INVESTIGATING THE ROLE OF THE ANAEROBIC PROTIST *BLASTOCYSTIS* IN THE GUT MICROBIOME BY METAGENOMIC ANALYSIS

by

Dandan Zhao

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
April 2020

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Blastocystis are amongst the most prevalent microbial eukaryotes inhabiting the gastrointestinal tracts of mammals. A bioinformatic workflow was developed to detect *Blastocystis* in gut metagenomic data and applied to 996 publicly available metagenomic sequencing datasets from fecal samples of humans and animals. *Blastocystis* incidence was determined to be 52.7% in human and 62.6% in animal samples. A *Blastocystis* subtype-specific distribution was observed both in human and animal carriers and associations between microbial community composition and subtypes was confirmed for humans. Specifically, the *Methanobrevibacter* genus, *Prevotella copri*, and species from the *Firmicutes* phylum were positively associated with the presence of *Blastocystis*. A tool, Eukfinder, was designed to recover protistan genome sequences from metagenomic data and successfully retrieved five near-complete nuclear genomes and mitochondrial genomes of *Blastocystis*. Overall, these bioinformatic workflows for analysis of metagenomic data performed well to detect difficult-to-cultivate protists, investigate their genomic diversity and their impact on prokaryotic microbiota.

LIST OF ABBREVIATIONS USED

ANOVA	ANalysis Of VAriance
BLAST	Basic Local Alignment Search Tool
BMI	Body Mass Index
bp	Base pair
BRIG	Blast Ring Image Generator
CD	Crohn's Disease
FMT	Fecal Microbiota Transplantation
GB	GigaByte
Gbp	Giga base pairs
GI	Gastrointestinal
HMP	Human Microbiome Project
HUMAnN2	HMP Unified Metabolic Analysis Network 2
IBD	Inflammatory Bowel Disease
IBS	Irritable Bowel Syndrome
IGV	Integrative Genomics Viewer
ITS	Internal Transcribed Spacer
LDA	Linear discriminant analysis
LEfSe	LDA Effect Size
M	Million
MAG	MetagenomeAssembled Genome
max.	Maximum
MetaHIT	Metagenomics of the Human Intestinal Tract

MetaPhlAn2	Metagenomic Phylogenetic Analysis 2
MB	Mega Byte
Mbp	Mega base pairs
MG	Metagenomic
MRO	Mitochondrion Related Organelles
mtDNA	Mitochondrial DNA
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
nt	Nucleotide
NW	Non-Westernized
OTU	Operational Taxonomic Unit
PCA	Principal Components Analysis
PCR	Polymerase Chain Reaction
PLAST	Parallel Local Alignment Search Tool
qPCR	Quantitative PCR
rCDI	Recurrent *Clostridioides difficile* infections
rRNA	Ribosomal RNA
SCG	Single Copy Gene
SRA	Sequence Read Archive
SSU rRNA	Small-subunit ribosomal RNA
ST	Subtype
STAMP	STatistical Analysis of Metagenomics Profiles
SVM	support-vector machine

tRNA	Transfer RNA
UC	Ulcerative colitis
W	Westernized
WGS	Whole Genome Shotgun

ACKNOWLEDGMENTS

First of all, I am highly grateful to my supervisor Andrew J. Roger for taking me into his lab, giving me the opportunity to explore a new topic, and supervising the work. I cannot express the gratitude I have towards him for the encouragement, support, and understanding that made it possible for me to successfully complete my research.

I would like to extend my gratitude to my co-supervisor Dayana Salas-Leiva for mentoring me these three years. She is the main developer for some of the bioinformatic tools and has provided kind assistance in all aspects of the project. Without her help I would not have been able to complete my work.

I would also like to thank my supervisory committee members John Archibald and Morgan Langille for their helpful guidance and invaluable insights on my project. Special thanks to Bruce Curtis who provided helpful input with many scripts and his expertise on dealing with sequencing data for some of the data analyses.

My special thanks go to all the members of the Roger lab who have been here over the last few years with all possible help.

Finally, I appreciate the help and support from everyone else in CGEB group for providing feedback on my work and the opportunity to hear inspiring talks. A special thank goes to Roisin McDevitt who helped me with all the paper works regarding my degree.

CHAPTER 1 INTRODUCTION

1.1 GUT MICROBIOME STUDIES IN THE NEXT-GENERATION SEQUENCING ERA

The microbial community that colonizes the animal gastrointestinal (GI) tract, known as the ‘gut microbiota’, is composed of bacteria, archaea, eukaryotes, and viruses. This community of approximately 10^{14} microorganisms 0.5%–30% and 30–76% in industrialized(Fujimura et al. 2010) is increasingly recognized for its important roles in host health and disease conditions (Clemente et al. 2012). The collection of genomes of these microbes, the ‘gut microbiome’, is relevant for the understanding of the structure, function and dynamics of the gut microbiota and their interactions with the host.

Studies investigating the importance of the gut microbiota in host health have gradually gained attention since the 1950s with the development of modern molecular and microbiological techniques (Savage 2001). Originally, most knowledge about gut microbes was gained from culture-based methods, which were laborious and time-consuming. This improved after the 1980s with the development of polymerase chain reaction (PCR) amplification and sequencing of the small subunit (SSU) ribosomal RNA (rRNA) gene that revealed large numbers of novel taxa in fecal samples. A majority of these 16S rRNA sequences belonged to uncultivated species and novel prokaryotes (Suau et al. 1999; Eckburg et al. 2005). In the last decade, culture-independent DNA sequencing technologies have revolutionized this field and next-generation sequencing (NGS) has emerged as a powerful approach to characterize microbial community composition with unprecedented resolution and throughput. The taxonomic profile of the microbiome composition can be obtained either by marker gene-based amplicon analysis or through whole-genome shotgun (WGS) metagenomics. Amplicon approaches typically sequence one or several of the marker genes including prokaryotic 16S rRNA, eukaryotic 18S rRNA, and internal transcribed spacers (ITS) for fungi. Due to the limitation on depth of sequence, the target of prokaryotic amplicon sequencing has shifted from full-length 16S rRNA gene to a shorter region of the gene that contains one or several of nine hypervariable regions of prokaryotic16S rRNA (V1-V9) (Mizrahi-Man et al. 2013). Differences in choice of primers used to amplify different regions can lead to bias with over- or under-representation and relative abundance of specific taxa (Comeau et al. 2017; Laudadio et al. 2018). In contrast, the WGS metagenomic approach sequences random DNA fragments isolated from the environmental samples and offers higher resolution and more sensitivity for studying the compositional and functional profiles of the microbial communities (Ranjan et al. 2016). However, it highly relies on the availability of diverse, well-annotated reference genomes for assignment of taxonomy to microbes inhabiting environments of interest.

Thanks to advances in DNA-sequencing and bioinformatics, a more complete picture of the importance and the role of the gut microbiota has been gained. The gut microbiota contributes to host wellness by supplying the host with nutrients and anti-pathogen substances, regulating and improving the immune system, and maintaining gut integrity and homeostasis (Clemente et al. 2012; Thursby & Juge 2017). Combined results from large gut microbiome research consortia such as the NIH Human Microbiome Project (HMP) and Metagenomics of the Human Intestinal Tract (MetaHIT), as well as, smaller scale studies have provided a more comprehensive view of the diversity and distribution of human-associated gut microbial communities (Qin et al. 2010; The Human Microbiome Project Consortium 2012). Thousands of new bacterial species have been identified and grouped within 12 different phyla, with most falling within the *Bacteroidetes*, *Firmicutes*, *Actinobacteria*, and *Proteobacteria* phyla (Donaldson et al. 2015). Many of the new species are of clinical interest due to their potential anti-inflammatory or anti-infectious roles (Hugon et al. 2015). Interestingly, the gut microbiota is not as diverse as microbial communities from other body sites and a high degree of functional redundancy and inter-individual variability has been observed across samples from different countries (Costello et al. 2009; Schluter & Foster 2012; Moya & Ferrer 2016).

The acquisition, diversification and maintenance of the gut microbiota is affected by multiple factors, as inferred from large-scale population-based metagenomics studies (Lozupone et al. 2012; Zhernakova et al. 2016; Falony et al. 2016). For example, newborns acquire different founder species depending on the delivery method (Rodríguez et al. 2015). Other microbes can rapidly colonize the GI tract under following life events like the introduction of solid food or antibiotic treatments (Rodríguez et al. 2015). Studies have shown that some taxa are inherited from the mother and that the microbiome’s composition is shaped by the host's genetic makeup (Goodrich et al. 2014). However, the composition and the general activity of the gut microbiome can also be influenced by short- and long-term dietary habits (e.g., animal-based vs. plant-based diets, the consumption of processed food, and dietary fibre) (Wu et al. 2011; David et al. 2013; Xu & Knight 2015), age, medical practices (e.g., use of pre-, pro- and antibiotics)(Francino 2016), and the environment (e.g., smoke exposure, hygiene practices and climate)(Lozupone et al. 2012; Chabé et al. 2017).

Although it remains unclear what constitutes a "healthy microbiome" (Zhernakova et al. 2016; Falony et al. 2016), it has been observed that heathier individuals often harbor greater gut microbial diversity and richness, and that the compositional changes in gut microbiota can be associated with illnesses that affect the digestive system and metabolism (e.g., obesity and type 2 Diabetes), immune system (e.g., irritable bowel syndrome (IBS), inflammatory bowel disease (IBD), Crohn's disease (CD), rheumatoid arthritis, etc), cancers (e.g., gastric cancer and colorectal cancer), and also neurological conditions (e.g., autism, anorexia, anxiety and depression, among others (Clemente et al. 2012; Schmidt et al. 2018). Table 1.1 lists a number of examples of changes in the gut microbiota associated with diseases). Experimental and clinical evidence has shown that suppression of dysbiosis, a state lacking microbial diversity and/or richness, together with the restoration of the altered microbiome represents a potential approach to improve host health and a promising avenue for the development of new therapies (Marchesi et al. 2016). Dietary intervention (Cotillard et al. 2013), probiotics (Table 1.2) (Madsen et al. 2001), and fecal microbiota transplantation (FMT) (Seekatz et al. 2014) are potential approaches to restore gut microbial health.

1.2 MICROBIAL EUKARYOTES IN THE GUT

Most gut microbiome studies have been focused on the prokaryotic component, leaving the eukaryotic component (e.g., fungi, helminths, and protists) incompletely charted. Historically, eukaryotes inhabiting the gut were generally assumed to be pathogens, but recent studies have shown that their relationship with the host varies from mutualistic to commensalistic to parasitic (Parfrey et al. 2011; Lukeš et al. 2015). Paleoparasitology

Table 1.1 Changes in the intestinal microbiota associated with human diseases.

Disease categories	Specific diseases	Changes[*] in Microbiota Presence/Function	References
Metabolic disorders	Obesity	↑ *Firmicutes*, *Actinobacter* ↑ *Lactobacillus reuteri* ↑ Glycoside hydrolase and SCPAs(butyrate and acetate) ↓ *Bacteroidetes*, *Bifidobacterium animalis*, *Methanobrevibacter smithii*	Turnbaugh et al. 2006; Million et al. 2012; Koliada et al. 2017
	Type-2 Diabetes	↑ *Lactobacillus* ↓ *Clostridium coccoides*, *Atopobium* cluster, *Prevotella* ↓ Butyrate biosynthesis	Qin et al. 2010; Sato et al. 2014
Immune-mediated /autoimmune diseases	IBS	↑*Escherichia coli* ↓ *Clostridium leptum*, *Bifidobacterium* ↓ Bile acid biotransformation	Duboc et al. 2012
	IBD	↑*Actinobacteria*, *Proteobacteria* ↓*Bifidobacteria*, *Clostridium leptum*, *Clostridium coccoides*, *Lachnospiraceae*, *Faecalibacterium prasnitzii, Roseburia hominis* ↓ Firmicutes/Bacteroidetes ratio	Spor et al. 2011; Perry et al. 2006; Machiels et al. 2014
	Crohn's Disease	↑*Bacteroides ovatus*, *Bacteroides vulgatus* ↓ *Bacteroides uniformis*	Dicksved et al. 2008
	Rheumatoid arthritis (RA)	↑*Prevotella copri* in new-onset RA ↑Microbiota diversity of *Lactobacillus* genus in early RA ↓ *Bacteroides* sp. in new-onset RA	Liu et al. 2013; Scher et al. 2013
Cancer	Gastric cancer	↑ *Helicobacter pylori*	Lathrop et al. 2011
	Colorectal cancer	↑ *Bacteroides fragilis*, *Fusobacterium*, *Campylobacter* sp. ↓ butyrate-producer (*Faecalibacterium*, *Roseburia*)	Wang et al. 2012; Ahn et al. 2013
Neuropsychiatric	Autism	↑ *Bacteroidetes*, *Clostridium* sp., *Lactobacillus*, *Desulfovibrio* ↓ *Bifidobacteria*	Song et al. 2004; Adams et al. 2011
	Depression	↑ *Eggerthella*, *Holdemania*, *Gelria*, *Turicibacter*, *Paraprevotella*, *Anaerofilum* ↓ gut microbiota diversity, *Prevotella* and *Dialister*	Kelly et al. 2016

* Changes relative to healthy individuals in control groups. Increase: ↑. Decrease: ↓.

Table 1.2 Examples of next generation probiotics, their function and potential weakness.

Next generation probiotics	Main functions and mechanisms	Potential weakness	references
Akkermansia muciniphila	Anti-obsogenicity and metabolic syndromes	Positive association with Parkinson disease and multiple sclerosis	Chang et al. 2019
Bacteroides fragilis	Anti-inflammations. It also may enhance efficacy of immune check point inhibitors cancer therapy.	Enterotoxin containing B. fragilis is closely related to colorectal cancer development.	
Bifidobacterium spp.	Some *Bifidobacterium* species strains may enhance the efficacy of Immune Checkpoint Inhibitors cancer therapy	The anti-cancer effects may be strain specific.	
Christensenella minuta	Anti-obsogenicity. Highly heritable in a lean host phenotype.	Not applicable.	
Faecalibacterium prausnitzii	Anti-inflammation by Butyrate production. May ameliorate IBD and CRC.	Not applicable.	
Parabacteroides goldsteinii	Anti-obsogenicity. Ameliorates prediabetes syndromes and liver inflammations.	Not applicable.	
Prevotella copri	Ameliorate prediabetes syndromes	Production of branch chain amino acids (BCAA) that may cause insulin resistance.	
Bacteroides uniformis	Anti-obsogenicity. Anti-inflammation	Not applicable.	Neef & Sanz 2013
Clostridia clusters IV, XIVa and XVIII	Anti-inflammation	Not applicable.	

studies have confirmed that helminths and protists were part of the ancestral human gut microbiota (reviewed by Frías et al. 2013). Taxonomic surveys from different datasets have revealed that the presence of microbial eukaryotes in the human gut is ubiquitous and the prevalence can sometimes be very high. For instance, fungal species were detectable in 98% of samples from an HMP study consisting of 317 fecal samples from 147 healthy volunteers (Nash et al. 2017) and the colonization frequency of certain protists approaches 100% in some rural communities (El Safadi et al. 2014; Morton et al. 2015). DNA-based detection and compositional profiling have revealed that the interplay between gut eukaryotes and gut bacteria is important in training the immune system and potentially causes variation in the virulence of gut-colonizing eukaryotes (Stensvold & van der Giezen 2018).

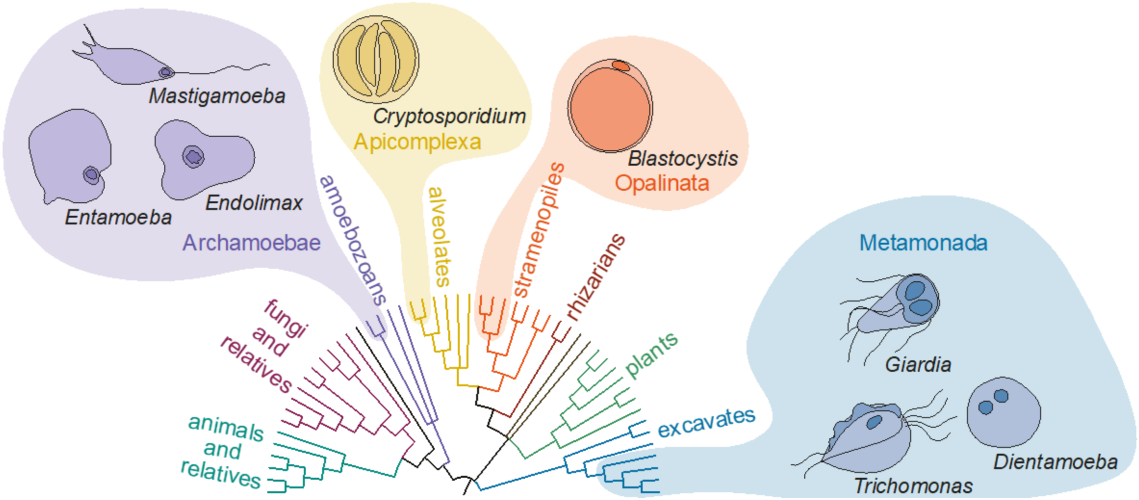
Among the gut-inhabiting eukaryotic groups, the unicellular protists are the most phylogenetically diverse with representation of several major groups of eukaryotes (Figure 1.1). The colonization of the gut by some protists can be stable and widespread in both healthy individuals and groups of patients with infectious bowel diseases (Scanlan & Marchesi 2008; Scanlan et al. 2014). Many gut protists show varied pathogenicity in hosts that can range from asymptomatic colonization to causing mild or severe symptoms, or even death (Lukeš et al. 2015). Such variability may be linked to the diversity in the gut microbial community, differences in host genetic and/or immune system, genotypes of the strain, or the interaction between protists and prokaryotes in the gut (Clemente et al. 2012). Indeed, several studies have demonstrated that protists/gut microbiota interactions are important factors affecting colonization and virulence. Experiments in mice showed that the presence of certain probiotic strains of *Lactobacillus* can inhibit the growth of *Giardia intestinalis*, a flagellated parasitic microorganism that cause diarrhea (giardiasis) in humans and other mammals throughout the world (Humen et al. 2005; Shukla et al. 2008). An *in vitro* experiment by Galván-Moroyoqui et al. (2008) showed co-cultivation of enteropathogenic bacteria strains with the potentially pathogenic protist *Entamoeba histolytica* can increase the frequency with which the protist invades epithelial cells. These results challenge the paradigm of "one microbe, one disease." Investigating how intestinal protists interact with prokaryotic microbiota and the host immune system is clearly important for a comprehensive understanding of the role of the gut microbiome on human health and diseases.

Gut-inhabiting protists possess key adaptions to low oxygen environments, including metabolically distinct mitochondria (e.g., mitosomes in *Giardia* and *Entamoeba*) and anaerobic ATP-generating pathways in the cytoplasm. These features that likely evolved in their free-living ancestors (reviewed in Stairs et al. 2015) have allowed them to colonize animal GI tracts (Mi-Ichi et al. 2009; Jedelský et al. 2011). Functional and comparative genomic studies can be very useful to provide information to understand the genetic diversity in protists, shedding important light on the pathogenesis of these organisms, as well as help in identifying potential interactions between the protists with other gut microbes and the host. Developments in DNA-sequencing technologies in the past two decades have enabled the characterization of the genomes of a variety of protists, especially those with biomedical relevance. For example, more than 12 draft genomes have been published from various *Giardia* isolates. Comparative genomic analyses have helped to identify genome variation between different isolates (Jerlström-Hultqvist et al. 2010) and revealed useful information regarding the diversity of metabolic pathways allowing pathogenic strains to be distinguished from their not-pathogenic counterparts and offering potential targets for drug development.

1.3 BLASTOCYSTIS

The various subtypes of *Blastocystis* sp. (referred to as *Blastocystis*) are amongst the most prevalent microbial eukaryotes colonizing the GI tracts of mammals, birds, reptiles, amphibians and cockroaches (Alfellani, Jacob, et al. 2013). *Blastocystis* are unicellular anaerobes belonging to the group Stramenopila, a major eukaryotic clade encompassing an extremely large diversity of heterotrophic and/or photosynthetic, unicellular and multicellular protists and algae (Derelle et al. 2016). *Blastocystis*, unlike many stramenopiles, lacks a flagellate stage but, in the common vacuolar form has a spherical shape ranging from 10-50 microns in diameter with a large central vacuole and organelles (e.g. nuclei and mitochondrion-related organelles (MROs)) organized around the periphery of the cell. Several less-common morphological forms, including granular,

Figure 1.1 A simplified tree of eukaryotes emphasizing the most commonly occurring anaerobic protists, emphasizing gut-colonizing genera (provided with the generosity of Sergio Muñoz-Goméz and Andrew Roger). Major eukaryote lineages are indicated on the schematic with different colours. The cells are not drawn to scale.



avacuolar, multivacuolar, ameboid and cyst stage, have been reported but there is no agreement on the significance of the different forms (Stensvold & Clark 2016).

Because diverse *Blastocystis* strains are indistinguishable under microscopic examination, it is difficult to assign them distinct species names. To address this problem a consensus terminology has been adopted: '*Blastocystis* sp.' is accompanied by a corresponding subtype (ST) number designation; subtypes are defined as *Blastocystis* clades made up of closely related strains that are more than 4% divergent in the 18S SSU rRNA gene from other subtypes (Stensvold et al. 2007; Clark et al. 2013). Up to 17 STs have been recognized from mammal and bird host using phylogenetic reconstruction with full length 18S SSU rRNA sequences. Of these ST1 - 9 and ST12 are found in humans, although ST1 - 4 are generally most common (Clark et al. 2013; Ramírez et al. 2016). Based on partial SSU rRNA gene sequences, 5 possible novel *Blastocystis* subtypes (ST18 to ST22) have been proposed from animals in wildlife parks in China (Zhao et al. 2017) plus 4 STs (ST23 to ST26) from dairy heifer calves from the USA (Maloney, Molokin, et al. 2019). However, Stensvold and Clark recently urged caution about designation of these isolates as new subtypes until full-length 18S rRNA genes are characterized (Stensvold & Clark 2020).

Blastocystis is transmitted by the fecal-oral route, i.e., by ingestion of cyst-containing water or food (Leelayoova et al. 2008; Caradonna et al. 2017). Animal handlers, pet owners, and people who are exposed to contaminated water are at a higher risk of possible infection (Stensvold et al. 2009; Lee et al. 2012; Nagel et al. 2012). It is estimated that one billion people are infected worldwide (Clark et al. 2013) but the prevalence of *Blastocystis* in human varies with geography and economic status; it is generally much higher in non-industrialized countries (Clark et al. 2013; Stensvold & Clark 2016). However, epidemiologic data gathered to date are heavily influenced by the methods used for detection; studies comparing different detection methods have shown that traditional laboratory technologies like microscopy are more likely to underestimate *Blastocystis* carriage (Roberts et al. 2011; Javanmard et al. 2018). Molecular approaches using PCR amplification of full-length or variable regions of 18S rRNA are now considered the most reliable detection approach (Stensvold & Clark 2016).

Blastocystis' pathogenicity is controversial. There are many reports of *Blastocystis* infections associated with diarrhea, abdominal pain, nausea, bloating, urticaria and various other symptoms (Roberts et al. 2014). Furthermore, experimental studies of ST7 isolates from Singapore have indicated that, *in vitro*, they secrete cysteine proteases that degrade secretory IgA, erode tight-junctions, induce NF-kB-mediated secretion of cytokines and can cause host-cell apoptosis (Stensvold, Tan, et al. 2020). However, recent population-wide studies using molecular markers rarely associate *Blastocystis* carriage with GI disease, and instead find associations with positive health indicators and/or high microbial diversity (Nieves-Ramírez et al. 2018; Tito et al. 2019). The potential pathogenicity of *Blastocystis* can also be obscured by errors in its diagnosis and the lack of comprehensive information on the existing genotypes of subtypes and variation within subtypes. Furthermore, more diverse *Blastocystis* strains are continually being discovered (Stensvold & Clark 2020). The detection of high intra- and inter-subtype genetic variability is suspected to be responsible for the ambiguity in the pathogenicity results in clinical studies (Wu et al. 2014), and experimental setting (Yason et al. 2019). There is a need for more accurate, sensitive, and practical approaches for detecting and genotyping of *Blastocystis*.

Even less is known about the genetic diversity of *Blastocystis* strains on the genomic level. Until recently, efforts to gather genomic information have resulted in the reconstruction and analyses of three *Blastocystis* complete genomes and a few draft genomes. The first characterized *Blastocystis* genome was from the ST7 clade (Denoeud et al. 2011), followed by published descriptions of the ST4 and ST1 genomes (Wawrzyniak et al. 2015; Gentekaki et al. 2017). Rough draft assemblies of ST2, ST3, ST6, ST8, and ST9 (Andersen et al. 2015) were obtained through whole genomic sequencing and deposited in databases, but gene predictions/annotations were not made or remain unpublished. In general, it appears that *Blastocystis* strains have reduced genomes, ranging from ~12 mega base pairs (Mbp) to ~ 18 Mbp displaying a huge range of GC content (39.6% - 54.6%) and significant differences in their gene contents, with ST4 having the least number of protein-coding genes and ST1 the mostb(Table 1.3). The closest relative of *Blastocystis* is *Proteromonas lacertae* that has a genome size of 52 Mbp with a 26.9% GC content. Genes acquired by *Blastocystis* via lateral gene transfer from both prokaryote and eukaryote donors were identified in the ST1 genome by Eme et al. (2017) and these genes seem to be crucial to its adaptation to the gut environment. *Blastocystis* also has a modified mitochondrion (a mitochondrion-related organelle: MRO) capable of anaerobic metabolism presumably for energy production (Gentekaki et al. 2017). Although gene content and order are conserved across mitochondrial genome sequences of the *Blastocystis* STs (Table 1.4), variation of genomic characteristics like number of overlapping genes and gains/losses of start and stop codons on certain genes make them genetically distinguishable (Stechmann et al. 2008). Comparative analyses of differences between nuclear and mitochondrial genomes of different *Blastocystis* STs can be useful to guide future experimental research to shed light on their potential for pathogenicity and the identification of potential targets for anti-protozoan drug development. Therefore, ample genomic information is required to better understand *Blastocystis'* ecological role and clinical significance.

1.4 AIMS OF THIS THESIS

Despite the application of state-of-the-art molecular and immunological methods to study *Blastocystis*, our knowledge of its pathogenicity and roles in the GI tract is still very poor. We lack a complete understanding of its geographic distribution, host specificity, genetic diversity, as well as its interactions with the prokaryotic gut flora. While WGS metagenomic sequencing is a promising means to investigate both compositional and functional aspects of the gut microbiome, the large data sizes and numerous tools available pose challenges for the computational analysis of WGS sequencing datasets. This is especially true for microbial eukaryotes since they usually are a less abundant component of gut metagenome (Laforest-Lapointe & Arrieta 2018) and most of the pipelines and databases developed thus far focus on the prokaryotic components. In this thesis, I describe the metagenomic 'pipelines' that have been developed and applied to gut metagenomes to profile the common intestinal protists, such as *Blastocystis*, to shed light on the role of microbial eukaryotes in the gut microbiome.

In Chapter 2, I describe and apply a metagenomic analysis approach to detect and assign *Blastocystis* STs and establish minimal thresholds to decide whether their presence in WGS fecal samples from humans and animals can be considered an infection or not. To complement these analyses, I examined the relationship between the presence/absence of *Blastocystis* and the composition of the gut prokaryotic microbial community. Since one of the difficulties associated with studying *Blastocystis* in the microbiome is the lack of genomic information for diverse isolates, in Chapter 3, I present Eukfinder, a bioinformatic pipeline to reconstruct draft genomes of microbial eukaryotes, and use *Blastocystis* as a study case to recover nuclear and mitochondrial genomes from human gut metagenomic datasets. Finally, in Chapter 4, I summarize the results from chapters 2 and 3 and discuss outstanding questions that future work should address. The metagenomic approaches developed in this thesis should aid future investigations into the prevalence, functions, physiologies, and evolutionary histories of eukaryotic microbes in the gut microbiome and a variety of other ecosystems.

Table 1.3 Genomic features of published *Blastocystis* reference genomes.

***Blastocystis* subtype & isolate**	**GenBank Accession Number**	**Size (Mbp)**	**Scaffolds**	**GC content (%)**	**protein-coding genes**	**single-coped genes (BUSCO)**
ST1 Nand II	GCA_001651215	16.4683	580	53.00	6544	171
ST4 WR1	GCA_000743755	12.9194	1301	39.70	5707	138
ST7 isolate B	GCA_000151665	18.8172	54	45.30	6020	138
ST2 Flemming	GCA_000963365	12.6931	969	54.00	N/A*	150
ST3 ZGR	GCA_000963385	11.6514	917	52.00	N/A*	140
ST4 BT1	GCA_000963395	11.5409	849	39.90	N/A*	124
ST6 SSI:754	GCA_000963415	15.4178	879	43.10	N/A*	135
ST7 ASY-1	GCA_003575125	10.4299	10257	52.00	N/A*	110
ST8 Dmp/ 08-128	GCA_000963455	12.2390	947	39.70	N/A*	113
ST9 F5323	GCA_000963465	11.7149	871	43.00	N/A*	111

Table 1.4 Genomic features of published *Blastocystis* mitochondrial genomes.

***Blastocystis* subtype & isolate**	**GenBank Accession Number**	**Size (bp)**	**Coding density (%)**	**protein-coding genes**	**Over-lapped genes**	**Total length of overlap (bp)**	**tRNAs**	**GC content (%)**
ST1 Nand II	EF494740	28,385	77.5	27	6	115	16	19.9
ST2 Flemming	KU900235	28,305	78.0	26	8	163	16	19.7
ST3 DMP/ 08-326	HQ909886	28,243	77.5	27	7	113	16	21.6
ST3 DMP/ 08-1043	HQ909887	28,268	77.2	27	7	86	16	21.4
ST4 DMP/ 02-328	EF494739	27,718	77.1	27	8	126	16	21.9
ST4 DMP/ 10-212	KU900236	27,817	76.9	27	8	126	16	21.6
ST6 SSI:754	KU900237	28,806	77.0	26	11	176	16	18.9
ST7 isolate B	CU914152	29,270	77.1	26	7	193	16	20.1
ST8 DMP/ 08-128	KU900238	27,958	77.0	27	9	237	16	22.7
ST9 F5323	KU900239	28,788	77.3	26	11	204	15	18.8

CHAPTER 2 EVALUATING THE ROLE AND IMPACT OF *BLASTOCYSTIS* IN GUT MICROBIAL COMMUNITIES

2.1 INTRODUCTION

2.1.1 The diversity and controversial role of *Blastocystis* in the gut microbiome

Blastocystis sp. is a genus of unicellular eukaryotes (protists) that frequently colonizes the guts of humans and animals. It is estimated that *Blastocystis* colonizes approximately one billion individuals worldwide (Clark et al. 2013). Over the years, *Blastocystis* has been associated with a variety of diseases, prominently GI disorders, including diarrhea, abdominal pain, vomiting and irritable bowel syndrome (IBS). However, evidence for direct pathology caused by *Blastocystis* is very sparse and a causal relationship between the presence of the organism and disease symptoms has not been established conclusively (Roberts et al. 2014).

Epidemiological surveys of the correlation between *Blastocystis* and gastrointestinal syndromic patients show controversial results. Yakoob et al. (2004) detected a high ratio of *Blastocystis* in IBS patients than in healthy controls from Pakistan. A higher prevalence of *Blastocystis* was observed in the IBS group (56 patients) compared to the control group (56 healthy individuals) in France (Nourrisson et al. 2014). A similar pattern was found in IBS patients from Turkey (Dogruman-Al et al. 2009) and Mexico (Jimenez-Gonzalez et al. 2012). However, many other studies found no correlation between *Blastocystis* infection and IBS or other GI disorders. Scanlan et al. (2014) found high *Blastocystis* prevalence (56%, n=105) in healthy adults in Ireland than previously reported from an industrialized country (0.5%–30% in industrialized countries) (Alfellani, Stensvold, et al. 2013) and temporal stability of the protist colonization with the same strain over a period of 6 to 10 years. An attempt to determine whether *Blastocystis* is associated with Crohn's disease (CD) or ulcerative colitis (UC) in a metagenomic survey by Andersen et al. (2015) found a higher positive rate in the healthy group compared to UC patients and no presence of the protist in CD patients. Such a negative correlation between the presence of *Blastocystis* and CD or inflammatory bowel disease (IBD) was also observed in two other more recent studies (Beghini et al. 2017; Tito et al. 2018). Recently, a study focusing on patients with multiple recurrent *Clostridium difficile* infections (rCDI) showed that

Blastocystis can be transmitted from healthy donor to rCDI recipients with fecal microbiota transplantation (FMT). This first-to-be-recorded human-to-human *Blastocystis* transmission did not influence the success rate of the FMT to treat rCDI, nor did it to lead to any GI symptoms in the recipients (Terveer et al. 2019). As evidence mounts for asymptomatic intestinal colonization with *Blastocystis*, it seems very likely that at least some subtypes of this protist may be components of a healthy human gut microbiota (Lukeš et al. 2015).

There are 10 different subtypes of *Blastocystis* that have been found to colonize humans: ST1 to ST9, and ST12. The conflicting reports regarding the pathogenicity of *Blastocystis* may therefore be related to inter-ST and intra-ST variation. *In vitro* and *in vivo* experiments have extensively studied some putatively pathogenic isolates, including those with published genomes like the ST7 isolate B (ST7-B) (from a symptomatic patient in Singapore) , the ST4 isolate WR-1 (ST4 WR-1) (from a laboratory rodent in Singapore), and a ST1 NandII strain from a symptomatic human (Denoeud et al. 2011; Wawrzyniak et al. 2015; Gentekaki et al. 2017). Studies of *Blastocystis* growing with nontransformed rat intestinal epithelial cell line (IEC-6) showed that ST4 WR-1 can induce contact independent apoptosis and increase epithelial permeability of the cell monolayers (Puthia et al. 2006). When incubated with a human colonic cell line (Caco-2), ST7-B significantly increased apoptosis, disrupted epithelial monolayer and increased membrane permeability, but not such effects by rodent isolate ST4 WR-1 (Wu et al. 2014). This difference in pathogenicity may be an indication of host specificity for different isolates. Hussein et al. (2008) observed that isolates from a symptomatic patient with ST3 colonization caused tissue damage in infected rats while isolates from asymptomatic carriers of the same ST had weakly pathogenic effects on infected rats. The existence of asymptomatic and symptomatic individuals with the same subtype could suggest high variation among intra-ST isolates in pathogenicity. Therefore, accurately identifying genotypes of *Blastocystis* infecting symptomatic patients is potentially important for clinical decisions of whether their presence is harmful or not.

2.1.2 The prevalence of *Blastocystis* and methods of detection

Knowledge of *Blastocystis'* pathogenicity and role in the gut microbiome can be hindered by variations in the sensitivity of various diagnosis methods. Traditional approaches to detect *Blastocystis* in stool samples employed microscopic observations, permanently stained smears, or culturing. These methods are time-consuming, highly depend on the type of preparation method and the expertise of the observer, and are unable to distinguish different STs and generally lack sensitivity. PCR assays and amplicon sequencing of the SSU rRNA marker gene have improved diagnostic properties in terms of sensitivity and consistency and allow the subtyping of strains. Consequently PCR-based assays are thought to be the state-of-the-art means for *Blastocystis* detection and subtyping. Conventional or real-time PCR amplification of a barcode region of the SSU rRNA followed by (Sanger) sequencing had been used as a screening tool in clinical microbiology laboratories (Andersen & Stensvold 2016). With the advances in next-generation sequencing (NGS), amplicon sequencing of the *Blastocystis* 18S rRNA gene has shown promising results in detecting *Blastocystis* from fecal or sewage samples with high sensitivity (Tito et al. 2018; Stensvold et al. 2020). The drastic improvement offered by the amplicon method was exemplified with a meta-analysis of the prevalence of *Blastocystis* that showed that up to 20.89% prevalence was detected with such a method while only 8.96 % prevalence was detected by microscopical examination (Javanmard et al. 2018). In recent studies, amplicon-based SSU rRNA gene sequencing identified more than 10 times mixed-subtype infections than PCR based Sanger sequencing (Maloney, Molokin, et al. 2019) and can detect *Blastocystis* from untreated wastewater samples (Stensvold et al. 2020). Despite these improvements, primer bias and chimeras are still unavoidable limitations of PCR and amplicon sequencing and may lead to incomplete detection of all *Blastocystis* in one sample when there are two or more distinct subtypes in the DNA sample (Stensvold & Clark 2020). Furthermore, amplicon-based NGS of 18S rRNA (and combing sequencing of it with 16S rRNA gene sequencing that enables compositional analysis of gut microbiota) provides no direct information on presence or absence of potential molecular determinants of pathogenicity in *Blastocystis* strains identified.

Fortunately, whole genome shotgun (WGS) metagenome sequencing has recently been applied to the gut microbiome for detection and profiling of intestinal protozoa.

Currently, only a handful of studies have used WGS sequencing to investigate *Blastocystis* in gut metagenome samples. For example, Beghini et al. (2017) and Lokmer et al. (2019) detected *Blastocystis* in human samples by mapping reads from metagenomic sequencing of fecal samples to publicly available genomes (STs 1-4 and 6-9) using several measures to avoid false positives. While useful, this approach potentially could miss human-infecting *Blastocystis* strains that currently do not have a reference genome (ST5) and also those that occur in animals, since there are no animal-specific STs (e.g., ST10 – ST26) with genome data available. With the advantages of WGS for analyses of both compositional and functional profiles in the gut microbiome and with the increase of available WGS data from gut metagenomic analyses, there is a strong demand to develop more effective bioinformatic approaches to explore the prevalence and association patterns of gut protozoa.

2.1.3 A strong demands for metagenomic methods to detect and characterize *Blastocystis*

With the advances in high-throughput DNA sequencing, researchers can begin to characterize the relationship between *Blastocystis'* colonization and the composition of the prokaryotic microbiota of the gut. So far, a number of reports have produced contradictory conclusions. For example, using an amplicon approach, Nourrisson et al. found an association of *Blastocystis* colonization with a decrease in protective bacteria in the gut (Nourrisson et al. 2014). In contrast, amplicon-based analyses by Nagel and colleagues suggested that there were no differences in microbiota between *Blastocystis*-positive and -negative patients (Nagel et al. 2016). Additional studies have added to the confusion by demonstrating that *Blastocystis* is a common member of the gut microbiota of healthy people and could be associated with increased prokaryotic diversity and species richness in the gut (Andersen et al. 2015; Audebert et al. 2016; Nieves-Ramírez et al. 2018; Tito et al. 2018).

Most of the studies discussed above were amplicon-based using *Blastocystis*-specific primers for the 18S rRNA gene. However, of these, only the study of Tito and colleagures (Tito et al. 2019) compared the prevalence of *Blastocystis* and its subtypes with prokaryotic microbiota profiles. For example, Tito and colleagues observed that the abundance of *Akkermansia*, a fecal isolate of clinical interest that has been linked to glucose homeostasis (Dao et al. 2016), correlates negatively with the abundances of *Blastocystis* ST3 and positively with ST4 (Tito et al. 2019). Some other studies have noticed that the presence of *Blastocystis* may positively correlate with certain groups of gut prokaryotic microbiota, like the microbes that are prevalent in the '*Bacteroides* enterotype' (Andersen et al. 2015) or the archaeon *Methanobrevibacter smithii* (Beghini et al. 2017), but they did not explain the effects of such correlation.

To address the need for a robust approach to detect *Blastocystis* in WGS metagenome sequencing data, I have developed a novel bioinformatic approach that I applied to 996 gut metagenome datasets from 10 human gut metagenome projects and 13 animal gut metagenome projects from hosts including primates (baboon), other mammals (pigs, cattle) and birds (chickens). Using this approach, I further investigated correlations between the presence/absence of this protist and the abundance of prokaryotic species or metabolic pathways on a 200-sample subset of the above-mentioned human gut metagenome data. This work establishes a new methodology for using WGS metagenomics to detect and analyze gut protists and investigate differences in composition and function of gut microbiome associated with the presence of particular protistan strains.

2.2 METHODS

2.2.1 Workflow for detecting and genotyping *Blastocystis*

A customized bioinformatics workflow for detecting and genotyping *Blastocystis* in gut metagenome sequencing data was developed (Figure 2.1). The raw gut metagenome read files were first processed with Trimmomatic v0.36 (Bolger et al. 2014) to trim adapters and filter out low-quality bases (Phred Q < 15) and short length reads (< 40 bp). Host DNA and Illumina spike-in DNA (Bacteriophage phiX174) were removed using Bowtie2 v2.3.1 (Langmead & Salzberg 2012) by mapping the reads against the host reference genomes downloaded from NCBI (Supplementary Table S1). The resulted metagenomic reads were used as input for read classification against specialized database 1 using Centrifuge v1.0.4 (Kim et al. 2016) and also assembled using MetaSPAdes v3.13.1 (Nurk et al. 2017) or MegaHit v1.1.1-2 (Li et al. 2016) in those cases where MetaSPAdes failed with default parameters.

Read classification was carried out using Centrifuge by mapping the preprocessed metagenomic reads to the specialized database 1 (described in detail below). For human samples, the minimum length of partial hit length and the number of distinct hits were set to 30 and 1, respectively (--min-hitlen 30, -k 1) and for animal samples, "--min-hitlen 25, -k 1".

The metagenome-assembled genomes (MAGs) were processed with Metaxa2 v2.2 beta 9 (Bengtsson-Palme, Hartmann, et al. 2015) to identify nuclear genome- and mitochondrial genome-encoded LSU/SSU rRNA gene sequences. All contigs detected as *Blastocystis* SSU rRNA gene sequences by Metaxa2 were extracted from the assemblies and assigned a *Blastocystis* subtype based on the best match by BLAST (Basic Local Alignment Search Tool (Altschul et al. 1990)) search in the GenBank database with the reference 18S ribosomal DNA sequences of *Blastocystis* STs defined based on Alfellani et al. (2013).

2.2.2 Construction of specialized databases for studying the gut metagenome

Centrifuge database

Centrifuge is a metagenomics taxonomy classification software tool that uses an optimized indexing scheme and contains built-in tools to download genomes from the National Center for Biotechnology Information (NCBI) website and to build custom databases. To maximize Centrifuge's ability to classify gut metagenome data, a custom database was built (here referred to as specialized database 1) by compiling newly published reference genomes from gut microbiota. Prior to constructing the database, all available genome sequences of *Blastocystis* were downloaded from NCBI (accession numbers listed in Table 1.3). Since they may contain contaminant sequences from other organisms, a decontamination step was carried out by mapping the *Blastocystis* genomes against the NCBI nucleotide (nt) database (up to Jan 2019) that did not include any known *Blastocystis* sequences. The contigs in the *Blastocystis* reference genomes that matched over 50% of the total of their length to a bacterium, archaeon, or viral sequence in the NT database with a nucleotide identity of at least 80% were considered as contaminants and were eliminated from the draft genomes (Supplementary Table S2).

Archaeal, bacterial, and viral genomes related to the gut microbiome or without any specific environment listed in the project names were downloaded from NCBI. Genomes with all four assembly levels – complete, chromosome, scaffolds, and contigs – were included. An in-house python script was applied to exclude genomes retrieved from environments other than the GI tract. In addition, bacterial and archaeal genomes of 4,930 species-level genome bins from >9000 human metagenomes (Pasolli et al. 2019) and 913 microbial genomes obtained from rumen metagenomic sequencing (Stewart et al. 2018) were downloaded. Redundant bacterial and archaeal genomes were removed with GTDB-Tk (Chaumeil et al. 2018) and Treemmer (Menardo et al. 2018). For viral genomes, MyCC (Lin & Liao 2016) was used to bin genomes into clusters and a proportion of contigs (40% of the total contigs with a minimal of 20) from each cluster were randomly chosen to be included in the database. Eukaryotic genomes from EupathDB Kraken2 Database (Lu & Salzberg 2018) were also included and any NCBI pre-downloaded genomes for the same species were excluded. Additional eukaryotic genomes for protists, fungi, and animals with complete or chromosome level genome assemblies and mitochondrial genomes were downloaded from NCBI Genbank. Several in-house python scripts were used to build the index files and the centrifuge-build command from Centrifuge was used to build the centrifuge database. The numbers of genomes in each category are listed in Supplementary Table S3.

PLAST database

PLAST (Parallel Local Alignment Search Tool) (Nguyen & Lavenier 2009) is a rapid sequence similarity search tool that is more sensitive than Centrifuge but is not as fast as the latter. To mitigate computational burden, a specialized PLAST database (here referred to as specialized database 2) was built with a subset of reference genomes from archaea, bacteria, eukaryotic, and mitochondrial genomes selected from the complete set of all the downloaded genomes. Specialized database 2 overlaps with specialized database 1 to some degree to enhance the sensitivity of the classification method (Supplementary Table S3). Viral genomes were downloaded from NCBI Refseq database (ftp.ncbi.nlm.nih.gov/refseq/release/viral/, Mar 2019). An in-house python script was used to create an index file. All the genome files were combined into a single fasta file, which was then formatted using the command makeblastdb from BLAST (Altschul et al. 1990).

2.2.3 Comparison of reads classification results between Centrifuge and Kraken2

To verify the read classification results generated by Centrifuge using specialized database 1, different minimal hit lengths (22, 25, 30, and 40) were applied to a subset of datasets and the numbers of reads that could be classified were compared to the results from Centrifuge (parameter "--minhitlen") using NCBI nt (Mar 2018) release as the database. Reads classified as originating from *Blastocystis* by Centrifuge using minhitlen22 were extracted with Recentrifuge (Martí 2019) and mapped against the specialized database 2 by PLAST. An in-house perl script was used to count number of reads hit *Blastocystis* genomes with at least 90% identity over at least 90% of the read length.

To compare the read classification results by Centrifuge and other read classification software, the genomes included in specialized database 1 were used to build a database for Kraken2 (Wood et al. 2019) and KrakenUniq (Breitwieser et al. 2018). KrakenUniq failed due to memory limitations and therefore was not used in the comparison. A subset of metagenomic datasets was run on Kraken2 using default parameters or the setting "-confidence 0.2".

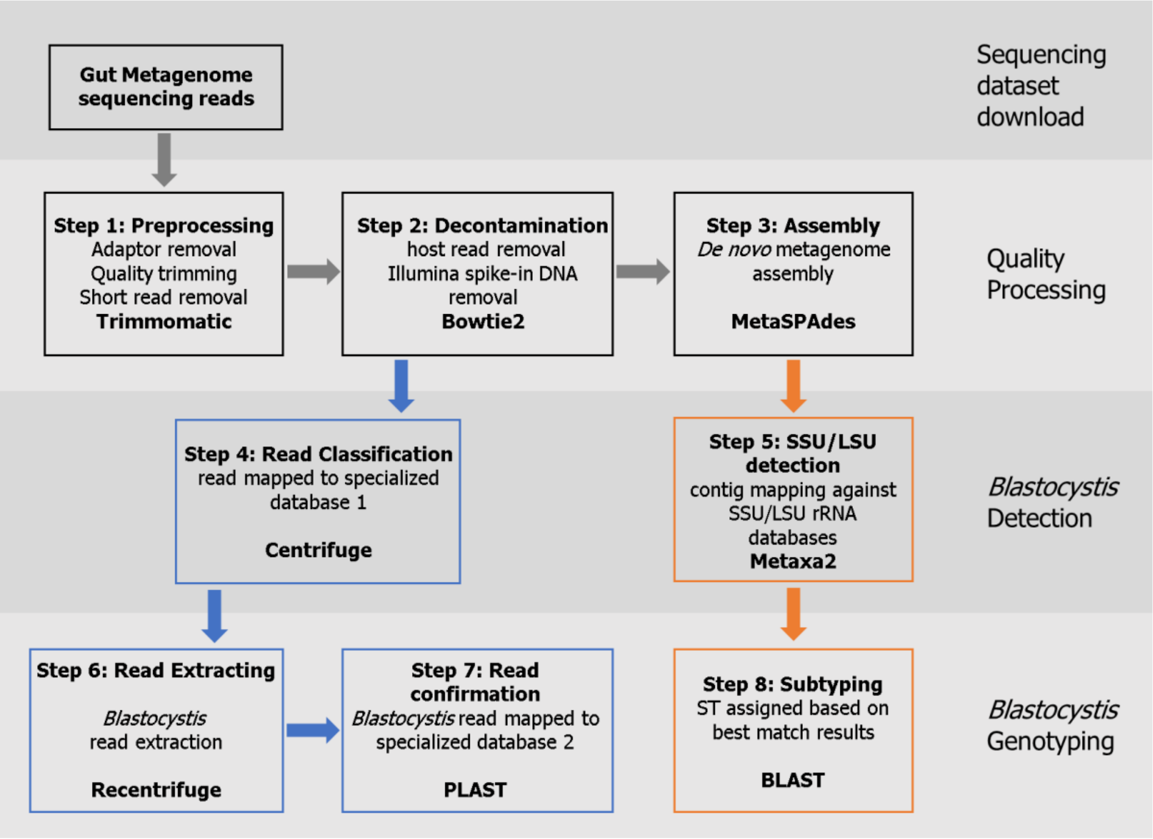
2.2.4 Gut metagenomic datasets

Human and animal gut metagenome samples were obtained from the NCBI Sequence Read Archive (SRA) database using the search terms "(gut metagenome) AND WGS[Strategy] AND METAGENOMIC[Source]". For human gut metagenome samples, projects focusing on infants or diseases unrelated to metabolism were excluded. Ten human gut metagenomic projects (Table 2.1) and 13 animal gut metagenomic projects (Table 2.2) were downloaded from NCBI and analyzed using the workflow described in Section 2.2.1. Due to limitations in time and computational resources, only random subsets of datasets in four of the animal projects (A3, A5, A7, and A9) were analyzed. All datasets consisted of paired-end Illumina sequencing files with an average of 33.4 million (M) reads per sample for human datasets and an average of 39.8 million reads per sample for animal datasets.

2.2.5 Compositional and functional profiling of metagenomes

To investigate the potential differences of taxonomic profiles and metabolic activities in a microbial community with or without *Blastocystis*, HMP Unified Metabolic Analysis Network 2 (HUMAnN2) (Franzosa et al. 2018) with MetaPhlAn2 (metagenomic phylogenetic analysis 2) (Truong et al. 2015) guided species-resolved functional profiling was applied to a subset of 200 human metagenomic datasets (Table 2.3). A series of alignment steps is implemented in HUMAnN2. In the first alignment, MetaPhlAn2 is employed to map reads to a set of ~ 1 million clade-specific marker genes from > 7,500 species and provides microbial taxonomies in the metagenomic samples. Then HUMAnN2 constructs a sample-specific database containing functionally annotated pangenomes of the identified species and maps reads to the pangenome database at the nucleotide-level. In the following step, unaligned reads are translated and undergo Diamond (Buchfink et al. 2015)

Figure 2.1 Bioinformatics workflow for detection and genotyping of *Blastocystis* in gut metagenomics.


