Infection, Genetics and Evolution 73 (2019) 119–125

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Next generation amplicon sequencing improves detection of *Blastocystis* mixed subtype infections

Jenny G. Maloney, Aleksey Molokin, Monica Santin*

USDA ARS, Environmental Microbial and Food Safety Laboratory, BARC, Beltsville, MD, USA

ARTICLE INFO

Keywords:
Blastocystis
Next Generation Sequencing (NGS)
Mixed infections
Sanger sequencing
SSU rRNA gene
Within-host diversity

ABSTRACT

Blastocystis is a highly prevalent enteric protist parasite of humans and animals. Transmission occurs via the fecal-oral route through ingestion of contaminated food or water. Genetic diversity studies have identified numerous subtypes (STs) within the genus *Blastocystis* based on polymorphism at the SSU rRNA gene. Although there is evidence of frequent mixed subtype infections, the extent of within-host subtype diversity remains largely unexplored. Accurate assessment of *Blastocystis* ST diversity is crucial to understand epidemiology and sources of *Blastocystis* transmission to humans. Here, we report the application of next generation sequencing (NGS) for detection and characterization of *Blastocystis* subtypes to investigate intra-host *Blastocystis* diversity. A total of 75 specimens obtained from cattle feces, previously identified as *Blastocystis* positive, were examined using next generation amplicon sequencing. A fragment of the SSU rRNA gene was amplified using *Blastocystis*-specific primers and resulting amplicons were used for NGS. Comparison of Sanger and NGS results suggest greater sensitivity using the NGS approach. Using Sanger sequencing, mixed infections were suspected in 18 specimens but only confirmed through cloning in three, while NGS identified 49 mixed infections (16 times more). In addition, NGS revealed greater diversity of subtypes with 14 detected compared to 11 by Sanger. Nine more infections with potentially zoonotic STs were detected by NGS than Sanger. Indeed, subtype 3, the most common subtype found in humans, was found in 37% (28) of specimens tested by NGS but in only four specimens using Sanger. Our findings indicate that mixed *Blastocystis* infections may be far more common than previously thought due to the limitations of current detection methods. This next generation amplicon sequencing strategy improves detection of mixed subtype infections and low abundance subtypes and represents a valuable resource for future *Blastocystis* studies to improve our understanding of its epidemiology.

1. Introduction

Blastocystis is a widespread unicellular eukaryotic parasite which infects humans and many other animals. It is transmitted via the fecal-oral route through direct contact or consumption of cysts in contaminated food and water. Infection is associated with intestinal illness including diarrhea and abdominal pain although asymptomatic carriage is thought to be common. *Blastocystis* has also been associated with irritable bowel syndrome and chronic spontaneous urticaria (hives) (Ajjampur and Tan, 2016). *Blastocystis* is perhaps the most common intestinal parasite of humans, yet many important aspects of its epidemiology remain unexplored.

Blastocystis is currently divided into 26 subtypes (STs) based on polymorphism at the small subunit (SSU) rRNA gene (Alfellani et al., 2013; Zhao et al., 2017; Maloney et al., 2019). Genetic variability between subtypes in the SSU rRNA gene is substantial (at least 4–5%) (Clark et al., 2013), while variability within most *Blastocystis* subtypes is 1–2% (Stensvold et al., 2007; Fayer et al., 2012). Nine subtypes, ST-1 to ST-8 and ST-12, have been reported in humans and animals and are potentially zoonotic (Ramírez et al., 2016; Stensvold and Clark, 2016). *Blastocystis* subtypes are morphologically indistinguishable, and subtype identification requires molecular characterization. However, mixed infections may not be readily identified by molecular diagnostic tools currently used to detect *Blastocystis* because of preferential PCR amplification of the predominant subtypes. Sanger sequencing coupled with cloning has been previously used to detect mixed infections when mixtures were suspected from chromatograms(Yoshikawa et al., 2004; Scicluna et al., 2006; Santín et al., 2011; Scanlan et al., 2015). But this approach risks overlooking subtypes present in low abundance causing genetic heterogeneity within a specimen to be underreported even

* Corresponding author at: Environmental Microbial and Food Safety Laboratory, Agricultural Research Service, United States Department of Agriculture, BARC-East, Building 173, 10300 Baltimore Avenue, Beltsville, MD 20705, USA.

E-mail addresses: jenny.maloney@ars.usda.gov (J.G. Maloney), aleksey.molokin@ars.usda.gov (A. Molokin), monica.santin-duran@ars.usda.gov (M. Santin).

https://doi.org/10.1016/j.meegid.2019.04.013
Received 30 November 2018; Received in revised form 15 March 2019; Accepted 18 April 2019
Available online 23 April 2019
1567-1348/ Published by Elsevier B.V.

when cloning is used. In most epidemiological studies where genetic heterogeneity was reported, it was not the focus of the study but rather a casual finding (Santín et al., 2011; Fayer et al., 2012; Alfellani et al., 2013). *Blastocystis* mixed subtype infections have been reported to range between 1.1 and 14.3% of human samples surveyed with an average of 6% globally (Tan, 2008; Alfellani et al., 2013). However, only two studies have ever sought to directly address mixed infections. The first found that a single human host harbored three subtypes of *Blastocystis* by analyzing 50 clones (Meloni et al., 2012). The other study used a nested PCR assay targeting the four most common subtypes in humans (ST-1 through ST-4) and found that 22% of tested samples contained mixed infection even though they had only been found to contain a single subtype previously (Scanlan et al., 2015). To better understand *Blastocystis* transmission dynamics and to answer many unresolved epidemiological questions, it is necessary to improve detection of mixed infections.

