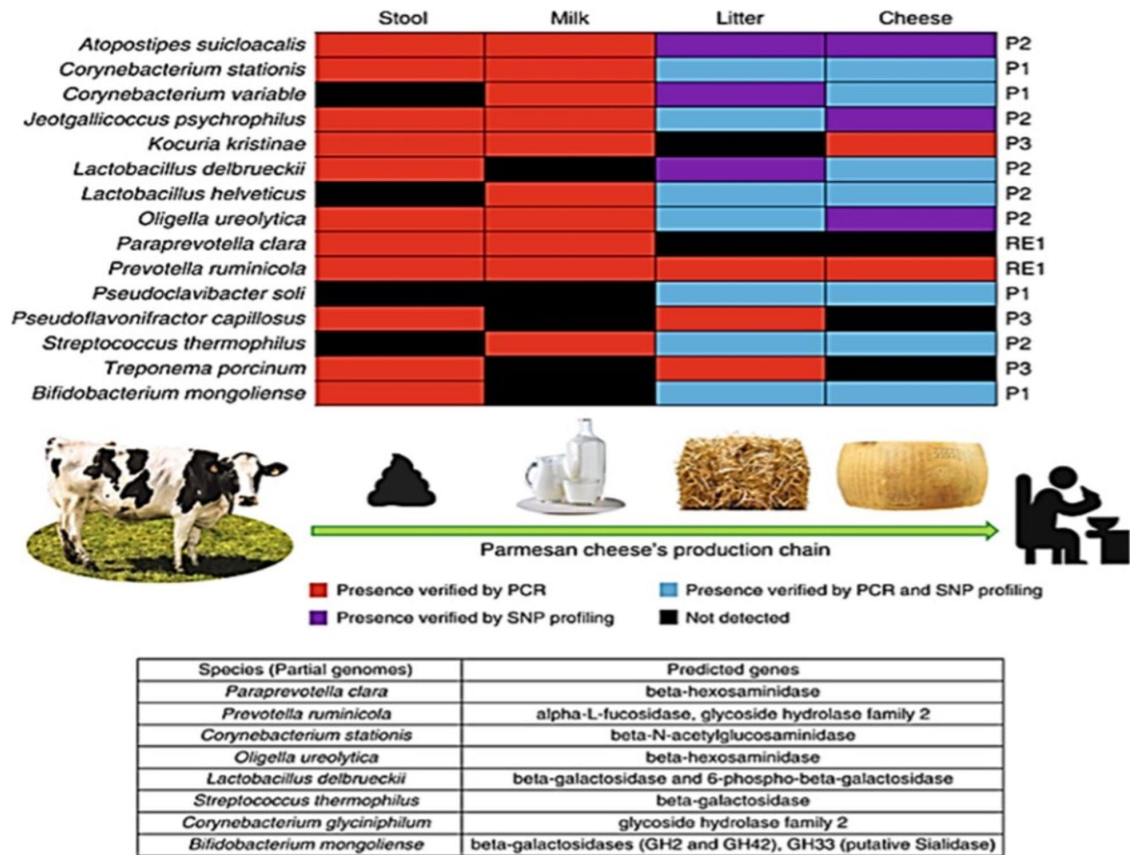


Figure 1.7 An example of colonization of the human gut by bovine bacteria present in Parmesan cheese showing the cross transfer of microbes from cattle to humans (adapted from Milani *et al*, 2019).

1.11 The Gut Microbiome Changes in Response to Pathogen infection

The changes in gut microbiome may change the infectivity and growth of parasites associated with intestinal infections. Germ-free or gnotobiotic animals lack a microbiome and are generally effective models for considering the importance of the microbiome in disease. Nevertheless, some harmful implications, like reduced mucosal linings due to an absence of microbial stimulation of mucosa production in germ-free sea bass [*Centropristis*] (Rekecki *et al*, 2009), have been observed. Moreover, in dogs and mice, the turnover rate of the intestinal epithelium was decreased due to reduced interaction with the bacteria that stimulates the immune system and histological development (Fasano and Shea-donoghue, 2005). In germ-free animals the physiological changes were also seen that decreased or increased the

size of the organs; as noted in germ-free rats, which have smaller hearts and caeca that are four to six times larger than rats with 'normal' microbiomes (Gordon *et al*, 1963).

The 'normal' microbiome can be disrupted by many factors of which colonisation of pathogens is the most important which results in a shift of microbiome population (Heselmans *et al*, 2004). Cryptosporidiosis is a gastrointestinal disease caused by apicomplexan parasite *Cryptosporidium* that affects young ruminal livestock, specifically neonatal calves, and can cause a disastrous effect upon animal health. In the 1970s, the first case of Cryptosporidiosis in cattle was detected (Panciera *et al*, 1971) and in 1983 infected calves with neonatal diarrhoea were experimentally detected that were attributed to *Cryptosporidium* species (Tzipori *et al*, 1983). Nowadays Cryptosporidiosis has been declared as an endemic disease in cattle which can cause neonatal enteritis (Cho and Yoon, 2014). Four species of *Cryptosporidium* have been detected of which *C. parvum* showed clinical symptoms in neonatal calves (Fayer *et al*, 2005). The age-related distribution of *Cryptosporidium* species may be suggested as changes in microflora of the host due to dietary habits that affect the ability of the parasite to infect the gut. This has been supported by the age-related susceptibility of pigs to *C. scrofarum*; while *C. suis* infected all age groups of pigs, *C. scrofarum* only infected older animals (Kvac *et al*, 2009). Pigs are thought to be one of the most important *Cryptosporidium* reservoir hosts among vulnerable animals as many species such as *C. suis*, *C. parvum*, *C. muris*, *C. andersoni*, *C. scrofarum*, and *C. tyzzeri* have been isolated from pigs (Yui *et al.*, 2014). However, there are reports which have highlighted that pigs are susceptible to *C. hominis* and *C. meleagridis* infections, implying that pigs could be a source of infection for humans and other animals, posing an unseen threat to human security (Sheoran *et al.*, 2012). It has been recently suggested that specific alterations in the host-microbiome allow for more favourable growth of cryptosporidial parasites including a significant decrease in anaerobes and overgrowth of Enterobacteriaceae leading to a reduction in butyrate production (Charania *et al*, 2020). These alterations in the microbiota provide the conditions for other bacterial pathogens or commensals to flourish. Therefore, cryptosporidiosis can lead to long-term dysbiosis of the gut (Partida Rodriguez *et al*, 2017). It has been reported that infections with *C. parvum* are associated with changes in the microbiota

that lead to a reduction in faecal indole (Chappel *et al*, 2016). These changes include a decrease in the proportion of *E. coli*, *Bacillus* spp and an increase in *Prevotella* spp and *Bacteroides*. These changes can completely disrupt the normal microbiome of the host and may thereby cause fatal interactions due to alterations in immune response (Peterson *et al*, 2007).

1.12 The Role of the Microbiome in Cryptosporidiosis

Compared to the human faecal microbiome and its association with various diseases such as diarrhoea, very little is known about microbiome changes during cryptosporidiosis. The primary shield between the host tissue and the microbiota is mucus which prevents microbial translocation (McGuckin *et al.*, 2011). Also, the antimicrobial peptides produced by intestinal lineages of cells provides an important role in limiting the exposure to microbiota. These antimicrobials enhance the enzymatic action on bacterial cell walls or the inner compartments of the cell as well (Hooper and Macpherson, 2010). RegIII γ , a mucosal peptide with antibacterial action is expressed either after birth or after the colonisation of the germ-free mice (Cash *et al.*, 2006). RegIII γ also maintains the segregation between microbiota and host intestine creating a visualised separation called a demilitarised zone (Vaishnavi *et al.*, 2011). Memory T cells reside mostly in the tissues that are harboured by microbiota like skin and GI tract. Mostly IL-17 (Th17) and IFN γ (Th1) T cells are found in the GI tract and develop in response to signals derived from the microbiota (Macpherson and Harris, 2004; Ivanov *et al.*, 2008; Gaboriau-Routhiau *et al.*, 2009).

In pre-weaned dairy calves, Firmicutes was the most prevalent phylum observed in the faecal samples with prevalence ranging from 63.84% to 81.90%, this was followed by phylum Bacteroidetes with prevalence ranging from 8.36% to 23.93%, Proteobacteria (23.72% to 9.75%), Fusobacteria (0.76% to 5.67%) (Oikonomou *et al*, 2013). However, in the early weeks of life, the incidence of diarrhoea was highly associated with a reduction of microbial diversity. It was observed that an increase in microbial diversity in the faecal samples after the second week was associated with higher weight gain. The higher prevalence of *Faecalibacterium* spp. were associated with higher weight gain and less diarrhoea

(Oikonomou *et al*, 2013). In another study, an increase in the *Fusobacterium* abundance was observed in the stool samples of animals infected with *C. parvum*. The severity of diarrhoea increased high with an increase of *Fusobacterium* and *C. parvum*, suggesting the role of *Fusobacterium* as an important factor of cryptosporidiosis in animals (Ichikawa-Seki *et al*, 2019).

Aims of the study presented

- 1) To assemble a pipeline for the simultaneous detection of *Cryptosporidium* and compositional analysis of metagenomes in genome shotgun sequences derived from bovine faecal DNA.
- 2) To utilize these data to understand the potential impact of *Cryptosporidium* infection on the faecal microbiome.

CHAPTER 2

Benchmarking methods for detecting *Cryptosporidium* using metagenomic shotgun methods

2.1 Introduction

In this chapter a pipeline is set out for *Cryptosporidium* detection in faecal DNAs using metagenomic shotgun sequencing to ascertain the possibility of identifying *Cryptosporidium*, evaluating the parasite load and determining what a cut-off for detection would be. While metagenomic sequencing is a very powerful method to determine the taxonomic content of a sample, there are potential sources of error from contamination or mismapping of reads. Therefore, initially it was set out to simply determine if it is possible to tell the difference between an un-infected sample and an infected sample, and what the threshold of detection is.

2.1.1 Mapping methods

Read mapping is the process by which short reads of DNA sequences are aligned to a reference genome (or genomes) in order to determine where the query sequence originated from and if the query genome has any differences from the reference. Mapping software represents a big challenge when mapping large numbers of short reads, since some software depends on shredding the reference genome sequence while other softwares chops read sequences for the scanning process which sometimes requires large virtual memory (Li and Durbin, 2009). Here in this study BWA (Li and Durbin, 2009) and CENTRIFUGE (Kim *et al*, 2016) were used to assess which mapping method would produce an accurate detection of *Cryptosporidium* derived reads from a mixed sample. BWA maps sequences to specific reference genomes whereas CENTRIFUGE maps sequences to a database consisting of many genomes with different lengths and checks which genomes are hit by the sequences.

2.1.2 Using metagenomics to find pathogens

Metagenomic shotgun sequencing has been used previously to find pathogens, for example (Yang *et al*. 2016) applied metagenomic methods across the beef production food chain to identify a number of different common pathogens. They used the log read count as a proxy for relative abundance of the pathogens (although the total cell count was not known) and demonstrated that the relative abundance for a number of pathogens increased. Huang *et al*. (2017) used a similar approach to accelerate the method for identifying pathogens in foodborne disease outbreaks. The use of this method in public health has been reviewed, metagenomic methods for pathogen detection can provide rapid information such as strain typing and potentially clinically important traits such as antibiotic resistance but it may lack

sensitivity for pathogens that are at low abundance compared to PCR detection (Miller *et al*, 2013).

Before applying these methods on ‘real world’ samples, first benchmarking samples were used to understand the threshold of detection. In this case samples of DNA from human faecal samples that have been ‘spiked’ with increasing quantities of *Cryptosporidium* DNA were used. These samples were sequenced and benchmarked with a variety of methods to find the detection threshold.

2.2 Material and methods

Previous approach was applied on two sets of spiked samples i.e samples that had a specific quantity of *Cryptosporidium* cells added to them. A single human faecal sample was split into thirteen equal (250 mg) parts with different numbers of *Cryptosporidium* oocysts (a log dilution, as described in Table 2.1) and using 2 different sets of extraction conditions. The second set comprised a single faecal rabbit sample similarly split into 5 (250 mg) parts spiked with *Cryptosporidium* oocysts (as described in Table 2.2).

The human samples were obtained from King’s College London where the DNA of the 13 samples was extracted from the stool of a patient with a non-enteric disorder using the Qiagen DNeasy Powersoil Pro extraction kit (according to the manufacturer’s instructions), in brief, stool samples were lysed by chemical and mechanical homogenization. Lysis buffer was added to a mixed zirconium bead tube containing 250 mg of feces and oocysts. Bead beating was performed using a bead beater (PowerLyser). DNA was purified from the lysates via the supplied silica spin filter membranes. DNA was eluted using a 10 mM Tris elution buffer and the Qubit dsDNA BR assay kit was used to calculate the DNA concentration in each sample. Six samples (S1T, S2T, S3T, S4T, S5T and S6T) out of the 13 underwent an additional heating step of 70° C for 5 min to encourage cell lysis, these are marked with a letter ‘T’ in their sample ID. One sample (S0) is a control sample with no *C. parvum* oocysts added. Each sample was sequenced using an Illumina NOVAseq 6000 system using 2×150bp paired-end reads.

Table 2.1 Spiked human faecal samples from King's College London. Number of *Cryptosporidium parvum* (IOWA) added to each 250 mg aliquot of faecal material. S1T to S6T were heat treated at 70°C for 5 min.

Sample ID	Total number of paired-end reads	Positive/ Negative PCR for <i>Cryptosporidium</i> after spike-in	Number of Spike-in <i>Cryptosporidium</i> Oocysts
S0	3,510,890	Pseudo-control	0
S1	3,911,740	Positive	10^7
S2	3,587,910	Positive	10^6
S3	3,869,916	Positive	10^5
S4	3,499,240	Positive	10^4
S5	3,573,980	Positive	10^3
S6	4,267,300	Positive	10^2
S1T	3,664,230	Positive	10^7
S2T	4,114,610	Positive	10^6
S3T	3,729,520	Positive	10^5
S4T	3,777,270	Positive	10^4
S5T	4,853,910	Positive	10^3
S6T	3,130,760	Positive	10^2

In comparison, the uninfected rabbit samples were acquired from the University of East Anglia, prepared by a different investigator and the kit that was used for the DNA extraction was the FastDNA SPIN Kit for Feces (MP biochemicals, according to the manufacturer's instructions). Briefly, samples were mixed with the lysing matrix supplied. Oocysts were disrupted using zirconium beads but utilising a vortex adapter rather than a dedicated bead-beater, no heat treatment was employed. Following centrifugation at $14,000 \times g$ for 5 min supernatant was transferred to a fresh tube and precipitated. The resulting supernatant was

mixed with binding solution and then applied to a spin column and DNA eluted with the tris buffer.

Table 2.2 Spiked rabbit samples from University of East Anglia. Number of *Cryptosporidium parvum* (IOWA) added to each 250 mg aliquot of faecal material. MMNC is the unspiked sample which served as a negative control.

Sample ID	Total number of paired-end reads	Positive/ Negative PCR for <i>Cryptosporidium</i> after spike-in	Number of Spike-in <i>Cryptosporidium</i> Oocysts
MMNC	3,495,590	Negative	0
MMPC	3,939,682	Positive	10 ⁷
MM6	562,221	Positive	10 ⁶
MM5	3,035,760	Positive	10 ⁵
MM4	211,433	Positive	10 ⁴

2.2.1 Genome sequencing

Sequencing was carried out using the LITE protocol by the Genomic Pipelines group at the Earlham Institute (Perez-Sepulveda *et al*, 2021). Briefly, A master mix containing Nextera buffer, Nextera enzyme, and DNase-free water was combined with of normalized DNA. This reaction was incubated at 56 °C for 10 min. Then, appropriately barcoded P7 adapter was added, and then a master mix containing appropriately barcoded P5 primer, Kapa Robust 2G 5× reaction buffer, dNTPs, Kapa Robust 2G polymerase, and DNase-free water were added to the tube. This PCR amplified as follows: 72 °C × 3 min, 98 °C for 2 min, then 14 cycles of 98 °C × 10 s, 62 °C × 30 s and 72 °C × 3 min, followed by a final incubation at 72 °C for 5 min.

The amplified library purification using magnetic beads had pelleted, the supernatant was removed and discarded, and the beads washed and re-suspended in 25 µL of 10 mM TRIS-HCl, pH 8 (Elution Buffer) and the supernatant containing the DNA was transferred. Libraries were pooled in equimolar concentrations and sequenced across three lanes of Illumina NovaSeq 6000 system.

2.2.2 K-mer Density Analysis

The K-mer analysis tool KAT (Mapleson *et al*, 2017) was used on read files generated from sequencing lanes to compare K-mer spectra, each set of reads were compared to their corresponding ones from the other lane. flag `-t` was set to 4 to use 4 threads on HPC and `-m 27` as K-mer length shown in the following command:

```
kat comp -t 4 -m 27 -o output_file Lane1_L001_R1.fastq.gz  
Lane2_L002_R1.fastq.gz #this tool compares and draws a  
histogram of the shared K-mer between reads
```

```
kat plot density ${file}/kat_density-main.mx -o  
${file}/kat.density.png #this tool is to visualise K-mer hashes  
of 2 reads
```

2.2.3 Taxonomic classification and quantification

CENTRIFUGE version 1.0.3 (Kim *et al*, 2016) was used on each reads from all lanes to label and quantify the species within all the reads using the *nt* database. The flag `-x` also was used to produce an output file with taxonomy information in Kraken format. In addition, the flags `-1` and `-2` were used because the read files were paired-end, and the flag `-q` because the input reads are in FASTQ format. The command was run as below.

```
srunk centrifuge -x nt databases -1  
R1_lane_1.fastq.gz,R1_lane_2.fastq.gz,R1_lane_3.fastq.gz -2  
R2_lane_1.fastq.gz,R2_lane_2.fastq.gz,R2_lane_3.fastq.gz -q --report-file  
centrifuge_output_for_sampleX.tsv  
  
#there should NOT be white spaces between read files, only just  
commas
```

2.2.4 Mapping and counting aligned *Cryptosporidium* reads

In order to perform this stage BWA version 0.7.13 (Li and Durbin, 2009), was implemented for each sample all read files for R1 from all the three lanes were combined into one R1 read file and the same applied on R2 read file as well. Then for each sample the R1 combined files and R2 combined files mapped to single database references genome consisted of 7 *Cryptosporidium* reference genome shown in Table 2.3 using the command:

```
bwa index Cryptosporidium_combined_reference_genomes.fasta
```

```
bwa mem crypto_combined_reference_genomes.fasta R1_cobined_sampleX.fastq.gz  
R2_combined_sampleX.fastq.gz > mapping_of_sampleX.sam
```

```
samtools sort -o mapping_of_sampleX.bam mapping_of_sampleX.sam  
samtools index mapping_of_sampleX.bam
```

Table 2.3 Details of 7 *Cryptosporidium* reference genomes used to build a single database to be used in the mapping phase with BWA.

Reference Genome	Accession number	Genome Assembly Length in Mb
<i>Cryptosporidium hominis</i> _30976	GCA_001483515.1	9.06
<i>Cryptosporidium hominis</i> _37999	GCA_000804495.1	9.05
<i>Cryptosporidium hominis</i> _TU502	GCA_000006425.1	8.74
<i>Cryptosporidium hominis</i> _TU502_2012	GCA_001593465.1	9.11
<i>Cryptosporidium hominis</i> _UKH1	GCA_001593475.1	9.16
<i>Cryptosporidium parvum</i> _IOWA	GCA_015245375.1	9.12
<i>Cryptosporidium parvum</i> _IowaII	GCA_000165345.1	9.10

Then the number of reads that hit *Cryptosporidium* from the mapped reads by BWA is calculated. Thus the command was used to get the number of reads

```
samtools view -b -F12 mapping_of_sampleX.bam | cut -f1 | sort | uniq >  
bothEndsMapped.txt  
Cat bothEndsMapped.txt | wc -l
```

The number of *Cryptosporidium* reads detected is normalised to the total number of reads generated by calculating the number of *Cryptosporidium* reads per million. This is calculated using the following formula:

$$\text{Cryptosporidium reads per million} = \frac{\text{Number of Cryptosporidium hits by BWA or CENTRIFUGE}}{\text{Number of Total Cryptosporidium reads form all lanes}} \times 10^6$$

2.3 Results

2.3.1 K-mer Density Analysis

As the samples were split across sequencing lanes it is crucial to ensure that the different lanes did not introduce batch effects by introducing different error profiles. To ensure this was the case, read files from different lanes were checked that they shared the same k-mer content. The result of k-mer density using KAT (Mapleson *et al*, 2017) in Figure (2.1) shows that the k-mer content from all these samples are consistent across lanes. It is noted that there is a positive correlation in k-mer content between read files from different lanes and there was no sign that k-mer content was mutually exclusive. Thus, the read files across all sequencing lanes can be confidently combined.