Gut Metagenome sequencing reads
Sequencing dataset download
Step 1: Preprocessing
Adaptor removal
Quality trimming
Short read removal
Trimmomatic
Step 2: Decontamination
host read removal
Illumina spike-in DNA removal
Bowtie2
Step 3: Assembly
De novo metagenome assembly
MetaSPAdes
Quality Processing
Step 4: Read Classification
read mapped to specialized database 1
Centrifuge
Step 5: SSU/LSU detection
contig mapping against SSU/LSU rRNA databases
Metaxa2
Blastocystis Detection
Step 6: Read Extracting
Blastocystis read extraction
Recentrifuge
Step 7: Read confirmation
Blastocystis read mapped to specialized database 2
PLAST
Step 8: Subtyping
ST assigned based on best match results
BLAST
Blastocystis Genotyping

search against UniRef databases (Suzek et al. 2007) to predict functions from gene families. Annotated metabolic enzymes from predicted gene families are reconstructed and quantified into complete metabolic pathways based on MetaCyc databases (Caspi et al. 2006). HUMAnN2 reports the percentage of unaligned reads after both steps. The output files were normalized to relative abundance by HUMAnN2 and processed with Microbiome Helper (Comeau et al. 2017) to generate stratified tables.

Westernization and urbanization are known to be important factors affecting the composition of the human gut microbiome (Yatsunenko et al. 2012). Therefore, samples from Europe and USA were categorized as 'westernized', and the rest of the samples from Africa, Asia, and South America were treated as the non-westernized cohort. In the American immigrant project, the European Americans in the control group were treated as 'westernized', while the newly and longer-term immigrants from Thailand were grouped into 'non-westernized'. Samples from children with ages under 10, carriers of helminths and *Blastocystis* carriers with a percentage of *Blastocystis* reads < 0.02% were excluded in this analysis and a total of 200 datasets were chosen for this analysis (Table 2.3).

2.2.6 Statistical analyses

Differences in frequencies for categorical and continuous variables between *Blastocystis* carriers and non-carriers were evaluated using Fisher's exact test and Student's *t*-test, respectively. The predicted differences on taxonomic levels and pathways were represented graphically using the STAMP v2.1.3 software (Parks et al. 2014) with removal of all unclassified reads. Two tailed Welch's *t*-tests, with the Welch's inverted CI method, were conducted in STAMP and used to evaluate differences in the relative abundances of microbial taxa and pathways with respect to presence or absence of *Blastocystis*. In comparisons exceeding two categories, Kruskal-Wallis tests were performed with Tukey-Kramer post hoc comparisons. Unless otherwise stated, a final FDR<0.05 based on Benjamini–Hochberg FDR multiple-test correction was used as a significance threshold. Categories with < 5 samples were excluded in the analysis. Enrichment of prokaryotic species associated with *Blastocystis* presence or absence to the categories of westernized versus non-westernized was performed using the Linear discriminant analysis (LDA) effect size (LEfSe) tool (Segata et al. 2011). Hierarchical clustering for microbial community difference was performed using script metaphlan_hclust_heatmap.py in HUMAnN2 whenever applicable, or using the online tool Heatmapper (Babicki et al. 2016).

Table 2.1 List and characteristics of the human gut metagenomic datasets analyzed in this study. Datasets from ten projects were downloaded from NCBI SRA databases.

Human project & country	Condition	# Subjects	# Samples analyzed	# reads per sample (M) mean ± std	Age (yrs) median (interquartile range)	Reference
H1_Cameroon	Rural population	57	57	23.3 ± 4.9	51 (37-65)	Lokmer et al., 2019
H2_Ethiopia	Healthy	50	50	26.8 ± 9.3	NA	Pasolli et al., 2019
H3_Indonesia, Liberia	Worm infected	24	24	72.1 ± 17.6	24 (12-36)	Rosa et al., 2018
H4_Madagascar	Healthy	111	111	25.6 ± 18.8	35 (25-45)	Pasolli et al., 2019
H5_Peru, USA	Low-income	58	58	20.9 ± 2.9	26 (17-35)	Obregon_Titl et al., 2015
H6_Sweden	Travelers	35	70	77.5 ± 25.8	26 (23-29)	Bengtsson-Palme et al.,
H7_Sweden	Obesity	21	21	10.2 ± 1.3	48 (40-56)	Tremaroli et al., 2015
H8_Tanzania, Italy	Hunter-gatherer	38	38	16.5 ± 9.8	30 (23-38)	Rampelli et al., 2015
H9_USA	Natives	36	36	22.6 ± 2.9	49 (33-65)	Sankaranarayanan et al.,
H10_USA	Immigrants	55	55	19.4 ± 2.9	NA	Vangay et al., 2018
Total		**485**	**520**	**33.4 ± 29.4**	**51 (37-65)**	

Table 2.2 List and characteristics of the animal gut metagenomic datasets analyzed in this study.

Animal project & host	Country	# Samples analyzed (Total samples)	# reads per sample (M) mean ± std	Reference
A1_Baboon	Kenya	48	11.6 ± 8.1	Tung et al.,2015
A2_Cattle	China	30	41.5± 1.8	PRJNA392516
A3_Cattle	France	25 (112) *	122.8 ± 60.2	Li et al., 2018
A4_Cattle	Italy	16	23.3± 5.4	Sandri et al., 2017
A5_Cattle	USA	29 (72) *	156.9 ±149.8	Rovira-Sanz 2017
A6_chicken, pig, cattle**	China	13	14.0 ± 5.0	PRJNA293646
A7_ chicken	China	20	22.9 ± 3.0	Huang et al., 2018
A8_Pig	China, Denmark & France	216 (295) *	28.9 ± 7.1	Xiao et al., 2016
A9_Pig	China	8	25.9 ± 2.0	PRJNA400119
A10_Pig	Denmark	35 (220) *	41.0 ± 29.7	PRJEB26961
A11_Pig	Japan, Gabon	6	58.8 ± 17.0	Ushida K et al., 2016
A12_Pig	Germany	22	13.5 ± 1.8	PRJNA373834
A13_Pig	Spain	8	15.6 ± 4.7	Lanza et al., 2018
Total		**476**	**39.8 ± 54.7**	

* Only a subset of datasets from the project was analyzed. Numbers in () are the total available samples in the project.

** In this project, there are four chicken samples, five cows and four pigs.

Table 2.3 List of metagenomic datasets used for compositional and functional profiling. The presence or absence of *Blastocystis* was based on the results from the detection workflow developed in this study.

Human Project & Country	**# samples Analyzed**	**Positive (n=119)**		**Negative (n=81)**	
		NonW	**W**	**NonW**	**W**
H1_Cameroon	46	41	0	5	0
H2_Ethiopia	28	18	0	10	0
H4_Madagascar	6	3	0	3	0
H5_Peru, USA	32	11	5	2	14
H6_Sweden	31	0	20	0	11
H7_Sweden	20	0	2	0	18
H8_Tanzania, Italy	15	11	0	1	3
H9_USA	9	0	0	0	9
H10_USA	13	7	1	2	3
TOTAL	**200**	**91**	**28**	**23**	**58**

* NonW: Non-westernized, W: westernized

2.3 RESULTS

2.3.1 Benchmarking on specialized databases for gut metagenomes

To improve detection of *Blastocystis* and other gut microbe sequences in gut metagenomic datasets, specialized databases containing representative genomes of gut microbes were constructed. Over 60,560 genomes were downloaded from various sources and selected based on factors including the environment from which the microbe was isolated, taxonomic redundancy, and genome diversity. The final specialized database 1 for Centrifuge contains 32,402 genomes (93.3 GB total) and the PLAST specialized database 2 contains 9,345 genomes (17.8 GB total; Supplementary Table S3). The newly built specialized database 1 dramatically improved the percentage of reads that can be assigned a taxonomy with Centrifuge when compared to the assignments made using the NCBI nt database (from 30% - 60% to 80% - 95%, Figure 2.2).

To minimize false results in taxonomy classification, the impact of the Centrifuge parameter, minimal hit length ("--minhitlen"), was examined by using 22 (default), 25, 30 and 40 bps for human gut metagenomic samples. The number of reads classified as *Blastocystis* was compared with the results using PLAST searches against the specialized database 2 (Figure 2.3) (PLAST is similar to but faster than BLAST yet much slower than Centrifuge). A read that aligns to contigs in *Blastocystis* genomes with > 90% identity over >90% of the read length was considered a real hit. Centrifuge results with minimal hit length of 30 bp generated results most similar to PLAST and therefore this parameter setting was chosen for subsequent analyses of human gut metagenome samples. Since most of the *Blastocystis* subtypes in animal hosts have no available genomes (except ST4 from rat), a less stringent Centrifuge parameter "--minhitlen 25" was used for detecting potential *Blastocystis* sequences in animal gut metagenome samples.

A subset of human samples was analyzed using Kraken2 with the same specialized database 1 used for Centrifuge to verify the accuracy and sensitivity of Centrifuge. The classification results by Centrifuge (with "--minhitlen 30") were very similar to Kraken2 (with the parameter "--confidence 0.2") (Figure 2.4). With similar computing time, Kraken2 required slightly more memory than centrifuge. Considering the computational resources and time for analysis, Centrifuge was therefore chosen for downstream analyses.

Figure 2.2 Comparison of percentage of reads that can be assigned a taxonomy by Centrifuge (default parameters) using the default database (NCBI nt) vs. the newly-built specialized database 1. The datasets were from the US immigrant gut microbiome project (Vangay et al., 2018)

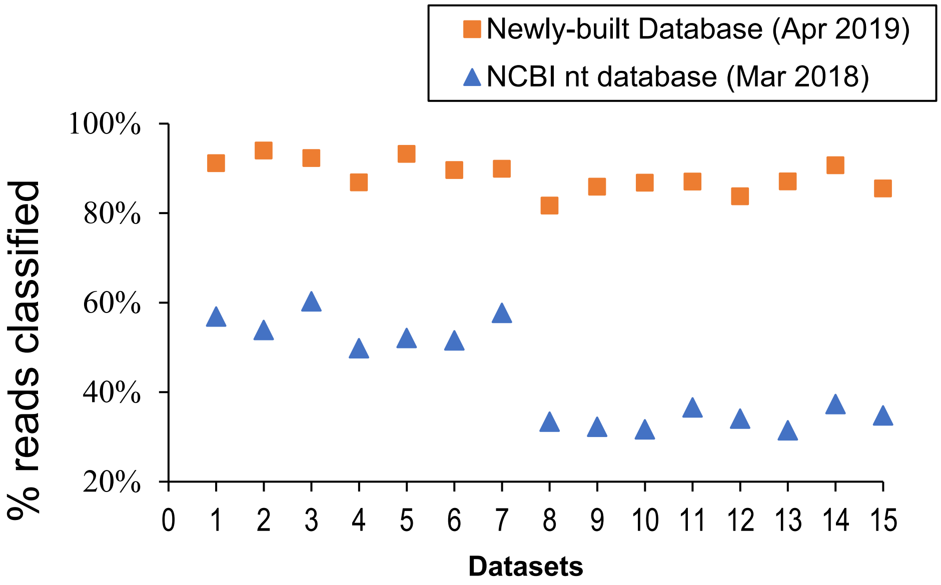


Figure 2.3 Number of reads classified as *Blastocystis* by Centrifuge using different parameter minimal hit length (minhitlen) and by PLAST with identity > 90% over >90% of the read length. The datasets were from Latin gut microbiome projects (Pehrsson et al., 2016).

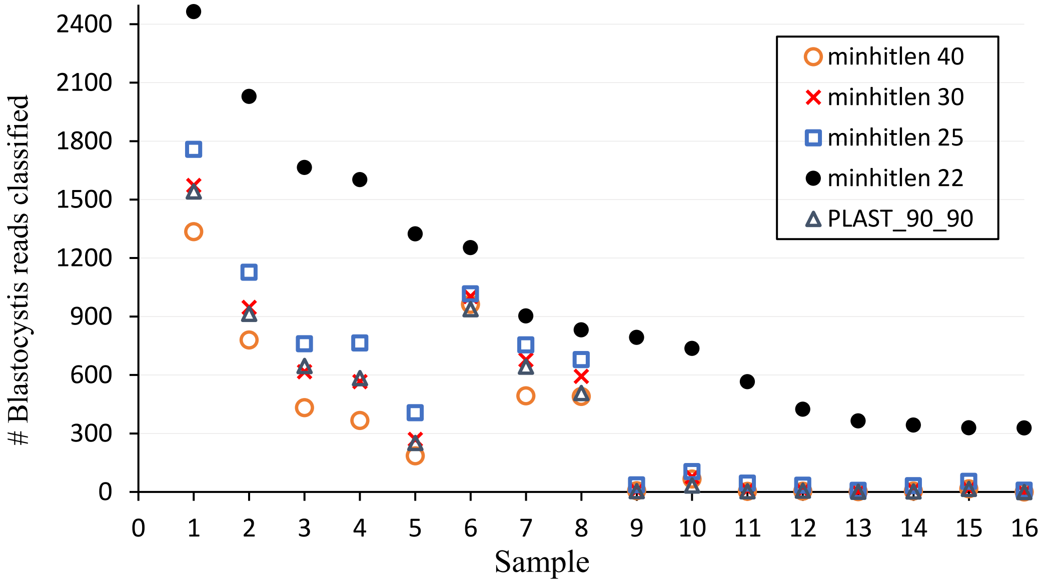
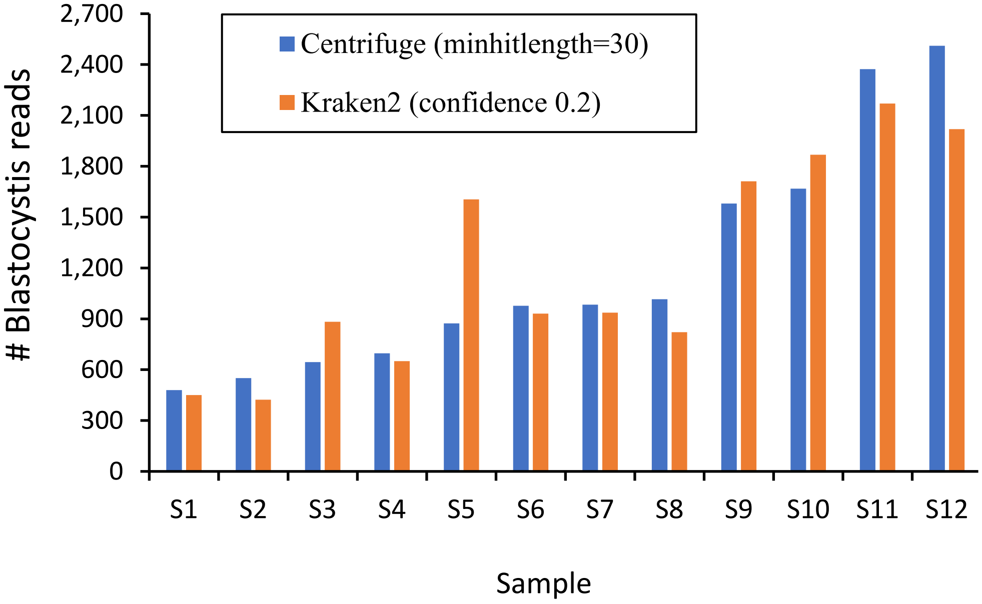


Figure 2.4 Comparison between the results from Centrifuge and Kraken2 on reads classified as *Blastocystis* using specialized database 1. The datasets were from Latin gut microbiome projects (Pehrsson et al., 2016).



2.3.2 Detection of *Blastocystis* using metagenomics

To facilitate large-scale investigation of the prevalence of *Blastocystis* in human and animal gut metagenomes, a bioinformatic workflow (Figure 2.1) was developed and applied to 23 published large metagenomic projects (Table 2.1 and 2.2). Overall, 996 metagenomic datasets from 961 subjects (485 humans and 476 animals) from 18 different countries in four continents (Africa, Asia, Europe, and South/North America) were analyzed. For human samples, this study focused on subjects from countries with potentially high infection rates like Africans or Asians or people from non-westernized backgrounds, such as new immigrants in America (Vangay et al. 2018) or native Americans (Sankaranarayanan et al. 2016). Most samples corresponded to healthy people, although some were from individuals with helminth infections (Rosa et al. 2018), obese patients with gastric bypass surgery (Tremaroli et al. 2015), and university student travelers (Bengtsson-Palme, Angelin, et al. 2015). Seven of the ten published human studies and most of the 12 animal projects were exclusively focused on investigating the prokaryotic components of the microbiome. Animal samples were mainly chosen from cattle and pigs that had several projects to compare the results.

To detect *Blastocystis*, the total number of *Blastocystis* reads in each sample were determined based on taxonomy classification and the LSU/SSU rRNA gene sequences were extracted from the MAG. I defined a human dataset as positive for a *Blastocystis* ST if the sample fulfilled one of the following criteria (Table 2.4):

(1) there were more than 300 pairs of reads that could be classified to the corresponding *Blastocystis* ST by Centrifuge and also had > 200bp of the large subunit (LSU) and/or small subunit (SSU) rRNA gene sequences detected in the MAG by Metaxa2. To diminish false positives caused by potential cross contamination during DNA extraction or errors during/after sequencing due to sample bleeding (or index mis-assignment), the minimum number of *Blastocystis* reads in each project must be larger than 0.015% × the maximum number of *Blastocystis* reads in the same project.

(2) For samples with >300 reads or >0.015% × max. # *Blastocystis* reads but no *Blastocystis* LSU/SSU rRNA sequences, all the paired-end *Blastocystis* reads extracted were mapped to *Blastocystis* genomes using PLAST. If there are more than 300 single

reads mapping to a *Blastocystis* genome (or genomes) with >90% identity over >90% of the read length, the sample was also defined as a positive carrier for *Blastocystis*. To diminish the potential cross contamination during DNA extraction or errors during/after sequencing due to sample bleeding, to qualify as positive the number of *Blastocystis* reads identified by PLAST had to be larger than 0.001% of the total reads of the metagenome sequencing for the sample.

For animal datasets, only the first criterion applied since there are very few reference genomes for animal-infecting *Blastocystis* subtypes and so the numbers of read 'hits' were consequently lower. Co-infections of *Blastocystis* were defined based only on criterion 1; i.e., if a dataset had >300 reads for both STs and also two different sets of LSU/SSU rRNA gene sequences were detected for both STs, this dataset was designated as co-infected with both STs.

The threshold number of 0.015% of *Blastocystis* reads discussed above was determined by considering both the false-assignment rate in sequencing platforms and the percentage of eukaryotic reads in metagenomic samples. The sample bleeding rate for single-indexing Illumina sequencing is ~0.3% and between 0.14% and 0.17% for double-indexing (Kircher et al. 2012). Typically, metagenomic samples contain less than 5% of eukaryotic reads. If we assume all the assigned eukaryotic reads were falsely assigned, that would correspond to 0.3% × 5% = 0.015% (assuming the highest possible bleeding rate of 0.3% for all the samples).

The threshold value of 300 reads came from the following observations. The top 3 largest numbers of *Blastocystis* reads per sample I detected were 2,633,558 reads, 1,647,190 reads, and 941,9492 reads, and the average of 0.015% of these numbers was 261, which, to be conservative, was rounded up to 300. Also based on observation, the least number of *Blastocystis* reads a sample had when it had at least a 200 bp SSU was around 300. For these reasons a sample with at least 300 reads assigned to a *Blastocystis* ST was defined as a potential positive sample.

Table 2.4 Definition of *Blastocystis* positive infection in gut metagenomic datasets.

Criterion	# Total *Blastocystis* reads based on Centrifuge results	Metaxa2 / PLAST results	*Blastocystis* presence or not?	Sample type
1	> max (300, 0.015% × Max # *Blastocystis* reads)	Nuclear LSU/SSU rRNA sequences by Metaxa2 (> 200bp)	Yes	Human Animal
2	> max (300, 0.015% × Max # *Blastocystis* reads)	# of *Blastocystis* reads by PLAST > max (300 single reads, 0.001%Total reads) (> 90% of the length & > 90% identity)	Yes	Human

2.3.3 Prevalence of *Blastocystis* in human gut metagenomes

Using the bioinformatic workflow developed in this study and the threshold defined in Table 2.4, at least one *Blastocystis* sp. ST was detected in 240 out of 484 (52.7%) individuals in the 10 studied projects from 10 countries (Figure 2.5 (a) and Supplementary Table S4). The prevalence was higher in African subjects (167 of 249 samples, 67%) and lower in European and North American ones (37.9% and 17.7%, respectively). There was only one project containing individuals for South America (Peru), with the highest infection rate (30 of 36 samples, 83.3%) among all the continents. The 20 Indonesian individuals with helminth infections were the only Asian population (Rosa et al. 2018). Despite the relatively small size of the dataset, a relatively high prevalence (65%) of *Blastocystis* was detected in these helminth-infected subjects.

To better detect the prevalence of *Blastocystis* in culturally diverse populations, the infection rates of the control groups were separated from the studied groups in projects H5 (Peru versus the USA) and H8 (Tanzania vs. Italy), which resulted in 12 populations based on country and conditions (Figure 2.5 (a)). *Blastocystis* had the highest prevalence in the Tanzania seasonal hunter and gatherer group (23 of the 27 subjects, 85.2%), while it was not detected in the native Americans from tribes in Oklahoma. For the other two USA projects, there was a similar overall prevalence rate: 27.3% in the H5 USA control group and 25.5% in the USA immigrant group. The results also showed a huge difference in the prevalence of the microorganism among European groups, with very low frequency in Italians (1 of 11 samples, 9.1%) and a group of Swedish obesity individuals (2 out of 20, 10.0%), but a relatively high frequency in a group of Swedish university students (22 of 35 samples, 62.9%).

2.3.4 Subtype identification and co-infections in human gut metagenomes

Regarding the distribution of *Blastocystis* subtypes amongst the positive samples (255 samples), the top three most common single STs were ST1 (78 samples, 31% of all the positive samples, found in all populations except Italians), ST3 (71 samples, 28%, in all groups except the H7 Swedish obesity individuals), and ST2 (43 samples, 17%, found in all populations except Italians) (Figure 2.5(b)). There was a very low frequency of ST7 and ST8 and no ST5, ST6 and ST9 were detected in these human samples. ST4 (9 samples, 3%) was only detected from American and European groups, which is consistent with the higher prevalence of ST4 in European populations (Beghini et al. 2017).

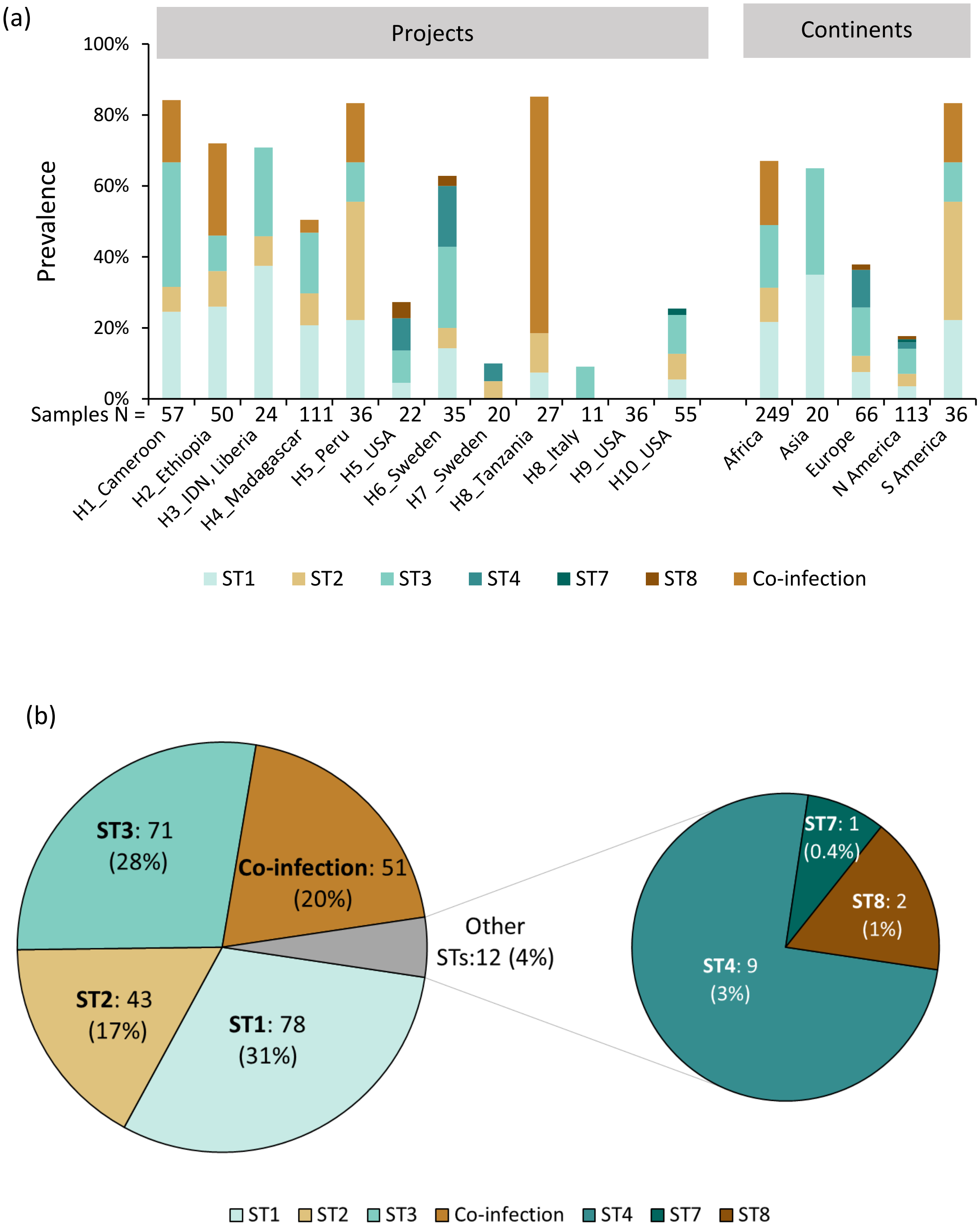
Blastocystis co-infections (carrying more than one ST) were found in 51 individuals (20% of all the positive samples) from African and South American groups (Figure 2.5 (c)), with ST1 + ST3 as the most frequent combination (23 samples, 45.1%), followed by ST2 + ST3 (25.5%) and ST1 + ST2 (19.6 %). Five co-infection samples (9.8%) had three STs: ST1, ST2, and ST3. There was no co-infection found in European, Asian or North American samples.

2.3.5 Relating *Blastocystis* prevalence to metadata in human gut metagenomes

Metadata including gender, age, and Body mass index (BMI) were available for up to 252 samples, and *Blastocystis* colonization was detected in 148 of these individuals, 51.3% of whom (76/148) were women. The numbers of female/male, the mean and standard deviation of age and BMI for each group were calculated (Table 2.5 for adults and Table 2.6 for children with age < 18 yr). The oldest individual detected with *Blastocystis* was a 72-year-old, while the youngest child was only 2 years old. The differences in gender or ages between carriers and non-carriers were not statistically significant. There was no difference in the mean ages between adult *Blastocystis* carriers and non-carriers (Student's *t*-test, p value = 0.609), however, the difference in the mean age of female carriers was significantly smaller than the male carriers (38.1 ± 13.6 years [mean ± standard deviation] versus 44.0 ± 14.4 years, respectively; Student's *t*-test, p value = 0.012). Females with *Blastocystis* ST3 colonization also had a significantly smaller mean age than the male ST3 carriers (36.9 ± 9.7 years versus 46.1 ± 12.2 years, respectively; Student's *t*-test, p value = 0.010). There was no difference in the mean age or mean BMI detected in the groups of children.

For adults with known BMI values, *Blastocystis* carriers tended to have a smaller mean BMI than non-carriers (Student's t-test, p value<0.0001; Table 2.5). For three projects with BMI values available, the prevalence of *Blastocystis* was higher in underweight and normal groups than in obesity groups in all three projects (Figure 2.5 (d)), but no statistical significance was found among different BMI classes. For the two western projects containing mainly obese individuals, overweight and obese individuals are less frequently colonized by *Blastocystis* (Figure 2.5 (a)). For the H7 Swedish obese project that contained 14 obese individuals after bariatric surgery and 7 obesity controls, the prevalence of *Blastocystis* was only 9.5%, lower than the average prevalence detected in this study and much lower than the prevalence in the H6 project with Swedish university students. For project H9 that contains 36 native Americans from western Oklahoma with 22% were overweight and 72% obesity, no *Blastocystis* was detected from this group.

Figure 2.5 Prevalence of *Blastocystis* and subtypes distribution (a) in the different human projects and different continents, (b) in percentage for all samples, (c) in projects with co-infection, and (d) in BMI classes.



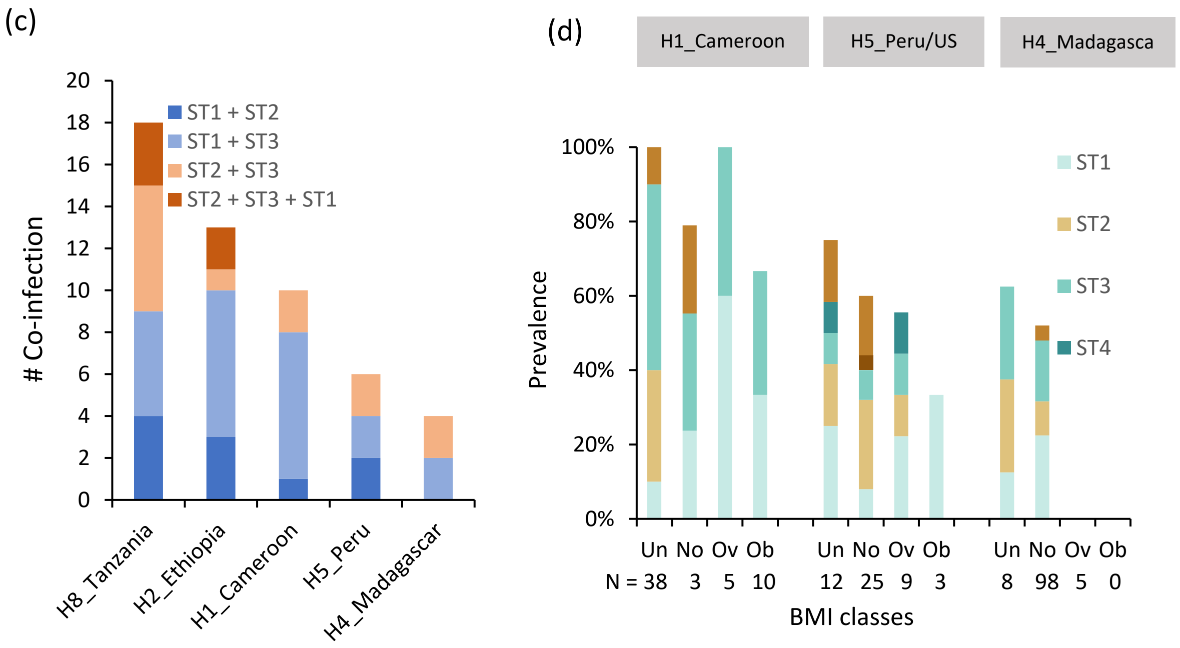


Table 2.5 Descriptive statistics of datasets grouped by *Blastocystis* adult carrier status and subtypes.

Metadata	Non-carriers	Carriers					
	n=104	Total n=148	ST1 n=45	ST2 n=23	ST3 n=46	ST4/ST8 n=2	Co-infection n=32
Gender (F/M)	67/37	76/72	26/19	17/6	18/28	1/1	14/18
Age (mean, SD)	40.4, 14.3	40.9, 14.2	40.6, 15.3	39.7, 13.5	42.5, 12.1	39.0, 15.6	40.3, 16.6
BMI (mean, SD)	27.8, 7.2	21.6, 3.1	22.0, 3.4	20.8, 2.6	21.3, 3.4	24.0, 2.6	21.7, 1.8

Table 2.6 Descriptive statistics of datasets with age younger than 18-year old, grouped by *Blastocystis* carrier status and subtypes.

Metadata	Non carriers	Carriers					
	n=10	Total n=25	ST1 n=8	ST2 n=6	ST3 n=4	ST4 n=1	Co-infection n=6
Gender (F/M/unknown)	2/6/2	15/7/3	4/4	3/3	4/0	1/0	3/0/3
Age (mean, SD)	8.9, 5.9	8.8, 5.2	9.8, 5.8	4.6, 2.8	10.4, 5.8	10, NA	11.3, 4.8
BMI (mean, SD)	18.5, 2.1	17.7, 2.2	17.2, 2.2	19.1, 1.6	NA	15.0, NA	NA

2.3.6 Comparison of results of *Blastocystis* prevalence with other studies

Some of the human gut metagenomic datasets analyzed here have also been used in previous studies to detect *Blastocystis* using different approaches, so I compared the results obtained by using the workflow developed in this study with the previous studies. Lokmer et al. (2019) applied a metagenomic (MG) approach, which mapped metagenome reads to reference genomes and retained only high-confidence alignments to detect *Blastocystis*, to 127 datasets from three projects (H1 Cameroon, H5 Peru/USA, and H8 Tanzania/Italy). A dataset with reads mapped to > 10% of the contigs in the genome of a certain *Blastocystis* ST and with a breadth of coverage > 0.001 was defined as a positive sample. In the H5 Peru/USA and the H8 Tanzania/Italy projects, individuals younger than 18 years of age were excluded and they only detected *Blastocystis* in 32.4% (12 out of 37) and 57.6% (19 out of 33) of datasets, respectively (Figure 2.6 (a)). For the H5 and H8 projects, I found the prevalence of *Blastocystis* was 56.8% (21/37 samples) and 60.6% (20/33 samples), respectively. Lomker and colleagues also compared the MG-based approach to quantitative PCR (qPCR) results in one of the projects (H1) and concluded that qPCR was at least as sensitive as metagenomics for *Blastocystis* diagnosis, although the MG-based method was more likely to detect co-infections that multiple subtypes colonized in the gut. For the total of 57 datasets in H1 that had results for both methods, they found *Blastocystis* was present in 49 samples with 4 co-infections using qPCR and 44 positives with 11 co-infections using the MG-based method. Using my workflow, I found *Blastocystis* in 48 samples with 10 having co-infections. Each ST and type of co-infections detected by Lokmer et al. (2019) were also found by my method.

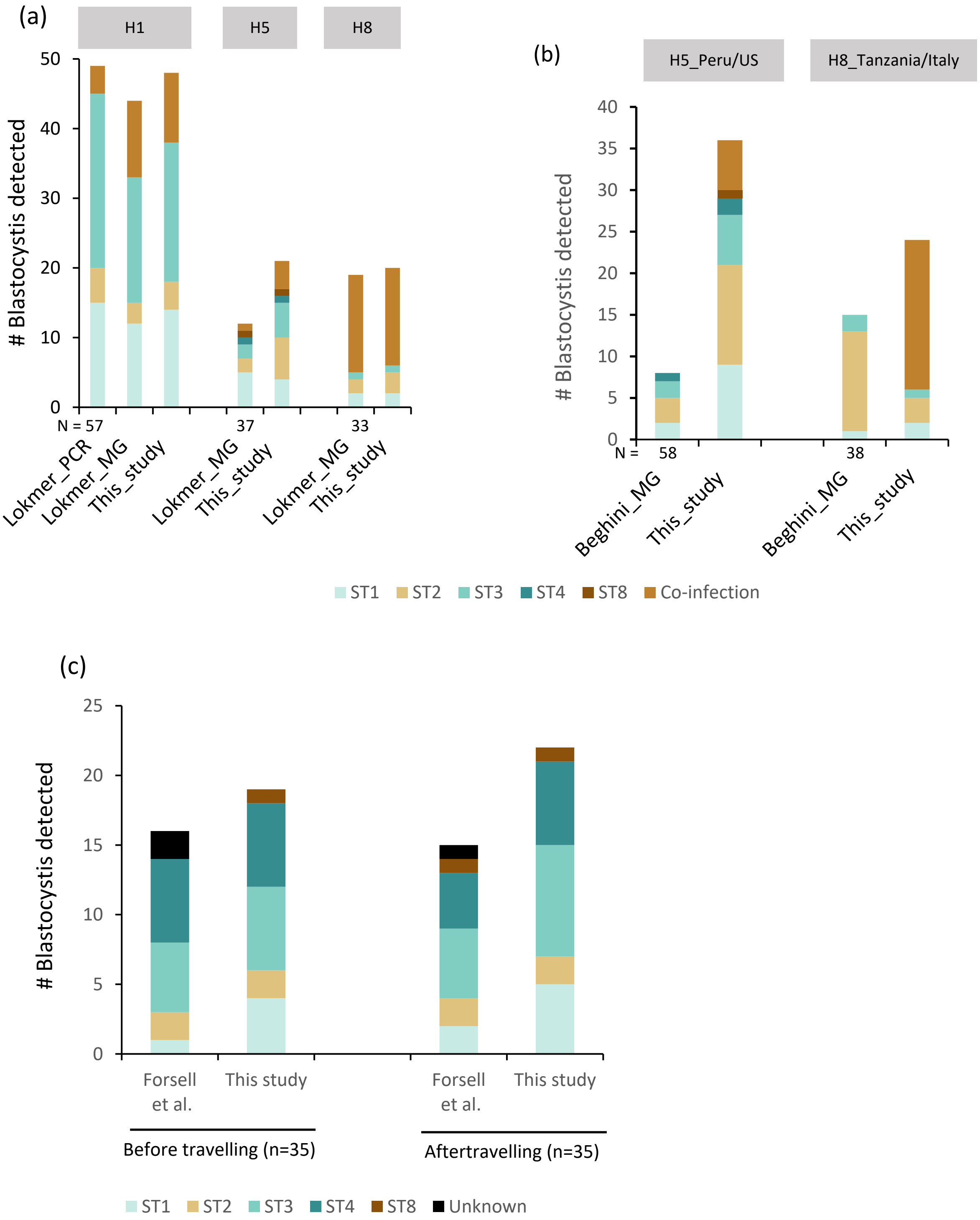
Beghini et al. (2017) used a similar metagenomic approach to detect the presence of *Blastocystis* in human gut metagenomes. To minimize the false positive rate, they removed the potentially contaminated contigs that had bacterial or archaeal alignments from the reference genomes before mapping the metagenome reads to reference genomes. They defined a sample as positive for a *Blastocystis* ST if the breadth of coverage of assembled reads to the *Blastocystis* ST genome was higher than 10%. Using their approach and definition of positive samples, they detected a prevalence of 13.8% (8/58) in the H5 Peru/USA project, with the presence of *Blastocystis* ST1, ST2, ST3, and ST4 but no co-infections (Figure 2.6 (b)). In my analyses, I separated the control groups from the study groups and found more positive samples in each group (data not shown). In the H5 project, I detected *Blastocystis* in 30 Peruvian samples and 6 American samples. Besides the STs they detected, I also found 1 sample with ST8 and 6 samples co-infected with multiple STs. For H8 the Tanzania/Italy project, they detected ST1, ST2, and ST3 with a total prevalence of 55.6% (15 of 27 samples) from Tanzanian subjects and none from Italian subjects, while I found a total prevalence of 85.2% (23/27 with 18 co-infections) in the Tanzania samples and one positive sample in the Italian group.

Forsell and colleagues detected the prevalence of *Blastocystis* in a group of Swedish university students (age 23-34 years) before and after traveling to the Indian peninsula or Central Africa with a median travel duration of 34 days (range 14 to 150) (Forsell et al. 2017). They used software called Metaxa2 to detect the partial sequence of SSU rRNA of *Blastocystis* and counted the number of reads that can be assigned to a ST. A prevalence of 16/35 (45.7%) before travel and 15/35 (42.8%) after travel was found in their study (Figure 2.6(c)). Amongst the positive samples, they found no co-infections and ten individuals with a typable ST before and after travel maintained the initial ST (2 positive datasets before travel and one after travel cannot be assigned to a specific STs). In contrast, I found more positive samples before and after the individuals traveled (19/35, 54.3% and 22/35, 62.9%, respectively) and 17 subjects maintained their initial STs. No co-infections were detected in this group.

2.3.7 The prevalence of *Blastocystis* in animal gut metagenomes

Thirteen animal gut metagenome projects were downloaded from NCBI and analyzed using my *Blastocystis* detection workflow (Table 2.2). These animal samples included several different hosts (baboon for non-human primates, chicken, cattle, and pigs for livestock) from four continents (Africa, Asia, Europe, and North America). An animal dataset was defined as positive with *Blastocystis* infection if it fulfilled the Criterion 1 described in Section 2.3.2 (Table 2.4). From the total number of 476 samples that were analyzed, 298 (62.6%) were carriers for *Blastocystis* (Figure 2.7 and Supplementary Table S5).

Figure 2.6 Comparison of prevalence of *Blastocystis* detected in this study and other studies by (a) Lokmer et al. (2019), (b) Begnini et al. (2017), and (c) Forsell et al. (2017) that also used metagenomic (MG) approach to detect *Blastocystis*.



Sample positivity across different animal groups was not evenly distributed. The prevalence of *Blastocystis* in baboons was high with 24/48 (50.0%) positive. By contrast, only 3/24 (12.5%) chickens and 16/105 (15.2%) cattle were positive. All three positive chicken samples (75%) in project A6 were from China while the other chicken project A5, also from China, had no positives. For cattle, the prevalence of Blastocystis did not vary much in three projects with positive samples (13.3%, 27.6%, and 20% for projects A2, A5, and A6, respectively). Two cattle projects had no positive samples (A3 and A4). However, some of these values should not be interpreted as reflective of the overall prevalence of *Blastocystis* in the projects since, for four of the projects (A3, A5, A8, and A10), not all the datasets in each project were analyzed. Pigs had the highest prevalence, with an average of 85.3% positive rate. However, one pig project, A12 consisting of 22 pigs from Germany showed a very low *Blastocystis* prevalence (2/22, 9.1%).

2.3.8 *Blastocystis* subtype dominance and host specificity in animal gut metagenomes

Among the 298 samples positive for *Blastocystis*, eight STs were detected in the 266 samples carrying only one ST (Figure 2.7). Two subtypes (ST1 and ST3) were detected in baboons, each having roughly half of the positive samples. In cattle, four STs were detected with ST1 as the most common (5/14, 35.7%) followed by ST10 (4/14, 28.6%). ST1 and ST3 were only detected in American cattle (project A5), while ST5 and ST10 were found in cattle from China (projects A2 and A7). ST6 and ST7, the commonly identified STs in the bird (Clark et al. 2013), were only found in chickens from project A7.

Pigs had the most subtypes detected in this study. Among five different STs in pigs, ST5 was detected with the highest frequency (70.8%) in pigs from all the projects containing pigs (except A7). ST15 was the second most common at 20.8% and detected in 4 projects. The rest STs found in pig samples were ST1(2.71%), ST3 (3.88%), and ST13 (0.39%).

Co-infections were found in five projects containing cattle and pigs (21.4% and 11.63%, respectively)(Figure 2.8). In cattle, there are two types of combinations (two samples for ST1+ST3 and one sample for ST1+ST5). The most common combination of co-infections in pigs was ST5+ST15 that was detected in 25 samples. The rest of the combinations in pigs were ST1+ST15, ST3+ST5, and ST1+ST3+ST5.

Figure 2.7 Prevalence of *Blastocystis* and subtypes in the different animal projects and different animal categories. (*) Only a subset of datasets from the project was analyzed.

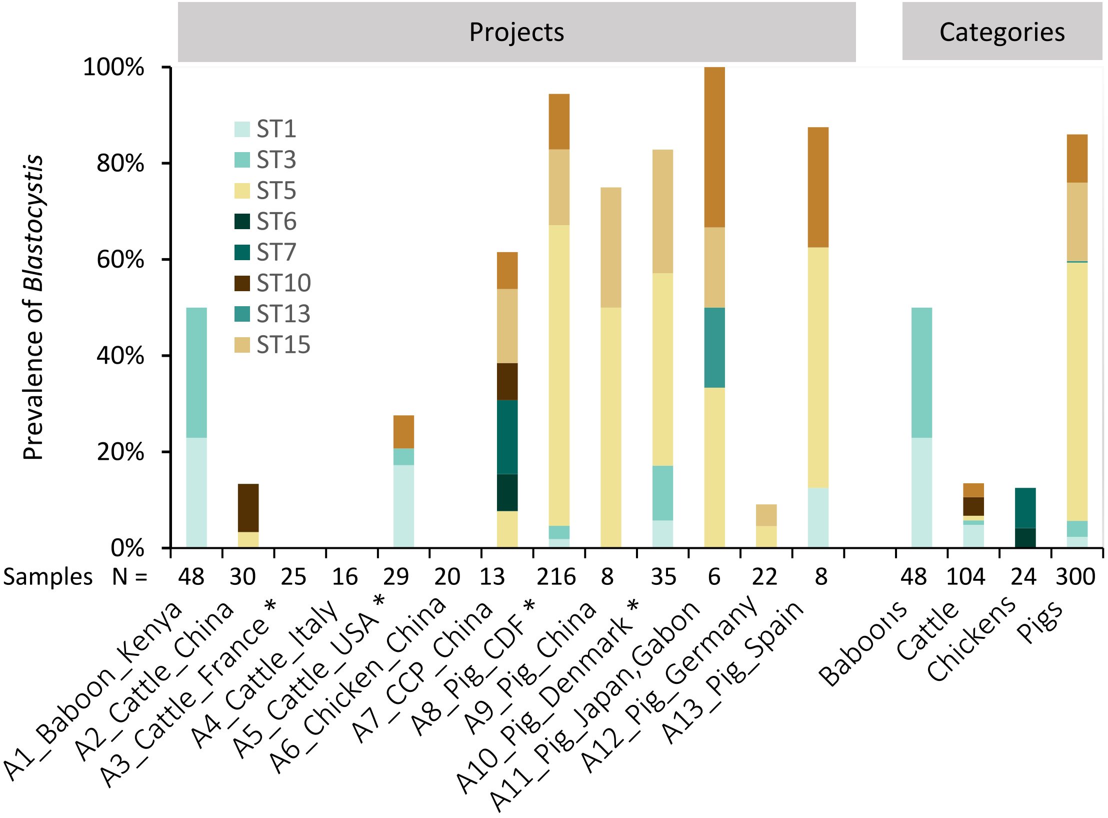
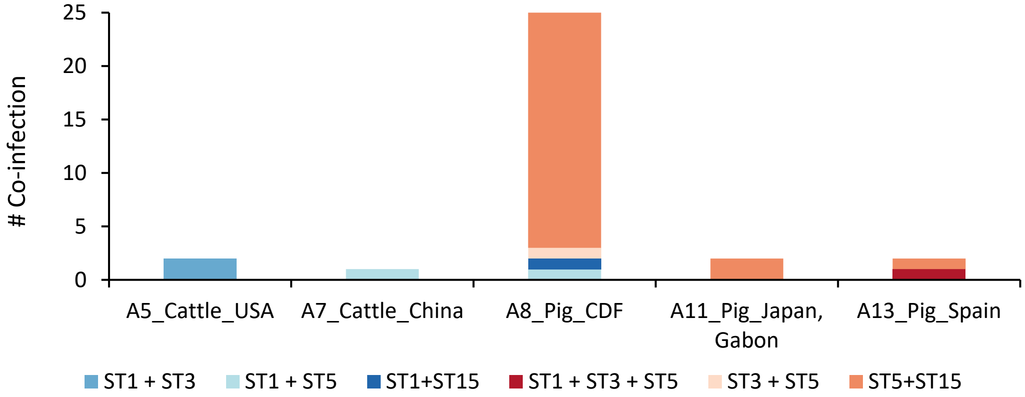


Figure 2.8 Combinations of *Blastocystis* co-infections detected in animal projects.



2.3.9 Gut microbiome composition associated with the presence of *Blastocystis*

To investigate how *Blastocystis* can affect prokaryote communities in the gut, compositional analyses were conducted on a subset of the human datasets including 119 *Blastocystis* carriers and 81 *Blastocystis* non-carriers (Table 2.3). The effects of westernization on the gut microbiome were also examined in *Blastocystis* positive and negative samples (see section 2.2.5 for the definition of westernized/non-westernized cohorts).

The abundance of archaea was strongly associated with the presence of *Blastocystis* (Welch's *t*-test, FDR=6.47e-3; Figure 2.9(a)), especially the species *Methanobrevibacter smithii* and an unclassified *Methanobrevibacter* (Welch's *t*-test, FDR=0.041 and FDR=3.82e-4, respectively; Figure 2.10). However, when subcategorized with westernization, both *Methanobrevibacter* species had a significantly high relative abundance only in non-westernized positive samples (Kruskal-Wallis H-tests, FDR = 4.39e-11 and FDR = 1.20e-11, respectively; Figure 2.9(b-e)).

Ten bacterial species, including *Faecalibacterium prausnitzii, Prevotella copri,* and *Treponema succinifaciens,* were found to be strongly associated with the presence of *Blastocystis*, while seven bacterial clades had higher abundance in *Blastocystis* negative samples (Welch's *t*-test with , FDR <0.05, effect size (difference in mean proportion) > 0.2; Figure 2.10). Bacteria in the *Firmicutes* were found abundant both in positive and negative samples, one member of the *Bacteroidetes* and one member of the *Spirochaetes* were significantly associated with *Blastocystis* colonization, while bacteria in the *Bacteroidetes* group were strongly associated with the absence of *Blastocystis* (Supplementary Table S6).

When expanding the association analysis between microbial composition and the presence of *Blastocystis* to the categories of westernized *versus* non-westernized based on countries of origin for individual samples, a principal components analysis (PCA) of microbiota community composition revealed that westernization had a more profound effect on bacteria and archaea species-level abundance than *Blastocystis* carriage; westernized samples tend to cluster in a small region while non-westernized samples were more spread out (Figure 2.11). A few exceptional species (e.g., *Prevotella copri*, and *Faecalibacterium prausnitzii*) had relatively high abundance in all categories. Hierarchical clustering performed on the microbiota community species composition differences showed that samples from western countries clustered together and as did the ones from non-westernized countries (Figure 2.12). LEfSe analysis showed that westernized *Blastocystis* positive samples versus non-westernized positive samples were enriched with different bacterial taxa (with effect size > 3.5; Supplementary Figure S1). Pair-wise comparison of species composition changes in *Blastocystis* positive/negative samples revealed that westernization has a bigger impact on the community variation than *Blastocystis* presence or absence (Supplementary Figure S2). Bacterial species in the *Bacteroidales* order had a significant association with westernized individuals, and some were more abundant in westernized *Blastocystis* positive samples (e.g., *Bacteroides caccae*, *Bacteroides ovatus* and *Alistipes putredinis*. Supplementary Table S7). For the *Firmicutes* phylum, some specific species strongly associated with westernized *Blastocystis*-negative samples (e.g. *Ruminococcus torques* and *Ruminococcus* sp 5_1_39BFAA).