Next generation sequencing (NGS) technologies provide powerful tools for the characterization and quantification of microbial communities. For *Blastocystis*, the SSU rRNA gene is currently the best phylogenetic marker available as it is the most commonly sequenced *Blastocystis* gene with publicly available reference sequences that include all subtypes previously reported. Next generation amplicon sequencing generates an immense number of individual sequences from a single sample allowing for highly sensitive population level analysis. In this study, we developed a next generation amplicon sequencing protocol targeting a fragment of the SSU rRNA gene that includes an analysis pipeline to efficiently detect both mixed subtype infections and low abundance subtypes in fecal samples. It also aims to compare conventional Sanger sequencing coupled with cloning and NGS for molecular subtype detection in *Blastocystis*-positive fecal samples.

2. Materials and methods

2.1. Source of samples

Seventy-five DNA samples previously screened for *Blastocystis* from cattle fecal specimens by a PCR that amplifies a ca 500 base pair fragment of the SSU rRNA gene were included in this study (Maloney et al., 2019). Briefly, each fecal specimen (15 g) was sieved to concentrate parasite forms by CsCl density centrifugation and total DNA was extracted from each CsCl-concentrated fecal sample using a modified version of the DNeasy Tissue Kit protocol (Qiagen, Valencia, CA) as previously described (Santin et al., 2004). Seventy-two samples were successfully identified as *Blastocystis*-positive and subtype identification was attained. Two additional samples were included that produced sequences of low quality that could not be subtyped or successfully cloned, and one additional sample was included that produced too little amplicon to be sequenced.

2.2. Sanger sequencing and cloning

All PCR products were purified using Exonuclease I/Shrimp Alkaline Phosphatase (Exo-SAP-IT™, USB Corporation, Cleveland, OH), and sequenced in both directions using the same PCR primers in 10 µl reactions, Big Dye™ chemistries, and an ABI 3130XL sequence analyzer (Applied Biosystems, Foster City, CA). Sequence chromatograms of each strand were aligned and examined with Lasergene software (DNASTAR, Inc., Madison, WI). If the sequence trace of a specimen indicated a potentially mixed infection, the SSU rRNA gene product was cloned using the TOPO TA cloning kit (Invitrogen Corp. Carlsbad, CA). Up to 16 clones per specimen were sequenced in both directions using M13 forward and reverse primers per the sequencing protocol previously described. Nucleotide sequences were compared with sequences in the GenBank database by BLAST analysis to identify *Blastocystis* subtypes.

2.3. Illumina MiSeq library preparation and sequencing

PCR was conducted using the same assay used for Sanger sequencing using primers Blast505_532F/Blast998_1017R as previously described (Santín et al., 2011), with the exception that primers were modified to contain the Illumina overhang adapter sequences on the 5′ end, forward primer ILMN_Blast505_532F 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGAGGTAGTGACAATAAATC -3′ and reverse primer ILMN_Blast998_1017R 5′- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGCTTTCGCACTTGTTCATC-3′ (adapter sequences underlined). These primers were selected because they amplify a variable region of the gene that is suitable for subtyping and have been shown to be highly sensitive and specific (Santín et al., 2011). To ensure primer sensitivity was not affected by the addition of the Illumina overhang adapter sequences, the primers were assayed using serially diluted *Blastocystis* DNA. The Illumina sequencing primers were able to detect a concentration as low as 0.0001 ng/ul which matches the reported sensitivity of the primers without the adapter sequence addition (Santín et al., 2011). The Illumina 16S Metagenomic Sequencing Library Preparation protocol (Part# 15044223 Rev. B) was used for library preparation with minor modifications to the amplicon PCR step. The amplicon PCR step used the following cycling conditions: 95 C for 4 min, 35 cycles of 95 C for 30 s, 54 C for 30s, and 72 C for 30 s, and a final elongation step at 72 C for 5 min and included 1.25 ul of BSA (0.1 g/10 ml). Following the amplicon PCR step, all samples were analyzed using a QIAxcel (Qiagen, Valencia, CA). A dual indexing strategy was used during preparation of sequencing libraries. This enabled sample multiplexing and ensured that each library contained a unique set of indexes by which they could be identified bioinformatically. Final libraries were quantified by Qubit fluorometric quantitation (Invitrogen, USA) prior to normalization. A final pooled library concentration of 8 pM with 20% PhiX control was sequenced on an Illumina MiSeq using 600 cycle v3 chemistry (300 base pair, paired-end reads) (Illumina, USA) following manufacturer's recommendations.