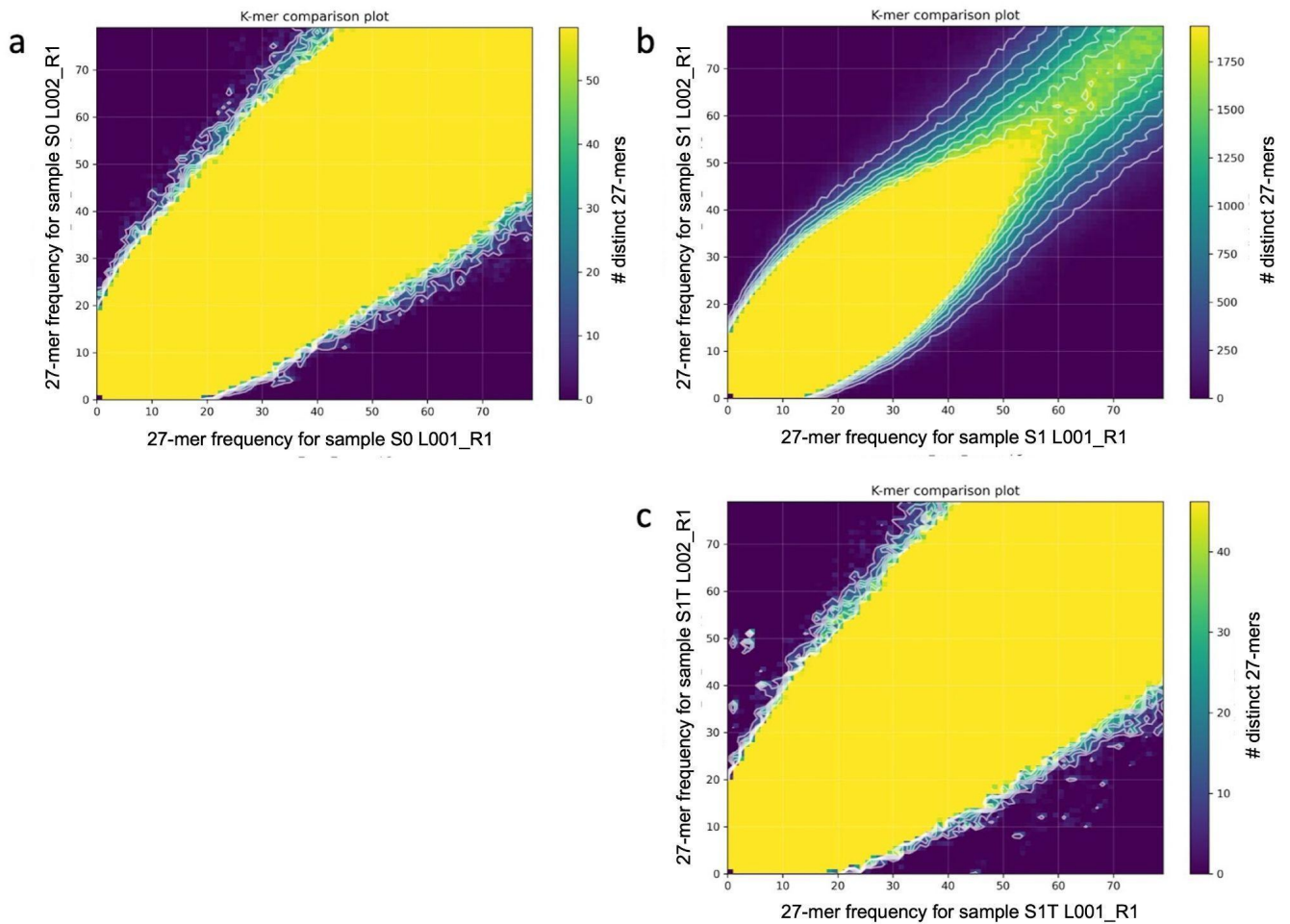


Figure 2.1 Kat density plot comparing read file 1(x-axis) from a single sequencing lane against read file 1 (y-axis) from another sequencing lane for samples (a) S0, (b) S1 and (c) S1T.

2.3.2 Identifying *Cryptosporidium* using BWA and CENTRIFUGE

Table (2.4) presents the total number of reads from all lanes alongside the number of *Cryptosporidium* reads detected using the two different methods, BWA (Li and Durbin, 2009) and CENTRIFUGE (Kim *et al*, 2016) in accordance with the earlier normalisation equation (*Cryptosporidium* reads per million).

Table 2.4 Comparison between *Cryptosporidium* reads per million from BWA and CENTRIFUGE for both, King's College London spiked human samples with *Cryptosporidium parvum* (IOWA) added to each 250 mg aliquot of faecal material. S1T to S6T were heat treated at 70°C for 5 min from. And rabbit samples acquired from University of East Anglia, where they are spiked with *Cryptosporidium parvum* (IOWA) added to each 250 mg aliquot of faecal material. Both samples S0 and MMNC served as a negative control.

Sample ID	Total number of paired-end reads in sample from all lanes	CENTRIFUGE <i>Cryptosporidium</i> Unique reads per Million	BWA reads mapped to Crypto per million	Number of Spike-in <i>Cryptosporidium</i> Oocysts
S0	3,510,890	9.40	1,937.40	0
S1	3,911,740	136.77	4,113.51	10 ⁷
S2	3,587,910	16.44	1,840.35	10 ⁶
S3	3,869,916	9.82	26,973.20	10 ⁵
S4	3,499,240	7.14	22,784.09	10 ⁴
S5	3,573,980	11.19	22,095.53	10 ³
S6	4,267,300	11.25	8,815.41	10 ²
S1T	3,664,230	112.17	14,099.82	10 ⁷
S2T	4,114,610	14.34	6,024.14	10 ⁶
S3T	3,729,520	13.94	24,321.09	10 ⁵
S4T	3,777,270	9.53	12,464.82	10 ⁴
S5T	4,853,910	8.86	1,299.57	10 ³
S6T	3,130,760	6.39	4,222.94	10 ²
MMNC	3,495,590	31.18	55,783.72	0
MMPC	3,939,682	248.24	4,783.38	10 ⁷
MM6	562,221	329.05	1,202.37	10 ⁶
MM5	3,035,760	61.27	5,793.28	10 ⁵
MM4	211,433	42.57	1,314.84	10 ⁴

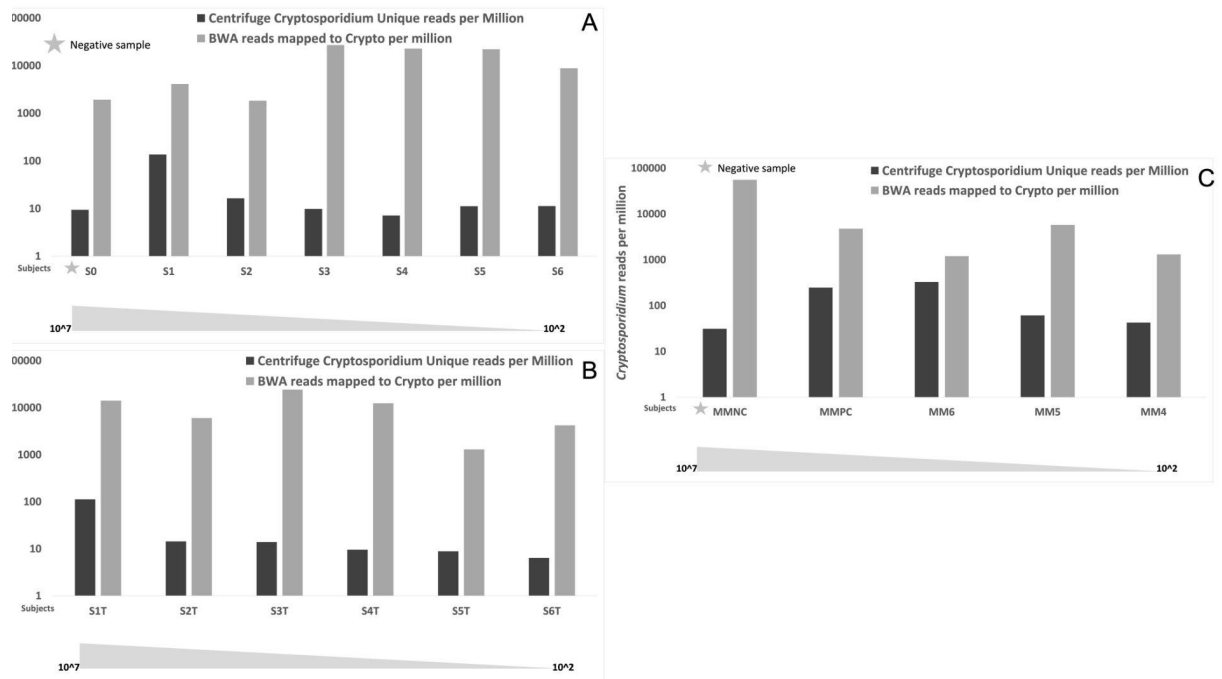


Figure 2.2 Number of *Cryptosporidium* read per million produced by both CENTRIFUGE (Black bars) and BWA (Gray bars) for human samples with no heat treatment (A), human samples with heat treatment (B) and rabbits samples (C)

Figure (2.2) shows *Cryptosporidium* reads per million generated by CENTRIFUGE in human derived samples with maximum of over 100 reads per million where 10^7 oocysts were spiked in 250 mg of feces. While the highest number of *Cryptosporidium* reads per million for the results generated by BWA is nearly 27000 reads per million, where the spik-in amount is 10^5 oocysts in 250 mg of feces. In the corresponding rabbit samples the highest number of *Cryptosporidium* reads per million produced by BWA is above 55000 reads per million in the negative sample. However, CENTRIFUGE gave a number of reads per million as almost 330 reads per million at spiking amount of 10^6 oocysts in 250 mg of feces. Overall, it is observed that the results from CENTRIFUGE show that sample S1 had the most reads classified as *Cryptosporidium* followed by S2. Samples S3, S4 and S5 had a nearly similar abundance to each other which correspond to the background detection rate. In the abundance of *Cryptosporidium* derived sequence reads while the results obtained from BWA (using a bespoke database of *Cryptosporidium* genomes) does not. Using both methods the control samples appear to give a surprisingly high level of reads that map *Cryptosporidium*. The read

hits were visually inspected to validate the results and therefore hits across different genes were identified.

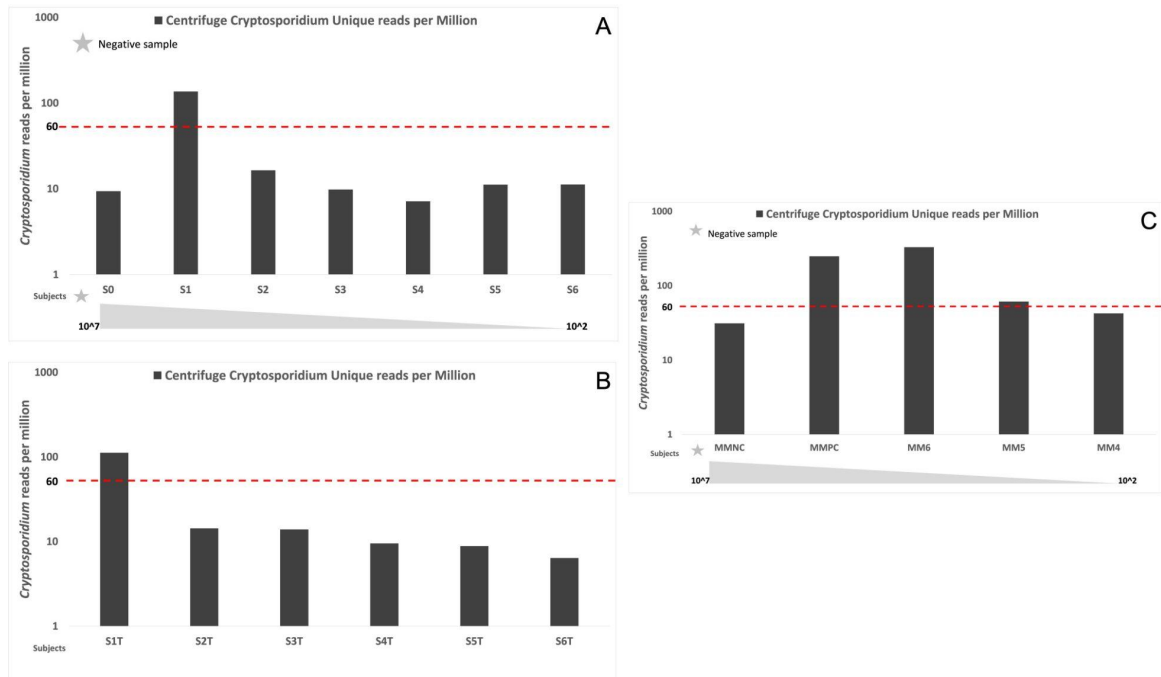


Figure 2.3 setting a detection cutoff at 60 reads per million (dotted Red line) for *Cryptosporidium* reads given by CENTRIFUGE. Five samples exceed the cutoff, one human sample with no heat treatment (A), one human sample with heat treatment (B) and three rabbit samples (C)

2.4 Discussion

From Figure (2.2) it seems that BWA might be giving inaccurate results because the human samples with no heat treatment S1, S2, S3, S4, S5 and S6 were spiked with different amount of *Cryptosporidium* however there is no clear difference between the control sample S0 and the rest of the samples from BWA. It was expected that samples with 10^7 *C.parvum* oocysts in 250 mg of feces spike-in would have more *Cryptosporidium* reads than samples with 10^6 *C.parvum* oocysts in 250 mg of feces and with 10^5 *C.parvum* oocysts in 250 mg of feces as illustrated in Figure (2.4). However, BWA gave a background level of detection that meant that it was not possible to observe the increase in *Cryptosporidium* in the spike in samples while CENTRIFUGE had a much lower background detection rate. It is likely that this is due to mismapping of reads to the *Cryptosporidium* reference genomes as my database only contained *Cryptosporidium* references and therefore any reads that are common to many genomes, such as low complexity regions and repeat regions will map, regardless of what organisms the originated from. Whereas, CENTRIFUGE appears to be working better than

BWA by giving results which are consistent with the number of *Cryptosporidium* cells in the spike in samples. Therefore the cutoff detection for *Cryptosporidium* would be 60 reads per million (Figure 2.3) this would equate to 10^7 oocysts in 250 mg of feces and should be identifying *Cryptosporidium* as it is double of the negative control reads in the rabbit derived sample.

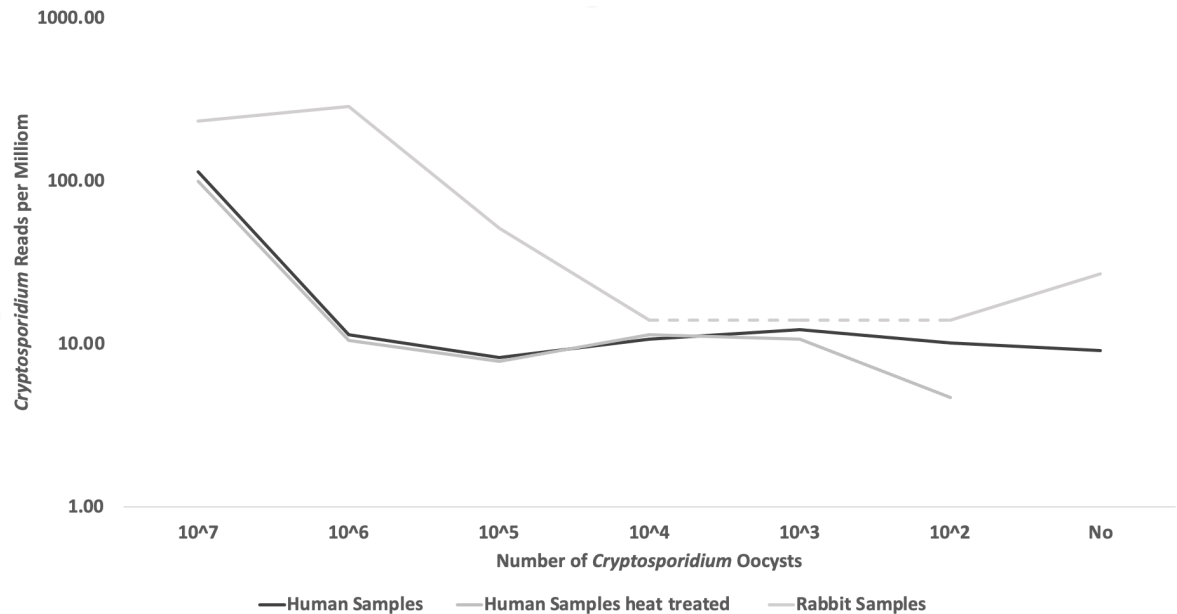


Figure 2.4 A comparison between the number of *Cryptosporidium* reads per million and the numbers of oocysts spiked into human and rabbit samples.

Interestingly, there was little variation in read numbers and hence sensitivity of detection between the two treatments of the human spiked samples indicating good consistency in the preparations and little effect from the heat treatment. By comparison rabbit spiked samples showed relatively higher read numbers for the same number of spiked oocysts indicating that the kind of host faeces and the exact sample preparation may have a considerable effect on the sensitivity of detection.

Surprisingly, there appears to be a large number of reads mapping to *Cryptosporidium* in the negative control samples and the samples with a low number of spike-in cells. This could have a number of explanations:

1. The negative control may have come from a host with a *Cryptosporidium* infection that was undetected. However this would seem unlikely as PCR is highly sensitive.
2. There may be other organisms present that are not in the CENTRIFUGE database but are similar enough to *Cryptosporidium* that the reads map to the *Cryptosporidium* reference.
3. There could be cross contamination of samples; whereby material from positive samples has contaminated the other samples during DNA prep or library construction.
4. Some of the molecular barcodes of the genomic library have been misassigned due to sequence errors. This is unlikely as the barcodes used are quite different to each other however it has been reported in other studies that this has occurred.

CHAPTER 3

Analysis of the interaction of *Cryptosporidium* and the gut microbiome in humans and cattle

3.1 Introduction

The previous chapter (Chapter 2) demonstrates how to detect *Cryptosporidium* presence in spiked samples and set a cutoff for detection. Therefore this approach should be able to simultaneously measure the presence of *Cryptosporidium* in a faecal sample and study the microbiome of the host from the same metagenomic shotgun dataset. In this chapter, previous method was applied to human and animal samples to test if the presence of *Cryptosporidium* in the gut impacts the microbiome. Here, the same pipeline that was developed in Chapter 2 was applied to human and two sets of cattle samples and also to investigate the wider eukaryotic and prokaryotic microbiome to ascertain if there are differences in the infected individuals.

