Finding *Sarcocystis* spp. on the Tioman Island: 28S rRNA gene next-generation sequencing reveals nine new *Sarcocystis* species

Florence C. H. Lee

ABSTRACT

The Tioman Island of Malaysia experienced acute muscular sarcocystosis outbreaks from 2011 to 2014. So far, a previous study based on the 18S rRNA gene sequencing has reported *S. singaporensis*, *S. nesbitti* and *Sarcocystis* sp. YLL-2013 in water samples acquired from the island, thus confirming the waterborne nature of this emerging parasitic disease. This study aimed to improve the detection methods for *Sarcocystis*, in order to have a clearer picture of the true diversity of *Sarcocystis* species in Tioman. A new primer set (28S R7F–28S R8 Deg R) was designed to amplify the 28S rRNA gene of *Sarcocystis*. Subsequently, *Sarcocystidae* was detected in 65.6% (21/32) of water samples and 28% (7/25) of soil samples acquired between 2014 and 2015 from Tioman. Next-generation sequencing (NGS) on 18 of the positive samples was then performed using amplicons generated from the same primer set. This yielded 53 potentially unique *Sarcocystidae* sequences (290 bp), of which nine of the most abundant, prevalent and unique sequences were named herein. In contrast, NGS of the 18S rRNA gene V9 hypervariable region of 10 selected samples detected only two *Sarcocystis* species (160 bp). *S. mantioni* was the most ubiquitous sequence found in this study.

Key words | 28S, NGS, *Sarcocystis*, sarcocystosis, Tioman, V9

Florence C. H. Lee
Environmental Health Research Centre,
Institute for Medical Research (IMR), Ministry of Health Malaysia,
Jalan Pahang, 50588 Kuala Lumpur,
Malaysia
E-mail: *florencelee@imr.gov.my*

INTRODUCTION

Sarcocystosis is a parasitic disease caused by members of the *Sarcocystis* genus. With more than 200 species, an intermediate–definitive host (diheteroxenous) life cycle (Fayer 2004), and occasionally being dihomoxenous (Bannert 1994; Koudela & Modrý 2000), the *Sarcocystis* genus, is ubiquitous in various environments and hosts. Numerous findings of sarcocystosis among wildlife and livestock have been reported (Latif *et al.* 1999; Dubey *et al.* 2015; Gjerde *et al.* 2018). In humans, it can take the form of intestinal or muscular sarcocystosis.

Between 2011 and 2014, a series of acute muscular sarcocystosis had occurred among tourists who travelled to the tropical islands of Tioman and Pangkor, Malaysia (CDC 2012; Esposito *et al.* 2012; Esposito *et al.* 2014; Italiano *et al.* 2014). With confirmation from just a handful of biopsy samples, *S. nesbitti* was found to be the only known species to cause human muscular sarcocystosis on these two islands (Tappe *et al.* 2013; Lau *et al.* 2014). Water and food are stipulated to be the transmission routes of sarcocystosis on the Malaysian islands (Dubey 2015). Early surveillance monitoring was carried out in Tioman, but finding *Sarcocystis* among the environmental samples was elusive (Husna Maizura *et al.* 2012). A subsequent study in Tioman that examined various water samples taken from Salang, Juara and Ayer Batang reported the detection of *S. singaporensis*,

doi: 10.2166/wh.2019.124

S. nesbitti, *Sarcocystis* sp. YLL-2013 and one unknown *Sarcocystis* species based on the 18S rRNA gene sequences (Shahari *et al.* 2016).

This study continues the quest of finding *Sarcocystis* in Tioman, using next-generation sequencing (NGS) to target a region in the 28S rRNA gene with a new primer set 28S R7F–28S R8 Deg R, and the V9 hypervariable region of the 18S rRNA gene. With the deep sequencing ability of NGS and identification of unique 28S rRNA gene sequences, the presence of more *Sarcocystis* species in the environmental samples from Tioman was revealed. The sequence of primary polymerase chain reaction (PCR) products from initial screening examinations, and cloning results when available are referred in the reporting of these NGS acquired sequences. In addition, the corresponding amplified regions (R7–R8 and V9) belonging to eight in-house unknown *Sarcocystidae* sequences with long nucleotides from the 18S to 28S rRNA gene were extracted for comparison to assign a full-length ribosomal DNA sequence to the matching NGS sequence.