2.4. Bioinformatic analysis

Raw bcl files were demultiplexed using the Illumina MiSeq Reporter software. Paired end reads were processed and analyzed with an in-house pipeline that used the BBTools package v38.22 (Brian Bushnell, 2014), VSEARCH v2.8.0 (Rognes et al., 2016), and BLAST+ 2.7.1. Demultiplexed reads were trimmed on both ends using BBduk to remove residual Illumina adapters, indices, and primer dimers. Reads below a length of 200 bases were discarded. BBmerge was used to merge read pairs based on overlap to extend read length and improve quality of overlapping bases. Pairs that were not merged successfully on the first attempt were quality-trimmed at the 3′ end using a minimum Phred score of 20 followed by a second merge attempt. Merged reads below a length of 400 bases were filtered out and the remaining reads were subjected to quality filtering using the VSEARCH fastx_filter command and a maximum expected error threshold of 0.5. Filtered reads were then dereplicated within each sample with VSEARCH derep_fulllength. The dereplicated (unique reads) from all samples were pooled together to increase the sensitivity of chimera detection (Edgar, 2016). Dereplication was performed once more on pooled reads to obtain unique sequence abundances across all samples and singletons were removed. The VSEARCH uchime_denovo command was used to perform de novo chimera detection on pooled reads. Chimera-free sequences were then extracted from each dereplicated sample. After removing singletons, clustering and the assignment of centroid sequences to operational taxonomic units (OTU) was performed within each sample at a 98% identity threshold using the VSEARCH cluster_size command. Low abundance centroids were filtered out with fastx_filter using a minimum threshold of 100. The filtered OTUs were blasted against a reference database consisting of all *Blastocystis* sequences available from NCBI with the BLAST+ blastn command. Blast results

were filtered to remove hits below an alignment length of 400. All OTUs were assigned a *Blastocystis* subtype based on the best match from the BLAST results using the consensus subtype terminology (Stensvold et al., 2007). All raw fastq files were deposited to the NCBI sequence read archive under the accession number PRJNA507584. The nucleotide sequences for unique OTUs obtained in this study have been deposited in GenBank under the accession numbers MK244898-MK244953.

3. Results

3.1. Screening of samples and Sanger sequencing

A total of 75 DNA samples were selected from cattle fecal samples previously screened by PCR amplification of the SSU rRNA gene (Maloney et al., 2019). Seventy-four samples were *Blastocystis*-positive by PCR. Amplicon for sample #75 produced a signal too weak to be sequenced using Sanger, but a small peak with the correct product size was observed during screening on the QIAxcel electropherogram. All PCR products were Sanger sequenced and any nucleotide sequence with the appearance of a mixed infection (mixed chromatograms) was cloned. Of the 75 samples assayed, 72 were successfully subtyped by Sanger sequencing and cloning (Table 1) and three samples (63, 64, and 75) could not be successfully sequenced by direct Sanger sequencing or after cloning.

3.2. Illumina MiSeq sequencing data

A total of 19,959,561 read pairs were generated with an 84.3% pass filter rate. After demultiplexing and PhiX filtering10,705,916 read pairs remained. The percentage of reads greater than Q30 was 80.6% for read 1 and 76.9% for read 2. The error rate reported for reads 1 and 2 were 2.99% and 3.12% respectively. The samples described here comprised 7,633,848 read pairs (average of 101,785 ± 51,168 pairs per sample). After trimming, pair merging, and quality filtering there was a total of 3,680,920 reads (average of 49,709 ± 32,460 per sample). Chimera filtering further reduced the number of reads to 3,479,249. VSEARCH detected a total of 137,834 chimeric or borderline chimeric reads (2.7% of total filtered/non-singleton reads). Clustering yielded a total of 274 OTUs and 56 unique OTUs across the 75 samples after low abundance filtering with an average of 4 ± 2 OTUs per sample and an average OTU abundance of 46,390 ± 31,223 per sample. A blast search of all OTUs revealed that 1.38% of reads were not attributable to *Blastocystis*.

3.3. Comparison of Sanger sequencing and cloning with NGS for detection of mixed infections and low abundance subtypes

There was consistency in subtypes identified by Sanger and NGS in samples included in this study (Table 1). Indeed, subtypes identified from the 72 samples by Sanger matched subtypes found by NGS. For the three samples not successfully subtyped by Sanger sequencing and cloning, NGS detected the presence of ST-4 in two of them and ST-1/ST-24 in the other demonstrating how its greater sensitivity enables successful subtyping in samples with low abundance of the parasite.

NGS detected more subtypes than Sanger sequencing. Fourteen subtypes were detected by NGS (ST-1 to ST-5, ST-10, ST-11, ST-14, ST-17, ST-21, and ST-23 to ST-26) while three of those subtypes (ST-1, ST-2, and ST-11) were not found by Sanger sequencing. NGS detected 16 times more mixed infections than Sanger sequencing and cloning in the same population. Following direct Sanger sequencing, cloning was performed on 18 samples that were suspected to have mixed infections based on sequence chromatograms. However, cloning was only able to confirm mixed infection in three (4%) of those samples (72, 73, and 74). Using NGS, mixed infections with more than one subtype were detected in 49 (65%) samples. For the three mixed infections identified by cloning, NGS revealed a relatively even distribution of each subtypes present in the samples. NGS could resolve mixed infection amidst an overwhelming majority of one subtype while cloning was only able to detect the predominant subtype (Table 1). The high sequencing depth per sample obtained by NGS also improved the detection of multiple subtypes within a single sample (Fig. 1). NGS identified up to eight subtypes within an individual sample (sample 56) while cloning only recovered up to three (sample 72) (Table 1).