3.1.1 The impact of infection on microbiomes

The microbiome plays a crucial role in a host's health, the gut metabolome seems to be influenced by changes that occur to gut microbiota composition (Corrêa-Oliveira *et al*, 2016), pathogen colonization affecting the natural pattern of the microbiome (Heselmans *et al*, 2004). Furthermore, pathogens may cause chronic gut microbiome imbalance (Partida Rodriguez *et al*, 2017). *Cryptosporidium* infection can lead to antimicrobial peptide deprivation which has an impact on the gut microbiota content, increasing gram-negative bacteria abundance (Rossignol *et al*, 2006). *Cryptosporidium* has been demonstrated to impact the host microbiome in a number of studies. For example Karpe *et al*. (2021) has shown in mice *Cryptosporidium* caused an increase in the proportion of bacteria associated with a stressed gut environment which in turn caused an increase in host proteins involved in energy pathways and *Lactobacillus Glyceraldehyde-3-phosphate* dehydrogenase were upregulated during infection. Also, in mice Charania *et al*. (2020) demonstrated that the microbiome of the host can influence the susceptibility to infection. Moreover, Ichikawa-Seki *et al* (2019) have shown that in cattle *Fusobacterium* (some of which are known to be pro-inflammatory) may be an important aggravating factor of cryptosporidiosis as it is associated with infection.

3.2 Methods

Human stool samples were acquired from Public Health Wales where all samples had tested positive for *Cryptosporidium* by PCR. The DNA had been extracted using Invitrogen purelink microbiome DNA purification kit and was sequenced at the Earlham Institute on an Illumina Novaseq on 3 lanes using paired end reads with a read length of 150bp. (Table 3.1).

3.2.1 Samples used in this study

Table 3.1 Human samples acquired from Public Health Wales where all samples had been tested positive for *Cryptosporidium* by PCR test.

Human samples		
Sample ID	Total number of paired-end reads	Positive/ Negative For <i>Cryptosporidium</i>
UKCU18	3,260,560	Positive
UKH101	2,974,560	Positive
UKH102	2,950,612	Positive
UKH104	212,252	Positive
UKH105	2,987,530	Positive
UKUB17	3,197,680	Positive

The first set of 11 cattle samples were taken from Calves and their mothers from a farm in Europe as part of the Help for Dairy Cows project (H4DC) and provided by the University of Kent where the DNA extraction was conducted by using Invitrogen purelink microbiome DNA purification kit. 2 samples out of the 11 were PCR negative to *Cryptosporidium* (B165, B174). While the remaining 9 samples out of 11 were positive to *Cryptosporidium* by PCR as shown in Table (3.2). These samples were sequenced on an Illumina Novaseq on 3 lanes using paired end reads with a read length of 150bp. At the Earlham institute, also the samples K-mer content was analysed as demonstrated in the previous chapter (2.2.1)

Table 3.2 Cattle samples acquired from University of Kent where 2 samples are negative and the rest are positive for *Cryptosporidium* by PCR test.

Cattle samples		
Sample ID	Total number of paired-end reads	Positive/ Negative PCR test
B165	3,350,120	Negative
B174	3,799,090	Negative
B19	3,126,340	Positive
B45	4,577,790	Positive
B50	3,967,128	Positive
B54	3,388,520	Positive
B8	4,106,290	Positive
N2	3,243,295	Positive
N3	3,495,561	Positive
N4	4,404,238	Positive
N5	4,393,070	Positive

The second set of Cattle samples were 36 pre existing dataset of cross-breed dairy adult African cattle faecal microbiome collected from two regions of Tanzania, Rungwe region and Lushoto region (Kibegwa *et al*, 2020) obtained from MG-RAST repository from a metagenomic project (<https://www.mg-rast.org/mgmain.html?mgpage=project&project=mgp81260>). QIAamp DNA Stool Mini Kit (Qiagen, USA) was used in DNA extraction. These had not been tested for *Cryptosporidium*. (Table 3.3). These samples were sequenced on an Illumina MiSeq using paired end reads with a read length of 200bp.

Table 3.3 Tanzanian Cattle samples downloaded from MG-RAST based on their collection location with their total number of reads.

Cattle samples from Rungwe region		Cattle samples from Lushoto region	
Sample ID	Total number of reads	Sample ID	Total number of reads
TMP017787	395,167	TRZ505	859,596
TMP017800	315,166	TRZ511	253,257
TMP017811	726,815	TRZ531	834,245
TMP017852	409,782	TRZ562	464,850
TMP017871	430,555	TRZ577	401,816
TMP017887	364,184	TRZ578	1,128,833
TMP017891	528,171	TRZ674	363,852
TMP017909	793,242	TRZ677	335,147
TMP017921	835,335	TRZ679	889,849
TMP017926	449,874	TRZ689	333,285
TMP017982	719,285	TRZ718	836,335
TMP018001	684,545	TRZ721	1,037,625
TMP018016	840,437	TRZ728	701,396
TMP018023	357,863	TRZ1429	338,294
TMP018027	310,963	TRZ1464	439,797
TMP018032	233,470	TRZ1469	419,444
TMP018046	311,466	TRZ1474	614,810
TMP018060	497,494	TRZ1480	914,889

3.2.2 Taxonomic classification and quantification of humans and University of Kent cattle samples

CENTRIFUGE version 1.0.3 (Kim *et al*, 2016) was run on all reads from all lanes to label and quantify the species within all the reads using the *nt* database. The flag `-x` also was used to produce an output file with taxonomy information in Kraken format. In addition, the flags `-1` and `-2` were used because the read files were paired-end, and the flag `-q` because the input reads are in FASTQ format. The command was run as below.

```
srun centrifuge -x nt databases -1
R1_lane_1.fastq.gz,R1_lane_2.fastq.gz,R1_lane_3.fastq.gz -2
R2_lane_1.fastq.gz,R2_lane_2.fastq.gz,R2_lane_3.fastq.gz -q --report-file
centrifuge_output_for_sampleX.tsv
```

3.2.3 Mapping and counting aligned *Cryptosporidium* reads of humans and University of Kent cattle samples

In order to perform this stage BWA version 0.7.13 (Li and Durbin, 2009), was used for each sample all read files for R1 from all the three lanes were combined into one R1 read file and the same applied on R2 read file as well. Then for each sample the R1 combined files and R2 combined files were mapped to single database references genome which consisted of 8 *Cryptosporidium* reference genomes (which are provided in Appendix II) using the command:

```
bwa index Cryptosporidium_combined_reference_genomes.fasta

bwa mem  crypto_combined_reference_genomes.fasta R1_cobined_sampleX.fastq.gz
R2_combined_sampleX.fastq.gz >  mapping_of_sampleX.sam

samtools sort -o mapping_of_sampleX.bam mapping_of_sampleX.sam
samtools index mapping_of_sampleX.bam
```

Then it was needed to calculate the number of reads that hit *Cryptosporidium spp* genomes from the mapped reads by BWA (Li and Durbin, 2009). Thus the command was used to get the number of reads

```
samtools view -b -F12 mapping_of_sampleX.bam | cut -f1 | sort | uniq >
bothEndsMapped.txt
Cat bothEndsMapped.txt | wc -l
```

3.2.4 Analysis of pre existing dataset

For the Tanzanian dataset the same previous approach was used to identify *Cryptosporidium* as previously, except in this attempt CENTRIFUGE was only employed to calculate and identify the reads that hit *Cryptosporidium* against the *nt* database. Some amendments were done to the use of flags of CENTRIFUGE due to the reads being interleaved, thus flag -U was used alongside with flag -q because input read files were also in FASTQ format. And to obtain an output file with taxonomy information the flag -x was used.

```
srunk centrifuge -x nt databases -U sample_read.fastq -q --report-file  
centrifuge_output_for_sampleX.tsv
```

3.3 Results

3.3.1 Identifying *Cryptosporidium* BWA vs. CENTRIFUGE for human and H4DC samples

In Tables (3.4 and 3.5) again the total number of reads from all lanes was employed. Subsequently, *Cryptosporidium* reads per million were calculated separately by the numbers that were generated by both BWA (Li and Durbin, 2009) and CENTRIFUGE (Kim *et al*, 2016) according to the following equation:

$$\frac{\text{Number of } \textit{Cryptosporidium} \text{ hits by BWA or CENTRIFUGE}}{\text{Number of Total } \textit{Cryptosporidium} \text{ reads form all lanes}} \times 10^6$$

Table 3.4 Comparison between *Cryptosporidium* reads per million from BWA and CENTRIFUGE for human samples from Public Health Wales

Sample ID	Total number of paired-end reads in sample from 3 lanes	CENTRIFUGE <i>Cryptosporidium</i> Unique reads per Million	BWA reads mapped to Crypto per million
UKCU18	3,260,560	7.36	1,233.53
UKH101	2,974,560	10.76	8,091.28
UKH102	2,950,612	31.18	2,920.41
UKH104	212,252	2,341.56	16,150.61
UKH105	2,987,530	51.88	17,870.95
UKUB17	3,197,680	254.87	11,273.80

As these human samples are from infected individuals which had PCR positive results for *Cryptosporidium* infection, Figure (3.1) shows that CENTRIFUGE detects infection in two of the samples where sample (UKH104) looks to have a much more severe infection with an extremely high number of reads hitting *Cryptosporidium*. Despite the fact that all of the 6 samples were positive for *Cryptosporidium* by PCR, confidently sample (UKH104) would be the one that is coming from a highly infected human as it shows that *Cryptosporidium* reads exceeded 1000 reads per million which may indicate that there is an acute infection

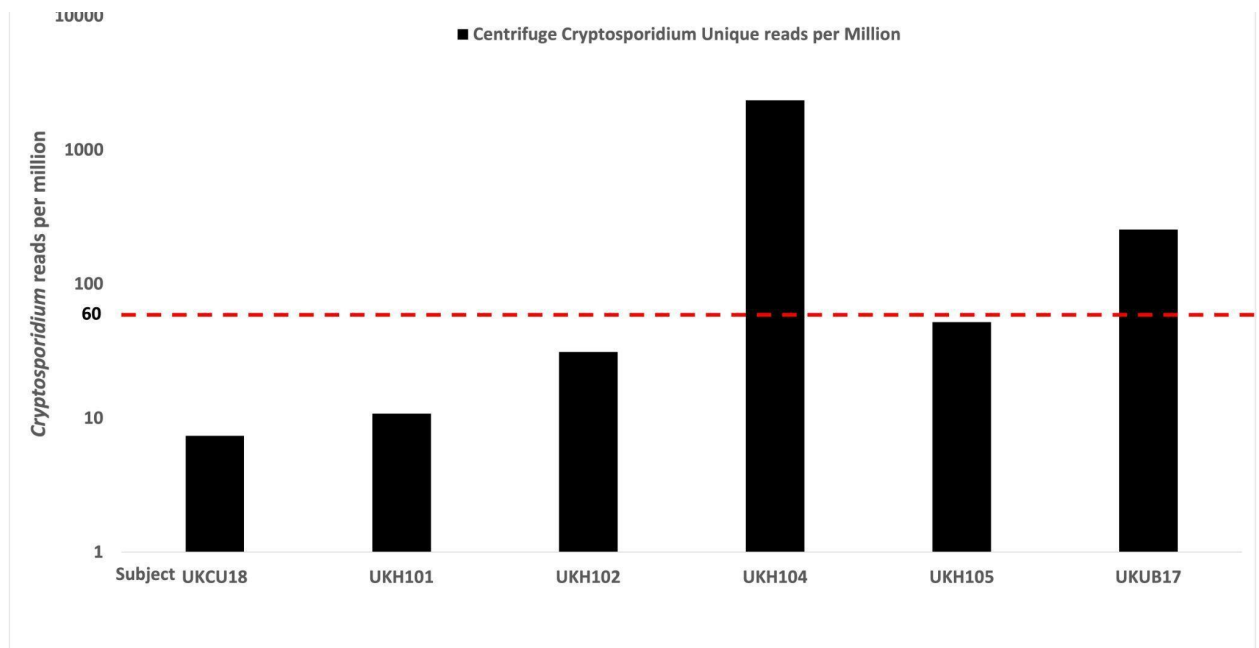


Figure 3.1 Number of *Cryptosporidium* reads per million produced by both CENTRIFUGE with a detection cutoff (dotted red line) where two samples exceed the cutoff.

Table 3.5 Comparison between *Cryptosporidium* reads per million from BWA and CENTRIFUGE for cattle samples from University of Kent

Sample ID	Total number of paired-end reads in sample from all lanes	CENTRIFUGE <i>Cryptosporidium</i> Unique reads per Million	BWA reads mapped to Crypto per million
B165	3,350,120	42.98	2,791.24
B174	3,799,090	31.06	21,996.85
B19	3,126,340	8.64	5,059.91
B45	4,577,790	39.76	17,988.16
B50	3,967,128	41.09	17,915.23
B54	3,388,520	38.36	21,734.56
B8	4,106,290	6.09	22,861.27
N2	3,243,295	28.67	26,778.94
N3	3,495,561	32.04	4,859.59
N4	4,404,238	2.50	24,116.32
N5	4,393,070	5,054.78	105,915.00

In European cattle (unspiked) samples that are PCR positive to *Cryptosporidium*, CENTRIFUGE still outperforms BWA in terms of *Cryptosporidium* read accuracy. Figure (3.2) shows *Cryptosporidium* reads given by CENTRIFUGE where sample (N5) has a much higher number of *Cryptosporidium* reads exceeding 1,000 reads per million compared to the negative samples (B165 and B174) that have less than 100 reads per million.

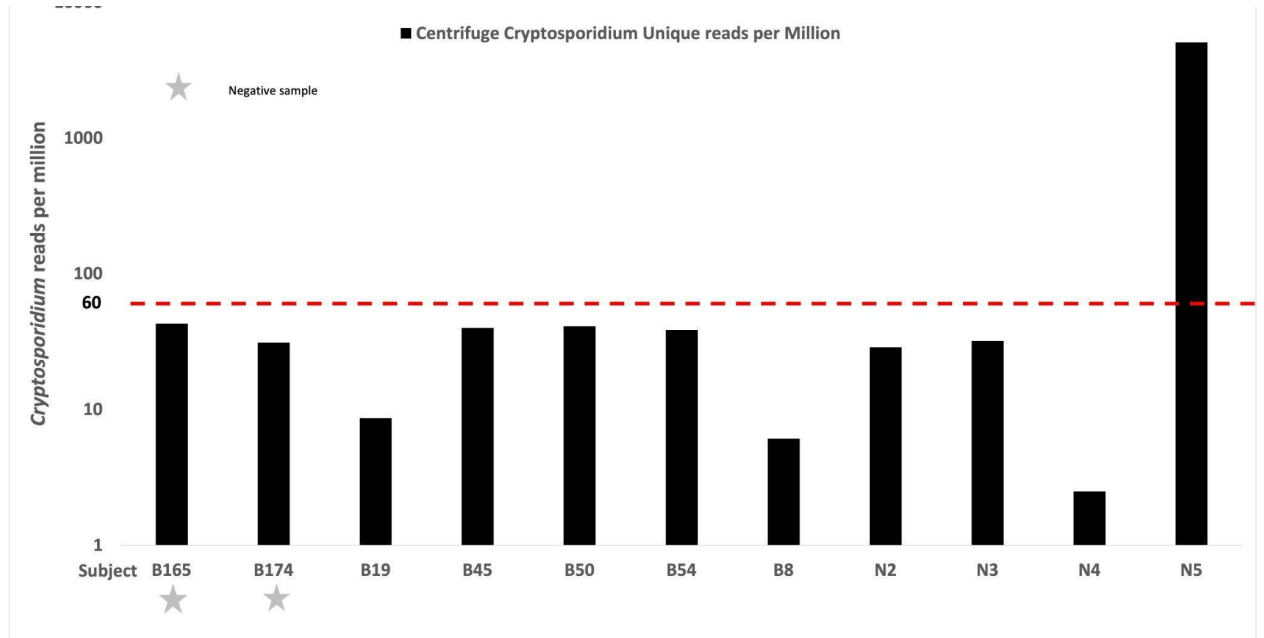


Figure 3.2 Number of *Cryptosporidium* read per million produce by both CENTRIFUGE and BWA with a detection cutoff (dotted red line) where only one sample exceeds the cutoff.

3.3.2 Analysis of the microbiota in *Cryptosporidium* infected human samples

The microbiota of the infected human samples was analysed to identify whether there were any changes associated with infection. As a control for this analysis the un-spiked control sample (S0) was used. Firstly, to identify any clusters, PCA analysis was used (Figure 3.3) to be able to distinguish the similarities and variations in the microbiome composition between all samples. PCA was performed based on the phylum assigned to each read, the clear outlier from the PCA analysis presented (brown dot) is sample (UKH104) which also has by far the most read mapping to *Cryptosporidium*. The other samples are separated along principle component 2. In both the eukaryotic microbiota and the bacterial microbiota UKH104 is an outlier.

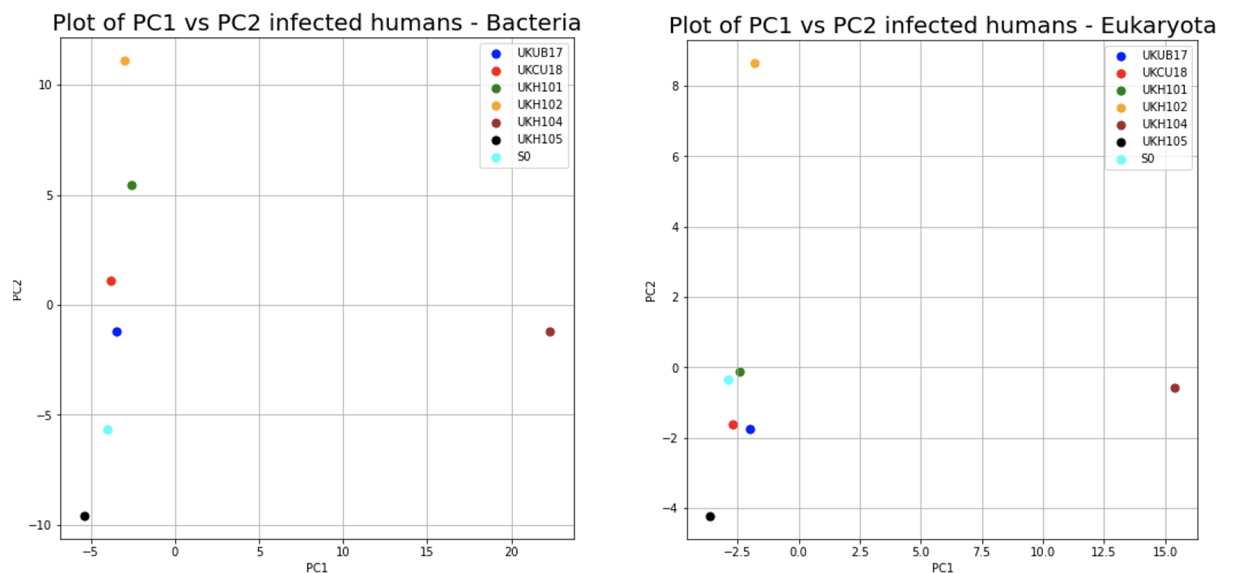


Figure 3.3 PCA analysis for both bacterial and eukaryotic components of infected bovine where it shows that samples UKUB17, UKCU18, UKH101, UKH102, UKH105 and S0 (control sample from King's College London) are clustering together. Whereas sample UKH104 is an outlier which suggests that there has been a change in its both eukaryotic and bacterial microbiota.

3.3.2.1 Bacterial microbiota changes associated with *Cryptosporidium* infection in human samples

Looking at Figure (3.4) sample (UKH104) has significant changes to its bacterial microbiome where there were over 80% *Bacteroidetes* while in the control sample (S0) there were less than 30% *Bacteroidetes*. In addition a pattern of changes in the microbiome in response to infection was observed, sample (UKH104) has much reduced microbiome diversity while it has an increased proportion of *Bacteroidetes*.