MATERIALS AND METHODS

Environmental sampling and DNA extraction

Samplings were conducted in four main locations on the Tioman Island, L1 to L4, in October 2014 and August 2015. Nine and 23 grab water samples and 19 and six grab soil samples were collected, respectively, in the two consecutive years, totalling up to 32 grab water samples and 25 grab soil samples. Samples were acquired from rivers, estuaries, mangrove and bay. The samples were kept in sterile Nasco Whirl-Pak bags, and stored on ice in cool boxes or refrigerated until being processed. Samples acquired in 2014 were processed within two months, and within two weeks for the 2015 samples. DNA was extracted from each water sample in duplicates; each replicate with 100 ml of water content aspirated from the sampling bags, filtered through 0.45 μm cellulose nitrate membrane filters (Sartorius, Germany) using the Millipore vacuum manifold system, and followed by genomic DNA extraction with PowerWater DNA Isolation Kit (Mo Bio, USA). DNA was extracted from each soil sample in triplicates; the first and second replicates each with genomic DNA extracted from 0.25 g of soil sample using the DNeasy PowerSoil DNA Isolation Kit (Mo Bio, USA), and the third replicate from 10 g of soil samples using the PowerMax Soil DNA Isolation Kit (Mo Bio, USA). Extracted DNA was stored at −20 °C. A muscle sample from water monitor lizard provided by The Tioman Island *Sarcocystosis* Investigation Team (Esposito *et al.* 2014) served as the positive control in this study.

DNA extraction from the domestic water filter

One domestic ceramic water filter was acquired from a household during the 2014 sampling trip. The device filtered natural surface water and groundwater collected for household usage. It was soaked overnight in a 1 L glass measuring cylinder containing 400 ml of sterilised artificial groundwater diluent. The diluent contained 0.284 g $MgSO_4{\cdot}7H_2O$, 0.032 g NaCl, and 0.266 g $CaCl_2{\cdot}2H_2O$ per litre of ultrapure water (Charles *et al.* 2009). The content in the cylinder was transferred into a sterile stainless steel container. The debris (filtered material) on the water filter was scrapped into the soaked content using a sterile stainless steel spatula. The content was then divided into eight 50 ml Corning centrifuge tubes and sonicated for 20 min to dislodge microorganisms that possibly had clumped up among the silt, followed by ultracentrifugation of the tubes at 13,000 g. The supernatant of two tubes was pooled together for genomic DNA extraction using the PowerWater DNA Isolation Kit (Mo Bio, USA). Soil pellet from each two centrifuge tubes was pooled into one PowerBead tube of the PowerMax Soil DNA Isolation Kit (Mo Bio, USA) using the PowerBead Solution, followed by genomic DNA extraction.

Sarcocystidae PCR screening targeting a new 28S rRNA gene region, R7–R8

The extracted DNA of all the technical replicates from their respective samples were screened for the presence of *Sarcocystidae* through PCR, using a newly designed primer pair (Primer3Plus, http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) consisting of forward primer 28S R7 F (5′-TCGCGTTCATCGGGATTTGA-3′) and reverse primer 28S R8 Deg R (5′-TTTCAAGACGGGTCGGTTRA-3′). The expected product size is about 480 bp, based on *Sarcocystidae* sequences retrieved from the GenBank

database. The PCR contained 25 μl of Q5 High Fidelity 2X Master Mix (New England Biolabs), 2.5 μl (10 μM) of each forward and reverse primer, 3 μl of genomic DNA, and 17 μl of Nuclease-Free Water (Qiagen, USA). PCR was performed with the following conditions: 98 °C for 3 min of initial denaturation, 40 cycles of amplification at 98 °C for 10 s, 65 °C for 45 s, 72 °C for 45 s; final elongation at 72 °C for 3 min and hold at 4 °C. The PCR products were screened with FlashGel DNA Cassette (1.2% agarose, Lonza, Switzerland). Presumptively positive samples (showing bands of about 500 bp in size) were electrophoresised on 1.5% agarose gel (HydraGene Instant Agarose Tablet, USA). The respective bands were excised and sent for Sanger sequencing (Genomics, Taiwan) to confirm the presence of *Sarcocystidae*. An environmental sample with at least one technical replicate confirmed as harbouring *Sarcocystidae* by Sanger sequencing was considered positive for *Sarcocystidae*.