One or more of the potentially zoonotic subtypes (ST-1 to ST-5) identified in this study were present in 79% (59) of the samples by NGS but only found in 67% (50) by Sanger sequencing. Low abundance subtypes identified in mixed infections were frequently zoonotic (Table 1). Potentially zoonotic ST-5 was the most frequently observed subtype in this study and was detected in similar numbers by Sanger and NGS (27 by Sanger and 33 by NGS). Furthermore, ST-5 was frequently the most abundant subtype in a mixed sample as revealed by NGS (Table 1). In contrast, ST-3, the subtype most frequently reported in humans, was found in 28 (37%) samples by NGS but only in four (5%) samples by Sanger sequencing. In the four samples in which ST-3 was detected by Sanger sequencing, it was the most abundant or lone subtype detected by NGS (Table 1). However, in the 24 samples in which ST-3 was only detected by NGS, it ranged in abundance from 0.3% to 20.9% of the sample. ST-3 was the most commonly detected low abundance subtype in this study.

NGS analysis detected 56 unique OTUs among the 14 subtypes detected in this study indicating a high degree of within-subtype variability for *Blastocystis* (Table 2). ST-10 displayed the greatest sequence variability with 11 unique OTUs attributed to this subtype (Supplementary Appendix 1). Multiple ST-10 OTUs were also detected in the same sample with up to three different ST-10 sequence variants in a single sample (sample#73; see Supplementary Appendix 2). ST-5 and ST-14 also had a high level of sequence variability and multiple variants were found to co-occur in the same sample (e.g. samples 41 and 50; see Supplementary Appendix 2). Two subtypes, ST-2 and ST-23, were only detected once, but displayed no subtype variability. ST-17 was the only subtype detected in multiple samples that displayed no subtype variability (Table 2). These data demonstrate that within-subtype variability is quite common in *Blastocystis* and can be captured using NGS.

4. Discussion

Blastocystis is currently divided into 26 subtypes based on polymorphism at the SSU rRNA gene (Maloney et al., 2019). Although there is ample evidence in the literature of mixed subtype infections, the extent of within-host *Blastocystis* diversity remains mostly unexplored. This is due mainly to the lack of sensitive molecular tools capable of accurately detecting the presence of *Blastocystis* mixed infections in a sample. However, an accurate assessment of diversity in *Blastocystis* STs is key to understanding *Blastocystis* transmission, potential animal and environmental reservoirs, and pathogenicity. In this study, we compared NGS and Sanger sequencing coupled with cloning to detect *Blastocystis* subtypes and within-host genetic diversity in 75 *Blastocystis* positive fecal samples. Our results showed agreement between the two sequencing technologies and demonstrated that NGS was as accurate as Sanger sequencing for subtype detection while being a far more sensitive tool for the identification of mixed infections and the detection of low abundance subtypes (Table 1). Sanger sequencing accurately detects the most abundant subtype present in a sample, but it often misses mixed infections. Sanger and cloning successfully resolved mixed infections in three samples while NGS detected 49 mixed infections (16 times more than Sanger) demonstrating its ability to detect low abundance subtypes that are below the resolving power of Sanger sequencing.

A total of fourteen subtypes were detected using NGS (ST-1 to ST-5, ST-10, ST-11, ST-14, ST-17, ST-21, and ST-23 to ST-26), three more than Sanger. We were also able to detect and subtype *Blastocystis* in three samples which could not be sequenced by Sanger sequencing.

Table 1
Blastocystis subtypes identified using Sanger and next generation sequencing indicating presence of mixed infections and of potentially zoonotic subtypes in each of the samples obtained from cattle feces.