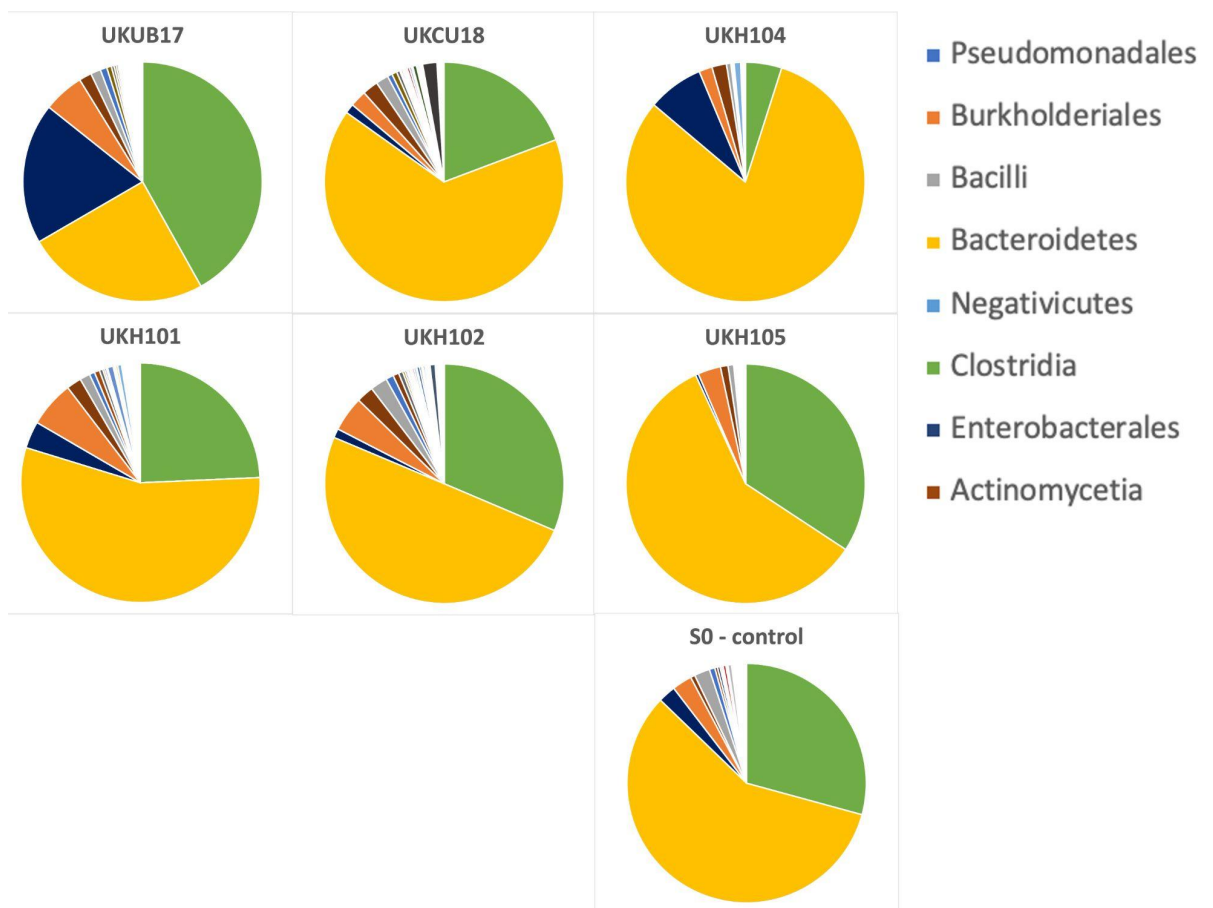


Figure 3.4 The bacterial component of human unspoked samples

3.3.2.2 Eukaryotic microbiota

By comparison in Figure (3.5) the eukaryotic microbiome content does not noticeably change in humans infected with *Cryptosporidium* as noticeably all samples have the same abundance of Eukaryota in the human gut although sample (UKH104) has marginally more Apicomplexa in it than the other samples, presumably due to the high presence of *Cryptosporidium*. However, unlike the observations in the bacterial microbiota, there is no clear evidence of major changes in the other phyla associated with the *Cryptosporidium* infection.

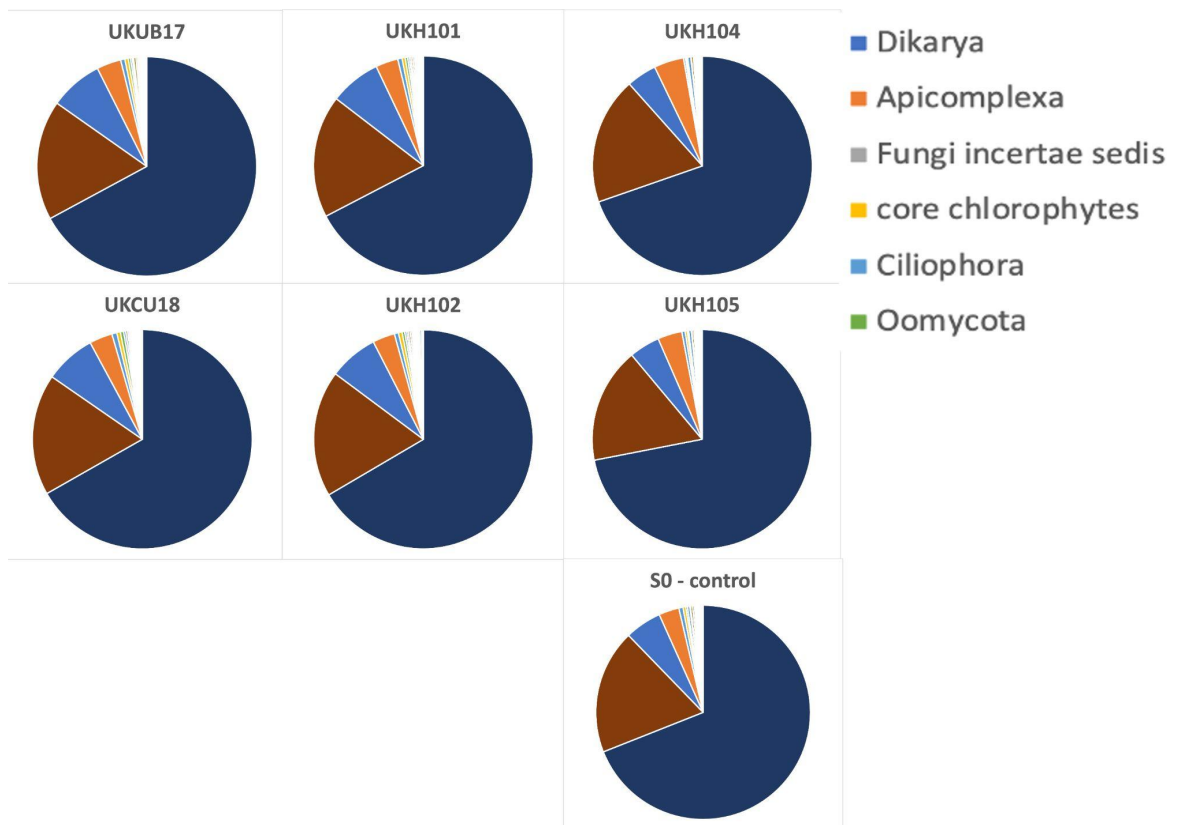


Figure 3.5 Eukaryotic microbiota a changes associated with *Cryptosporidium* infection in human samples

3.3.3 Analysis of the microbiota in *Cryptosporidium* infected European bovine samples

Principal component analysis was used to identify any clustering of the samples based on the eukaryotic or prokaryotic component of their microbiota. Based on this analysis (Figure 3.6), there is a small cluster of five samples in the PCA analysis of the bacterial content but this does not specifically align with the *Cryptosporidium* content and sample B165 is an outlier. There is no clustering from the eukaryotic content. The two negative control samples B165 and B174 do not cluster together on either sample.

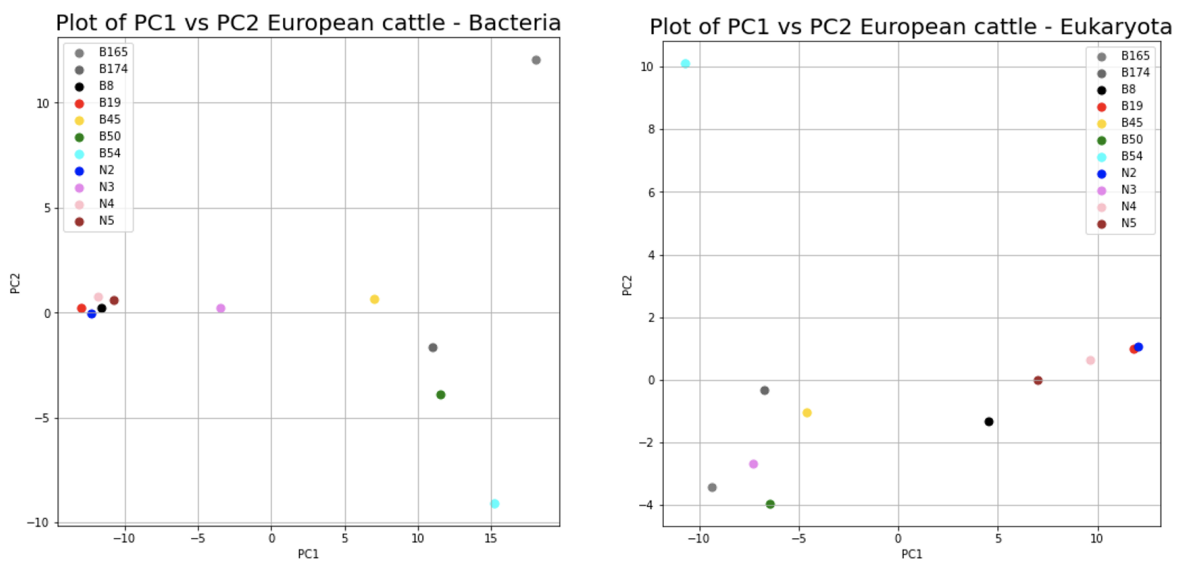


Figure 3.6 PCA analysis for both bacterial and eukaryotic components of infected bovine

3.3.3.1 Bacterial microbiota of European cattle from University of Kent

Looking at Figure (3.7) it seems that the negative control samples (B165 and B174) are showing a standard bacterial diversity whereas sample (N5) shows what a severe *Cryptosporidium* case looks like and it is also observed that there is a considerable loss in bacterial diversity in it.

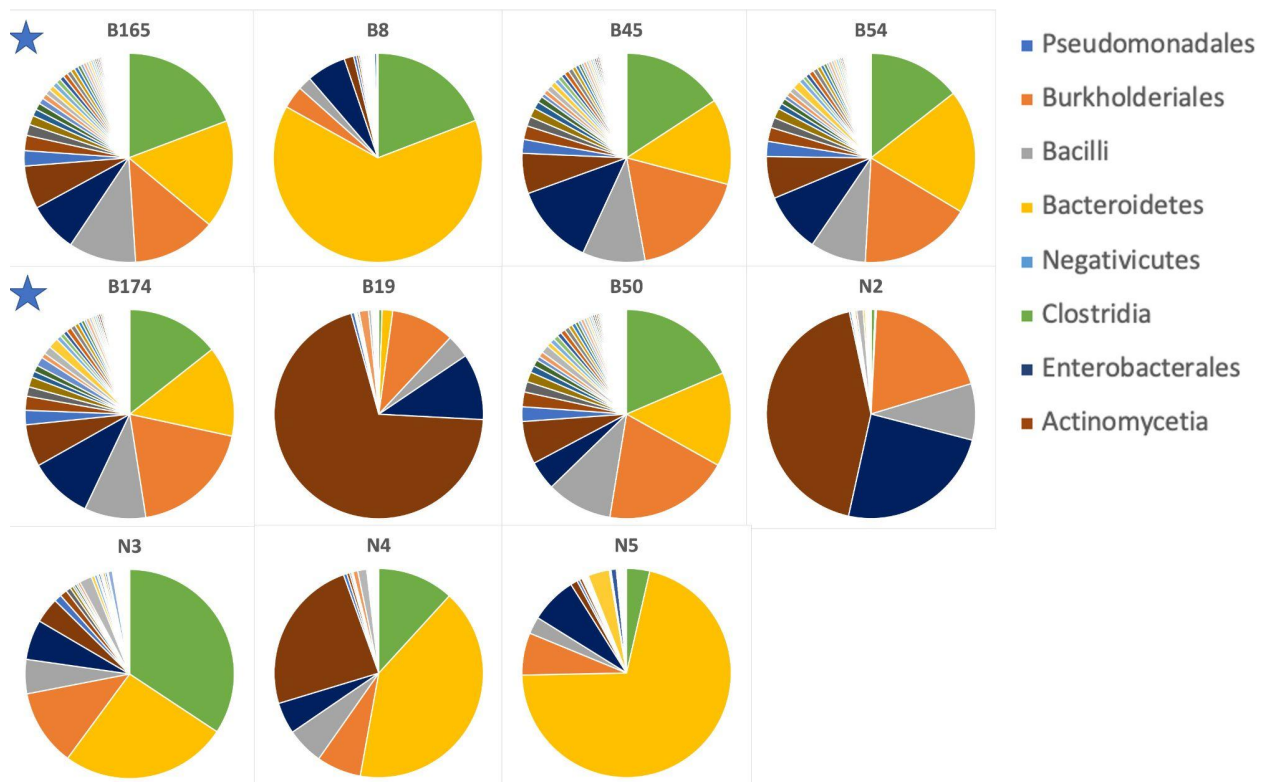


Figure 3.7 The bacterial component of European cattle samples

3.3.3.2 Eukaryotic microbiota

The results of the eukaryotic component in the unspiked cattle samples in Figure (3.8) contrasts with the unspiked human samples in that there is much greater eukaryotic diversity as one might expect in the faecal DNA of a ruminant. However there is a reduction in diversity of the eukaryotic component in some infected samples of cattle than in the control samples, such as in samples (B19, N2, N4 and N5). These samples, although they do not all have a large number of *Cryptosporidium* derived reads, they all have an interestingly increase in reads classified as *Apicomplexa*, as all *Apicomplexa* are parasites it is possible that these animals have another gastrointestinal infection, such as *Coccidia*.

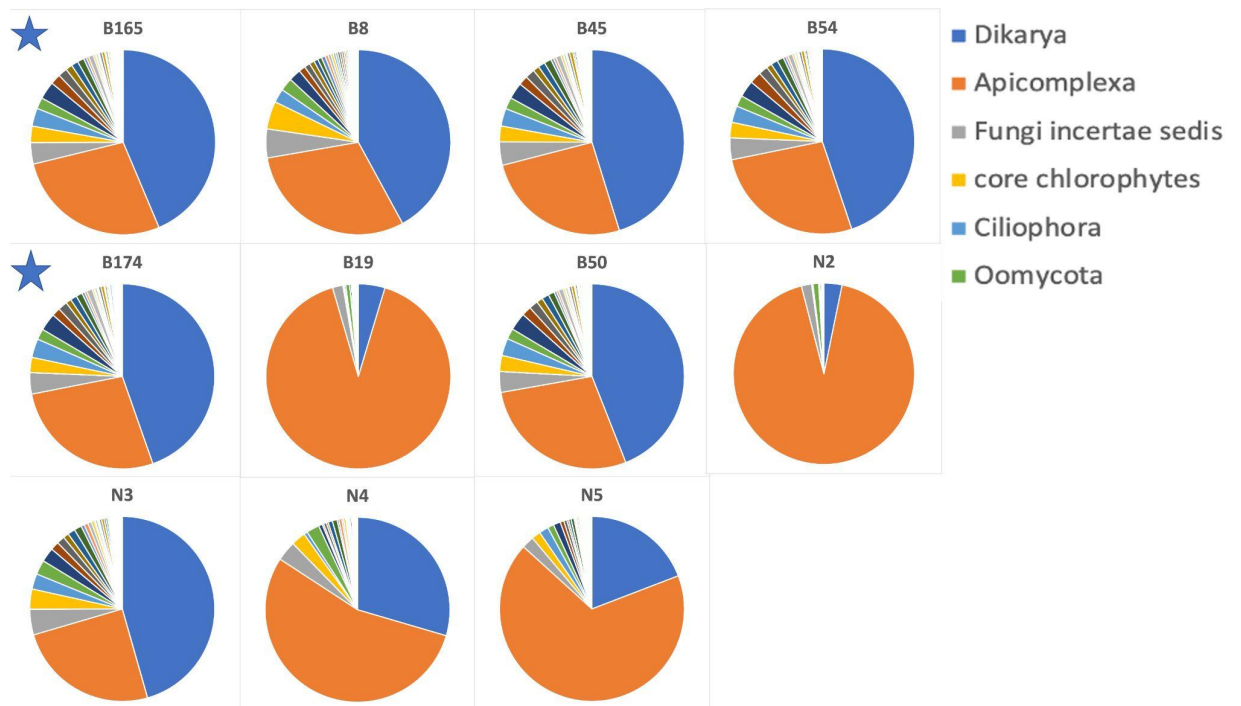


Figure 3.8 The Eukaryotic component of European cattle samples

3.3.4 Analysis of the microbiota in Tanzanian pre existing dataset

Pre-existing faecal DNA metagenomic dataset from Tanzanian cattle (Kibegwa *et al*, 2020) was used. These samples have not been tested for *Cryptosporidium* hence the aim was to identify if any cattle were positive using my metagenomic detection method and study the impact on the microbiome. These cattle are from two different regions of Tanzania and are likely to have a more diverse diet and less antibiotic intake than European cattle. Also samples were tested for the *Cryptosporidium* reads using CENTRIFUGE as shown in Figure (3.9). It was not able to use the detection threshold described in chapter 3 as none of these samples have more than 45 reads per million. This is likely because they have been sampled using different methods and use slightly different sequencing chemistry. However it was still able to compare the samples using the relative abundance of reads mapping to *Cryptosporidium*.

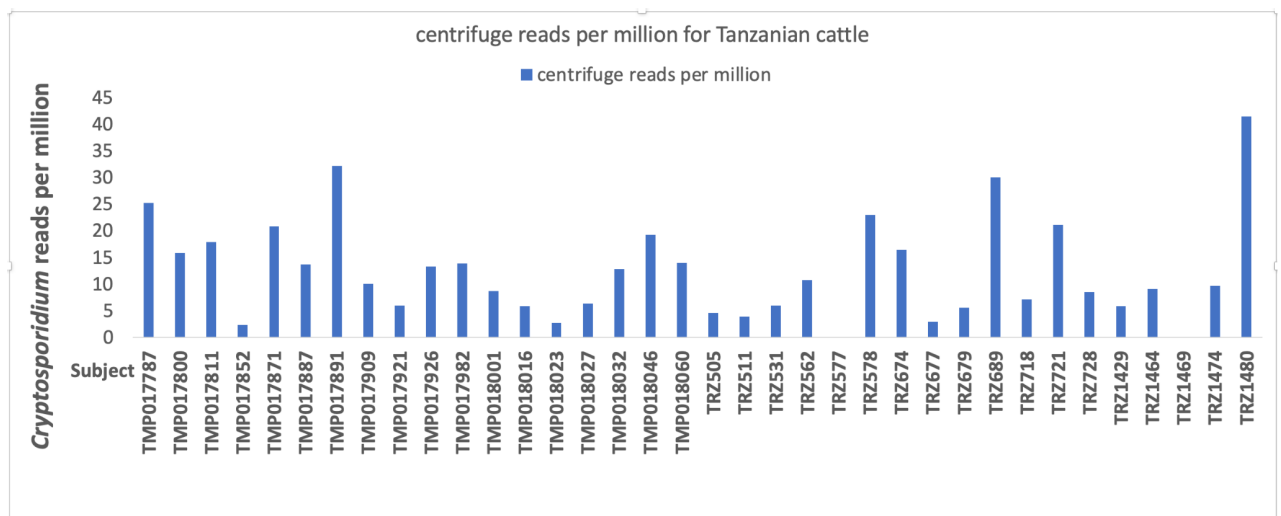


Figure 3.9 *Cryptosporidium* reads for Tanzanian cattle produced by CENTRIFUGE

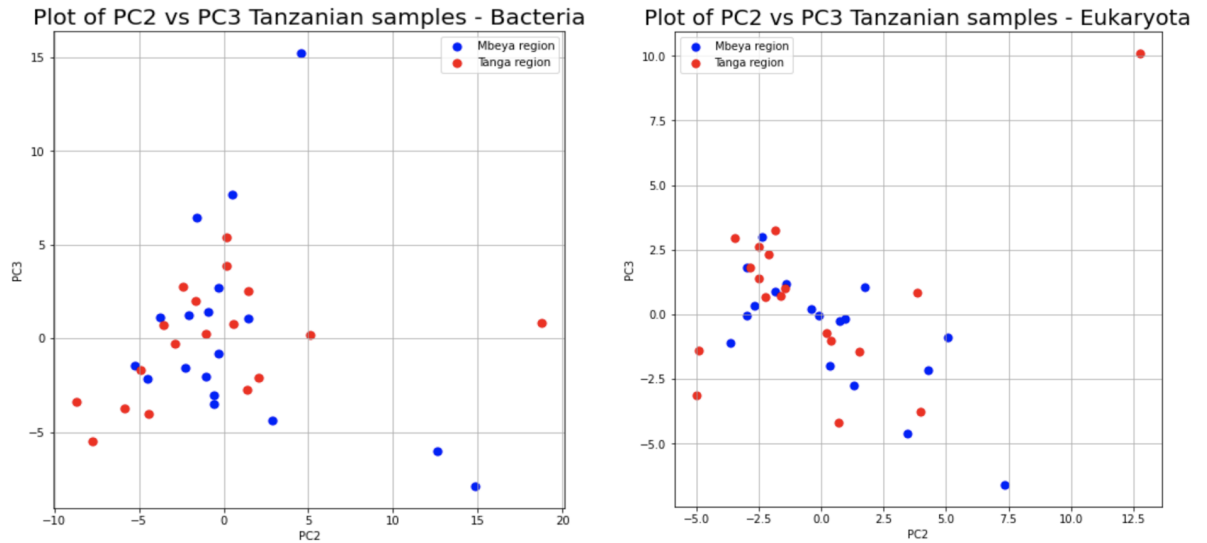


Figure 3.10 PCA analysis for both bacterial and eukaryotic components of Tanzanian cattle

PCA analysis carried out of the prokaryotic and eukaryotic microbiota to see if the samples clustered by region or by infection status. There is no observation of any clustering by region and therefore the dataset was treated as a single group for subsequent analysis.