NGS targeting the R7–R8 region of the 28S rRNA gene

PCR using primers 28S R7F–28S R8 Deg R that contained the Illumina NGS platform adapter, 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTTCAAGACGGGTCGGTTRA-3′ and 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCGCGTTCATCGGGATTTGA-3′ were carried out. About 36 μl of each of the PCR products was electrophoresised. Bands of the expected size were excised, purified with Monarch DNA Gel Extraction Kit (New England Biolabs) and sequenced to confirm that *Sarcocystidae* has been amplified. The purified PCR products were then subjected to NGS library preparation using the Nextera XT Index Kit (Illumina, USA) and HiFi HotStart ReadyMix PCR Kit (Kapa Biosystems, USA), undergoing a reduced PCR cycle of eight. The NGS libraries were then quantified with the Kapa Library Quantification Kit (Kapa Biosystems, USA) using the Applied Biosystems 7500 Real-Time PCR Systems, normalised to 1.4 nM working concentration, and loaded into the Illumina Miseq System following the manufacturer's instruction, using the Miseq Reagent Nano Kit, v2 (300 cycles) and the PhiX Control Kit v3 (Illumina, USA). The reverse primer 28S R8 Deg R was assigned to the Read 1 adapter to enable single-end reading from the antisense direction, as public database information reveals a more diverse region from this direction among *Sarcocystidae*. For quality assurance, four replicates were also sent for paired-end 500 cycles NGS (BioBasic, Canada).

NGS targeting the V9 hypervariable region of the 18S rRNA gene, PCR product cloning, and in-house data reference

Ten candidates were sent for paired-end NGS of the V9 hypervariable region in the 18S rRNA gene (1st Base, Singapore), using the 1380F and 1510R universal primers. Seven of these 10 candidates had also been used for 28S rRNA gene NGS, either as the same technical replicate or different replicates of the same sample.

PCR product cloning was performed using either the pGC Blue Cloning & Amplification Kits (Lucigen, USA) with the *E. cloni* 10G Supreme Electrocompetent Cells, or NEB PCR Cloning Kit with NEB 10-Beta chemically competent cells (New England Biolabs). Samples R-2015, J-2015, Sarco, H-2015 and I-2015 were cloned through technical replicates R2-2015, J2-2015, Sarco, H2-2015 and I2-2015, and the products were sent for Sanger sequencing (Genomics, Taiwan). The first three technical replicates have also undergone 28S NGS, and replicate H1-2015 for 18S V9 NGS.

There were eight in-house unknown *Sarcocystidae* sequences, namely Combi, WFMax1B, WF2B, C1.9, C2.1, C2.5, 50.20Jun and J1.11. These sequences were acquired through the cloning of PCR products that amplified genomic DNA from the 18S to the 28S rRNA gene (submitted), using the same samples from this study. Their nucleotide fragments delineated by primers 28S R7F–28S R8 Deg R and 1380F–1510R were extracted for matching with the NGS reads.

NGS data analysis

The NGS data output was analysed using the USEARCH commands available from https://www.drive5.com/usearch/manual/cmds_all.html (accessed between July 2017 and April 2018). The decompressed fastq files were filtered with the criteria of maximum expected error rate (maxee) of 1.0, truncated into a suitable length of nucleotides and saved as fasta files. Paired-end fastq files were merged first before the similar quality filtration step. All quality filtered fasta files belonging to the 28S rRNA gene were pooled into a single fasta file to produce unique

reads from the total libraries, followed by generation of Zotus (Zero-radius Operational Taxonomic Unit) or Otus (Operational Taxonomic Unit), and mapping of the respective raw fastq files to the Zotus or Otus. Similar steps were performed for the paired-end reads of 18S V9 region. The command lines for NGS data processing from fastq files to alpha and beta analysis based on Zotus generation are presented in Supplementary Material S1 (available with the online version of this paper). Otus were produced using the cluster_otus command (not shown).

