Meta-Analysis of *Blastocystis* Subtype Distribution and Prevalence Across Hosts and Geographies: Advancing Knowledge Through Data Analysis and Visualisation

Abi Girl Sanda

School of Biosciences

University of Kent

Thesis for the MSc in Microbiology

2024

Supervisors: Anastasious Tsaousis and Marta Farre Belmonte

Dry-lab Project

University of Kent

DECLARATION

I declare that the wording of this report is completely my own. No part of the report has been copied by scientific journals, web sites or any other sources. For a detailed statement on plagiarism the student is referred to the guidelines for preparing project reports.

Signature of the student:

Abi Girl Sanda

Date: 17/09/2024

Acknowledgements

I would like to thank Dr. Anastasios Tsaousis and every member from his lab for his help and constant support during the duration of this project. I also want to extend this gratitude to Eleni Gentekaki for also guiding me through my research and data collection. Finally, I would also like to thank my friends and family for the moral support and for their massive help in the proofreading process.

Abstract

Blastocystis is a controversial parasite found in the intestinal tract of vertebrate hosts. The diversity of Blastocystis has allowed for scientists to classify numerous subtypes (STs) over the past few years. Interdisciplinary studies have significantly advanced our understanding of the parasite, elucidating some of its behavioural patterns and genetic characteristics. Over the last decade there has not been a study focusing on the published and submitted ST. Such studies are informative as they provide a comprehensive understanding of ST prevalence and distribution across countries and hosts. This research is a meta-analysis of 18S rRNA sequences submitted to GenBank, whereby STs are classified according to country/continent and host characteristics such as class and diet. RStudio was used to provide visual representations of trends and patterns detected hence allowing easy recognition of subtype prevalence patterns in each country already studied.

The analysis revealed subtype-host specificity, with ST1 being the most prevalent in Egypt, ST2 in Senegal and ST3 in Denmark, Mammalian hosts exhibited the highest diversity of subtypes, including novel ones such as ST24a-c, primarily in Europe. Statistical tests demonstrated a significant difference between subtype and host/environmental characteristics. Predictive modelling was attempted, however data limitations made the results unreliable. These findings provide valuable insights into the geographical distribution and host specificity of *Blastocystis* subtypes, with potential public health implications presented by the presence of pathogenic subtypes in environmental samples such as wastewater and soil.

Abbreviations Used

ST - Subtype

STs - Subtypes

CA - Correspondence Analysis

MCA - Multiple Correspondence Analysis.

EDA - Exploratory Data Analysis

GLM - Generalised Linear Models

AIC - Akaike Information Criterion

H(0) - Null hypothesis

H(1) - alternative hypothesis

Df - degree of freedom

PCR - Polymerase chain reaction

IBS - Irritable bowel syndrome

Introduction

Overview of *Blastocystis*

Blastocystis is a microscopic parasite that can be found in the digestive tract of both humans and animals which currently has over 17 confirmed different subtypes. It exhibits multiple reproductive strategies, including schizogeny and endodyogeny, as shown in **Figure 2**, which illustrates distinct morphological stages of the organism. Recent advancements in research surrounding *Blastocystis* have facilitated the expansion of our knowledge of the microorganism. However, for decades since its discovery, *Blastocystis* was thought to be a non-pathogenic fungus-like organism (5). It was not until the 20th century that scientists finally recognised it as a protozoan parasite and its involvement in gastrointestinal symptoms was suggested (1), after its detection in faecal samples. Despite the years of controversy, it is now widely accepted that *Blastocystis* is a stramenopile (4). **Figure 1** illustrates the life cycle of *Blastocystis*, including its main morphological forms and transmission routes between humans and animal hosts. Interestingly, members of this group possess flagella, however, *Blastocystis* does not present any tubular hairs (6). It was suggested that this was due to a secondary loss of these characteristics. Stramenopiles are known for thriving in environments varying from as parasites on plants to soils and, as seen with *Blastocystis*, in the guts of humans(8). One Health is a concept that highlights the need for interdisciplinary cooperation to address public health issues and it states that the health of humans, animals and the environment is interconnected. With this knowledge, the research surrounding *Blastocystis* continues and every day new insights into the parasites are discovered.

Pathogenicity

As aforementioned, *Blastocystis* is found in the gut microbiome, however, its pathogenicity is still an inconclusive topic. It has been speculated that *Blastocystis* may be related to certain

gastrointestinal symptoms, disorders (which have been named Blastocystosis) and conditions such as irritable bowel syndrome (IBS)(15)(13). This has however been a topic of debate as some studies conducted have put the pathogenicity of *Blastocystis* into question. In a study conducted, scientists looked at faecal smears from two separate groups made up of asymptomatic and symptomatic patients (control) (16). No statistical significance was found between the two groups, they also did not find an association between faecal the presence of *Blastocystis* and fecal leukocytes, which can be used as indicators of intestinal inflammation. After considering the clinical profile of the people infected with the parasite and those who weren't, they also concluded that they were similar. These conclusions all point to *Blastocystis* being non-pathogenic, however, they also found that in symptomatic patients, there was a high concentration of *Blastocystis* along with other non-pathogenic protozoa. The later discovery that some *Blastocystis* strains are more infectious than others, could explain these results(17). It has in fact been hypothesised that proteases produced by *Blastocystis* function as virulence factors allowing host immune evasion, which may explain why some strains are more pathogenic than others(18).

The possibility of its pathogenicity together with the fact that, like many other parasites, *Blastocystis* is transmitted between individuals via the faecal-oral route(14), puts underdeveloped communities at more risk. Individuals from developing countries are more susceptible to the spreading of infectious disease, which is also a leading cause of death(19), due to the tendency of having inadequate living conditions and less knowledge of health and how to prevent disease spread. A lot of these communities simply lack proper access to healthcare providers. In a study conducted in Indonesia, *Blastocystis* was found in diarrhoea-experiencing patients suffering from HIV/AIDS which is a condition that unfortunately many people suffer from, especially in poorer countries(21). In Colombia, parasitism was studied in preschool children from a rural area(20). 36.4% had *Blastocystis* and the researchers suggested an association between the presence of this parasite and close contact with animals. Given the fact that sharing living quarters with domestic and non-domestic animals is common practice in certain countries (especially poorer ones)(13),

this also increases infection risk(22). Some treatments have been suggested, however, there are multiple cases of failure and resistance. Understanding *Blastocystis* pathogenicity in its entirety is imperative to understand the extent of its symptoms in humans and begin to provide people who may suffer because of it with appropriate treatment.

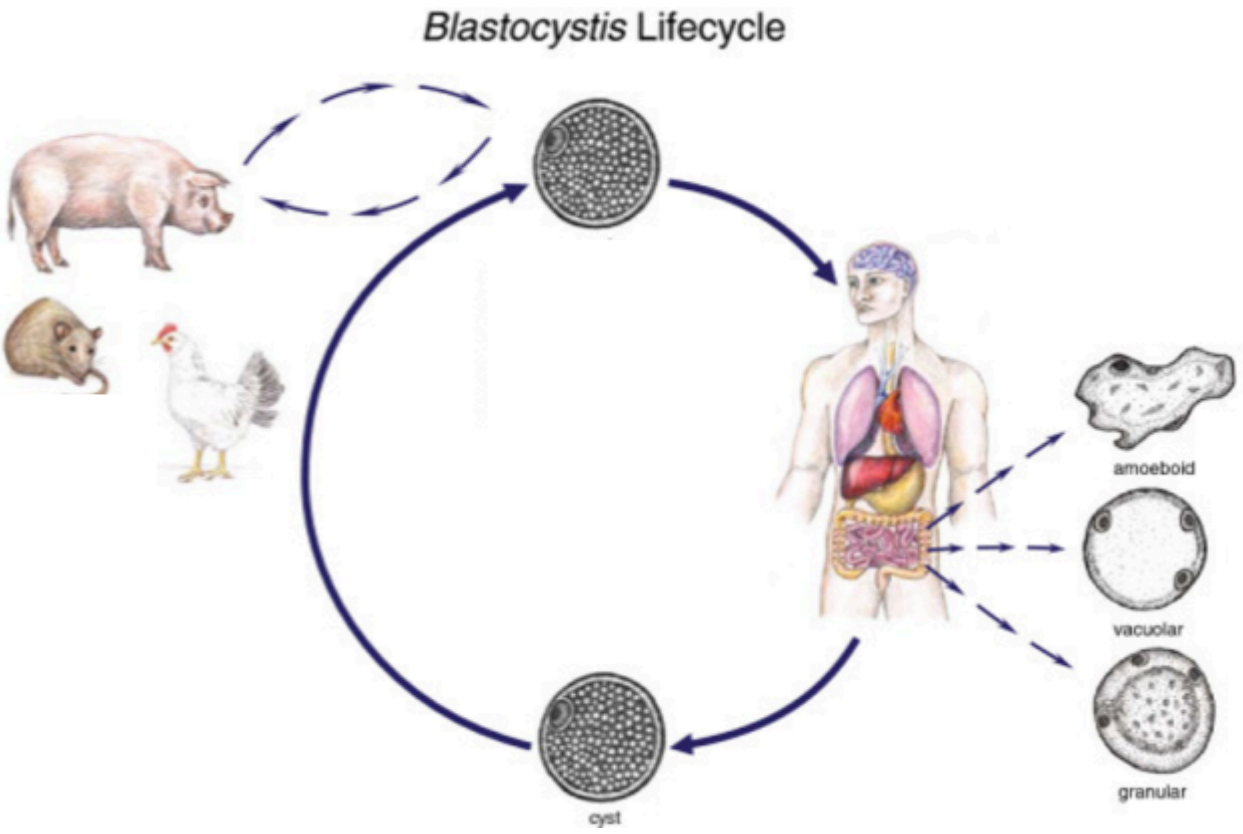


Figure 1. Image taken from Roberts, T et al.'Update on the pathogenic potential and treatment options for *Blastocystis Sp*.'. Figure illustrating *Blastocystis* life cycle (on the right) and potential animal hosts (on the left). When ingested via contaminated sources such as water, food or direct fecal-oral transmission, *Blastocystis* at its cyst stage can lead to infection. Once in the intestine, excystation takes place, during which the cyst develops into vacuolar forms.The cycle carries on through encystation and the cysts are later shed in feces. Rats, pigs and chickens represent possible animal hosts.

Molecular Advancements

Many molecular advances have been made since the discovery of *Blastocystis* which have helped scientists understand its structure and homogeneity. When *Blastocystis* was first recognised as a protozoan, electron microscopy was used to study its structure in much

greater detail than what had been done before(2), however, this technique comes with its limitations. Some researchers state that in epidemiology, PCR should always be used along with other methods(9). This is due to microscopy-negative samples later being found as positive when PCR was used. Microscopy is still widely used in pathology and *Blastocystis* analysis, sometimes due to the elevated cost of other equipment.

Due to the morphological variations that happen within the faecal samples(3), recognising differences in *Blastocystis* strains has been difficult. However, the development of various laboratory techniques allowed for the discovery of even more details, including the genetic and antigenic heterogeneity of *Blastocystis,* which is illustrated by the phylogenetic clades and subtypes shown in **Figure 3**(4). Isoenzyme analysis allowed the identification of *Blastocystis* strain differences(11) through the use of agarose gels as a medium to exploit the mobility of intracellular enzymes to evaluate cell purity and identify the species of origins of cell lines(10). Immunoblotting is another technique that supported these findings; it revealed that only specific strains of *Blastocystis,* particularly those associated with symptomatic infections, express a 29-30 kDa protein that reacts with cytotoxic monoclonal antibody mAb 1D5, suggesting a protein role in pathogenicity(12). Finally, another research technique that was critical to the placing of different groups within *Blastocystis sp.* was the random amplification of polymorphic DNA (RAPD) with the use of different PCR primers. The discovery that the amplification of different strains was not possible with the same primers, led to the sub-grouping of the parasite(7). These are just some examples of how scientists came to the conclusion that there was more to *Blastocystis* when it came to the different strains.

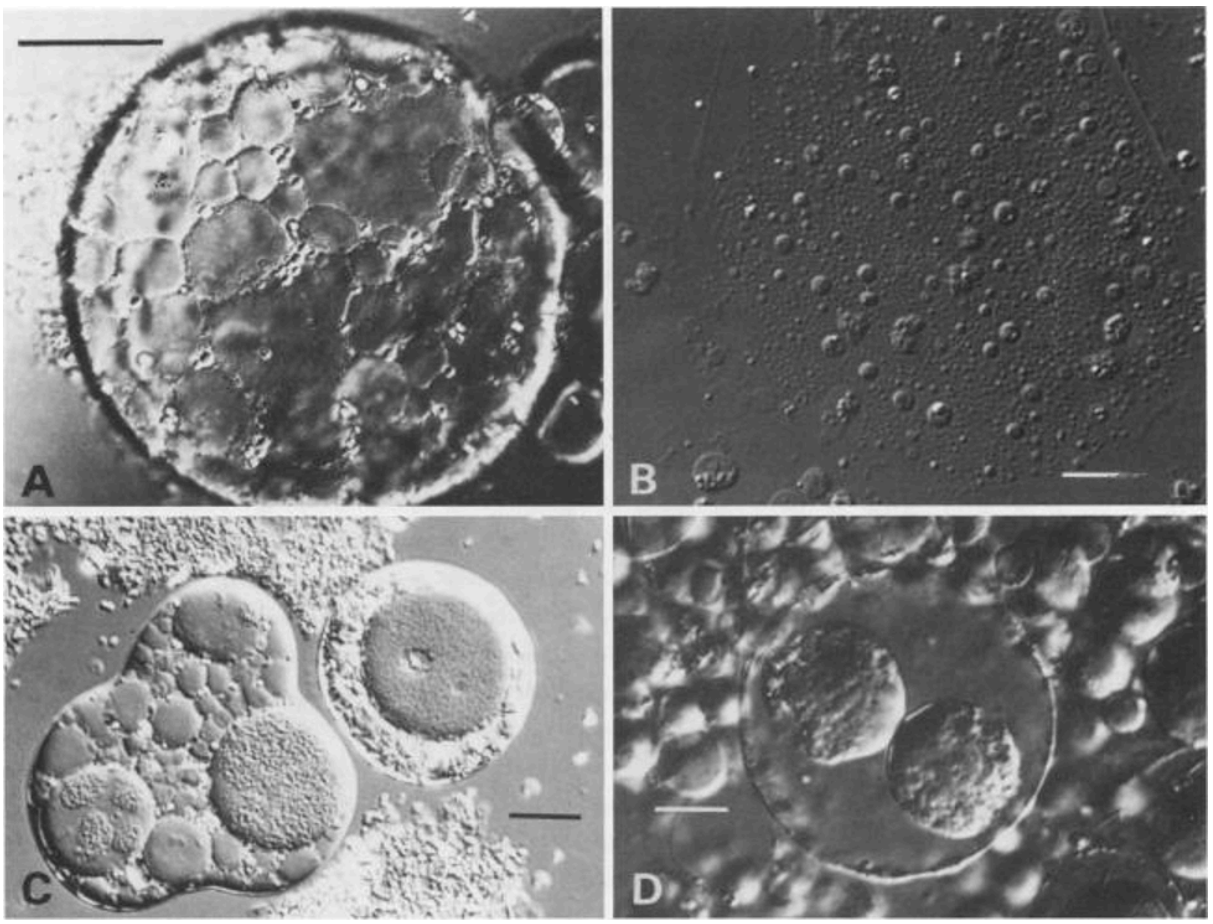


Figure 2. *B. hominis* DIC optics image representing Schizogony from Zierdt's lab(2). A) Schizont filled with progeny, or daughter cells. (B) Ruptured schizont has released small, condensed, brownish progeny characteristic of only rare strains wherein a higher proportion of cells undergo this asexual division, resulting in smaller progeny, perhaps due to competitive nutrition. Bar, 10 μm. (C) Schizont with progeny in varying stages of maturity. Bar, 5 μm. (D) Cell in endodyogeny, the creation of two progeny within the parent cell. Bar, 5 um.(Descriptions from original article)

Subtype Identification, Distribution and Importance

The knowledge and research continued to grow and it was increasingly agreed that *Blastocystis* subtypes needed to be divided into smaller subgroups, due to their apparent sequence heterogeneity. However, it wasn't till 2007 that scientists reached a consensus for *Blastocystis* subtype terminology, thanks to a publication by Stensvold et al.(24). In this article they made a significant contribution by developing a standardised method of subtype identification and classification. They had samples from rats, humans and reptiles and amplified the sample using eukaryote-specific primers A and B. The amplification of SSU

rRNA gene coding regions resulted in fragments of ~1.8kb in length, which was the expected size. After the elution and cloning of the fragments produced into a pCR 2.1-TOPO vector, they were compared to the clones sequenced from the isolate. These were always identical. The new SSU rRNA gene sequences had a similarity of minimum 82.6% to the homologous sequences of other isolates. They also created a rooted maximum-likelihood tree with the backing of Bayesian posterior probabilities to identify clades called groups I to VII. Therefore, it was stated that groups I to VII correspond to ST1, ST5, ST3, ST7, ST6, ST4 and ST2. This not only confirmed the presence of multiple *Blastocystis* subtypes, but the standardised protocol created by them has been widely accepted since then and it has allowed for a more homogenous and clearer way to report and compare findings across different populations and geographic regions.

However, it is to be noted that subtypes (STs) were not ‘created’ by Tensvold et al., but just standardised. In the late 90s, scientists identified and proposed two subtypes - which were later identified as ST1 and ST2 - from 8 different strains of *Blastocystis* (7). Another study by Noël et al. in 2005(25) used the comparative analysis of SSU rRNA-and EF-1α-based trees methods as well as Bayesian analysed phylogenetic trees to identify 7 distinct groups and other potential clades.

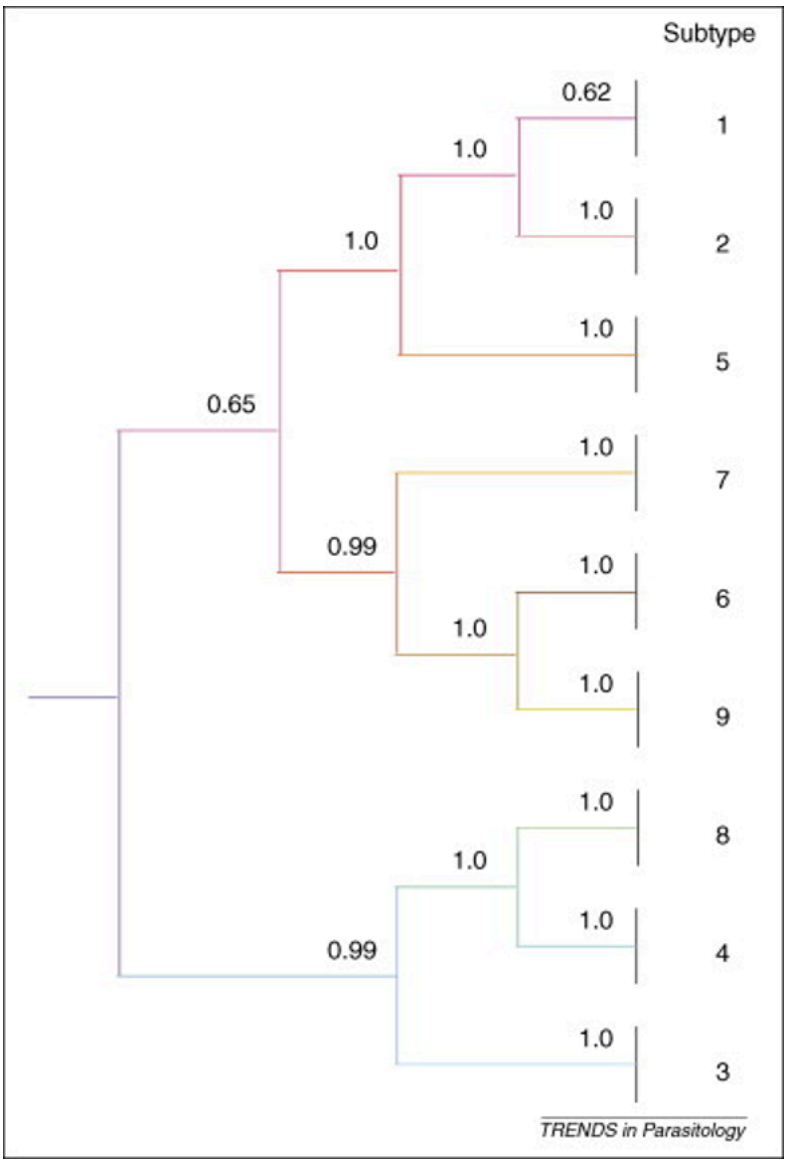


Figure 3. Proposed clades and subtype terminology for *Blastocystis,* based on phylogenetic analysis of SSU rRNA gene sequences. Clade structure follows Noël et al.(2005), with subtype designation adopted from the consensus proposed by Stensvold et al.(2007). Reproduced from Stensvold et al. (2007)(24). Numerical values on branches represent Bayesian posterior probabilities, indicating the level of statistical support.