Figure 2.9 The presence of *Blastocystis* and certain STs is associated with high abundance in (a) Archaea domain and two archaeal species in *Methanobrevibacter* genus, (b-c) Methanobrevibacter smithii and (c-d) unclassified *Methanobrevibacter*, using the LEfSe biomarker discovery tool and STAMP. In (a), yellow bar represents *Blastocystis* positive samples and blue bar represents *Blastocystis* negative samples.

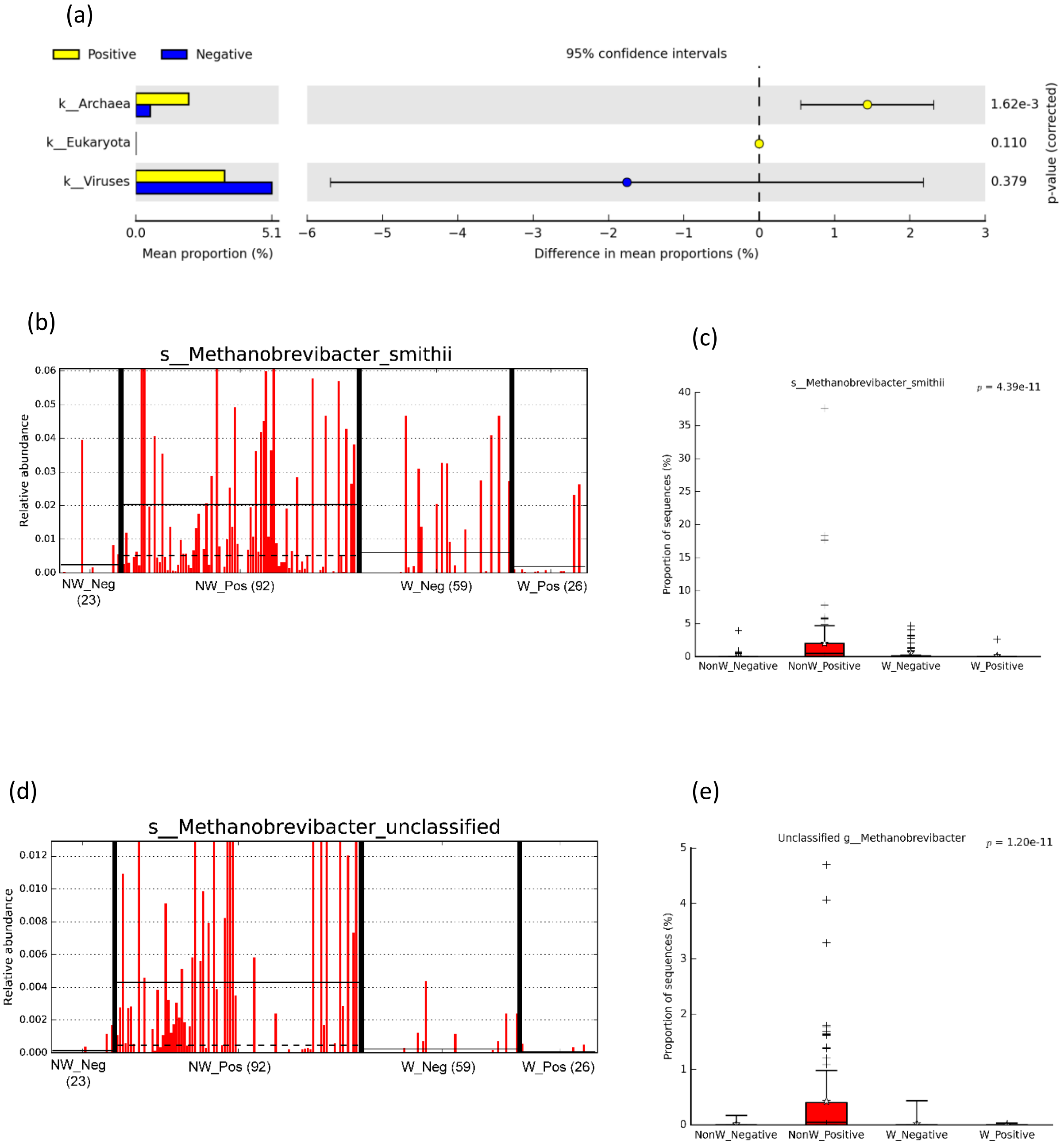


Figure 2.10 Enrichment of microbial species with *Blastocystis* presence (yellow bar) or absence (blue bar). Between-group differences were evaluated with two-tailed Welch's *t*-test with Storey FDR corrections (FDR<0.05) and only difference in mean proportion > 0.2% were shown.

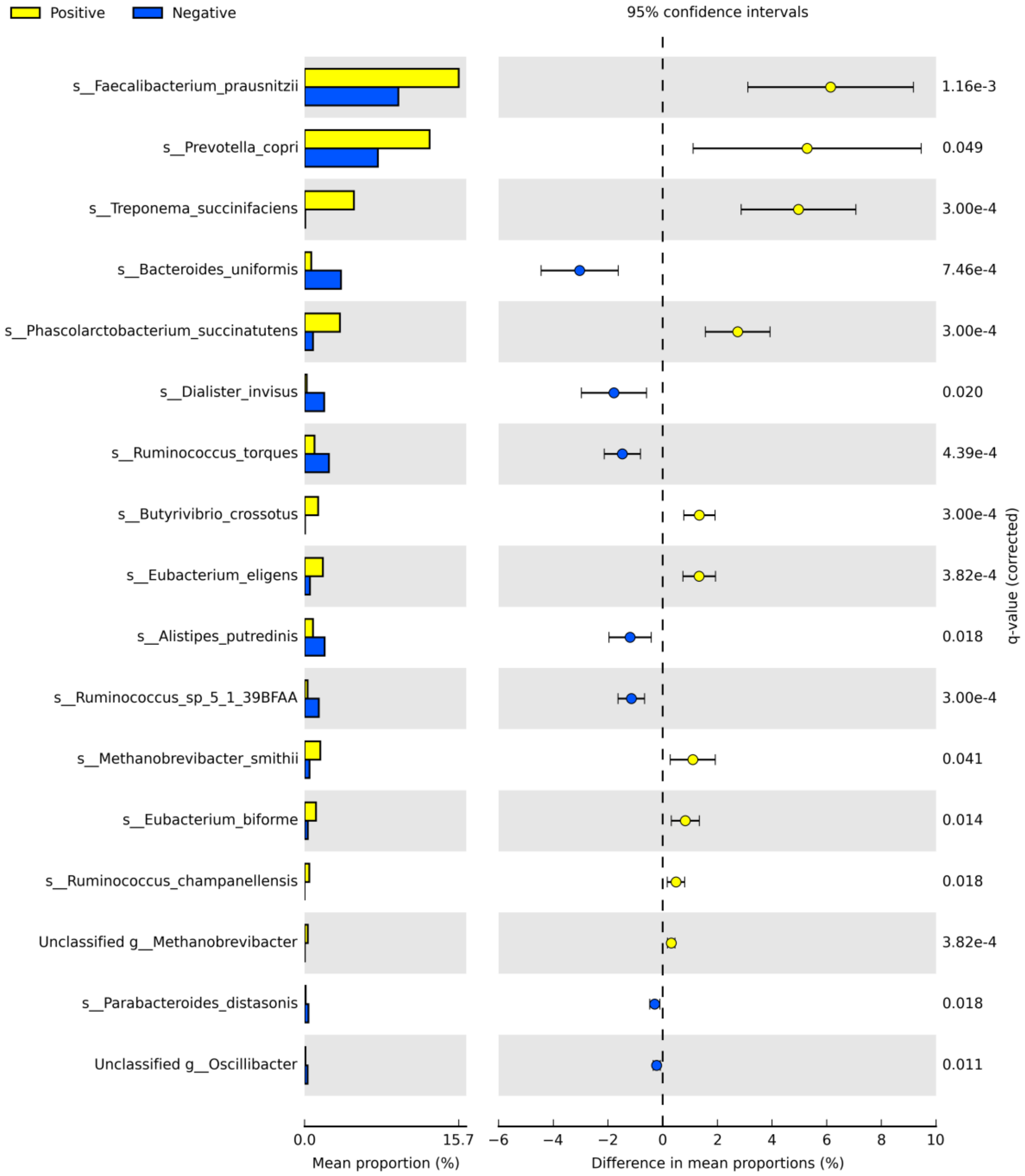


Figure 2.11 Principal Components Analysis of gut micribal community compostion for samples with or without *Blastocystis* (a) between positive and negative samples and (b) positive and negative samples with non-westernized (NonW) or westernized (W) cohorts.

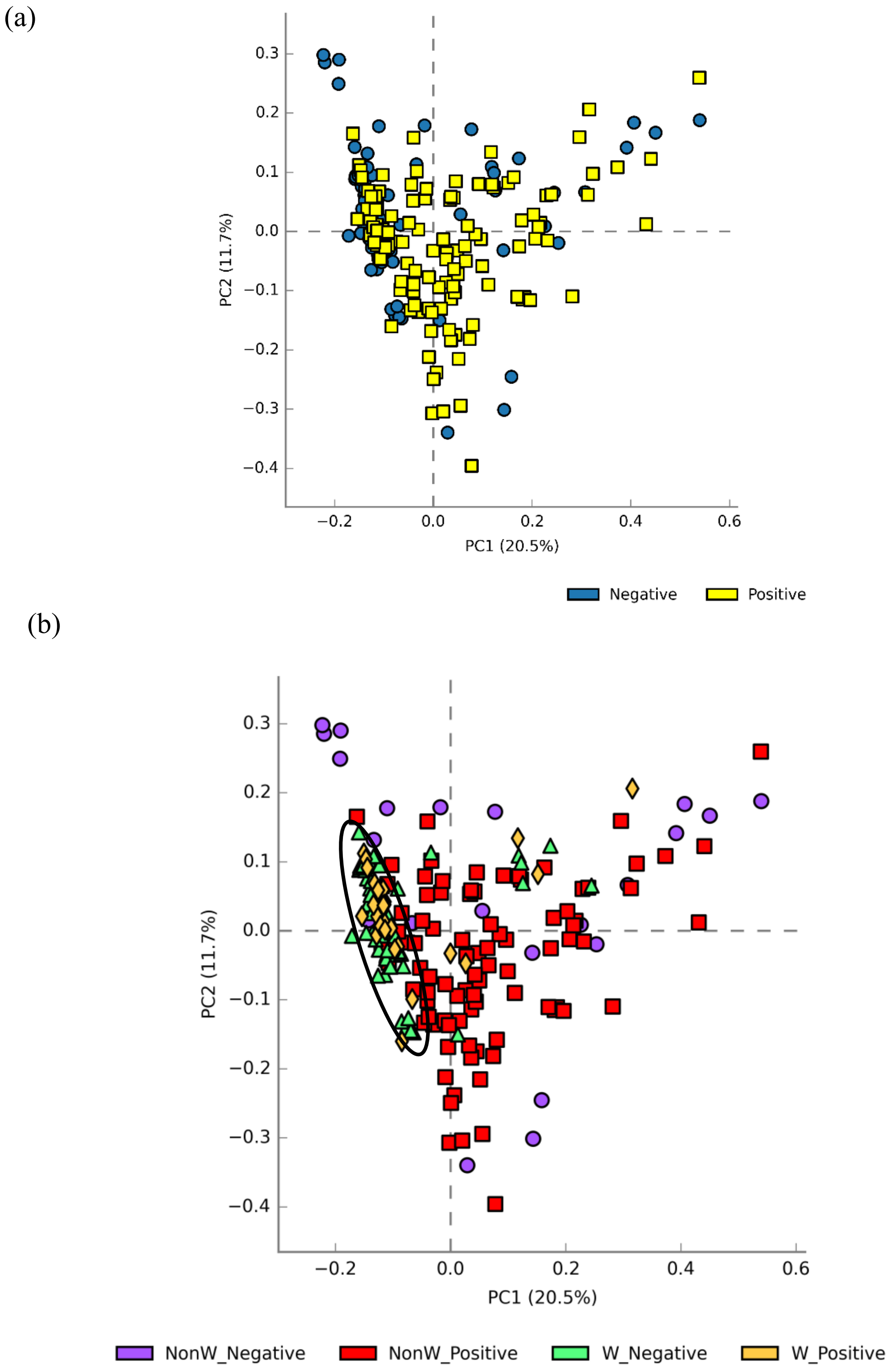
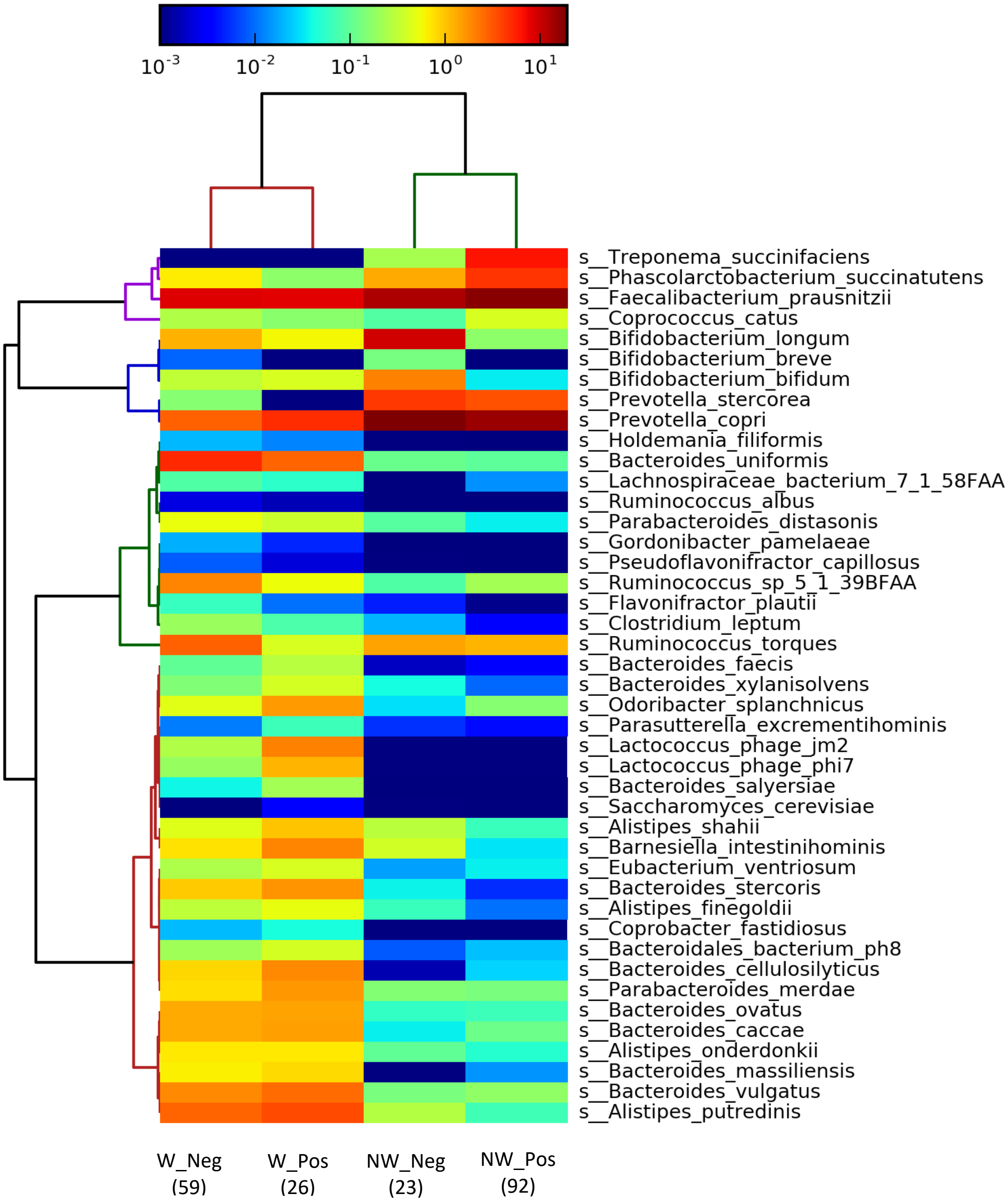


Figure 2.12 Heatmap of enrichment or depletion of microbial species with *Blastocystis* presence or absence in non-westernized (NW) or westernized (W) individuals. The rows and columns were clustered using complete linkage clustering of similarities in similarity microbial species abundance using the correlation distance function and the Bray-Curtis distance metric, respectively. Neg: *Blastocystis* absent, Pos: *Blastocystis* present. The number of samples in each group was labelled in the brackets at the bottom of each column. Statistical test used: Kruskal-Wallis test with Benjamini–Hochberg FDR corrections (FDR < 0.05).



2.3.10 *Blastocystis* subtypes correlate differentially with intestinal prokaryotic microbiota

Potential interactions among *Blastocystis* subtypes and gut microbiota constituents were assessed in the subset of samples. Examined *Blastocystis* subtypes included ST1, ST2, ST3, ST4, and co-infection which is the combination of detected co-infection types. *Blastocystis* subtypes were also differentiated into westernized or non-westernized cohorts. All co-infection samples were from non-westernized countries, ST4 (n=9) was only detected in westernized samples, and the remaining STs and negative samples can be further divided into western or non-western groups. ST groups with samples <5 were excluded from the analyses.

A total of 22 bacterial species (Figure 2.13; Supplementary Table S8) and two archaeal species from the genus *Methanobrevibacter* were found to have distinct relative abundance distributions among *Blastocystis* subtypes. Gut microbes with a significant positive association with *Blastocystis* infection relative to the 'negative' samples tended to have high abundance in ST1-, ST2-, ST3-, and co-infection samples but not in ST4 (Figure 2.14 (a) - (g)). A number of other species appeared to be instead positively associated with ST4 infections relative to *Blastocystis* negative samples and the other subtypes (Figure 2.14 (h)-(k)), with only one exception, *Eubacterium eligens*, that had relatively high abundance in ST3 and ST4 (Figure 2.14 (l)). One possible explanation of this phenomenon was the effect of westernization in different samples. A total of 53 taxa were found to have significantly high abundance among some of the *Blastocystis* subtypes in westernized or non-westernized cohorts (Figure 2.15). Hierarchical clustering also revealed that non-westernized samples tend to have more similar gut microbiome composition to non-westernized infected samples (for ST1, ST2, ST3, and mixed infections), while westernized negative samples tend to cluster with westernized subtypes (Figure 2.15). Different subtypes in the westernized cohort had dramatically varied gut microbiota community 'signals' compared to non-westernized STs, although this may, in part, reflect variation caused by the small sample size in westernized ST1, ST2, and ST8 cohorts (Supplementary Figure S3). Some bacterial species that were significantly associated with *Blastocystis*-positive samples had high relative abundance in most *Blastocystis* STs, with less effect by westernization (e.g., *Faecalibacterium prausnitzii*; Figure 2.16 (a)), while some had high abundance only in non-westernized positive samples (e.g., *Phascolarctobacterium succinatutens* and *Prevotella copri*; Figure 2.16 (b) – (d)). Surprisingly, some bacteria which were previously found to be significantly correlated with *Blastocystis*-negative samples (Supplementary Table S6) had a relatively high abundance in some westernized individuals infected with *Blastocystis* subtypes, mostly ST3 and ST4, (e.g., *Alistipes putredinis* and *Bacteroides uniformis*; Figure 2.16 (e) – (g)), or only ST4 (*Akkermansia muciniphila*; Figure 2.17).

Figure 2.13 Heatmap of enrichment or depletion of microbial species associated with *Blastocystis* STs. The rows and columns were clustered using complete linkage clustering of similarities in similarity microbial species abundance using the correlation distance function and the Bray-Curtis distance metric, respectively. Neg: *Blastocystis* absent. The number of samples in each group was labelled in the brackets at the bottom of each column. Statistical test used: Kruskal-Wallis test with Benjamini–Hochberg FDR corrections (FDR < 0.05).

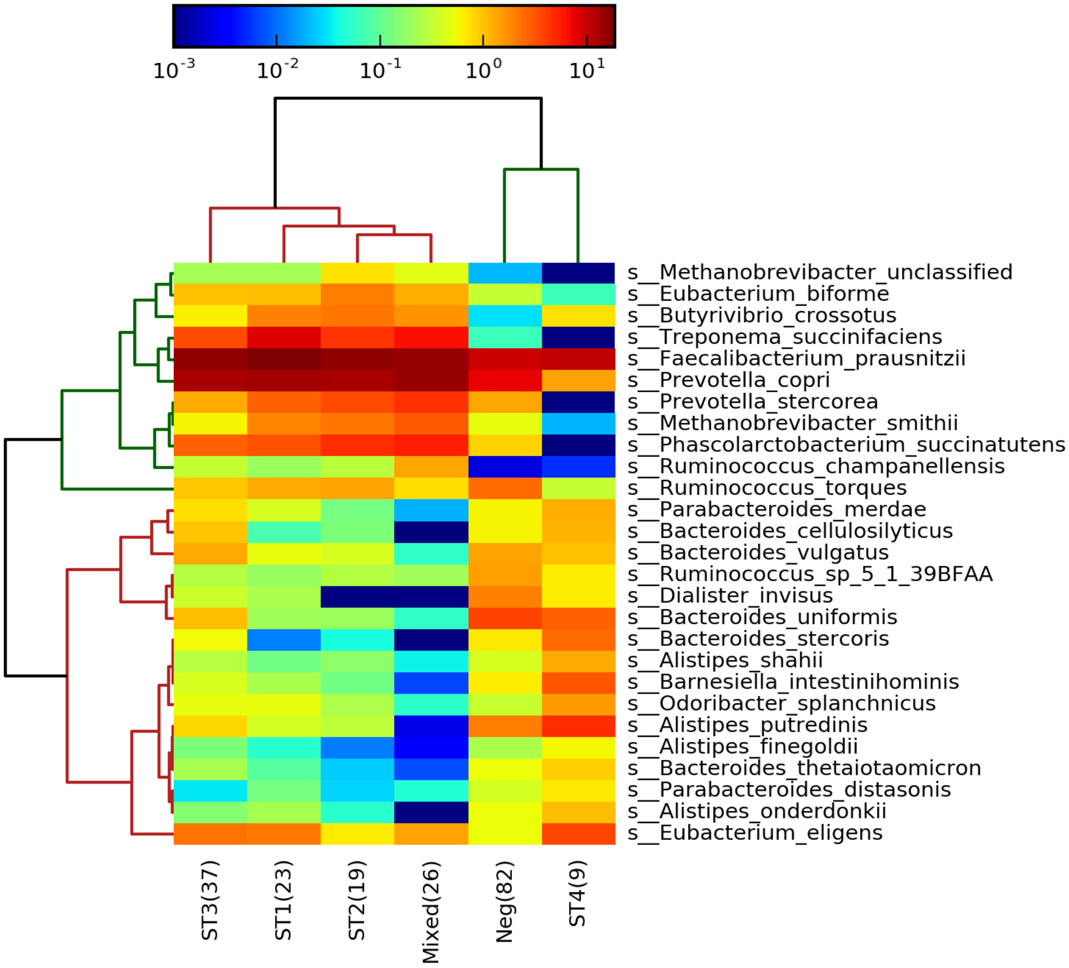
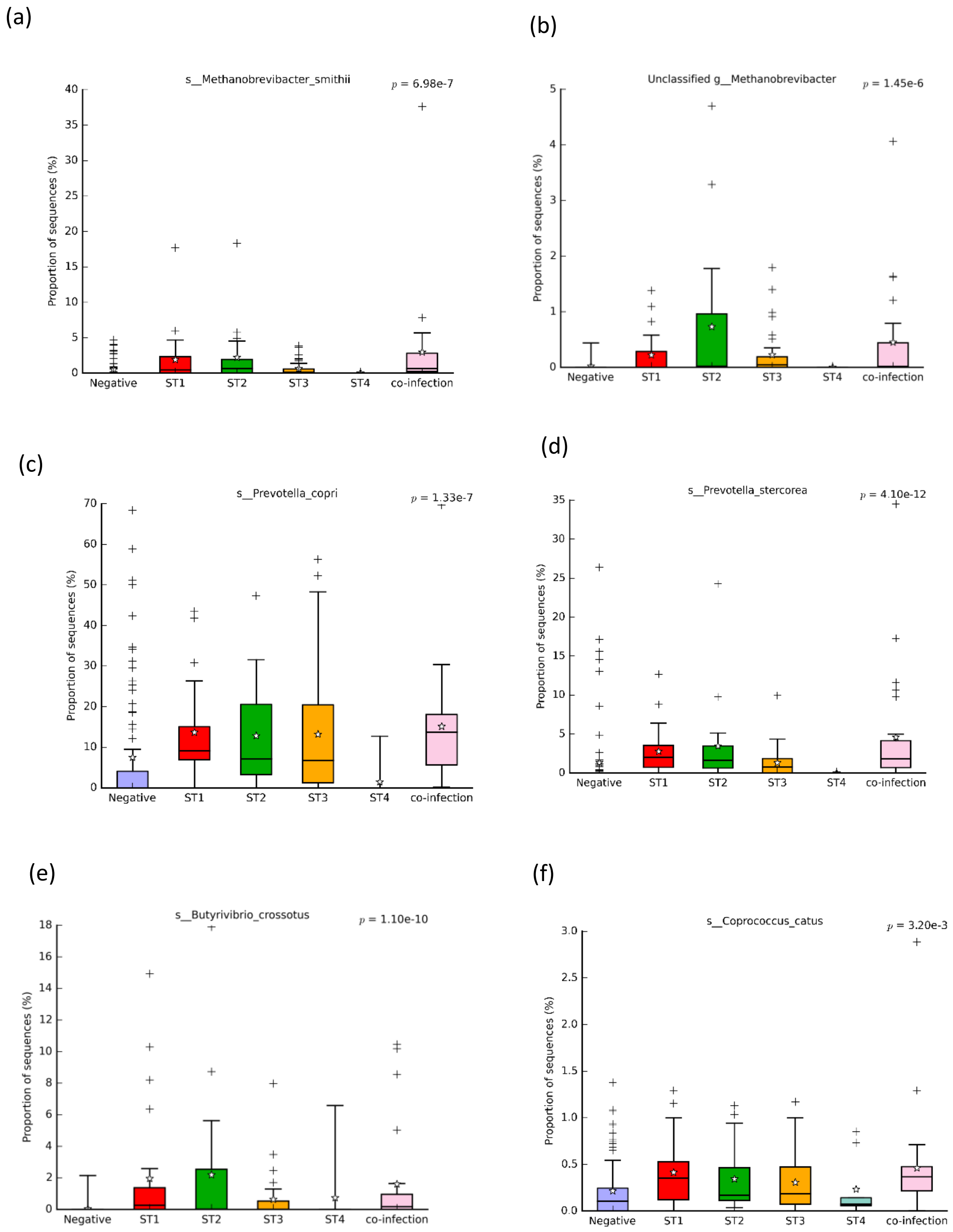


Figure 2.14 Relative abundance of prokaryotic species in the gut microbiome varied significantly in *Blastocystis* STs. Some species are strongly associated with all or most of *Blastocystis* ST1, ST2, ST3, and co-infections (a-g), while others tend to correlated to ST4 (h-l). The number of samples in each group was labelled in the brackets. Statistical test used: Kruskal-Wallis test with Benjamini–Hochberg FDR corrections (FDR < 0.05).



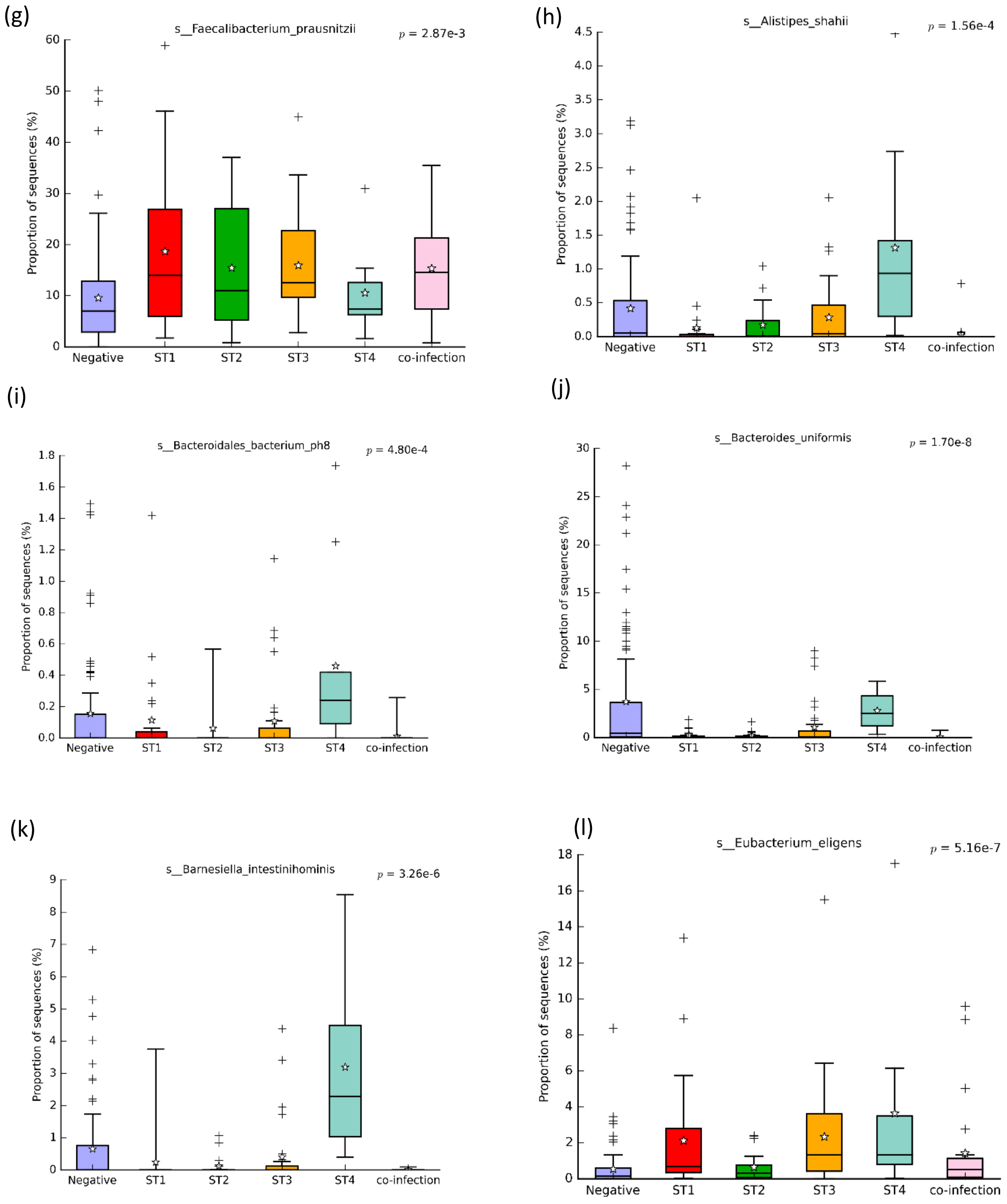

(g)
s__Faecalibacterium_prausnitzii
p = 2.87e-3
Proportion of sequences (%)
0
10
20
30
40
50
60
Negative
ST1
ST2
ST3
ST4
co-infection
(h)
s__Alistipes_shahii
p = 1.56e-4
Proportion of sequences (%)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Negative
ST1
ST2
ST3
ST4
co-infection
(i)
s__Bacteroidales_bacterium_ph8
p = 4.80e-4
Proportion of sequences (%)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Negative
ST1
ST2
ST3
ST4
co-infection
(j)
s__Bacteroides_uniformis
p = 1.70e-8
Proportion of sequences (%)
0
5
10
15
20
25
30
Negative
ST1
ST2
ST3
ST4
co-infection
(k)
s__Barnesiella_intestinihominis
p = 3.26e-6
Proportion of sequences (%)
0
1
2
3
4
5
6
7
8
9
Negative
ST1
ST2
ST3
ST4
co-infection
(l)
s__Eubacterium_eligens
p = 5.16e-7
Proportion of sequences (%)
0
2
4
6
8
10
12
14
16
18
Negative
ST1
ST2
ST3
ST4
co-infection

Figure 2.15 Heatmap of enrichment or depletion of microbial species for bacterial or archaeal species among groups of *Blastocystis* STs and *Blastocystis*-negative samples in non-westernized (NW) or westernized (W) individuals. The rows and columns were clustered using complete linkage clustering of similarities in similarity microbial species abundance using the correlation distance function and the Bray-Curtis distance metric, respectively. Neg: *Blastocystis* absent. The number of samples in each group was labelled in the brackets at the bottom of each column. Groups with less than 5 samples were excluded in this analysis. Statistical test used: Kruskal-Wallis test with Benjamini–Hochberg FDR corrections (FDR < 0.05).

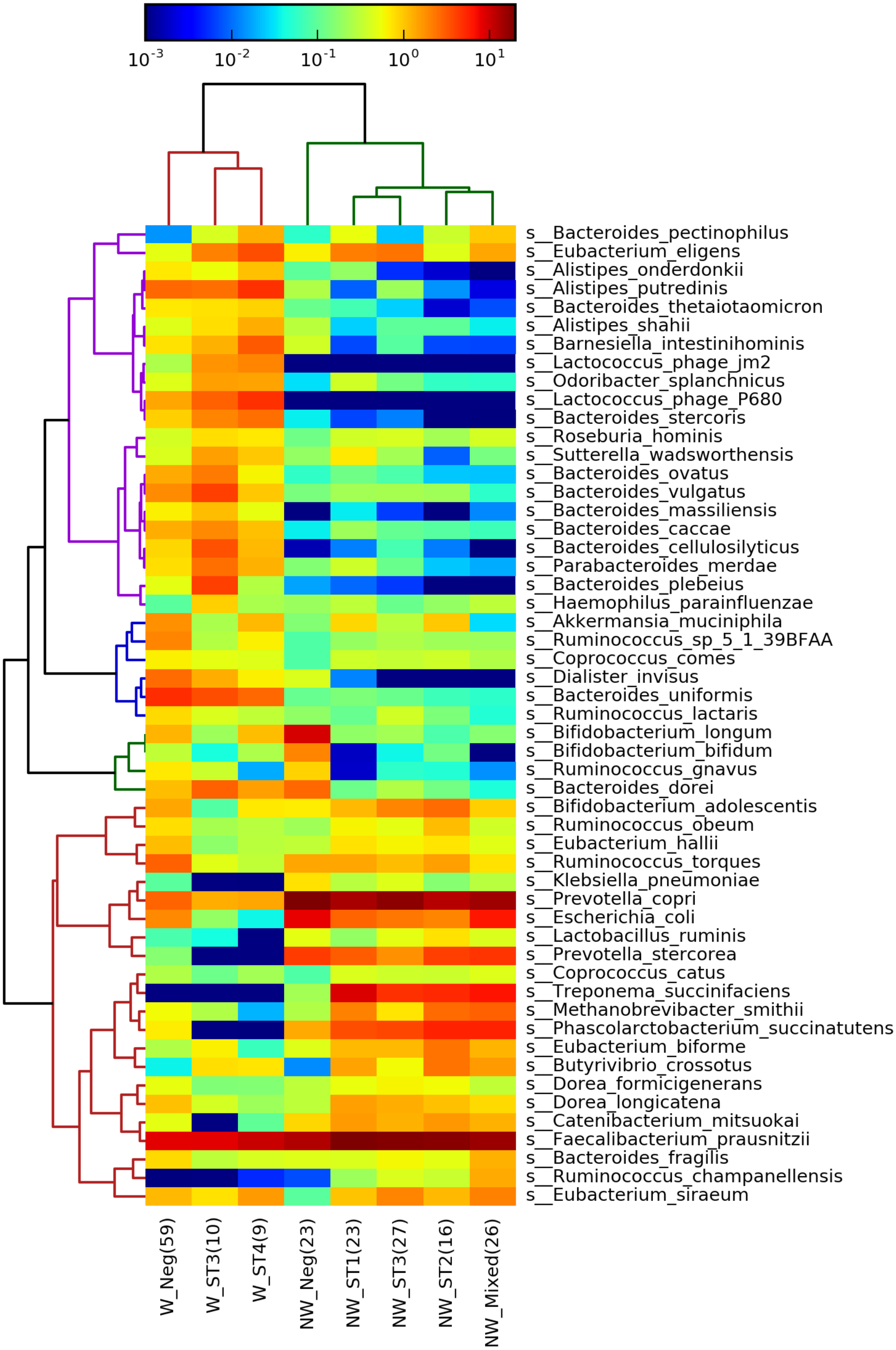
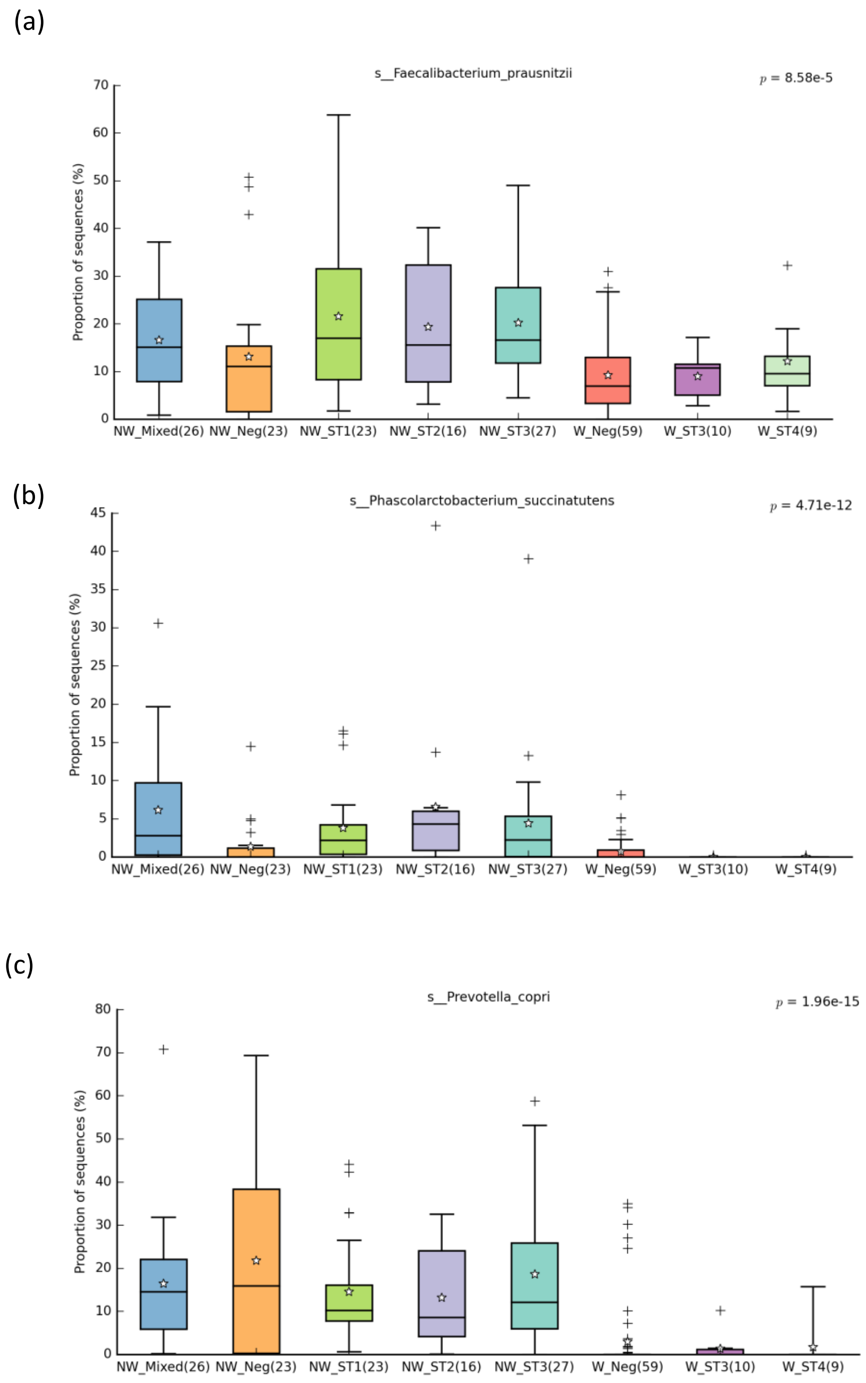

10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
s__Bacteroides_pectinophilus
s__Eubacterium_eligens
s__Alistipes_onderdonkii
s__Alistipes_putredinis
s__Bacteroides_thetaiotaomicron
s__Alistipes_shahii
s__Barnesiella_intestinihominis
s__Lactococcus_phage_jm2
s__Odoribacter_splanchnicus
s__Lactococcus_phage_P680
s__Bacteroides_stercoris
s__Roseburia_hominis
s__Sutterella_wadsworthensis
s__Bacteroides_ovatus
s__Bacteroides_vulgatus
s__Bacteroides_massiliensis
s__Bacteroides_caccae
s__Bacteroides_cellulosilyticus
s__Parabacteroides_merdae
s__Bacteroides_plebeius
s__Haemophilus_parainfluenzae
s__Akkermansia_muciniphila
s__Ruminococcus_sp_5_1_39BFAA
s__Coprococcus_comes
s__Dialister_invisus
s__Bacteroides_uniformis
s__Ruminococcus_lactaris
s__Bifidobacterium_longum
s__Bifidobacterium_bifidum
s__Ruminococcus_gnavus
s__Bacteroides_dorei
s__Bifidobacterium_adolescentis
s__Ruminococcus_obeum
s__Eubacterium_hallii
s__Ruminococcus_torques
s__Klebsiella_pneumoniae
s__Prevotella_copri
s__Escherichia_coli
s__Lactobacillus_ruminis
s__Prevotella_stercorea
s__Coprococcus_catus
s__Treponema_succinifaciens
s__Methanobrevibacter_smithii
s__Phascolarctobacterium_succinatutens
s__Eubacterium_biforme
s__Butyrivibrio_crossotus
s__Dorea_formicigenerans
s__Dorea_longicatena
s__Catenibacterium_mitsuokai
s__Faecalibacterium_prausnitzii
s__Bacteroides_fragilis
s__Ruminococcus_champanellensis
s__Eubacterium_siraeum
W_Neg(59)
W_ST3(10)
W_ST4(9)
NW_Neg(23)
NW_ST1(23)
NW_ST3(27)
NW_ST2(16)
NW_Mixed(26)

Figure 2.16 Bar plot for comparison of proportion of sequences in bacterial species that had significant association with *Blastocystis* STs in non-westernized or westernized individuals. The number of samples in each group was labelled in the brackets at the bottom of each column. Statistical test used: Kruskal-Wallis test with Benjamini–Hochberg FDR corrections. NW: non-westernized, W: westernized, Neg: *Blastocystis* absent.



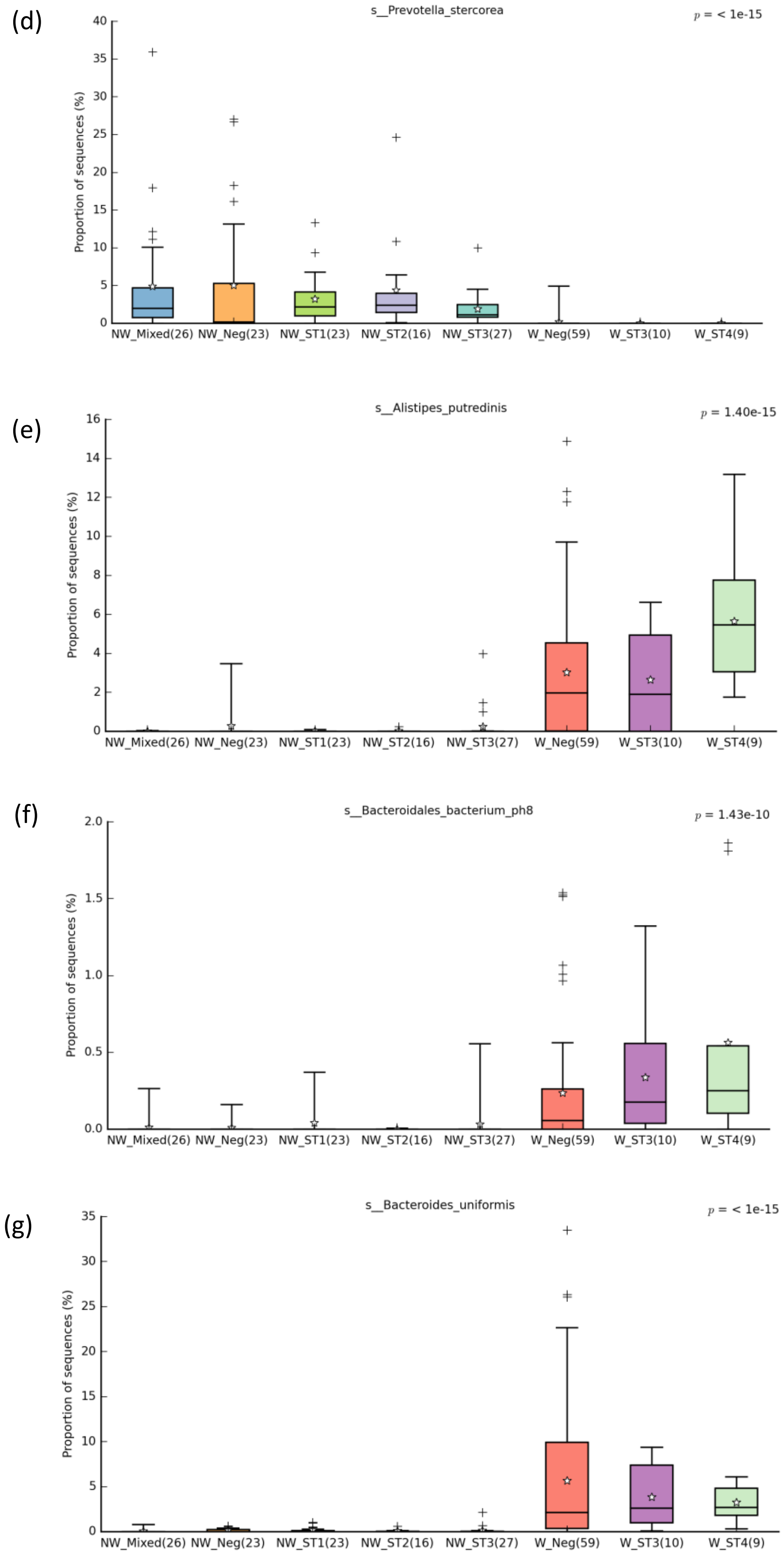
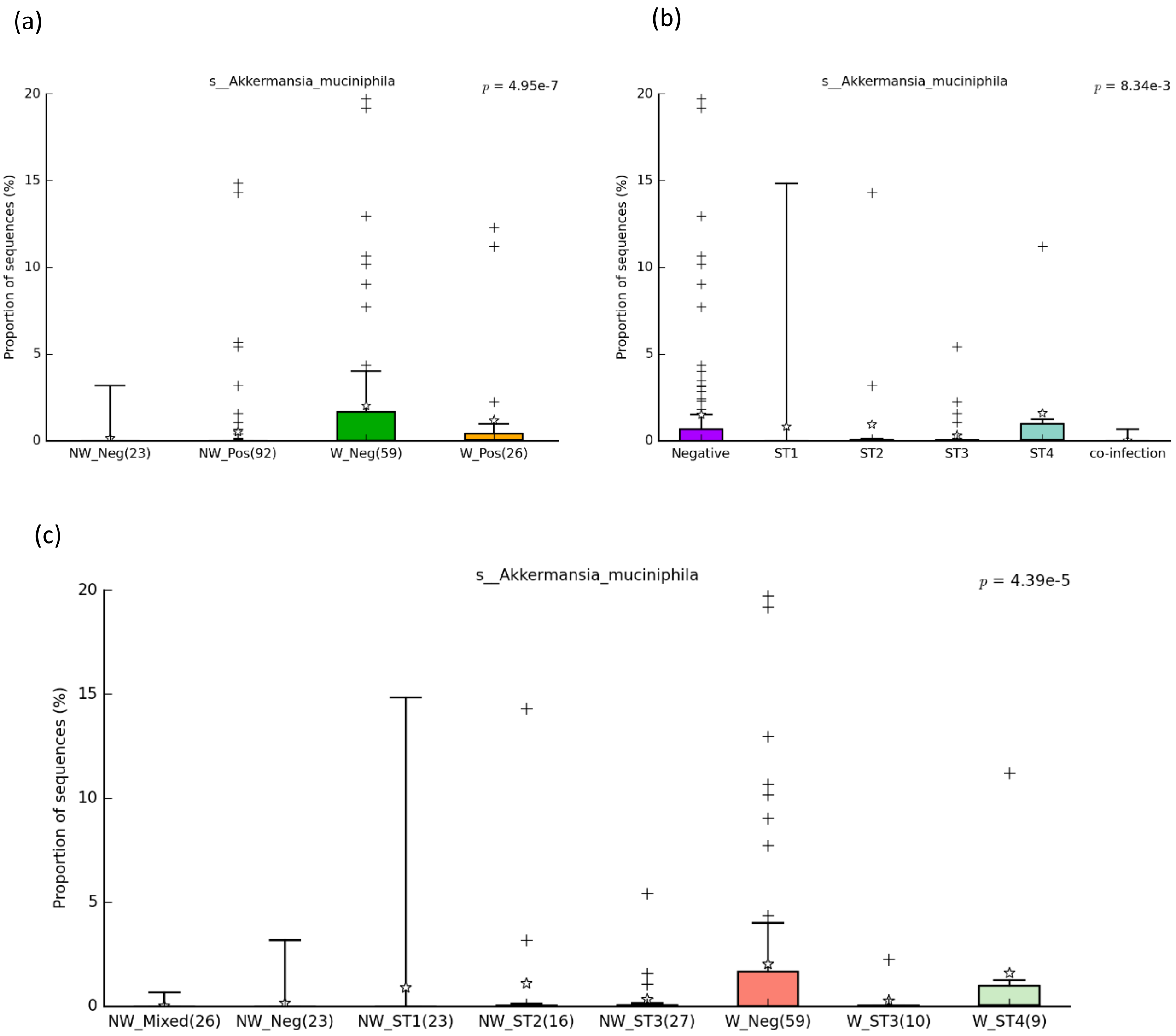

(d)
s__Prevotella_stercorea
p = < 1e-15
Proportion of sequences (%)
0
5
10
15
20
25
30
35
40
NW_Mixed(26)
NW_Neg(23)
NW_ST1(23)
NW_ST2(16)
NW_ST3(27)
W_Neg(59)
W_ST3(10)
W_ST4(9)
(e)
s__Alistipes_putredinis
p = 1.40e-15
Proportion of sequences (%)
0
2
4
6
8
10
12
14
16
NW_Mixed(26)
NW_Neg(23)
NW_ST1(23)
NW_ST2(16)
NW_ST3(27)
W_Neg(59)
W_ST3(10)
W_ST4(9)
(f)
s__Bacteroidales_bacterium_ph8
p = 1.43e-10
Proportion of sequences (%)
0.0
0.5
1.0
1.5
2.0
NW_Mixed(26)
NW_Neg(23)
NW_ST1(23)
NW_ST2(16)
NW_ST3(27)
W_Neg(59)
W_ST3(10)
W_ST4(9)
(g)
s__Bacteroides_uniformis
p = < 1e-15
Proportion of sequences (%)
0
5
10
15
20
25
30
35
NW_Mixed(26)
NW_Neg(23)
NW_ST1(23)
NW_ST2(16)
NW_ST3(27)
W_Neg(59)
W_ST3(10)
W_ST4(9)

Figure 2.17 Bar plot for significant association between *Akkermansia muciniphila* and *Blastocystis* in (a) westernized non-carriers, (b) and (c) ST4 infections. Statistical test used: Kruskal-Wallis test with Benjamini–Hochberg FDR corrections. NW: non-westernized, W: westernized, Neg: *Blastocystis* negative, Pos: *Blastocystis* positive.