Sequence display, alignment, analysis and editing were carried out using Mega 7 (Kumar *et al.* 2016), Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/), NCBI Tree Viewer (https://www.ncbi.nlm.nih.gov/projects/treeview/) and iTOL (Letunic & Bork 2011). Four alpha diversity metrics were reported, namely Richness, Chao1, Simpson and Berger-Parker. The Richness metrics of the amplified 28S and 18S regions are portrayed as rarefaction curves with Excel. Bray-Curtis that is based on abundance was used as the main beta diversity analysis metric, supplemented by the Jaccard metric that is based on presence/absence. The phylogenetic tree was produced from aligned sequences with the Neighbour Joining method and 100 bootstrap replicates.

RESULTS

Sarcocystidae prevalence and NGS results for the 28S R7–R8 region

Sarcocystidae prevalence among the environmental samples acquired between 2014 and 2015 showed that 65.6% of the 32 grab water samples (5/9 in 2014 and 16/23 in 2015) and 28% of the 25 grab soil samples (6/19 in 2014 and 1/6 in 2015) contained *Sarcocystidae*. Note that the water filter did not count towards these prevalence statistics. Due to resource limitation, not all *Sarcocystidae*-positive technical replicates were subjected to NGS. Eighteen samples were chosen for the 28S rRNA gene NGS library preparation, producing 25 libraries, among which were three pairs of sample replicates, J1-2015 and J2-2015; S1-2015 and S2-2015; Sarco_S10 and Sarco.mar; and one group of related replicates that were processed from the water filter, namely WF-1A, WF-1B, WF-2A, Dmax.mar and WF-Sup. The reads and diversity metrics of these libraries are presented in Table 1. Reads from libraries D7.mar, Dmax.mar, Sarco.mar and D2.mar (paired-end sequencing) were trimmed down to 480 bp, whereas reads from the other 21 libraries (single-end sequencing) were trimmed down to 290 bp. Libraries Sarco_S10 (300 cycles) and Sarco.mar (500 cycles) served as the positive controls. D2.mar served as a 'negative control', as the Sanger sequencing result of the PCR product (with Illumina adaptors) was ineligible.

85.8% of the 28S rRNA gene NGS reads achieved a Phred quality score of Q30 and above. The number of raw reads pooled from all the 25 libraries was 447,328, which produced a total of 283,933 reads after applying the filtration criteria of maxee 1.0. This subsequently yielded 39,854 unique reads; among them 30,903 reads (77.5%) were singletons. Sixty-two Zotus were produced from the unique reads (Supplementary Material S2, available with the online version of this paper), of which 15 chimeras were removed. A total of 367,095 raw reads were successfully mapped to the Zotus. At least 84.2% of raw reads per sample could be mapped to the Zotus. In comparison, from the 39,854 unique reads, 33 Otus were produced, whereby 65 chimeras were removed (results not shown). The rarefaction curves of the 28S rRNA gene NGS were presented as $\log_{10}$ trend lines of the richness index, as shown in Figure 1. From the figure, it seems that the 1.4 nM of NGS libraries concentration could still be increased under this experimental workflow, but is nonetheless sufficient.

The use of Zotu (UNOISE algorithm) was considered more appropriate in this study as compared to Otu (UPARSE algorithm) because the heterogeneity of sequences in the amplified 28S region is unknown. Hence, it is better to apply the 'all correct biological sequences are assigned' principle of the UNOISE algorithm, rather than the 'no two sequences are >97% identical' principle of the UPARSE algorithm (Edgar & Flyvbjerg 2015; Edgar 2016). The 97% similarity threshold could be insufficient to differentiate *Sarcocystis* species (Edgar 2018). On the other hand, a sequence that is produced by both the UNOISE and UPARSE algorithm may be more credible as a correct biological sequence.