The two aforementioned studies, also gave an insight into subtype specificity and distribution. Primers used to subtype strains from human hosts did not work on non-human hosts, not only suggesting the existence of ST1 and ST2, but also that the subtypes found in humans are potentially different from the ones found in animals(7). It is now widely known that the distribution of *Blastocystis* STs varies significantly across regions and hosts. ST1, ST2, ST3 and ST4 have been found to be the most common in human hosts(24). This

varies depending on the region. ST3 tends to be the most frequently present across various studies, this is especially the case in Africa and Asia. To be more specific, studies have found allele 34, 36 and 37 to be the most common alleles found in *Blastocystis* ST3, with allele 34 appearing the most across all continents(27). Other studies conducted in India, South America and Iran have identified allele 34 as the most present in the samples they used(26)(28). ST4 is often referred to as the European ST, this is due to its high prevalence in the continent, but quite rare in others(24). These findings all point to *Blastocystis* spread being heavily influenced by geographical or host-related factors.

The subtype diversity found in both humans and animals also highlighted the zoonotic potential of *Blastocystis*. While certain STs such as ST10, ST14 and ST17 seem to be animal-specific(30)(31). The presence of ST5, ST6 and ST7 has been related to samples belonging to animal hosts such as pigs and birds as well as human hosts(25)(29). ST1-5 and ST8 were also identified as zoonotic from samples collected from zoos in South-west China(30). This is an area of concern as it suggests that close contact with animals increases exposure to different *Blastocystis* STs. This is not only a concern for communities in underdeveloped countries (as suggested earlier), but also for anyone that may own these animals as pets or come into contact with them in public spaces such as petting zoos.

With the knowledge and acceptance of the existence of different *Blastocystis* subtypes, it is easier to also address related topics such as clinical implications of the parasite. *Blastocystis* infections are often subtype-dependent, meaning that certain STs can be related to human diseases more than others. Because ST4 has frequently been identified in human populations (33), especially in countries such as Denmark and the UK, one could assume that this specific ST could play a part in human infection. However, studies are yet to decisively link ST4 to either human or animal diseases. In fact, a study conducted with mice did not find any harmful effects on mice when they were colonised with *Blastocystis* ST4. Surprisingly, ST4 colonisation seemed to promote T helper 2 (Th2) response defined by

interleukin (IL)-5 and IL-13 cytokine production, and T regulatory (Treg) induction from colonic lamina propria in normal healthy mice(34). Other studies have suggested that ST1 may be pathogenic (17)(35), and others that allele 34 of ST3, a commonly found subtype, is connected to urticaria (32), but again no certain links have been made. Although incontestable conclusions are still lacking, ongoing research continues to shape our understanding of the clinical impact of the parasite, which in turn may also inform improved strategies for diagnosis, treatment and control.

Research Gaps and Future Directions

Gaps in our understanding of *Blastocystis* are highlighted by several still unanswered questions. One of the main questions concerns *Blastocystis'* geographical distribution; why are certain subtypes more prevalent in specific regions. The 'European' ST4 is a prime example, its frequency in the continent has been recognised multiple times, but this subtype is rarely seen in African or Asian countries. This differs from ST3 behaviour as it is more likely to be detected globally. This variability paired with the fact that new subtypes are still being identified, encourages further investigations in order to detect the explanation for this behaviour. Areas of interest could be genetic variability, socially-based geographical factors or even ecological.

Another gap in our understanding is the extent and explanation for host specificity. As discussed earlier in this introduction, certain subtypes, like ST4, are generally related to human samples, whereas others, such as ST5, ST6 and ST17 tend to be more related to animal samples. We do not know exactly why certain subtypes seem to have a higher affinity for certain hosts and less for others. In addition, information surrounding the zoonotic potential of *Blastocystis* subtypes is also limited. For example, regardless of the hosts they are mostly related to, ST4 and ST5 have in fact previously been detected in both human and animal samples. This indicates that there are still zoonotic pathways that we are still

unaware of. ST4's potential link to infection also sheds light on possible transmission routes and interactions.

In accordance with One Health, answering questions surrounding environmental factors are also imperative for a full understanding of *Blastocystis'* behaviour. Bodies of water have the capacity of acting as reservoirs for microorganisms, in fact ST1-ST8, ST10, ST23 and ST26 have already been identified in different bodies of water across the globe(36). This may also affect subtype prevalence - are countries surrounded by bodies of water more susceptible to certain subtypes? Having an understanding of which environmental conditions, such as potable water contamination or availability, can contribute to a more controlled infection spread in different populations. Again, new subtypes are continuously being found, and environmental factors could also offer an explanation for this.

Aims and Objectives of this Thesis

This thesis is heavily influenced by the findings of Alfellani et al. (2013)(37). In their molecular epidemiology study, they aimed to investigate the distribution of different *Blastocystis* subtypes across various geographic regions and hosts. The work published by Alfellani et al. did in fact do that by conducting a cross-sectional study in order to show how significantly varied the prevalence of subtypes such as ST1, ST2, ST3 and ST4, was across different countries. The study used a large cohort from multiple continents, including Africa, Europe and Asia, which analysed both animal and human hosts in order to comprehend how subtype prevalence is affected by regional and host-specific factors. Although not the first to do so, this study restated the understanding that certain subtypes are geographically specific, such is the case for ST3 in African communities and ST4 in European populations. ST1 was also found at a high frequency in Nigeria and Libya. Interestingly, ST2 and ST4 were either absent or misrepresented in these countries. This research, along with prior knowledge, inspired questions on the possible genetic composition of the host being the

reason for the subtype specificity and fueled the curiosity surrounding environmental and epidemiological factors in relation to *Blastocystis.*

Alfellani et al.'s publication was inspiring, and with this also came more questions attempting to understand *Blastocystis* even further. For example, this study was conducted over a decade ago, has this reflected on subtype trends and patterns since then resulting in a shift in subtype prevalence? Have novel subtypes been found in the same countries/regions Alfellani had previously explored? In addition, Alfellani et al.'s cohort mainly focused on subtypes from human samples; the extension to animal and environmental samples, such as water or soil, could highlight new information on subtype prevalence, infection risk factors and zoonotic transmission routes.

Building upon these insights, the main objective of this thesis is to explore and evaluate the current distribution of *Blastocystis* subtypes. This will be done using human, animal and environmental samples. The specific aims of this research are as follows:

1. Explore *Blastocystis'* geographic and host subtype specificity and assess whether significant changes can be observed in countries from Africa, Asia and Europe.
2. Analyse the relationship between host factors that could affect subtype distribution These include location, taxonomy and diet.
3. Identifying connections between factors influencing subtype prevalence and their potential for zoonotic transmission and pathogenicity .

By addressing these questions, this thesis seeks to expand on the collective knowledge surrounding *Blastocystis*, especially pertaining to the relationship between subtype distribution and host specificity. This will then also contribute to the comprehension of *Blastocystis* epidemiology.

Method

Literature Search Strategy

The primary database used for this meta-analysis was GenBank, which is a comprehensive public database of nucleotide sequences with supporting bibliographic and biological annotations. All the relevant information was collected from the 'FEATURES' section. To perform the research, specific keywords related to *Blastocystis* were used. These consisted of the word 'Blastocystis' followed by the country of interest, for example when the UK was being researched the search keyword was 'Blastocystis UK'. As this study aimed to collect all the existing data, things such as hosts or dates were not specified.

In the case of samples derived from published studies, this was also accessed (usually through the PubMed code) to gain more context to aid the discussion.
Other search engines were also accessed, these being 'Google Scholar' and 'PubMLST'. These were used to gain further context on the samples submitted in GenBank.

Inclusion and Exclusion Criteria

Samples and studies that did not specify the ST found were deemed as incomplete data and not included in the research. Clones of samples were also excluded to avoid overrepresentation. The samples taken from other studies were also excluded for the same reason.

Data Extraction

A standardised Excel sheet was used to extract the data, the characteristics extracted were as follows: title, author/s, year published, link to the study (if available), country, host, ST and accession code.
When the data was collected directly from the study, it was sometimes impossible to also note the accession code as allocating it to a specific sample was not possible.

Coding and Statistical Analysis

1. Software Environment

Once collected, data from studies was treated as individual data points, including the ones deriving from the same study. This is due to a large number of direct submissions to NCBI that appear to originate from the same study (usually by the presence of the same host, country or authors), but had no way of confirming it. All analysis was done through the use of the RStudio software with the 4.41 R version. Key packages used:

- 'install.packages()' and 'library()': Install and call packages.
- 'ggplot2': visualisation of barplots, mosaic plots, etc.
- 'tidyverse': a collection of R packages aiding data manipulation, exploration and visualisation.
- 'dplyr': Set of functions that allow data to be transformed (row naming or removal etc.).
- 'ca': Used for CA.
- 'sjPlot': models visualisation.
- 'FactoMineR': used to perform MCA.

Full list of packages used available in Appendix I.

2. Data Processing

All the data used was filtered using the 'dplyr' package and percentages were calculated using a simple formula (figure one and two). When calculating counts and percentages, data was first grouped by the desired variable to isolate and analyse the prevalence of subtypes in it. For example, when looking at subtype prevalence in different host classes, subsets were created that focused on specific combinations within the 'class' and the 'subtypes' category. Then these were used to calculate the relative frequency of each subtype within the total occurrences of subtypes in each class. Therefore, referring to the previous

example, once the groups were made the percentage of each subtype in these was calculated.

```
subtype_percentage <- EU_AF_data %>%
  group_by(Subtypes) %>%
  summarise(count = n()) %>%
  mutate(total = sum(count),
         percentage = (count / total) * 100)
```

Figure 4. Example code of data percentage calculation. ‘EU_AF_data’ refers to the name of the name given to the data including all the continents. ‘Subtypes’ refers to the column containing the *Blastocystis sp.* Subtypes found in samples collected. The percentage of each subtype was calculated using the total number of subtypes using the ‘group_by’ function. The final percentage is calculated by dividing the count of each subtype by the total count of subtypes within the same class and then multiplying this by 100.

```
Euro_hd <- Europe_Blasto_data %>%
  filter(Subjects == "Humans")

# Removal of rows containing Subject 'human' as subject
animonly_data <- EU_AF_data %>%
  filter(!grepl("Humans", Subjects))
```

Figure 5. Example code of data selection/manipulation. ‘Euro_Blasto_data’ refers to all the European *Blastocystis sp.* Data collected. Data containing only animal data selected using ‘filter’ command.

3. Exploratory Data Analysis (EDA)

EDA was conducted with the use of the ‘ggplot2’ package. Stacked bar charts were created to visualise subtype prevalence across continents for human, animal and non-living samples. Stacked bar charts were also created to show subtype distribution based on the host family. Animated bar charts were created to show the change in subtype spread between countries. Interactive Sankey diagram showing subtype, continent and host family relationship with counts of each category.

A mosaic plot was created to show the relationship between *Blastocystis sp.* Subtypes and

host diet.

Dot plot created to visualise sample host class spread based on subtypes. World maps created to demonstrate subtype spread across countries 1, 2, 3, 4, 5, 6, 7 and 14.

4. Statistical Analysis

Contingency tables created which show the frequency distribution of two or more categorical variables. The number of observations into each category is displayed to visualise the relationship between these. Chi squared tests carried out using the function ‘chisq.test()’ to assess the significance of the relationship between subtypes and host diet, subtype and host class, subtype and country sample was collected from and subtypes and continent sample was found in. H0 in all 4 tests stated that there is no relationship between the subtype found in the sample and the other variable (Class, Continent, Country, Subtype and Diet) and they are therefore independent of each other.

Prediction model creation was attempted in order to create data that could predict future behaviours in subtype prevalence. The model was created by using subtype as the response model, therefore the variable of which pattern prediction is being attempted. Continent, Class and Diet were used as covariates, therefore as explanatory variables for the prediction model.Unfortunately, the model failed the validation test. A confusion matrix was created to test the model and the predictions created were not reliable.

GLM models fitted using the ‘nnet::multinom()’ function. Subtype was used as the response variable, explanatory variables used were continent, host diet and class. These models were assessed using AIC and p-values. MCA plots were created to further visualise relationships between continent, host family and subtypes as well as continent, host diet and subtypes.
A MCA test can be considered a statistical test, however in this context it was used as an exploratory technique and was carried out to assess whether a pattern and relationship could be detected between categories of multiple variables. The aim was to make the

interpretations of the relationships previously spotted by presenting the categorical data of this study in a lower-dimensional space. Therefore in summary, the closer data points on a MCA plot are, the closer the categories that they represent are to each other.

Results

The original aim of this research was to provide a comprehensive meta-analysis of the majority (if not all) of the data available on *Blastocystis sp.,* however, this was not achieved due to time constraints and changes that had to be done to the data post collection - such as the omission of data sets due to the possible presence of duplicates, which would lead to a steep over-estimation of the data. Consequently, this research will now focus on the relationship between *Blastocystis* subtypes, sample source (including human, animal or environmental origin) and region specificity.

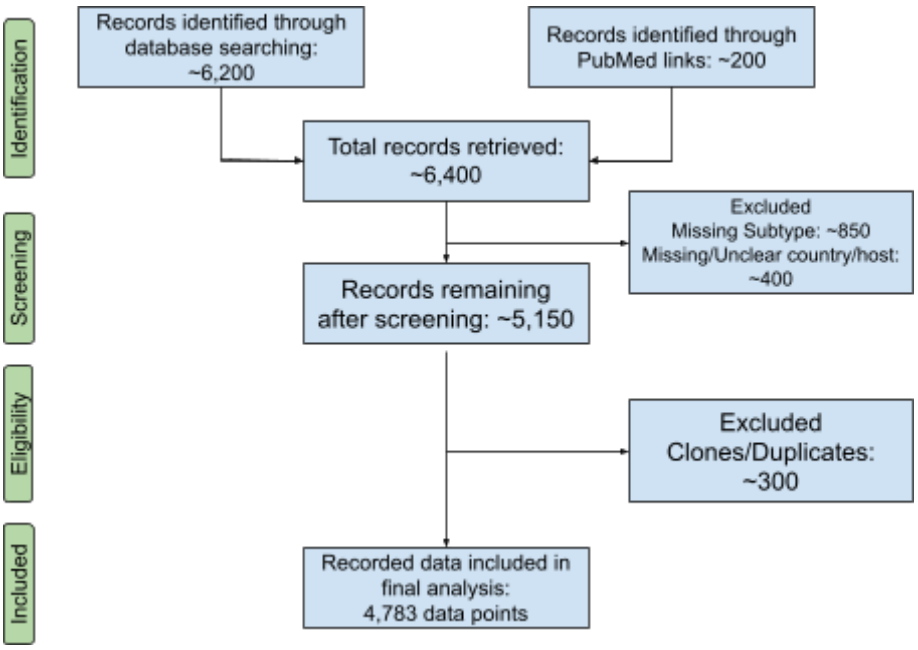


Figure 6. PRISMA flowchart depicting the identification screening, eligibility and inclusion of studies.

44 Studies from Europe, 13 from Africa and 10 from Asia were included in this research. This excludes the untitled direct submissions deposited on GenBank. In total, this section presents the analysis of 4,783 *Blastocystis* data points. Results are grouped thematically by geography, subtype and sample category to provide a coherent narrative.

The names of the countries presented in this research and the percentage of each *Blastocystis* subtype found in them is shown below (**Table Appendix I**).

Global Overview of Subtype Distribution

This section begins by providing an overview of the most commonly found *Blastocystis* subtypes globally. This was done to establish a foundation for a complete understanding of how subtype prevalence differs between regions and sample sources.

It can be seen that certain subtypes are more homogeneously present across all countries. This is especially true for ST1, ST2 and ST3, which were the most widespread, with samples found across various countries and sample types. This is visualised in the heat maps in **Figure 7A-C**, which present subtype frequencies grouped by country. These maps were created to facilitate the visualisation of the data provided in summary tables (**Appendix I**). For instance, the ST1 map shows the subtype as being most prevalent in Egypt (>15%), followed by Guinea, Senegal and Spain, which all have a prevalence above 10%. All the other countries show an ST1 prevalence of around 6% or lower, save for Thailand which sits just below 10%. ST2 has a notable presence in Spain compared to the other countries, however; it is highest in Senegal (~30%). ST2 is also present in Turkey, Egypt and Guinea at around 12%. ST3 is most prominent in Denmark at above 25% followed by Egypt at around 15%. Maps for subtypes 4, 5, 6, 7 and 14 are available in **Appendix II**.

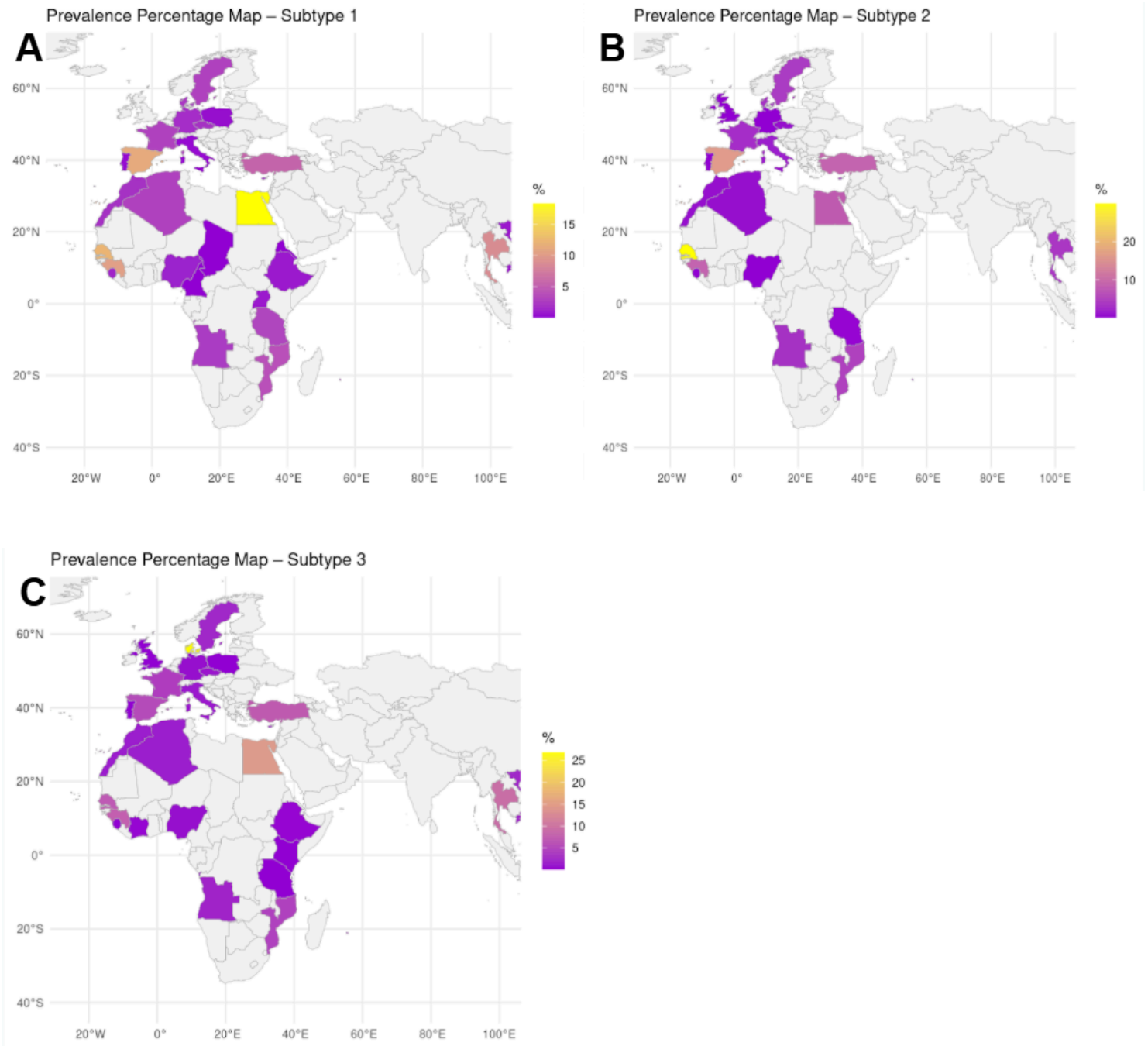


Figure 7. Global distribution of *Blastocystis* subtypes ST1-ST3 by country. A) Prevalence of ST1; **B**) ST2; **C**) ST3. Each panel shows a heat map which represents the frequency of each subtype based on the data grouped by country. The colour gradient displayed in the legend indicates the proportion of each subtype, with proximity to the colour yellow signifying a higher presence of the subtype and proximity to dark purple indicating a lower presence of the subtype. Geographic coordinates (latitude and longitude, in degrees) can be seen in the X and Y axes, representing orientation. A number of countries for which no data was available for ST1, ST2 and ST3 were excluded to enhance visual clarity.