2.3.11 The presence of *Blastocystis* is highly correlated with certain metabolic pathways in the gut microbiome

The relationship between the relative abundance of metabolic pathways and *Blastocystis* infection was assessed for the subsets of samples listed in Table 2.3. A principal components analysis of pathway abundance variation revealed that pathways in *Blastocystis*-positive samples tended to cluster separately with the negative samples (Figure 2.18 (a)). Moreover, when considering the western/non-western origin of the sample, there was a much clearer clustering of non-westernized *Blastocystis*-positive samples versus westernized positive samples (red dots *versus* orange dots: Figure 2.18 (b)), but no clear separation between non-westernized positive vs. negative samples (red vs. purple) or westernized positive vs. negative samples (orange vs. green), which again indicates that westernization has a more profound impact on cellular functions of the gut microbiome than colonization of *Blastocystis*. However, the impact of westernized vs. non-westernized origins appears to be greater for positive *Blastocystis* samples.

The relative abundance of genes for 12 cellular pathways was significantly associated with the presence of *Blastocystis* (Figure 2.19 (a)). Three of these were involved in tRNA charging, processing, or synthesis, three for phospholipid biosynthesis, two for fatty acid biosynthesis, and two involved in gluconeogenesis or glycogen biosynthesis. Twenty-six pathways were correlated with the absence of *Blastocystis*, most of them involved in carbohydrate metabolism (six for glycolysis, two for sugar degradation, and two for pyruvate fermentation) and amino acid biosynthesis. Pathway coverage calculated by HUMAnN2 indicated the presence of eight pathways significantly associated with the presence of *Blastocystis* (Figure 2.19 (b)).

Hierarchical clustering on pathway abundance performed on *Blastocystis* positive and negative samples, taking into account westernization confirmed that the greatest difference occurs between westernized and non-westernized samples. However, colonization of *Blastocystis* does affect the abundance of many pathways between carriers and non-carriers (Figure 2.20, Supplementary Figure S4). For example, two pathways related to amino acid synthesis and one for propanediol degradation appear to be specifically depleted in both positive sets of samples relative to negative samples (Figure 2.21 (a)-(c); the opposite pattern is observed for two pathways (lysine biosynthesis and chitin derivative degradation; Figure 2.21 (d)-(e)). On the subtype level, hundreds of pathways had significant associations with different STs, but the differences in the proportions of sequences among different ST groups were very small (Figure 2.22), ranging from 0.0001% to 0.2%. To gain a more robust picture of the real differences occurring between STs, analyses on many more metagenomic samples are required.

Figure 2.18 Principal Components Analysis of cellular pathway abundance for samples with or without *Blastocystis* (a) between positive and negative samples and (b) positive and negative samples separated based on whether westernized or not.

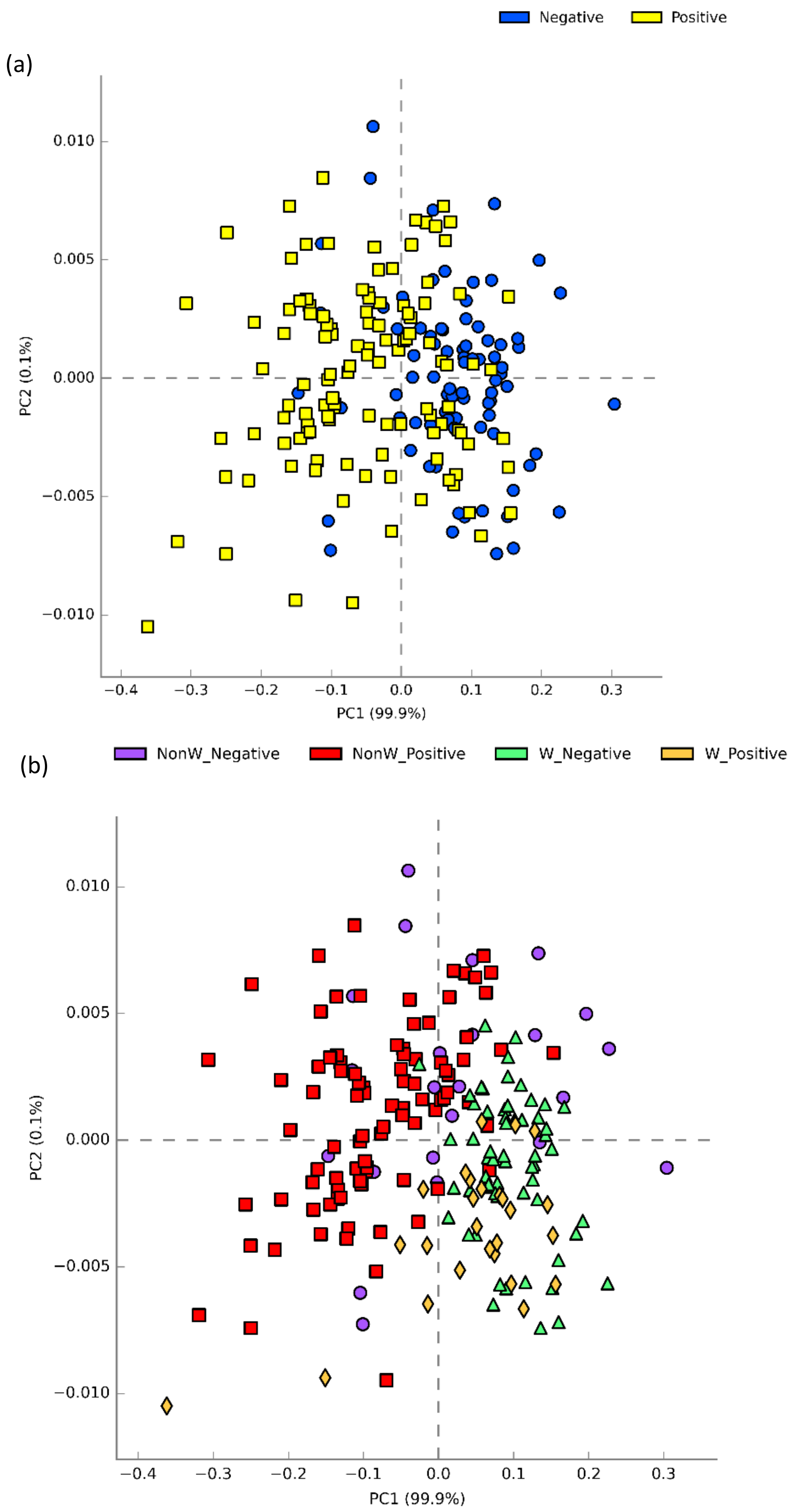
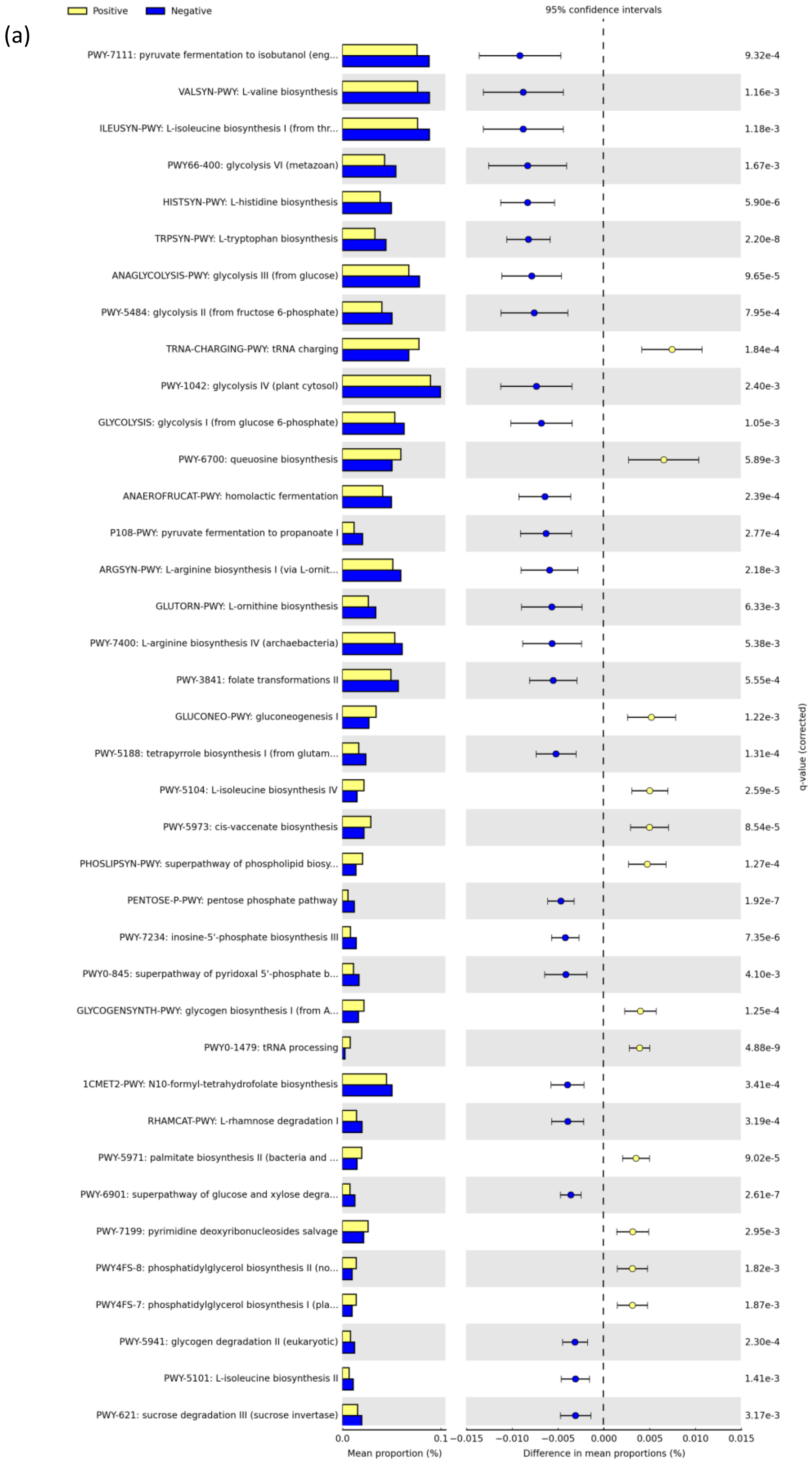

Negative
Positive
(a)
0.010
0.005
0.000
−0.005
−0.010
PC2 (0.1%)
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
PC1 (99.9%)
NonW_Negative
NonW_Positive
W_Negative
W_Positive
(b)
0.010
0.005
0.000
−0.005
−0.010
PC2 (0.1%)
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
PC1 (99.9%)

Figure 2.19 Gut microbiome cellular pathways significantly associated with the presence or absence of *Blastocystis*. Cellular pathways with relatively (a) high abundances and (b) high coverages in the samples with *Blastocystis* infection (yellow bar) *versus* absence (blue bar). Between-group differences were evaluated with Welch's *t*-test. The corrected p-values (q-values) were controlled for multiple testing according to Benjamini–Hochberg FDR corrections (FDR < 0.01). For pathway abundance, only the top 38 pathways of the results with effect size (ratio of proportion) > 1% are shown here. Groups of "UNINTEGRATED" and "UNMAPPED" were filtered before running the analysis.


(a)
Positive
Negative
95% confidence intervals
PWY-7111: pyruvate fermentation to isobutanol (eng...
VALSYN-PWY: L-valine biosynthesis
ILEUSYN-PWY: L-isoleucine biosynthesis I (from thr...
PWY66-400: glycolysis VI (metazoan)
HISTSYN-PWY: L-histidine biosynthesis
TRPSYN-PWY: L-tryptophan biosynthesis
ANAGLYCOLYSIS-PWY: glycolysis III (from glucose)
PWY-5484: glycolysis II (from fructose 6-phosphate)
TRNA-CHARGING-PWY: tRNA charging
PWY-1042: glycolysis IV (plant cytosol)
GLYCOLYSIS: glycolysis I (from glucose 6-phosphate)
PWY-6700: queuosine biosynthesis
ANAEROFRUCAT-PWY: homolactic fermentation
P108-PWY: pyruvate fermentation to propanoate I
ARGSYN-PWY: L-arginine biosynthesis I (via L-ornit...
GLUTORN-PWY: L-ornithine biosynthesis
PWY-7400: L-arginine biosynthesis IV (archaebacteria)
PWY-3841: folate transformations II
GLUCONEO-PWY: gluconeogenesis I
PWY-5188: tetrapyrrole biosynthesis I (from glutam...
PWY-5104: L-isoleucine biosynthesis IV
PWY-5973: cis-vaccenate biosynthesis
PHOSLIPSYN-PWY: superpathway of phospholipid biosy...
PENTOSE-P-PWY: pentose phosphate pathway
PWY-7234: inosine-5'-phosphate biosynthesis III
PWY0-845: superpathway of pyridoxal 5'-phosphate b...
GLYCOGENSYNTH-PWY: glycogen biosynthesis I (from A...
PWY0-1479: tRNA processing
1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis
RHAMCAT-PWY: L-rhamnose degradation I
PWY-5971: palmitate biosynthesis II (bacteria and ...
PWY-6901: superpathway of glucose and xylose degra...
PWY-7199: pyrimidine deoxyribonucleosides salvage
PWY4FS-8: phosphatidylglycerol biosynthesis II (no...
PWY4FS-7: phosphatidylglycerol biosynthesis I (pla...
PWY-5941: glycogen degradation II (eukaryotic)
PWY-5101: L-isoleucine biosynthesis II
PWY-621: sucrose degradation III (sucrose invertase)
0.0
0.1
Mean proportion (%)
−0.015
−0.010
−0.005
0.000
0.005
0.010
0.015
Difference in mean proportions (%)
9.32e-4
1.16e-3
1.18e-3
1.67e-3
5.90e-6
2.20e-8
9.65e-5
7.95e-4
1.84e-4
2.40e-3
1.05e-3
5.89e-3
2.39e-4
2.77e-4
2.18e-3
6.33e-3
5.38e-3
5.55e-4
1.22e-3
1.31e-4
2.59e-5
8.54e-5
1.27e-4
1.92e-7
7.35e-6
4.10e-3
1.25e-4
4.88e-9
3.41e-4
3.19e-4
9.02e-5
2.61e-7
2.95e-3
1.82e-3
1.87e-3
2.30e-4
1.41e-3
3.17e-3
q-value (corrected)

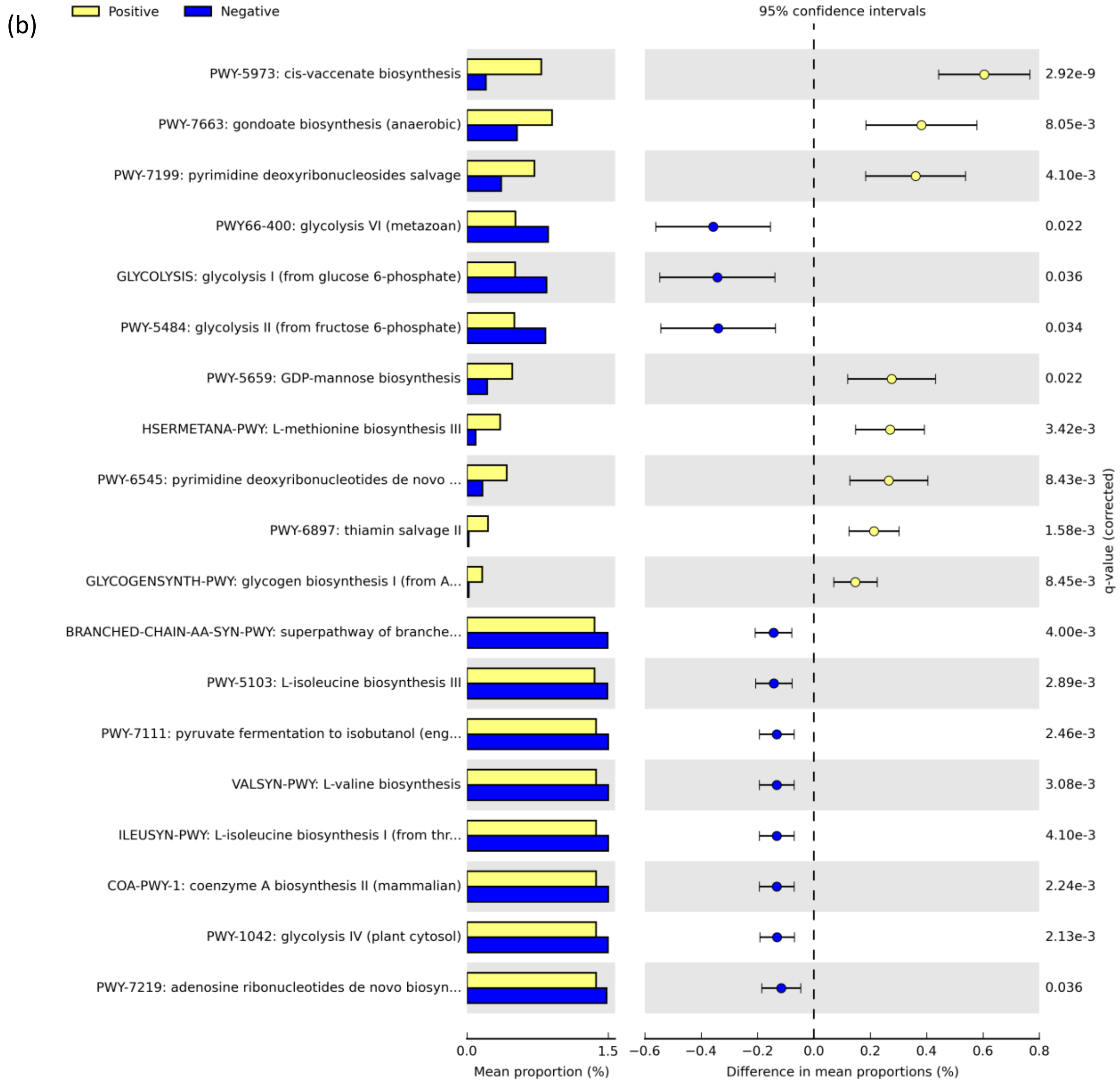

(b)
Positive
Negative
95% confidence intervals
PWY-5973: cis-vaccenate biosynthesis
PWY-7663: gondoate biosynthesis (anaerobic)
PWY-7199: pyrimidine deoxyribonucleosides salvage
PWY66-400: glycolysis VI (metazoan)
GLYCOLYSIS: glycolysis I (from glucose 6-phosphate)
PWY-5484: glycolysis II (from fructose 6-phosphate)
PWY-5659: GDP-mannose biosynthesis
HSERMETANA-PWY: L-methionine biosynthesis III
PWY-6545: pyrimidine deoxyribonucleotides de novo ...
PWY-6897: thiamin salvage II
GLYCOGENSYNTH-PWY: glycogen biosynthesis I (from A...
BRANCHED-CHAIN-AA-SYN-PWY: superpathway of branche...
PWY-5103: L-isoleucine biosynthesis III
PWY-7111: pyruvate fermentation to isobutanol (eng...
VALSYN-PWY: L-valine biosynthesis
ILEUSYN-PWY: L-isoleucine biosynthesis I (from thr...
COA-PWY-1: coenzyme A biosynthesis II (mammalian)
PWY-1042: glycolysis IV (plant cytosol)
PWY-7219: adenosine ribonucleotides de novo biosyn...
2.92e-9
8.05e-3
4.10e-3
0.022
0.036
0.034
0.022
3.42e-3
8.43e-3
1.58e-3
8.45e-3
4.00e-3
2.89e-3
2.46e-3
3.08e-3
4.10e-3
2.24e-3
2.13e-3
0.036
q-value (corrected)
0.0
1.5
Mean proportion (%)
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Difference in mean proportions (%)

Figure 2.20 Heatmap of enrichment or depletion of cellular pathways associated with *Blastocystis* STs. The rows and columns were clustered using complete linkage clustering of similarities in similarity microbial species abundance using the correlation distance function and the Bray-Curtis distance metric, respectively. Neg: *Blastocystis* absent. The number of samples in each group was labelled in the brackets at the bottom of each column. Statistical test used: Kruskal-Wallis test with Benjamini–Hochberg FDR corrections (FDR < 0.05).

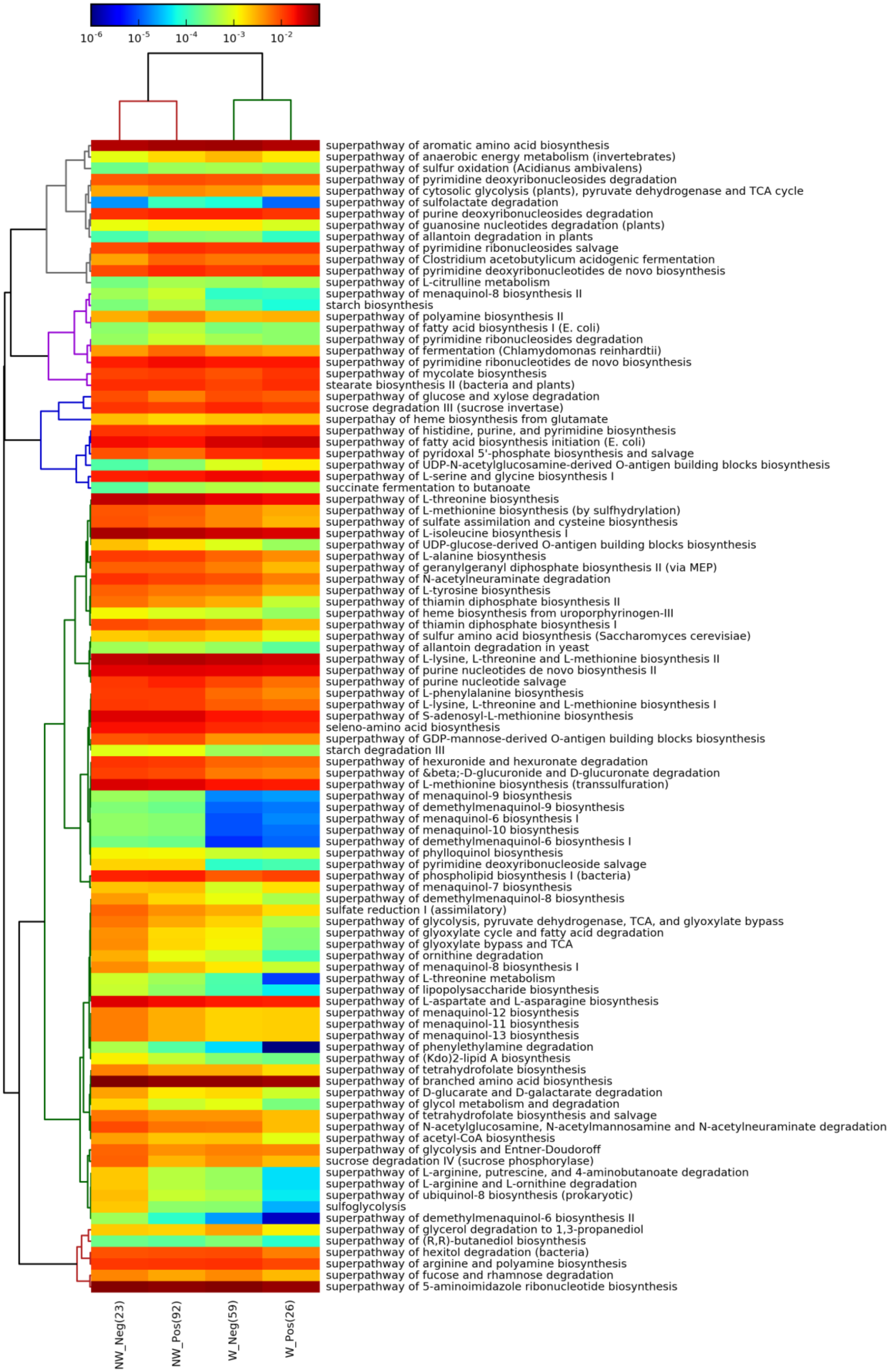
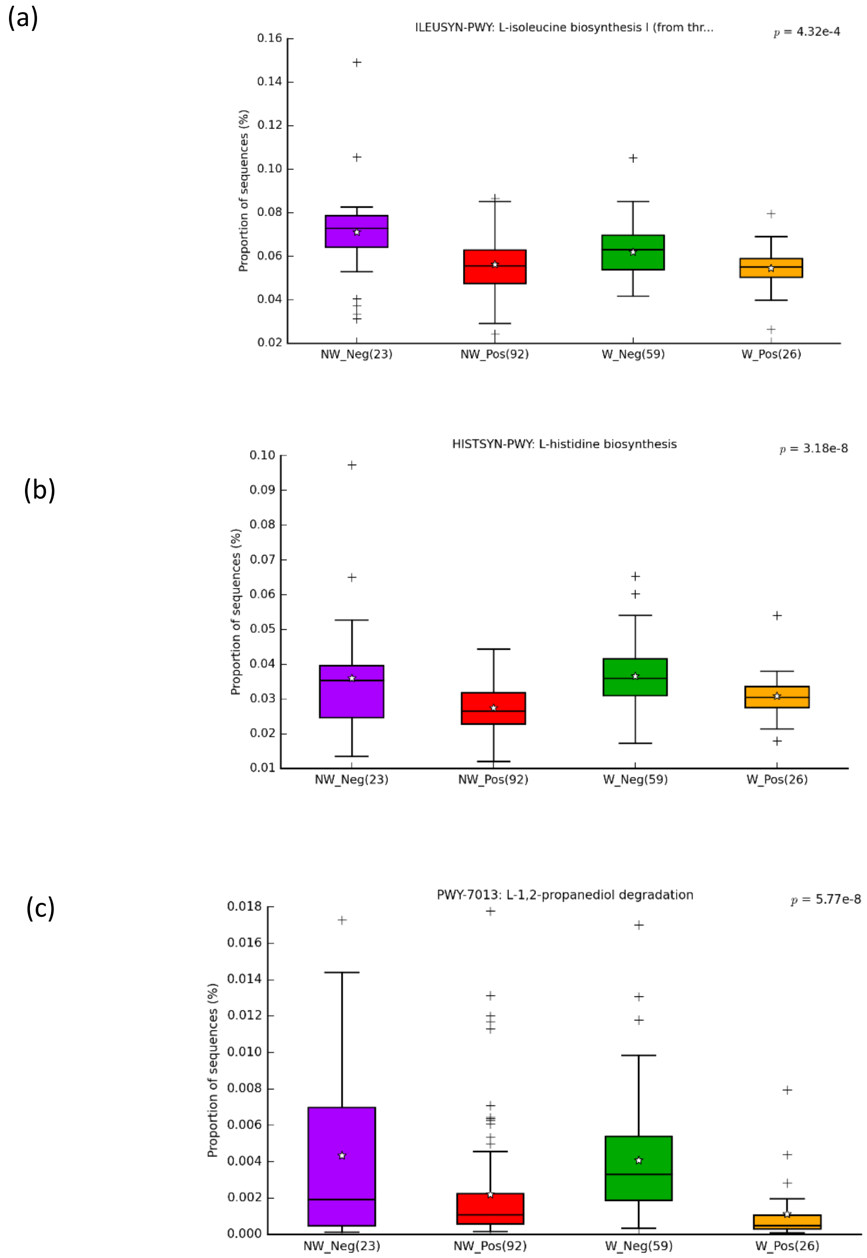

10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
superpathway of aromatic amino acid biosynthesis
superpathway of anaerobic energy metabolism (invertebrates)
superpathway of sulfur oxidation (Acidianus ambivalens)
superpathway of pyrimidine deoxyribonucleosides degradation
superpathway of cytosolic glycolysis (plants), pyruvate dehydrogenase and TCA cycle
superpathway of sulfolactate degradation
superpathway of purine deoxyribonucleosides degradation
superpathway of guanosine nucleotides degradation (plants)
superpathway of allantoin degradation in plants
superpathway of pyrimidine ribonucleosides salvage
superpathway of Clostridium acetobutylicum acidogenic fermentation
superpathway of pyrimidine deoxyribonucleotides de novo biosynthesis
superpathway of L-citrulline metabolism
superpathway of menaquinol-8 biosynthesis II
starch biosynthesis
superpathway of polyamine biosynthesis II
superpathway of fatty acid biosynthesis I (E. coli)
superpathway of pyrimidine ribonucleosides degradation
superpathway of fermentation (Chlamydomonas reinhardtii)
superpathway of pyrimidine ribonucleotides de novo biosynthesis
superpathway of mycolate biosynthesis
stearate biosynthesis II (bacteria and plants)
superpathway of glucose and xylose degradation
sucrose degradation III (sucrose invertase)
superpathay of heme biosynthesis from glutamate
superpathway of histidine, purine, and pyrimidine biosynthesis
superpathway of fatty acid biosynthesis initiation (E. coli)
superpathway of pyridoxal 5'-phosphate biosynthesis and salvage
superpathway of UDP-N-acetylglucosamine-derived O-antigen building blocks biosynthesis
superpathway of L-serine and glycine biosynthesis I
succinate fermentation to butanoate
superpathway of L-threonine biosynthesis
superpathway of L-methionine biosynthesis (by sulfhydrylation)
superpathway of sulfate assimilation and cysteine biosynthesis
superpathway of L-isoleucine biosynthesis I
superpathway of UDP-glucose-derived O-antigen building blocks biosynthesis
superpathway of L-alanine biosynthesis
superpathway of geranylgeranyl diphosphate biosynthesis II (via MEP)
superpathway of N-acetylneuraminate degradation
superpathway of L-tyrosine biosynthesis
superpathway of thiamin diphosphate biosynthesis II
superpathway of heme biosynthesis from uroporphyrinogen-III
superpathway of thiamin diphosphate biosynthesis I
superpathway of sulfur amino acid biosynthesis (Saccharomyces cerevisiae)
superpathway of allantoin degradation in yeast
superpathway of L-lysine, L-threonine and L-methionine biosynthesis II
superpathway of purine nucleotides de novo biosynthesis II
superpathway of purine nucleotide salvage
superpathway of L-phenylalanine biosynthesis
superpathway of L-lysine, L-threonine and L-methionine biosynthesis I
superpathway of S-adenosyl-L-methionine biosynthesis
seleno-amino acid biosynthesis
superpathway of GDP-mannose-derived O-antigen building blocks biosynthesis
starch degradation III
superpathway of hexuronide and hexuronate degradation
superpathway of &beta;-D-glucuronide and D-glucuronate degradation
superpathway of L-methionine biosynthesis (transsulfuration)
superpathway of menaquinol-9 biosynthesis
superpathway of demethylmenaquinol-9 biosynthesis
superpathway of menaquinol-6 biosynthesis I
superpathway of menaquinol-10 biosynthesis
superpathway of demethylmenaquinol-6 biosynthesis I
superpathway of phylloquinol biosynthesis
superpathway of pyrimidine deoxyribonucleoside salvage
superpathway of phospholipid biosynthesis I (bacteria)
superpathway of menaquinol-7 biosynthesis
superpathway of demethylmenaquinol-8 biosynthesis
sulfate reduction I (assimilatory)
superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass
superpathway of glyoxylate cycle and fatty acid degradation
superpathway of glyoxylate bypass and TCA
superpathway of ornithine degradation
superpathway of menaquinol-8 biosynthesis I
superpathway of L-threonine metabolism
superpathway of lipopolysaccharide biosynthesis
superpathway of L-aspartate and L-asparagine biosynthesis
superpathway of menaquinol-12 biosynthesis
superpathway of menaquinol-11 biosynthesis
superpathway of menaquinol-13 biosynthesis
superpathway of phenylethylamine degradation
superpathway of (Kdo)2-lipid A biosynthesis
superpathway of tetrahydrofolate biosynthesis
superpathway of branched amino acid biosynthesis
superpathway of D-glucarate and D-galactarate degradation
superpathway of glycol metabolism and degradation
superpathway of tetrahydrofolate biosynthesis and salvage
superpathway of N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminate degradation
superpathway of acetyl-CoA biosynthesis
superpathway of glycolysis and Entner-Doudoroff
sucrose degradation IV (sucrose phosphorylase)
superpathway of L-arginine, putrescine, and 4-aminobutanoate degradation
superpathway of L-arginine and L-ornithine degradation
superpathway of ubiquinol-8 biosynthesis (prokaryotic)
sulfoglycolysis
superpathway of demethylmenaquinol-6 biosynthesis II
superpathway of glycerol degradation to 1,3-propanediol
superpathway of (R,R)-butanediol biosynthesis
superpathway of hexitol degradation (bacteria)
superpathway of arginine and polyamine biosynthesis
superpathway of fucose and rhamnose degradation
superpathway of 5-aminoimidazole ribonucleotide biosynthesis
NW_Neg(23)
NW_Pos(92)
W_Neg(59)
W_Pos(26)

Figure 2.21 Bar plot for comparison of proportion of sequences in bacterial species that had significant association with *Blastocystis* STs in non-westernized or westernized individuals. The number of samples in each group was labelled in the brackets at the bottom of each column. Statistical test used: Kruskal-Wallis test with Benjamini–Hochberg FDR corrections. NW: non-westernized, W: westernized, Pos: *Blastocystis* present, Neg: *Blastocystis* absent.



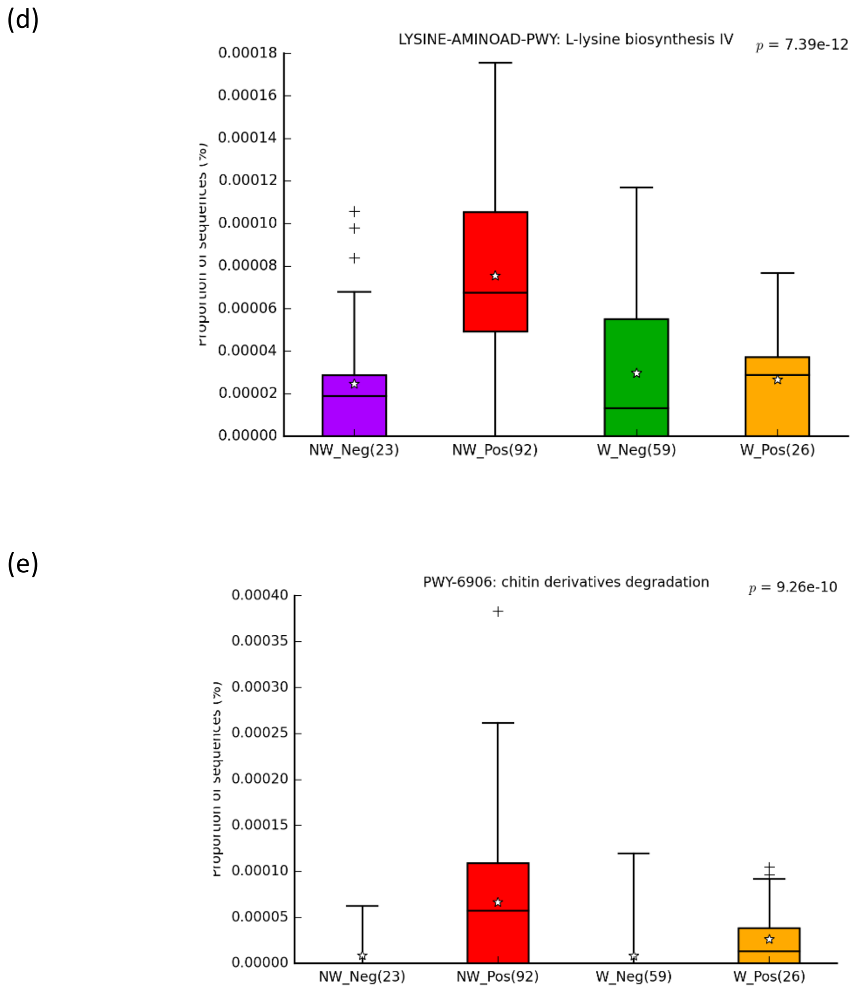
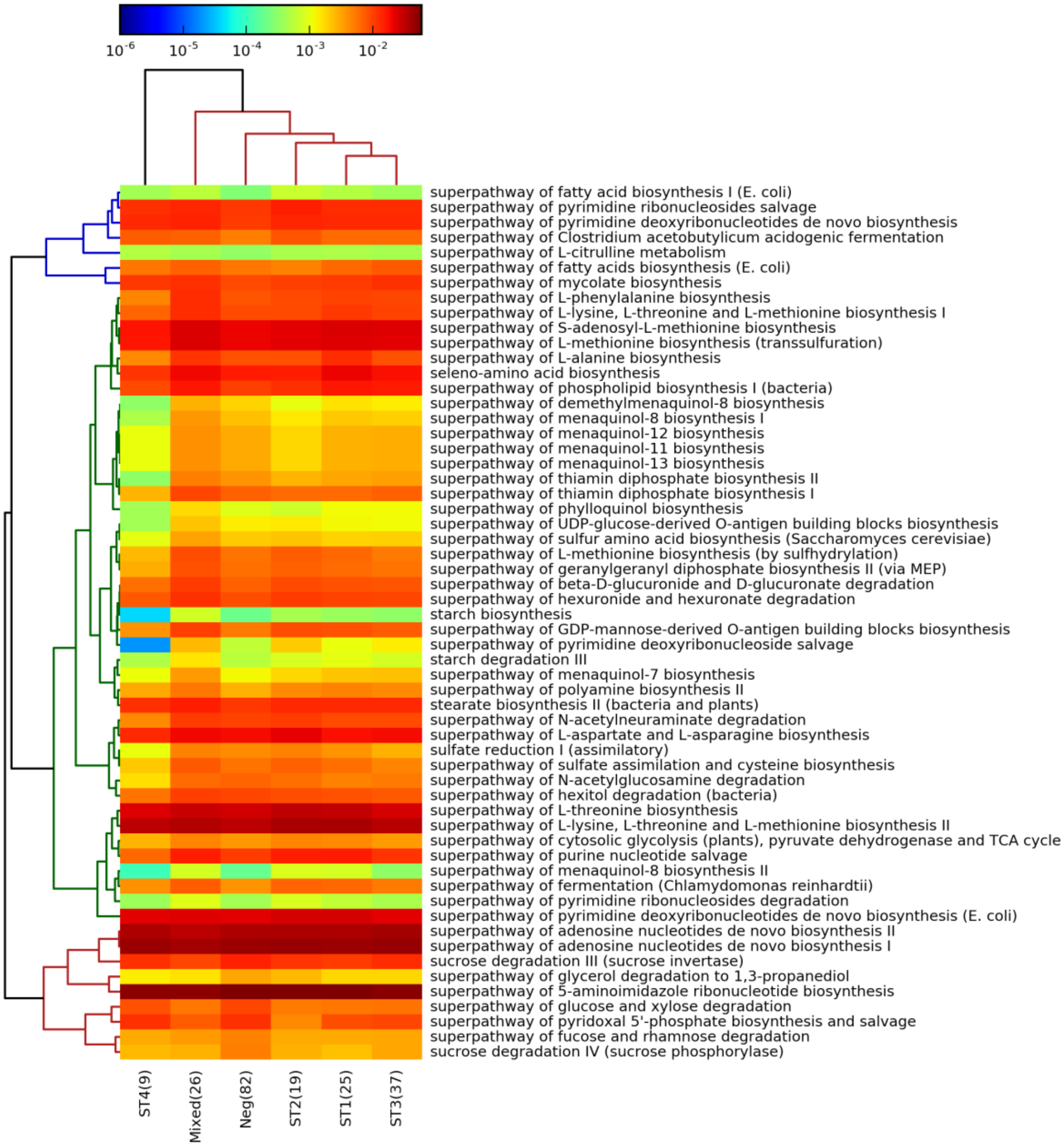

(d)
LYSINE-AMINOAD-PWY: L-lysine biosynthesis IV
p = 7.39e-12
Proportion of sequences (%)
0.00018
0.00016
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
NW_Neg(23)
NW_Pos(92)
W_Neg(59)
W_Pos(26)
(e)
PWY-6906: chitin derivatives degradation
p = 9.26e-10
Proportion of sequences (%)
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
NW_Neg(23)
NW_Pos(92)
W_Neg(59)
W_Pos(26)

Figure 2.22 Heatmap of enrichment or depletion of cellular pathways associated with *Blastocystis* STs. The rows and columns were clustered using complete linkage clustering of similarities in similarity microbial species abundance using the correlation distance function and the Bray-Curtis distance metric, respectively. Neg: *Blastocystis* absent. The number of samples in each group was labelled in the brackets at the bottom of each column. Statistical test used: Kruskal-Wallis test with Benjamini–Hochberg FDR corrections (FDR < 0.05). Hierarchical clustering solution UPGMA dendrogram was based on Bray-Curtis distance metric and complete clustering method



2.4 DISCUSSION

2.4.1 Advantages and limitations of microbial eukaryotic diagnosis using a metagenomic approach

Blastocystis is the most common eukaryotic microbe inhabiting human and animal guts yet, until recently, it was not considered in most gut microbiome studies. Various diagnostic methods have been used to detect *Blastocystis* in stool samples, but DNA sequencing-based methods are culture-independent, more sensitive and have the advantage that the diversity, composition, and putative biological functions of the gut microbiome can be investigated. The WGS metagenomic approach can be as accurate and sensitive as PCR (Lokmer et al. 2019), and it allows for better taxonomic annotation and abundance estimation when compared to PCR or 18S rRNA amplicon gene sequencing (Khachatryan et al. 2020). WGS-based metagenomic analyses have been shown to have advantages over 16S amplicon sequencing analyses including increased accuracy, the capability of detecting more genera of eukaryotes and viruses, and enabling prediction of putative functional genes (Brumfield et al. 2020).

In this study, a taxonomy classification-based bioinformatic workflow for detection and characterization of *Blastocystis* was developed and applied to 996 human and animal WGS metagenomic datasets. With a focus on the effect of *Blastocystis* colonization in the gut microbiome, a large proportion of the datasets chosen for the study were from healthy individuals living in the non-westernized developing countries. This is the reason for the high overall prevalence detected in human samples in this study. With the rapid increases in the number of WGS metagenomic sequencing datasets for gut microbiomes published on NCBI that were not used for microbial eukaryotic studies, my workflow can easily be applied to these datasets to investigate the prevalence of *Blastocystis* in different cohorts.

Unlike marker gene studies, WGS metagenomic analyses also allow profiling of the changes in the composition of the community and metabolic pathways in the gut microbiome upon *Blastocystis* colonization. Several bacterial and archaeal species and cellular pathways were significantly associated with the presence of *Blastocystis*. The associations were, in some cases, subtype-dependent. However, despite the richer information afforded by these kinds of analyses an important limitation of WGS metagenomic analysis is the heavy requirement of computational resources, like memory and disk space, and computation time. This computational burden limited the number of samples and kinds of analyses that could be completed in this study. However, with computational efficiency of methods and advances in computer software and hardware, WGS metagenomic methods and the analysis workflow presented here could become the method-of-choice for *Blastocystis* diagnostics and scientific investigations in future.

2.4.2 Many factors can affect detection results in metagenomic analyses

The database that contains genomes from all taxonomy groups and specialized for the gut microbiome can significantly improve the sensitivity and accuracy of read classification in the taxonomy classification-based detection workflow. Choosing taxonomy classifiers that can build a custom database is essential for the performance of the workflow and Centrifuge was chosen because of its capacity for using custom databases and low memory requirements (Méric et al. 2019; Watts et al. 2019). By replacing the fungal genomes with genome sequences of amphibians and reptiles and showing that turtles, frogs, and snakes as the most abundant species in the gut microbiome from Tanzania hunter-gatherers, Marcelino et al. (2020) demonstrated the importance of reference databases containing genomes from all major taxonomic groups (bacteria, archaea, eukaryotes, and viruses) for metagenomic classification. Larger custom databases containing as many representative genomes as possible compared to NCBI Refseq nt database can dramatically improve taxonomy classification performance and accuracy in a metagenomic analysis (Méric et al. 2019; Ye et al. 2019). This was also clearly demonstrated in my analyses where the development and use of specialized database 1 showed clear advantages in detecting microbial taxa over the NCBI nt database (Figure 2.2).

When detecting eukaryotes or pathogens from shotgun metagenomic datasets, the majority of studies use reference genome mapping and then calculate the breadth of coverage (Beghini et al. 2017; Olm et al. 2019). One big limitation for reference genome mapping-based detection is that it highly depends on the availability of the reference genomes, which restrict its application to many eukaryotes (e.g., there are no genomes available for most non-human-infecting *Blastocystis* STs). For instance, some pig samples

(Table 2.7) would be negative for *Blastocystis* colonization based on breadth of coverage defined by Beghini et al. (2017), but when using taxonomy classification and marker gene detection from the MAG assembly, they have a clear signal for carrying *Blastocystis* ST5 (Table 2.8).

The definition of the threshold for positive samples is critical for not only *Blastocystis* profiling in the gut and but also for finding any target species in environmental samples. So far, there is no consensus on the threshold to be used to define a positive sample in metagenome datasets. For analyses that used reference genome mapping methods, the cutoff values for the breadth of coverage for defining a positive samples have varied substantially. For example, Beghini et al. used a threshold of 10% genome coverage for *Blastocystis* (Beghini et al. 2017) whereas Olm and colleagues used 50% for fungi (Olm et al. 2019). Another study of *Blastocystis* by Lokmer et al. (2019) used a cutoff of 10% of positive contigs that were detected from the reference genome with a breadth of coverage >0.001. Cutoff values from relative abundance estimation by taxonomy classifiers like Centrifuge or Kraken can also be used for deciding positive samples (Ye et al. 2019). In this study, because almost no reference genomes are available for animal-specific *Blastocystis* STs, the relative abundances found by the reference-mapping method were extremely precluding their use for defining a threshold for scoring positive samples. Instead, I opted to use number of reads aligned to contigs in *Blastocystis* genomes when defining the positive samples. This number excluded all possible artificial *Blastocystis* reads related to contaminations or errors before or after sequencing.

Due to the difference in methodologies and cutoff values, the prevalence of *Blastocystis* from the same datasets but detected by different studies can vary accordingly (Figure 2.6). With my approach, I detected similar numbers of positive and co-infected samples to Lokmer et al., (2019) for two out of three projects, markedly more positive samples compared to the analyses by Beghini et al. (2017) and Forsell et al. (2017). Marcelino et al. (2020) used the kingdom-agnostic CCMetagen metagenomic pipeline with the NCBI Refseq nt database as a reference and detected *Blastocystis* in 15/27 Tanzanian samples but none in Italian samples. Although they did not specify the STs, the total number of positive samples was the same as that found by Beghini et al. (2017). Both the Marcelino *et al.* and Beghini *et al.* studies excluded the possibility of co-infection in their study. My approach is designed to be more sophisticated and robust, but further study is required to know to what extent false positives and false negatives are occurring. When using a metagenomic approach to detect *Blastocystis*, the choice of tools and databases, as well as the thresholds for defining positive samples are essential and studies using pre-constructed mock communities with known species abundance to test different tools, databases, and threshold values may help to come to a consensus.

2.4.3 Comparison of *Blastocystis* prevalence and subtype distribution in recent studies

Blastocystis was detected in human samples from 9 of the 10 projects which is consistent with previous results reporting a global distribution of *Blastocystis* in all continents (Alfellani, Stensvold, et al. 2013). Generally, it has a lower prevalence in westernized countries, i.e., in Europe and North America with frequencies ranging from 0% to 30%, with the exception of some European cohorts where the prevalence was >30% or even exceeding 50% (e.g., 35.2% in a study in northern Spain (Paulos et al. 2018) and 62.8% for Swedish university students in this study). The prevalence of *Blastocystis* in non-westernized countries is moderate to high and may exceed 60% in Africa, Asia, and South America, which is consist with current literature (El Safadi et al. 2014; Forsell et al. 2016; Jiménez et al. 2019).

For individual STs, in contrast to previous studies showing ST3 as the most common subtype (Alfellani, Stensvold, et al. 2013; Tito et al. 2018), this is the first large-scale study to show that, for cases of single subtype colonization, ST1-positive samples exceeded ST3 samples. One possible reason for this observation is that ST3 is more frequently found in co-infections than ST1. A geographically structured distribution of subtypes was also confirmed in this study. ST1, ST2, and ST3 had a broad distribution, ST4 and ST8 are more common in western countries, and ST5, ST6, ST7, and ST9 are rare human subtypes (I did not detect any of these STs in human samples except 1 sample carrying ST7). The low infection rate of ST4 detected in this study (3.5%, 9 of 255 positive samples) compared to the results of Beghini et al. (2017) that had a 31.1% of ST4 (99 of 318 positive samples) is explained by the fact that there were relatively few European samples (37%, 179 of total 485 individuals) in this study compared to the samples they analyzed (71%, 1204 of total 1689 individuals were from Europe or USA). The overall prevalence of co-infections was very similar to the study by Lokmer et al. (2019) who also used metagenomic analysis. In agreement with current literature, I found the prevalence of *Blastocystis* is high in the healthy population and low in obese or overweight individuals (Beghini et al., 2017; Tito et al, 2018).

For animal samples, the prevalence of *Blastocystis* varied in different countries and the distribution of subtypes also depended on hosts and geographical region (Alfellani, Taner-Mulla, et al. 2013). Here I reported higher colonization of *Blastocystis* in baboons (50%, 24/48) than previous studies: Legesse and Erko (2004) found a frequency of 3.3% in Ethiopian baboons using microscopical examination (Legesse & Erko 2004), whereas no baboons were found to be infected in the Bangladesh National Zoo by Li et al. (2019) as assayed by PCR amplification (Li et al. 2019). Only ST1 and ST2 were detected from baboon samples and the common ST3, ST5, and ST8 existing in non-human primates (Alfellani, Jacob, et al. 2013; Betts et al. 2020) was not found in baboons in my study.