Sample ID	Mixed Infection identified by sanger	Subtype/s identified by sanger	Mixed infection identified by NGS	Subtype/s identified by NGS (% of reads)	Presence of potentially zoonotic ST
1	No	ST3	No	ST3(100)	Yes
2	No	ST3	No	ST3(100)	Yes
3	No	ST3	No	ST3(100)	Yes
4	No	ST4	No	ST4(100)	Yes
5	No	ST4	No	ST4(100)	Yes
6	No	ST4	No	ST4(100)	Yes
7	No	ST4	No	ST4(100)	Yes
8	No	ST4	No	ST4(100)	Yes
9	No	ST4	No	ST4(100)	Yes
10	No	ST4	No	ST4(100)	Yes
11	No	ST5	No	ST5(100)	Yes
12	No	ST5	No	ST5(100)	Yes
13	No	ST5	No	ST5(100)	Yes
14	No	ST5	No	ST5(100)	Yes
15	No	ST5	No	ST5(100)	Yes
16	No	ST5	No	ST5(100)	Yes
17	No	ST14	No	ST14(100)	No
18	No	ST17	No	ST17(100)	No
19	No	ST3	Yes	ST3(99.3)/ST5(0.4)/ST11(0.3)	Yes
20	No	ST4	Yes	ST4(98.9)/ST5(1.1)	Yes
21	No	ST4	Yes	ST4(97.2)/ST26(2.8)	Yes
22	No	ST4	Yes	ST4(99.9)/ST14(0.1)	Yes
23	No	ST4	Yes	ST4(96)/ST26(1.6)/ST10[a](1.5)/ST24(1)	Yes
24	No	ST4	Yes	ST4(99.6)/ST3(0.4)	Yes
25	No	ST5	Yes	ST5(93)/ST3(7)	Yes
26	No	ST5	Yes	ST5(98.9)/ST3(1.1)	Yes
27	No	ST5	Yes	ST5(89.7)/ST3(10.3)	Yes
28	No	ST5	Yes	ST5(88.6)/ST3(11.1)/ST26(0.3)	Yes
29	No	ST5	Yes	ST5(90.2)/ST3(9.8)	Yes
30	No	ST5	Yes	ST5(79.1)/ST3(20.9)	Yes
31	No	ST5	Yes	ST5(98.6)/ST3(1.4)	Yes
32	No	ST5	Yes	ST5(92.5)/ST3(7.5)	Yes
33	No	ST5	Yes	ST5(97.4)/ST3(2.6)	Yes
34	No	ST5	Yes	ST5(95.5)/ST3(3.5)/ST1(1)	Yes
35	No	ST5	Yes	ST5(99.4)/ST3(0.6)	Yes
36	No	ST5	Yes	ST5(94.5)/ST3(5.5)	Yes
37	No	ST5	Yes	ST5(93.4)/ST3(6.6)	Yes
38	No	ST5	Yes	ST5(94.4)/ST3(4.4)/ST2(1.2)	Yes
39	No	ST5	Yes	ST5(94)/ST3(6)	Yes
40	No	ST5	Yes	ST5(98.8)/ST3(1.2)	Yes
41	No	ST5	Yes	ST5[a](98)/ST3(2)	Yes
42	No	ST5	Yes	ST5(98.8)/ST14(1.2)	Yes
43	No	ST10	Yes	ST10(99)/ST24(1)	No
44	No	ST10	Yes	ST10(97.7)/ST24(1.8)/ST14(0.5)	No
45	No	ST10	Yes	ST10(100)	No
46	No	ST14	Yes	ST14(98.9)/ST5(1.1)	Yes
47	No	ST14	Yes	ST14(84.6)/ST25(11.5)/ST21(3.2)/ST24(0.7)	No
48	No	ST14	Yes	ST14(92.3)/ST21(3)/ST24(2.7)/ST5(1.6)/ST26(0.4)	Yes
49	No	ST14	Yes	ST14(94.6)/ST25(2.2)/ST21(1.2)/ST24(1.9)/ ST26(0.4)/ST10(0.3)	No
50	No	ST14	Yes	ST14[a](79.2)/ST25(13.6)/ST21(4.6)/ST3(1)/ ST10[a](1.3)/ ST26(0.5)	Yes
51	No	ST17	Yes	ST17(96.5)/ST10(3.5)	No
52	No	ST21	Yes	ST21(80.5)/ST25(18.5)/ST14(0.5)/ST10(0.4)	No
53	No	ST21	Yes	ST21(99.8)/ST24(0.2)	No
54	No	ST24	Yes	ST24(98.3)/ST10(1.4)/ST26(0.4)	No
55	No	ST25	Yes	ST25(72.7)/ST14(25.1)/ST21(1)/ST24(0.8)/ ST26(0.4)	No
56	No	ST26	Yes	ST26(84.6)/ST10(10.3)/ST25(1.7)/ST21(1.3)/ ST14(1.2)/ ST24(0.4)/ST11(0.3)/ST3(0.3)	Yes
57	Suspected/not confirmed	ST4	No	ST4(100)	Yes
58	Suspected/not confirmed	ST4	No	ST4(100)	Yes
59	Suspected/not confirmed	ST4	No	ST4(100)	Yes
60	Suspected/not confirmed	ST4	No	ST4(100)	Yes
61	Suspected/not confirmed	ST5	No	ST5(100)	Yes
62	Suspected/not confirmed	ST26	No	ST26(100)	No
63	Suspected/not confirmed	untypable	No	ST4(100)	Yes
64	Suspected/not confirmed	untypable	No	ST4(100)	Yes
65	Suspected/not confirmed	ST26	Yes	ST26(94.6)/ST5(5.4)	Yes
66	Suspected/not confirmed	ST4	Yes	ST4(98.9)/ST3(0.8)/ST10(0.4)	Yes
67	Suspected/not confirmed	ST4	Yes	ST4(96.8)/ST3(3.2)	Yes
68	Suspected/not confirmed	ST5	Yes	ST5(99.7)/ST3(0.3)	Yes

(*continued on next page*)

Table 1 (*continued*)

Sample ID	Mixed Infection identified by sanger	Subtype/s identified by sanger	Mixed infection identified by NGS	Subtype/s identified by NGS (% of reads)	Presence of potentially zoonotic ST
69	Suspected/not confirmed	ST5	Yes	ST5(87.2)/ST3(11.8)/ST1(1)	Yes
70	Suspected/not confirmed	ST14	Yes	ST14(96.8)/ST10(2.1)/ST4(0.7)/ST5(0.4)	Yes
71	Suspected/not confirmed	ST17	Yes	ST17(85.6)/ST14(10.4)/ST1(4)	Yes
72	Suspected/confirmed	ST10/ST14/ST26	Yes	ST10(77)/ST26(18)/ST14(3.9)/ST24(1.1)	No
73	Suspected/confirmed	ST23/ST26	Yes	ST10[a](49.9)/ST23(29.6)/ST26(15.4)/ST14(3.6)/ ST24(1.5)	No
74	Suspected/confirmed	ST10/ST24	Yes	ST10[a](62.5)/ST24(37.5)	No
75	Not sequenced	untypable	Yes	ST1(52.3)/ST24(47.7)	Yes

[a] Denotes intra-subtype variability.