From the map chart portraying ST2, it is interesting to see that the subtype's prevalence is 30% in Senegal, as it was expected to be higher in other European countries. Given the fact that the total number of data collected from African countries is about 30% of the one from Europe, this result seems to not be due to a sampling bias. This result can also be seen in the grouped bar charts (**Figure 7B),** where ST2 in Senegal is shown to account for over 25% of the subtype in the whole data. Over 50% of the data from ST2 found in humans is also accounted for by Senegal (**Figure 12)**. In contrast, The map in **Figure 7A** shows Egypt as being the country with the most concentration of ST1, however, this could be due to sampling bias. In the bar plots grouped by country, Egypt is shown to account for less than 25% of ST1. In the animal subtype prevalence stacked graphs, only a small number of samples came from Egypt, however, when human data were analysed the majority of ST1 samples came from Egypt. This doesn't necessarily mean that ST1 isn't indeed most present in Egypt, but this conclusion can be questioned.

Subtype Distribution by Sample Source

Following the global overview of the subtype distribution, a deeper understanding of subtype behaviour was reached by analysing the subtype distribution with a focus on the sample source. These findings are summarised in the following section.

An initial exploration was conducted using grouped bar charts (**Figure 8)**, the X-axis represents the sample source, including samples derived from living organisms (for which animal classes were used), environmental samples and entries with unknown origin (N/A). The bar colour represents each *Blastocystis* subtype found within the specific source. The percentages on the Y-axis show the proportion of each subtype among the total sample. Multiple bars in a group reach 100% because each bar represents the distinct subtype distribution within the specified source.

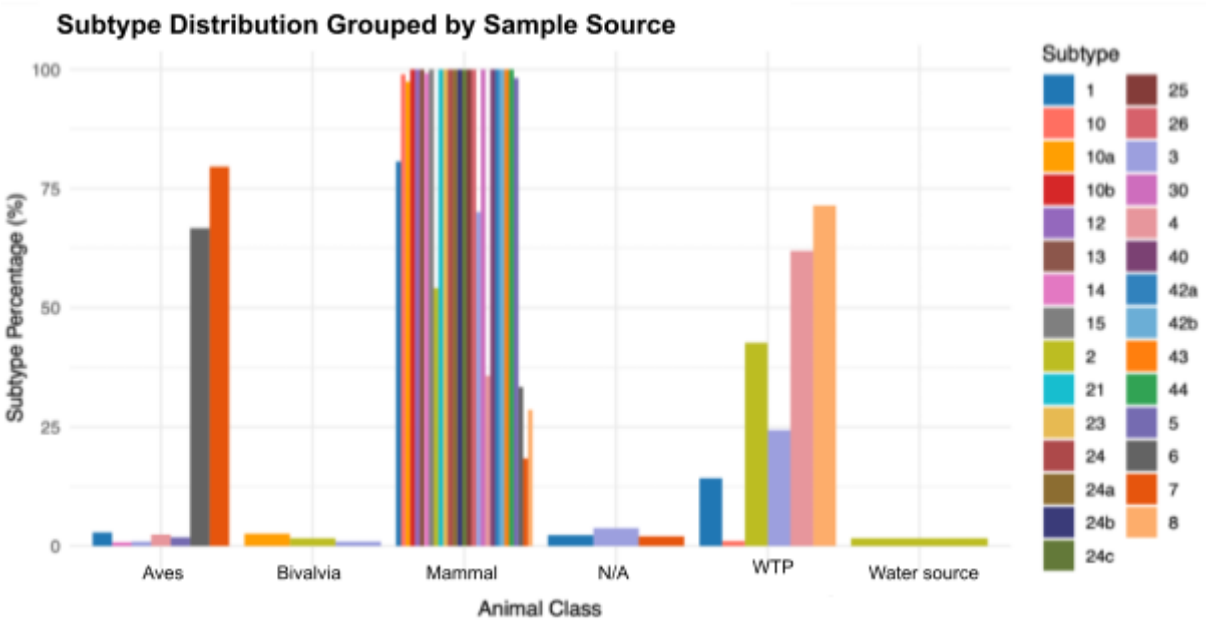


Figure 8. Grouped bar chart showing the distribution of *Blastocystis* subtypes across sample sources. The sample sources are represented on the X-axis; taxonomic class was used for living organisms in order to divide the data in a way that would show a widespread representation of the datapoints, while also allowing a clear visual of the subtypes present. N/A represents pre-diagnosed samples with an unknown source and 'WTP' refers to data collected from wastewater treatment plants. The Y-axis shows the proportion of each subtype within that sample source, bars that total 100% indicate that the presence of a specific subtype was exclusive to that source. This format was chosen to represent a large number of results while also allowing visual clarity due to the large number of data points.

The grouped graph shows a clear disparity in the diversity of subtypes across sample sources. The Mammalia group displays the most broad range of subtypes, including ST14, ST21, ST23 and ST24b. In addition, these subtypes along with ST10b, ST12, ST13 and ST23 are examples of subtypes that were only detected in this sample source, as they reach 100%. This suggests that mammals may possess characteristics that encourage the formation of a wider ecological range for *Blastocystis* subtypes, however, this could also be due solely to sampling bias created by the sample size of this study.

In contrast, the Aves class shows a predominance in the presence of subtypes ST6 and ST7, accounting for approximately 70% and 75% of all the subtypes found in birds, respectively.

ST2 also shows an association with water as it is not only present in both samples from water sources, but also in those collected from a wastewater treatment plant at around 45%.

ST2 is also present in the class Bivalvia; however, not significantly as it is only ~10%. In the N/A grouped bars only ST1, ST3 and ST7 are shown to be present, however at very low frequencies (<6%). This is likely to be due to the small sample size of this specific source category. However, considering the presence of the aforementioned subtypes and the lack of subtype ST2, an initial assumption can be made that this sample source is not related to water and does not belong in the Bivalvia class. It is interesting to see how research into subtype source/host specificity can lead to the recognition of the origin of the sample.

The N/A category is made up of samples of which the source was not specified, while these cannot be definitely allocated to a specific sample source, they are still important in highlighting gaps and the lack of homogeneity in *Blastocystis* data reporting and may also point to the need for improvement in sample annotation databases such as GenBank.

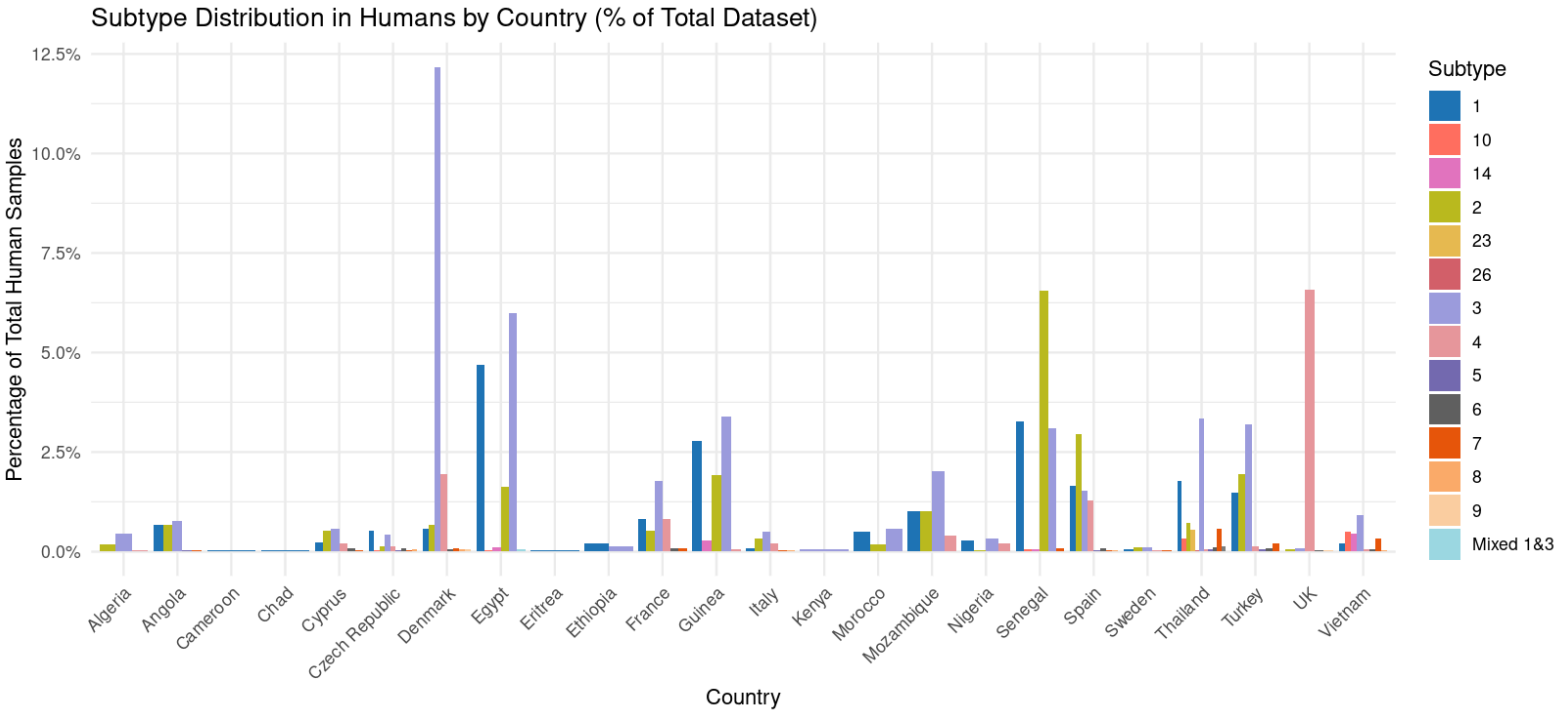


Figure 9. Grouped bar chart showing *Blastocystis* subtype distribution of human-derived samples in Africa and Europe, grouped by country. The X-axis shows the specific countries analysed with each bar representing the percentage contribution of a specific subtype to the whole dataset. The Y-axis shows the proportion of each subtype within that country. The Y-axis portrays the proportion of total human samples associated with the countries and subtypes. The bar colour associated with each subtype can be seen in the legend on the right of the chart. The width of each bar group reflects the subtype diversity within a country; narrower bars indicate a higher number of different subtypes present.

Given the large amount of data in the mammals group, which is primarily composed of human samples, a bar chart focusing on human subtype distribution was created. The grouping of the bars by country also provided further insight into *Blastocystis* subtype behaviours and patterns. For instance, the presence of ST3 is notable in several countries; however, Denmark stands out with the highest frequency of ST3 (~12.5%), followed by Egypt (~6.3%). ST3 being present in humans across different countries and two continents suggests that this specific subtype may exhibit less regional specificity compared to others. At the same time, the significantly higher frequencies of ST3 in Denmark and Egypt may also indicate that, although its distribution is not strictly country-specific, these countries may share characteristics that promote its prevalence. The UK has a significantly higher frequency of ST4 than any other country with it making up ~6.5% of the whole human dataset, in comparison, the second highest frequency belongs to Denmark at ~1.7%. Interestingly, ST4 has previously been referred to as the European ST(24), which can be seen in this chart as the subtype is detected in all the European countries analysed in this dataset. However, it can also be spotted in non-European countries such as Vietnam, Algeria and Angola at a higher frequency than in some European countries. This could be due to a sampling bias; however it could also indicate a change in transmission channels and how these are affecting regional specificity.

Another notable observation concerns Senegal which is one of the only African countries with over three STs having a frequency above 2.5%. ST2 in Senegal accounts for over 7% of all human samples in this dataset, making it the country in which this subtype was detected the most. This is significant because if we consider that the total number of data collected from African countries is about 30% of the data from Europe, this result does not appear to be due to sampling bias alone. This, in turn, indicates a possible geographic or population-specific association. The Czech Republic and Denmark are responsible for most ST8 observations. Egypt is the only country that presents a mix of two subtypes (ST1 and ST3), however this can be attributed to sampling error in this research. ST23, ST26 and

ST14 are present in very small amounts and only Cyprus and Thailand. There is no clear reason why this is, however, it could also point to the need for standardised annotation of *Blastocystis* subtype data and the use of genetic evidence such as sample sequences to confirm results.

Regional and Country-Level Patterns

Building on previous findings that suggest *Blastocystis* subtypes display regional and host specificity, this relationship was further explored by analysing subtype prevalence based on host class and family (or by overall sample source as in the case of non-living sources such as water). In order to facilitate the detection and understanding of possible trends, data from Europe, Africa and Asia was charted separately. Focusing on the host family instead of the class, allowed for a more detailed analysis by enabling comparisons between different organisms within the same class.

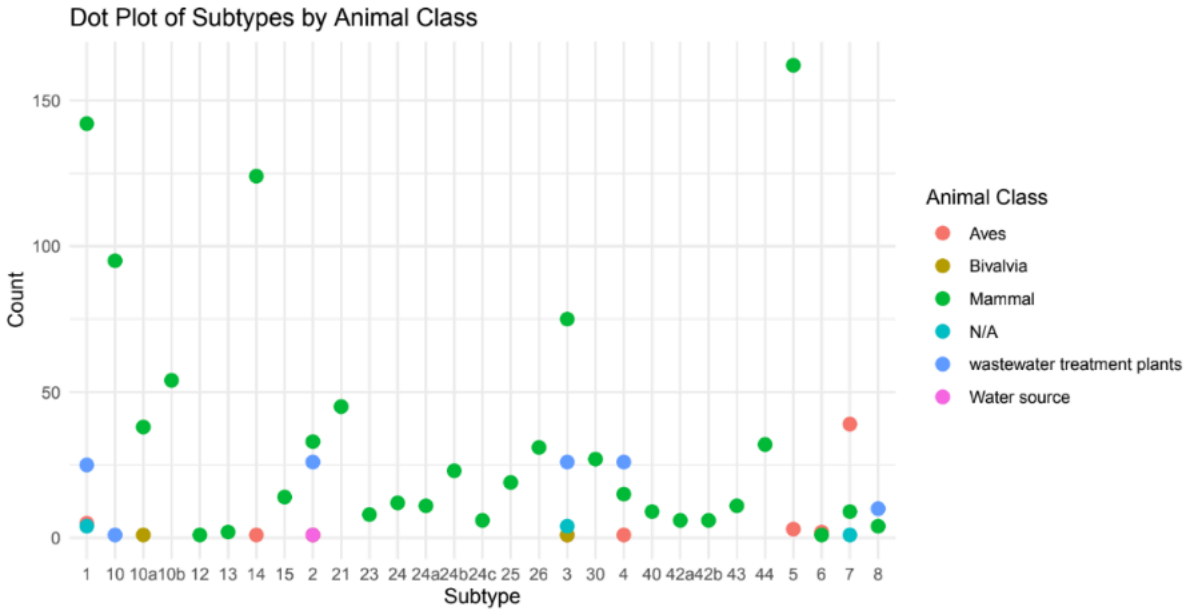


Figure 10. Dot plot showing the count of total samples per subtype, based on host taxonomic class. X-axis shows a list of the subtypes present in this dataset. Y-axis represents the total count of subtype occurrences. Dot colours represent the host class; assigned colours are specified in the legend on the right of the graph. Non-living organisms (e.g. water) added to aid a visual understanding of patterns found within this data.

Creating an initial graph that portrayed host classes and subtypes (**Figure 10**), allowed there to be an initial understanding of certain patterns in the data collected. For example, when looking at the taxonomic class of the host, Aves showed a high presence of ST7 and ST6, with the highest being ST7 at above 75%, however, it was not the highest across every sample. It also had one of the lowest ST3 presence. In the taxonomic class Bivalvia class had the lowest prevalence of ST10a not exceeding ~5%. Mammal was the host class with the highest diversity of subtypes. ST10B, ST12, ST13, ST15, ST21, ST24, ST24a, ST24b, ST24c, ST25, ST30, ST40, ST42a, ST42b, ST43, ST44, ST9 and a sample of mixed ST1&3 were only present in this class. This graph gave an initial insight on overrepresentation of certain host sources (in this case mammals) in *Blastocystis* research, however it left the need for a more in-depth analysis on which hosts demonstrated host specificity.

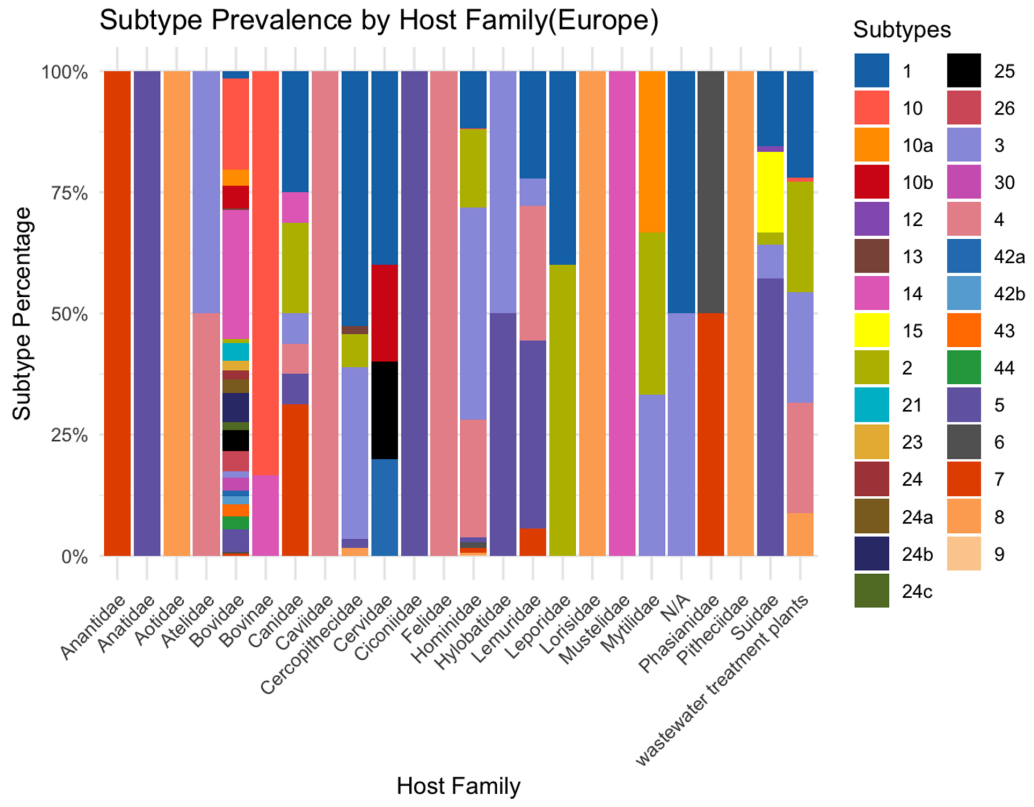


Figure 11. Stacked bar chart showing *Blastocystis* subtype distribution across host families in Europe. The X-axis represents different host families including non-living sources (wastewater treatment plants). Y-axis indicated the percentage of each subtype within that host group. Each bar is divided by colour to represent the relative contribution of

subtypes found within the specific host family. Subtype colours are shown in the legend on the right. N/A represents samples with unidentified sources.