On average, 15% of cattle samples in this study were positive for *Blastocystis*, a relatively low overall infection rate compared to previous findings (Aynur et al. 2019) but a higher rate compared to American cattle samples (Maloney, Lombard, et al. 2019). Four subtypes (ST1, ST3, ST5, and ST10) and two types of mixed infection (ST1 + ST3 and ST1 + ST5) were identified. The most common cattle subtype ST10 was only found in one of two projects and other common subtypes like ST4 and ST14 were not found (Aynur et al. 2019; Greige et al. 2019). The mixed subtype of *Blastocystis* infection in American cattle was only found in 25% of the positive samples (2/8) with only one combination of co-infection (ST1 + ST3). This is a much lower rate and a less complex situation than that described in the recent study by Maloney et al. (2019). These researchers compared NGS amplicon sequencing with Sanger sequencing and defined a sample positive for a ST if >= 0.1% of merged contigs mapped to the 18S rRNA gene sequence of specific *Blastocystis* ST (the length of merged contigs were between 400 to 600 bps; the average pairs of reads in their study was 101,785). Using this threshold, they detected a total of 14 subtypes (ST1 to ST5, ST10, ST11, ST14, ST17, ST21, and ST23 to ST26) from 75 amplicon sequencing datasets from cattle feces and found 65% (49/75) of the positive samples were mixed infections, 41% (20/49) of which contained ≥ 3 STs with one sample being infected with 8 different subtypes. This demonstrates that NGS amplicon sequencing is a powerful tool to detect low abundant subtypes and mixed infections of *Blastocystis*. However, it also suggests that careful interpretation is necessary for using NGS sequencing to detect *Blastocystis*.

For pig samples in my study, the overall prevalence (85.3%) and individual prevalence are very high except for one German project (9.1%, 2/22). A total of five subtypes (ST1, ST3, ST5, ST13, and ST15) were identified from pig positive samples. ST5 was the most common subtype and unexpected ST15 was the second most common ST (19.2%, 29/255) in all pig projects. Previously, ST15 was detected in artiodactyls (camels, cattle, and sheep) and non-human primates (chimpanzees and gibbons) (Alfellani, Jacob, et al. 2013; Betts et al. 2020) but was first reported in pig feces by Wylezich et al. (2019). The high prevalence of ST15 in pig samples as individual ST and mixed infections with ST5 reveals that it is a previously underappreciated ST colonizing pigs. The origin of ST15 in mammals is unusual because phylogenies of SSU rRNA show that it is only distantly related to other main mammalian *Blastocystis* STs, branching instead within a clade otherwise made up solely of lineages from reptiles or amphibians (Alfellani, Jacob, et al. 2013).

2.4.4 Compositional and functional profiling of *Blastocystis* in the gut microbiome

Several studies have demonstrated that colonization with *Blastocystis* is strongly associated with broad shifts in gut microbial communities (Beghini et al. 2017; Nieves-Ramírez et al, 2018; Tito et al., 2018), but many of these studies report different gut bacterial taxa in the gut that correlate with *Blastocystis* in different populations or cohorts with diseases like IBS patients. *In vitro* and *in vivo* studies showing that different *Blastocystis* subtypes had different effects on gut microbiota suggested that the interactions between *Blastocystis* and gut bacteria are likely subtype-dependent and need to be analyzed at the level of subtype (Yason et al. 2019). For this reason, I choose to focus my investigations of the association between *Blastocystis* colonization and the gut microbiome on the subtype level while taking into account the differences between westernized and non-westernized carriers.

In this study, I found that the presence of *Blastocystis* was strongly associated with an increase in two *Methanobrevibacter* species: *M. smithii* and an unclassified species, a result concordant with the finding of Beghini et al. (2017). Interestingly, when separating samples into westernized and non-westernized groups, I found that the high abundance of these two *Methanobrevibacter* species was only associated with non-westernized *Blastocystis* carriers, especially amongst individuals with ST1 and ST2. The biological significance of this association is unclear. *Methanobrevibacter smithii* is a dominant species of the methanogenic archaea fund in the human gut and can comprise up to 10% of all the anaerobic microorganisms in the colon (Samuel et al. 2007). It is capable of converting bacterial fermentation products like H_2 and CO_2 to methane that makes it essential for syntrophic associations within the gut microbial community (Bang et al. 2014). One possibility is that *Blastocystis* ST1 and ST2 produce significant amounts of H_2 and CO_2 and this could be 'feeding' the growth of the *Methanobrevibacter* species. This does make sense as *Blastocystis* is an anaerobic fermenter with a hydrogenosome-like MRO that may be producing H_2 and CO_2 (Gentekaki et al. 2017), although production of hydrogen has not been observed at least for ST7 (Lantsman et al. 2008).

The medical implications for the association between *Blastocystis* and *Methanobrevibacter* are also unclear. Previous studies have showed an increase of *M. smithii* in IBS patients (Kim et al. 2012; Nagel et al. 2016) and a potential association with diet-induced weight gain and obesity (Mathur et al. 2012; Mbakwa et al. 2015). One study even suggested it may induce an inflammatory response (Bang et al. 2014). However, *Blastocystis* colonization has not been associated with any of these factors in recent large-scale population studies (e.g., Tito et al. 2018), with the possible exception of one study that found an association with one type of IBS (Nourrisson et al., 2014). On the other hand, a recent study showed that *M. smithii* was significantly decreased in IBD patients compared to the healthy control group and this reverse association suggested it might be a biomarker for IBD (Ghavami et al. 2018). This fits with the fact that *Blastocystis* has also been found to have a clearly lowered prevalence in IBD cohorts (Andersen et al. 2015; Tito et al. 2018). In any case, results in this study indicate that the association of *Blastocystis* with *Methanobrevibacter* is likely subtype specific and mostly confined to non-westernized populations. Further studies of this association and its clinical relevance should take this into account.

At the bacterial phylum level, a high abundance of bacterial taxa within *Firmicutes* and low abundance of *Bacteroidetes* were found in *Blastocystis*-positive samples, in line with previous reports (Andersen et al, 2015; Beghini et al. 2017). Two species from the *Firmicutes* phylum, *Faecalibacterium prausnitzii* and *Eubacterium eligens*, are both commensal bacteria found in the healthy intestine; in fact, *F. prausnitzii* can be used as a biomarker to assist in gut diseases diagnostics (Chung et al. 2016; Lopez-Siles et al. 2017). These species are both strongly associated with *Blastocystis* colonization in general and also specifically with ST1-3. In contrast to the finding by Nieves-Ramírez et al (2018) that *Blastocystis* colonization was strongly associated with a decrease in *Precotella copri* in healthy individuals from a semi-industrialized region in rural Mexico, two *Precotella* species, *P. copri* and *P. stercorea*, were predicted to be enriched in *Blastocystis* carriers (Figure 2.14 (c)-(d) and Figure 2.16 (c)-(d)), especially in non-westernized carriers. This association suggests that *Blastocystis* may play a role in the abundance of *Prevotella* species besides a potential richness of plant-rich diets in non-western populations (De Filippo et al. 2010; Yatsunenko et al. 2012). The previous finding of an association between '*Ruminococcus*-enterotype' individuals and *Blastocystis* colonization was only observed in westernized non-carriers in this study (Andersen et al, 2015). *Akkermansia muciniphila*, a potential probiotic with a positive effect on metabolic syndrome in obese humans, showed a positive association with *Blastocystis* in ST4 samples, but a negative association for ST1-3 (Figure 2.17), which extends the finding by Tito et al. (2018).

It should be noted that the level of unmapped reads could affect the significance of relative abundancies between samples since there were wide variations in the percentages of unaligned reads across samples, ranging from 20% to 95% with an average of 69% (median 72%) for mapping reads to pangenome databases at the species-level (Section 2.2.5), and from 15% to 55% (mean 33% and median 32%) for mapping after translation (Supplementary Figure S4 (a)). When separating these samples into *Blastocystis* positive and negative groups in terms of westernized and non-westernized, the rates of unaligned reads in non-western groups were significantly higher than those in western groups, which suggests an underrepresentation of gut species in non-westernized populations in the current databases (Supplementary Figure S4 (b) and (c)) (Ayeni et al. 2018; Brewster et al. 2019).

Nevertheless, the results observed here are consistent with the hypothesis that *Blastocystis* and the other microbial inhabitants of the gut influence each other. However, the extent of the interactions, and the extent to which they vary amongst subtypes, requires many more samples of WGS metagenomic sequencing data from many more individuals associated with various types of metadata. With the increase of publicly available metagenomes in number and size from diverse populations worldwide, the diagnosis of *Blastocystis* together with the compositional and functional profiling in the gut microbiome should greatly improve. These data, coupled with experimental in vitro and in vivo studies of the physiological relevance of genomic differences between subtypes, will help resolve unanswered questions about the pathogenicity and physiological role of *Blastocystis*.

Table 2.7 Detection of *Blastocystis* in a pig sample (DRR025071) by using the reference genome mapping method (Bowtie2 and SAMtools; Beghini et al., 2017) compared to taxonomy classification methods (Centrifuge).

Subtype of reference genome	Reference genome mapping results		Centrifuge results
	base covered at X=1	Breadth of coverage	num Reads mapped
ST1	18889	0.11%	2706
ST2	414	0.00%	1142
ST3	366	0.00%	1575
ST4	2567	0.02%	2227
ST6	757	0.01%	774
ST7	48793	0.27%	1587
ST8	1627	0.01%	813
ST9	470	0.00%	803
Total	-	-	11627

Table 2.8 *Blastocystis* marker genes detected by Metaxa2 from metagenome assembly of the pig sample (DRR025071).

Contigs	Metaxa2 results			ST detected by BLAST
	rRNA gene	aligned length	identity	
Contig_8895 length_5442 cov_156.468907	LSU	1121	85.0%	ST5
Contig_8895 length_5442 cov_156.468907	SSU	1410	99.3%	ST5
Contig_931 length_16912 cov_34.616836	Mitochondrial LSU	2365bp	89.26%	ST5

CHAPTER 3 EUKFINDER: A BIOINFORMATIC WORKFLOW TO RETRIEVE MICROBIAL EUKARYOTE GENOMES FROM ENVIRONMENTAL METAGENOMIC SEQUENCING DATA

3.1 INTRODUCTION

Unicellular protists are ubiquitous and inhabit every global ecosystem including freshwater, marine, terrestrial and the GI tracts of humans and animals. The genome sequences of these microbial eukaryotes inform us of their physiology capacities, evolutionary histories, as well as their interactions with other microbes and/or host and their environment. Unfortunately, unlike for prokaryotes, there is still a lack of genome information for diverse protistan species (Sibbald & Archibald 2017). Many protists are difficult to bring into culture; fewer can be cultivated in pure axenic conditions and, for those that can be, scaling up cultures and extracting pure DNA is laborious and time-consuming. For these reasons, high-throughput analyses of population genomics of protists have lagged behind those of prokaryotes.

Whole genome shotgun (WGS) metagenomic sequencing is a technology that could make it possible to characterize protistan genomes in the environment without the need for cultivation. WGS metagenomic approaches enable the simultaneous sequencing of multiple genomes from microorganisms living in the communities of complex ecosystems. In WGS metagenomics, DNA from all the microorganisms in the community within an environmental sample is extracted and sequenced to generate millions of short-length reads (100 – 300 bp) that are assembled into continuous genome fragments (i.e., contigs) to allow the recovery of full-length gene sequences or even longer gene clusters. In addition, sorting the assembled contigs into categories (commonly called bins) separates fragments that likely originated from different taxa by grouping them into species (or closely related strains) based on their genome composition (e.g., k-mers) and/or depth of coverage, leading to partial or even complete reconstruction of their genomes. This computational method has been standardized and applied to various environmental samples. Since the first near-complete bacterial genomes were reconstructed by Tyson et al. (2004), using metagenomic sequencing from a low-complexity microbial environment, thousands of high-quality complete or near-complete genomes for bacteria and archaeal species have been reconstructed, which is the main reason for the dramatic growth of available prokaryotic genome data on NCBI (Figure 3.1).

Although the gut microbiome is one of a few heavily studied microbial environments with thousands of novel bacterial genomes sequenced using cultured-based and increasingly metagenomic approaches each year, the number of published high-quality draft genomes for gut microbial eukaryotes remain very few (Table 3.1). The application of WGS metagenomics to eukaryotic microbes is not well-established due to the large size, complexity and repetitive nature of eukaryotic genomes. In addition, the fact that eukaryotic reads are usually only found in a very small proportion (generally < 5%) in the metagenomic sequencing data with low coverage makes the recovery of eukaryotic genomes even more challenging. To date, only a handful of investigations have used a metagenomic approach to reconstruct eukaryotic genomes. For example, Beghini et al (2017) used a bioinformatic approach to map reads to *Blastocystis* reference genomes from gut metagenomic data and extract reads that align to the reference genome to do *de novo* assembly. With this approach, they were able to assemble 43 draft *Blastocystis* genomes from 2154 publicly available gut metagenomic datasets. Among these genomes, 19 had sizes > 5 Mb with a completeness of 33% to 85% based on the assembly size estimation. West and colleagues developed a k-mer-based tool, EukRep, for separating eukaryotic genomes from prokaryotic ones in MAGs (West et al. 2018). EukRep employed a machine-learning strategy with linear support-vector machine (SVM) classifiers to detect and select eukaryotic contigs based on k-mer frequencies. They trained the SVM classifier with 5-mer frequencies that they extracted from a database of reference genomes they constructed from several sources. When EukRep was applied to metagenomic assemblies from infant fecal samples, six near-complete genomes of fungi were retrieved. A recent study applied EukRep to 1174 infant fecal metagenomes and 24 metagenomes from hospital rooms and in total 14 eukaryotic metagenome assembled genomes (MAGs) were recovered (12 fungi, one belonging to the clade of Diptera and one Nematoda) with a median estimated completeness of 91% (Olm et al. 2019). Analyses of these genomes allowed detailed genomic comparisons and detection of population micro-diversity among different fungi. The foregoing studies demonstrate that it is possible to reconstruct microbial eukaryotic genomes without cultivation and targeted DNA isolation work. However, each of these pipelines has limitations. For example, the reference genome mapping approach cannot apply to organisms without available reference genomes and the performance of the machine-learning approach EukRep can be affected by the accuracy and consistency of the assembly tools and the training reference genome sets. Furthermore, for the latter, there are no published instructions on how to build a custom training genome set.

To circumvent the limitations of the foregoing metagenomic analysis pipelines, I developed a bioinformatic tool, Eukfinder, to recover and assemble eukaryotic nuclear and mitochondrial genomes from environmental metagenomes. Eukfinder improves upon the existing pipelines by adding a pre-selection step classifying reads based on taxonomy and two specialized databases that can be built by users to include the reference genomes from the representative organisms in the environment. To demonstrate its utility, I have applied it to human gut metagenomic datasets to recover nuclear and mitochondrial (MRO) genomes, focusing on *Blastocystis* genomes from human gut metagenomic datasets as a test case.

Blastocystis is a good example of a gut-dwelling protist that is extremely common in human populations but very difficult to bring into stable culture. As a result, relatively little is known about the genetic diversity among and between *Blastocystis* subtypes and how this may affect their potential for pathogenicity. The few publicly available *Blastocystis* nuclear genomes range from 12.9 Mbp to 18.8 Mbp in size and vary markedly in their GC content (39.6% - 54.6%) and gene content (Denoeud et al., 2011; Wawrzyniak et al., 2015; Gentekaki et al., 2017). They also have a number of notable features including genes that require mRNA polyadenylation to create functional termination codons (Klimeš et al. 2014; Gentekaki et al. 2017) as well as genes acquired by lateral gene transfer that allow them to thrive in the gut environment, evade the immune system and potentially modulate the growth of other gut microbes (Eme et al., 2017). *Blastocystis* also have genome-containing mitochondrion-related organelles (MROs) (Jacob et al., 2016) that are adapted to function in anaerobic conditions of the animal gut (Tan et al., 2008). Both nuclear and MRO genomes potentially offer insights to help us understand differences between *Blastocystis* STs that can guide future experimental investigations into their pathogenicity, as well as, to detect possible targets for anti-protozoan drug development.

Figure 3.1 Dramatic growth of published genomes in NCBI GenBank that include all assembly levels: complete, chromosome, scaffold, contig, based on data from ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/.

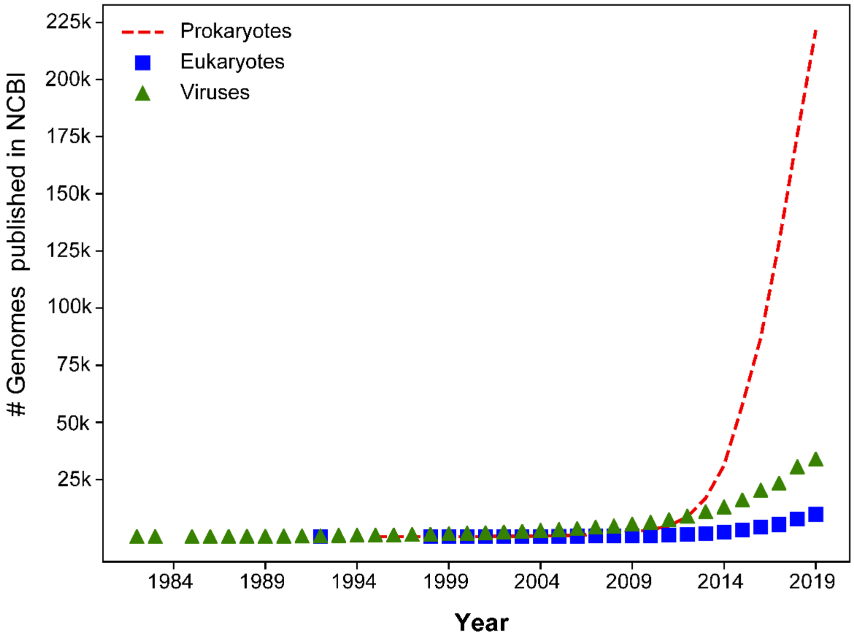


Table 3.1 List of numbers of published genomes for common gut protists up to September 2019.

Protist	# species /subtypes	Available genomes	
		# species/subtypes have genomes	# Total
Blastocystis	17	8	11
Cryptosporidium	10	10	37
Dientamoeba	1	0*	0
Endolimax	2	0	0
Entamoeba	7	5	10
Giardia	9	2	13

* There is only one transcriptome published for *Dientamoeba*.

3.2 IMPLEMENTATION

3.2.1 Overview of the Eukfinder approach

Eukfinder is a taxonomy-classification based workflow to recover microbial eukaryotic genomes from WGS metagenomic sequencing datasets. It can be applied to short or long metagenomic sequencing reads directly, or MAG contigs. First it separates these reads or contigs into different taxonomy groups using Centrifuge and then further refine the unclassified reads or contigs by conducting PLAST searches. This results in candidate eukaryotic reads that can be assembled or eukaryotic contigs that can be extracted and subjected to a series of supervised binning steps to retrieve eukaryotic genomes.

The workflow, which is described below, requires at least one of the following files as input:

1. Paired-end short-read "raw" sequences in FASTQ format. These are processed (see Section 3.2.2) and subjected to iterative taxonomic classification (Figure 3.2(a) and Section 3.2.3).

2. A *de novo* metagenome assembly (MAG) in FASTA format. This is subjected to a simplified taxonomic classification workflow (Figure 3.2(b) and Section 3.2.3).

3. Long-read sequence data in FASTA or FASTQ format (Figure 3.2(b) and Section 3.2.3)

3.2.2 Databases

Two databases must be built before running the Eukfinder workflow: one compatible with the software called Centrifuge (Kim et al., 2016) and the other one with PLAST (Nguyen and Lavenier 2009). The Centrifuge database is built using *centrifuge-download* and *centrifuge-build* commands as implemented in Centrifuge. The PLAST database is built with the BLAST *makeblastdb* command and a simplified index file containing information that cross-references each sequence accession entry in the database to its respective taxonomic group (bacteria, archaea, eukaryote and virus). In order to demonstrate the applicability of Eukfinder, I used the gut microbiome-focused specialized "database 1" and "database 2" (described in Chapter 2.2.1). Specialized database 1 was designed for the taxonomy classifier Centrifuge and contained 32,000 genomes (Supplementary Table S3) that were selected from the NCBI Genbank database and represented common species of gut microbiota. Specialized database 2 was for the more sensitive alignment tool PLAST and included 9,000 representative genomes overlapping with specialized database 1. To carry out supervised binning, I employed the NCBI-NT database.

3.2.3 Short-read pre-processing

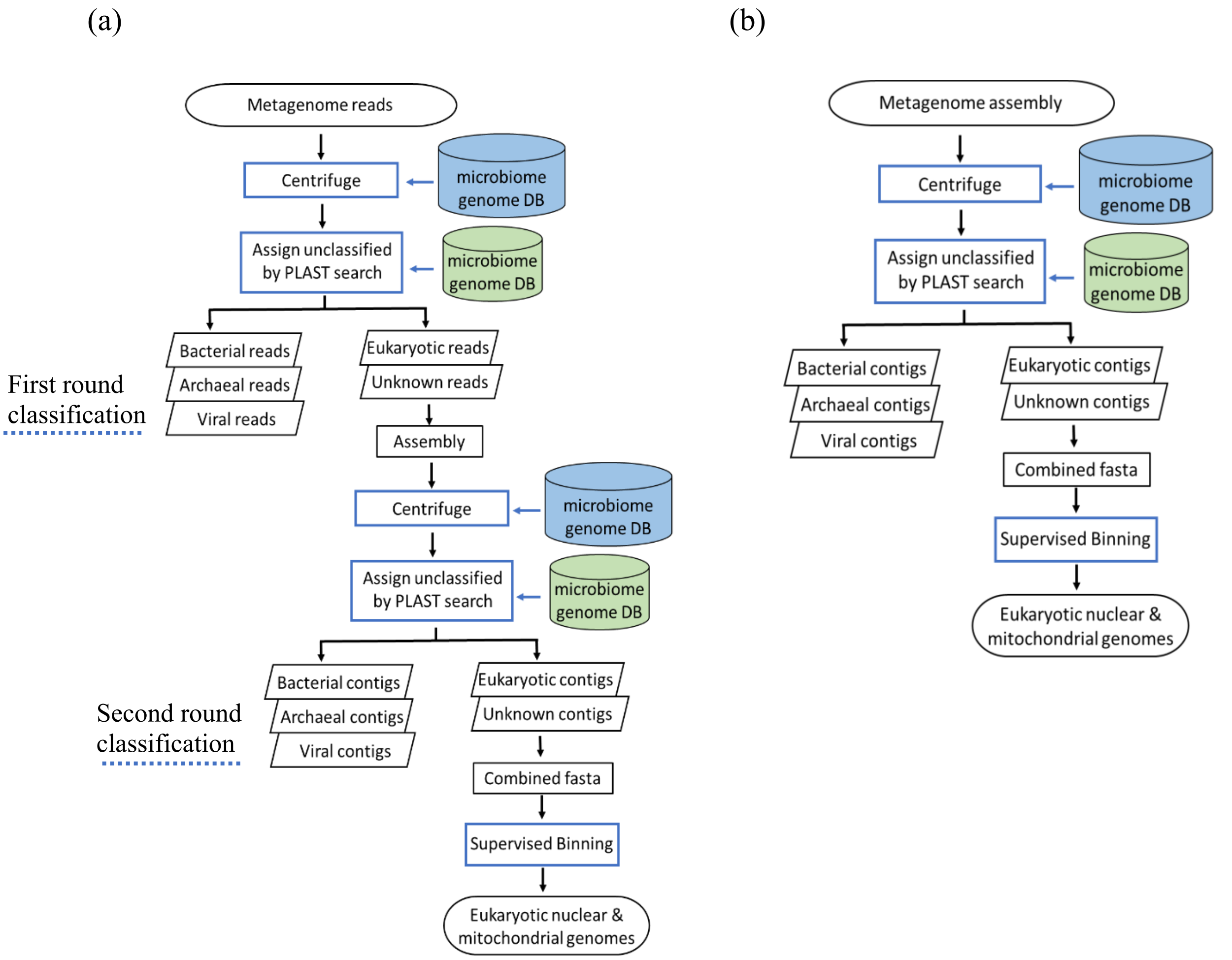
The pre-processing steps include (1) removal of low-quality reads, sequencing adapters (using Trimmomatic v0.36) and host reads (using Bowtie2 v2.3.1) and (2) first round of taxonomic classification using Centrifuge v1.0.4 and the specialized "database 1". For centrifuge, the default minimal hit length is set to 40 bp but the user can modify this parameter. The pre-processing produces five files that are required for downstream analysis: three cleaned FASTQ files (two paired-end and one unpaired-end read files) and two Centrifuge results files (a single file for paired-end reads and a second one for unpaired-end reads).

3.2.4 Eukfinder workflow for short-read sequence files

Eukfinder takes in pre-processed short-read sequences (generally up to 150 bp long; referred to as Eukfinder_reads in this thesis) with their respective taxonomic pre-classification files as mentioned above. As not all the reads are classified at this point, a second attempt to classify them is carried out by PLAST v2.3.2 searches against "specialized database 2". After that, all reads are separated into five groups: Archaea, Bacteria, Eukaryotes, Virus, and Unknown. Reads in Eukaryotes and Unknown groups are used in combination for assembly with SPAdes v3.13.1 (Nurk et al. 2017). The resulting assembly (with minimum contig length of 1000 bp) goes through a new round of taxonomic classification with Centrifuge and PLAST search to separate contigs into the five groups described above (at this point the sequences/contigs are at least ~6.6 to 25 times longer than the original reads, thereby increasing the resolution of the taxonomic searches). Contigs from 'Eukaryotes' and 'Unknown' groups are combined again into one FASTA file for supervised binning as described in Section 3.2.5.

Eukfinder can also accept contigs from MAGs or long-read sequences (length >= 1000bp, referred to as Eukfinder_contigs) and carry out one round of taxonomic classification using Centrifuge and PLAST as described above (Figure 3.2(b)).

Figure 3.2 Schematic representation of Eukfinder workflows. Eukfinder is a taxonomic classification-based bioinformatics approach to retrieve microbial eukaryotic nuclear and mitochondrial genomes from WGS metagenomic sequencing data. Eukfinder has two different workflows based on the input files, (a) Eukfinder_reads using Illumina short reads, or (b) Eukfinder_contigs using MAG assembled contigs or long-read sequencing data generated by Nanopore or Pacbio platforms.



3.2.5 Supervised binning

The assembly from the previous step containing eukaryotic and unknown contigs (i.e., EUnk assembly) is pre-binned with MyCC (Lin & Liao, 2016) using the 4mer, 5mer, and 56mer (a combination of 5mer and 6mer) parameters and additional evidence is collected to assist with the final binning. First, the read-coverage depth for EUnk assembly is calculated by mapping the cleaned short-reads to the contigs with Bowtie2, sorting and indexing them with SAMtools v1.9 (Li et al. 2009), and the script jgi_summarize_bam_contig_depths from MetaBat2. Second, Metaxa2 is used to identify the LSU/SSU rRNA sequences in the EUnk-assembly using its default databases. Third, a nucleotide-based PLAST search is conducted using the contigs as queries against the NCBI-NT database (Jan 2019) and the taxonomy of the best hits' is obtained with acc2tax v0.6 (github.com/richardmleggett). All these results are collected and sorted based on their corresponding MyCC bins. For a contig to be included or excluded in the final eukaryotic bin (or bins), the following rules are applied:

A contig is excluded from eukaryotic bin(s) if:

1) Its depth of coverage exceeds that of the mitochondrial contigs or that of the SSU rRNA gene.

2) Its best PLAST 'hit' shows that it is a sequence from a prokaryote or virus with > 80% identity over an aligned length > 1000 bp.

A contig is kept in the eukaryotic bin(s) if:

1) It hits mitochondrial sequences by Metaxa2 and the best PLAST hit is mitochondrial. These contigs will be marked as mitochondrial genomes.

2) It hits eukaryotic LSU or SSU rRNA as reported by Metaxa2, centrifuge, and /or PLAST.

3) After binning by MyCC, each contig will be assigned to a cluster based on marker genes, the k-mer usage and the depth of coverage. By default, three k-mers (4mer, 5mer, and 56mer) are used and three cluster maps are generated (Supplementary Figure S5). Based on the Centrifuge and PLAST results, some

contigs can be classified as eukaryotic. Some clusters can be marked as potential eukaryotic clusters based on the percentage of the contigs classified as eukaryotes in one cluster. Contigs that appear at least twice in potential eukaryotic clusters are included in the eukaryotic genome.

It is important to mention that, although the supervised binning step is part of the classification workflow, it is not currently implemented in the Eukfinder program due to software incompatibility with the programs required.

3.3 BENCHMARKING METHODS

3.3.1 WGS metagenomic samples

Gut metagenome samples used for retrieving eukaryotic genomes were preselected based on the detection method described in Section 2.2. Human samples with more than 200,000 reads classified as *Blastocystis* based on Centrifuge results (minimal hit length 30) were deemed likely useful for *Blastocystis* genome reconstruction. Six human WGS datasets (SRA accession numbers: ERR321560, ERR636351, ERR636359, ERR636373, ERR636397, and ERR636414) with total sizes ranging from 8.8 giga base pairs (Gbp) to 23.3 Gbp and number of raw reads from 48.9 million (M) to 119.7 M (Table 3.2) were used in this study. The sample ERR321560 was from a Danish individual from a study aiming to characterize metabolic markers from the gut microbiome (Le Chatelier et al. 2013). The remaining five samples were from the gut microbiome datasets of Swedish university students that had traveled to the Indian peninsula or Central Africa (Forsell et al. 2017). All of the datasets were pre-processed as described in section 3.2.2 to generate cleaned short-read files and assembled using MetaSPAdes v3.13.1 (Nurk et al., 2017) to generate MAG assemblies.

3.3.2 Comparison of Eukfinder with existing methods for eukaryotic genome recovery

The performance of recovering nuclear genomes using Eukfinder was compared with those of the machine-learning based software called EukRep (West et al., 2018) and the reference genome mapping method used by Beghini et al., 2017. In the case of mitochondrial genomes, Eukfinder was compared with EukRep, NOVOplasty (Dierckxsens et al. 2017), and the reference genome mapping method.

Since EukRep uses MAG assemblies as input, metaSPAdes assemblies from each dataset were used by EukRep to get eukaryotic contigs with parameter "--tie euk" that treats a contig as eukaryotic when an equal number of sequence chunks were predicted to be of eukaryotic and prokaryotic origin. The resulting eukaryotic contigs underwent supervised binning as described above.

For the reference genome mapping method, metagenomic short reads are extracted from the dataset by mapping them to eukaryotic reference genomes, followed by the assembly of these reads to generate draft eukaryotic genomes. Here, the pre-processed metagenomic reads were mapped to the cleaned reference genome of *Blastocystis* with the same subtype (described in section 2.2.1) using Bowtie2 in local mode. All the mapped reads were assembled using SPAdes (v3.13.1) with default parameters and contigs shorter than 1000 bp were discarded. To explore the mitochondrial genomes with NOVOplasty, a tool for *de novo* assembly of organelle genomes from WGS (meta)genome data, the raw metagenomic reads (as required by NOVOPlasty) were used as input and the corresponding *Blastocystis* mitochondrial SSU rRNA sequence was used as seed. The resulted MRO genomes were BLAST against *Blastocystis* mitochondrial reference genomes for comparison.

3.3.3 Assessment of genome completeness

The evaluation of nuclear genomes recovered from each dataset using Eukfinder, EukRep, and reference-genome mapping methods was performed using BUSCO v3.0.2 (Simão et al. 2015) with eukaryota_odb9 lineage data and Quast v5.0.2 (Gurevich et al. 2013) and compared with cleaned *Blastocystis* genomes that served as references (see section 2.2.1). The shared BUSCO genes among the recovered genomes in each sample were visualized using the Upset Shiny App (Conway et al. 2017). The evaluation of mitochondrial genomes was performed using Quast v5.0.2. If the mitochondrial genome was recovered as one single contig, it was circularized using the overlapping ends and annotated by the online tool Mfannot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl), converted to GenBank format by NCBI software Sequin, and visualized by OGDRAW v1.3.1 (Greiner et al. 2019). The comparison of genome maps between recovered genome fragments and the reference genome was generated using Blast

Ring Image Generator (BRIG)(Alikhan et al. 2011). The sequence coverage BAM files were generated by mapping metagenomic reads to the genome by Bowtie2, sorting and indexing by SAMTools (Li et al., 2009), and visualized in Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al. 2013).

Table 3.2 Sequence features of tested metagenome datasets. *Blastocystis* sequence reads estimated with Centrifuge as described in section 2.2.

Name	Dataset	Size (Gbp)	#Total Reads (M)	MAG size (MB)	Blastocystis subtype
MH0206	ERR321560	8.8	48. 9	263	ST2
TRV13	ERR636373	15.2	77.8	427	ST2
TRV02	ERR636351	11.1	56.7	401	ST3
TRV33	ERR636414	23.3	119.7	401	ST3
TRV06	ERR636359	16.5	84.2	520	ST4
TRV25	ERR636397	13.6	75.3	406	ST4

3.4 RESULTS

3.4.1 Recovered *Blastocystis* genomes by Eukfinder_reads workflow

Six human gut metagenomic datasets were processed by Eukfinder to recover *Blastocystis* nuclear genomes following a benchmarking protocol that allowed me to compare the two methods available within Eukfinder (Figure 3.2) against EukRep and reference-mapping method. This yielded a total of four reconstructed genomes for each dataset (Table 3.3). The cleaned reads from each dataset after pre-processing steps were input into the Eukfinder_reads workflow and after the first round of classification by Centrifuge (Figure 3.2(a)), reads classified as eukaryotic were only a very small proportion, ranging from less than 1% to 4%, while reads that could not be classified at this stage (Unk) ranged from 6% to 17 % (Figure 3.3(a)). After the second round of classification, assembly and binning, the resulting *Blastocystis* draft nuclear genomes recovered were 8 Mbp to 13

Mbp in size corresponding to 60% to 97% complete based on the reference genomes (Table 3.3 Column “Eukfinder_reads”).

Dataset MH0206 was the smallest sequencing file in this study with 512,475 *Blastocystis* reads identified by Centrifuge; its recovered *Blastocystis* ST2 genome was also the smallest in total contig length (8.92 Mbp). The GC content was much markedly lower (51.91%) than the reference genome (54.07%), and the recovered draft genome (3650 contigs with N50 = 2817) was more fragmented than that of the reference genome (969 contigs with N50 = 20102).

The number of reads from dataset TRV13, an ST2 positive dataset, was twice the number of MH0206 and contained more *Blastocystis* reads (888,692). Therefore, the recovered genome (13.36 Mbp) was even larger than the reference genome (12.66). Due to the nature of short metagenomic sequencing, the recovered genome was still more fragmented (1816 contigs and N50=12863bp) than the reference genome. The GC content (53.97%) was much closer to the reference genome (54.07%) than the genome recovered from dataset MH0206.

For the two datasets with *Blastocystis* ST3, the recovered nuclear genomes showed similar trends. The one with more *Blastocystis* reads from dataset TRV33 had a length of 12.27 Mbp, which is 0.68 Mbp larger than the ST3 reference genome and 1.60 Mbp greater than the one from TRV02 (10.67 Mbp). The TRV33 draft genome was less fragmented (1614contigs and N50=13076 bp) than the one from TRV02 (3597 contigs and N50 = 3616). The GC content of the genome recovered from TRV33 (51.93%) was closer to the ST3 reference genome (52.1%) than that of TRV02 (51.51%).

Of the datasets with *Blastocystis* ST4 sequences, the TRV25 dataset had more than 2 million reads classified as *Blastocystis* by Centrifuge. Therefore, the genome recovered from this dataset were more complete (12.26 Mbp) and more similar in GC content (39.82%) to the ST4 JPUL02 reference genome (12.92 Mbp, GC 39.72%) than the genome reconstructed from TRV06 dataset (11.79 Mbp, GC 39.88%). The TRV06 *Blastocystis* genome had 1979 contigs with much shorter contigs (N50 = 9529 bp), while the TRV25 genome had fewer contigs (1106) but a slightly smaller N50 (27653 bp) than the ST4 reference genome (1301 contigs, N50 =29931bp).

3.4.2 Recovered *Blastocystis* genomes by Eukfinder_contigs workflow

The MAGs from six human gut metagenome datasets, size ranging from 260 metabytes (MB) to 520 MB with an average of 265290 contigs, were input into the Eukfinder_contigs workflow (Figure 3.2(b)) and after classification by Centrifuge, about 2.5% to 4.4% of the nucleotides were designated as eukaryotic (Figure 3.3(b)). The proportion of nucleotides without any taxonomy assignment on average was 1.2%. After supervised binning, six *Blastocystis* nuclear genomes were recovered (see Table 3.3 Column "Eukfinder_contigs") and possessed similar features to the ones recovered by Eukfinder_reads.

The *Blastocystis* ST2 nuclear genome recovered from dataset MH0206 MAG using Eukfinder_contigs was the only one that had a smaller size (8.79 Mbp) than the genomes recovered from the same dataset by Eukfinder_reads (8.92 Mbp) (Table 3.3). All the rest of the *Blastocystis* nuclear genomes generated by Eukfinder_contigs had a slightly larger size (0.02 ~0.06 Mbp) and, on average, more contigs than the ones by Eukfinder_reads, with the exception of the genome from TRV06, where Eukfinder_contigs generated a genome with 53 fewer contigs and a larger N50 than the one from Eunfinder_reads. The differences in GC content between the assemblies from Eukfinder_contigs and from Eukfinder_reads were no more than 0.052%. The N50 values from three of the genomes (TRV06, TRV13, and TRV33) recovered from Eukfinder_contigs were larger than the ones from Eukfinder_reads.

3.4.3 Comparing the performance of Eukfinder with EukRep

The MAG from each dataset was also used as input for the machine-learning-based method, EukRep, to recover *Blastocystis* genomes. For each assembly, EukRep will generate a fasta file containing all the eukaryotic contigs. For the six human gut metagenomic datasets used in this study, the total nucleotides in the contigs classified as eukaryotes in each dataset by EukRep ranged from 6% to 9 %, which was 1.5% to 3.5% more than the percentages obtained by Eukfinder_contigs (Figure 3.3 (b)). However, after supervised binning of the output, the resulting *Blastocystis* nuclear genomes (ranging from 8.27 Mbp to 13.07 Mbp, see Table 3.3 Column "EukRep") were smaller in size (0.3 Mbp to 0.9 Mbp) than those obtained from Eukfinder_contigs or Eukfinder_reads. The assembled *Blastocystis* genomes recovered by EukRep had 150 ~ 460 fewer contigs than those from the Eukfinder approaches.

With the exception of dataset MH0206, all the other genomes recovered by EukRep had a larger N50 than those recovered from either of the Eukfinder approaches. One possible reason for this was that most of the contigs that were missed by EukRep were relatively short (between 1,000 bp and 3000 bp; Supplementary Figure S6). The GC contents of the genomes reconstructed by EukRep were very close to those recovered by Eukfinder.

3.4.4 Comparing the performance of Eukfinder with the reference-genome mapping method

The pre-processed metagenomic sequencing files from each human gut metagenome were mapped against the *Blastocystis* reference genomes with the same subtypes to reconstruct draft genomes (referred to as Ref_mapping; see section 3.3.2) and the resulting genomes (see Table 3.3 Column "Ref_mapping") was compared with those recovered by Eukfinder approaches. All six genomes generated this way were smaller than either of genomes retrieved by Eukfinder approaches from the same dataset, in particular, the *Blastocystis* genomes from datasets MH0206 and TRV33 were ~ 1 Mbp less than those recovered by Eukfinder. As was the case for EukRep, genomes with smaller sizes generated by the reference-mapping method also had fewer contigs compared to Eukfinder results, except for TRV06; the latter genome was only 0.2 Mbp smaller but had more contigs (2182) than those retrieved by Eukfinder_reads and Eukfinder_contigs (1979 and 1926 contigs, respectively). The N50 values for Ref_mapping genomes were greater in two datasets, TRV02 and TRV33, and lower in the rest of datasets than those recovered using Eukfinder workflows. The GC contents inferred from the Ref_mapping genomes were within 0.5% of the reference genomes, except for the dataset MH0206 for which the difference was 2%.

3.4.5 The completeness of the *Blastocystis* nuclear genomes

To test the quality and completeness of the recovered *Blastocystis* nuclear genomes, Quast and BUSCO analysis with eukaryotic single-copy genes (SCGs) were applied to the four genomes recovered from each dataset. For dataset MH0206, which had the fewest total reads and generated the smallest *Blastocystis* genome, the genome fractions assessed by Quast all hovered around 60%; the genome recovered by Eukfinder_reads had the largest (63%) and the one from EukRep smallest (58%) (Figure 3.4 (a)). For the other ST2 sample, TRV13, for which most of the recovered genomes were larger than the reference genome, the genome fractions for each of the four nuclear genomes were 95%, with the one from Ref_mapping having the lowest value (94.5%). For *Blastocystis* ST3 genomes from TRV02, although the genome recovered from Ref_mapping did not have the largest size, it had the highest genome fraction (87%), followed by two genomes from Eukfinder (both 85%), and the lowest was the one from EukRep (78%). For the remaining three datasets, the genomes fractions from Eukfinder and Ref_mapping were very similar (> 90%, 93%, and 97% for genomes from TRV06, TRV25, and TRV33, respectively) whereas the ones recovered by EukRep had the lowest completeness (80%, 89%, and 95% for genome from TRV06, TRV25, and TRV33, respectively).

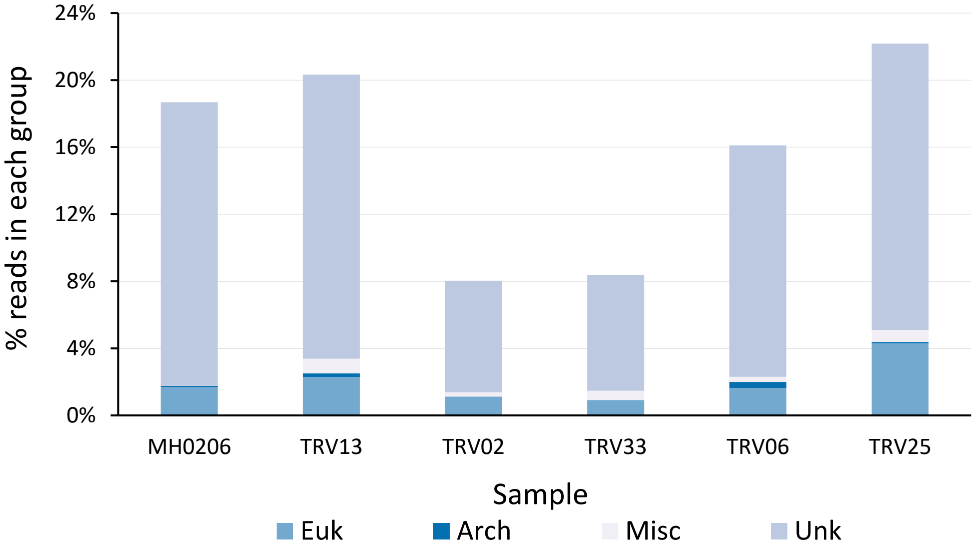
Due to the genome diversity among different *Blastocystis* subtypes, the presence or absence of 303 eukaryotic single-copy genes (SCGs) in the reference genomes of ST2, ST3, and ST4 JPUL02 was identified by BUSCO (Figure 3.4 (b)). This was treated as the baseline to assess the genome completeness for the recovered *Blastocystis* genomes from each sample. *Blastocystis* ST2 nuclear genomes recovered from dataset MH0206 were the least complete compared to those from other datasets. The genome reconstructed from this dataset by Rep_mapping had only 85 SCGs detected compared to the reference genome that had 149. The Eukfinder_contigs retrieved genome had the most SCGs (114/149), followed by the genome from Eukfinder_reads (113/149), and the one from EukRep (110/149). Among the detected single-copy genes, 68 were shared by all four newly recovered genomes and reference genome, while 48 were only found in the reference genome (Figure 3.5 (a)). The *Blastocystis* ST2 genomes recovered from dataset TRV13 were more complete: the genomes generated by Eukfinder_contigs, EukRep and Eukfinder_reads had more SCGs (168, 168, 167 genes respectively) than the reference genome (149 SCGs) and the genome reconstructed by Ref_mapping (149 SCGs) had the same number as the reference genome. Whereas 109 SCGs were shared by all the genomes recovered from TRV13 and the reference genome, less than 34SCGs were shared by all four recovered genomes but not detected in the reference genome while 29 SCGs were only detected in the reference genome(Figure 3.5 (b)).

For *Blastocystis* ST3, genomes recovered from TRV02 by Eukfinder_reads and Eukfinder_contigs and recovered from TRV33 by Eukfinder_reads, Eukfinder_contigs, and EukRep were more complete than the ST3 reference genome and had more SCGs detected (Figure 3.4 (b)). In both datasets, genomes recovered using the Ref_mapping method were the least complete. Of these SCGs, 102 were shared by all these genomes recovered from TRV02 and 124 from TRV33; 20 of the genes were shared by at least three recovered genomes (two by Eukfinder and one by EukRep) from TRV02 and 18 were shared by at least three in TRV33 (Figure 3.5 (c) and (d)).

The *Blastocystis* ST4 reference genome was a complete genome with gene annotation, so it had more detected SCGs than all the newly-recovered genomes from dataset TRV06 and RV25 (Figure 3.5 (e) and (f)). For dataset TRV06, the genome recovered by Eukfinder_reads was the most complete (126 SCGs), followed by Eukfinder_contigs and Ref_mapping (both had 120), and the one by EukRep was the least complete (114 SCGs). In these detected SCGs, 95 were shared by all four genomes and the reference genome. For dataset TRV25, the genome recovered by Eukfinder_contigs was the most complete and had 132 SCGs compared to the 138 of the reference genome. The genome by Eukfinder_reads had 129 SCGs and the ones by EukRep and Rep_mapping had the least (both had 124). About 106 SCGs were shared by all four genomes and the reference genome. SCGs detected only in the reference genome but not in any of the recovered genomes from TRV06 and TRV25 were 10 and 7 respectively, which is less than the corresponding numbers detected for ST2 and ST3 samples.

Figure 3.3 Proportion of reads/nucleotides classified to each group (excluding the bacterial ones) by (a) Eukfinder_reads and (b) Eukfinder_contigs vs. EukRep (Only contigs with >= 1000 bp were included in the calculation) for all datasets. Euk: eukaryotic; Arch: archaeal; Misc: viral; Unk: unclassified.

(a)



(b)

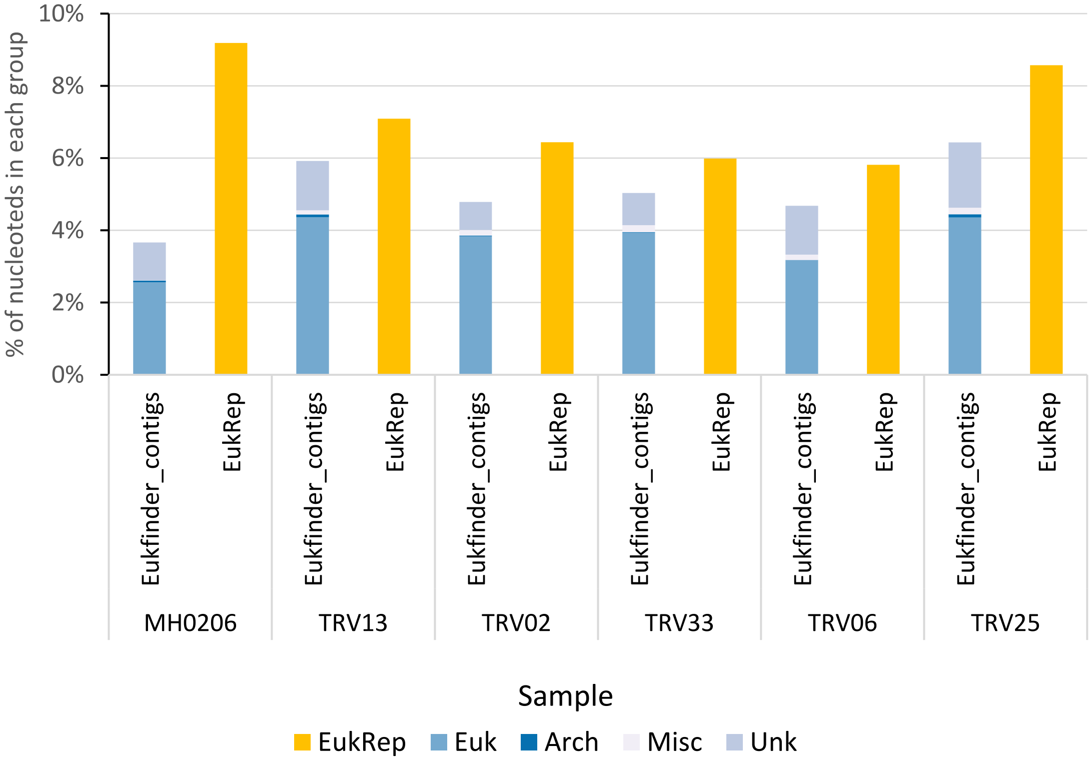


Table 3.3 Summary of genome features for all recovered *Blastocystis* nuclear genomes by four different methods. Features of reference genomes are shaded with light gray, and the repeated numbers are in dark gray. The largest genome in each dataset is in bold.

Genome (*Blastocysti s* reads)	Eukfinder_ reads				Eukfinder_contigs				EukRep				Reference-mapping			
	Size (Mb)	# Contigs	N50	% GC	Size (Mb)	# Contigs	N50	% GC	Size (Mb)	# Contigs	N50	% GC	Size (Mb)	# Contigs	N50	% GC
ST2 Ref genome	12.66	969	20102	54.07	12.66	969	20102	54.07	12.66	969	20102	54.07	12.66	969	20102	54.07
MH0206 (512 K)	**8.92**	3650	2817	51.91	8.79	3714	2720	51.86	8.27	3412	2801	51.81	7.7	3343	2601	52.04
TRV13 (888 K)	13.36	1816	12863	53.97	**13.38**	1822	13077	53.99	13.07	1665	13412	53.99	12.21	1571	11938	54.09
ST3 Ref genome	11.59	909	20816	52.10	11.59	909	20816	52.10	11.59	909	20816	52.10	11.59	909	20816	52.10
TRV02 (415 K)	10.67	3597	3616	51.51	**10.71**	3641	3577	51.52	9.76	3178	3786	51.59	10.12	3190	3890	51.65
TRV33 (819 K)	12.27	1614	13076	51.93	**12.31**	1615	13283	51.88	11.78	1419	13727	52.00	11.37	1257	14119	52.17
ST4 JPUL02 Ref genome	12.92	1301	29931	39.72	12.92	1301	29931	39.72	12.92	1301	29931	39.72	12.92	1301	29931	39.72
TRV06 (587 K)	11.79	1979	9529	39.88	**11.81**	1926	9800	39.89	10.42	1572	10488	39.89	11.59	2182	8039	39.91
TRV25 (2633 K)	12.26	1106	27653	39.82	**12.32**	1120	27601	39.83	11.58	973	28885	39.82	12.08	1090	25423	39.86

Figure 3.4 Genome completeness for recovered *Blastocystis* nuclear genomes compared to the reference genomes by (a) Quast genome fraction and (b) BUSCO single-copy genes detected in genome. The x-axis describes the most likely *Blastocystis* subtype from each sample.