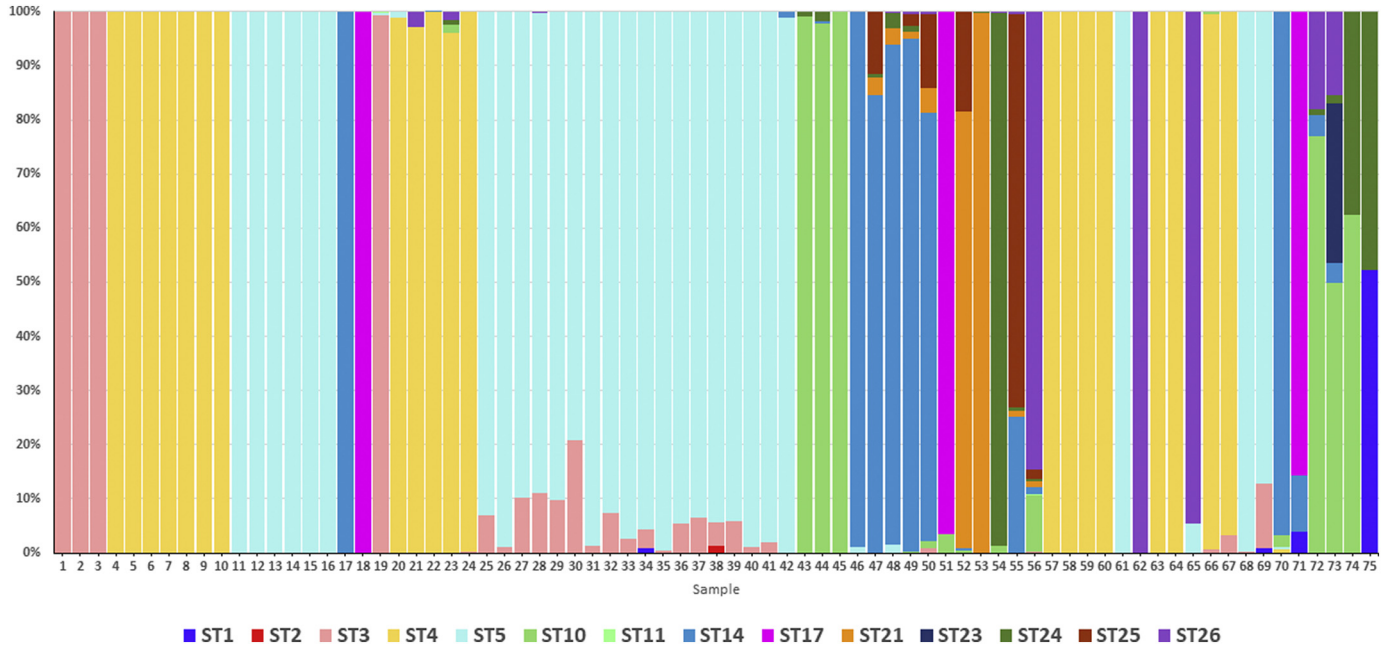

Fig. 1. *Blastocystis* subtype distribution identified by next generation sequencing in each sample included in this study.

Table 2
Blastocystis subtypes identified in this study including number of samples for which each subtype was identified and number of unique operational taxonomic units (OTUs) among the subtypes identified using next generation sequencing.

Subtype	No. of samples	No. of unique OTU
ST1	4	4
ST2	1	1
ST3	28	4
ST4	21	2
ST5	32	6
ST10	15	11
ST11	2	2
ST14	16	5
ST17	3	1
ST21	8	2
ST23	1	1
ST24	14	6
ST25	6	2
ST26	13	9

Two of these samples produced robust amplicons by PCR, but could never be successfully sequenced or cloned. Interestingly, both samples were found to only contain a single subtype of *Blastocystis* by NGS. Off target amplification is a common problem when performing PCR on fecal and environmental samples as the content of the DNA extracted from these materials can be quite complex. NGS and data processing allowed contaminant sequences to be removed while retaining the sequences of interest. This provides a great advantage when working with difficult source material. The third sample produced too little amplicon to be adequate for Sanger sequencing. However, this sample was successfully sequenced using NGS, demonstrating its superior sensitivity compared to Sanger sequencing.

In this study, many of the low abundance subtypes detected by NGS were one of the potentially zoonotic subtypes (ST-1 to ST-5). ST-5 was the most frequently observed subtype in this study by both Sanger and NGS (27 and 33 samples, respectively). Whereas ST-3 was detected in 4 samples by Sanger but in 28 samples by NGS. This is of interest, as ST-3 is the most common subtype observed in human fecal samples and it has been associated with human disease (Jones et al., 2009; Wang et al., 2014). The improved detection of ST-3 by NGS demonstrates the value of utilizing a more sensitive detection method in studies that seek to determine the molecular epidemiology of *Blastocystis* and could improve our understating of host range and transmission dynamics. Some of the potentially zoonotic subtype infections reported here were found in low relative abundance. As these data represent only one point in time, it is difficult to interpret the relevance of these infections. However, it is clear that mixed subtype infections and low abundance subtypes should be better studied and characterized in the future to better understand *Blastocystis* epidemiology.