Alpha and beta diversity of the samples based on the 28S R7–R8 region

The highest observed diversity among the libraries, as portrayed by the 'Richness' index, was reported by the sample

Table 1 | NGS results and diversity analysis for 25 sample replicates, 28S R7–R8 region

Sample replicate	Site	Sample type	No. of raw reads	No. of filtered reads	No. of mapped reads	% of mapped reads	Richness	Chao 1	Simpson	Berger-Parker
J1-2015_S7	River 3, L1	Water	17,617	12,170	15,446	87.7	30	33.1	0.185	0.35
J2-2015_S15[a]	River 3, L1	Water	14,777	10,177	13,178	89.2	27	31.5	0.139	0.25
S1-2015_S8	Estuary, L3	Water	23,011	16,507	21,337	92.7	16	18.0	0.426	0.49
S2-2015_S17	Estuary, L3	Water	14,006	9,430	12,449	88.9	13	15.0	0.335	0.39
WF-1A_S13		Domestic filter	14,792	10,043	13,080	88.4	13	17.5	0.613	0.74
WF-1B_S5		Domestic filter	15,764	10,574	13,824	87.7	22	32.1	0.489	0.54
WF-2A _S14		Domestic filter	12,316	7,891	10,803	87.7	15	17.0	0.994	1.00
WF-Sup _S21		Domestic filter	13,463	8,954	11,870	88.2	12	12.1	0.495	0.51
1A-2014_S2	River 1, L2	Water	17,224	9,523	15,023	87.2	16	18.0	0.342	0.41
2A-2014_S3	River 1, L2	Water	15,621	11,389	14,206	90.9	11	11.9	0.996	1.00
G1-2015_S6	River 1, L1	Water	19,041	11,190	16,034	84.2	21	27.1	0.338	0.51
3B-2014_S11	River 1, L1	Water	13,932	10,034	12,809	91.9	16	22.1	0.983	0.99
8B-2014_S12	River 3, L1	Water	15,492	11,143	14,251	92.0	12	12.5	0.382	0.50
R2-2015_S16[a]	Estuary, L1	Water	17,252	12,256	15,722	91.1	19	37.0	0.31	0.33
T2-2015_S18	River 1, L3	Water	17,229	10,569	14,796	85.9	23	26.6	0.278	0.36
V1-2015_S20	Mangrove, L3	Water	14,116	9,402	12,484	88.4	12	12.0	0.935	0.97
B2-2015_S22	River 1, L2	Water	20,645	14,073	18,843	91.3	13	25.5	0.514	0.66
P2-2014_S1	Bay, L4	Soil	13,091	8,554	11,416	87.2	20	38.0	0.211	0.34
2A-2015_S9	Bridge, L2	Soil	14,593	10,177	12,943	88.7	12	12.5	0.436	0.58
C3-2014_S19	River 4, L1	Soil	18,737	13,147	17,076	91.1	15	47.0	0.997	1.00
Sarco_S10[a]		Positive control	13,841	9,299	12,175	88.0	15	27.5	0.471	0.53
Sarco.mar[b]		Positive control	25,552	21,785	24,830	97.2	5	5.0	0.479	0.54
D7.mar[b]	River 4, L1	Water	20,515	16,378	19,537	95.2	7	7.0	0.372	0.48
Dmax.mar[b]		Domestic filter	23,794	19,212	22,914	96.3	6	6.0	0.832	0.91
D2.mar[b]	River 1, L2	'Negative' control	40,907	56	49	0.1	5	7	0.741	0.86

[a]PCR products of these samples were also subjected to cloning.
[b]Paired-end sequencing at 500 cycles.

replicate pair of J1-2015 and J2-2015, and also T2-2015, which were water samples acquired from two different locations, L1 and L3 (Table 1). However, the 'Chao 1' diversity index that also captures low-frequency reads in a library, estimated that P2-2014, C3-2014 and R2-2015, the first two being soil samples, would have the highest diversity. Based on the 'Richness' and 'Chao1' indices, water samples 2A-2014, 2A-2015, 8B-2014, V1-2015 and D7.mar were constantly reported as having low diversity. This implies the presence of dominant species in the samples, which is also reflected by their Simpson values of about 0.4 or higher.