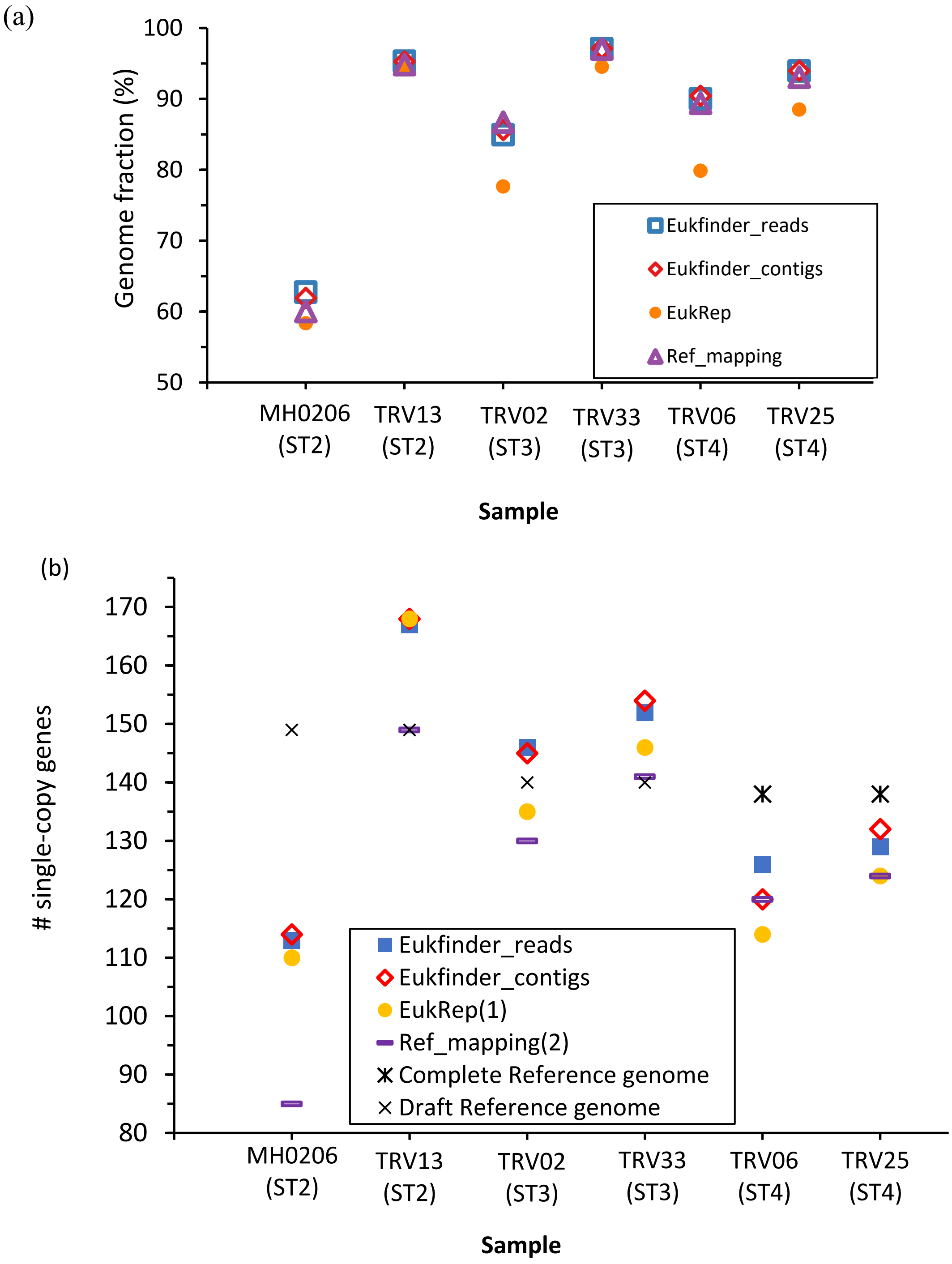
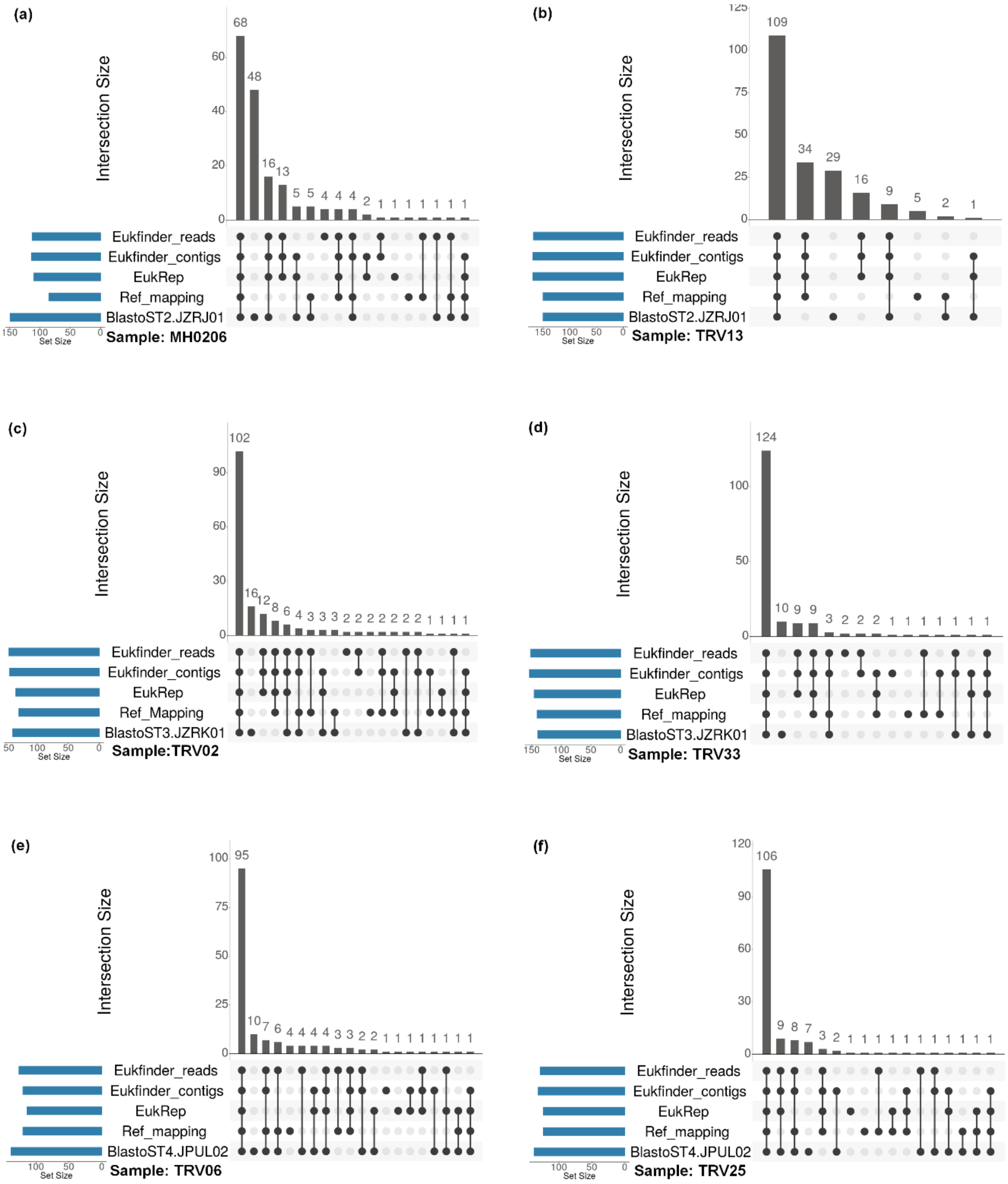


Figure 3.5 Number of BUSCO genes shared in the recovered *Blastocystis* nuclear genomes from six human gut metagenomes and the reference genomes. The dark gray vertical bar represents the number of shared genes by intersections, the blue horizontal bar shows the total number of genes detected in each genome.



3.4.6 *Blastocystis* MRO genomes

To benchmark the performance of the Eukfinder workflows on recovering organelle genomes, *Blastocystis* MRO genomes reconstructed by using either reads or contigs as input for Eukfinder were compared to genomes generated by EukRep, reference genome mapping, and NOVOPlasty. The genome length, number of contigs and GC content for each recovered MRO genome were listed in Table 3.4. The only dataset with incomplete MRO genomes was MH0206, for which Eukfinder, EukRep, and reference-mapping method yielded genomes with 7 contigs and around 11 Kbp, and NOVOPlasty recovered only a short contig with 787 bases. These recovered fragments included regions encoding SSU and LSU rRNAs, the proteins nad1, nad4, nad5, nad7, rpl16, and a region containing two genes and several tRNAs (Supplementary Figure S7). *Blastocystis* MRO genomes recovered from the remaining five samples by each method tested were complete with one contig, except EukRep did not recover the MRO genome for dataset TRV33 and NOVOPlasty that generated a genome with two contigs for TRV33. The MRO genome from TRV13 had seven nucleotides more than the reference genome, the genomes from datasets TRV02 and TRV33 both had a size very close to the ST3 DMP/08-326 MRO genome. For ST4 samples, the MRO genome recovered from TRV06 had only two more nucleotides than the DMP/02-328 reference genome, while the one from TRV25 had 11 more nucleotides than the reference genome.

All the MRO genomes reconstructed by Eukfinder, EukRep, and reference-mapping methods were manually circularized. NOVOPlasty generated circularized organelle genomes for TRV02, TRV06, TRV13, and TRV25. For TRV33, a complete genome was manually circularized from the two contigs generated by NOVOPlasty after removing a repetitive region. The recovered MRO genomes were aligned to reference genomes by BLAST and four circularized complete genomes from each dataset of TRV02, TRV13, TRV25, and TRV33 were identical to each other so the results were combined into one line for each dataset. For sample TRV06, NOVOPlasty generated an MRO genome lacking 11 nucleotides (possibly due to mapping or assembly mistakes; Supplementary Figure S8) compared to the ones recovered by other methods. This error-containing MRO genome was not used for comparison with the reference genomes.

Annotation of all of the draft MRO genomes showed conservation in gene content and synteny (Supplementary Figure S9).

When compared to the reference MRO genomes, these retrieved MRO genomes shared high identities and few gaps/mismatches to most reference genomes. The recovered partial MRO genomes from MH0206 aligned to the ST2 reference genomes with identities of 99% (Table 3.5) and 6 gaps, except for NOVOPlasty that recovered a contig with only 1 gap. Of these gaps, three were located in the LSU/SSU rRNA coding region, one in tRNA and two in protein-coding regions. For the TRV13 MRO genome, it was 99% identical to the reference genome with 158 mismatches. It has thirteen gaps relative to the reference with three occurring in the LSU/SSU rRNA coding region, six in protein-coding regions and four in intergenic regions. Because there are three reference genomes for both *Blastocystis* ST3 and ST4, the resulting genomes from the datasets with the corresponding subtype were compared to all the reference genomes. The recovered ST3 MRO genomes from TRV02 and TRV33 were 99% identical to DMP/08-326 and DMP/IH478 reference genomes but only 96% identical to DMP/08-1043 reference genome (Table 3.6). There were fewer than five gaps between these ST3 MRO genomes and the reference genomes of DMP/08-326 or DMP/IH478, with at least half of them occurring in intergenic regions.

Similarly, the ST4 mitochondrial genomes from TRV06 and TRV25 were 99% identical to the DMP/02-328 and WR1 reference genomes, respectively, but shared only 88% identity to the DMP/10-212 reference genome with about 600 gaps (Table 3.7). There were only two gaps and four mismatches found between MRO genome recovered from the TRV06 dataset and the WR1 reference genome, with two gaps and one mismatch located in the LSU rRNA coding region and three mismatches in the protein-coding regions (Table 3.8). But TRV06 MRO genome had a difference of two gaps and eight mismatches compared to the ST4 reference genome from strain DMP/02-328. Although sharing the same percentage of identity, the TRV25 MRO genome had more gaps when aligned with the DMP/02-328 and WR1 reference genomes; there were 14 and 16 gaps, respectively, with most of them (>10) in intergenic regions.

Table 3.4 Summary of genome features for all recovered *Blastocystis* MRO genomes by five different methods. Features of reference genomes are shaded with light gray, and the repeated numbers are in dark gray. The genomes with the least bases or no base were highlighted in yellow.

MRO Genome (*Blastocystis* reads)	Eukfinder_reads			Eukfinder_contigs			EukRep			Reference-mapping			NOVOPlasty		
	Length (bp)	Contigs	% GC	Length (bp)	Contigs	% GC	Length (bp)	Contigs	% GC	Length (bp)	Contigs	% GC	Length (bp)	Contigs	% GC
ST2 Ref genome	28,305	1	19.7	28,305	1	19.7	28,305	1	19.7	28,305	1	19.7	28,305	1	19.7
MH0206 (512 K)	11,030	7	26.5	11,117	7	26.7	11,117	7	26.7	11,203	7	26.6	787	1	27.1
TRV13 (888 K)	28,312	1	19.7	28,312	1	19.7	28,312	1	19.7	28,312	1	19.7	28,312	1	19.7
ST3 DMP /08-326 Ref genome	28,242	1	21.6	28,242	1	21.6	28,242	1	21.6	28,242	1	21.6	28,242	1	21.6
TRV02 (415 K)	28,245	1	21.5	28,245	1	21.5	28,245	1	21.5	28,245	1	21.5	28,245	1	21.5
TRV33 (819 K)	28,240	1	21.6	28,240	1	21.6	0	0	0	28,240	1	21.6	28,240	2	21.6
ST4 DMP /02-328 Ref genome	27,717	1	21.9	27,717	1	21.9	27,717	1	21.9	27,717	1	21.9	27,717	1	21.9
TRV06 (587 K)	27,719	1	21.9	27,719	1	21.9	27,719	1	21.9	27,719	1	21.9	27,708	1	21.9
TRV25 (2633 K)	27,728	1	21.8	27,728	1	21.8	27,728	1	21.8	27,728	1	21.8	27,728	1	21.8

Table 3.5 Comparison of *Blastocystis* ST2 MRO genomes recovered by all the methods used in this study with the reference genome. Results from TRV13 were the same for all methods and are, therefore, only shown in a single line.

Sequence	Size (bp)	Comparison to reference genome				Location of gaps/mismatches in the genome				
		Aligned length	% Identity	# Mismatch	# Gaps	SSU rRNA	LSU rRNA	tRNA	Protein coding	Intergenic
ST2 Flemming	28,305	10,967	99%	57	6	2	2		2	1
MH0206 Eukfinder_reads	11,030	11,054	99%	57	6	2	2	-	2	1
MH0206 Eukfinder_contigs	11,117	11,054	99%	57	6	2	2	-	2	1
MH0206 EukRep	11,117	11,161	99%	36	6	2	2	-	2	1
MH0206 Ref_mapping	11,203	786	99%	0	1	-	-	-	-	1
MH0206 NOVOPlasty	787	28,141	99%	158	13	2	1	-	6	4
TRV13	28,306	10,967	99%	57	6	2	2	-	2	1

Table 3.6 Comparison of *Blastocystis* ST3 MRO genomes recovered in this study with the reference genomes. Genomes recovered by different methods had the same sequence for ST3 samples, except EukRep did not recover an MRO genome for TRV33.

Sequence	Size (bp)	Comparison to reference genome				Location of gaps/mismatches in the genome				
		Aligned length	% Identity	# Mismatch	# Gaps	SSU rRNA	LSU rRNA	tRNA	Protein coding	Intergenic
ST3 DMP/08-326	28,242	-	-	-	-	-	-	-	-	-
TRV02	28,245	28,206	99%	36	3	-	-	-	-	3
TRV33	28,240	28147	99%	92	4	-	-	-	1	3
ST3 DMP/IH478	28242	-	-	-	-	-	-	-	-	-
TRV02	28,245	28200	99%	40	5	-	-	-	1	4
TRV33	28,240	28154	99%	82	4	-	-	2	-	2
ST3 DMP/08-1043	28,268	-	-	-	-	-	-	-	-	-
TRV02	28,245	27166	96%	1057	67	-	-	-	-	-
TRV33	28,240	27153	96%	1066	70	-	-	-	-	-

Table 3.7 Comparison of *Blastocystis* ST4 MRO genomes recovered in this study with the reference genomes. MSH: mismatch

Sequence	Size (bp)	Comparison to reference genome				Location of gaps/mismatches in the genome				
		Aligned length	% Identity	# Mismatch	# Gaps	SSU rRNA	LSU rRNA	tRNA	Protein coding	Intergenic
ST4 WR1	27,717	-	-	-	-	-	-	-	-	-
TRV06	27,719	27,713	99%	4	2		2 gaps 1 MSH		3 MSH	
TRV25	27,728	27,629	99%	83	16	1 gap	3 gaps	1 gap		10 gaps
ST4 DMP/02-328	27,719	-	-	-	-	-	-	-	-	-
TRV06	27,719	27,710	99%	8	2	2 MSH	1 MSH	1 MSH	1 gap 4 MSH	1 gap
TRV25	27,728	27627	99%	87	14	1 gap	1 gap	1 gap	1 gap	10 gaps
ST4 DMP/10-212	27,817	-	-	-	-	-	-	-	-	-
TRV06	27,719	24616	88%	2836	597	-	-	-	-	-
TRV25	27,728	24639	88%	2811	605	-	-	-	-	-

3.5 DISCUSSION

Eukfinder is a taxonomy-classification based workflow for microbial eukaryotic genome recovery from environmental metagenomes that was developed and applied to six human gut metagenomic datasets. For five of these datasets, near-complete (>= 85% completeness) *Blastocystis* nuclear genomes were generated. The smallest nuclear genome recovered by Eukfinder was an ST2 genome with 8.8 Mbp (60% completeness) and the largest genome was also an ST2 genome with 13.4 Mbp, which is larger than the reference *Blastocystis* ST2 genome. Two ST3 genomes and two ST4 genomes were also recovered with 85% to 97% completeness. The recovered genomes with >90% completeness also had a small difference (< 0.1%) in GC contents when compared to reference genomes. Due to the nature of short-read sequencing and relatively low fold-coverage, the recovered genomes tended to be more fragmented than the reference genomes. For the metagenomic samples with a higher number of eukaryotic reads (e.g. sample TRV25), the newly assembled genomes had numbers of contigs and large N50 values that are comparable to the reference genomes.

Blastocystis nuclear genomes recovered by Eukfinder using WGS reads or MAG contigs as input were benchmarked with the genomes reconstructed using EukRep or reference genome mapping methods. Genome completeness assessment showed that genomes recovered by Eukfinder were more generally complete than those generated by EukRep or the reference-mapping method. In some of the datasets (TRV13 and TRV33), genomes reconstructed using Eukfinder had larger sizes and more essential single-copy genes than the corresponding reference genome. Recovering larger genomes with better genome completion than the currently available reference genomes highlights the potential benefits of using Eukfinder for robust genome reconstruction, as the *Blastocystis* ST2 and ST3 reference genomes are rough drafts known to be incomplete. Using the specialized databases containing as many representative genomes as possible, Eukfinder pre-selects the reads from the metagenomic sequencing data or contigs from MAGs based on taxonomy classification and retains all the possible eukaryotic-origin reads and contigs along with the those that cannot be classified so far to maximize the yields for recovering eukaryotic genomes. It does not require a direct reference genome to do the alignment so it has the potential to recover genomes from organisms without closely related reference genomes, making it more sensitive than reference mapping methods (Beghini et al, 2017) in some situations (e.g., in recovering genomes for *Blastocystis* ST5). My results also showed that the genomes recovered by Eukfinder can generate more complete genomes than those recovered by EukRep (West et al., 2018), which uses a machine-learning approach to identify contigs from eukaryotes based on k-mer usage. Missing representative genomes in their training reference set of genomes for k-mer frequency analysis and/or the use of only 5-mers may be potential reasons why EukRep fails to capture some eukaryotic contigs found in this analyses. The taxonomic classification of contigs based on alignment of homologous genes in the databases (that are not necessarily identical) likely improve the chance of these contigs being included in the eukaryotic group by Eukfinder.

One surprising difference between the *Blastocystis* genomes reconstructed in this study by reference-mapping and the genome recovered by using reference-mapping method of Beghini et al. (2017) comes from analyses of the sample MH0206. They mapped the MH0206 metagenomic sequencing file to their cleaned ST2 genome (11.45 Mbp with 854 contigs) and obtained a genome with 10.13 Mbp (1671 contigs and N50=8777 bp. Table 3.8). Their recovered genome was more complete than the one generated in this study (7.70 Mbp, 3343 contigs and N50= 2601 bp) by reference mapping, although the reference genome was larger (12.66 Mbp in size). Besides the differences in the pre-processing (adapter trimming, quality control trimming, and host reads removal), the Bowtie2 alignment mode was different between the two studies: Beghini and colleagues used end-to-end alignment, whereas local alignment was used in this study. This marked difference in performance between these two alignment modes in this (and potentially other) cases should be carefully investigated in further studies. Meanwhile, this serves as a warning that the use of different parameter settings in component software tools of these pipelines can lead to very different results, so caution is warranted in interpreting the results obtained in this study.

Blastocystis MRO genomes were also recovered by Eukfinder from six metagenomic datasets. Five of these genomes were complete with one circularized contig. Although many tools that can reconstruct organelle genomes from metagenomic datasets like NOVOPlasty, the benchmark results show that Eukfinder can efficiently recover near-complete nuclear and complete mitochondrial genomes at the same time for microbial eukaryotes whereas for some datasets NOVOPlasty or EukRep failed to recover partial or complete mitochondrial genomes.

The sizes of the recovered nuclear genomes are related to the number of eukaryotic reads that can be detected from the metagenomic dataset. For each subtype that has genomes recovered in this study (ST2, ST3, ST4), there were two sets of genomes recovered from two datasets. The ones with more *Blastocystis* reads (TRV13, TRV33, and TRV25) always yielded a larger and more complete genome, with GC content more similar to the reference genome, relative to samples with fewer reads (e.g., MH0206, TRV02, and TRV06). To get near-complete genome recovery, a 90% breadth of coverage with at least X=1 (total number of bps mapped to reference genome divided by the genome size of the target organism) or at least 400,000 reads that can be classified as originated from the organism by Centrifuge is recommended, which is the choosen standard I used for recovering genomes from *Blastocystis*. For other organisms, due to differences in genome size, GC content, the minimum numbers of reads or breadth of coverage recommended for near-complete genome recovery varies. The total numbers of eukaryotic reads in the metagenomic sequencing datasets depend on the sequencing depth, species diversity, and the number of eukaryotes in the sample. Observations from the application of Eukfinder to animal samples (data not shown) indicate that samples with only a few (<5) eukaryotes and one or no unknown eukaryotic species are ideal candidates to recover genomes. With the continued decrease in sequencing cost, metagenomic sequencing with deeper sequencing depth is possible allowing recovery of more microbial eukaryotic genomes.

In general, Eukfinder performs well in recovering microbial eukaryotic genomes from human gut metagenome datasets. This culture-free genome reconstruction tool can be very useful for recovering genomes from difficult-to-culture eukaryotic microbes. This method is more time-efficient than culture-based isolation and genome sequencing. It only requires DNA extraction from environmental samples and WGS metagenomic sequencing. Applying Eukfinder to the large amount of published human/animal gut metagenomes currently existing offers potential to rapidly and efficiently reconstruct genomes of microbial eukaryotes that colonize the GI tract, even for species or genera whose genomes have not been previously characterized.

Eukfinder can also be used as a decontamination tool for *de novo* genome assemblies of genomic data obtained from non-axenic cultures of protists. Many microbial eukaryotes, like *Blastocystis*, live in an environment where they closely interact with bacteria. It can be very hard to eliminate all the bacterial species before DNA extraction and virtually impossible to remove all bacterial DNA afterwards. In these cases, Eukfinder can be used to clean the genome assembly from contaminating sequences from bacteria or other eukaryotes. Indeed, it has been used for this purpose in genome assemblies produced in the Roger Lab and proved to be an efficient tool (data not shown). With the cost decrease in metagenomic sequencing using the Nanopore or PacBio platforms, long-read sequencing has started to be employed in metagenomic sequencing for environmental samples like the human gut or wastewater treatment samples (Suzuki et al. 2019; Che et al. 2019). These long-read metagenomic datasets can be directly analyzed with the Eukfinder_ contigs workflow to facilitate the reconstruction of genomes from microbial eukaryotes. However, for this to be maximally effective, the two databases employed by Eukfinder should be populated with as many previously characterized genomes from these environments as possible.

Overall, with the increase in number and size in publicly available metagenomes, the bioinformatic workflow in Eukfinder can be applied to diverse metagenomic samples to retrieve high-quality microbial eukaryotic genomes. This will increase the numbers of reference genomes available to aid future metagenomic investigations into the functions, physiologies, and evolutionary histories of eukaryotic microbes in the gut microbiome and a variety of other ecosystems.

Table 3.8 Differences of genome features for *Blastocystis* ST2 reference genomes and genomes recovered from MH0206 dataset between Beghini et al. (2017) and this study.

Source	**Genome**	**Total size (Mbp)**	**# contigs**	**N50**	**GC content**
Beghini et al. (2017)	cleaned ST2 Reference genome	11.45	854	20,462	53.98
	MH0206	10.13	1671	8,777	54.07
This study	cleaned ST2 Reference genome	12.66	967	20,102	54.2
	MH0206	7.7	3343	2,601	52.04

CHAPTER 4 CONCLUSIONS

WGS metagenome-based bioinformatic workflows were developed to investigate the prevalence of the gut protist *Blastocystis* in human and animal samples and to retrieve genomes of microbial eukaryotes from environmental metagenomic sequencing data, e.g., *Blastocystis* from human gut metagenomes. The detection workflow was applied to 996 human and animal gut metagenome samples. *Blastocystis* was highly prevalent in non-industrialized human populations with a specific subtype distribution and a high rate of co-infections. Amongst animals, *Blastocystis* had higher colonization frequencies (>50%) in baboons and pigs *versus* chickens and cattle. The compositional and functional changes in the human *Blastocystis* carriers compared to non-carriers were also analyzed. Both westernization and subtypes can affect the gut microbiota species composition and abundance of their metabolic pathways. Finally, the genome recovery tool, Eukfinder, was applied to six human gut metagenomic datasets carrying *Blastocystis* and retrieved six near-complete nuclear genomes and five full-length MRO genomes. From all of these analyses a number of interesting novel findings were mad that are discussed in more detail below.

The presence of multiple *Blastocystis* subtypes (co-infection) in gut samples is often underestimated, if not completely neglected (Maloney et al. 2019; Betts et al. 2020). The detection workflow developed in this study was able to detect co-infections and found an average of 20% and 10% co-infection rates in human and animal positive samples respectively, consistent with previous analyses of co-infection rates (Scanlan et al. 2015; Betts et al. 2020). Of all cohorts analyzed, the most co-infection occurred in the Tanzanian datasets, accounting for ¾ of all positive in these samples. The numbers of each mixed type were too few in this study to be permit statistically robust analyses of their impact. The workflows described here will be useful for future association analyses of much larger metagenome datasets to distinguish the impact of specific mixed *Blastocystis* subtypes co-infections on the gut microbiome compared to single-ST infections.

This was the first study that has investigated *Blastocystis* in animal gut metagenomic samples and the methods developed herein can be applied to other types of animals. Besides the epidemiological information on prevalence of *Blastocystis* subtypes and co-infections, the detection workflow can also obtain full-length SSU rRNA gene sequences for phylogenetic analysis. If a sample contains enough *Blastocystis* reads and fewer than three other microbial eukaryotes lacking available reference genomes, the Eukfinder workflow can be used to retrieve more genomes of *Blastocystis* STs even if they lack available reference genomes. The detection workflow also has the potential to be used for to other microbial eukaryotes, e.g., *Giardia* and *Entamoeba*, in gut metagenomic samples.

However, many components of the detection workflow may need to be updated/expanded before it can be broadly applied to larger numbers of datasets and different organisms. First of all, the quality of the databases still needs to improve since some of the eukaryotic reference genomes contain a large number of contaminating bacteria reads that can cause taxonomy assignment mistakes for some species (Steinegger & Salzberg 2020). Second, the cutoff value for defining a *Blastocystis*-positive sample in human samples needs to be verified using mock community sequencing data with a known number of *Blastocystis* reads to ensure its sensitivity and accuracy. Some of *Blastocystis* positive rates detected by this study differed considerably from the results of Lokmer et al. (2019) both in the overall prevalence and in the numbers of different subtypes detected. This large difference occurred for one of the three projects analyzed by both studies and suggests that applying a universal threshold value for designation of positives may be problematic considering difference in the sequencing depths amongst different projects. Lastly, and most importantly, when expanding the workflow to other protists, ‘positive’ threshold cutoff values should be investigated and optimized, as each protistan species (or genus) having different numbers of available reference genomes and degrees of sequence and genome content difference amongst species or strains.

The difference I found in gut microbiota composition and functional profiles in *Blastocystis* carrier samples compared to non-carriers revealed potential interactions between *Blastocystis* and gut prokaryotes. Archaeal species like *Methanobrevibacter smithii* were significantly associated with presence of *Blastocystis*, especially in non-westernized samples. This association is worth further investigation because of *M. smithii*’s important role in hydrogen consumption and methanogenesis. It remains to be determined if the associations between specific gut bacteria and presence of *Blastocystis* in western carriers are really significant since most of these groups I investigated contained only about 10 samples. A larger scale comparative analysis between non-western and western carriers is needed to investigate this further.

Recovering genomes of microbial eukaryotes using a metagenomic approach is extremely challenging in comparison to recovery of prokaryotic genomes from these kinds of data. Eukfinder is an attempt to combine state-of-the-art software tools to generate draft eukaryotic genomes of reasonable quality with culture-independent methods. The full-length *Blastocystis* MRO genomes produced in this study can be used to build phylogenetic trees with highly conserved genes (e.g., the *nad* gene, Jacob et al., 2016) and investigate the phylogenetic relationships amongst different strains of the same ST. The near-complete *Blastocystis* nuclear genomes can be used for gene prediction (although this will be more challenging without (meta-)transcriptome sequencing data). If shared genes can be found in genomes from different strains from the same or different STs, there is a potential to reveal genomic diversity and pathogenicity determinants using phylogenomic analysis. Finally, the accumulation of more genome data from the foregoing analyses can help build better databases and aid in better detection of *Blastocystis* subtypes with no reference genomes and further retrieval of more genomes.

In summary, the WGS metagenomic workflows developed in this thesis may prove useful in studying the prevalence and genomic diversity of *Blastocystis* and other protists from environmental metagenome sequencing data. Preliminary results show that the workflows are sensitive and effective, but this has to be confirmed by further mock community analyses and test on much larger datasets. As we continue to improve these workflows, it is possible to develop them into an easy-to-install and easy-to-use software tools with capability to automatically handle metagenomic data and generate prevalence reports and candidate eukaryotic genome assemblies.

APPENDIX A – SUPPLEMENTARY TABLES

Table S1. The host reference genomes downloaded from NCBI.

Host	Genome accession number
Human	GCF_000001405.37
Baboon	GCF_000264685.3
Cattle	GCF_000003055.6
Chicken	GCF_000002315.4
Pig	GCF_000003025.6
Bacteriophage phiX174	NC_001422.1

Table S2. Contigs removed from *Blastocystis* reference genomes. MRO genomes were labelled with a asterisks.

ST2	JZRJ01000088 *	ST6	JZRM01000240
ST2	JZRJ01000923	ST6	JZRM01000270
ST3	JZRK01000047 *	ST6	JZRM01000285
ST3	JZRK010000726	ST6	JZRM01000317
ST3	JZRK01000382	ST6	JZRM01000353
ST3	JZRK01000623	ST6	JZRM01000399
ST3	JZRK01000726	ST6	JZRM01000403
ST3	JZRK01000754	ST6	JZRM01000411
ST3	JZRK01000820	ST6	JZRM01000419
ST3	JZRK01000826	ST6	JZRM01000431
ST3	JZRK01000839	ST6	JZRM01000456
ST6	JZRM01000006	ST6	JZRM01000464
ST6	JZRM01000011	ST6	JZRM01000480
ST6	JZRM01000013	ST6	JZRM01000507
ST6	JZRM01000016	ST6	JZRM01000535
ST6	JZRM01000019	ST6	JZRM01000542
ST6	JZRM01000022	ST6	JZRM01000570
ST6	JZRM01000023	ST6	JZRM01000598
ST6	JZRM01000027	ST6	JZRM01000617
ST6	JZRM01000029	ST6	JZRM01000659
ST6	JZRM01000030	ST6	JZRM01000755
ST6	JZRM01000035 *	ST6	JZRM01000790
ST6	JZRM01000036	ST6	JZRM01000826
ST6	JZRM01000052	ST6	JZRM01000830
ST6	JZRM01000058	ST6	JZRM01000879
ST6	JZRM01000081	ST8	JZRN01000022 *
ST6	JZRM01000084	ST8	JZRN01000233
ST6	JZRM01000095	ST8	JZRN01000747
ST6	JZRM01000101	ST8	JZRN01000879
ST6	JZRM01000108	ST9	JZRO01000015 *
ST6	JZRM01000117	ST9	JZRO01000142
ST6	JZRM01000122	ST9	JZRO01000228
ST6	JZRM01000123	ST9	JZRO01000234
ST6	JZRM01000137	ST9	JZRO01000235
ST6	JZRM01000150	ST9	JZRO01000417
ST6	JZRM01000163	ST9	JZRO01000444
ST6	JZRM01000189	ST9	JZRO01000788
ST6	JZRM01000198	ST9	JZRO01000789
ST6	JZRM01000214	ST9	JZRO01000859
ST6	JZRM01000230	ST9	JZRO01000871

Table S3. Numbers of genomes in each group in the specialized databases.

Group	# genomes in centrifuge DB "database 1"	# genomes in PLAST DB "database 2"
Archaea	3,662	211
Bacteria	10,488	747
Eukaryotes	1,036	126
EukPathDB	244	61
Mitochondria	10,830	111
Virus	6,136	8,100
Total	**32,402**	**9,345**

Table S4. Prevalence for each *Blastocystis* subtype in human samples.

	# samples with *Blastocysits*								# Total samples
Projects	ST1	ST2	ST3	ST4	ST7	ST8	Mix	All	
H1_Cameroon	14	4	20				10	48	57
H2_Ethiopia	13	5	5				13	36	50
H3_IDN, Liberia	9	2	6					17	24
H4_Madagascar	23	10	19				4	56	111
H5_Peru	8	12	4				6	30	36
H5_USA	1		2	2		1		6	22
H6_Sweden	5	2	8	6		1		22	35
H7 _Sweden		1		1				2	21
H8_Tanzania	2	3					18	23	27
H8_Italy			1					1	11
H9_USA								0	36
H10_USA	3	4	6		1			14	55
Continents									
Africa	54	24	44	0			45	167	249
Asia	7	0	6	0				13	20
Europe	5	3	9	7		1		25	67
N America	4	4	8	2	1	1		20	113
S America	8	12	4				6	30	36

Table S5. Prevalence for each *Blastocystis* subtype in animal samples.

	# samples with *Blastocysits*										# Total samples
Projects	ST1	ST3	ST5	ST6	ST7	ST10	ST13	ST15	Mix	All	
A1_Baboon Kenya	11	13								24	48
A2_Cattle China			1			3				4	30
A3_Cattle France										0	25
A4_Cattle Italy										0	16
A5_Cattle USA	5	1							2	8	29
A6_Chicken China										0	20
A7_CCP China			1	1	2	1		2	1	8	13
A8 _Pig CDF	4	6	135					34	25	204	216
A9_Pig China			4					2		6	8
A10_Pig Denmark	2	4	14					9		29	35
A11_Pig Japan, Gabon			2				1	1	2	6	6
A12_Pig German			1					1		2	22
A13_Pig Spain	1		4					2		7	8
Animals											
Baboons	11	13								24	48
Cattle	5	1	1			4			3	14	104
Chickens				1	2					3	24
Pigs	7	10	161				1	49	29	255	300

Table S6. Bacterial species abundances which differed in *Blastocystis* carriers (positive) and non-carriers (negative). Between group differences were evaluated with two-tailed Wilch's *t*-tests with Storey's FDR corrections(FDR<0.05). Rows shaded in yellow represent species enriched in *Blastocystis* carrier group, while rows shaded in blue represent those enriched in non-carriers.

Phylum	Species	Proportion of sequences	
		Positive Mean ± std. dev (%)	Negative Mean ± std. dev (%)
Firmicutes	*Butyrivibrio crossotus*	1.37 ± 3.10	0.03 ± 0.24
Firmicutes	*Eubacterium eligens*	1.87 ± 2.96	0.54 ± 1.14
Firmicutes	*Faecalibacterium prausnitzii*	15.72 ± 11.46	9.57 ± 9.99
Firmicutes	*Phascolarctobacterium succinatutens*	3.60 ± 5.98	0.86 ± 2.09
Bacteroidetes	*Prevotella copri*	12.76 ± 13.99	7.47 ± 15.03
Firmicutes	*Ruminococcus champanellensis*	0.49 ± 1.74	0.00 ± 0.01
Spirochaetes	*Treponema succinifaciens*	5.04 ± 11.44	0.06 ± 0.51
Firmicutes	*Eubacterium biforme*	1.15 ± 2.66	0.32 ± 0.73
Euryarchaeota	*Methanobrevibacter smithii*	1.60 ± 4.29	0.50 ± 1.17
Euryarchaeota	Unclassified *Methanobrevibacter*	0.33 ± 0.75	0.02 ± 0.07
Firmicutes	Unclassified *Oscillibacter*	0.08 ± 0.22	0.30 ± 0.56
Bacteroidetes	*Alistipes putredinis*	0.86 ± 2.14	2.05 ± 3.04
Bacteroidetes	*Bacteroides uniformis*	0.68 ± 1.65	3.72 ± 6.25
Firmicutes	*Dialister invisus*	0.21 ± 1.03	1.99 ± 5.32
Bacteroidetes	*Parabacteroides distasonis*	0.11 ± 0.39	0.40 ± 0.77
Firmicutes	*Ruminococcus* sp. 5_1_39BFAA	0.30 ± 0.54	1.44 ± 2.14
Firmicutes	*Ruminococcus torques*	1.00 ± 1.44	2.48 ± 2.77

Table S7. Species abundances which differed in *Blastocystis* carriers (positive) and non-carriers (negative) with regarding to non-westernized (NonW) or westernized (W) individuals. Between group differences were evaluated with Kruskal-Wallis tests with Benjamini-Hochberg FDR corrections (FDR<0.05). Pos: positive, Neg: Negative.

Phylum	Species	Mean of proportion of sequences (%)			
		NonW _Pos	NonW _Neg	W_Pos	W _Neg
Actinobacteria	*Bifidobacterium bifidum*	0.034	2.021	0.414	0.32
Actinobacteria	*Bifidobacterium breve*	0	0.13	0	0.009
Actinobacteria	*Bifidobacterium longum*	0.17	9.488	0.581	1.227
Actinobacteria	*Gordonibacter pamelaeae*	0.001	0.001	0.005	0.019
Actinobacteria	*Unclassified Olsenella*	0.024	0.005	0	0.001
Ascomycota	*Saccharomyces cerevisiae*	0	0	0.003	0
Bacteroidetes	*Alistipes finegoldii*	0.010	0.064	0.515	0.302
Bacteroidetes	*Alistipes onderdonkii*	0.050	0.099	0.699	0.701
Bacteroidetes	*Alistipes putredinis*	0.068	0.269	3.651	2.740
Bacteroidetes	*Alistipes shahii*	0.062	0.297	1.034	0.459
Bacteroidetes	*Bacteroidales bacterium* ph8	0.022	0.008	0.397	0.214
Bacteroidetes	*Bacteroides caccae*	0.119	0.036	1.510	1.306
Bacteroidetes	*Bacteroides cellulosilyticus*	0.027	0.002	1.863	0.840
Bacteroidetes	*Bacteroides faecis*	0.003	0.002	0.290	0.098
Bacteroidetes	*Bacteroides massiliensis*	0.014	0	0.814	0.639
Bacteroidetes	*Bacteroides ovatus*	0.066	0.058	1.451	1.341
Bacteroidetes	*Bacteroides salyersiae*	0.000	0.001	0.228	0.038
Bacteroidetes	*Bacteroides stercoris*	0.005	0.036	1.671	0.930
Bacteroidetes	*Bacteroides uniformis*	0.097	0.111	2.735	5.123
Bacteroidetes	*Bacteroides vulgatus*	0.181	0.138	2.536	1.859
Bacteroidetes	*Bacteroides xylanisolvens*	0.010	0.042	0.395	0.147
Bacteroidetes	*Barnesiella intestinihominis*	0.032	0.391	1.935	0.750
Bacteroidetes	*Coprobacter fastidiosus*	0	0	0.043	0.021
Bacteroidetes	*Odoribacter splanchnicus*	0.161	0.031	1.537	0.462
Bacteroidetes	*Parabacteroides distasonis*	0.036	0.088	0.357	0.518
Bacteroidetes	*Parabacteroides merdae*	0.138	0.153	1.611	0.786
Bacteroidetes	*Prevotella copri*	14.982	19.386	4.883	2.830
Bacteroidetes	*Prevotella stercorea*	3.293	4.332	0.000	0.159
Bacteroidetes	*Unclassified Paraprevotella*	0.015	0.037	0.367	0.084
Euryarchaeota	*Methanobrevibacter smithii*	2.023	0.268	0.114	0.590
Euryarchaeota	*Unclassified Methanobrevibacter*	0.427	0.014	0.002	0.023
Firmicutes	*Clostridium leptum*	0.003	0.020	0.080	0.202
Firmicutes	*Coprococcus catus*	0.416	0.085	0.165	0.267
Firmicutes	*Eubacterium ventriosum*	0.036	0.016	0.424	0.256
Firmicutes	*Faecalibacterium prausnitzii*	17.893	12.781	8.018	8.321

Phylum	Species	Mean of proportion of sequences (%)			
		NonW _Pos	NonW _Neg	W_Pos	W _Neg
Firmicutes	*Flavonifractor plautii*	0	0.005	0.010	0.061
Firmicutes	*Holdemania filiformis*	0	0	0.012	0.020
Firmicutes	*Lachnospiraceae bacterium* 7_1_58FAA	0.014	0	0.053	0.082
Firmicutes	*Phascolarctobacterium succinatutens*	4.569	1.348	0.173	0.668
Firmicutes	*Pseudoflavonifractor capillosus*	0	0	0.002	0.008
Firmicutes	*Ruminococcus albus*	0	0	0.002	0.002
Firmicutes	*Ruminococcus sp* 5_1_39BFAA	0.226	0.082	0.547	1.973
Firmicutes	*Ruminococcus torques*	1.163	1.411	0.425	2.891
Firmicutes	*Unclassified Oscillibacter*	0.041	0.037	0.234	0.406
Proteobacteria	*Parasutterella excrementihominis*	0.004	0.006	0.065	0.012
Spirochaetes	*Treponema succinifaciens*	6.458	0.231	0	0

Table S8. Species abundances which differed in *Blastocystis* ST infections and non-carriers (Neg: Negative). Between group differences were evaluated with Kruskal-Wallis tests with Benjamini-Hochberg FDR corrections (FDR<0.05). Species names are shaded based on the phylum groups: *Bacteroidetes* (green), *Firmicutes* (yellow), *Proteobacteria* (orange), *Spirochaetes* (blue), *Actinobacteria* (gray).

Species	Mean of proportion of sequences (%)					
	Neg	ST1	ST2	ST3	ST4	Mixed
Barnesiella intestinihominis	0.650	0.237	0.122	0.402	3.186	6.72E-03
Alistipes putredinis	2.047	0.398	0.296	0.825	4.813	2.42E-03
Alistipes shahii	0.414	0.124	0.171	0.283	1.312	0.036
Bacteroides uniformis	3.717	0.213	0.196	1.056	2.770	0.056
Bacteroides stercoris	0.679	0.012	0.042	0.552	2.475	9.54E-05
Prevotella intermedia	0	0	0	0	4.39E-04	0
Parabacteroides distasonis	0.398	0.130	0.026	0.032	0.678	0.048
Odoribacter splanchnicus	0.341	0.485	0.251	0.500	1.491	0.054
Clostridium leptum	0.151	1.79E-03	3.70E-03	0.024	0.026	2.19E-03
Ruminococcus sp 5_1_39BFAA	1.443	0.189	0.267	0.266	0.656	0.215
Eubacterium eligens	0.536	2.115	0.641	2.329	3.630	1.414
Phascolarctobacterium succinatutens	0.859	3.259	4.809	2.763	6.39E-04	5.587
Ruminococcus torques	2.476	1.296	1.375	0.950	0.330	0.768
Butyrivibrio crossotus	0.030	1.954	2.187	0.632	0.732	1.601
Ruminococcus flavefaciens	1.28E-03	2.01E-02	1.69E-03	1.99E-03	1.80E-04	3.38E-03
Ruminococcus champanellensis	2.29E-03	0.193	0.292	0.322	5.26E-03	1.357
Parasutterella excrementihominis	0.010	3.40E-03	5.65E-03	0.022	0.113	2.43E-04
Burkholderiales bacterium 1_1_47	7.31E-03	4.54E-03	0.017	7.83E-03	0.200	0
Desulfovibrio piger	0.026	0.328	0.084	0.215	0.026	0.098
Treponema succinifaciens	0.065	8.312	4.455	3.545	0	6.560
Unclassified *Brachyspira*	1.57E-03	0.032	0.036	7.41E-03	4.62E-03	0.026
Unclassified *Olsenella*	2.09E-03	0.024	0.023	0.015	0	0.022
Unclassified *Methanobrevibacter*	0.020	0.220	0.730	0.223	0	0.450

APPENDIX B – SUPPLEMENTARY FIGURES

Figure S1. Enrichment of microbial species when *Blastocystis* presence or absence in westernized or non-westernized groups. Analyzed using LEfSe tool at effect size of 3.5. Purple colour for westernized carriers (W_Positive), blue for westernized non-carriers (W_Negative), green for non-westernized carriers (NonW_Positive), and red for non-westernized non-carriers (NonW_Negative).

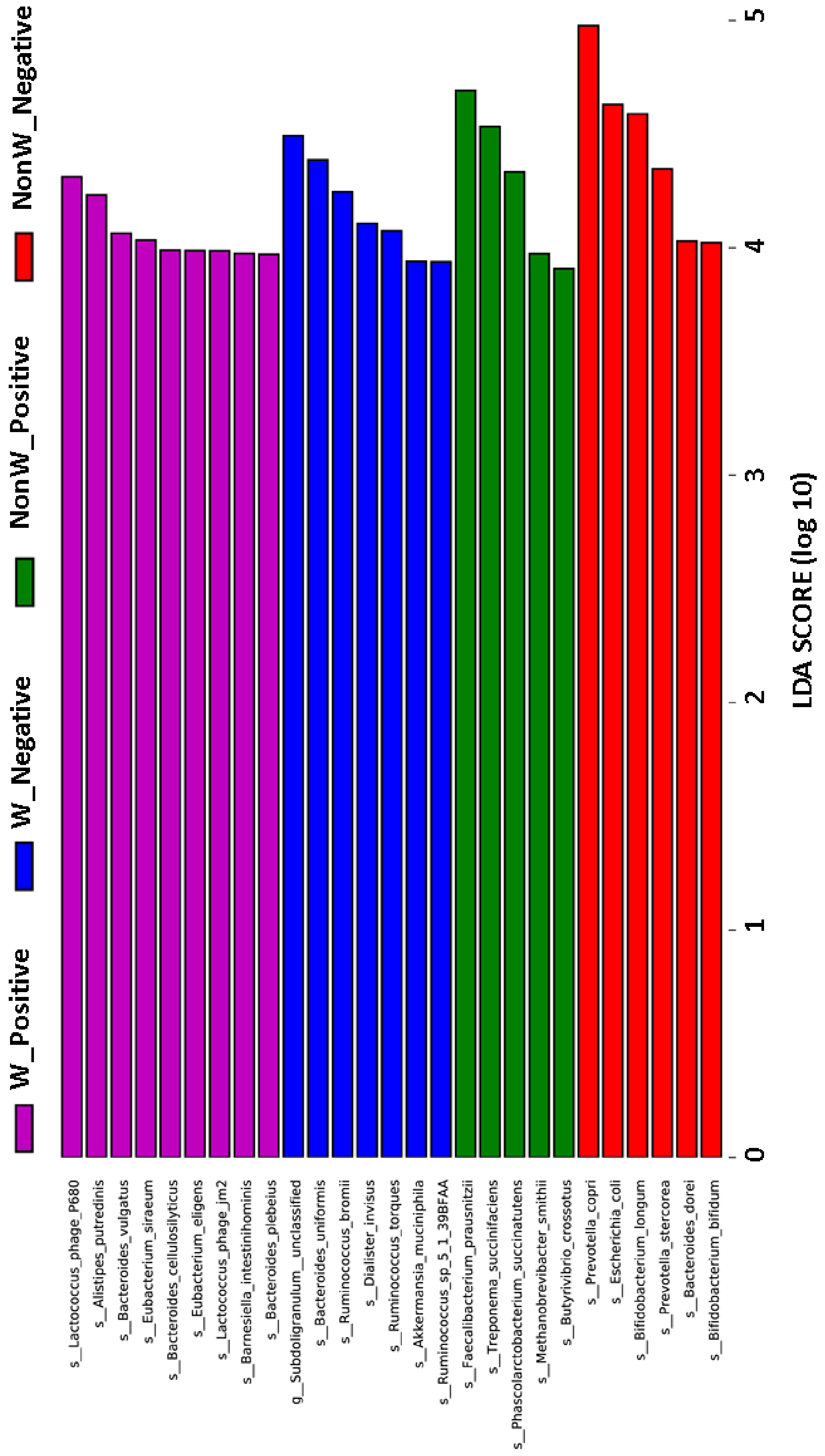

W_Positive
W_Negative
NonW_Positive
NonW_Negative
s__Lactococcus_phage_P680
s__Alistipes_putredinis
s__Bacteroides_vulgatus
s__Eubacterium_siraeum
s__Bacteroides_cellulosilyticus
s__Eubacterium_eligens
s__Lactococcus_phage_jm2
s__Barnesiella_intestinihominis
s__Bacteroides_plebeius
g__Subdoligranulum__unclassified
s__Bacteroides_uniformis
s__Ruminococcus_bromii
s__Dialister_invisus
s__Ruminococcus_torques
s__Akkermansia_muciniphila
s__Ruminococcus_sp_5_1_39BFAA
s__Faecalibacterium_prausnitzii
s__Treponema_succinifaciens
s__Phascolarctobacterium_succinatutens
s__Methanobrevibacter_smithii
s__Butyrivibrio_crossotus
s__Prevotella_copri
s__Escherichia_coli
s__Bifidobacterium_longum
s__Prevotella_stercorea
s__Bacteroides_dorei
s__Bifidobacterium_bifidum
0
1
2
3
4
5
LDA SCORE (log 10)

Figure S2. Pairwise comparison of enrichment or depletion of microbial species when *Blastocystis* presence or absence between (a) non-westernized (NonW) positive and westernized positive individuals, (b) non-westernized positive and non-westernized negative, and (c) westernized positive and westernized negative. Statistic test used: Welch's *t*-test with Storey FDR (FDR < 0.05).