The samples examined in this study frequently contained multiple subtypes. Of the 75 samples examined, 49 (65%) contained two or more subtypes. However, this degree of subtype mixing could only be

detected by NGS. Only 3 mixed subtype infections were successfully detected by Sanger sequencing and cloning, although more were suspected from the Sanger sequencing chromatograms. It has been hypothesized that subtype level differences could be responsible for the variability in infection outcomes. One of the advantages of using NGS for subtyping is that it would allow for a more complete picture of intra-host subtype variability and shed light on the potential role of mixed subtype infections on host infection outcome.

The primers used for Sanger sequencing and NGS amplify the same region of the *Blastocystis* SSU rRNA gene. The use of these primers in NGS requires the addition of Illumina overhang adapter sequences so that amplicons can be indexed. Primer sensitivity and PCR reaction efficiency can be affected by altering primer sequence. However, the NGS primers performed as well as unaltered primers, and sensitivity was shown to be unaffected by the addition of the overhang. For this reason, the NGS primers could be used for both screening and NGS library preparation, allowing for both sample and reagent resources to be conserved. These primers are also ideal for sequencing on the Illumina MiSeq as they produce an amplicon that is ~500 bp and particularly well suited for the 600 cycle v3 chemistry that generates paired end reads that are 300 bp in length. The 3′ ends of each read pair overlap to produce high quality, full-length sequences which include a variable region that can be used for subtyping.

Bioinformatic analysis of NGS sequences provided a challenge in this study as no comprehensive reference database for *Blastocystis* currently exists. For this reason, we chose to be extremely conservative in our methodology for OTU generation and subtype assignment. Only long ($\geq$400 bp), high quality, chimera-free sequences were retained for OTU generation. A minimum depth of at least 100 sequences further limited the likelihood of retaining erroneous OTUs. A similarity threshold of 98% for sequence clustering was chosen as it most accurately recapitulates the subtype sequences obtained by Sanger sequencing while minimizing the occurrence of low frequency OTUs of questionable validity. This is an important concern as intra-isolate heterogeneity in the SSU rRNA gene has been postulated (Poirier et al., 2014). It was reported that *Blastocystis* ST-7 strain B had 17 non-identical copies of the SSU rRNA gene present in the whole genome sequence (Denoeud et al., 2011; Poirier et al., 2014). The sequence identity between copies was reported to be between 99.9% and 98.1% (Poirier et al., 2014). Therefore, the 98% similarity threshold used for clustering in this study should mollify the concern that the subtypes reported here are attributable to intra-isolate sequence variability.

Intra-subtype variability was common in this study, although it varied widely by subtype (Table 2). ST-10 had the most unique OTUs, and ST-5, ST-10, ST-14 were the only subtypes with multiple variants present in a single sample (Tables 1 and 2). Intra-subtype variability has been reported in *Blastocystis* (Clark, 1997; Fayer et al., 2012, 2014). In fact, different *Blastocystis* subtypes likely represent separate species (Yoshikawa et al., 2016). However, the importance of this variability in infection outcome is still unclear. A lower level of intra-subtype variability has been reported for some *Blastocystis* subtypes such as ST-3 and ST-4 (Beghini et al., 2017), and indeed we observe that same trend in this study. Only four unique OTUs were observed among the ST-3 sequences in this study despite it being one of the most commonly observed subtypes. ST-4 and ST-17 also displayed low intra-subtype variability with only two and one unique OTUs respectively.

A sensitive and accurate detection method is crucial for the identification of parasites from fecal matter or environmental samples. NGS provides an important tool for better characterization of the true diversity of *Blastocystis* from these types of samples by generating a massive number of sequences in a high-throughput manner. We have demonstrated that NGS is useful for detecting low abundance subtypes and mixed subtype *Blastocystis* infections. Traditionally, suspected mixed infections are resolved through cloning of PCR amplicons and Sanger sequencing of multiple clones. This process is labor intensive, time consuming, and expensive. As the cost of NGS technology continues to decrease, it is more cost effective than cloning when mixed subtype detection is needed for many samples. It also saves time and conserves valuable samples. It is evident from this study that mixed subtype infections may be far more common than previously thought. The sensitive detection method described here could aid in improving our understanding of *Blastocystis* epidemiology.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of interest

None.

Disclaimer

USDA is an equal opportunity provider and employer.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2019.04.013.

References

Ajjampur, S.S.R., Tan, K.S.W., 2016. Pathogenic mechanisms in Blastocystis spp. — interpreting results from in vitro and in vivo studies. Parasitol. Int. 65, 772–779. https://doi.org/10.1016/j.parint.2016.05.007.

Alfellani, M.A., Stensvold, C.R., Vidal-Lapiedra, A., et al., 2013. Variable geographic distribution of Blastocystis subtypes and its potential implications. Acta Trop. 126, 11–18. https://doi.org/10.1016/J.ACTATROPICA.2012.12.011.

Beghini, F., Pasolli, E., Truong, T.D., et al., 2017. Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. ISME J. 11, 2848–2863. https://doi.org/10.1038/ismej.2017.139.

Brian Bushnell, 2014. BBMap Download | SourceForge.net.

Clark, C.G., 1997. Extensive genetic diversity in Blastocystis hominis. Mol. Biochem. Parasitol. 87, 79–83. https://doi.org/10.1016/S0166-6851(97)00046-7.