Based on the Bray-Curtis beta analysis of the libraries as being depicted in Figure 2, water sample 8B-2014 that was acquired from location L1, which had common sightings of water monitor lizards, was closely related to the muscle sample of water monitor lizard (Sarco_S10 and Sarco.mar, positive controls). Two water samples, V1-2015 and D7.mar, were clustered under the same clade with four replicates of the water filter, WF-1A, WF-1B, WF-2A, and Dmax.mar. Interestingly, V1-2015, which was acquired from a mangrove area in L3, had the most similar content to WF-2A, surpassing the other three

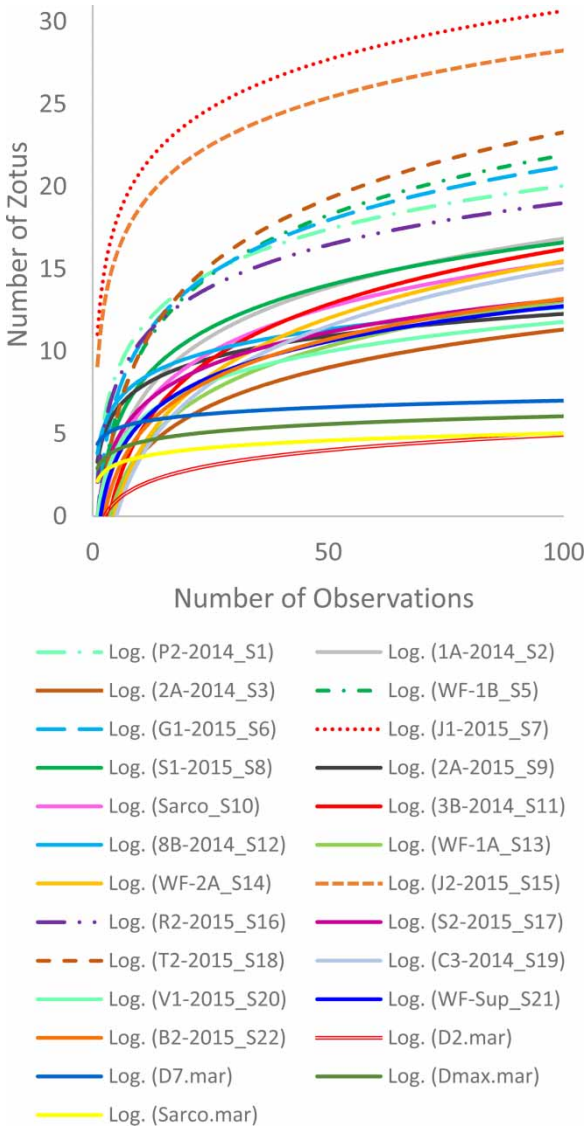

Figure 1 | Rarefaction curves of the 25 sample replicates, 28S R7–R8 region. The 1.4 nM of NGS libraries working concentration was sufficient but could still be increased accordingly.

replicates of the water filter, which was taken from location L1 in 2014. Furthermore, the supernatant content of the water filter, WF-Sup, were related to 2A-2015, a soil sample taken from L2 in 2015. The Jaccard beta analysis of the libraries presented similar relationships as the Bray-Curtis analysis of Figure 2 (results not shown). These overall findings suggest that there is no specific clustering of *Sarcocystidae* based on different locations of the island, over the two subsequent years of 2014 and 2015. Nevertheless, sampling from the

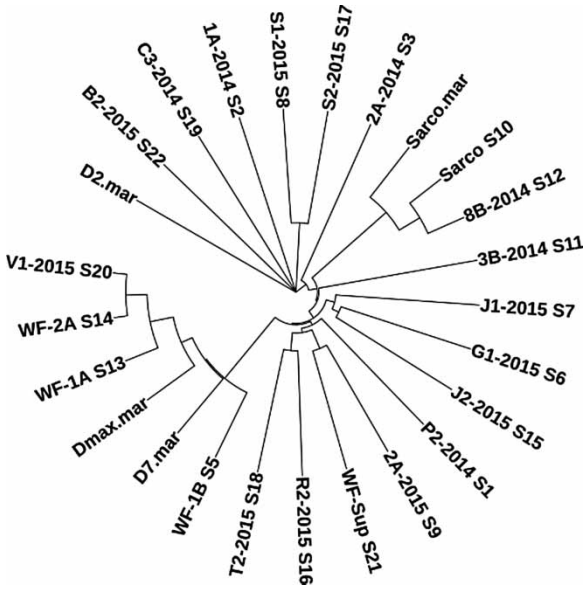

Figure 2 | Bray-Curtis beta analysis of 25 sample replicates, 28S R7–R8 region. There was no specific clustering of *Sarcocystidae* based on location or the year of sampling.

mangrove area and water filter appears to be useful, besides sampling different sections of rivers.