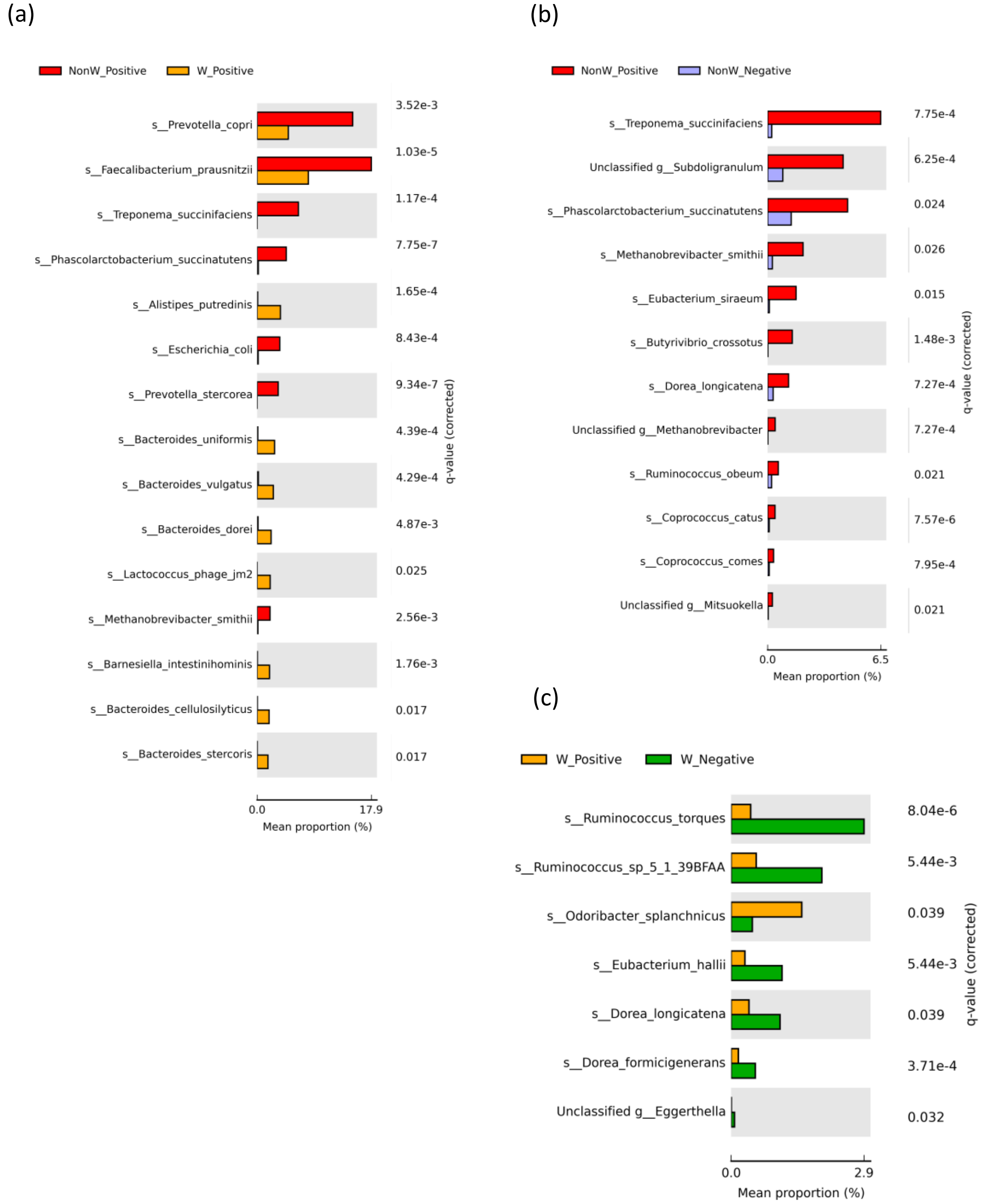


Figure S3 Heatmap of enrichment or depletion of microbial species for bacterial or archaeal species among groups of *Blastocystis* STs and *Blastocystis*-negative samples in non-westernized (NW) or westernized (W) individuals. The rows and columns were clustered using complete linkage clustering of similarities in similarity microbial species abundance using the correlation distance function and the Bray-Curtis distance metric, respectively.Groups with less than five samples were included. Among group differences were evaluated with ANVOA test without corrections (p-value<0.05). The number of samples in each group was labelled in the brackets. Neg: *Blastocystis* absent.

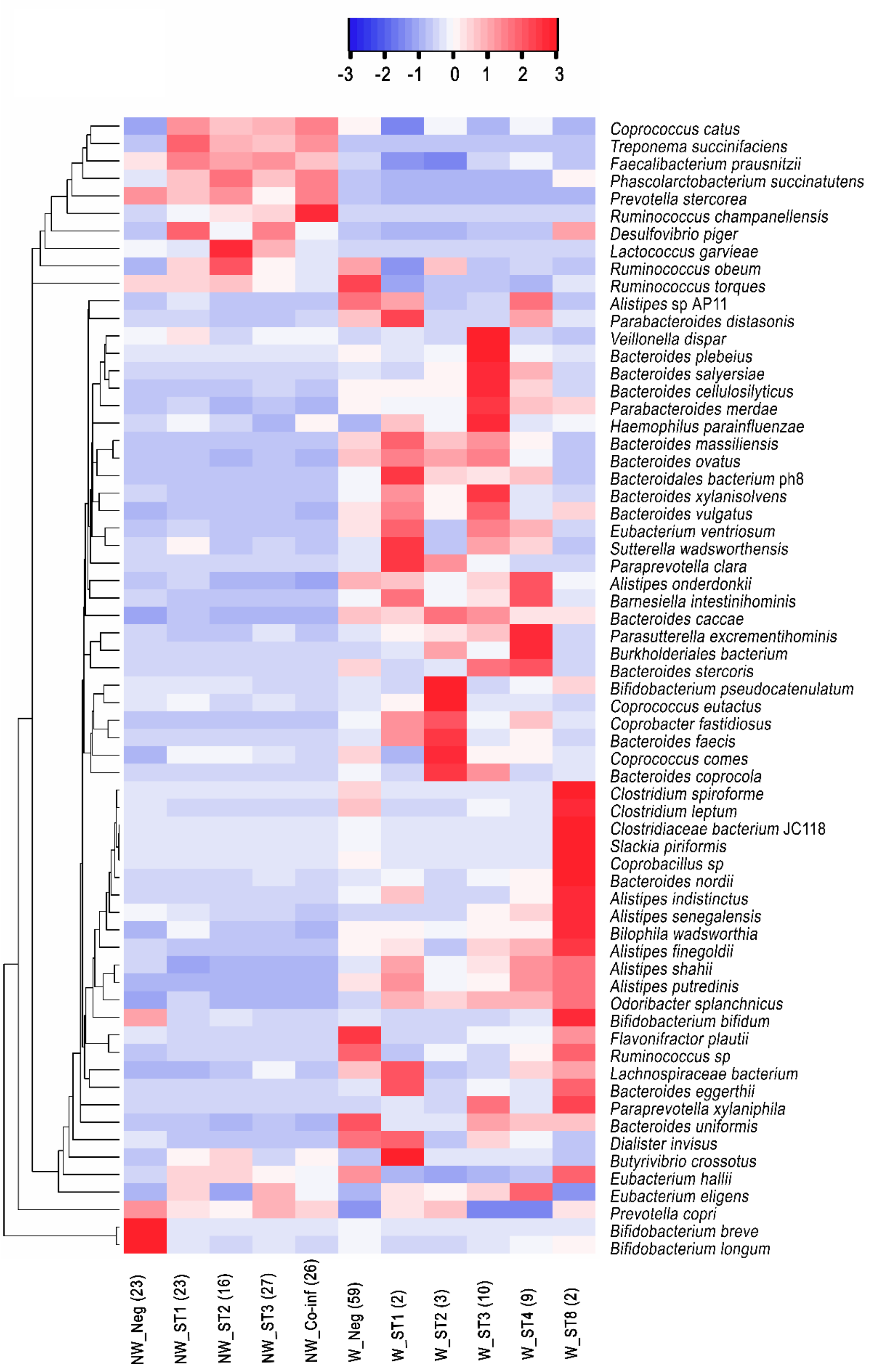
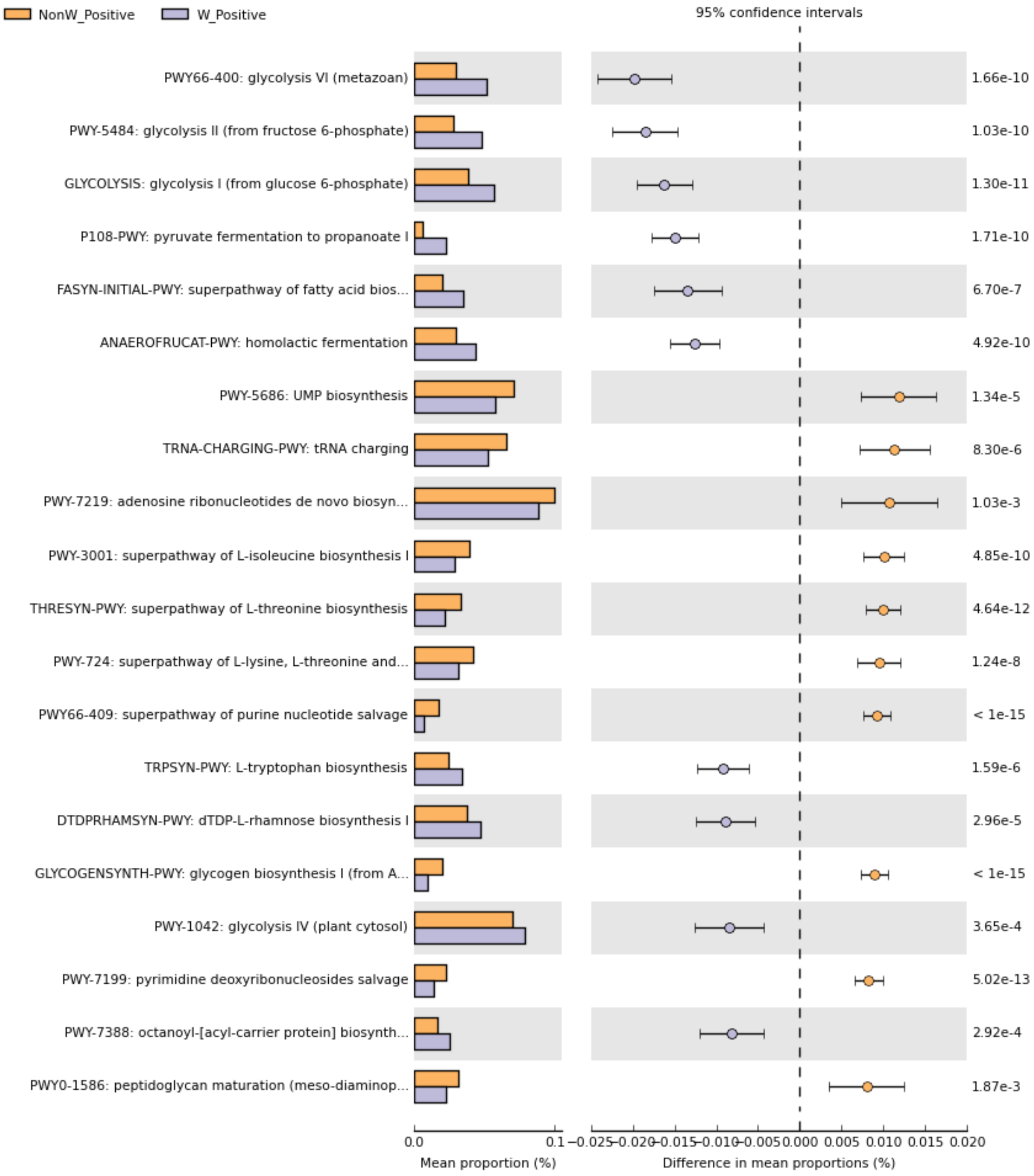
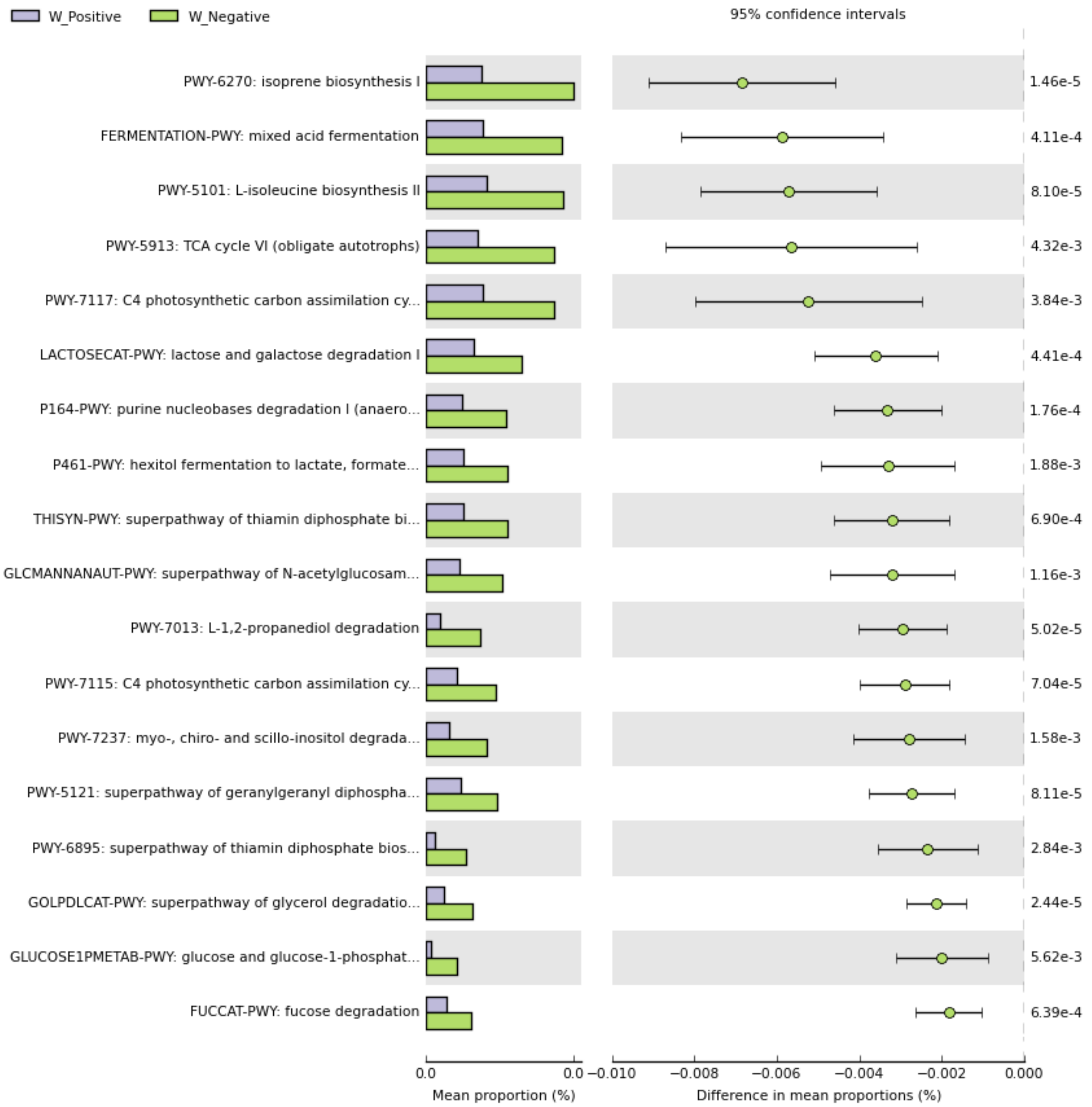

-3
-2
-1
0
1
2
3
Coprococcus catus
Treponema succinifaciens
Faecalibacterium prausnitzii
Phascolarctobacterium succinatutens
Prevotella stercorea
Ruminococcus champanellensis
Desulfovibrio piger
Lactococcus garvieae
Ruminococcus obeum
Ruminococcus torques
Alistipes sp AP11
Parabacteroides distasonis
Veillonella dispar
Bacteroides plebeius
Bacteroides salyersiae
Bacteroides cellulosilyticus
Parabacteroides merdae
Haemophilus parainfluenzae
Bacteroides massiliensis
Bacteroides ovatus
Bacteroidales bacterium ph8
Bacteroides xylanisolvens
Bacteroides vulgatus
Eubacterium ventriosum
Sutterella wadsworthensis
Paraprevotella clara
Alistipes onderdonkii
Barnesiella intestinihominis
Bacteroides caccae
Parasutterella excrementihominis
Burkholderiales bacterium
Bacteroides stercoris
Bifidobacterium pseudocatenulatum
Coprococcus eutactus
Coprobacter fastidiosus
Bacteroides faecis
Coprococcus comes
Bacteroides coprocola
Clostridium spiroforme
Clostridium leptum
Clostridiaceae bacterium JC118
Slackia piriformis
Coprobacillus sp
Bacteroides nordii
Alistipes indistinctus
Alistipes senegalensis
Bilophila wadsworthia
Alistipes finegoldii
Alistipes shahii
Alistipes putredinis
Odoribacter splanchnicus
Bifidobacterium bifidum
Flavonifractor plautii
Ruminococcus sp
Lachnospiraceae bacterium
Bacteroides eggerthii
Paraprevotella xylaniphila
Bacteroides uniformis
Dialister invisus
Butyrivibrio crossotus
Eubacterium hallii
Eubacterium eligens
Prevotella copri
Bifidobacterium breve
Bifidobacterium longum
NW_Neg (23)
NW_ST1 (23)
NW_ST2 (16)
NW_ST3 (27)
NW_Co-inf (26)
W_Neg (59)
W_ST1 (2)
W_ST2 (3)
W_ST3 (10)
W_ST4 (9)
W_ST8 (2)

Figure S4. Pairwise comparison of enrichment or depletion of gut microbiome pathways associated with *Blastocystis* presence or absence between (a) non-westernized (NonW) positive and westernized (W) positive individuals, (b) westernized positive and westernized negative, and (c) non-westernized positive and non-westernized negative. Statistic test used: Welch's *t*-test with Benjamini-Hochberg FDR correction (FDR < 0.05).

(a)



(b)



W_Positive
W_Negative
95% confidence intervals
PWY-6270: isoprene biosynthesis I
FERMENTATION-PWY: mixed acid fermentation
PWY-5101: L-isoleucine biosynthesis II
PWY-5913: TCA cycle VI (obligate autotrophs)
PWY-7117: C4 photosynthetic carbon assimilation cy...
LACTOSECAT-PWY: lactose and galactose degradation I
P164-PWY: purine nucleobases degradation I (anaero...
P461-PWY: hexitol fermentation to lactate, formate...
THISYN-PWY: superpathway of thiamin diphosphate bi...
GLCMANNANAUT-PWY: superpathway of N-acetylglucosam...
PWY-7013: L-1,2-propanediol degradation
PWY-7115: C4 photosynthetic carbon assimilation cy...
PWY-7237: myo-, chiro- and scillo-inositol degrada...
PWY-5121: superpathway of geranylgeranyl diphospha...
PWY-6895: superpathway of thiamin diphosphate bios...
GOLPDLCAT-PWY: superpathway of glycerol degradatio...
GLUCOSE1PMETAB-PWY: glucose and glucose-1-phosphat...
FUCCAT-PWY: fucose degradation
1.46e-5
4.11e-4
8.10e-5
4.32e-3
3.84e-3
4.41e-4
1.76e-4
1.88e-3
6.90e-4
1.16e-3
5.02e-5
7.04e-5
1.58e-3
8.11e-5
2.84e-3
2.44e-5
5.62e-3
6.39e-4
0.0
0.0
Mean proportion (%)
−0.010
−0.008
−0.006
−0.004
−0.002
0.000
Difference in mean proportions (%)

(c)

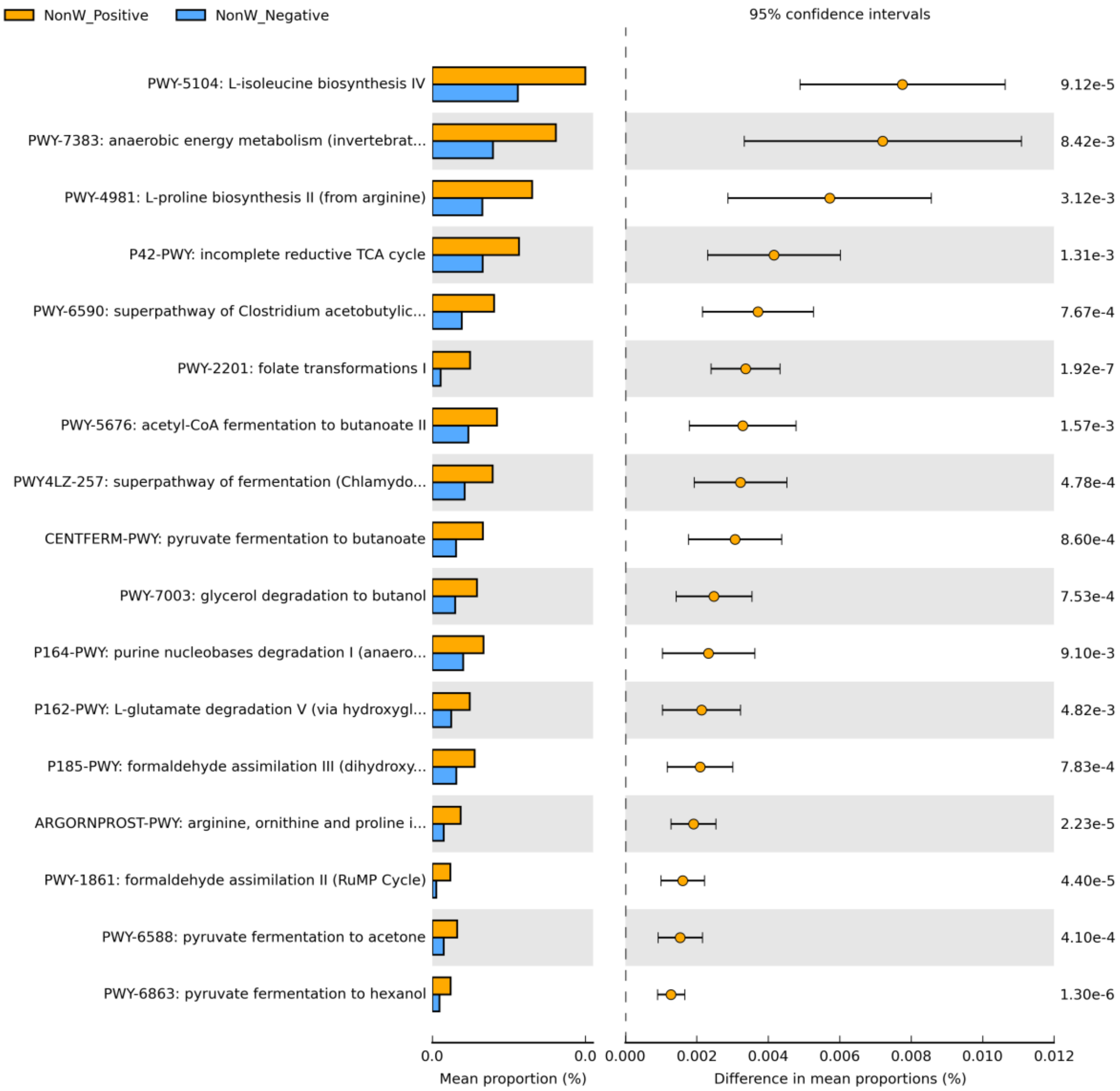

NonW_Positive
NonW_Negative
95% confidence intervals
PWY-5104: L-isoleucine biosynthesis IV
PWY-7383: anaerobic energy metabolism (invertebrat...
PWY-4981: L-proline biosynthesis II (from arginine)
P42-PWY: incomplete reductive TCA cycle
PWY-6590: superpathway of Clostridium acetobutylic...
PWY-2201: folate transformations I
PWY-5676: acetyl-CoA fermentation to butanoate II
PWY4LZ-257: superpathway of fermentation (Chlamydo...
CENTFERM-PWY: pyruvate fermentation to butanoate
PWY-7003: glycerol degradation to butanol
P164-PWY: purine nucleobases degradation I (anaero...
P162-PWY: L-glutamate degradation V (via hydroxygl...
P185-PWY: formaldehyde assimilation III (dihydroxy...
ARGORNPROST-PWY: arginine, ornithine and proline i...
PWY-1861: formaldehyde assimilation II (RuMP Cycle)
PWY-6588: pyruvate fermentation to acetone
PWY-6863: pyruvate fermentation to hexanol
9.12e-5
8.42e-3
3.12e-3
1.31e-3
7.67e-4
1.92e-7
1.57e-3
4.78e-4
8.60e-4
7.53e-4
9.10e-3
4.82e-3
7.83e-4
2.23e-5
4.40e-5
4.10e-4
1.30e-6
0.0
0.0
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Mean proportion (%)
Difference in mean proportions (%)

Figure S5. Percentage of unaligned reads in the MetaPhlan2 and HUMAnN2 analyses, (a) comparison of two steps, nucleotide-level mapping and translation-level mapping and comparison between *Blastocystis* carriers (Positive) and non-carriers in westernized and non-westernized samples in (b) nucleotide-level mapping step and (c) translation-level mapping step. Statistical test:Student's *t*-test

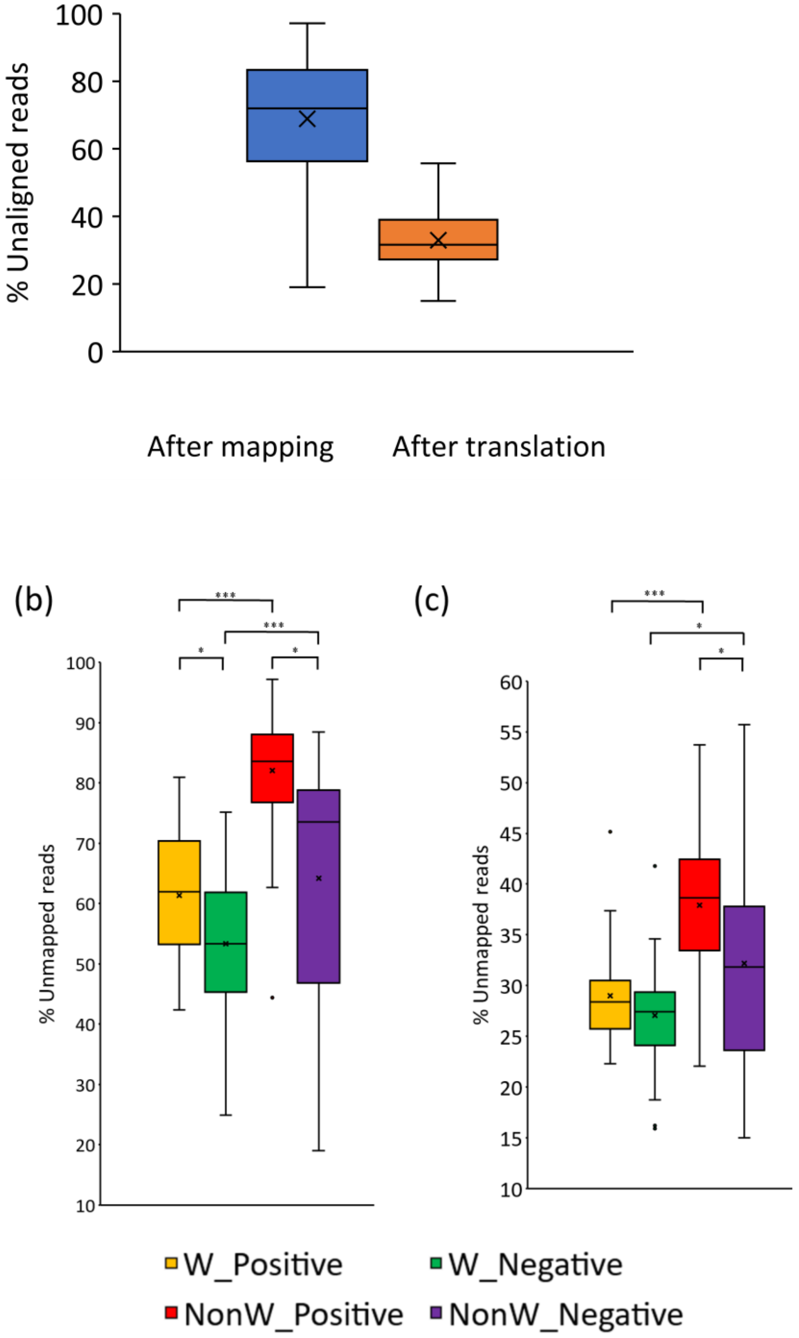
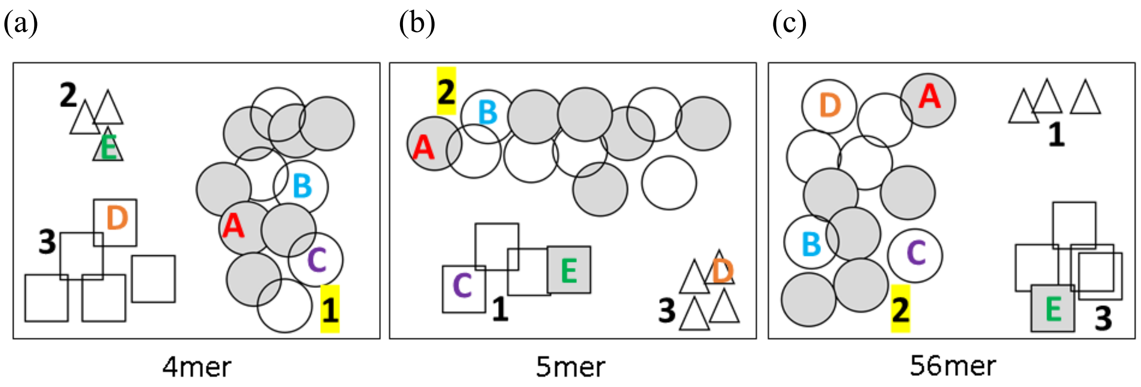


Figure S6. The schematic explanation of how eukaryotic contigs are selected based on MyCC binning, Centrifuge, and PLAST results. (a) – (c) represent the plots of cluster maps generated by MyCC based on marker genes, k-mer usage and depth of coverage for each k-mer (marked under the box). The geometric shapes (triangles, squares, and circles) represent contigs in different clusters. Contigs with a hit to eukaryotes by Centrifuge or PLAST are shaded in gray. Digital numbers in each plot represent the cluster number. The numbers of potential eukaryotic clusters are highlighted yellow. Alphabet letters A – E represent the contigs that appeared at least once in the potential eukaryotic clusters. To be included in a eukaryotic genome, a contig has to appear in at least twice in the potential eukaryotic clusters across different values of k-mers (Contigs A-C). Note that 56mer represents a combination of 5mer and 6mer.



Contig	Centrifuge results	PLAST results	Cluster Number of the contig in MyCC			Times hit potential Euk bin	Included/Excluded from final Euk genome
			4mer	5mer	56mer		
A	Euk	-	1	2	2	3	Included
B	-	-	1	2	2	3	Included
C	-	-	1	1	2	2	Included
D	-	-	3	3	2	1	Excluded
E	-	Euk	2	1	3	0	Excluded

Figure S7. The size distribution of the contigs from the draft genomes generated by Eukfinder approaches and EukRep for (a) TRV13 and (b) TRV25samples.

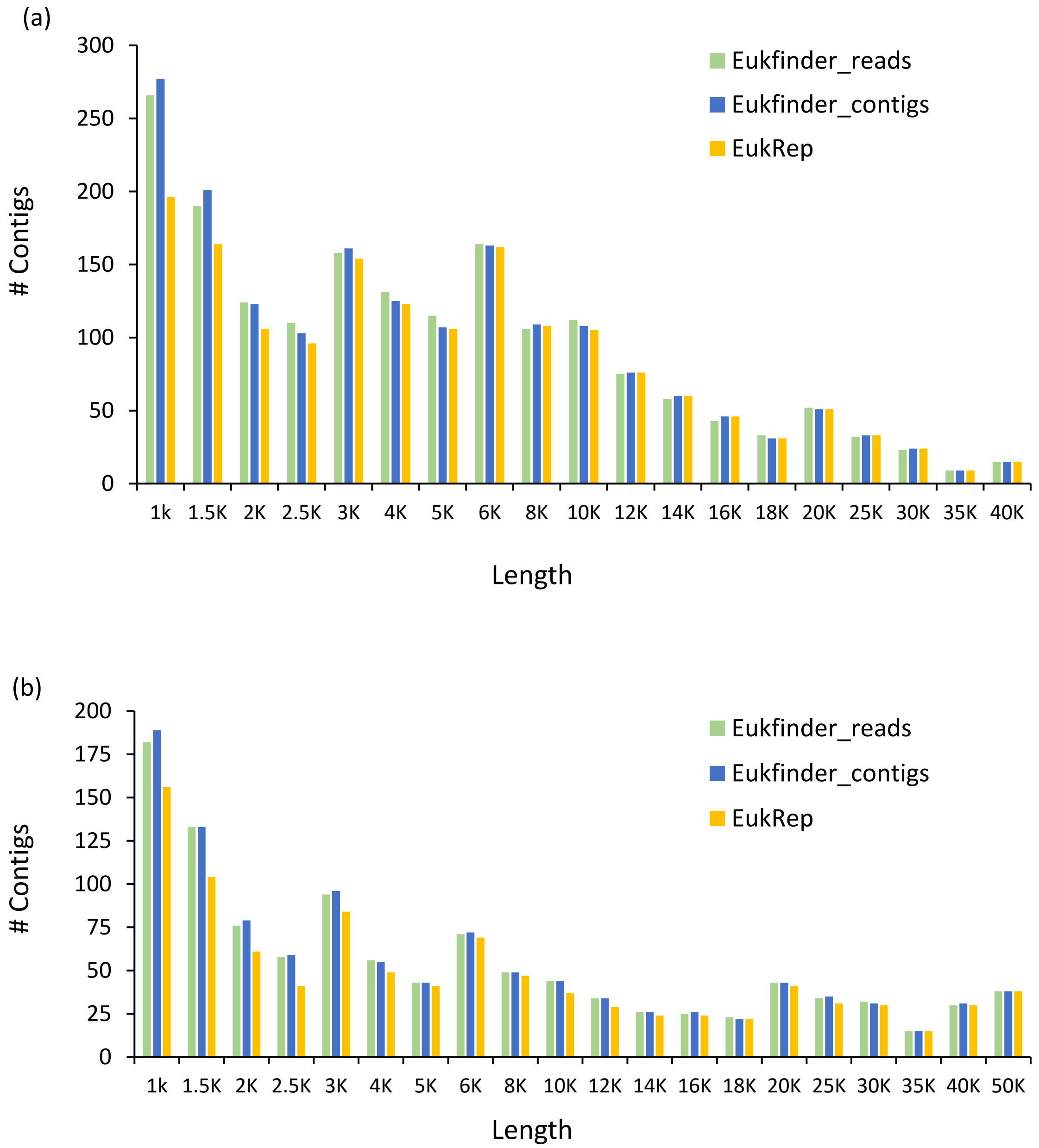


Figure S8. BRIG BLAST analysis of the MRO genomes recovered by difference methods from dataset MH0206 against ST2 reference genome. The inner-most ring represents the ST2 Fleming reference genome with length, followed by the GC content. Outer rings with colours show the recovered genome fragments. Black labels and arcs indicate the rRNA and protein-coding regions. Gray labels and arcs indicate tRNA regions.

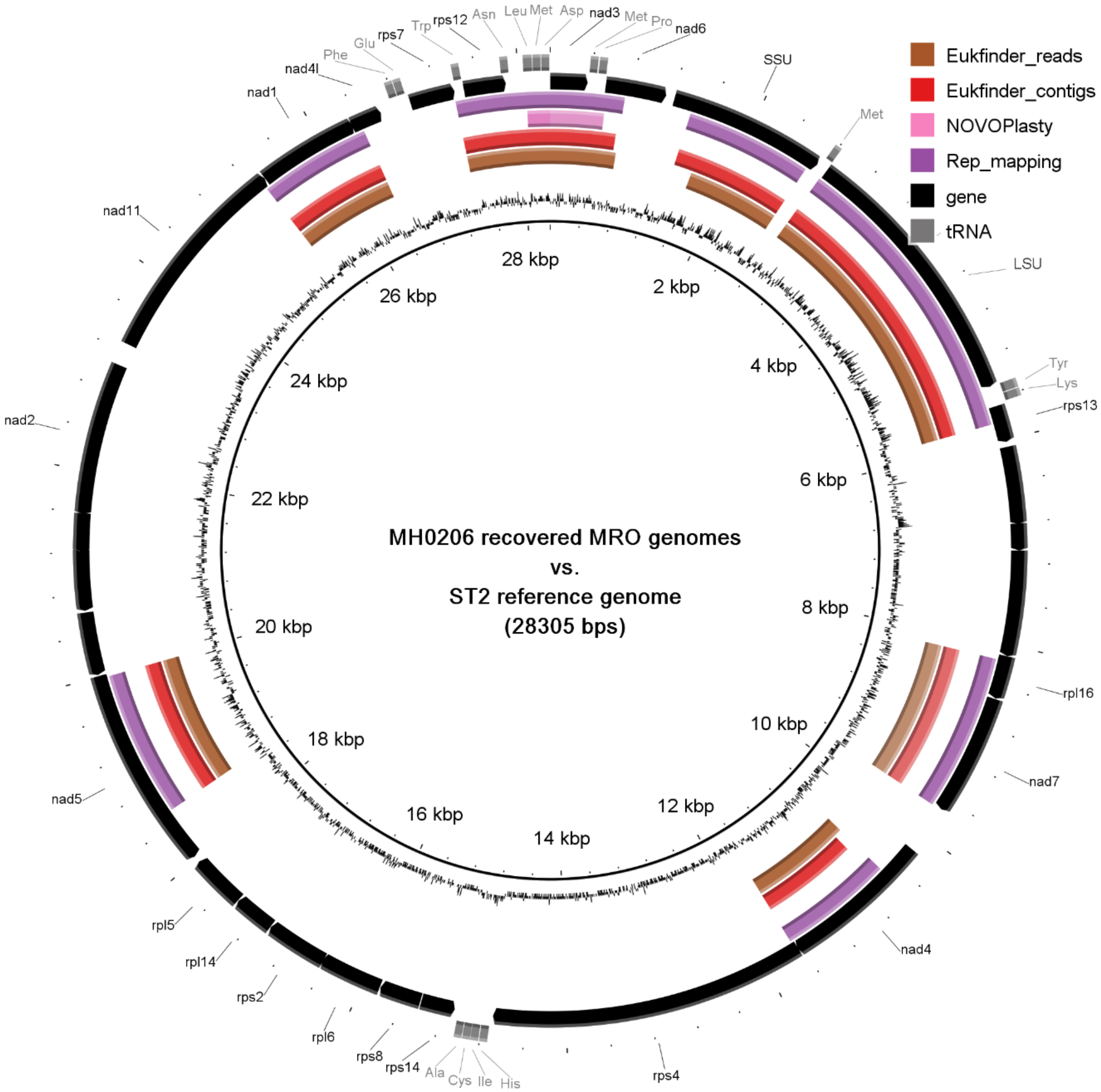
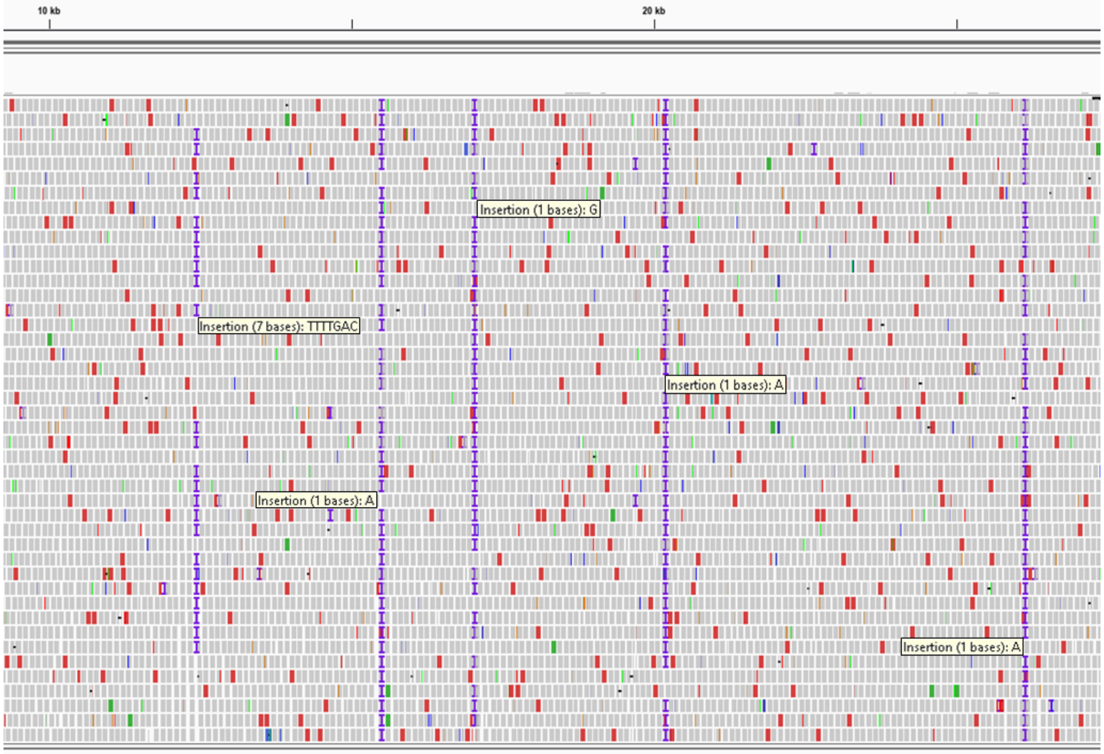
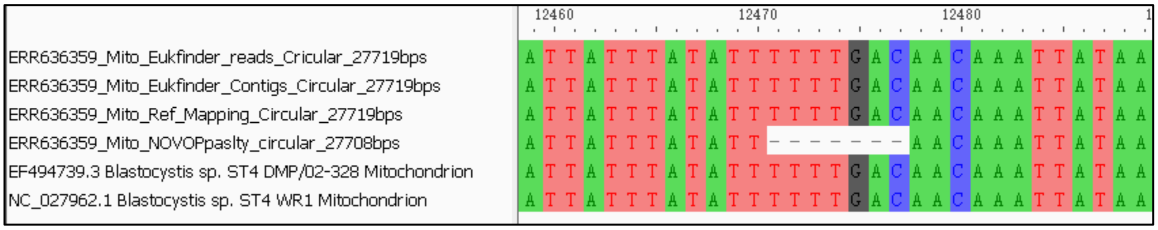


Figure S9. The sequence reads mapping against the TRV06 MRO genome recovered by NOVOPlasty with five regions of insertion to the genomes indicating the potential errors in the recovered genome. (a) IGV view of read mapping against the NOVOPlasty generated MRO genome. "I" indicates an insertion. (b) Alignment of four recovered TRV06 MRO genomes with two reference genomes shows the gaps with missing 7 bps on the genome generated by NOVOPlasty. A zoom-in view for mapped reads to MRO genome recovered by (c) NOVOPlasty and (d) Eukfinder_reads.

(a)

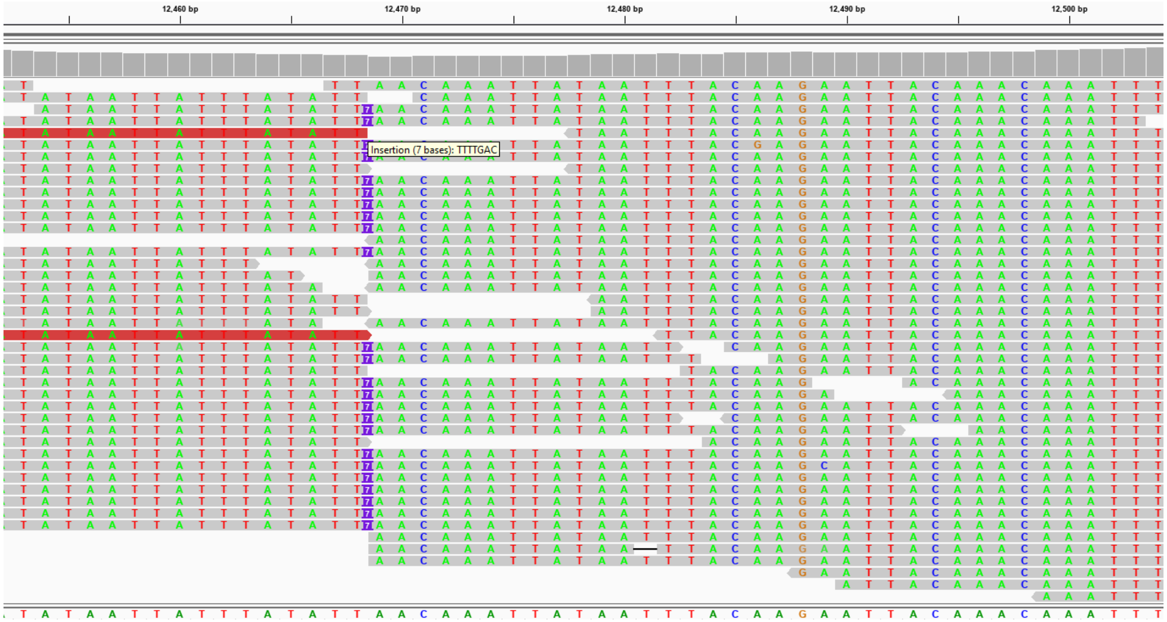


(b)



```
ERR636359_Mito_Eukfinder_reads_Cricular_27719bps      ATTATTTATATTTTTTGACAACAAATTATAA
ERR636359_Mito_Eukfinder_Contigs_Circular_27719bps    ATTATTTATATTTTTTGACAACAAATTATAA
ERR636359_Mito_Ref_Mapping_Circular_27719bps          ATTATTTATATTTTTTGACAACAAATTATAA
ERR636359_Mito_NOVOPpaslty_circular_27708bps          ATTATTTATATT-------AACAAATTATAA
EF494739.3 Blastocystis sp. ST4 DMP/02-328 Mitochondrion  ATTATTTATATTTTTTGACAACAAATTATAA
NC_027962.1 Blastocystis sp. ST4 WR1 Mitochondrion    ATTATTTATATTTTTTGACAACAAATTATAA
```

(c)


12,460 bp
12,470 bp
12,480 bp
12,490 bp
12,500 bp
Insertion (7 bases): TTTTGAC
T A T A A T T A T T T A T A T T A A C A A A T T A T A A T T T A C A A G A A T T A C A A A C A A A T T T

(d)

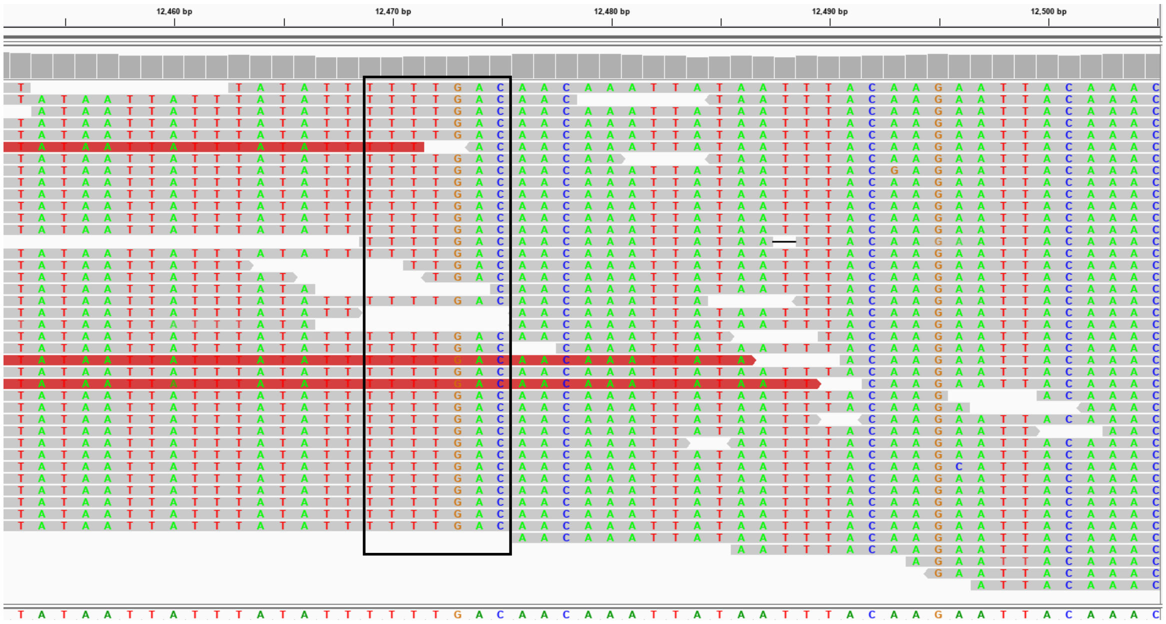
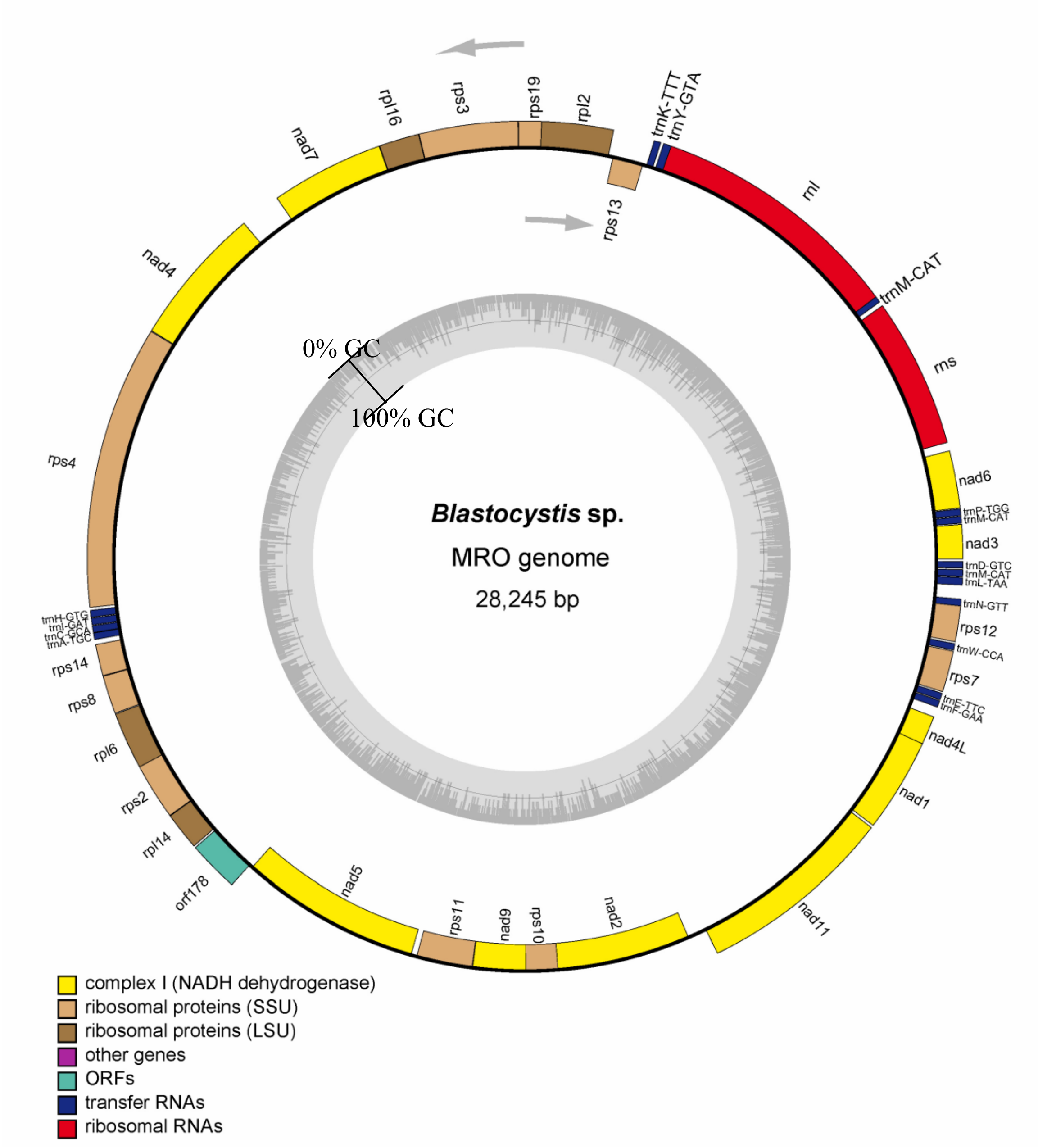
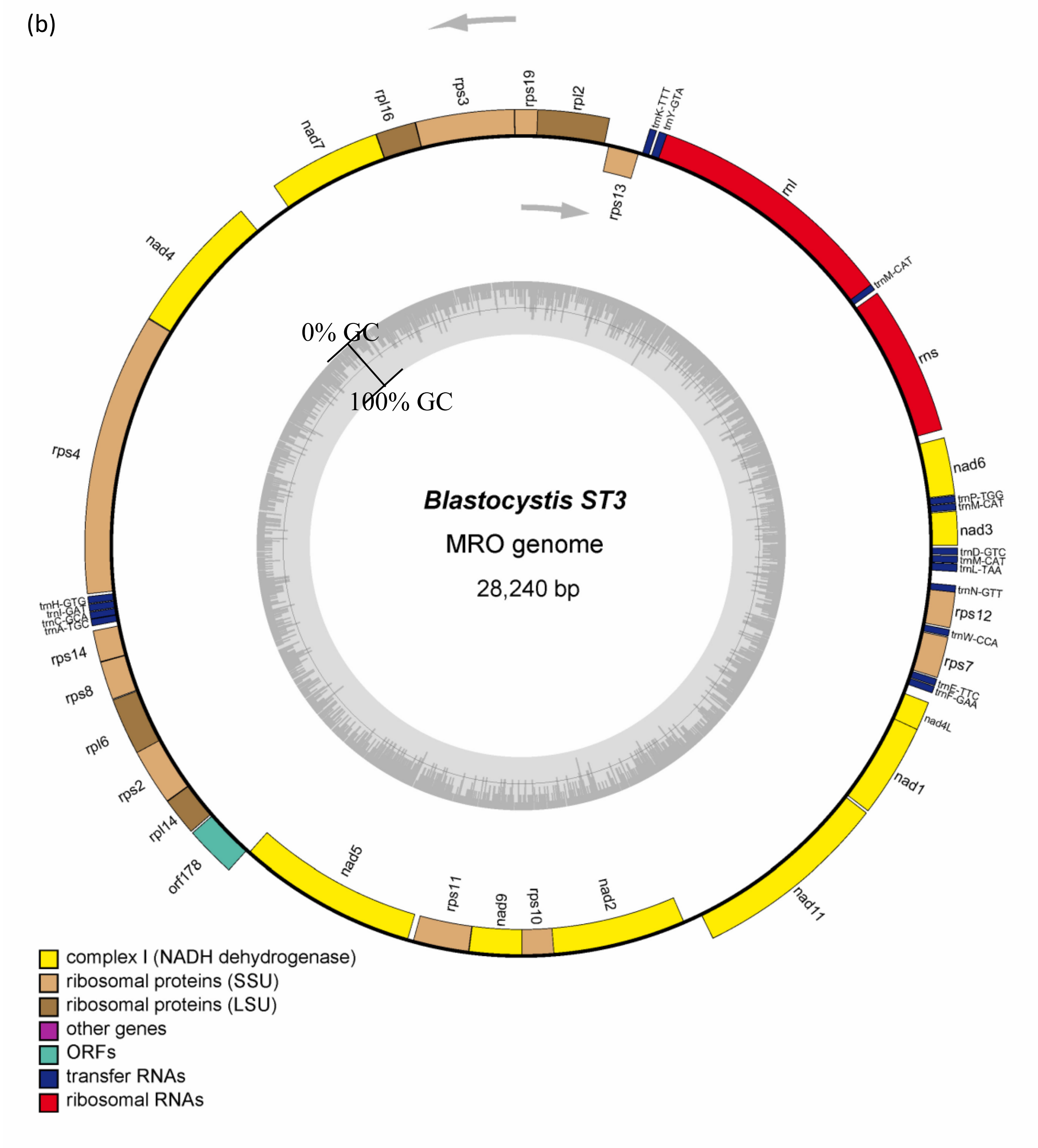

12,460 bp
12,470 bp
12,480 bp
12,490 bp
12,500 bp
T A T A A T T A T T T A T A T T T T T T G A C A A C A A A T T A T A A T T T A C A A G A A T T A C A A A C

Figure S10. Genome maps of *Blastocystis* ST3 MRO genomes recovered from (a) TRV02 and (b) TRV33 datasets. OGDraw v1.2 was used to draw the annotated MRO genome. The inner gray circular graph shows GC content with 0% on the outside and 100% on the inside and the central line represents 50% GC. Genes on the outer circle are transcribed in an anticlockwise direction.