Both Figures (3.11 and 3.12) indicate that there are much more diverse bacterial microbiomes in cattle from both regions of Tanzania. There are also greater differences between individuals including some individuals where a few phyla dominate which may reflect that these cattle being out-breed and a free ranging population (with a more diverse diet). Noteworthy, the eukaryotic component in the Tanzanian cattle is much more stable as shown in Figures (3.13 and 3.14).

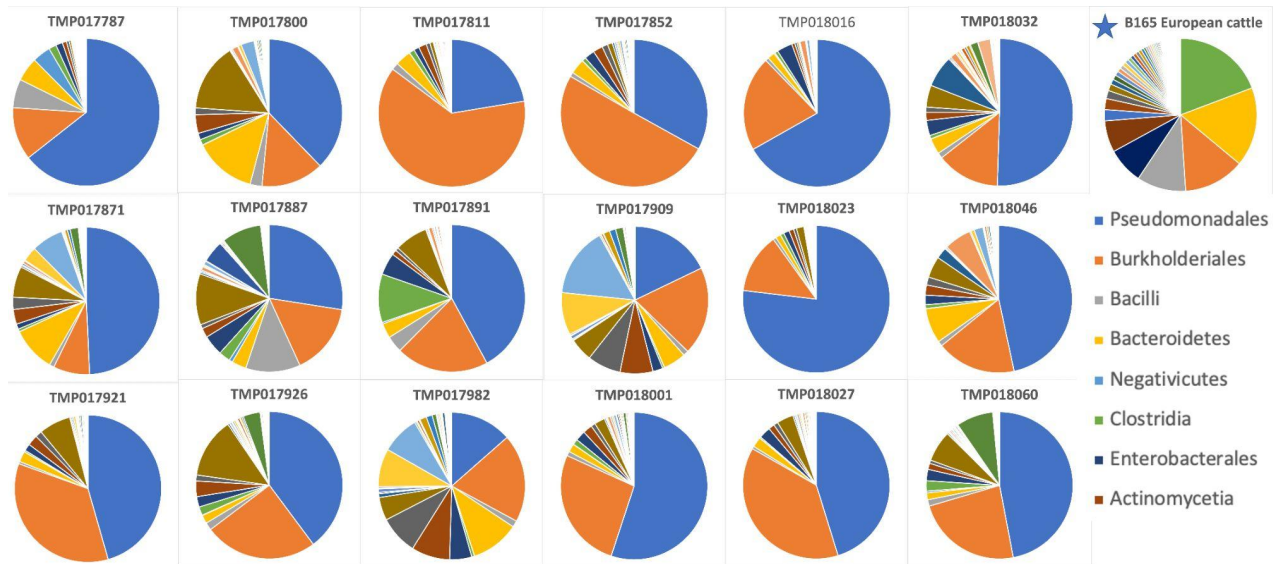


Figure 3.11 Bacterial component in Tanzanian cattle from Rungwe region

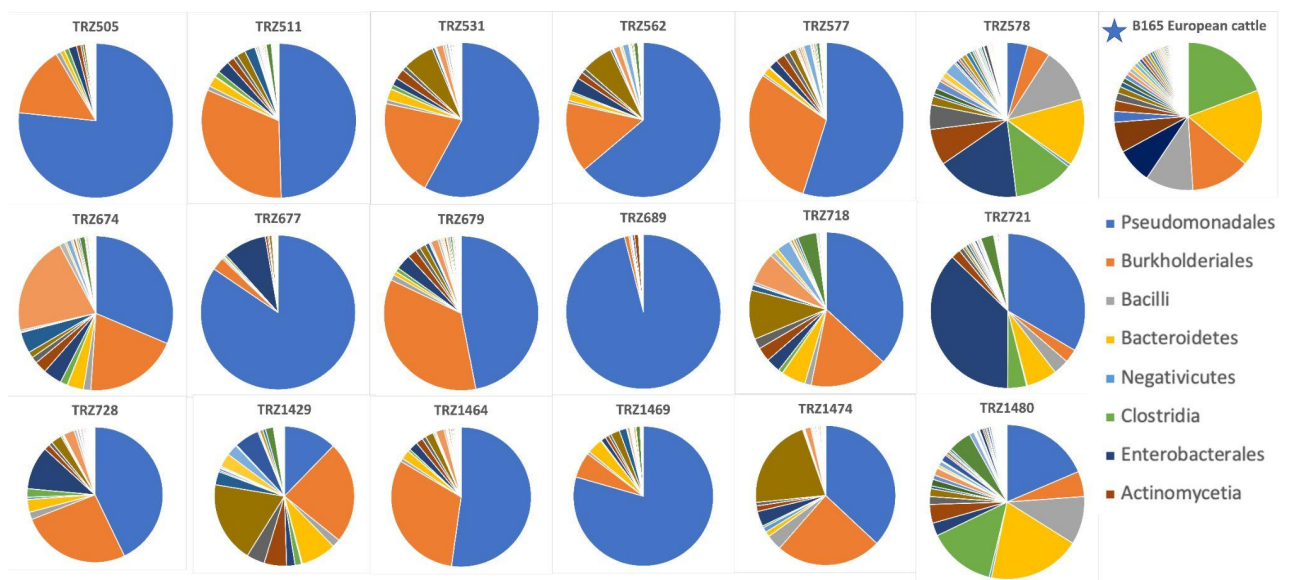


Figure 3.12 Bacterial component in Tanzanian cattle from Lushoto region

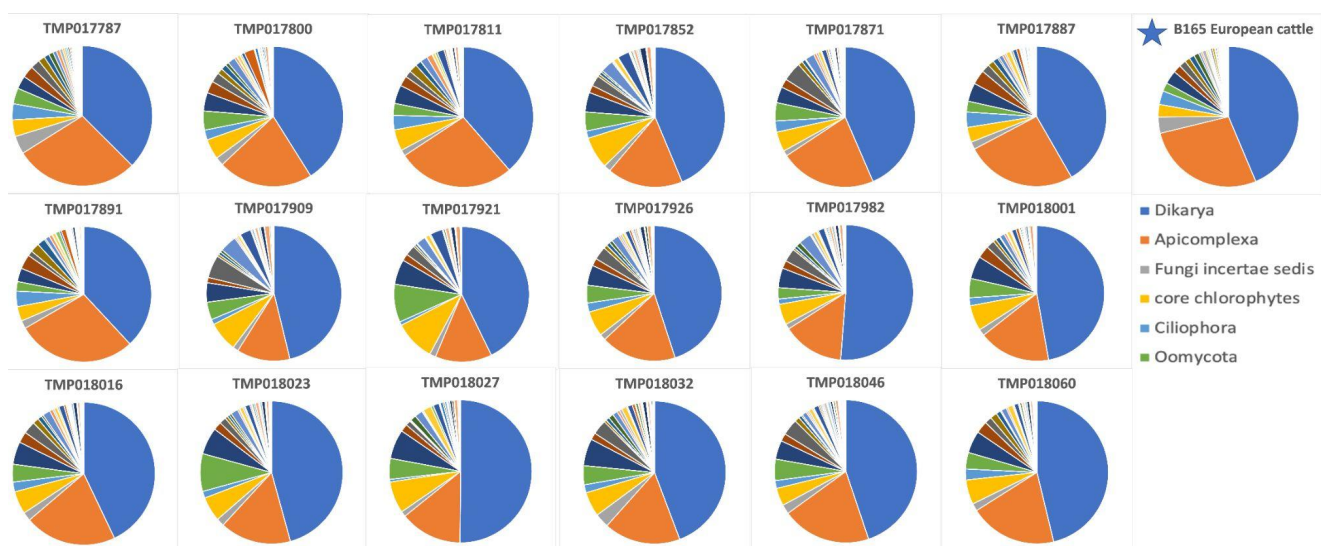


Figure 3.13 Eukaryotic component in Tanzanian cattle from Rungwe region

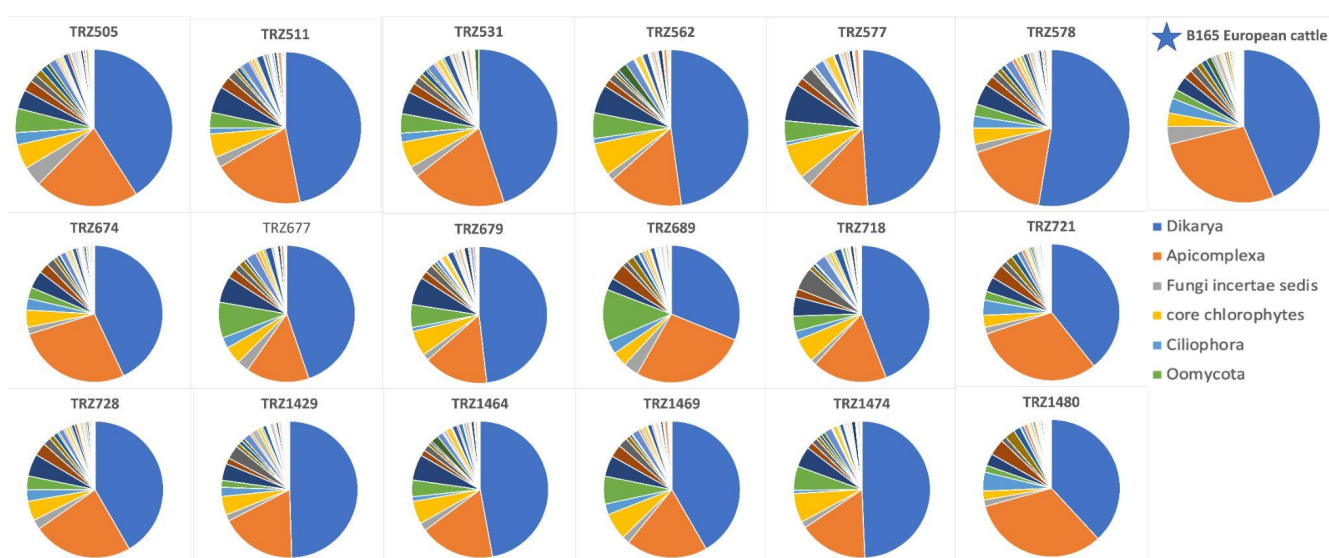


Figure 3.14 Eukaryotic component in Tanzanian cattle from Lushoto region

3.3.5 The interaction between *Cryptosporidium* and *Fusobacterium*

In 2019 it was reported that *Cryptosporidium* infection is associated with an increase of *Fusobacterium* (Ichikawa-Seki *et al*, 2019). Therefore the human samples from Public Health Wales and the bovine samples from the University of Kent were observed, and interestingly in Figure (3.15) on the right hand side there is the human samples where it is observed that there is some fluctuation in the amount of *Fusobacterium*, however it seems to be fairly consistent and it is possible to pick up its variation which could be similar to what reported by (Ichikawa-Seki *et al*, 2019), whereas both samples (UKH104) and (UKUB17) seem to be at odds with the rest of the human samples where *Cryptosporidium* reads per million exceed *Fusobacterium* reads per million. On the left hand side it seems that in the bovine samples there is no any particular relationship in the amount of *Cryptosporidium* and *Fusobacterium*. But looking at sample (N5) it seems it contrasts with the rest of the bovine samples where *Cryptosporidium* reads per million exceed *Fusobacterium* reads per million. However, in cattle samples there does appear to be some level of correlation there than in humans.

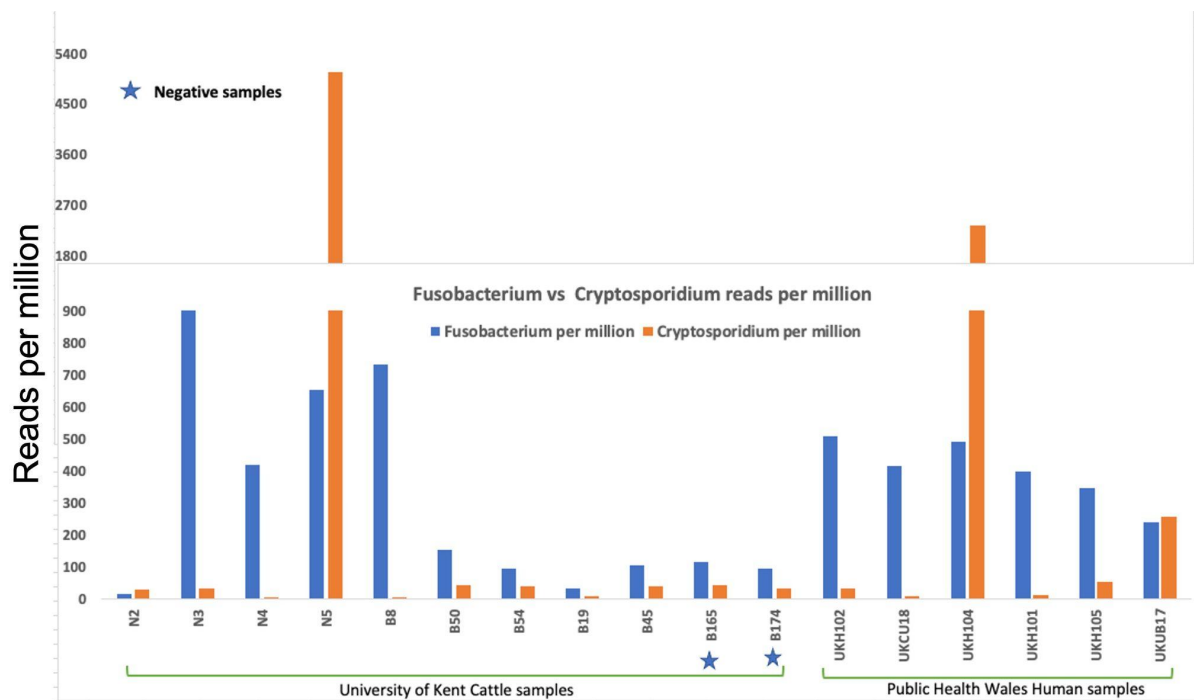


Figure 3.15 Comparison between *Cryptosporidium* and *Fusobacterium* in samples of humans from Public Health Wales and cattle from University of East Anglia

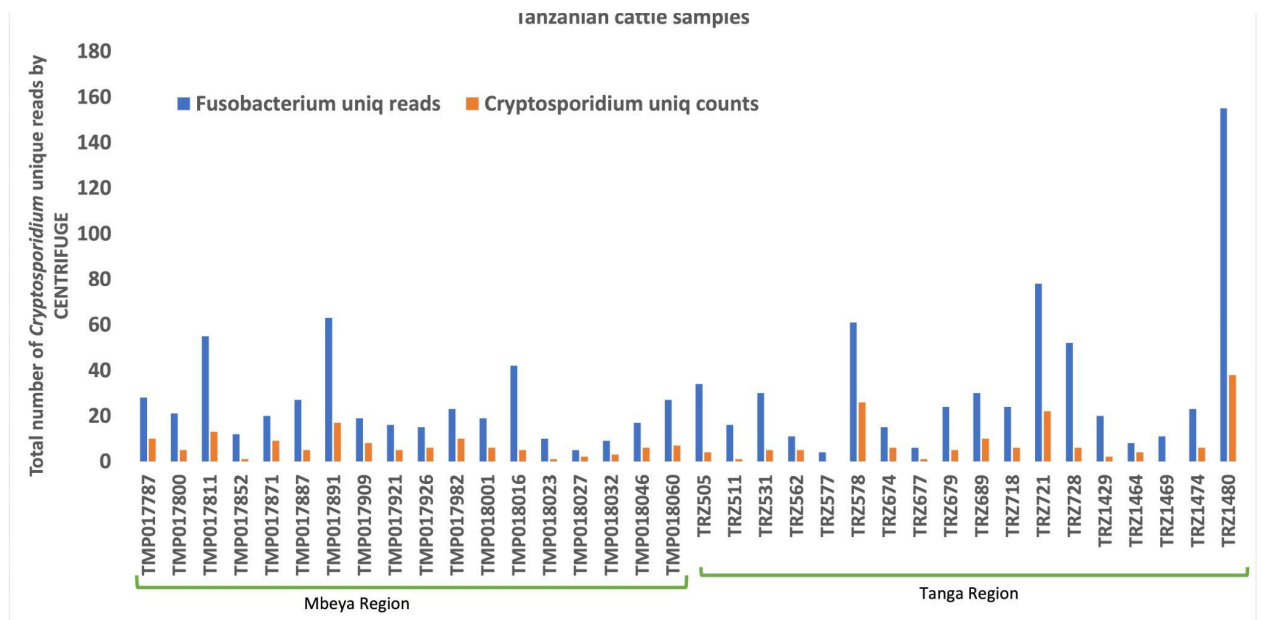


Figure 3.16 Comparison between *Cryptosporidium* and *Fusobacterium* in Tanzanian cattle samples

A comparison was carried out for *Cryptosporidium* to *Fusobacterium* in the Tanzanian cattle samples as presented in Figure (3.16), it looks like there is a strong correlation. Cattle that are less infected with *Cryptosporidium* have lower levels of *Fusobacterium*. Whereas higher levels of *Cryptosporidium* is accompanied by an increase in the level of *Fusobacterium*. Thus, Figure (3.17) shows a strong correlation between *Cryptosporidium* to *Fusobacterium* in the Tanzanian cattle with $R^2 = 0.83$.