Identity of Zotus based on the 28S R7–R8 region

The potential identity of the 62 Zotus produced from the 25 libraries and their frequency of presence are shown in Supplementary Material S3 (available online). Forty-six of the Zotus made up five main 'clusters' of Zotus, showing various percentages of similarity to *Sarcocystis* sp. No. 5, *S. singaporensis*, *S. speeri*, *S. zuoi*, and *S. zamani*. Among them, Zotu 28 and Zotu 41 matched perfectly to *S. zamani* (GenBank KU244528.1) and *S. singaporensis* (GenBank KU341123.1), respectively. The identity of the remaining Zotus could be *S. cymruensis*, *S. lari*, *S. halieti*, *Eimeria gruis* and *Goussia janae*. The Zotus related to *S. singaporensis*, *S. zamani* and *S. cymruensis* showed at least 90% of similarities, with 100% query coverage, except for Zotu 47, 59, and 7 that had lower query coverage. Other Zotus have a poor similarity of below 90% to their related species. If the six Zotus that returned no matches in GenBank Megablast search and Zotu.20 were excluded, it is possible that this study has identified 53 unique *Sarcocystidae* species (Zotu.28 and Zotu.41 not included) through the NGS approach. Among these unique Zotu sequences, 36 had a similarity of less than the

97% threshold to their related species, and thus likely are different *Sarcocystidae* species.

There were a total of 67 *Sarcocystis* species with 28S rRNA gene information in the Genbank database as of July 2018. All these were initially used to produce a phylogeny tree together with the 62 Zotus. Species that did not show additional phylogenetical information were removed, leaving 17 representative species. The phylogenetic relationships of these 17 *Sarcocystidae* species, 62 Zotus and eight in-house unknown *Sarcocystidae* sequences (R7–R8 region of the 28S rRNA gene, Supplementary Material S2) are delineated in Figure 3. Three Zotus turned out to be identical: Zotu.27 to J1.11, Zotu.12 to Combi, and Zotu.58 to 50.20Jun.

The 'cluster' of Zotus related to *Sarcocystis singaporensis* (Supplementary Material S3) was phylogenetically divided into one monophyletic group consisting of Zotu 36, 56, 38, 60, 25, 10, 11, 4, 23, 8, 49, and 27, which also included J1.11 and *S. singaporensis* KU341122.1; and polyphyletic groups of 16 other Zotus (Zotu 17, 37, 41, 43, 44, 13, 59, 33, 40, 35, 51, 21; 47, 24, 34, 62) and *S. singaporensis* KU341123.1 (Figure 3). The 'cluster' of Zotus related to *Sarcocystis* sp. No. 5 (Supplementary Material S3) was grouped into a single monophyletic group consisting of Zotu 26, 55, 42, 50, 3, 45, 32, 1, 14, together with WF2B. Zotu 5, 58, 39 and 12 are in the same monophyletic group with 50.20Jun and Combi that also includes *Sarcocystis* species isolated from mammals and ducks. This group shares the same clade with another monophyletic group that comprises Zotu 31 and 61, which were related to *Sarcocystis* species of viper and cattle origins. Zotu 22 that has a poor query coverage with Megablast search appears together with three other Zotus that have no result in Megablast search, the Zotu 15, 48 and 54, in the monophyletic group that also comprises of Zotu 46, 53 and 19, C1.9, C2.1, C2.5, and the paraphyletic group of Zotu 6 and 18. These 12 taxons may be related to non-*Sarcocystis Sarcocystidae* species.