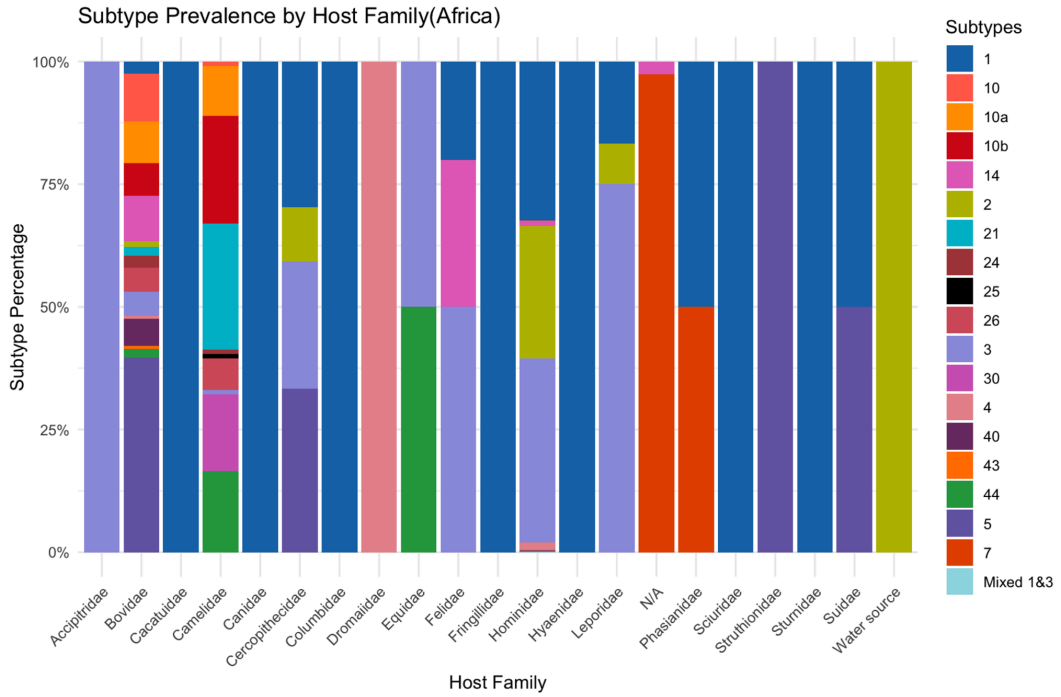


Figure 12. Stacked bar chart showing *Blastocystis* subtype distribution across host families in Africa. The X-axis represents different host families including non-living sources (water source). Y-axis indicated the percentage of each subtype within that host group. Each bar is divided by colour to represent the relative contribution of subtypes found within the specific host family. Subtype colours are shown in the legend on the right. N/A represents samples with unidentified sources.

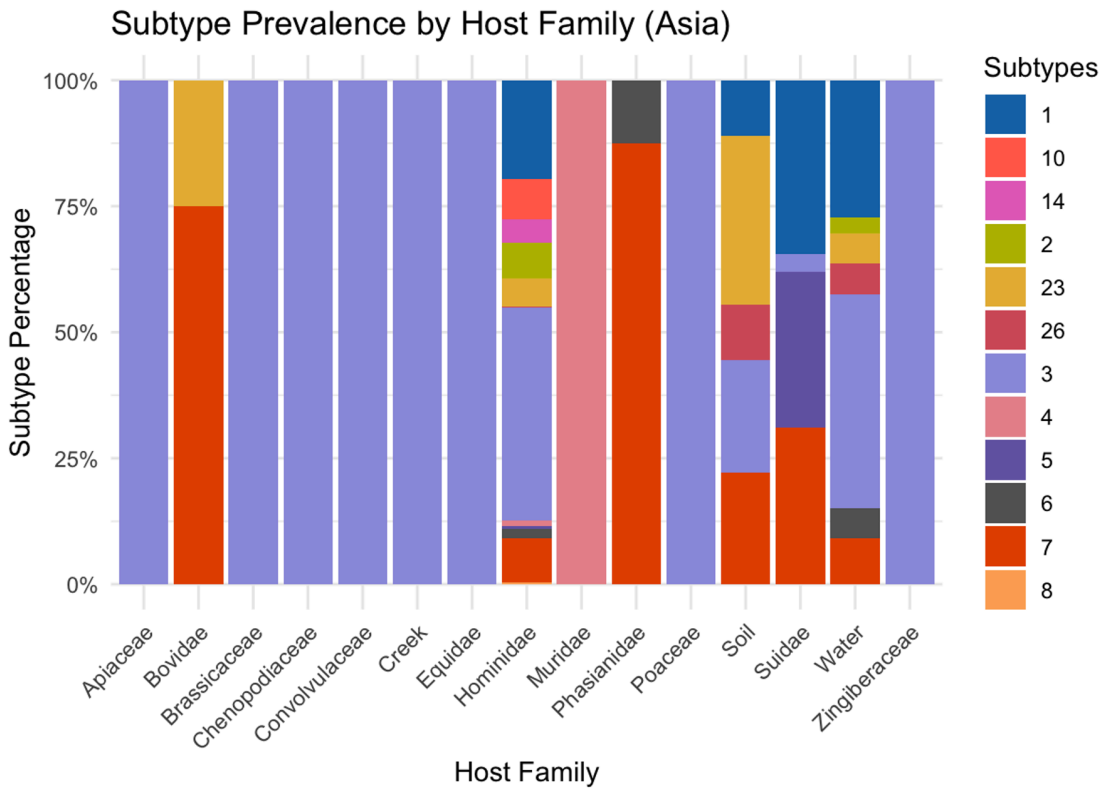


Figure 13. Stacked bar chart showing *Blastocystis* subtype distribution across host families in Asia. The X-axis represents different host families including non-living sources (water and creek). Y-axis indicated the percentage of each subtype within that host group. Each bar is divided by colour to represent the relative contribution of subtypes found within the specific host family. Subtype colours are shown in the legend on the right. N/A represents samples with unidentified sources.

When focusing on samples originating from Europe (**Figure 11**), members of the family Bovinidae display a subtype diversity that is significantly higher than the one shown in the other families, with over 15 subtypes associated with this taxonomic family. Bovidae family samples from Africa (**Figure 12**), show a similarity with European ones as they too present the highest subtype diversity compared to other sample sources. These shared results suggest that members of the family Bovidae may possess characteristics that encourage *Blastocystis* biodiversity. ST1 was detected in the taxonomic families Hominidae and Suidae across Europe, Africa and Asia (**Figure 13**); Cercopithecidae and Leporidae in both Europe and Africa; Phasianidae and Felinidae in African samples**.** Interestingly, ST1 cannot be observed in taxonomic families such as Camelidae, Mytilidae and Equidae. As well as Bovidae, Phasianida and Meridae members deriving from Asia. These results create further

theories surrounding ST1, they suggest that one or multiple characteristics associated with a sample source can either encourage growth and proliferation of *Blastocystis* with this subtype or prevent it completely.

Hominidae is the second most diversified taxonomic family in European and Asian samples. This was expected as humans are more likely to interact with different organisms and environments through activities such as travelling, facilitating transmission. It was unexpected, however, that family Camelidae in African samples would show the second highest subtype diversity in the continent. Upon closer inspection of the subtypes found in this family, ST10a and ST10b can be seen. The use of this subtype nomenclature is relatively new in *Blastocystis* research, which affects the reliability of comparisons involving these specific samples and further highlights the need for globally standardised reporting methods. This is also true for some of the subtypes only detected in the Bovidae family, some examples being ST24a, ST24b and ST24c. Some of these subtypes can be considered novel (post ST17). They have been accepted by a few scientists, hence why they are present in this current research, however there is not much data on these to allow for proper sample analysis. In this study alone, most of these subtypes have only been identified in one or two countries (notably Portugal) and only in animals (sheep and camels). One could blame the sample size of this research, but these subtypes are mostly observed in Europe, of which data is quite rounded. Therefore at the current time, even if the sample size of this study was more appropriate, it would be hard to use these as a justification for any conclusions surrounding the identification of a relationship between host characteristics and subtype specificity. The same conclusion applies to the African Camelidae samples.

ST4 was only detected in three families between Africa and Asia, but in six different families in Europe. This supports the idea of it being the European subtype. Future research using disciplines surrounding zoonotic behaviours may provide impactful insights on reasons for this geographical specificity.

ST7 shows an affinity for the taxonomic family of Phasianidae, as it is detected across all three continents at above 45%, suggesting a subtype specificity that is not affected by

region, but by its sample source -more specifically, characteristics of a host.

It is difficult to draw any comparisons between sample sources (such as water) and living organisms when analysing taxonomic families. However, environmental sources can still provide insight into how external factors may affect subtype diversity. For instance, all water sources show a consistent presence of ST2 and this subtype is also found across multiple host types in the dataset. In both Europe and Asia, the percentage of ST2 in water samples is similar to that found in human samples, which could indicate a possible transmission through frequent contact between humans and water sources.

While the previous *Blastocystis* sample source-focused analysis offered insights on subtype host specificity, it also suggested the possibility of certain subtypes being more prevalent in certain regions. Therefore, the geographical origin of the samples was analysed further with the use of stacked bar charts, the split of the bars was used to represent the frequency of a subtype in a specific country. Due to the vast amount of data, charts were again made based on continent and split between human and animal data, for an easier comparison.

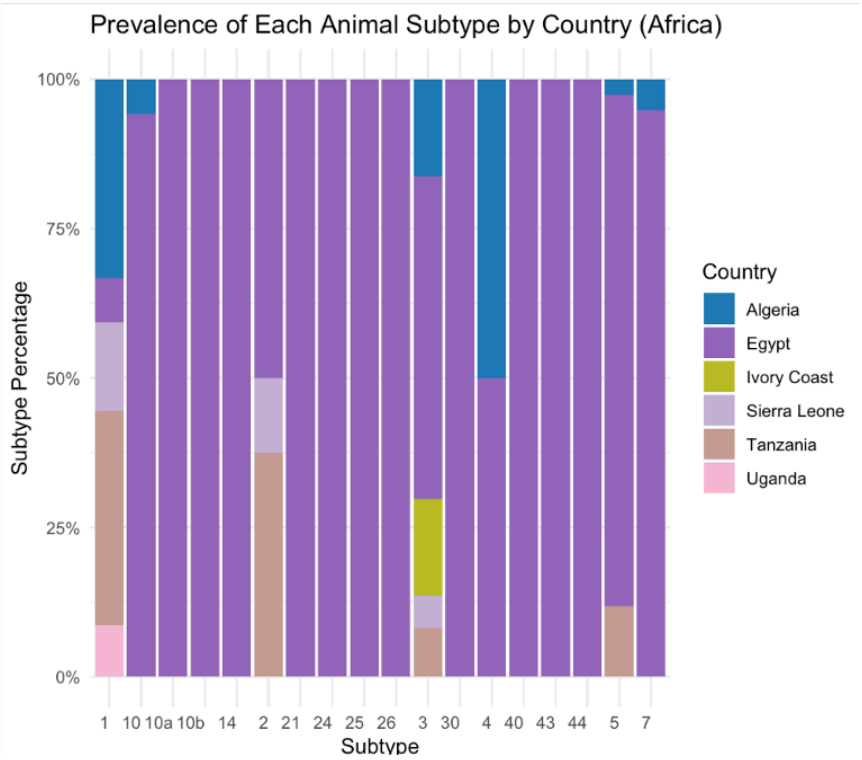


Figure 14. Stacked bar chart showing *Blastocystis* subtype distribution of animal data across African countries. The X-axis represents individual *Blastocystis* subtypes. The Y-axis indicates the percentage of each subtype attributed to specific countries. Each bar is divided into colours which represent the contribution of each country to the occurrence of that subtype. A longer length of a coloured section within a bar, represents higher amounts of occurrences of that subtype in that country. Country colour assignments are shown in the legend on the right.

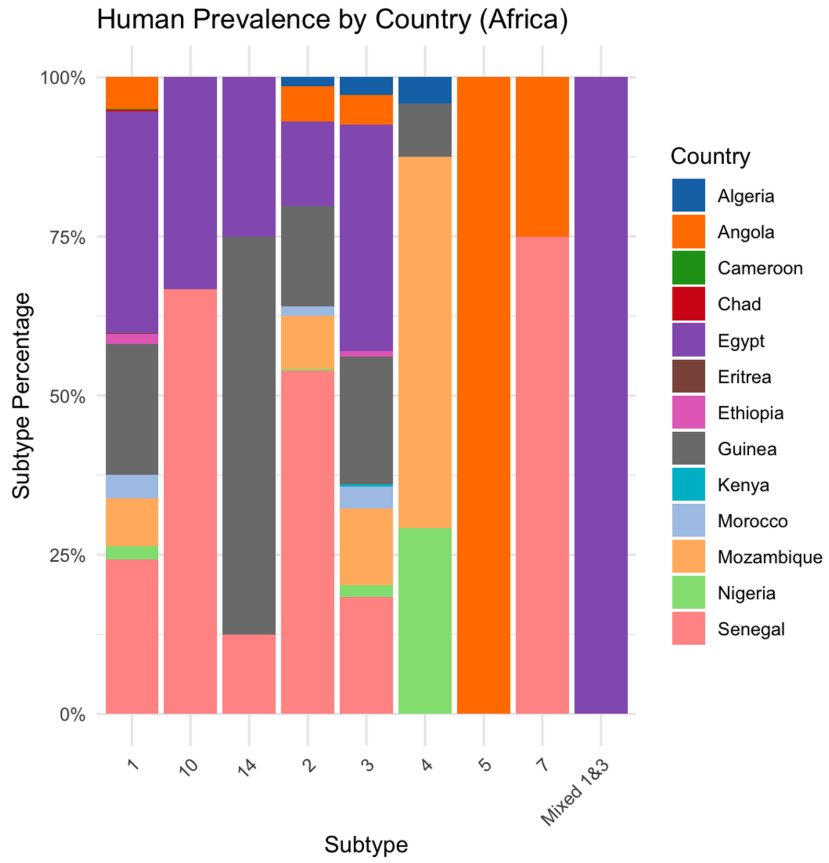


Figure 15. Stacked bar chart showing *Blastocystis* subtype distribution of human data across African countries. The X-axis represents individual *Blastocystis* subtypes. The Y-axis indicates the percentage of each subtype attributed to specific countries. Each bar is divided into colours which represent the contribution of each country to the occurrence of that subtype. A longer length of a coloured section within a bar, represents higher amounts of occurrences of that subtype in that country. Country colour assignments are shown in the legend on the right.

A few of the African countries in both the animal and human data, showed subtype diversity as they appear in various bars, representing various subtypes. This is especially true for Egypt and Senegal, from which samples contained six different subtypes in the human data (including ST1, ST10 and ST14). However, data from Senegal did not include any animal samples. Egypt on the other hand, showed a high subtype diversity as it can be seen

present in every bar representing different subtypes, with its frequency being above 50% in most cases. This, along with the presence of over five subtypes found only in Egypt, can be used to argue the reliability of the results as it could be an indicator of a sampling bias. The presence of these subtypes in Egypt is not in question; however, comparing these results with those from other countries could lead to an unfair and incorrect analysis. Ivory Coast, Uganda, Tanzania and Sierra Leone only had samples containing animal data, for the latter two there is some evidence of subtype diversity as they appear in multiple bars including ST1, ST2 and ST3. As expected, ST1 and ST3 were the subtypes found in most countries, which was also previously observed in heatmaps (**Figure 7A-C**).

Overall, the chart created with African human data, gave a more complete picture of subtype trends in this continent compared to animals-only focused data. For example, highest detection of ST10, ST2 and ST7 was in Senegal (>50%) and for ST4 it was in Mozambique (>65%). ST5 only appeared in Angolan data. The differences in trends observed in **Figure 14** and **15**, highlight the need for research that incorporates samples from a broader range of sources in order to develop a more complete understanding of *Blastocystis* subtype distribution.

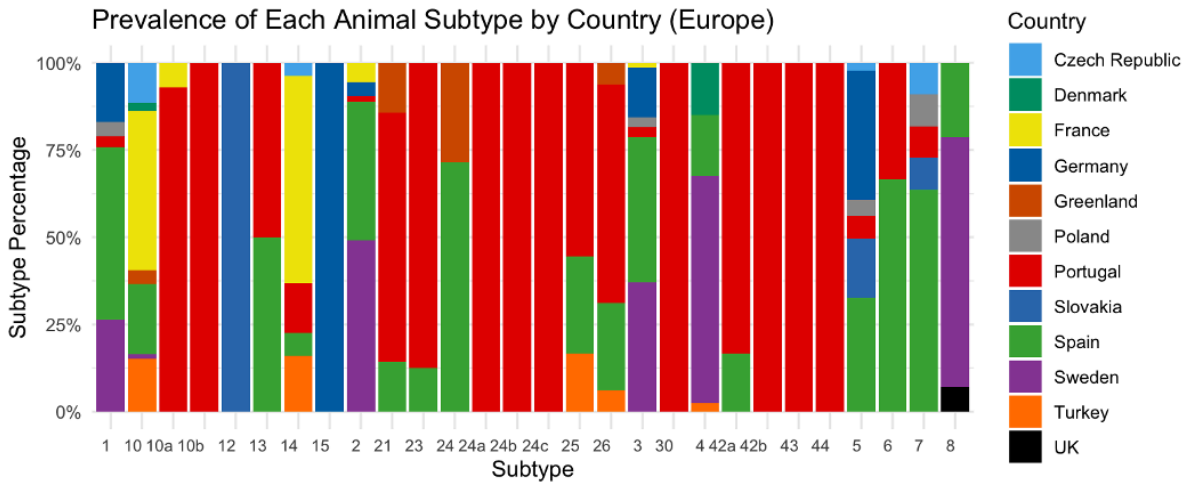


Figure 16. Stacked bar chart showing *Blastocystis* subtype distribution of animal data across European countries. The X-axis represents individual *Blastocystis* subtypes. The Y-axis indicates the percentage of each subtype attributed to specific countries. Each bar is divided into colours which represent the contribution of each country to the occurrence of that subtype. A longer length of a coloured section within a bar, represents higher amounts of occurrences of that subtype in that country. Country colour assignments are shown in the

legend on the right.