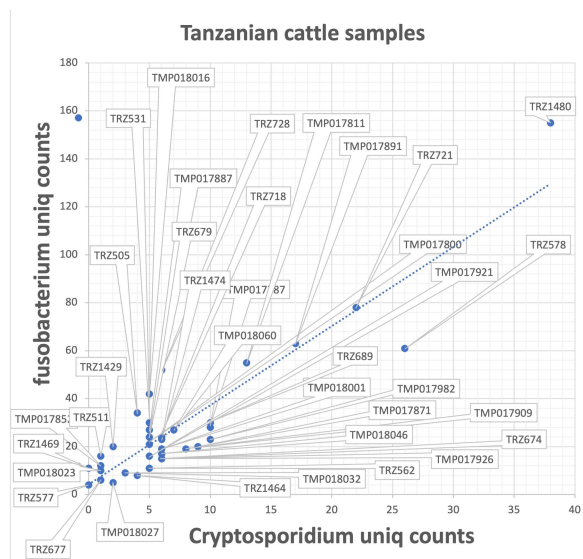


Figure 3.17 correlation between *Cryptosporidium* and *Fusobacterium* Tanzanian cattle samples.

3.4 Discussion

Cryptosporidium infection seems to have a different impact on humans and cattle in both bacterial and eukaryotic components. In both humans and cattle it was possible to note that there can be either a change in the components of the bacterial microbiome or a loss in the bacterial diversity due to severity of *Cryptosporidium* infection as shown in Figures (3.4 , 3.7 , 3.11 and 3.12). However, the eukaryotic component in the human gut tends to have little change but there could be an increase in abundance of *Apicomplexa* (Figure 3.5) which is directly correlated with the *Cryptosporidium* as one would expect. While in cattle, Figures (3.8 , 3.13 and 3.14) show the eukaryotic component to be much more stable which may reflect that European cattle are highly inbred and have a shared controlled diet. However in the case of these cattle the number of reads classified as *Apicomplexa* did not directly reflect the abundance of *Cryptosporidium*, suggesting they could have other protist infections. The *Apicomplexa* did associate with a reduction in bacterial diversity which has previously been associated with gastrointestinal infection. .

The interaction between *Cryptosporidium* and *Fusobacterium* in this study differs between cattle and humans. It was demonstrated that there was a clear correlation between *Cryptosporidium* and *Fusobacterium* in Tanzanian cattle (Figure 3.17) where it is perceived that whenever there is high abundance of *Cryptosporidium* it is followed by high levels of *Fusobacterium*. It was not possible to identify any sign of a relationship between *Cryptosporidium* and *Fusobacterium* in either of the infected human samples from Public Health Wales and the infected European cattle samples from the University of Kent as there was no clear variation of *Fusobacterium* observed (3.18). In the phylum *Firmicutes*, commensal *Clostridia* consist of gram-positive, rod-shaped bacteria and therefore make up a substantial part of the total bacteria in the gut microbiota (Lopetuso *et al*, 2013). They are known to be involved in regulating immune processes and therefore a reduction in *Clostridia* and the associated increase in the proportion of *Bacteroidetes* could be a sign of disbiosis either causing or leading to *Cryptosporidium* infection.

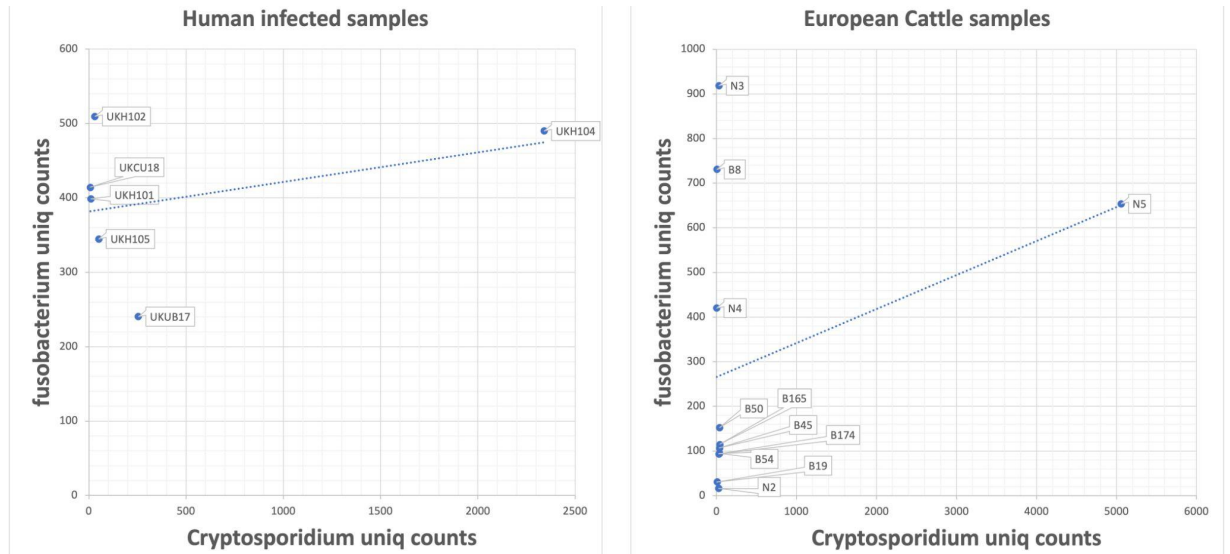


Figure 3.18 Correlation between *Cryptosporidium* and *Fusobacterium* in infected humans and European cattle samples. Where samples (UKH104) from Public Health Wales infected humans and (N5) from the University of Kent infected cattle seem to be severely infected.

An attempt was conducted to clear away samples with reduced bacterial diversity that were outliers in Figure (3.18) for the infected human samples and the infected European cattle samples in order to identify any possible correlation between *Cryptosporidium* and *Fusobacterium*. Therefore samples (UKH104) from infected humans samples and sample (N5) from infected cattle samples were excluded from the correlation chart as shown in Figure (3.19) this analysis then shows no correlation between *Cryptosporidium* and *Fusobacterium* in either the human samples or the European cattle samples.

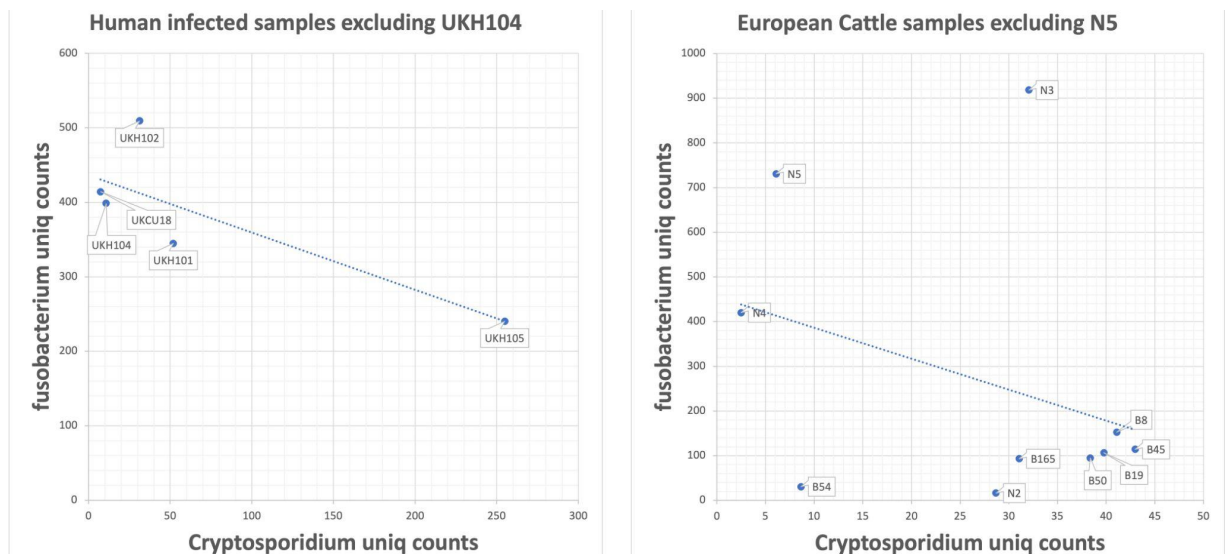


Figure 3.19 Correlation between *Cryptosporidium* and *Fusobacterium* in infected humans without sample UKH104 and European cattle samples without sample N5.

Although *Fusobacterium* seems to be increasing in cattle infected with *Cryptosporidium* (Ichikawa-Seki *et al*, 2019) and my analysis on Tanzanian cattle tends to correspond, the infected European cattle samples provide little support for this relationship except insofar that the highly infected N5 does show a relatively high level of *Fusobacterium*. Future analysis of the metadata for the European cattle such as whether they have been treated with antibiotics etc. may provide additional insight when these are made available. As the data suggests that there could be other apicomplexan parasites present, it is possible that this is confounding the results.

CHAPTER 4

General discussion

Cryptosporidium is an enteric infection in which the parasite damages the surface of the gut epithelium. Disease pathology reflects not just direct damage and dysfunction of the enteric epithelium caused by the parasite itself but the dysbiosis generated in the gut lumen as a result of dysfunction. Here the complexity was addressed by forging a pipeline with bespoke genomic tools and methods which are able to evaluate both the level of parasite present (parasite load) and the extent of dysbiosis (microbiota composition) and these were utilised to provide an initial analysis and optimisation using datasets obtained from veterinary and medical samples in the form of faecal DNAs from *Cryptosporidium* infections.

Initially, commonly employed microbiome bioinformatic tools were evaluated to assess which gave the most accurate results. There have been many studies that focused on the gut microbiome (Kim *et al*, 2017; Richardson *et al*, 2018; Kibegwa *et al*, 2020). However, in this study the effect of *Cryptosporidium* infection on humans and cattle was analysed to compare the microbiome changes between infected and uninfected hosts. As can be seen in chapters 2 and 3, CENTRIFUGE best met the requirements in terms of creating credible results for *Cryptosporidium* detection.

In the new data sets analysed, two heavily infected samples were apparent - sample (UKH104) from the infected humans and sample (N5) from infected calves. Principal component analysis highlighted similarities between these samples - Figures (3.3) and (3.6) - in terms of the changes to their microbiome which indicates that there has been a change in the microbiome diversity associated with the severity of Cryptosporidiosis where both samples (UKH104) and (N5) happen to have an increase in *Bacteroides*. In the outbred, adult, free-ranging Tanzanian cattle it was observed from the PCA (Figure 3.10) that there is more microbiome diversity. This is unsurprising given the age and conditions. Interestingly, the specific diet, cattle type and conditions from the two regions of Tanzania appear to be similar enough that although there are outliers which are quite different, the populations do not appear to have diverged dramatically. The outliers may be due to the increase in *Bacteroides* associated with *Cryptosporidium* infection.

In this study, demonstration has shown that the results from CENTRIFUGE are more useful than the results from BWA, CENTRIFUGE appears to have greater accuracy in quantifying the parasite load. This is reflected in the fact that CENTRIFUGE results show a positive correlation between the number of *Cryptosporidium* reads and *Cryptosporidium* spike-in

levels. However, the results from CENTRIFUGE were slightly different in rabbit samples which could be due to the fact that rabbit faecal microbiome composition is very different to human faecal microbiome composition. However in the control samples, there was a significant number of reads that mapped to *Cryptosporidium*, this may be due to lack of other organisms that may be similar to *Cryptosporidium* in the CENTRIFUGE database causing false mapping, or possibility of contamination in the negative control samples from positive samples. If this work were to be repeated it should be ensured that care is taken to prevent cross contamination between samples.

The human samples and the cattle samples showed corresponding results with previous studies (Ras *et al*, 2015 ; McKenney *et al*, 2017) in terms of changes in the microbiome in cases of *Cryptosporidium* infection. McKenny *et al*. (2017) demonstrated in mice that infection with *Cryptosporidium* caused a decrease in microbial diversity and that recovery time was inversely proportional to age. Ras *et al*. (2015) also demonstrated changes in microbiota in mice and that the changes were similar regardless of *Cryptosporidium* species used and that infected individuals had a microbiota that was more similar to each other than uninfected individuals. The different host species will have different microbiota and will likely respond differently to infection. The data demonstrated that there are very different microbiota in the European cattle compared to the African cattle which likely reflect the different genetics, environment and diet. In the African cattle a strong interaction between *Fusobacterium* and *Cryptosporidium* was observed, while this was not observed in the European cattle where it may indicate that *Fusobacterium* variability could be much less in adults than in calves. Furthermore, other factors for which there was no data, such as antibiotic treatment, could play a role that makes some of these interactions. If this study was to be repeated, collection of more metadata for the samples could shed more light on analysing different hypotheses that could account for the differences in microbiota observed. Such as the age of the cattle, antibiotic treatments and diet.

While the read based methods used here have been useful, there are other bioinformatic methods that could be used to analyse the data if more time was available. Assembly methods attempts can be used to generate draft genomes of the microbiota, this would enable to annotate the microbiomes more effectively and study changes in metabolic pathways present. It could also include doing more analyses of the *Cryptosporidium* genome to characterise what species were present and also identify specific strains.

In summary, this study was undertaken to construct a single CENTRIFUGE based pipeline for the analysis of *Cryptosporidium* or other enteric parasite infections which has given the ability to correlate the loss of bacterial diversity with high levels of infection typically characterized by increases in bacteroides and in cattle with *Fusobacterium*. Previous studies have shown that the microbiota in the gut may be important for susceptibility to *Cryptosporidium* infection.

The results are consistent with this observation as it has been specifically observed an increase in *Fusobacterium* associated with infection. This observation is most apparent in the African cattle samples.

The scope of this study was subjected to certain limitations. Firstly, it is unfortunate that the study was accessing one moment in time during infection and testing if the microbiome of an infected host is different to a non infected one, however it is not possible to know from this study if the changes are due to infection or if they are the cause. Also another limitation of this study, is the lack of enough metadata of cattle samples from the University of Kent such as possibility of Antibiotic treatment, gender and age. Furthermore, an additional uncontrolled factor due to the COVID-19 restriction in the UK is the possibility that cross contamination between samples may occur and thus this could be the reason for having backgrounds in the negative samples. Finally, an important limitation may lie in the fact that faecal samples were used in this study. Although it may not exactly reflect what is going on inside the gut, it is a non invasive and convenient mechanism which can correlate with diseased gut and allows working on more individuals with no requirement for experimentation paperworks.

A Future study could also be conducted to determine the difference in the gut microbiome composition between infected and non infected individuals with cryptosporidiosis with the use of longitudinal study, whereby individuals gut microbiota is tested and sequenced on a monthly basis before infection, during infection and after infection. This can contribute towards understanding if a host had unusual microbiota which therefore was susceptible to infection or that host had normal microbiota and then got infected.

In this study the analysis was carried out on the sequence reads directly but recently more and more studies of metagenomes use a genome assembly step to resolve the genomes of individual species in the microbial community. In order to use this approach deep sequencing would be needed to enable assembly approaches that would provide a much more detailed understanding of community structure and enable identification of genes that may be important to colonisation with *Cryptosporidium*.

References

- Abera, B. *et al.* (2020) Aetiology of acute diarrhoea and antimicrobial usage among children aged under five years at health centres in Bahir Dar, Ethiopia, *Tropical Doctor*, 50(3), 190-194.
- Abrahamsen, M. S. *et al.* (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*, 304(5669), 441-445.
- Acinas, S. G. *et al.* (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons, *Journal of Bacteriology*, 186(9), 2629-2635.
- Alvarez-Pellitero, P. *et al.* (2014) *Cryptosporidium scophthalmi* n. sp. (Apicomplexa: Cryptosporidiidae) from cultured turbot *Scophthalmus maximus*. Light and electron microscope description and histopathological study, *Diseases of Aquatic Organisms*, 62(1-2), 133-145.
- Alves, M. *et al.* (2003) Subgenotype analysis of *Cryptosporidium* isolates from humans, cattle, and zoo ruminants in Portugal, *Journal of Clinical Microbiology*, 41(6), 2744-2747.
- Areeshi, M. Y. *et al.* (2007) Cryptosporidiosis in Saudi Arabia and neighboring countries, *Annals of Saudi medicine*, 27(5), 325–332.
- Balvočiūtė, M. and Huson, D. H. (2017) SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare?, *BMC genomics*, 18(2), 1-8.
- Bedi, B. *et al.* (2015) Il-18 cytokine levels modulate innate immune responses and cryptosporidiosis in mice, *Journal of Eukaryotic Microbiology*, 62(1), 44–50.
- Biesbroek, G. *et al.* (2012) Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection, *PLoS One*, 7(3), e32942.
- Blaser, M. *et al.* (2013) The microbiome explored: recent insights and future challenges, *Nature Reviews Microbiology*, 11(3), 213-217.
- Bones, A. J. *et al.* (2019) Past and future trends of *Cryptosporidium in vitro* research. *Experimental Parasitology*, 196, 28-37.
- Boughner, L. A. and Singh, P. (2016) Microbial ecology: where are we now?. *Postdoc Journal: a journal of postdoctoral research and postdoctoral affairs*, 4(11),3.
- Bouzig, M. *et al.* (2013) *Cryptosporidium* pathogenicity and virulence, *Clinical Microbiology Reviews*, 26(1), 115–134.

- Callahan, B.J. *et al.* (2017) Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women, *Proceedings of the National Academy of Sciences USA*, 114(37), 9966-9971.
- Callaway, T. *et al.* (2010) Evaluation of bacterial diversity in the rumen and feces of cattle fed different levels of dried distillers grains plus solubles using bacterial tag-encoded FLX amplicon pyrosequencing, *Journal of Animal Science*, 88(12), 3977–83.
- Canard, B. and Sarfati, R.S. (1994) DNA polymerase fluorescent substrates with reversible 3'-tags, *Gene*, 148(1), 1-6.
- Carberry, C.A., Kenny, D.A., Han, S., McCabe, M.S. and Waters, S.M., 2012. Effect of phenotypic residual feed intake and dietary forage content on the rumen microbial community of beef cattle. *Applied and Environmental Microbiology*, 78(14), 4949-4958.
- Carey, C. M. *et al.* (2004) Biology, persistence and detection of *Cryptosporidium parvum* and *Cryptosporidium hominis* oocyst, *Water Research*, 38(4), 818-862.
- Cash, H. L. *et al.* (2006) Symbiotic bacteria direct expression of an intestinal bactericidal lectin, *Science*, 313(5790), 1126–1130.
- Chalmers, R. M. *et al.* (2019) *Cryptosporidium*, *Microbiology*, 165(5), 500-502.
- Chappell, C.L. *et al.* (2016) faecal indole as a biomarker of susceptibility to *Cryptosporidium* infection, *Infection and Immunity*, 84(8), 2299–2306.
- Charania, R. *et al.* (2020) Changes in the Microbiome of *Cryptosporidium* infected mice correlate to differences in susceptibility and infection levels, *Microorganisms*, 8(6), 879.
- Cho, Y.I. and Yoon, K.J. (2014) An overview of calf diarrhea—infectious etiology, diagnosis, and intervention, *Journal of Veterinary Science*, 15(1), 1–17.
- Choudhury, P. K. *et al.* (2012) Harnessing the diversity of rumen microbes using molecular approaches. *Livestock Green House Gases: emission and options for mitigation*. Satish Serial Publishing House, Delhi, Section-B, 65-82.
- Clemmons, B. A. *et al.* (2019) Altering the gut microbiome of cattle: considerations of host-microbiome interactions for persistent microbiome manipulation. *Microbial Ecology*, 77(2), 523-536.
- Corrêa-Oliveira, R. *et al.* (2016) Regulation of immune cell function by short-chain fatty acids, *Clinical & Translational Immunology*, 5(4), e73.