Clark, C.G., van der Giezen, M., Alfellani, M.A., Stensvold, C.R., 2013. Recent developments in *Blastocystis* research. Adv. Parasitol. 82, 1–32. https://doi.org/10.1016/B978-0-12-407706-5.00001-0.

Denoeud, F., Roussel, M., Noel, B., et al., 2011. Genome sequence of the stramenopile Blastocystis, a human anaerobic parasite. Genome Biol. 12, R29. https://doi.org/10.1186/gb-2011-12-3-r29.

Edgar, R.C., 2016. UNOISE2: Improved Error-correction for Illumina 16S and ITS Amplicon Sequencing.

Fayer, R., Santin, M., Macarisin, D., 2012. Detection of concurrent infection of dairy cattle with Blastocystis, Cryptosporidium, Giardia, and Enterocytozoon by molecular and microscopic methods. Parasitol. Res. 111, 1349–1355. https://doi.org/10.1007/s00436-012-2971-1.

Fayer, R., Elsasser, T., Gould, R., et al., 2014. Blastocystis tropism in the pig intestine. Parasitol. Res. 113, 1465–1472. https://doi.org/10.1007/s00436-014-3787-y.

Jones, M.S., Whipps, C.M., Ganac, R.D., et al., 2009. Association of Blastocystis subtype 3 and 1 with patients from an Oregon community presenting with chronic gastrointestinal illness. Parasitol. Res. 104, 341–345. https://doi.org/10.1007/s00436-008-1198-7.

Maloney, J.G., Lombard, J.E., Urie, N.J., et al., 2019. Zoonotic and genetically diverse subtypes of Blastocystis in US pre-weaned dairy heifer calves. Parasitol. Res. 118, 575–582. https://doi.org/10.1007/s00436-018-6149-3.

Meloni, D., Poirier, P., Mantini, C., et al., 2012. Mixed human intra- and inter-subtype infections with the parasite Blastocystis sp. Parasitol. Int. 61, 719–722. https://doi.org/10.1016/J.PARINT.2012.05.012.

Poirier, P., Meloni, D., Nourrisson, C., et al., 2014. Molecular subtyping of Blastocystis spp. using a new rDNA marker from the mitochondria-like organelle genome. Parasitology 141, 670–681. https://doi.org/10.1017/S0031182013001996.

Ramírez, J.D., Sánchez, A., Hernández, C., et al., 2016. Geographic distribution of human Blastocystis subtypes in South America. Infect. Genet. Evol. 41, 32–35. https://doi.org/10.1016/j.meegid.2016.03.017.

Rognes, T., Flouri, T., Nichols, B., et al., 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4, e2584. https://doi.org/10.7717/peerj.2584.

Santin, M., Trout, J.M., Xiao, L., et al., 2004. Prevalence and age-related variation of Cryptosporidium species and genotypes in dairy calves. Vet. Parasitol. 122, 103–117. https://doi.org/10.1016/j.vetpar.2004.03.020.

Santín, M., Gómez-Muñoz, M.T., Solano-Aguilar, G., Fayer, R., 2011. Development of a new PCR protocol to detect and subtype Blastocystis spp. from humans and animals. Parasitol. Res. 109, 205–212. https://doi.org/10.1007/s00436-010-2244-9.

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

Scicluna, S.M., Tawari, B., Clark, C.G., 2006. DNA barcoding of Blastocystis. Protist 157, 77–85. https://doi.org/10.1016/J.PROTIS.2005.12.001.

Stensvold, C.R., Clark, C.G., 2016. Current status of Blastocystis: a personal view. Parasitol. Int. 65, 763–771. https://doi.org/10.1016/j.parint.2016.05.015.

Stensvold, C.R., Suresh, G.K., Tan, K.S.W., et al., 2007. Terminology for Blastocystis subtypes – a consensus. Trends Parasitol. 23, 93–96. https://doi.org/10.1016/J.PT.2007.01.004.

Tan, K.S.W., 2008. New insights on classification, identification, and clinical relevance of Blastocystis spp. Clin. Microbiol. Rev. 21, 639–665. https://doi.org/10.1128/CMR.00022-08.

Wang, W., Owen, H., Traub, R.J., et al., 2014. Molecular epidemiology of Blastocystis in pigs and their in-contact humans in Southeast Queensland, Australia, and Cambodia. Vet. Parasitol. 203, 264–269. https://doi.org/10.1016/J.VETPAR.2014.04.006.

Yoshikawa, H., Wu, Z., Kimata, I., et al., 2004. Polymerase chain reaction-based genotype classification among human Blastocystis hominis populations isolated from different countries. Parasitol. Res. 92, 22–29. https://doi.org/10.1007/s00436-003-0995-2.

Yoshikawa, H., Koyama, Y., Tsuchiya, E., Takami, K., 2016. Blastocystis phylogeny among various isolates from humans to insects. Parasitol. Int. 65, 750–759. https://doi.org/10.1016/J.PARINT.2016.04.004.

Zhao, G.H., Hu, X.F., Liu, T.L., et al., 2017. Molecular characterization of Blastocystis sp. in captive wild animals in Qinling Mountains. Parasitol. Res. 116, 2327–2333. https://doi.org/10.1007/s00436-017-5506-y.