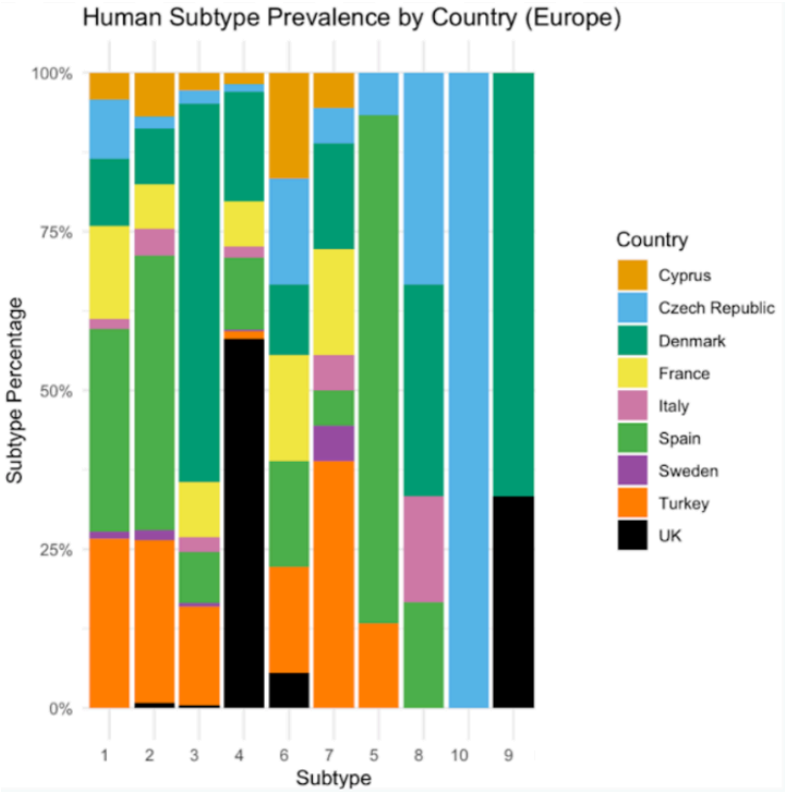


Figure 17. Stacked bar chart showing *Blastocystis* subtype distribution of human data across European countries. The X-axis represents individual *Blastocystis* subtypes. The Y-axis indicates the percentage of each subtype attributed to specific countries. Each bar is divided into colours which represent the contribution of each country to the occurrence of that subtype. A longer length of a coloured section within a bar, represents higher amounts of occurrences of that subtype in that country. Country colour assignments are shown in the legend on the right.

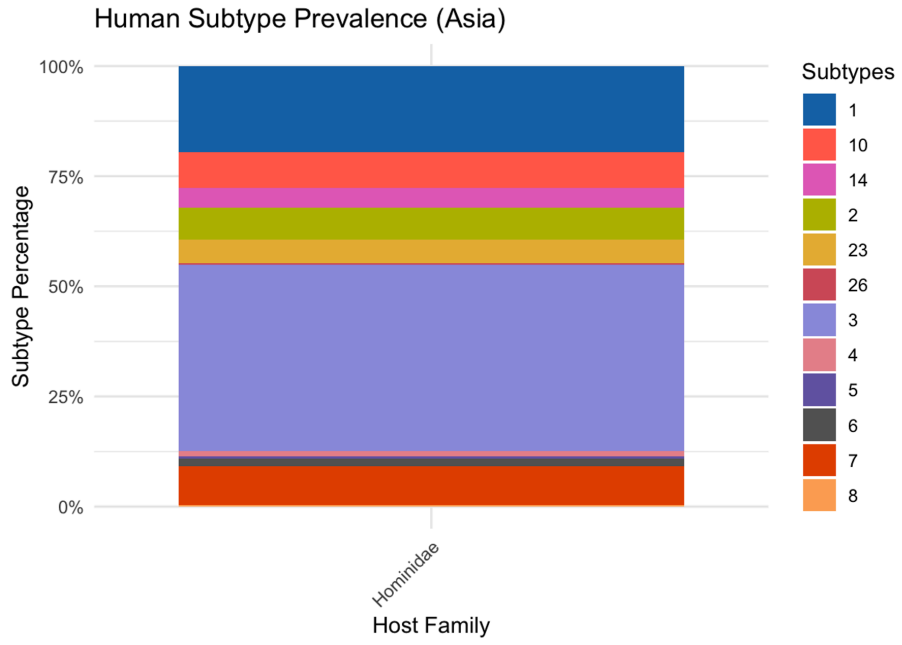


Figure 18. Stacked bar chart showing *Blastocystis* subtype distribution of human data across Asian countries. The X-axis represents individual *Blastocystis* subtypes. The Y-axis indicates the percentage of each subtype attributed to specific countries. Each bar is divided into colours which represent the contribution of each country to the occurrence of that subtype. A longer length of a coloured section within a bar, represents higher amounts of occurrences of that subtype in that country. Country colour assignments are shown in the legend on the right.

In Europe, there was a significant difference in the number and spread of subtypes found in human and animal data (**Figures 16 and 17)**. ST9 is only present in human samples, with Denmark still showing the highest percentage and the only other country it is present in is the UK. Portugal does not have any human data, however it has the highest number of ST24a, ST10b, ST24b, ST24c, ST42a, ST42b, ST6, ST13, ST23, ST25, ST30, ST43 and ST44 in animal data compared to the other countries (**Figure 16**). Sweden is one of the European countries that has both animal and human data and it has quite a high percentage of ST1-4 and ST8 in the animal samples. France is another country with both human and animal data with a high percentage of ST1, ST2, ST6 and ST7 (>50%) in the human samples and the highest presence of both ST2 and ST3 in the animal samples. France is also the only other European country to show presence of ST10a. The most prevalent subtype in Asian human data (**Figure 18)** is ST3 followed by ST1. The least prevalent is ST4. This concords with the previously discussed findings declaring ST4 as rarely found outside Europe.

When the human-only data from each continent is looked at, there are more cross-regional conclusions that can be drawn. For example, ST1, ST2 and ST3 suggest low regional subtype specificity as they are not only present in every continent, but they also tend to come from various countries. This is also reflected in the presence of these particular subtypes in some of the water samples. This conclusion is not definitive, however, due to the large presence of human samples in multiple countries/continents, cross-regional comparisons are made easier. However, the downfall of this is that the uneven distribution of host types can also lead to misinterpretation of the data, due to an overrepresentation of

certain categories. Looking at the taxonomic family as a whole is also detrimental because this includes both animals and humans which have very different habits and diets that could affect ST prevalence. For example, if we were only to look at European human data for ST5 (**Figure 14**), the conclusion made is that this subtype possesses strong regional specificity as it was only found in three different countries (Spain, Czech Republic and Turkey) and mostly observed in Spain. However if one were to look at the animal data from the same continent, it can be seen that ST5 is already present in more countries and, although still very present in Spain, it is also quite prevalent in Germany. This hints at a difference in subtype prevalence patterns between animals and humans in the Hominidae family. A few of the subtypes also suggest low zoonotic potential by being mostly (or only) found in humans. This is true for ST1-3 in every continent, ST9 in Europe, ST10 and ST14 in Asia. The low reporting of the last three subtypes however suggests that this could be due to the small sample size.

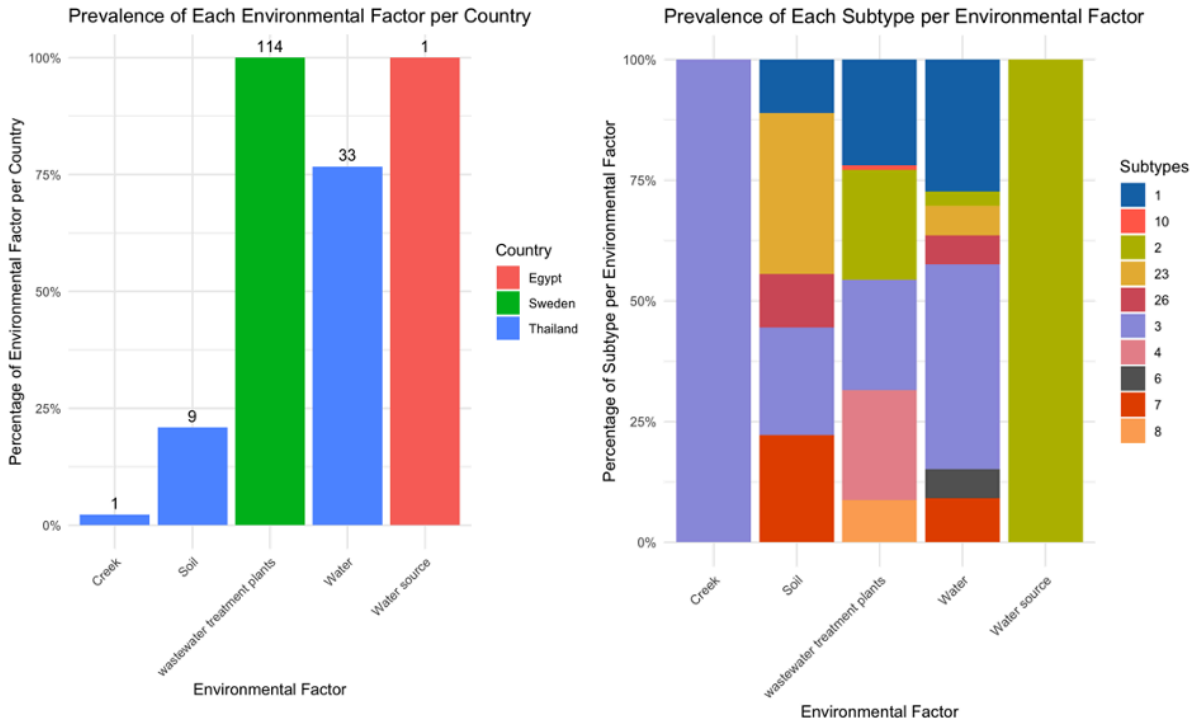


Figure 19. Bar chart showing the percentage of each environmental factor analysed in the dataset per country (on the left); the X-axis shows the environmental source the data was collected from, while the Y-axis represents percentage of data from each environmental factor based on country. Counts giving the number of data points belonging to that specific sample source are displayed at the top of each bar. The country the sample was found in is represented by the colour of the bar; these are specified in the colour legend at

the left of the chart. **Bar chart showing the percentage of each subtype in each environment (on the right).** The X-axis shows the environmental source the data was collected from, while the Y-axis represents the subtype percentage in that specific environmental source. Each bar is divided into colours that represent the subtype detected. Colour assigned to each subtype is specified in the legend on the right of the chart.

Figure 19 shows the subtype prevalence in each environmental factor, consisting of water from a creek in Thailand, drinking water from school buildings in Thailand, water from wastewater treatment plants in Sweden, water from a waterborne protozoan parasite and soil from Thailand. Only one case of both ST3 and ST2 was recorded from Thailand and Egypt respectively. ST3 was the most recurrent subtype, found in the creek, soil, wastewater plants and water samples, with the highest frequency being in the latter.This is consistent with previous results in **Figure 11**, where ST3 was detected in high frequencies across various host families. ST23 was detected in the water, but it was significantly higher in the soil samples. ST7 and ST26 are other subtypes that were only found in these two sample sources. ST2 was found in the water source, water and wastewater treatment plants. Considering the fact that ST2 was only found in human samples (**Figures 15, 17** and **18**), the presence of this subtype in wastewater was expected as this usually comes from human waste. A result that was not anticipated was the bar representing soil samples; this showed diversity in subtypes present which could be explained by the possible constant contact and interactions between soil and different organisms. As research surrounding *Blastocystis* increases, it would be interesting to see if (like with the N/A data in **Figure 8**) researchers may be able to deduce what kind of population lived in an area based on soil subtype diversity.

Subtype Specificity Based on Host Diet

After theorising that proximity and interaction between different sample sources (for example wastewater and humans) could explain subtype prevalence similarities, The relationship between *Blastocystis* subtypes and the host's diet was also explored and represented in the form of a faceted mosaic plot (**Figure 20**).

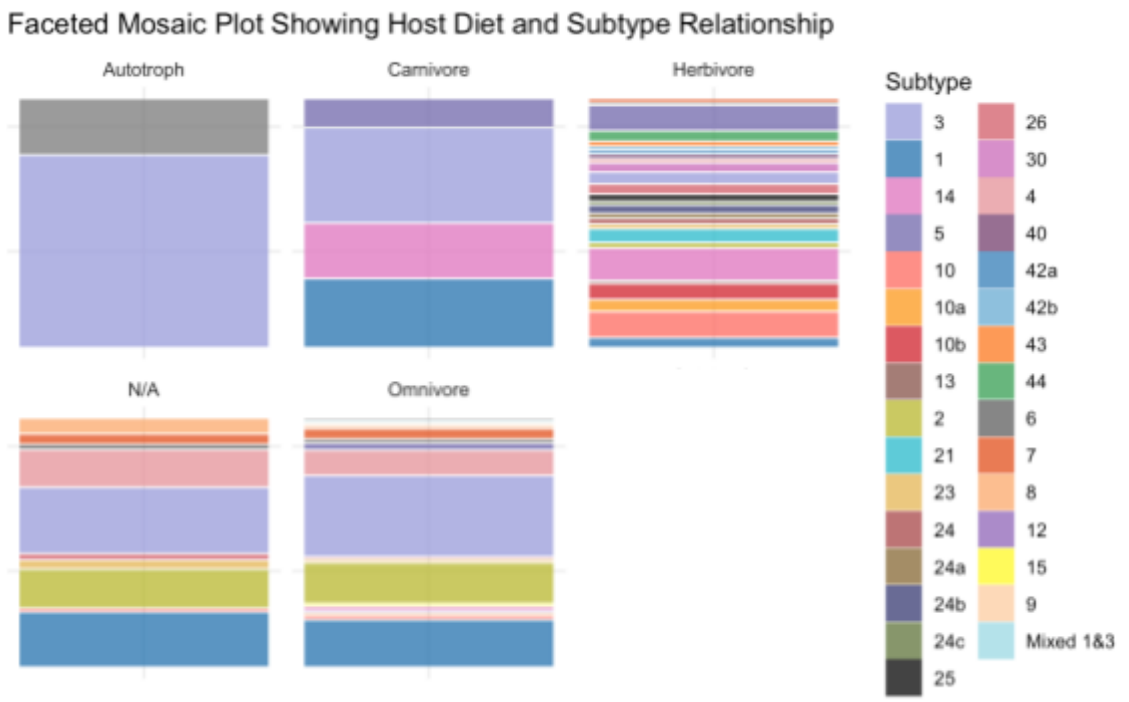


Figure 20. Multi-faceted mosaic plot exploring the relationship between host's diet and the subtype found. Each panel represents a different host diet, consisting of autotrophs, carnivores, herbivores, omnivores and N/A. N/A refers to samples from non-living objects (such as water) and pre-diagnosed samples. Each band represents a specific subtype in that group of data; a wider band indicates a higher number of subtype cases. Colours assigned to each subtype can be seen on the legend to the right of the chart.

ST3 is present in every host regardless of the diet and it is especially prevalent in the samples collected from autotrophs (which refers to only plants in this study), carnivores and omnivores. The only other subtype detected in autotrophs was ST6 which was detected in the omnivore and herbivore host samples in very small quantities compared to the rest of the subtypes and not at all in carnivores. All the subtypes present in carnivore host samples (1, 3, 5 and 14) are also shared by both omnivore and herbivore host samples. Herbivore host samples have the most variation when it comes to subtypes. This agrees with the findings discussed earlier in **Figures 11** and **12**, which showed a significantly higher amount of samples being taken from members of the Bovidae family. This pattern is also reflected in the presence of a multitude of different subtypes in omnivore host samples, of which a lot are from humans in this study, as samples from members of the Hominidae family were the second highest. A lot of samples from omnivores have ST1 which consists of the data found in Africa and Asia(**Figure 13**) as the stacked graphs showed ST1 as being the most

prevalent subtype in Hominidae and Suidae, however it disagrees with those found in Europe, where ST7 was found to be the most prevalent. ST7 is present in the omnivores data, but not at the high amount expected after the findings in **Figure 11**. In addition, ST10 was found to be the most prevalent in European and African Bovidae samples, this can be seen in the herbivore panel, however unexpectedly the width of the stripe representing ST10 is similar to that for ST14. This suggests that regardless of the unequal sample size of the host family, the amount of ST14 across all animals is quite high.

The relationship between subtypes, continents and diet can also be further seen in the Sankey diagram (**Figure 21**). The Sankey diagram takes data points as a whole into consideration and creates a visual representation of relationships through the connection and size of nodes and lines. In the diagram the Hominidae node is significantly wider than any other one, including Bovidae. This shows that the visualisation created by groupings based on host characteristics are very different to those representing data as a whole. The diagram also demonstrates host specificity when it comes to continents. Subtypes 10a, 10b and 44 are shown to be specific to the samples derived from Africa and Europe. Finally, this diagram further confirmed ST4 as being more prevalent in Europe, compared to other continents, given the thickness of the link.

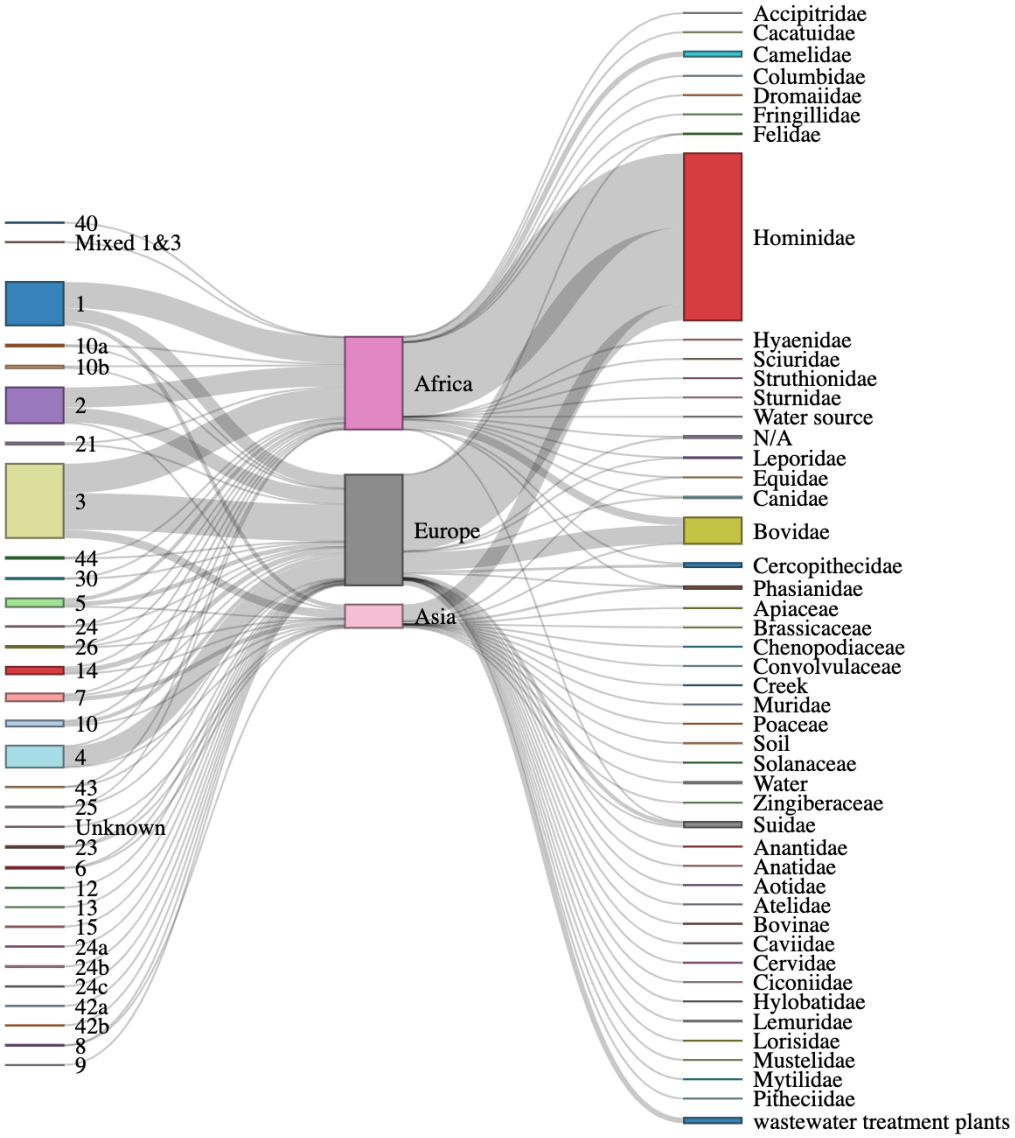


Figure 21. Sankey diagram showing relationship between subtypes (left nodes), continent (middle nodes) and host taxonomic family (right nodes). Wider links (lines) suggest higher occurrence of that specific parameter. For more information and interactive counts refer to appendix IV.

Statistical Analysis

To understand the relationship between host diet, sample geographical origin and subtypes, an MCA plot was created to assess whether the data would automatically show relatedness through grouping.