- Das, P. *et al.* (2006) Molecular characterization of *Cryptosporidium* spp. from children in Kolkata, India, *Journal of Clinical Microbiology*, 44(11), 4246-4249.
- De León, K. B. *et al.* (2013) Archaeal and bacterial communities in three alkaline hot springs in Heart Lake Geyser Basin, Yellowstone National Park, *Frontiers in Microbiology*, 4, 330.
- Dehority, B. and Orpin, C. (1988) Development of, and natural fluctuations in, rumen microbial populations, In: Hobson, P. (eds.) *Rumen Microbial Ecosystem*. London: Elsevier Applied Science, 151–83.
- DeMenezes, A. B. *et al.* (2011) Microbiome analysis of dairy cows fed pasture or total mixed ration diets, *FEMS Microbiology Ecology*, 78(2), 256–265.
- Deng, W. *et al.* (2008) The use of molecular techniques based on ribosomal RNA and DNA for rumen microbial ecosystem studies: A review, *Molecular Biology Reports*, 35(2), 265–274.
- Dethlefsen, L. *et al.* (2007) An ecological and evolutionary perspective on human–microbe mutualism and disease, *Nature*, 449(7164), 811-818.
- Dillies, M. A. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina High-throughput RNAsequencing data analysis, *Briefings Bioinformatics*, 14(6), 671–683.
- Dong, S. *et al.* (2020) Prevalence of *Cryptosporidium* infection in the global population: A systematic review and meta-analysis. *Acta Parasitologica*, 65(4), 882-889.
- Dowd, S. E. *et al.* (2008) Evaluation of the bacterial diversity in the feces of cattle using 16S rRNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP), *BMC Microbiology*, 8(1), 125.
- Eisenhofer, R. *et al.* (2019) Contamination in low microbial biomass microbiome studies: issues and recommendations, *Trends in Microbiology*, 27(2), 105-117.
- Eisenstein, M. (2020) The hunt for a healthy microbiome, *Nature*, 577(7792), S6-S6.
- Ellis, R. J. *et al.* (2013) Comparison of the distal gut microbiota from people and animals in Africa. *PloS One*, 8(1), e54783.
- El-Sayed, N. M. *et al.* (2005) Comparative genomics of trypanosomatid parasitic protozoa, *Science*, 309(5733), 404-409.
- Fan, J.B. *et al.* (2006) Highly parallel genomic assays, *Nature Reviews Genetics*, 7(8), 632-644.

- Fan, Y. *et al.* (2019) Comparative genomics: how has it advanced our knowledge of cryptosporidiosis epidemiology?, *Parasitology Research*, 118(12), 3195–3204.
- Fasano, A. and Shea-donoghue T. (2005) Mechanisms of disease: the role of intestinal barrier function in the pathogenesis of gastrointestinal autoimmune diseases, *Nature Clinical Practice Gastroenterology and Hepatology*, 2(9), 416–422.
- Fayer, R. *et al.* (2000) Epidemiology of *Cryptosporidium*: transmission, detection, and identification, *International Journal for Parasitology*, 30(12-13), 1305-1322.
- Fayer, R. *et al.* (2005) *Cryptosporidium* bovis n. sp. (Apicomplexa: Cryptosporidiidae) in cattle (*Bos taurus*), *Journal of Parasitology*, 91(3), 624–629.
- Feng, Y. *et al.* (2018) Genetic diversity and population structure of *Cryptosporidium*, *Trends in Parasitol*, 34(11), 997–101.
- Foo, J. L. *et al.* (2017) Microbiome engineering: Current applications and its future. *Biotechnology Journal*, 12(3), 1600099.
- Forrester, S. J. and Hall, N. (2014) The revolution of whole genome sequencing to study parasites, *Molecular and biochemical parasitology*, 195(2), 77-81.
- Frénal, K., *et al.* (2017) Gliding motility powers invasion and egress in Apicomplexa. *Nature Reviews Microbiology*, 15(11), 645-660.
- Fritzler, J. M. and Zhu, G. (2012) Novel anti-*Cryptosporidium* activity of known drugs identified by high-throughput screening against parasite fatty acyl-CoA binding protein (ACBP), *Journal of Antimicrobial Chemotherapy*, 67(3), 609–617.
- Fuks, G. *et al.* (2018) Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome*, 6(1), 1-13.
- Funkhouser-Jones, L. J. *et al.* (2020) Defining stage-specific activity of potent new inhibitors of *Cryptosporidium parvum* growth *in vitro*, *Mbio*, 11(2), e00052-20.
- Gaboriau-Routhiau, V. *et al.* (2009) The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses, *Immunity*, 31(4), 677–689.
- Gardner, M. J. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*, *Nature*, 419(6906), 498-511.
- Gkantiragas, A. and Gabrielli, J. (2021) A Meta-Analysis of the 16S-rRNA Gut Microbiome Data in Honeybees (*Apis Mellifera*). *BioRxiv*.
- Gnu, P. (2007) Free Software Foundation. Bash (3.2. 48)[Unix shell program]

- Gordon, H. A. *et al.* (1963) Effects of microbial flora on cardiac outputs and other elements of blood circulation, *Proceedings of the Society for Experimental Biology and Medicine*, 114(2), 301–304.
- Griffith, G. W. *et al.* (2010) Anaerobic fungi: Neocallimastigomycota, *IMA Fungus*, 1(2), 181–185.
- Hadfield, S. J. *et al.* (2015) Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples, *Bmc Genomics*, 16(1), 1-12.
- Hart, E. H. *et al.* (2018) Meta-proteomics of rumen microbiota indicates niche compartmentalisation and functional dominance in a limited number of metabolic pathways between abundant bacteria, *Scientific reports*, 8(1), 1-11.
- Heiges, M. *et al.* (2006) CryptoDB: a *Cryptosporidium* bioinformatics resource update, *Nucleic Acids Research*, 34(1), 419–422.
- 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(1), 1-15.
- Heselmans, M. *et al.* (2005) Gut flora in health and disease: potential role of probiotics, *Current Issues in Intestinal Microbiology*, 6(1), 1–7.
- Hooper, L. V. and Macpherson, A. J. (2010) Immune adaptations that maintain homeostasis with the intestinal microbiota, *Nature reviews Immunology*, 10(3), 159–169.
- Huang, A. D. *et al.* (2017) Metagenomics of two severe foodborne outbreaks provides diagnostic signatures and signs of coinfection not attainable by traditional methods, *Applied and environmental microbiology*, 83(3), e02577-16.
- Huerta-Cepas, J. *et al.* (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6), 1635-1638.
- Hunter, P. R. and Nichols, G. (2002) Epidemiology, and clinical features of *Cryptosporidium* infection in immunocompromised patients, *Clinical Microbiology Reviews*, 15(1), 145-54.
- Ichikawa-Seki, M. *et al.* (2019) Specific increase of *Fusobacterium* in the faecal microbiota of neonatal calves infected with *Cryptosporidium parvum*, *Scientific Reports*, 9(1), 12517.
- Ivanov, I. I. *et al.* (2008) Specific microbiota direct the differentiation of IL-17-producing T-helper cells in the mucosa of the small intestine, *Cell host & microbe*, 4(4), 337–349.

- Jami, E. and Mizrahi, I. (2012) Composition and similarity of bovine rumen microbiota across individual animals. *PloS One*, 7(3), e33306.
- Jami, E. *et al.* (2013) Exploring the bovine rumen bacterial community from birth to adulthood. *ISME Journal*, 7(6), 1069-1079.
- Jami, E. *et al.* (2014) Potential role of the bovine rumen microbiome in modulating milk composition and feed efficiency, *PLoS One*, 9(1), e85423.
- Janda, J. M. and 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), 2761–2764.
- Janssen, B. and Snowden, J., 2020. Cryptosporidiosis. In *StatPearls [Internet]*. StatPearls Publishing.
- Janssen, P. H. and Kirs, M. (2008) Structure of the archaeal community of the rume, *Applied and Environmental Microbiology*, 74(12), 3619–3625.
- Jewell, K. A. *et al.* (2015) Ruminal bacterial community composition in dairy cows is dynamic over the course of two lactations and correlates with feed efficiency, *Applied and environmental microbiology*, 81(14), 4697-4710.
- Jex, A. R. and Gasser, R.B. (2010) Genetic richness and diversity in *Cryptosporidium hominis* and *C. parvum* reveals major knowledge gaps and a need for the application of next generation technologies--research review, *Biotechnology Advances*, 28(1), 17-26.
- Johnson, D. E. and Ward, G. M. (1996) Estimates of animal methane emissions. *Environmental Monitoring and Assessment*, 42(1), 133-141.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. second edition. New York: Springer-Verlag New York, Inc.
- Karpe, A. V. *et al.* (2021) Cryptosporidiosis modulates the gut microbiome and metabolism in a murine infection model. *Metabolites*, 11(6), 380.
- Kav, A. B. *et al.* (2012) Insights into the bovine rumen plasmidome, *Proceeding of the National Academy of Sciences USA*, 109(14), 5452–5457.
- Khan, A. *et al.* (2018) Genomics and molecular epidemiology of *Cryptosporidium* species, *Acta Tropica*, 184, 1–14.
- Khurana, S. and Chaudhary, P. (2018) Laboratory diagnosis of cryptosporidiosis, *Tropical Parasitology*, 8(1), 2–7.

- Kibegwa, F. M. *et al.* (2020) A Comparison of Two DNA Metagenomic Bioinformatic Pipelines While Evaluating the Microbial Diversity in Feces of Tanzanian Small Holder Dairy Cattle, *BioMed Research International*, 2000, 2348560.
- 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 phylogenetic diversity census of ruminal microbiomes *FEMS Microbiology Ecology*, 76(1), 49–63.
- Kinnebrew, M. A. *et al.* (2012) Interleukin 23 production by intestinal CD103 (+) CD11b (+) dendritic cells in response to bacterial flagellin enhances mucosal innate immune defense *Immunity*, 36(2), pp:276–287.
- Kirkpatrick, B.D. *et al.* (2002) Cryptosporidiosis stimulates an inflammatory intestinal response in malnourished Haitian children, *Journal of Infectious Diseases*, 186(1), 94–101.
- Kissinger, J. C. and DeBarry, J. (2011) Genome cartography: charting the apicomplexan genome. *Trends in Parasitology*, 27(8), 345-354.
- Kittelman, S. *et al.* (2014) Two different bacterial community types are linked with the low-methane emission trait in sheep. *PLoS One*, 9(7), e103171.
- Klatt, N. R. *et al.* (2013) Microbial translocation, immune activation, and HIV disease, *Trends in microbiology*, 21(1) 6–13.
- Krause, D. O. *et al.* (2013) Callaway. Board-invited review: Rumen microbiology: Leading the way in microbial ecology, *Journal of Animal Science*, 91(1), 331– 341.
- Kumar, S. *et al.* (2014) New aspects and strategies for methane mitigation from ruminants, *Applied Microbiology and Biotechnology*, 98(1), 31–44.
- Kvác, M. *et al.* (2009) Prevalence and age-related infection of *Cryptosporidium suis*, *C. muris* and *Cryptosporidium* pig genotype II in pigs on a farm complex in the Czech Republic, *Veterinary Parasitology*, 160(3-4), 319–322.
- Landegren, U. *et al.* (1988) A ligase-mediated gene detection technique, *Science*, 241(4869), 1077-1080.
- Larue, R. *et al.* (2005) Novel microbial diversity adherent to plant biomass in the herbivore gastrointestinal tract, as revealed by ribosomal intergenic spacer analysis and rrs gene sequencing, *Environmental Microbiology*, 7(4), 530–543.

- Lebbad, M. *et al.* (2021) High Diversity of *Cryptosporidium* Species and Subtypes Identified in Cryptosporidiosis Acquired in Sweden and Abroad, *Pathogens*, 10(5), 523.
- Levine, N. D. (2017) *The Protozoan Phylum Apicomplexa*. Published by CRC Press. 218 Page ISBN 9781315897004.
- Ley, R. E. *et al.* (2005) Obesity alters gut microbial ecology. *Proceedings of The National Academy of Sciences USA*, 102(31), 11070-11075.
- Li, F. *et al.* (2019) Comparative metagenomic and metatranscriptomic analyses reveal the breed effect on the rumen microbiome and its associations with feed efficiency in beef cattle, *Microbiome*, 7(1), 1-21.
- Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform, *Bioinformatics*, 25(14), 1754-1760.
- Li, H. *et al.* (2009) The Sequence alignment/map (SAM) format and SAMtools, *Bioinformatics*, 25(16), 2078-2079.
- Li, M. and Wen, J. (2021) Recent progress in the application of omics technologies in the study of bio-mining microorganisms from extreme environments. *Microbial Cell Factories*, 20(1), 1-11.
- Li, R. W. *et al.* (2011) Metagenome plasticity of the bovine abomasal microbiota in immune animals in response to *Ostertagia ostertagi* infection, *PLoS One*, 6(9), e24417.
- Liggenstoffer, A. S. *et al.* (2010) Phylogenetic diversity and community structure of anaerobic gut fungi (phylum Neocallimastigomycota) in ruminant and non-ruminant herbivores, *ISME Journal*, 4(10), 1225–1235.
- Liu, Z. *et al.* (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers, *Nucleic Acids Research*, 36(18), e120.
- Logares, R. *et al.* (2014) Metagenomic 16S rDNA I llumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities, *Environmental Microbiology*, 16(9), 2659-2671.
- Lopetuso, L. R. *et al.* (2013) Commensal *Clostridia*: leading players in the maintenance of gut homeostasis. *Gut Pathogens*, 5(1), 1-8.
- Lozupone, C. A. *et al.* (2007) Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities, *Applied and environmental microbiology*, 73(5), 1576-1585.

- Lundberg, D.S. *et al.* (2012) Defining the core *Arabidopsis thaliana* root microbiome, *Nature*, 488(7409), 86-90.
- Macpherson, A. J. and Harris, N. L. (2004) Interactions between commensal intestinal bacteria and the immune system, *Nature reviews Immunology*, 4(6), 478–485.
- Mapleson, D. *et al.* (2017) KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies, *Bioinformatics*, 33(4), 574-576.
- McCliment, E.A. *et al.* (2006) Colonization of nascent, deep-sea hydrothermal vents by a novel Archaeal and Nanoarchaeal assemblage. *Environmental Microbiology*, 8(1), 114-125.
- McDonald, V. *et al.* (2006) potential role for interleukin-18 in inhibition of the development of *Cryptosporidium parvum*, *Clinical and Experimental Immunology*, 145(3), 555–562.
- McGarvey, J. A. *et al.* (2004) Identification of bacterial populations in dairy wastewaters by use of 16S rRNA gene sequences and other genetic markers, *Applied and Environmental Microbiology*, 70(7), 4267–4275.
- McGuckin, M.A. *et al.* (2011) Mucin dynamics and enteric pathogens, *Nature reviews Microbiology*, 9(4), 265–278.
- McKenney, E. A. *et al.* (2017) Down for the count: *Cryptosporidium* infection depletes the gut microbiome in Coquerels sifakas, *Microbial ecology in health and disease*, 28(1), 1335165.
- McKerr, C. *et al.* (2019) Cross-sectional investigation of household transmission of *Cryptosporidium* in England and Wales: the epiCrypt study protocol. *BMJ Open*, 9(6), e026116.
- Metzker, M.L. *et al.* (1994) Termination of DNA synthesis by novel 3-modifieddeoxyribonucleoside 5-triphosphates, *Nucleic Acids Research*, 22(20), 4259-4267.
- Metzker, M.L. (2005) Emerging technologies in DNA sequencing, *Genome research*, 15(12), 1767-1776.
- Metzker, M.L. (2010) Sequencing technologies—the next generation, *Nature Reviews Genetics*, 11(1), 31-46.
- Milani, C. *et al.* (2019) Colonization of the human gut by bovine bacteria present in Parmesan cheese. *Nature Communications*, 10(1), 1-12.
- Miller, R. R. *et al.* (2013) Metagenomics for pathogen detection in public health, *Genome Medicine*, 5(9), 1-14.

- Minh, B. Q. *et al.* (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5), 1530-1534.
- Mizrahi, I. (2012) The rumen plasmidome: a genetic communication hub for the rumen microbiome, *Mobile Genetic Elements*, 2(3), 152-153.
- Mizrahi, I. *et al.* (2021) The rumen microbiome: balancing food security and environmental impacts, *Nature Reviews Microbiology*, 1-14.
- Mølbak, K. *et al.* (1997) *Cryptosporidium* infection in infancy as a cause of malnutrition: a community study from Guinea- Bissau, West Africa, *American Journal of Clinical Nutrition*, 65(1), 149–152.
- Moraïs, S. and Mizrahi, I. (2019) Islands in the stream: from individual to communal fiber degradation in the rumen ecosystem, *FEMS Microbiology Reviews*, 43(4), 362-379.
- Mu, J. *et al.* (2007) Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome, *Nature Genetics*, 39(1), 126-130.
- Myer, P. R. *et al.* (2015) Rumen microbiome from steers differing in feed efficiency. *PLoS One*, 10(6), p.e0129174.
- Myer, P. R. *et al.* (2016) Microbial community profiles of the jejunum from steers differing in feed efficiency. *Journal of Animal Science*, 94(1), 327-338.
- Myer, P. R. *et al.* (2017) Analysis of the gut bacterial communities in beef cattle and their association with feed intake, growth, and efficiency. *Journal of Animal Science*, 95(7), 3215-3224.
- Myer, P. R. (2019) Bovine genome-microbiome interactions: metagenomic frontier for the selection of efficient productivity in cattle systems. *Msystems*, 4(3), e00103-19.
- Nader, J. L. *et al.* (2019) Evolutionary genomics of anthroponosis in *Cryptosporidium*, *Nature Microbiology*, 4(5), 826-836.
- Nime, F. A. *et al.* (1976) Acute enterocolitis in a human being infected with the protozoan *Cryptosporidium*, *Gastroenterology*, 70(4), 592-598.
- Norman, J. M. *et al.* (2014) Kingdom agnostic metagenomics and the importance of complete characterization of enteric microbial communities, *Gastroenterology*, 146(6), 1459–1469.
- O'Donoghue, P. J. (1995) *Cryptosporidium* and cryptosporidiosis in man and animals, *International Journal for Parasitology*, 25(2), 139-195.

- Oikonomou, G. *et al.* (2013) faecal microbial diversity in pre-weaned dairy calves as described by pyrosequencing of metagenomic 16S rDNA. Associations of Faecalibacterium Species with health and growth, *PLoS One*, 8(4), e63157.
- Omoruyi, B. E. *et al.* (2014) Comparative diagnostic techniques for *Cryptosporidium* infection, *Molecules*, 19(2), 2674–2683.
- Panciera, R. J. *et al.* (1971) Cryptosporidial infection in a calf, *Veterinary Pathology*, 8(5-6), 479–484.
- Parada Venegas, D. *et al.* (2019) Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases, *Frontiers Immunology*, 10, 277.
- Partida-Rodriguez, O. *et al.* (2017) Human intestinal microbiota: Interaction between parasites and the host immune response, *Archives of Medical Research*, 48(8), 690-700.
- Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space, *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572.
- Perez-Sepulveda, B. M. *et al.* (2021) An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. *Genome Biology*, 22(1), 1-18.
- Peterson, D. A. *et al.* (2007) IgA response to symbiotic bacteria as a mediator of gut homeostasis, *Cell host & microbe*, 2(5), 328–339.
- Pimentel, M. *et al.* (2013) Gas and the microbiome, *Current Gastroenterology Reports*, 15(12), 356.
- Pitta, D. W. *et al.* (2010) Rumen bacterial diversity dynamics associated with changing from bermudagrass hay to grazed winter wheat diets, *Microbial Ecology*, 59(3), 511–522.
- Poulsen, M. *et al.* (2013) Methylophilic methanogenic Thermoplasmata implicated in reduced methane emissions from bovine rumen, *Nature Communications*, 4(1), 1428
- Preidis, G. A. and Hotez, P. J. (2015) The newest omics—metagenomics and metabolomics—enter the battle against the neglected tropical diseases, *PLoS Neglected Tropical Diseases*, 9(2), e0003382.
- Pumipuntu, N. and Piratae, S. (2018) Cryptosporidiosis: A zoonotic disease concern, *Veterinary World*, 11(5), 681-686.
- Python Software Foundation. Python Language Reference, version 2.7. Available at <http://www.python.org> (accessed 17/06/2020).