(a)




(b)
nad7
rpl16
rps3
rps19
rpl2
trnK-TTT
trnY-GTA
rps13
rnl
trnM-CAT
rns
nad6
trnP-TGG
trnM-CAT
nad3
trnD-GTC
trnM-CAT
trnL-TAA
trnN-GTT
rps12
trnW-CCA
rps7
trnE-TTC
trnF-GAA
nad4L
nad1
nad11
nad2
rps10
nad9
rps11
nad5
orf178
rpl14
rps2
rpl6
rps8
rps14
trnH-GTG
trnI-GAT
trnC-GCA
trnA-TGC
rps4
nad4
0% GC
100% GC
Blastocystis ST3
MRO genome
28,240 bp
🟨 complex I (NADH dehydrogenase)
🟫 ribosomal proteins (SSU)
🟫 ribosomal proteins (LSU)
🟪 other genes
🟩 ORFs
🟦 transfer RNAs
🟥 ribosomal RNAs

REFERENCES

Adams JB, Johansen LJ, Powell LD, Quig D, Rubin RA. 2011. Gastrointestinal flora and gastrointestinal status in children with autism--comparisons to typical children and correlation with autism severity. BMC Gastroenterol. 11:22. doi: 10.1186/1471-230X-11-22

Ahn J, Sinha R, Pei Z, Dominianni C, Wu J, Shi J, Goedert JJ, Hayes RB, Yang L. 2013. Human gut microbiome and risk for colorectal cancer. J Natl Cancer Inst. 105:1907–1911. doi: 10.1093/jnci/djt300

Alfellani MA, Jacob AS, Perea NO, Krecek RC, Taner-Mulla D, Verweij JJ, Levecke B, Tannich E, Clark CG, Stensvold CR. 2013. Diversity and distribution of *Blastocystis* sp. subtypes in non-human primates. Parasitology. 140:966–971. doi: 10.1017/S0031182013000255

Alfellani MA, Stensvold CR, Vidal-Lapiedra A, Onuoha ESU, Fagbenro-Beyioku AF, Clark CG. 2013. Variable geographic distribution of *Blastocystis* subtypes and its potential implications. Acta Trop. 126:11–18. doi: 10.1016/j.actatropica.2012.12.011

Alfellani MA, Taner-Mulla D, Jacob AS, Imeede CA, Yoshikawa H, Stensvold CR, Clark CG. 2013. Genetic Diversity of *Blastocystis* in Livestock and Zoo Animals. Protist. 164:497–509. doi: 10.1016/j.protis.2013.05.003

Alikhan N-F, Petty NK, Zakour NL Ben, Beatson SA. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 12:402. doi: 10.1186/1471-2164-12-402

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410. doi: 10.1016/S0022-2836(05)80360-2

Andersen LO brie., Stensvold CR. 2016. *Blastocystis* in Health and Disease: Are We Moving from a Clinical to a Public Health Perspective? J Clin Microbiol. 54:524–528. doi: 10.1128/JCM.02520-15

Andersen LOB, Bonde I, Nielsen HBHB, Stensvold CR. 2015. A retrospective metagenomics approach to studying *Blastocystis*. FEMS Microbiol Ecol. 91:1–9. doi: 10.1093/femsec/fiv072

Audebert C, Even G, Cian A, Blastocystis Investigation Group, Loywick A, Merlin S, Viscogliosi E, Chabé M, El Safadi D, Certad G, et al. 2016. Colonization with the enteric protozoa *Blastocystis* is associated with increased diversity of human gut bacterial microbiota. Sci Rep. 6:1–11. doi: 10.1038/srep25255

Ayeni FA, Biagi E, Rampelli S, Fiori J, Soverini M, Audu HJ, Cristino S, Caporali L, Schnorr SL, Carelli V, et al. 2018. Infant and Adult Gut Microbiome and Metabolome in Rural Bassa and Urban Settlers from Nigeria. Cell Rep. 23:3056–3067. doi: 10.1016/j.celrep.2018.05.018

Aynur ZE, Güçlü Ö, Yıldız İ, Aynur H, Ertabaklar H, Bozdoğan B, Ertuğ S. 2019. Molecular characterization of *Blastocystis* in cattle in Turkey. Parasitol Res. 118:1055–1059. doi: 10.1007/s00436-019-06243-8

Babicki S, Arndt D, Marcu A, Liang Y, Grant JR, Maciejewski A, Wishart DS. 2016. Heatmapper: web-enabled heat mapping for all. Nucleic Acids Res. 44:W147–W153. doi: 10.1093/nar/gkw419

Bang C, Weidenbach K, Gutsmann T, Heine H, Schmitz RA. 2014. The intestinal archaea *Methanosphaera stadtmanae* and *Methanobrevibacter smithii* activate human dendritic cells. PLoS One. 9:1–9. doi: 10.1371/journal.pone.0099411

Beghini F, Pasolli E, Truong TD, Putignani L, Cacciò SM, Segata N. 2017. Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. ISME J. 11:2848–2863. doi: 10.1038/ismej.2017.139

Bengtsson-Palme J, Angelin M, Huss M, Kjellqvist S, Kristiansson E, Palmgren H, Joakim Larsson DG, Johansson A. 2015. The human gut microbiome as a transporter of antibiotic resistance genes between continents. Antimicrob Agents Chemother. 59:6551–6560. doi: 10.1128/AAC.00933-15

Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH. 2015. metaxa2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. Mol Ecol Resour. 15:1403–1414. doi: 10.1111/1755-0998.12399

Betts EL, Gentekaki E, Tsaousis AD. 2020. Exploring Micro-Eukaryotic Diversity in the Gut: Co-occurrence of *Blastocystis* Subtypes and Other Protists in Zoo Animals. Front Microbiol. 11:288. doi: 10.3389/fmicb.2020.00288

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120. doi: 10.1093/bioinformatics/btu170

Breitwieser FP, Baker DN, Salzberg SL. 2018. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol. 19:198. doi: 10.1186/s13059-018-1568-0

Brewster R, Tamburini FB, Asiimwe E, Oduaran O, Hazelhurst S, Bhatt AS. 2019. Surveying Gut Microbiome Research in Africans: Toward Improved Diversity and Representation. Trends Microbiol.:1–12. doi: 10.1016/j.tim.2019.05.006

Brumfield KD, Huq A, Colwell RR, Olds JL, Leddy MB. 2020. Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data. PLoS One. 15:1–21. doi: 10.1371/journal.pone.0228899

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12:59-60. doi: 10.1038/nmeth.3176

Caradonna T, Marangi M, Del Chierico F, Ferrari N, Reddel S, Bracaglia G, Normanno G, Putignani L, Giangaspero A. 2017. Detection and prevalence of protozoan parasites in ready-to-eat packaged salads on sale in Italy. Food Microbiol. 67:67–75. doi: 10.1016/j.fm.2017.06.006

Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Paley S, Pick J, Rhee SY, others. 2006. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. 34:D511--D516. doi: 10.1093/nar/gkj128

Chabé M, Lokmer A, Ségurel L. 2017. Gut Protozoa: Friends or Foes of the Human Gut Microbiota? Trends Parasitol. 33:925–934. doi: 10.1016/j.pt.2017.08.005

Chang C-J, Lin T-L, Tsai Y-L, Wu T-R, Lai W-F, Lu C-C, Lai H-C. 2019. Next generation probiotics in disease amelioration. J Food Drug Anal. 27(3):615-622. doi: 10.1016/j.jfda.2018.12.011

Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, et al. 2013. Richness of human gut microbiome correlates with metabolic markers. Nature. 500:541–546. doi: 10.1038/nature12506

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2018. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. doi: 10.1093/bioinformatics/btz848

Che Y, Xia Y, Liu L, Li A-D, Yang Y, Zhang T. 2019. Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. Microbiome. 7:44. doi: 10.1186/s40168-019-0663-0

Chung WSF, Walker AW, Louis P, Parkhill J, Vermeiren J, Bosscher D, Duncan SH, Flint HJ. 2016. Modulation of the human gut microbiota by dietary fibres occurs at the species level. BMC Biol. 14:1–13. doi: 10.1186/s12915-015-0224-3

Clark CG, van der Giezen M, Alfellani MA, Stensvold CR. 2013. Recent Developments in *Blastocystis* Research. Adv Parasitol. 82:1-32. doi: 10.1016/B978-0-12-407706-5.00001-0

Clemente JC, Ursell LK, Parfrey LW, Knight R. 2012. The impact of the gut microbiota on human health: An integrative view. Cell. 148:1258–1270. doi: 10.1016/j.cell.2012.01.035

Comeau AM, Douglas GM, Langille MGI. 2017. Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. mSystems. 2:1–11. doi: 10.1128/mSystems.00127-16

Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 33:2938–2940. doi: 10.1093/bioinformatics/btx364

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science. 326:1694–1697. doi: 10.1126/science.1177486

Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, Le Chatelier E, Almeida M, Quinquis B, Levenez F, Galleron N, others. 2013. Dietary intervention impact on gut microbial gene richness. Nature. 500:585–588. doi: 10.1038/nature12480

Dao MC, Everard A, Aron-Wisnewsky J, Sokolovska N, Prifti E, Verger EO, Kayser BD, Levenez F, Chilloux J, Hoyles L, et al. 2016. *Akkermansia muciniphila* and improved metabolic health during a dietary intervention in obesity: Relationship with gut microbiome richness and ecology. Gut. 65:426–436. doi: 10.1136/gutjnl-2014-308778

David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling A V, Devlin AS, Varma Y, Fischbach MA, et al. 2013. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 505:559. doi: 10.1038/nature12820

Denoeud F, Roussel M, Noel B, Wawrzyniak I, Da Silva C, Diogon M, Viscogliosi E, Brochier-Armanet C, Couloux A, Poulain J, et al. 2011. Genome sequence of the stramenopile *Blastocystis*, a human anaerobic parasite. Genome Biol. 12(3):R29. doi: 10.1186/gb-2011-12-3-r29

Derelle R, López-García P, Timpano H, Moreira D. 2016. A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). Mol Biol Evol. 33:2890–2898. doi: 10.1093/molbev/msw168

Dicksved J, Halfvarson J, Rosenquist M, Järnerot G, Tysk C, Apajalahti J, Engstrand L, Jansson JK. 2008. Molecular analysis of the gut microbiota of identical twins with Crohn’s disease. ISME J. 2:716–727. doi: 10.1038/ismej.2008.37

Dierckxsens N, Mardulyn P, Smits G. 2017. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 45(4):e18. doi: 10.1093/nar/gkw955

Dogruman-Al F, Kustimur S, Yoshikawa H, Tuncer C, Simsek Z, Tanyuksel M, Araz E, Boorom K. 2009. *Blastocystis* subtypes in irritable bowel syndrome and inflammatory bowel disease in Ankara, Turkey. Mem Inst Oswaldo Cruz. 104:724–727. doi: 10.1590/s0074-02762009000500011

Donaldson GP, Lee SM, Mazmanian SK. 2015. Gut biogeography of the bacterial microbiota. Nat Rev Microbiol. 14:20–32. doi: doi: 10.1038/nrmicro3552

Duboc H, Rainteau D, Rajca S, Humbert L, Farabos D, Maubert M, Grondin V, Jouet P, Bouhassira D, Seksik P, others. 2012. Increase in fecal primary bile acids and dysbiosis in patients with diarrhea-predominant irritable bowel syndrome. Neurogastroenterol Motil. 24(6):513-20, e246-7. doi: 10.1111/j.1365-2982.2012.01893.x

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. 2005. Diversity of the human intestinal microbial flora. Science. 308(5728):1635-8. doi: 10.1126/science.1110591

Eme L, Gentekaki E, Curtis B, Archibald JM, Roger AJ. 2017. Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite *Blastocystis* to the Gut. Curr Biol. 27:807–820. doi: 10.1016/j.cub.2017.02.003

Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, et al. 2016. Population-level analysis of gut microbiome variation. Science. 352(6285):560-4. doi: 10.1126/science.aad3503

De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P. 2010. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A. 107:14691–14696. doi: 10.1073/pnas.1005963107

Forsell J, Bengtsson-Palme J, Angelin M, Johansson A, Evengård B, Granlund M. 2017. The relation between *Blastocystis* and the intestinal microbiota in Swedish travellers. BMC Microbiol. 17(1):231. doi: 10.1186/s12866-017-1139-7

Forsell J, Granlund M, Samuelsson L, Koskiniemi S, Edebro H, Evengård B. 2016. High occurrence of *Blastocystis* sp. subtypes 1-3 and *Giardia intestinalis* assemblage B among patients in Zanzibar, Tanzania. Parasites and Vectors. 9:1–12. doi: 10.1186/s13071-016-1637-8

Francino MP. 2016. Antibiotics and the human gut microbiome: Dysbioses and accumulation of resistances. Front Microbiol. 6:1543. doi: 10.3389/fmicb.2015.01543

Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 15:962–968. doi: 10.1038/s41592-018-0176-y

Frías L, Leles D, Araújo A. 2013. Studies on protozoa in ancient remains - A Review. Mem Inst Oswaldo Cruz. 108:1–12. doi: 10.1590/s0074-02762013000100001

Fujimura KE, Slusher NA, Cabana MD, Lynch S V. 2010. Role of the gut microbiota in defining human health. Expert Rev Anti Infect Ther. 8:435–454. doi: 10.1586/eri.10.14

Galván-Moroyoqui JM, del Carmen Domínguez-Robles M, Franco E, Meza I. 2008. The interplay between *Entamoeba* and enteropathogenic bacteria modulates epithelial cell damage. PLoS Negl Trop Dis. 2(7):e266. doi: 10.1371/journal.pntd.0000266

Gentekaki E, Curtis BA, Stairs CW, Klimeš V, Eliáš M, Salas-Leiva DE, Herman EK, Eme L, Arias MC, Henrissat B, et al. 2017. Extreme genome diversity in the hyper-prevalent parasitic eukaryote *Blastocystis*. PLoS Biol. 15(9):e2003769. doi: 10.1371/journal.pbio.2003769

Ghavami SB, Rostami E, Sephay AA, Shahrokh S, Balaii H, Aghdaei HA, Zali MR. 2018. Alterations of the human gut *Methanobrevibacter smithii* as a biomarker for inflammatory bowel diseases. Microb Pathog. 117:285–289. doi: 10.1016/j.micpath.2018.01.029

Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, et al. 2014. Human genetics shape the gut microbiome. Cell. 159:789–799. doi: 10.1016/j.cell.2014.09.053

Greige S, El Safadi D, Khaled S, Gantois N, Baydoun M, Chemaly M, Benamrouz-Vanneste S, Chabé M, Osman M, Certad G, et al. 2019. First report on the prevalence and subtype distribution of *Blastocystis* sp. in dairy cattle in Lebanon and assessment of zoonotic transmission. Acta Trop. 194:23–29. doi: 10.1016/j.actatropica.2019.02.013

Greiner S, Lehwark P, Bock R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 47(W1):W59-W64. doi: 10.1093/nar/gkz238

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29:1072–1075. doi: 10.1093/bioinformatics/btt086

Hugon P, Dufour JC, Colson P, Fournier PE, Sallah K, Raoult D. 2015. A comprehensive repertoire of prokaryotic species identified in human beings. Lancet Infect Dis. 15:1211–1219. doi: doi: 10.1016/S1473-3099(15)00293-5

Humen MA, De Antoni GL, Benyacoub J, Costas ME, Cardozo MI, Kozubsky L, Saudan KY, Boenzli-Bruand A, Blum S, Schiffrin EJ, Pérez PF. 2005. *Lactobacillus johnsonii* La1 antagonizes *Giardia intestinalis* in vivo. Infect Immun. 73:1265–1269. doi: 10.1128/IAI.73.2.1265-1269.2005

Hussein E, Hussein A, Eida M, Atwa M. 2008. Pathophysiological variability of different genotypes of human *Blastocystis hominis* Egyptian isolates in experimentally infected rats. Parasitol Res. 102:853–860. doi: 10.1007/s00436-007-0833-z

Javanmard E, Niyyati M, Ghasemi E, Mirjalali H, Asadzadeh Aghdaei H, Zali MR. 2018. Impacts of human development index and climate conditions on prevalence of *Blastocystis*: A systematic review and meta-analysis. Acta Trop. 185:193–203. doi: 10.1016/j.actatropica.2018.05.014

Jedelský PL, Doležal P, Rada P, Pyrih J, Šmíd O, Hrdý I, Šedinová M, Marcinčiková M, Voleman L, Perry AJ, et al. 2011. The minimal proteome in the reduced mitochondrion of the parasitic protist *Giardia intestinalis*. PLoS One. 6:15–21. doi: 10.1371/journal.pone.0017285

Jerlström-Hultqvist J, Franzén O, Ankarklev J, Xu F, Nohýnková E, Andersson JO, Svärd SG, Andersson B. 2010. Genome analysis and comparative genomics of a *Giardia intestinalis* assemblage E isolate. BMC Genomics. 11. doi: 10.1186/1471-2164-11-543

Jimenez-Gonzalez DE, Martinez-Flores WA, Reyes-Gordillo J, Ramirez-Miranda ME, Arroyo-Escalante S, Romero-Valdovinos M, Stark D, Souza-Saldivar V, Martinez-Hernandez F, Flisser A, others. 2012. *Blastocystis* infection is associated with irritable bowel syndrome in a Mexican patient population. Parasitol Res. 110:1269–1275. doi: 10.1007/s00436-011-2626-7

Jiménez PA, Jaimes JE, Ramírez JD. 2019. A summary of *Blastocystis* subtypes in North and South America. Parasit Vectors. 12:1–9. doi: 10.1186/s13071-019-3641-2

Kelly JR, Borre Y, O'Brien C, Patterson E, El Aidy S, Deane J, Kennedy PJ, Beers S, Scott K, Moloney G, others. 2016. Transferring the blues: depression-associated gut microbiota induces neurobehavioural changes in the rat. J Psychiatr Res. 82:109–118. doi: 10.1016/j.jpsychires.2016.07.019

Khachatryan L, de Leeuw RH, Kraakman MEM, Pappas N, te Raa M, Mei H, de Knijff P, Laros JFJ. 2020. Taxonomic classification and abundance estimation using 16S and WGS—A comparison using controlled reference samples. Forensic Sci Int Genet. 46:102257. doi: 10.1016/j.fsigen.2020.102257

Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and accurate classificaton of metagenomic sequences. Genome Res. 26(12):1721-1729. doi: 10.1101/gr.210641.116

Kim G, Deepinder F, Morales W, Hwang L, Weitsman S, Chang C, Gunsalus R, Pimentel M. 2012. *Methanobrevibacter smithii* is the predominant methanogen in patients with constipation-predominant IBS and methane on breath. Dig Dis Sci. 57:3213–3218. doi: 10.1007/s10620-012-2197-1

Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40:1–8. doi: 10.1093/nar/gkr771

Klimeš V, Gentekaki E, Roger AJ, Eliaš M. 2014. A large number of nuclear genes in the human parasite *blastocystis* require mRNA polyadenylation to create functional termination codons. Genome Biol Evol. 6:1956–1961. doi: 10.1093/gbe/evu146

Koliada A, Syzenko G, Moseiko V, Budovska L, Puchkov K, Perederiy V, Gavalko Y, Dorofeyev A, Romanenko M, Tkach S, others. 2017. Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. BMC Microbiol. 17:120. doi: 10.1186/s12866-017-1027-1

Laforest-Lapointe, I., & Arrieta, M. C. 2018. Microbial eukaryotes: a missing link in gut microbiome studies. mSystems, 3(2). doi: 10.1128/mSystems.00201-17

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357–359. doi: 10.1038/nmeth.1923

Lantsman Y, Tan KSW, Morada M, Yarlett N. 2008. Biochemical characterization of a mitochondrial-like organelle from *Blastocystis* sp. subtype 7. Microbiology. 154:2757–2766. doi: 10.1099/mic.0.2008/017897-0

Lathrop SK, Bloom SM, Rao SM, Nutsch K, Lio C-W, Santacruz N, Peterson DA, Stappenbeck TS, Hsieh C-S. 2011. Peripheral education of the immune system by colonic commensal microbiota. Nature. 478:250–254. doi: 10.1016/j.smim.2013.10.002

Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C. 2018. Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome. Omi A J Integr Biol. 22:248–254. doi: 10.1089/omi.2018.0013

Lee LI, Chye TT, Karmacharya BM, Govind SK. 2012. *Blastocystis* sp.: waterborne zoonotic organism, a possibility? Parasit Vectors. 5:130. doi: 10.1186/1756-3305-5-130

Leelayoova S, Siripattanapipong S, Thathaisong U, Naaglor T, Taamasri P, Piyaraj P, Mungthin M. 2008. Drinking water: a possible source of *Blastocystis* spp. subtype 1 infection in schoolchildren of a rural community in central Thailand. Am J Trop Med Hyg. 79:401–406.

Legesse M, Erko B. 2004. Zoonotic intestinal parasites in *Papio anubis* (baboon) and

Cercopithecus aethiops (vervet) from four localities in Ethiopia. Acta Trop. 90:231–236. doi: 10.1016/j.actatropica.2003.12.003

Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1. 0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 102:3–11. doi: 10.1016/j.ymeth.2016.02.020

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25:2078–2079. doi: 10.1093/bioinformatics/btp352

Li J, Karim MR, Li D, Rahaman Sumon SMM, Siddiki SHMF, Rume FI, Sun R, Jia Y, Zhang L. 2019. Molecular characterization of *Blastocystis* sp. in captive wildlife in Bangladesh National Zoo: Non-human primates with high prevalence and zoonotic significance. Int J Parasitol Parasites Wildl. 10:314–320. doi: 10.1016/j.ijppaw.2019.11.003

Lin HH, Liao YC. 2016. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Sci Rep. 6:12–19. doi: 10.1038/srep24175

Liu X, Zou Q, Zeng B, Fang Y, Wei H. 2013. Analysis of fecal *Lactobacillus* community structure in patients with early rheumatoid arthritis. Curr Microbiol. 67:170–176. doi: 10.1007/s00284-013-0338-1

Lokmer A, Cian A, Froment A, Gantois N, Viscogliosi E, Chabé M, Ségurel L. 2019. Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels. PLoS One. 14(2):e0211139. doi: 10.1371/journal.pone.0211139

Lopez-Siles M, Duncan SH, Garcia-Gil LJ, Martinez-Medina M. 2017. *Faecalibacterium prausnitzii*: From microbiology to diagnostics and prognostics. ISME J. 11:841–852. doi: 10.1038/ismej.2016.176

Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. 2012. Diversity, stability and resilience of the human gut microbiota. Nature. 489:220–230. doi: 10.1038/nature11550

Lu J, Salzberg SL. 2018. Removing contaminants from databases of draft genomes. PLoS Comput Biol. 14:e1006277. doi: 10.1371/journal.pcbi.1006277

Lukeš J, Stensvold CR, Jirků-Pomajbíková K, Wegener Parfrey L. 2015. Are Human Intestinal Eukaryotes Beneficial or Commensals? PLoS Pathog. 11:7–12. doi: 10.1371/journal.ppat.1005039

Madsen K, Cornish A, Soper P, McKaigney C, Jijon H, Yachimec C, Doyle J, Jewell L, De Simone C. 2001. Probiotic bacteria enhance murine and human intestinal epithelial barrier function. Gastroenterology. 121:580–591. doi: 10.1053/gast.2001.27224

Maloney JG, Lombard JE, Urie NJ, Shivley CB, Santin M. 2019. Zoonotic and genetically diverse subtypes of *Blastocystis* in US pre-weaned dairy heifer calves. Parasitol Res. 118:575–582. doi: 10.1007/s00436-018-6149-3

Maloney JG, Molokin A, Santin M. 2019. Next generation amplicon sequencing improves detection of *Blastocystis* mixed subtype infections. Infect Genet Evol. 73:119–125. doi: 10.1016/j.meegid.2019.04.013

Marcelino VR, Holmes EC, Sorrell TC. 2020. The use of taxon-specific reference databases compromises metagenomic classification. BMC Genomics. 21:1–5. doi: 10.1186/s12864-020-6592-2

Marchesi JR, Adams DH, Fava F, Hermes GDA, Hirschfield GM, Hold G, Quraishi MN, Kinross J, Smidt H, Tuohy KM, et al. 2016. The gut microbiota and host health: A new clinical frontier. Gut. 65:330–339. doi: 10.1136/gutjnl-2015-309990

Martí JM. 2019. Recentrifuge: Robust comparative analysis and contamination removal for metagenomics. PLoS Comput Biol. 15:e1006967. doi: 10.1371/journal.pcbi.1006967

Mathur R, Kim G, Morales W, Sung J, Rooks E, Pokkunuri V, Weitsman S, Barlow GM, Chang C, Pimentel M. 2012. Intestinal *Methanobrevibacter smithii* but Not Total Bacteria Is Related to Diet-Induced Weight Gain in Rats. Obesity. 21:748–754. doi: 10.1002/oby.20277

Mbakwa CA, Penders J, Savelkoul PH, Thijs C, Dagnelie PC, Mommers M, Arts ICW. 2015. Gut colonization with *methanobrevibacter smithii* is associated with childhood weight development. Obesity. 23:2508–2516. doi: 10.1002/oby.21266

Menardo F, Loiseau C, Brites D, Coscolla M, Gygli SM, Rutaihwa LK, Trauner A, Beisel C, Borrell S, Gagneux S. 2018. Treemmer: A tool to reduce large phylogenetic datasets with minimal loss of diversity. BMC Bioinformatics. 19. doi: 10.1186/s12859-018-2164-8

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

Mi-Ichi F, Yousuf MA, Nakada-Tsukui K, Nozaki T. 2009. Mitosomes in *Entamoeba histolytica* contain a sulfate activation pathway. Proc Natl Acad Sci. 106:21731–21736. doi: 10.1073/pnas.0907106106

Million M, Maraninchi M, Henry M, Armougom F, Richet H, Carrieri P, Valero R, Raccah D, Vialettes B, Raoult D. 2012. Obesity-associated gut microbiota is enriched in *Lactobacillus reuteri* and depleted in *Bifidobacterium animalis* and

Methanobrevibacter smithii. Int J Obes. 36:817–825. doi: 10.1038/ijo.2011.153

Mizrahi-Man O, Davenport ER, Gilad Y. 2013. Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs.White BA, editor. PLoS One. 8:e53608. doi: 10.1371/journal.pone.0053608

Morton ER, Lynch J, Froment A, Lafosse S, Heyer E, Przeworski M, Blekhman R, Ségurel L. 2015. Variation in Rural African Gut Microbiota Is Strongly Correlated with Colonization by *Entamoeba* and Subsistence. PLoS Genet. 11(11):e1005658. doi: 10.1371/journal.pgen.1005658.

Moya A, Ferrer M. 2016. Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance. Trends Microbiol. 24:402–413. doi: 10.1016/j.tim.2016.02.002.

Nagel R, Cuttell L, Stensvold CR, Mills PC, Bielefeldt-Ohmann H, Traub RJ. 2012. *Blastocystis* subtypes in symptomatic and asymptomatic family members and pets and response to therapy. Intern Med J. 42:1187–1195. doi: 10.1111/j.1445-5994.2011.02626.x

Nagel R, Traub RJ, Allcock RJN, Kwan MMS, Bielefeldt-Ohmann H. 2016. Comparison of faecal microbiota in *Blastocystis*-positive and *Blastocystis*-negative irritable bowel syndrome patients. Microbiome. 4:1–9. doi: 10.1186/s40168-016-0191-0

Nash AK, Auchtung TA, Wong MC, Smith DP, Gesell JR, Ross MC, Stewart CJ, Metcalf GA, Muzny DM, Gibbs RA, et al. 2017. The gut mycobiome of the Human Microbiome Project healthy cohort. Microbiome. 5:153. doi: 10.1186/s40168-017-0373-4

Neef A, Sanz Y. 2013. Future for probiotic science in functional food and dietary supplement development. Curr Opin Clin Nutr Metab Care. 16:679–687. doi: 10.1097/MCO.0b013e328365c258

Nguyen VH, Lavenier D. 2009. PLAST: Parallel local alignment search tool for database comparison. BMC Bioinformatics. 10:1–13. doi: 10.1186/1471-2105-10-329

Nieves-Ramírez ME, Partida-Rodríguez O, Laforest-Lapointe LA, Reynolds LA, Brown EM, Morien E, Parfrey LW, Jin M, Walter J, Torres J, et al. 2018. Asymptomatic intestinal colonization with protist *Blastocystis* is strongly associated with distinct microbiome ecological patterns. Host-Microbe Biol. 3:1–18. doi: 10.1128/mSystems.00007-18

Nourrisson C, Scanzi J, Pereira B, NkoudMongo C, Wawrzyniak I, Cian A, Viscogliosi E, Livrelli V, Delbac F, Dapoigny M, Poirier P. 2014. *Blastocystis* is associated with decrease of fecal microbiota protective bacteria: Comparative analysis between patients with irritable bowel syndrome and control subjects. PLoS One. 9(11):e111868. doi: 10.1371/journal.pone.0111868

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27:824–834. doi: 10.1101/gr.213959.116

Olm MR, West PT, Brooks B, Firek BA, Baker R, Morowitz MJ, Banfield JF. 2019. Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. Microbiome. 7:1–16. doi: 10.1186/s40168-019-0638-1

Parfrey LW, Walters WA, Knight R. 2011. Microbial eukaryotes in the human microbiome: Ecology, evolution, and future directions. Front Microbiol. 2:1–6. doi: 10.3389/fmicb.2011.00153

Parks DH, Tyson GW, Hugenholtz P, Beiko RG. 2014. STAMP: Statistical analysis of taxonomic and functional profiles. Bioinformatics. 30:3123–3124. doi: 10.1093/bioinformatics/btu494

Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, et al. 2019. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 176:649-662.e20. doi: 10.1016/j.cell.2019.01.001

Paulos S, Köster PC, de Lucio A, Hernández-de-Mingo M, Cardona GA, Fernández-Crespo JC, Stensvold CR, Carmena D. 2018. Occurrence and subtype distribution of *Blastocystis* sp. in humans, dogs and cats sharing household in northern Spain and assessment of zoonotic transmission risk. Zoonoses Public Health. 65:993–1002. doi: 10.1111/zph.12522

Puthia MK, Sio SWS, Lu J, Tan KSW. 2006. *Blastocystis ratti* induces contact-independent apoptosis, F-actin rearrangement, and barrier function disruption in IEC-6 cells. Infect Immun. 74:4114–4123. doi: 10.1128/IAI.00328-06

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 464:59–65. doi: 10.1038/nature08821

Ramírez JD, Sánchez A, Hernández C, Flórez C, Bernal MC, Giraldo JC, Reyes P, López MC, García L, Cooper PJ, et al. 2016. Geographic distribution of human Blastocystis subtypes in South America. Infect Genet Evol. 41:32–35. doi: 10.1016/j.meegid.2016.03.017

Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. 2016. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. 469:967–977. doi: 10.1016/j.bbrc.2015.12.083

Roberts T, Barratt J, Harkness J, Ellis J, Stark D. 2011. Comparison of microscopy, culture, and conventional polymerase chain reaction for detection of *Blastocystis* sp. in clinical stool samples. Am J Trop Med Hyg. 84:308–312. doi: 10.4269/ajtmh.2011.10-0447

Roberts T, Stark D, Harkness J, Ellis J. 2014. Update on the pathogenic potential and treatment options for *Blastocystis* sp. Gut Pathog. 6:1–9. doi: 10.1186/1757-4749-6-17

Rodríguez JM, Murphy K, Stanton C, Ross RP, Kober OI, Juge N, Avershina E, Rudi K, Narbad A, Jenmalm MC, et al. 2015. The composition of the gut microbiota throughout life, with an emphasis on early life. Microb Ecol Heal Dis. 26:26050. doi: 10.3402/mehd.v26.26050

Rosa BA, Supali T, Gankpala L, Djuardi Y, Sartono E, Zhou Y, Fischer K, Martin J, Tyagi R, Bolay FK, et al. 2018. Differential human gut microbiome assemblages during soil-transmitted helminth infections in Indonesia and Liberia. Microbiome. 6:1–19. doi: 10.1186/s40168-018-0416-5

El Safadi D, Gaayeb L, Meloni D, Cian A, Poirier P, Wawrzyniak I, Delbac F, Dabboussi F, Delhaes L, Seck M, et al. 2014. Children of Senegal River Basin show the highest prevalence of *Blastocystis* sp. ever observed worldwide. BMC Infect Dis. 14:1–11. doi: 10.1186/1471-2334-14-164

Samuel BS, Hansen EE, Manchester JK, Coutinho PM, Henrissat B, Fulton R, Latreille P, Kim K, Wilson RK, Gordon JI. 2007. Genomic and metabolic adaptations of *Methanobrevibacter smithii* to the human gut. Proc Natl Acad Sci U S A. 104:10643–10648. doi: 10.1073/pnas.0704189104

Sankaranarayanan K, Ozga AT, Warinner C, Tito RY, Obregon-tito AJ, Xu J, Gaffney PM, Jervis LL, Stephens L, Foster M, et al. 2016. Gut microbiome diversity among Cheyenne and Arapaho individuals from western Oklahoma. 25:3161–3169. doi: 10.1016/j.cub.2015.10.060

Sato J, Kanazawa A, Ikeda F, Yoshihara T, Goto H, Abe H, Komiya K, Kawaguchi M, Shimizu T, Ogihara T, others. 2014. Gut dysbiosis and detection of “live gut bacteria” in blood of Japanese patients with type 2 diabetes. Diabetes Care. 37:2343–2350. doi: 10.2337/dc13-2817

Savage DC. 2001. Microbial biota of the human intestine: a tribute to some pioneering scientists. Curr Issues Intest Microbiol. 2 1:1–15.

Scanlan PD, Marchesi JR. 2008. Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. ISME J. 2:1183–1193. doi: 10.1038/ismej.2008.76

Scanlan PD, Stensvold CR, Cotter PD. 2015. Development and application of a *Blastocystis* subtype-specific PCR assay reveals that mixed-subtype infections are common in a healthy human population. Appl Environ Microbiol. 81:4071–4076. doi: 10.1128/AEM.00520-15

Scanlan PD, Stensvold CR, Rajilić-Stojanović M, Heilig HGHJ, De Vos WM, O'Toole PW, Cotter PD. 2014. The microbial eukaryote *Blastocystis* is a prevalent and diverse member of the healthy human gut microbiota. FEMS Microbiol Ecol. 90:326–330. doi: 10.1111/1574-6941.12396

Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, Rostron T, Cerundolo V, Pamer EG, Abramson SB, others. 2013. Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. Elife. 2:e01202. doi: 10.7554/eLife.01202

Schluter J, Foster KR. 2012. The Evolution of Mutualism in Gut Microbiota Via Host Epithelial Selection. PLOS Biol. 10:e1001424. doi: 10.1371/journal.pbio.1001424

Schmidt TSB, Raes J, Bork P. 2018. The Human Gut Microbiome: From Association to Modulation. Cell. 172:1198–1215. doi: 10.1016/j.cell.2018.02.044

Seekatz AM, Aas J, Gessert CE, Rubin TA, Saman DM, Bakken JS, Young VB. 2014. Recovery of the gut microbiome following fecal microbiota transplantation. mBio. 5(3):e00893-14. doi: 10.1128/mBio.00893-14

Shukla G, Devi P, Sehgal R. 2008. Effect of *Lactobacillus casei* as a probiotic on modulation of Giardiasis. Dig Dis Sci. 53:2671–2679. doi: 10.1007/s10620-007-0197-3

Sibbald SJ, Archibald JM. 2017. More protist genomes needed. Nat Ecol Evol. 1(5):145. doi: 10.1038/s41559-017-0145

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212. doi: 10.1093/bioinformatics/btv351.

Song Y, Liu C, Finegold SM. 2004. Real-time PCR quantitation of *Clostridia* in feces of autistic children. Appl Environ Microbiol. 70:6459–6465. doi: 10.1128/AEM.70.11.6459-6465.2004

Stairs CW, Leger MM, Roger AJ. 2015. Diversity and origins of anaerobic metabolism in mitochondria and related organelles. Philos Trans R Soc B Biol Sci. 370. doi: 10.1098/rstb.2014.0326

Stechmann A, Hamblin K, Pérez-Brocal V, Gaston D, Richmond GSS, van der Giezen M, Clark CG, Roger AJ. 2008. Organelles in *Blastocystis* that Blur the Distinction between Mitochondria and Hydrogenosomes. Curr Biol. 18:580–585. doi: 10.1016/j.cub.2008.03.037

Steinegger M, Salzberg SL. 2020. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. bioRxiv.:2020.01.26.920173. doi: 10.1101/2020.01.26.920173

Stensvold CR, Alfellani MA, Nørskov-Lauritsen S, Prip K, Victory EL, Maddox C, Nielsen H V., Clark CG. 2009. Subtype distribution of *Blastocystis* isolates from synanthropic and zoo animals and identification of a new subtype. Int J Parasitol. 39:473–479. doi: 10.1016/j.ijpara.2008.07.006

Stensvold CR, Clark CG. 2016. Current status of *Blastocystis*: A personal view. Parasitol Int. 65:763–771. doi: 10.1016/j.parint.2016.05.015

Stensvold CR, Clark CG. 2020. Pre-empting Pandora's Box: *Blastocystis* Subtypes Revisited. Trends Parasitol. 36:229–232. doi: 10.1016/j.pt.2019.12.009

Stensvold CR, van der Giezen M. 2018. Associations between Gut Microbiota and Common Luminal Intestinal Parasites. Trends Parasitol. 34:369–377. doi: 10.1016/j.pt.2018.02.004

Stensvold CR, Lebbad M, Hansen A, Beser J, Belkessa S, O'Brien Andersen L, Clark CG. 2020. Differentiation of *Blastocystis* and parasitic archamoebids encountered in untreated wastewater samples by amplicon-based next-generation sequencing. Parasite Epidemiol Control. 9:e00131. doi: 10.1016/j.parepi.2019.e00131

Stensvold CR, Suresh GK, Tan KSW, Thompson RCA, Traub RJ, Viscogliosi E, Yoshikawa H, Clark CG. 2007. Terminology for *Blastocystis* subtypes – a consensus. Trends Parasitol. 23:93–96. doi: 10.1016/j.pt.2007.01.004

Stensvold CR, Tan KSW, Clark CG. 2020. Blastocystis. Trends Parasitol. 36(3):315-316. doi: 10.1016/j.pt.2019.12.008

Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, Liachko I, Snelling TJ, Dewhurst RJ, Walker AW, et al. 2018. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nat Commun. 9:1–11. doi: 10.1038/s41467-018-03317-6

Suau A, Bonnet R, Sutren M, Godon JJ, Gibson GR, Collins MD, Doré J. 1999. Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. Appl Environ Microbiol. 65:4799–807.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 23:1282–1288.

doi: 10.1093/bioinformatics/btm098

Suzuki Y, Nishijima S, Furuta Y, Yoshimura J, Suda W, Oshima K, Hattori M, Morishita S. 2019. Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. Microbiome. 7:1–16. doi: 10.1186/s40168-019-0737-z

Terveer EM, van Gool T, Ooijevaar RE, Sanders IMJG, Boeije-Koppenol E, Keller JJ, Bart A, Kuijper EJ, Vendrik KEW, Ooijevaar R, others. 2019. Human transmission of *Blastocystis* by fecal microbiota transplantation without development of gastrointestinal symptoms in recipients. Clin Infect Dis. pii: ciz1122. doi: 10.1093/cid/ciz1122

The Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature. 486:207–214. doi: 10.1038/nature11234.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14:178–192. doi: 10.1093/bib/bbs017

Thursby E, Juge N. 2017. Introduction to the human gut microbiota. Biochem J. 474:1823–1836. doi: 10.1042/BCJ20160510

Tito RY, Chaffron S, Caenepeel C, Lima-Mendez G, Wang J, Vieira-Silva S, Falony G, Hildebrand F, Darzi Y, Rymenans L, et al. 2018. Population-level analysis of *Blastocystis* subtype prevalence and variation in the human gut microbiota. Gut. 68(7):1180-1189. doi: 10.1136/gutjnl-2018-316106

Tremaroli V, Karlsson F, Werling M, Ståhlman M, Kovatcheva-Datchary P, Olbers T, Fändriks L, Le Roux CW, Nielsen J, Bäckhed F. 2015. Roux-en-Y Gastric Bypass and Vertical Banded Gastroplasty Induce Long-Term Changes on the Human Gut Microbiome Contributing to Fat Mass Regulation. Cell Metab. 22:228–238. doi: 10.1016/j.cmet.2015.07.009

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 12:902–903. doi: 10.1038/nmeth.3589

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 444:1027–1031.doi: 10.1038/nature05414

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, JF B. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 428:37–43. doi: 10.1038/nature02340

Vangay P, Johnson AJ, Ward TL, Al-Ghalith GA, Shields-Cutler RR, Hillmann BM, Lucas SK, Beura LK, Thompson EA, Till LM, et al. 2018. US Immigration Westernizes the Human Gut Microbiome. Cell. 175:962-972.e10. doi: 10.1016/j.cell.2018.10.029

Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, Jia W, Cai S, Zhao L. 2012. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. ISME J. 6:320–329. doi: 10.1038/ismej.2011.109

Watts GS, Thornton JE, Youens-Clark K, Ponsero AJ, Slepian MJ, Menashi E, Hu C, Deng W, Armstrong DG, Reed S, et al. 2019. Identification and quantitation of clinically relevant microbes in patient samples: Comparison of three k-mer based classifiers for speed, accuracy, and sensitivity. PLoS Comput Biol. 15:1–27. doi: 10.1371/journal.pcbi.1006863

Wawrzyniak I, Courtine D, Osman M, Hubans-Pierlot C, Cian A, Nourrisson C, Chabe M, Poirier P, Bart A, Polonais V, et al. 2015. Draft genome sequence of the intestinal parasite *Blastocystis* subtype 4-isolate WR1. Genomics Data. 4:22–23. doi: 10.1016/j.gdata.2015.01.009

West PT, Probst AJ, Grigoriev I V., Thomas BC, Banfield JF. 2018. Genome-reconstruction for eukaryotes from complex natural microbial communities. Genome Res. 28:569–580. doi: 10.1101/gr.228429.117

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20:257. doi: 10.1186/s13059-019-1891-0

Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, et al. 2011. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science. 334(6052):105-8. doi: 10.1126/science.1208344

Wu Z, Mirza H, Tan KSW. 2014. Intra-Subtype Variation in Enteroadhesion Accounts for Differences in Epithelial Barrier Disruption and Is Associated with Metronidazole Resistance in *Blastocystis* Subtype-7. PLoS Negl Trop Dis. 8:27–31. doi: 10.1371/journal.pntd.0002885

Wylezich C, Belka A, Hanke D, Beer M, Blome S, Höper D. 2019. Metagenomics for broad and improved parasite detection: a proof-of-concept study using swine faecal samples. Int J Parasitol. 49:769–777. doi: 10.1016/j.ijpara.2019.04.007

Xu Z, Knight R. 2015. Dietary effects on human gut microbiome diversity. Br J Nutr. 113 Suppl:S1–S5. doi: 10.1017/S0007114514004127

Yakoob J, Jafri W, Jafri N, Khan R, Islam M, Beg MA, Zaman V. 2004. Irritable bowel syndrome: in search of an etiology: role of *Blastocystis hominis*.Yakoob J, editor. Am J Trop Med Hyg. 70:383–385.

Yason JA, Liang YR, Png CW, Zhang Y, Tan KSW. 2019. Interactions between a pathogenic *Blastocystis* subtype and gut microbiota: In vitro and in vivo studies. Microbiome. 7:1–13. doi: 10.1186/s40168-019-0644-3

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, et al. 2012. Human gut microbiome viewed across age and geography. Nature. 486:222–227. doi: 10.1038/nature11053

Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell. 178:779–794. doi: 10.1016/j.cell.2019.07.010

Zhao G, Hu X, Liu T, Hu R, Yu Z, Yang W, Wu Y, Yu S, Song J. 2017. Molecular characterization of *Blastocystis* sp. in captive wild animals in Qinling Mountains. Parasitol Res. 116:2327–2333. doi: 10.1007/s00436-017-5506-y

Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, Mujagic Z, Vila AV, Falony G, Vieira-Silva S, et al. 2016. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science. 352:565–569. doi: 10.1126/science.aad3369