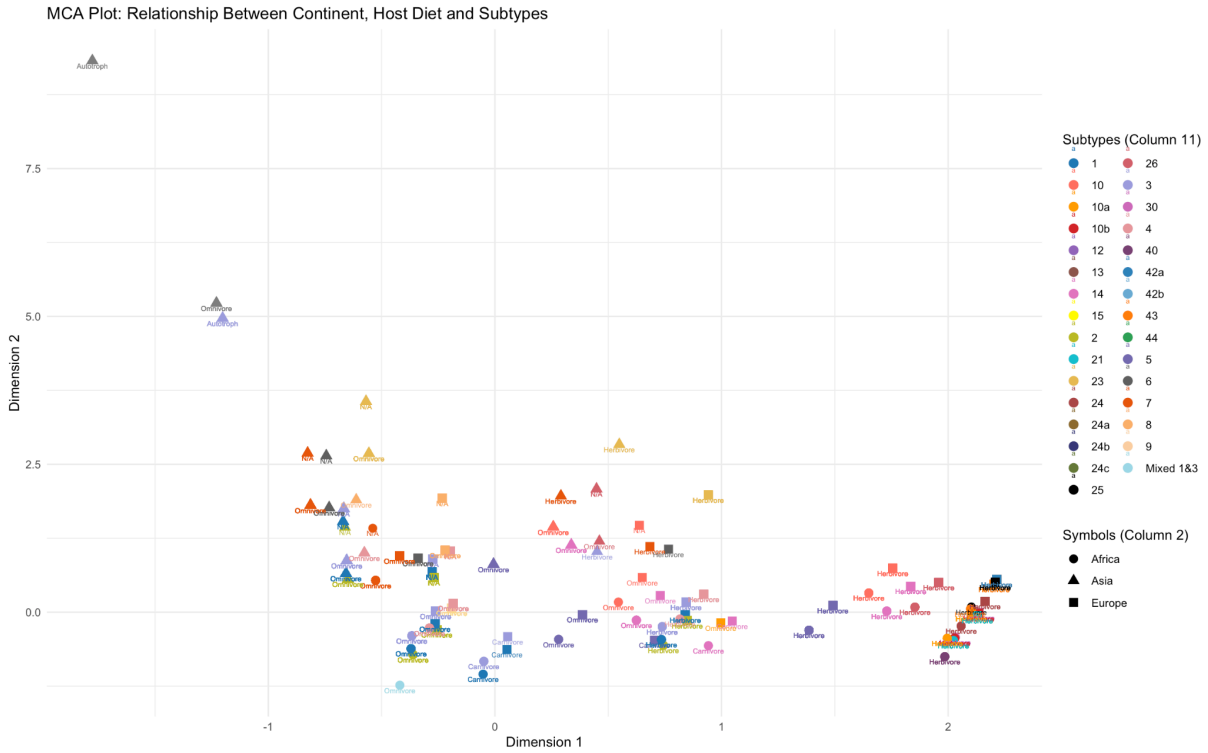


Figure 22. MCA plot visualising data point relationship between *Blastocystis* subtype, host and diet. Dimensions calculated using RStudio algorithm in MCA package. Each point represents a specific combination of categorical variables; proximity between points indicates greater similarity or association. Overlapping of points shows similar behaviour. Diet is indicated by captions underneath the point; the continent the sample was found in is represented by shapes and subtypes are shown by the colours of the datapoints. Assigned colours and shapes detailed in the legend on the right of the plot.

In **Figure 22**, the self-grouping of some data points suggests potential associations between *Blastocystis* subtype, sample source and host diet. Data that stands out the most are autotrophs from Asia, the plot suggests that the autotrophs with ST6 do not have much in common with any other data point, however the autotrophs with ST3 have a lot in common with omnivores that have the same subtype. Asian data points in general show the presence of a stronger geographical relationship with their subtypes as the points from this continent are mostly present together. Interestingly, Africa and Europe subtypes have more in common as these data points are more mixed together across the plot. However there is a clear relationship between the host diet and the subtype found as omnivore and herbivore data from Africa and Europe is plotted closer together. Some omnivores and herbivores from these two continents are quite mixed together (~1,0). For example ST10, ST14 and ST1 omnivore samples show a close relationship with ST1, ST3 and ST2 herbivore samples.

This relationship is interesting because it suggests that when it comes to Africa and Europe, these specific subtypes are more likely to be found in both herbivores and omnivores, while ST2 from omnivores is more likely to be found in other omnivores.

From both the data in the mosaic plot and the MCA plot, we can deduce a relationship between subtype specificity and diet. As aforementioned, human data was overwhelmingly more present than animal data, for this study all humans were recorded as omnivores. From the higher availability of human data, it was expected to see omnivorous data sets as the most equally diverse in terms of subtype concentrations, however in the omnivore panel from the faceted mosaic plot, it can be seen that this is the case for herbivorous hosts. Omnivores are still quite diversified however they show a clear relationship with ST1-3. Herbivores, on the other hand, do show a higher concentration of ST14 and ST10 within their samples, however, every other subtype present is quite varied. As we know some humans can be vegetarian, so it would be interesting to know whether a person's diet is what affects subtype specificity or even pathogenicity. Interestingly, carnivores show quite a high percentage of ST14 samples, whereas omnivores do not. Unfortunately, the carnivorous sample size is significantly smaller than that of omnivorous hosts, therefore a comparison is hard. The trends from the pre-diagnosed samples “N/A” are very similar to that of the omnivores, suggesting that they might come from an omnivorous living organism. This is an interesting finding because it points to the host’s diet and subtype relationship being so 47 specific that the subtype composition of an unknown sample can be used to deduce the diet of the host. The MCA plot reveals a relationship between the host diet and the subtypes identified, as omnivore and herbivore data from Africa and Europe are plotted closely together. Interestingly, ST10, ST14 and ST1 from omnivorous hosts show a close association with ST1, ST3 and ST2 from herbivorous hosts. This suggests that, in Africa and Europe, these subtypes are not strictly tied to host types. In addition, it hints at host specificity levels depending on the host diet. In this case, ST2 from omnivore hosts was

seen plotted with other omnivores, therefore, something within the omnivorous diet results in a higher host specificity.

```
1   Pearson's Chi-squared test

data:  hum_v_an
X-squared = 2746.1, df = 217, p-value < 2.2e-16

2   Pearson's Chi-squared test

data:  con_s
X-squared = 1518.9, df = 62, p-value < 2.2e-16

3   Pearson's Chi-squared test

data:  coun_s
X-squared = 10649, df = 1023, p-value < 2.2e-16

4   Pearson's Chi-squared test

data:  diet_s
X-squared = 3736.7, df = 124, p-value < 2.2e-16
```

Figure 23. Chi-Square of independence carried out to evaluate the relationship between categories. Calculated using chisq.test function in RStudio. 1:Subtype vs. Class; 2:Subtype vs. Continent; 3:Subtype vs. Country; 4: Subtype vs. Diet. P-value significance level = <0.05. Df= (number of rows - 1) x (number of columns - 1). X-squared = level of association between variables.

For all the tests the p-value was below the significance level of 0.05 (**Figure 23**), therefore the H0 was rejected as the chi-squared test determined that there is in fact a relationship between the variables and the subtypes found. The high df of values shows that the data is quite complex as it contains a lot of combinations of different categories. The presence of a relationship between the subtype and the variables is also further confirmed by the high x-squared values, as they show that there is a strong association between variables.

```
Overall Statistics

               Accuracy : 0.3856
                 95% CI : (0.3716, 0.3998)
    No Information Rate : 0.3381
    P-Value [Acc > NIR] : 7.346e-12

                  Kappa : 0.1369

 Mcnemar's Test P-Value : NA
```

Figure 24. Result from multinomial logistic regression model.

In the confusion matrix seen in **Figure 24**, the p-value shows significance, however accuracy is 38%, the confidence interval is between 37-40% and the no information rate is 33.8%. All of these values are below 50% which suggests the model not being reliable.

Discussion

The original aim of this thesis was to provide an updated overview of subtype distribution by building on the previous study conducted by Alfellani et al.(2013)(37). Unfortunately, the desired sample size and variety was not achieved. Therefore this study does not include data from all the regions available (notably North and South America) that were also previously addressed by Alfellani et al.'s study. In addition, only a small number of studies from Africa and Asia (13 and 10 respectively) were included compared to the 44 European studies used in the final calculation of subtype distribution, therefore hinting at sampling bias. However this was amended by a shift in perspective and methodologies, this study aimed still to shed some light on subtype distribution and host specificity, however, this was done by solely analysing the relationship between categorical factors, such as host location, taxonomic family/class and diet, and not by focusing on the amounts of each subtype in each region. This was achieved by the grouping of the categories during the data analytical stage.

Comparison with Existing Literature

Because it was decided that the focus would be on relationships between categories due to the desired sample size not being achieved, the total number of subtypes per country was not looked at like in Alfellani et al.'s publication(37). However, some of the knowledge on subtype regional specificity found by them can be expanded by the results of this study. Alfellani had previously stated that ST2 was not detected in Nigeria, however, this seems to not be the case anymore as a small amount (<5%) of African human ST2 came from Nigerian hosts (**Figure 12, Appendix IV**). In addition, It was also stated that ST7 was not found in any of the African samples, however, its presence was seen in the data collected (**Appendix II-c**). These came from Egyptian and Algerian animal data as well as Angola and Senegal in the human data (**Figures 11 and 12**). Due to the small presence of ST2 found so

far this change could simply be due to a different sampling area of Nigeria. If the same areas are re-sampled in the future and show an increase in ST2 prevalence, then this can be considered as an indicator of geographical changes in subtype distribution. The idea of this happening is not far-fetched as there have recently been various changes to environmental factors such as agriculture and soil quality in countries like Nigeria due to climate change(38).

Certain widely accepted subtype host specificity facts have also been supported by this research. ST1-3 were previously found to be predominantly found in human hosts (39), this was also observed in the data analysis conducted, with over 85% of all ST1-3 samples from Asian, European and African countries belonging to human hosts. ST4 has also been referred to as human-specific (40), this theory can also be deduced by the results obtained and the subtype's affinity with the family Hominidae, however, other families were also detected with this subtype. As well as Hominidae hosts, ST4 was found in Muridae hosts in Asia, Dromaiidae and Bovidae hosts in Africa and Lemuridae and Caviidae hosts in Europe. Although previous studies have not explicitly stated that ST4 is exclusively found in humans, its occurrence in non-human samples has been reported as minimal. However, due to the limited sample size in this study, the presence of ST4 across a variety of host families suggests that it may not be as human-specific as previously thought. This conclusion is put into question by the knowledge that (in this research) when Hominidae was related to ST4, it was automatically assumed that this concerned the human data due to the significantly larger amount of data collected from human hosts. Unfortunately, which species within Hominidae are most related to ST4 was not analysed. In addition, both Muridae and Caviidae family members are rodents and these have previously also been related to ST4(41).

There are a few widely considered animal-specific subtypes that have also been detected in this research. ST5, ST6 and ST7 are frequently related to livestock and poultry(42). ST6 and

ST7 have also previously been suggested to be avian-adapted subtypes(45). This was supported by some data across the regions analysed, such as the high prevalence of Suidae in ST5 in Europe and Asia. As can be seen by the grouped bar plots depicting the animal class, ST6 and ST7 were also highly correlated to Aves in these results. However these subtypes have been also found in various human hosts from all of the continents, this includes (and not limited to) Angola, Senegal, Sweden, Denmark and France. These results speak for the zoonotic potential of these subtypes, in fact, it has been previously found that there might be an increase in these STs in humans, especially in people who are in close contact with animals such as pigs (43)(44). The ST5-8 examples along with ST1-3 being increasingly found in both humans and animals, suggest that the boundaries between human and animal hosts are not as rigid as it was once thought. However a silver lining seems to be that for cross-infection to happen, multiple interactions between the human and the animal are required (45), therefore although potentially zoonotic, these *Blastocystis* subtypes are not to be considered highly contagious yet.

The *Blastocystis* subtype relationship with the host diet is not a novel area of research. Various studies have demonstrated that diet can influence the diversity and abundance of *Blastocystis* subtypes found in a host.

Studies have shown that animals, such as carnivores, that feed on prey are more likely to harbour a higher microbiota diversity(47), including *Blastocystis.* A wide cross-sectional study conducted in 2024(46) also found that individuals with increased adherence to a healthful plant-based diet were also more likely to host *Blastocystis*, supporting the notion that dietary fibre may promote colonisation.

Additionally, several studies have indicated *Blastocystis* as being [resent in food sources. For example, the parasite was found in vegetable samples from Syria(49) and vegetables sold in street markets in Thailand(48), further supporting the link between diet and *Blastocystis* transmission. This reinforces the idea that humans can not only obtain *Blastocystis* from animals but this can be transferred to them through the consumption of

plant-based foods. This might be especially true in countries where hygiene practices and standards are lacking. Furthermore, this observation is consistent with the conclusion previously made that herbivores present a higher variety of *Blastocystis* subtypes, likely due to the high consumption of vegetables creating a more favourable environment for parasites in the gut microbiome. It is well established that microbial growth is facilitated by the consumption of fibre-rich foods. Consequently, this might also facilitate the contribution to the increase of subtype prevalence among herbivores.

This conclusion on diet-related subtype specificity also explains the findings from the MCA plot. The higher likelihood of an omnivore and herbivore sharing the same subtype prevalence than two omnivores could be due to the specific hosts those samples came from sharing environmental ecological factors. This is interesting because it points to the possibility that one or more countries in Europe and Africa share enough ecological similarities that the hosts that inhabit them have similar subtype prevalence.

Blastocystis' presence in various environments and hosts does not spare water sources, which can be blamed for playing a part in the transmission of the parasite. Once again, underdeveloped countries suffer most from the lack of inadequately sanitised water and are therefore more susceptible to the transmission of parasites present in such environments. Several studies have found several water sources across the world to be directly responsible for *Blastocystis* infection (36), with ST1, ST3 and ST4 being more commonly found in drinking water sources(50). An extensive literature review on water contamination in African countries also attests to protozoans' tendency to be found in bodies of water(51).

As we established earlier, the parasite has still mostly unknown pathology methods and zoonotic potential. This makes its presence in sites such as wastewater treatment plants or school water sources very significant because even if it has not been proven beyond doubt, *Blastocystis* is thought to cause or even just increase gastrointestinal conditions. In rural areas where this parasite-infested water is used for consumption by both humans and animals and agriculture, the spreading of this parasite is facilitated even further. Curiously,

subtype regional specificity also applies to the environment as different subtypes were observed in different bodies of water which in some cases reflected the country's subtype prevalence patterns.

Potential Mechanisms and Explanations

After observing these relationships between variables relating to *Blastocystis*, several questions concerning underlying mechanisms that could affect subtype specificity are raised. After analysing different diets and subtype diversity further, it comes to mind that host-specificity could be potentially explained by the already existing gut microbiome of certain hosts. The gut microbiome is known to differ between hosts as it is heavily influenced by their lifestyle, in turn, the microbiome also shapes parasite colonisation. Several immunological or physiological factors determine the survival rate of different parasites in the microbiome. As concluded earlier, herbivores are more likely to retain different *Blastocystis* subtypes such as ST6 that are more frequently related to plants(48). This would explain some of the patterns observed and the relationship between subtype diversity and diet. Zoonotic transmission is another very favourable explanation for the relationship between specific subtype distribution and hosts. The fact that subtypes such as ST1 and ST3 are often seen in both animals and humans is evidence of the sharing of the parasite between individuals. In some places, it is common practice to live in close contact with livestock (37), in a lot of these populations it is also normal to eat most meals without the use of utensils, but just the hand. This paired with the fact that these individuals might not recognise the importance of common health practices such as the washing of hands throughout the day, serves as an explanation for the easy oral-faecal transfer. This also builds on the point made above on subtype prevalence being a result of one's gut microbiome. Living in close quarters suggests a shared environment, which in turn points to a similar microbiome.
Different living practices and a country's economic state can also be an explanation for geographical differences in *Blastocystis* subtype presence. The economic state of a country

dictates the resources that are spent to provide or allow access to environmental factors such as clean water(52), in addition, in some countries the sources of drinking water (such as rivers) are shared with some animals. It can then be deduced that if the water source has a specific subtype, this will also be transferred to the populations that make use of it. In turn, practices like bathing or washing clothes in bodies of water would create a repeated transmission cycle between living organisms and the environment. Another environmental factor that could affect geographical subtype prevalence is air pollution. It has been found that exposure to air pollutants can affect the health and diversity of the gut microbiome (53), including *Blastocystis.* No literature has been found to support this point, but some strains of *Blastocystis* subtypes may be more resistant to pollutants, resulting in significant differences in subtype prevalence.

The difference in subtypes found in different regions could also be due to temperature and climate-related factors. Warmer climates are known to be preferred by a lot of parasites, these favourable conditions could encourage the proliferation of specific subtypes. Seasonal changes in subtype prevalence have already been observed before(54), however, this was attributed to a change in population due to people wanting to be closer to bodies of water when the weather was warmer. These two factors together could explain *Blastocystis* specific subtype distribution in a specific region.

Limitations of the Study

Data Limitations

As aforementioned, this study was conducted to achieve a large-scale meta-analysis of most (if not all) the 18s rRNA sequences submitted to GenBank to provide an insightful overview of all the present *Blastocystis* subtype prevalence trends. This sample size however was not achieved. A consequence of a significantly smaller sample size is the limitation of the extent to which these results can represent different regions and populations. The majority of the

data was taken from European samples, creating an automatic bias in the conclusions drawn from the findings. This means that some of the samples reported as the most prevalent in a category, could not be such. The contrary is also possible. Sample bias also concerns the trends observed due to the difference in hosts. Many hosts that are exotic to the countries used could have been missed and with that the insight they could have given on the topic.

Methodological Constraints

There are a few limitations that could have led to the misrepresentation of the data. The choice to solely rely on GenBank for data collection was made based on the knowledge that it would contain most of the data found on *Blastocystis.* However, focusing on one source comes with the possibility that data is again missed. In addition, the keywords used in the search for datasets also limit the results that are given.

When sampling another limitation is the intentional omission of certain datasets. Data that lacked the indication of that subtype was found, was not used. It is also possible that some samples were filed under the wrong category such as country or type of host it was found in due to human error. Another possible mistake due to human error is the inclusion of clones in the data. If the samples were not clearly labelled as clones, they could have been added as separate, overrepresenting certain values.

Data such as diet was annotated with assumptions based on general host behaviour, for example, all humans were noted as omnivores and all sheep were noted as herbivores. This method does not account for special occurrences and unusual behaviours. In addition, sequencing was not used to confirm that the reported subtypes were correct, the subtype reported was completely reliant on the study, therefore if this was to be incorrect, so would the analysis be.

During analytical analysis, various commands could have hidden/unnoticed commands and functions that changed the data without the knowledge of the researcher. Commands such as ‘filter()’ and ‘mutate()’ from the ‘dplyr’ package were used; these can change data sets. If this happened the subsequent graphs created would be an incorrect representation of the data.

For visual clarity, the host category that was plotted was mainly the taxonomic family, however, this does not provide a detailed enough visual of the species present in a study and which subtype they can be related to. For example, there are many members of the Hominidae family, using only this limited the analysis between humans and animals.

In the case of data collected from direct submission, context was limited. This makes certain sets harder to generalise or apply to populations outside the regions studied due to potential differences in factors.

Implication For Future Research

During this study, it became increasingly apparent that more work is needed to achieve a full understanding of *Blastocystis* subtypes and their relationship with environmental and host factors. This can only be achieved through the application of One Health principles and the cooperation of experts from different disciplines.

The host specificity of various *Blastocystis* subtypes remains in question. From the data in this research and supporting literature, it was established that some subtypes specific to a host are increasingly seen in others. This is the case for subtypes such as ST1-ST4 being thought of being human-specific being found in animal hosts and ST5-7 being found in humans. Despite these findings, the extent of the zoonotic transmission of *Blastocystis* subtypes. The observations of animal subtypes in people such as zoo keepers, who have frequent livestock interactions, suggest that constant interaction with different hosts could

facilitate the transmission of new subtypes. However, the time frame for such exchanges is not well-defined.