- Ras, R. *et al.* (2015) Perturbation of the intestinal microbiota of mice infected with *Cryptosporidium parvum*, *International Journal for Parasitology*, 45(8), 567-573.
- Rekecki, A. *et al.* (2009) Effects of germ-free rearing environment on larval sea bass (*Dicentrarchus labrax* L.), *Aquaculture*, 293(1), 8–15.
- Ripple, W. J. *et al.* (2014) Ruminants, climate change and climate policy. *Nature Climate Change*, 4(1), 2-5.
- Robertson, L.J. and Gjerde, B.K. (2007) *Cryptosporidium* oocysts: challenging adversaries?, *Trends in parasitology*, 23(8), 344-347.
- Rossignol, J. F. *et al.* (2006) Effect of nitazoxanide in diarrhea and enteritis caused by *Cryptosporidium* species, *Clinical Gastroenterology Hepatology*, 4(3), 320–324.
- Rothschild, D. *et al.* (2018) Environment dominates over host genetics in shaping human gut microbiota, *Nature*, 555(7695), 210-215.
- Ryan, U. N. A. *et al.* (2014) *Cryptosporidium* species in humans and animals: current understanding and research needs, *Parasitology*, 141(13), 1667-1685.
- Saladié, M. *et al.* (2020) Microbiomic analysis on low abundant respiratory biomass samples; improved recovery of microbial DNA from bronchoalveolar lavage fluid, *Frontiers in Microbiology*, 11, 2477.
- Sanger, F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors, *Proceedings of the National Academy Sciences USA*, 74(12), 5463–5467.
- Sasson, G. *et al.* (2017) Heritable bovine rumen bacteria are phylogenetically related and correlated with the cows capacity to harvest energy from its feed. *MBio*, 8(4), e00703-17.
- Santín, M. (2013) Clinical and subclinical infections with *Cryptosporidium* in animals, *New Zealand Veterinary Journal*, 61(1), 1-10.
- Scallan, E. *et al.* (2011) Foodborne illness acquired in the United States—major pathogens, *Emerging Infectious Diseases*, 17(1), 7–15.
- Scheifinger, C. C. and Wolin, M. J. (1973) Propionate formation from cellulose and soluble sugars by combined cultures of *Bacteroides succinogenes* and *Selenomonas ruminantium*, *Applied microbiology*, 26(5), 789-795.
- Schnittger, A. R. *et al.* (2013) Current and prospective tools for the control of cattle-infecting *Babesia* parasites. In *Protozoa: Biology, classification and role in disease* (1-44). Nova Publishers Hauppauge, New York, USA.

- Shahi, S. K. *et al.* (2017) Gut microbiome in multiple sclerosis: the players involved and the roles they play, *Gut Microbes*, 8(6), 607-615.
- Sharpton, T. J. (2014) An introduction to the analysis of shotgun metagenomic data, *Frontiers in plant science*, 5, 209.
- Sheoran, A. *et al.* (2012) Infection with *Cryptosporidium hominis* provides incomplete protection of the host against *Cryptosporidium parvum*. *Journal of Infectious Diseases*, 205(6), 1019-1023.
- Shirley, D. A. T. *et al.* (2012) Burden of disease from cryptosporidiosis, *Current Opinion in Infectious Diseases*, 25(5), 555–563.
- Shoultz, D. A. *et al.* (2016) Addressing *Cryptosporidium* infection among young children in low- income settings: the crucial role of new and existing drugs for reducing morbidity and mortality, *PLoS Neglected Tropical Diseases*, 10, e0004242.
- Soergel, D. A. *et al.* (2012) Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences, *ISME Journal*, 6(7), 1440–1444.
- Srivastava, A. *et al.* (2021) Influence of gut microbiome on the human physiology. *Systems Microbiology and Biomanufacturing*, 1-15.
- Stecher, G. *et al.* (2020). Molecular Evolutionary Genetics Analysis (MEGA) for macOS, *Molecular Biology and Evolution*, 37(4), 1237–1239.
- Sun, Y. *et al.* (2012) A large-scale benchmark study of existing algorithms for taxonomy independent microbial community analysis, *Briefings Bioinformatics*, 13(1), 107–121.
- Suzuki, K. *et al.* (2004) Aberrant expansion of segmented filamentous bacteria in IgA-deficient gut, *Proceedings of the National Academy of Sciences USA*, 101(7), 1981–1986.
- Sylvester, J. T. *et al.* (2004) Development of an assay to quantify rumen ciliate protozoal biomass in cows using real-time PCR, *Journal of Nutrition*, 134(12), 3378–3384.
- Talavera-López, C. and Andersson, B. (2017) Parasite genomics—Time to think bigger, *PLoS neglected tropical diseases*, 11(4), e0005463.
- Thompson, J. D. *et al.* CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic acids research*, 22(22), 4673-4680.

- Tomkinson, A. E. *et al.* (2006) DNA ligases: structure, reaction mechanism, and function, *Chemical reviews*, 106(2), 687-699.
- Tuomisto, H. (2010) A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity, *Ecography*, 33(1), 2–22.
- Tyzzer, E. E. (1907) A sporozoan found in the peptic glands of the common mouse, *Proceedings of the Society Experimental Biology and Medicine*, 5(1), 12–13.
- Tzipori, S. *et al.* (1983) Experimental cryptosporidiosis in calves: clinical manifestations and pathological findings, *The Veterinary Record*, 112(6), 116–120.
- Vaishnava, S. *et al.* (2011) The antibacterial lectin Reg III gamma promotes the spatial segregation of microbiota and host in the intestine, *Science*, 334(6053), 255–258.
- Van Metre, D. C. *et al.* (2008) Infectious diseases of the gastrointestinal tract, *Rebhuns Diseases of Dairy Cattle*, 200-294.
- Vangay, P. *et al.* (2018) US immigration westernizes the human gut microbiome, *Cell*, 175(4), 962-972.
- Wang, W. *et al.* (2015) Metagenomic analysis of microbiome in colon tissue from subjects with inflammatory bowel diseases reveals interplay of viruses and bacteria, *Inflammatory Bowel Diseases*, 21(6), 1419–1427.
- Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews Genetics*, 10(1), 57-63.
- Warrenfeltz, S. *et al.* (2020). Accessing *Cryptosporidium* Omic and Isolate Data via CryptoDB.org, *Methods in molecular biology*, 139-192.
- Weimer, P. J. (1998) Manipulating ruminal fermentation: a microbial ecological perspective, *Journal of Animal Science*, 76(12), 3114–3122.
- Weimer, P. J. *et al.* (2010) Host specificity of the ruminal bacterial community in the dairy cow following near-total exchange of ruminal contents, *Journal of Dairy Science*, 93(12), 5902–5912.
- Whitford, M. F. (2001) Identification of bacteriocin-like inhibitors from rumen *Streptococcus* spp., and isolation and characterisation of Bovicin 255, *Applied and Environmental Microbiology*, 67(2), 569–574.
- Wold, B. and Myers, R.M. (2008) Sequence census methods for functional genomics, *Nature methods*, 5(1), 19-21.

- Wright, A. D. G. *et al.* (2007) Molecular diversity of methanogens in feedlot cattle from Ontario and Prince Edward Island, Canada, *Applied and Environmental Microbiology*, 73(13), 4206–4210.
- Wylie, K. M. *et al.* (2012) Novel bacterial taxa in the human microbiome, *PLoS One*, 7(6), e35294.
- Xiao, L. and Feng, Y. (2017) Molecular epidemiologic tools for waterborne pathogens *Cryptosporidium* spp. and *Giardia duodenalis*, *Food and Waterborne Parasitology*, 8, 14–32.
- Xiao, L. *et al.* (2001) Molecular characterization of *Cryptosporidium* oocysts in samples of raw surface water and wastewater, *Applied and Environmental Microbiology*, 67(3), 1097–1101.
- Xu, P. *et al.* (2004) The genome of *Cryptosporidium hominis*. *Nature*, 431(7012), 1107-1112.
- Yang, X. *et al.* (2016) Use of metagenomic shotgun sequencing technology to detect foodborne pathogens within the microbiome of the beef production chain, *Applied and Environmental Microbiology*, 82(8), 2433-2443.
- Yatsunenko, T. *et al.* (2012) Human gut microbiome viewed across age and geography, *nature*, 486(7402), 222-227.
- Young, V. B. (2017) The role of the microbiome in human health and disease: an introduction for clinicians, *The BMJ*, 356.
- Yui, T. (2014) Age-related detection and molecular characterization of *Cryptosporidium suis* and *Cryptosporidium scrofarum* in pre-and post-weaned piglets and adult pigs in Japan. *Parasitology Research*, 113(1), 359-365.
- Zhang, Y. *et al.* (2018) Caecal infusion of the short chain fatty acid propionate affects the microbiota and expression of inflammatory cytokines in the colon in a fistula pig model, *Microbial Biotechnology*, 11(5), 859-868.

Appendix I: Python script to count elements in one column based on another column in CSV file

```
# This script counts elements in one column based on a specific word given in
another column in CSV file

import pandas as pd

df = pd.read_csv("input_file.txt") # read csv file

df = pd.DataFrame(df[df['SK'] == 'specific word'] ['column
name'].value_counts().reset_index()\
.rename(columns={'index':'phy', 'column name': 'counts'})) #first select name based
on the Superkingom column then use .values_counts() to count the phylum column

df.to_csv (r'output_file.csv')
```

Appendix II: Python script for Principal Component Analysis

```
from sklearn.decomposition import IncrementalPCA
from sklearn.decomposition import PCA
from sklearn.preprocessing import scale
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score
from matplotlib.colors import ListedColormap

dataset = pd.read_csv(r'inputfile.csv') #input file

dataset.head()

x = dataset.iloc[:,0:48].values #selecting phyla columns
y = dataset.iloc[:,50].values

x = StandardScaler().fit_transform(x)

pca = PCA(n_components=10) #number of principal components

principalComponents = pca.fit_transform(x)

principalDataframe = pd.DataFrame(data = principalComponents,
                                  columns =
                                  ['PC1','PC2','PC3','PC4','PC5','PC6','PC7','PC8','PC9','PC10']) #
creating dataframe PC1 to PC10

targetDataframe = dataset[['location']] #creating dataframe for
location

newDataframe = pd.concat([principalDataframe, targetDataframe],axis
= 1) #combining the two dataframes together
```

```

fig = plt.figure(figsize = (8,8))
ax = fig.add_subplot(1,1,1)
ax.set_xlabel('PC1')
ax.set_ylabel('PC2')

ax.set_title('Plot of PC1 vs PC2', fontsize = 20)

targets = ['Mbeya Region', 'Tanga Region']

colors = ['Blue','Red']

for target, color in zip(targets,colors):
    indicesToKeep = newDataframe['location'] == target
    ax.scatter(newDataframe.loc[indicesToKeep, 'PC1']
               , newDataframe.loc[indicesToKeep, 'PC2']
               , c = color
               , s = 50)

ax.legend(targets)
ax.grid()

#drawing explained variance
import numpy as np
import matplotlib
import matplotlib.pyplot as plt

PC_values = np.arange(pca.n_components) + 1
plt.plot(PC_values,          pca.explained_variance_ratio_,          'ro-',
linewidth=2)
plt.title('Scree Plot')
plt.xlabel('Principal Components')
plt.ylabel('Proportion of Variance Explained')
plt.show()

```

Appendix III: CENTRIFUGE output report table showing *Cryptosporidium* reads found in the Tanzanian cattle samples, where it contains 8 columns: Sample name, Correspond name to the taxonomic ID, Taxonomic rank, Length of the genome sequence, Number of reads classified to this genomic sequence, Number of reads uniquely classified to this genomic sequence and Proportion of the genome normalised by its genomic length.

Sample	name	taxID	taxRank	genomeSize	numReads	numUniqueReads	abundance
TMP017787	Cryptosporidium parvum	5807	species	3376882	6	1	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	4	2	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	5	3	0
	Cryptosporidium muris RN66	441375	leaf	6945638	4	4	0
	Cryptosporidium sp. hedgehog/DEU/2008	655746	species	13966	3	0	0
	Cryptosporidium erinacei	1429248	species	10259	1	0	0
TMP017800	Cryptosporidium parvum	5807	species	3376882	2	0	2.93E-244
	Cryptosporidium hominis TU502	353151	leaf	5283275	3	3	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	1	1	0
	Cryptosporidium muris RN66	441375	leaf	6945638	1	1	0
	Cryptosporidium sp. skunk genotype	564991	species	11591	1	0	0
TMP017811	Cryptosporidium parvum	5807	species	3376882	7	2	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	8	3	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	10	3	0
	Cryptosporidium muris RN66	441375	leaf	6945638	5	5	0
TMP017852	Cryptosporidium hominis TU502	353151	leaf	5283275	2	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	2	0	0
	Cryptosporidium muris RN66	441375	leaf	6945638	1	1	0
TMP017871	Cryptosporidium parvum	5807	species	3376882	4	2	0
	Cryptosporidium wrairi	35083	species	12758	2	0	0

	Cryptosporidium hominis	237895	species	883816	2	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	4	2	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	4	1	0
	Cryptosporidium muris RN66	441375	leaf	6945638	4	4	0
TMP017887	Cryptosporidium parvum	5807	species	3376882	2	1	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	1	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	4	1	0
	Cryptosporidium muris RN66	441375	leaf	6945638	4	3	0
TMP017891	Cryptosporidium parvum	5807	species	3376882	4	1	3.88E-285
	Cryptosporidium hominis	237895	species	883816	1	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	5	2	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	4	1	0
	Cryptosporidium muris RN66	441375	leaf	6945638	14	13	0
TMP017909	Cryptosporidium parvum	5807	species	3376882	5	3	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	1	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	3	2	0
	Cryptosporidium muris RN66	441375	leaf	6945638	4	3	0
	Cryptosporidium sp. 192-Guppy	886674	leaf	404	1	0	0
	Cryptosporidium sp. 193 MSM-2012	1237895	species	404	1	0	0
TMP017921	Cryptosporidium parvum	5807	species	3376882	6	2	5.05E-127
	Cryptosporidium hominis TU502	353151	leaf	5283275	4	1	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	5	1	0
	Cryptosporidium muris RN66	441375	leaf	6945638	1	1	0
TMP017926	Cryptosporidium parvum	5807	species	3376882	2	0	0
	Cryptosporidium sp. 1040	202485	species	3846	1	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	1	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	1	0	0
	Cryptosporidium muris RN66	441375	leaf	6945638	7	6	0
TMP017982	Cryptosporidium parvum	5807	species	3376882	3	1	0

	Cryptosporidium sp. #691	128788	species	1969	1	0	0
	Cryptosporidium cf. molnari	271621	species	2656	1	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	5	2	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	6	3	0
	Cryptosporidium muris RN66	441375	leaf	6945638	4	4	0
	Cryptosporidium sp. AHCF-3	495339	species	507	1	0	0
	Cryptosporidium sp. piscine genotype 7	1740366	species	1876	1	0	0
TMP018001	Cryptosporidium parvum	5807	species	3376882	2	2	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	1	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	3	2	0
	Cryptosporidium muris RN66	441375	leaf	6945638	3	2	0
TMP018016	Cryptosporidium parvum	5807	species	3376882	2	1	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	4	1	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	5	2	0
	Cryptosporidium muris RN66	441375	leaf	6945638	2	1	0
TMP018023	Cryptosporidium parvum	5807	species	3376882	1	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	1	1	0
TMP018027	Cryptosporidium hominis TU502	353151	leaf	5283275	1	1	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	1	0	0
	Cryptosporidium muris RN66	441375	leaf	6945638	1	1	0
TMP018032	Cryptosporidium parvum	5807	species	3376882	1	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	2	1	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	2	0	0
	Cryptosporidium muris RN66	441375	leaf	6945638	2	2	0
TMP018046	Cryptosporidium parvum	5807	species	3376882	1	1	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	2	2	0
	Cryptosporidium muris RN66	441375	leaf	6945638	3	3	0
TMP018060	Cryptosporidium parvum	5807	species	3376882	6	2	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	3	0	0

	Cryptosporidium parvum Iowa II	353152	leaf	6825652	6	1	0
	Cryptosporidium muris RN66	441375	leaf	6945638	4	4	0
TRZ505	Cryptosporidium hominis TU502	353151	leaf	5283275	1	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	2	1	0
	Cryptosporidium muris RN66	441375	leaf	6945638	3	3	0
TRZ511	Cryptosporidium parvum	5807	species	3376882	3	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	3	1	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	2	0	0
TRZ531	Cryptosporidium hominis TU502	353151	leaf	5283275	3	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	6	3	0
	Cryptosporidium muris RN66	441375	leaf	6945638	2	2	0
TRZ562	Cryptosporidium parvum	5807	species	3376882	4	3	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	2	1	0
	Cryptosporidium muris RN66	441375	leaf	6945638	1	1	0
TRZ577	Cryptosporidium parvum	5807	species	3376882	1	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	1	0	0
TRZ578	Cryptosporidium parvum	5807	species	3376882	1	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	10	4	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	18	11	0
	Cryptosporidium muris RN66	441375	leaf	6945638	12	11	0
TRZ674	Cryptosporidium parvum	5807	species	3376882	2	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	4	2	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	5	1	0
	Cryptosporidium muris RN66	441375	leaf	6945638	4	3	0
TRZ677	Cryptosporidium parvum	5807	species	3376882	3	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	2	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	1	0	0
	Cryptosporidium muris RN66	441375	leaf	6945638	2	1	0
TRZ679	Cryptosporidium meleagridis	93969	species	201489	2	0	0

	Cryptosporidium hominis TU502	353151	leaf	5283275	3	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	5	3	0
	Cryptosporidium muris RN66	441375	leaf	6945638	2	1	0
TRZ689	Cryptosporidium parvum	5807	species	3376882	7	5	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	3	1	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	2	1	0
	Cryptosporidium muris RN66	441375	leaf	6945638	4	3	0
TRZ718	Cryptosporidium parvum	5807	species	3376882	2	1	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	6	2	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	6	2	0
	Cryptosporidium muris RN66	441375	leaf	6945638	2	1	0
TRZ721	Cryptosporidium parvum	5807	species	3376882	5	2	0
	Cryptosporidium wrairi	35083	species	12758	1	1	0
	Cryptosporidium hominis	237895	species	883816	1	1	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	14	3	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	20	7	0
	Cryptosporidium sp. Taiwan A-4	403649	species	661	1	0	0
	Cryptosporidium sp. Taiwan B2-21	403650	species	695	1	0	0
	Cryptosporidium muris RN66	441375	leaf	6945638	11	8	0
TRZ728	Cryptosporidium parvum	5807	species	3376882	2	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	2	1	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	4	2	0
	Cryptosporidium muris RN66	441375	leaf	6945638	4	3	0
TRZ1429	Cryptosporidium parvum	5807	species	3376882	4	1	0
	Cryptosporidium meleagridis	93969	species	201489	1	0	0
	Cryptosporidium hominis	237895	species	883816	1	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	2	1	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	2	0	0
	Cryptosporidium muris RN66	441375	leaf	6945638	1	0	0

TRZ1464	Cryptosporidium parvum Iowa II	353152	leaf	6825652	2	2	0
	Cryptosporidium muris RN66	441375	leaf	6945638	2	2	0
	Cryptosporidium sp. 192-Guppy	886674	leaf	404	1	0	0
	Cryptosporidium sp. 193 MSM-2012	1237895	species	404	1	0	0
TRZ1469	Cryptosporidium hominis TU502	353151	leaf	5283275	1	0	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	1	0	0
TRZ1474	Cryptosporidium parvum	5807	species	3376882	1	0	1.31E-76
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	2	0	0
	Cryptosporidium muris RN66	441375	leaf	6945638	3	3	0
TRZ1480	Cryptosporidium parvum	5807	species	3376882	3	0	0
	Cryptosporidium andersoni	117008	species	363612	1	0	0
	Cryptosporidium hominis	237895	species	883816	1	0	0
	Cryptosporidium hominis TU502	353151	leaf	5283275	13	4	0
	Cryptosporidium parvum Iowa II	353152	leaf	6825652	18	9	0
	Cryptosporidium muris RN66	441375	leaf	6945638	27	25	0
	Cryptosporidium cuniculus	564424	species	72991	1	0	0