Moreover, the specific genetic characteristics of an individual or *Blastocystis* subtype strains that may increase affinity to particular subtypes are yet to be fully understood. There are yet to find definitive genetic markers or mechanisms that explain why specific subtypes seem to be more attracted to specific hosts.

In this study, this was tried to be explained on a dietary level, through the analysis of different host diets. It has been theorised that this may be due to differences in the microbiome resulting from an individual's diet. Although plausible, this remains an hypothesis. In addition, in the MCA plot, the proximity of subtypes such as ST2 from herbivores being plotted closer to ST10, ST14 and ST1 from omnivores, than to ST2 from omnivores could be hinting to specific strains of the subtype being more host specific that others. This idea should also be explored further.

Extensive research on the relationship between the microbiome and *Blastocystis* subtypes should be carried out under various conditions to answer all the remaining questions on this relationship. Future research should also focus on identifying genetic markers and mechanisms that may explain why specific markers are more prevalent in specific hosts. This should include longitudinal studies of hosts in contact with livestock and cross-sectional studies in various regions.

The observed differences in regional subtype distribution could also be explored further. Large-scale multinational research is needed to understand the intricate web of connections between environmental factors such as climate, agricultural practices or urbanisation that ultimately contribute to changes in subtype prevalence. One of the relationships that are believed to have changed since Alfellani et al.'s study in 2013 is the presence of ST2 in Nigeria, as it was previously undetected in the country. In cases like this, the reason as to why this change has happened can only be speculated because there is no definite evidence that strongly suggests a factor/s that drives subtype distribution. Large-scale

multinational research would provide a clear indication of driving factors in region-dependent subtype prevalence and distribution. In addition, a longitudinal study would provide a timeline and context to the events that seem to affect *Blastocystis* subtypes as changes in the environment due to factors like climate change can often be gradual and therefore hard to identify as relevant to the answer the researcher is trying to answer. For example, low exposure to environmental pollutants such as pesticides or microplastics in the environment can eventually accumulate and lead to an altered gut microbiota through the selective effect of the survival and growth of certain microbes. Readings taken in intervals from an early stage of this exposure could provide most of the context needed to conclude the effects of environmental factors on *Blastocystis.*

Given the reported increasing zoonotic potential of *Blastocystis* transmission, especially in humans, future work should be carried out to tackle this topic. The parasite's potential link to IBS is not to be ignored, while scientists work on definitive answers, a public surveillance system could be set up in areas from which people tend to have close contact with animals, such as rural areas with wildlife prevalence, agricultural zones or livestock farms. This surveillance could detect the presence of subtypes of even specific strains that have been related to pathogenicity and prevent their spread throughout the communities. Regular check services made available for people for individuals would also help the containment of this spread. The previous findings of ST3 allele 34 being connected to urticaria and ST1 also being potentially pathogenic, make the fact that these subtypes have been found in bodies of water even more concerning. Although the control of subtype prevalence in large bodies of water may be close to impossible due to people's traffic, incorporating *Blastocystis* screening water quality monitoring systems may be a good idea to prevent the spread of pathogens through things such as drinking water. Investing in providing potable water everywhere would not only positively affect the countries it is provided in, but the world population as a whole as it would significantly decrease the risk of anyone retracting pathogenic parasites such as some *Blastocystis* subtypes. In addition, close attention should be paid to water

used for agriculture as *Blastocystis* can also be transferred through plant consumption.

Finally, the emerging subtypes were not extensively analysed in this study, however, there is a possibility that these could be a more resistant version of the old ones, therefore a monitoring system for these is highly recommended. A standardised international database should be created via the use of regular cross-regional cooperation. This should be made with globally agreed nomenclatures and accession codes.

Conclusions

At the beginning of this research primary objectives were to explore geographic and host *Blastocystis* subtype specificity and then subsequently analyse whether a change could be observed between the continents of Africa, Asia and Europe. Another objective was to analyse the relationship between factors that could affect subtype distribution including location, taxonomy and diet. The final objective was to identify connections between subtype prevalence affecting factors and the zoonotic transmission potential and pathogenicity.

This study confirmed that host specificity is still true for ST-3 as over 85% of the data belonging to these subtypes came from humans. ST4 still showed high affinity to humans, however it also appeared in multiple animal host families, two of which are rodents and have been related to ST4 before. ST2 was detected in Nigeria, contrary to the findings of Alfellani et al., who previously reported the subtype was not present in the region. ST5-7 showed animal host specificity as they were identified in animal families such as Aves. However, their high frequency in some of the human samples suggests they possess zoonotic transmission potential.

Host diet was confirmed to affect host specificity as data collected from herbivores showed a significantly greater presence of subtype diversity than any other diet and it was related to microbiome diversity. The data from the MCA plot hinted at the possibility of hosts from different regions sharing similar subtype prevalence due to their habitat having shared subtype prevalence affecting factors, such as food availability. From this plot, it could also be concluded that the host diet can affect host specificity.

After environmental *Blastocystis* subtype prevalence analysis, it was concluded that bodies of water reflect the regional subtype prevalence trends and patterns.

These findings underscore the importance of considering geographic and environmental factors to understand and monitor *Blastocystis* infections. An even higher level by experts from all disciplines is imperative to prevent any harmful pathogenic conditions. While this study presents an interesting analysis of the relationship between *Blastocystis* subtype

prevalence and other factors, it failed to provide a reliable overview of the current trends, a future attempt of this study would include a wider geographical sample pool, more representative of the whole world.

To fully understand the relationship between region, host and pathogenicity concerning subtype prevalence, it is essential to continue research from all angles. As the concept of One Health states, it is impossible to understand the trends and behaviours of different strains of parasites such as *Blastocystis* without interdisciplinary cooperation from different professionals. In this thesis, I have presented how different experts can contribute to a shared area of interest by presenting opinions from topics such as diet, climate and economy along with microbiology to explain certain subtype behaviours. Much progress has been made, yet there remains a great deal to explore and accomplish in this field.

References

1. Zierdt CH, Rude WS, Bull BS. Protozoan characteristics of Blastocystis hominis. American Journal of Clinical Pathology. 1967;48(5):495–501.

2. Zierdt CH. Blastocystis hominis--past and future. Clinical Microbiology Reviews [Internet]. 1991 Jan 1;4(1):61–79. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC358179/

3. Stenzel DJ, Boreham PF. Blastocystis hominis revisited. Clinical Microbiology Reviews. 1996 Oct;9(4):563–84.

4. Tan KSW. New Insights on Classification, Identification, and Clinical Relevance of Blastocystis spp. Clinical Microbiology Reviews [Internet]. 2008 Oct 1;21(4):639–65. Available from: https://cmr.asm.org/content/21/4/639

5. Alexeieff A. Sur la nature des formations dites kystes de Trichomonas intestinalis. CR Soc Biol. 1911;71:296–8.

6. Silberman JD, Sogin ML, Leipe DD, Clark CG. Human parasite finds taxonomic home. Nature. 1996 Apr;380(6573):398–8.

7. Yoshikawa H, Nagano I, Wu Z, Eu Hian Yap, Singh M, Takahashi Y. Genomic polymorphism among strains and development of subtype-specific diagnostic primers. Molecular and Cellular Probes [Internet]. 1998 Jun 1;12(3):153–9. Available from: https://pubmed.ncbi.nlm.nih.gov/9664577/

8. Yoon HS, Andersen RA, Boo SM, Bhattacharya D. Stramenopiles [Internet]. Schaechter M, editor. ScienceDirect. Oxford: Academic Press; 2009. p. 721–31. Available from: https://www.sciencedirect.com/science/article/pii/B9780123739445002534

9. Roberts T, Barratt J, Harkness J, Ellis J, Stark D. Comparison of Microscopy, Culture, and Conventional Polymerase Chain Reaction for Detection of Blastocystis sp. in Clinical Stool Samples. The American Journal of Tropical Medicine and Hygiene [Internet]. 2011 Feb 4;84(2):308–12. Available from:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3029188/

10. Dexler HG. Isoenzyme Analysis - an overview | ScienceDirect Topics [Internet]. www.sciencedirect.com. 2001. Available from: https://www.sciencedirect.com/topics/medicine-and-dentistry/isoenzyme-analysis

11. Mansour NS, Mikhail EM, El A, Sabry AG, Mohareb EW. Biochemical characterisation of human isolates of Blastocystis hominis. Journal of Medical Microbiology. 1995 Apr 1;42(4):304–7.

12. Tan K, Ibrahim M, Ng G, Nasirudeen A, Ho L, Yap E, et al. Exposure of Blastocystis species to a cytotoxic monoclonal antibody. Parasitology Research. 2001 Jun 19;87(7):534–8.

13. Scanlan PD, Stensvold CR. Blastocystis: getting to grips with our guileful guest. Trends in Parasitology. 2013;29(11).

14. Coyle CM, Varughese J, Weiss LM, Tanowitz HB. Blastocystis: To Treat or Not to Treat... Clinical Infectious Diseases [Internet]. 2011 Nov 10;54(1):105–10. Available from: https://academic.oup.com/cid/article/54/1/105/369396

15. Wawrzyniak I, Poirier P, Viscogliosi E, Dionigia M, Texier C, Delbac F, et al. Blastocystis, an unrecognized parasite: an overview of pathogenesis and diagnosis. Therapeutic Advances in Infectious Disease. 2013 Sep 12;1(5):167–78.

16. Udkow MP, Markell EK. Blastocystis hominis: Prevalence in Asymptomatic versus Symptomatic Hosts. Journal of Infectious Diseases. 1993 Jul 1;168(1):242–2.

17. Hussein EM, Hussein AM, Eida MM, Atwa MM. Pathophysiological variability of different genotypes of human Blastocystis hominis Egyptian isolates in experimentally infected rats. Parasitology Research. 2008 Jan 11;102(5):853–60.

18. Abdel-Hameed DM, Hassanin OM. Proteaese activity of Blastocystis hominis subtype3 in symptomatic and asymptomatic patients. Parasitology Research. 2011 Jan 29;109(2):321–7.

19. Rashid J, Taiwo OO, Ahluwalia I, Chungong S. Disparities in Infectious Diseases among Women in Developing Countries1. Emerging Infectious Diseases [Internet]. 2004 Nov;10(11):e24–4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3329039/

20. Kurniawan A, Karyadi T, Dwintasari SW, Sari IP, Yunihastuti E, Djauzi S, et al. Intestinal parasitic infections in HIV/AIDS patients presenting with diarrhoea in Jakarta, Indonesia. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2009 Sep;103(9):892–8.

21. Londoño ÁL, Mejía S, Gómez-Marín JE. Prevalencia y Factores de Riesgo Asociados a Parasitismo Intestinal en Preescolares de Zona Urbana en Calarcá, Colombia. Revista de Salud Pública. 2009 Feb;11(1):72–81.

22. Roberts T, Stark D, Harkness J, Ellis J. Update on the pathogenic potential and treatment options for Blastocystis sp. Gut Pathogens [Internet]. 2014 May 28;6(28):17. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4039988/

23. Tan KSW. Blastocystis in humans and animals: new insights using modern methodologies. Science Direct [Internet]. 2004;126(1-2). Available from: https://www.sciencedirect.com/science/article/pii/S0304401704004091#bib141

24. Stensvold CR, Suresh GK, Tan KSW, Thompson RCA, Traub RJ, Viscogliosi E, et al. Terminology for Blastocystis subtypes – a consensus. Trends in Parasitology. 2007 Mar;23(3):93–6.

25. Noël C, Dufernez F, Gerbod D, Edgcomb VP, Delgado-Viscogliosi P, Ho LC, et al. Molecular Phylogenies of Blastocystis Isolates from Different Hosts: Implications for Genetic Diversity, Identification of Species, and Zoonosis. Journal of Clinical Microbiology [Internet]. 2005 Jan 1;43(1):348–55. Available from: https://jcm.asm.org/content/43/1/348.short

26. Pandey PK, Verma P, Marathe N, Shetty S, Bavdekar A, Patole MS, et al. Prevalence and subtype analysis of Blastocystis in healthy Indian individuals. Infection, Genetics and Evolution. 2015 Apr;31:296–9.

27. Nemati S, Falahati Anbaran M, Mohammad Rahimi H, Hosseini MS, Aghaei S, Khalili N, et al. Evolutionary and phylogenetic analyses of the barcoding region suggest geographical relationships among Blastocystis sp., ST3 in humans. Infection, Genetics and Evolution. 2021 Dec;96:105151.

28. Jiménez PA, Jaimes JE, Ramírez JD. A summary of Blastocystis subtypes in North and South America. Parasites & Vectors. 2019 Jul 29;12(1).

29. C. Randall Clark, Mark, Alfellani MA, Christen Rune Stensvold. Recent Developments in Blastocystis Research. Advances in Parasitology. 2013 Jan 1;82:1–32.

30. Deng L, Yao J, Chen S, He T, Chai Y, Zhou Z, et al. First identification and molecular subtyping of Blastocystis sp. in zoo animals in southwestern China. Parasites & Vectors. 2021 Jan 6;14(1).

31. Alfellani MA, Taner-Mulla D, Jacob AS, Imeede CA, Yoshikawa H, Stensvold CR, et al. Genetic Diversity of Blastocystis in Livestock and Zoo Animals. Protist. 2013 Jul;164(4):497–509.

32. Casero RD, Mongi F, Sánchez A, Juan David Ramírez. Blastocystis and urticaria: Examination of subtypes and morphotypes in an unusual clinical manifestation. Acta Tropica. 2015 Aug 1;148:156–61.

33. Stensvold CR, Arendrup MC, Jespersgaard C, Mølbak K, Nielsen HV. Detecting Blastocystis using parasitologic and DNA-based methods: a comparative study. Diagnostic Microbiology and Infectious Disease. 2007 Nov;59(3):303–7.

34. Deng L, Wojciech L, Png CW, Koh EY, Aung TT, Kioh DYQ, et al. Experimental colonization with Blastocystis ST4 is associated with protective immune responses and modulation of gut microbiome in a DSS-induced colitis mouse model. Cellular and Molecular Life Sciences [Internet]. 2022 Apr 18;79(5). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9016058/pdf/18_2022_Article_4271.pdf

35. Yan YB, Su S, Lai R, Lu H, Ye Jinhua, Li X, et al. Genetic variability of Blastocystis hominis isolates in China. Parasitology Research. 2006 May 11;99(5):597–601.

36. Attah AO, Sanggari A, Li LI, Nik Him NAII, Ismail AH, Meor Termizi FH. Blastocystis occurrence in water sources worldwide from 2005 to 2022: a review. Parasitology Research. 2022 Nov 25;122.

37. Alfellani MA, Stensvold CR, Vidal-Lapiedra A, Onuoha ESU, Fagbenro-Beyioku AF, Clark CG. Variable geographic distribution of Blastocystis subtypes and its potential implications. Acta Tropica. 2013 Apr;126(1):11–8.

38. Ebele N, Emodi N. Climate Change and Its Impact in Nigerian Economy. Journal of Scientific Research and Reports. 2016 Jan 10;10(6):1–13.

39. Dima El Safadi, Amandine Cian, Céline Nourrisson, Pereira B, Morelle C, Bastien P, et al. Prevalence, risk factors for infection and subtype distribution of the intestinal parasite Blastocystis sp. from a large-scale multi-center study in France. BMC Infectious Diseases. 2016 Aug 26;16(1).

40. Popruk S, Adao DEV, Rivera WL. Epidemiology and subtype distribution of Blastocystis in humans: A review. Infection, Genetics and Evolution. 2021 Nov;95:105085.

41. Zhao W, Ren G, Wang L, Xie L, Wang J, Mao J, et al. Molecular prevalence and subtype distribution of *Blastocystis* spp. among children who have diarrheia or are asymptomatic in Wenzhou, Zhejiang Province, China. Parasite. 2024 Jan 1;31(Parasite):12–2.

42. Rudzińska M, Kowalewska B, Szostakowska B, Grzybek M, Sikorska K, Świątalska A. First Report on the Occurrence and Subtypes of Blastocystis in Pigs in Poland Using Sequence-Tagged-Site PCR and Barcode Region Sequencing. Pathogens. 2020 Jul 21;9(7):595.

43. Wang W, Owen H, Traub RJ, Cuttell L, Inpankaew T, Bielefeldt-Ohmann H. Molecular epidemiology of Blastocystis in pigs and their in-contact humans in Southeast Queensland, Australia, and Cambodia. Veterinary Parasitology. 2014 Jul;203(3-4):264–9.

44. ALFELLANI MA, JACOB AS, PEREA NO, KRECEK RC, TANER-MULLA D, VERWEIJ JJ, et al. Diversity and distribution of *Blastocystis* sp. subtypes in

non-human primates. Parasitology. 2013 Apr 8;140(8):966–71.

45. Greige S, El Safadi D, Bécu N, Gantois N, Pereira B, Chabé M, et al. Prevalence and subtype distribution of Blastocystis sp. isolates from poultry in Lebanon and evidence of zoonotic potential. Parasites & Vectors. 2018 Jul 4;11(1).

46. Piperni E;Nguyen LH;Manghi P;Kim H;Pasolli E;Andreu-Sánchez S;Arrè A;Bermingham KM;Blanco-Míguez A;Manara S;Valles-Colomer M;Bakker E;Busonero F;Davies R;Fiorillo E;Giordano F;Hadjigeorgiou G;Leeming ER;Lobina M;Masala M;Maschio A;McIver LJ;Pala M;Pitzalis M;Wolf J;Fu J;Zhernakova A;Cacciò SM;Cucca F;Berry SE;Ercolini D;Chan AT;Huttenhower C;Spector TD;Segata N;Asnicar F. Intestinal Blastocystis is linked to healthier diets and more favorable cardiometabolic outcomes in 56,989 individuals from 32 countries. Cell [Internet]. 2024;187(17). Available from: https://pubmed.ncbi.nlm.nih.gov/38981480/

47. Robertson LJ, Clark CG, Debenham JJ, Dubey JP, Kváč M, Li J, et al. Are molecular tools clarifying or confusing our understanding of the public health threat from zoonotic enteric protozoa in wildlife? International Journal for Parasitology: Parasites and Wildlife. 2019 Aug;9:323–41.

48. Jinatham V;Wandee T;Nonebudsri C;Popluechai S;Tsaousis AD;Gentekaki E. Blastocystis subtypes in raw vegetables from street markets in northern Thailand. Parasitology research [Internet]. 2023;122(4). Available from: https://pubmed.ncbi.nlm.nih.gov/36658225/#:~:text=Fresh%20produce%20(n%20%3D%2020)

49. Al Nahhas S, Aboualchamat G. Investigation of parasitic contamination of salad vegetables sold by street vendors in city markets in Damascus, Syria. Food and Waterborne Parasitology. 2020 Dec;21:e00090.

50. McCain A, Lucsame Gruneck, Siam Popluechai, Tsaousis AD, Gentekaki E. Circulation and colonisation of Blastocystis subtypes in schoolchildren of various ethnicities in rural northern Thailand. Epidemiology & Infection [Internet]. 2023;151:e77. Available from: https://www.cambridge.org/core/journals/epidemiology-and-infection/article/circulation-and-colonisation-of-blastocystis-subtypes-in-schoolchildren-of-various-ethnicities-in-rural-northern-thailand/77F42F6BBA552255CC5F381F16EED832

51. Siwila J, Mwaba F, Chidumayo N, Mubanga C. Food and waterborne protozoan parasites: The African perspective. Food and Waterborne Parasitology. 2020 Sep;20:e00088.

52. Global Affairs Canada. Water in developing countries [Internet]. GAC. 2017. Available from: https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/environmental_protection-protection_environnement/water-eau.aspx?lang=eng

53. Rio P, Gasbarrini A, Gambassi G, Cianci R. Pollutants, microbiota and immune system: frenemies within the gut. Frontiers in Public Health. 2024 May 10;12.

54. Ithoi I, Jali A, Mak JW, Wan Sulaiman WY, Mahmud R. Occurrence of*Blastocystis*in Water of Two Rivers from Recreational Areas in Malaysia. Journal of Parasitology Research. 2011;2011:1–8.

Appendix

Appendix I: Table with data percentages

Country	ST1	ST10	ST2	ST3	ST4
Algeria	43.5483871	1.61290323	9.67741935	35.483871	3.22580645
Angola	30.6666667	0	30.6666667	36	0
Cameroon	100	0	0	0	0
Chad	100	0	0	0	0
Cyprus	14.0350877	0	31.5789474	35.0877193	12.2807018
Czech Republic	27.2727273	13.6363636	7.57575758	22.7272727	7.57575758
Denmark	3.65630713	0.36563071	4.2047532	76.7824497	13.345521
Egypt	22.1052632	2.23684211	7.89473684	29.8684211	0.13157895
Eritrea	100	0	0	0	0
Ethiopia	58.3333333	0	0	41.6666667	0
France	11.4285714	14.6938776	8.57142857	25.3061224	11.4285714
Germany	21.3333333	0	2.66666667	13.3333333	0
Greenland	0	37.5	0	0	0
Guinea	32.9896907	0	22.6804124	40.2061856	0.68728522
Italy	7.5	0	27.5	42.5	17.5
Ivory Coast	0	0	0	100	0
Kenya	0	0	0	100	0
Morocco	39.5348837	0	13.9534884	46.5116279	0
Mozambique	22.7272727	0	22.7272727	45.4545455	9.09090909
Nigeria	34.4827586	0	3.44827586	37.9310345	24.137931
Poland	36.3636364	0	0	18.1818182	0
Portugal	1.66666667	0	0.55555556	1.11111111	0
Senegal	24.9448124	0.4415011	49.8896247	23.6203091	0
Sierra Leone	80	0	6.66666667	13.3333333	0
Singapore	0	0	0	0	100
Slovakia	0	0	0	0	0

Country	ST1	ST10	ST2	ST3	ST4
Spain	23.1625835	3.56347439	27.3942094	18.2628062	11.3585746
Sweden	21.4285714	0.79365079	23.8095238	23.8095238	21.4285714
Tanzania	65.9090909	0	6.81818182	6.81818182	0
Thailand	20.0992556	2.72952854	6.4516129	34.9875931	0.49627792
Turkey	18.2795699	4.30107527	24.0143369	39.4265233	2.15053763
UK	0	0	0.85106383	1.27659574	96.5957447
Uganda	100	0	0	0	0
Vietnam	7.95454545	19.3181818	0	36.3636364	2.27272727

Country	ST5	ST7	ST6	ST14	ST8
Algeria	3.22580645	3.22580645	0	0	0
Angola	1.33333333	1.33333333	0	0	0
Cameroon	0	0	0	0	0
Chad	0	0	0	0	0
Cyprus	0	1.75438596	5.26315789	0	0
Czech Republic	4.54545455	3.03030303	4.54545455	6.06060606	3.03030303
Denmark	0	0.54844607	0.36563071	0	0.36563071
Egypt	8.55263158	4.73684211	0	3.02631579	0
Eritrea	0	0	0	0	0
Ethiopia	0	0	0	0	0
France	0	1.2244898	1.2244898	25.7142857	0
Germany	44	0	0	0	0
Greenland	0	0	0	0	0
Guinea	0	0	0	3.43642612	0
Italy	0	2.5	0	0	2.5
Ivory Coast	0	0	0	0	0
Kenya	0	0	0	0	0
Morocco	0	0	0	0	0
Mozambique	0	0	0	0	0

Country	ST5	ST7	ST6	ST14	ST8
Nigeria	0	0	0	0	0
Poland	36.3636364	9.09090909	0	0	0
Portugal	3.33333333	0.55555556	0.55555556	8.33333333	0
Senegal	0	0.66225166	0	0.4415011	0
Sierra Leone	0	0	0	0	0
Singapore	0	0	0	0	0
Slovakia	88.2352941	5.88235294	0	0	0
Spain	6.68151448	1.78173719	1.11358575	1.55902004	0.8908686
Sweden	0	0.79365079	0	0	7.93650794
Tanzania	20.4545455	0	0	0	0
Thailand	2.72952854	20.3473945	2.97766749	0	0
Turkey	0.71684588	2.50896057	1.07526882	6.09318996	0
UK	0	0	0.42553191	0	0.42553191
Uganda	0	0	0	0	0
Vietnam	0	12.5	2.27272727	18.1818182	1.13636364

Country	ST9	ST10a	ST10b	ST21	ST24
Algeria	0	0	0	0	0
Angola	0	0	0	0	0
Cameroon	0	0	0	0	0
Chad	0	0	0	0	0
Cyprus	0	0	0	0	0
Czech Republic	0	0	0	0	0
Denmark	0.36563071	0	0	0	0
Egypt	0	3.28947368	4.60526316	4.07894737	0.65789474
Eritrea	0	0	0	0	0
Ethiopia	0	0	0	0	0
France	0	0.40816327	0	0	0
Germany	0	0	0	0	0
Greenland	0	0	0	25	25
Guinea	0	0	0	0	0

Country	ST9	ST10a	ST10b	ST21	ST24
Italy	0	0	0	0	0
Ivory Coast	0	0	0	0	0
Kenya	0	0	0	0	0
Morocco	0	0	0	0	0
Mozambique	0	0	0	0	0
Nigeria	0	0	0	0	0
Poland	0	0	0	0	0
Portugal	0	7.22222222	10.5555556	5.55555556	0
Senegal	0	0	0	0	0
Sierra Leone	0	0	0	0	0
Singapore	0	0	0	0	0
Slovakia	0	0	0	0	0
Spain	0	0	0	0.4454343	1.11358575
Sweden	0	0	0	0	0
Tanzania	0	0	0	0	0
Thailand	0	0	0	0	0
Turkey	0	0	0	0	0
UK	0.42553191	0	0	0	0
Uganda	0	0	0	0	0
Vietnam	0	0	0	0	0

Appendix II: RStudio Packages Used

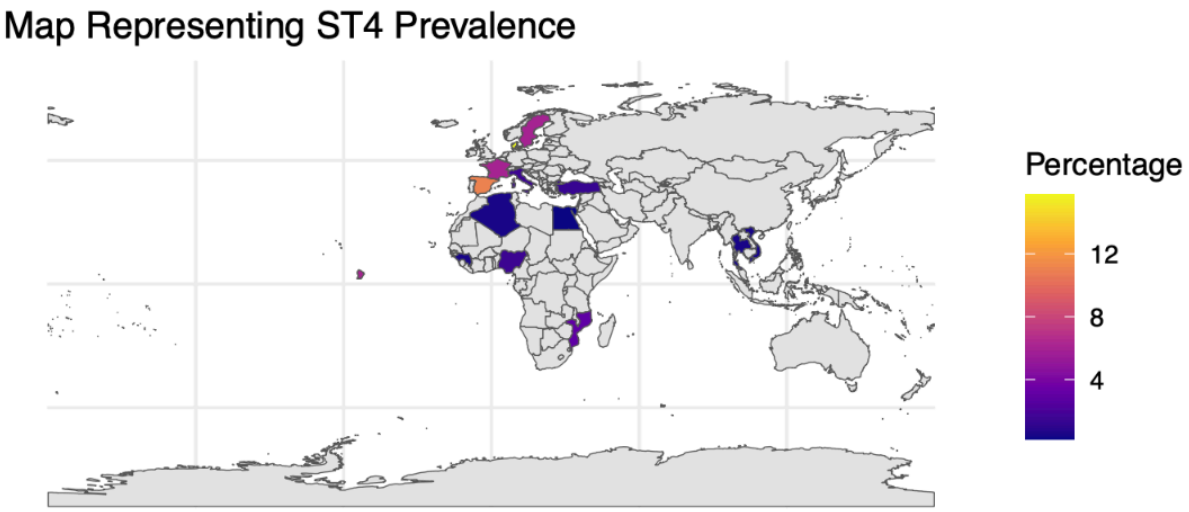
- Tidyverse: data manipulation
- Dplyr: data manipulation
- ggplot2: Creating plots
- ggrepel: Avoid data labels overlapping in plots.
- rnaturalearth: map data for continents and countries graphs
- ggmosaic: mosaic plot creation
- gganim: graph animation

Statistical analysis packages:

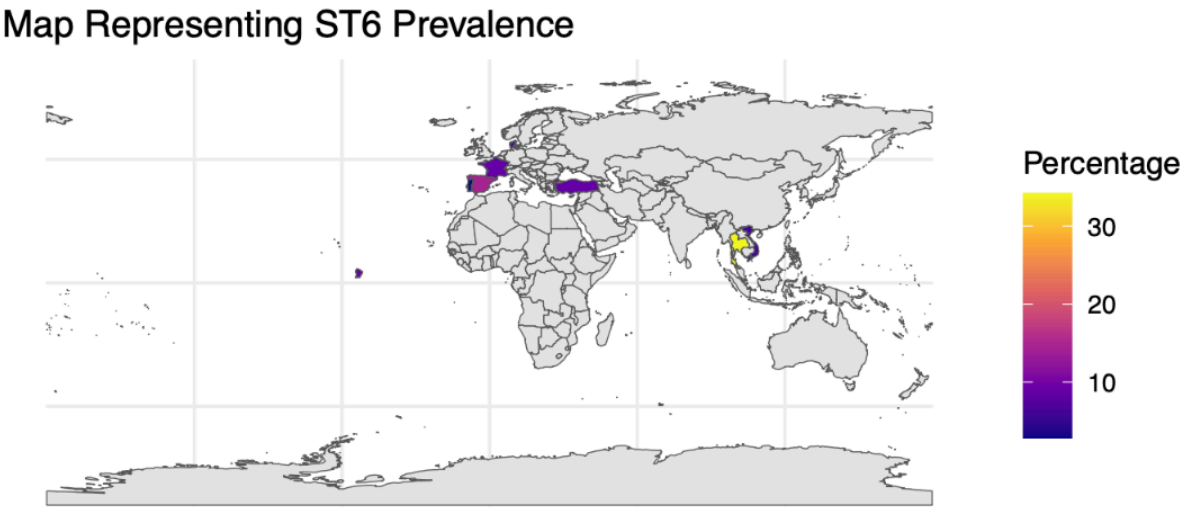
- Chisq.test(): evaluate association between two categorical variables.
- FactoMineR: ‘MCA()’ used to perform MCA.
- car: regression models creations
- ca: ‘ca()’ used to perform CA

Appendix III: World Maps

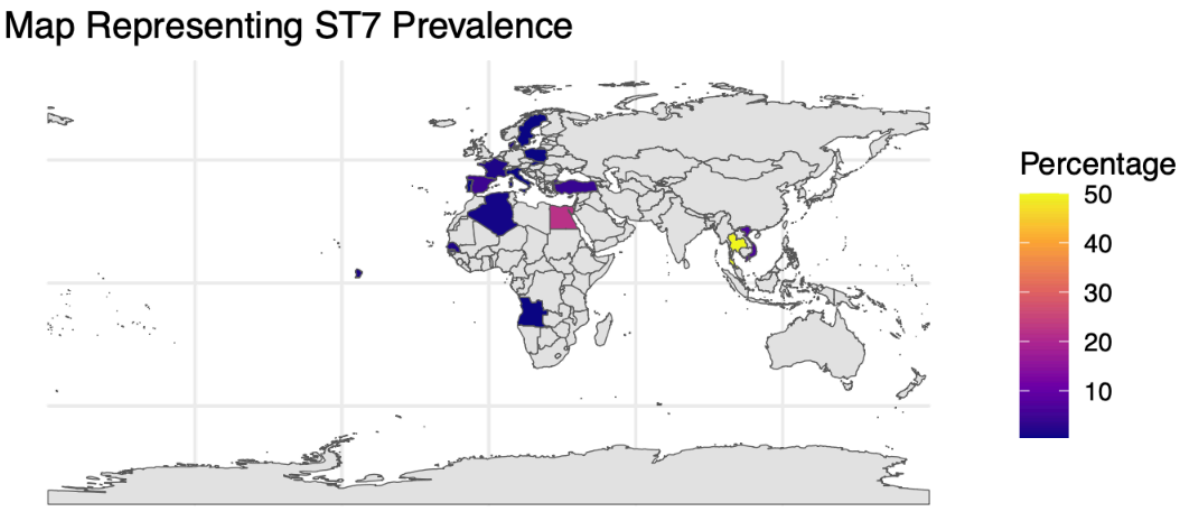
a. ST4 map



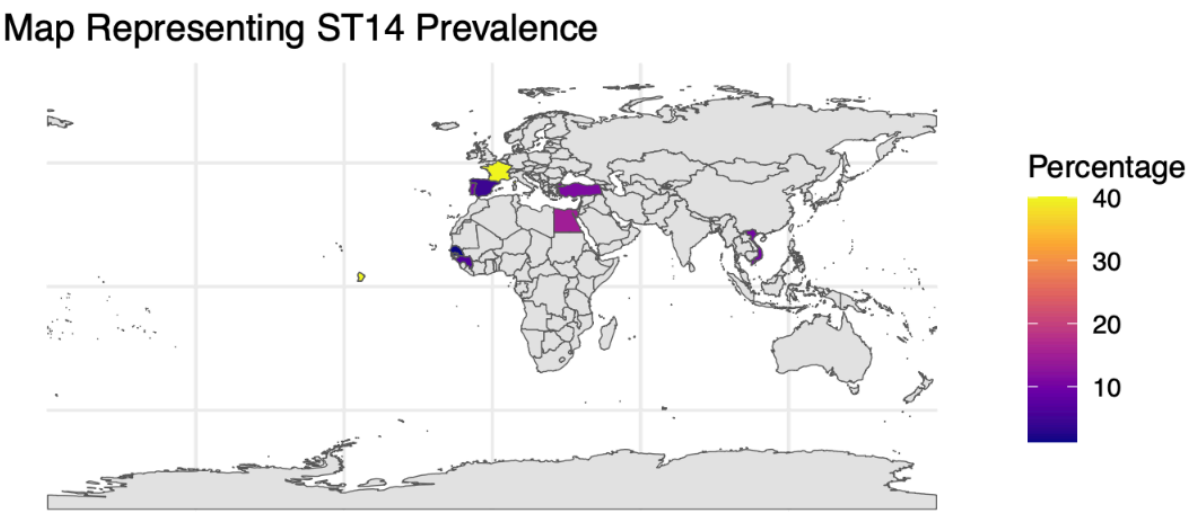
b. ST6 map



c. ST7 map



d. ST14 map

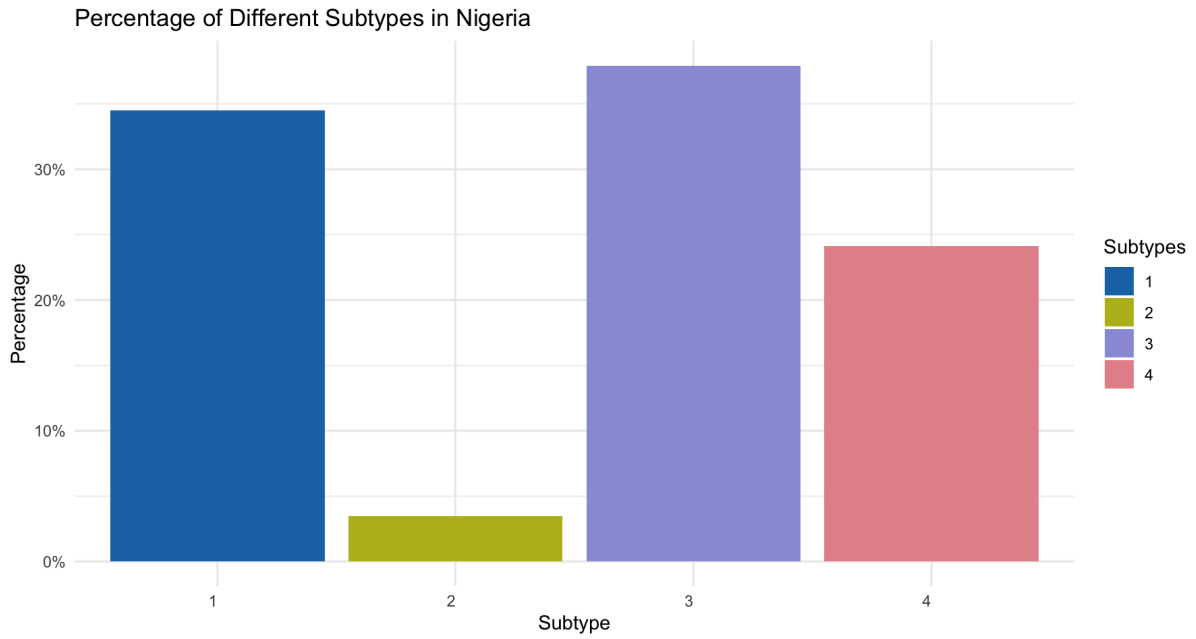


Appendix IV: Sankey Diagram link

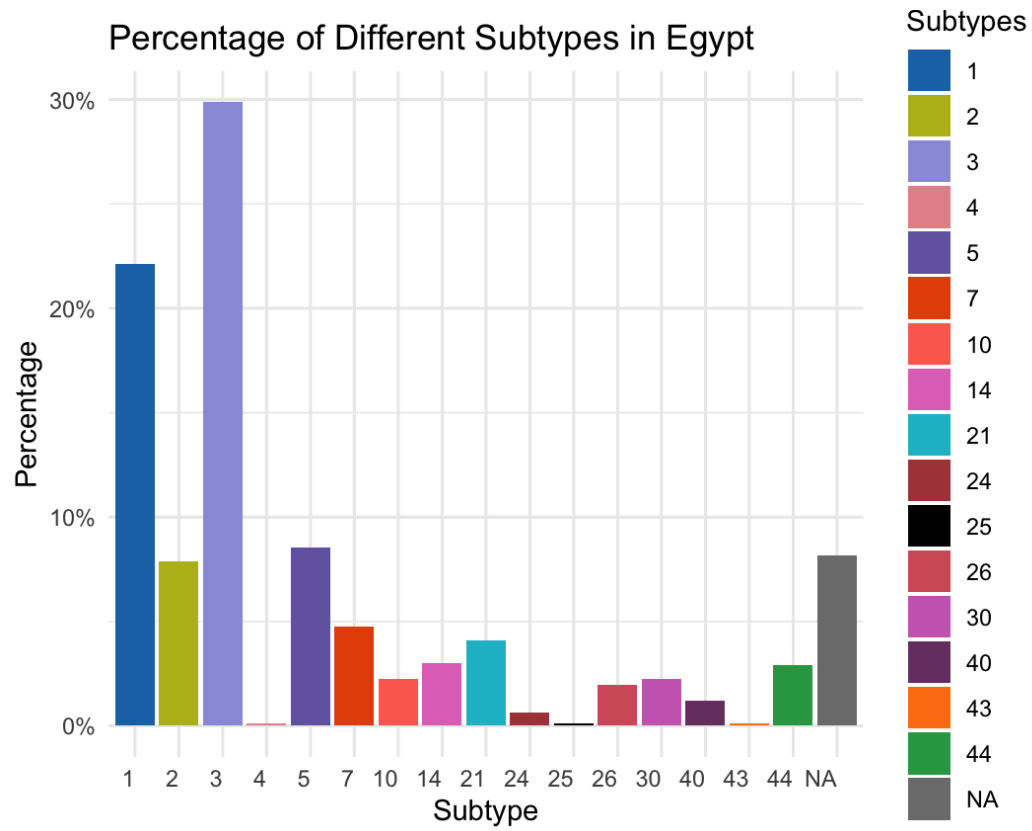
file:///Users/kanker/Downloads/thesis%20AS/interactive_sankey_subtype_continent_family.html

File is interactive and needs to be opened in a separate window. Sankey diagram showing connections between different categorical data; subtypes, continents and family. Hovering over nodes will give exact counts of the data and explain links. On the left, a list of all subtypes in this study. At the centre, a list of the continents and on the right, a list of the different host families detected.

Appendix IV: Supplemental Distribution Barplots



Plot showing Nigeria's subtype prevalence.



Plot showing Egypt's subtype prevalence.