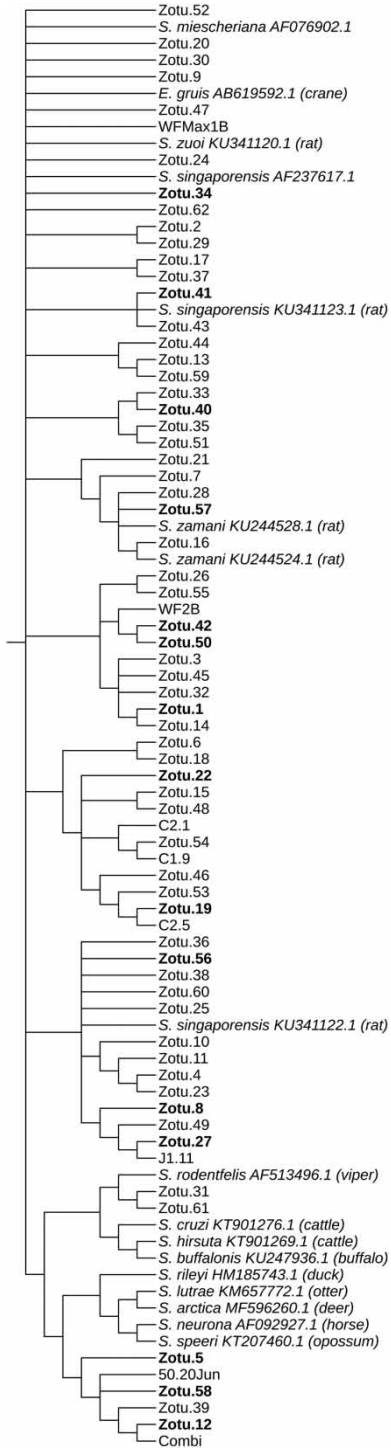

Figure 3 | Phylogenetic tree of 62 Zotus (28S R7-R8 region), eight in-house *Sarcocystidae* species, and 17 *Sarcocystis* species. Sequence of bold taxons were verified by primary PCR product and/or cloning.

Verification of Zotu sequences (28S R7–R8 region) based on Sanger sequencing of primary PCR and/or cloned products, or paired-end sequencing data subset

Eight Zotus have sequences identical to primary PCR products; four are identical to cloned sequences; three appeared both as primary PCR product and clones. These 15 Zotus with their presence verified by additional Sanger

sequencings are in bold in Figure 3. The four sample replicates (D7.mar, Dmax.mar, Sarco.mar and D2.mar) subjected to paired-end (500 cycles) NGS were re-analysed as a data subset. This produced seven Zotus with 480 bp. Five of them matched to Zotu 12, 31, 42, 58 and 61 at 290 bp length. They were denoted as Zotu.*n*.480 in parenthesis next to the respective Zotu headings in Supplementary Material S2. The remaining two Zotus with 480 bp, Zotu5.480 and Zotu7.480, were reported as the last entries in the same reference.

Naming of the important Zotu sequences

Based on the frequency of presence, abundance, and identity uniqueness (Nguyen *et al.* 2015), nine Zotus are named, as shown in Table 2. Each of these Zotus has no higher than 98% similarity to a known species (Supplementary Material S3). The multiple sequence alignment of the named Zotus is reported in Supplementary Material S4 (available online). *Sarcocystis varanus* (Zotu 12) and *Sarcocystis biawaki* (Zotu 58) were a 'doppelganger' pair. Both were found as dominant sequences of the positive control (sample replicate S10 and Sarco.mar), which also appeared in samples from two different rivers of location L1, and in one of the water filter replicate (D.max.mar) (refer Supplementary Material S5, available online). The 'doppelganger' pair has only one nucleotide of difference, but their simultaneous presences in other samples and replicates suggest that their identities are best kept distinguished. The Otu algorithm tolerates a higher level of nucleotide difference, thus did not distinguish the pair (Table 2). *Sarcocystis J1.11* (Zotu 27) and *Sarcocystis rompini* (Zotu 8) were clustered under the same monophyletic group (Figure 3), with a difference of three nucleotides. Seven of the named Zotus were detected in primary PCR and/or cloning, and hence supplemented with full-length sequence of 480 bp (Table 2, Supplementary Material S2). The remaining two, *Sarcocystis malayani* (Zotu 13) and *Sarcocystis pahangi* (Zotu 14), were at least produced by both the Zotu and Otu algorithm, and have only 290 bp of sequence length.

NGS results for the 18S rRNA gene V9 hypervariable region

Table 3 presents the NGS sequencing results of the 18S rRNA gene V9 hypervariable region of 10 environmental samples. 71.6% of the reads achieved a Phred quality score of Q30 and above. The total number of merged reads from the 10 samples was 2,003,590. This produced 132,734 unique reads; among them 81,160 reads (61.1%) were singletons. 6,995 Zotus were yielded, with 15 chimeras removed. Values of the Richness metrics and the estimated Chao1 are very close, indicating saturation of detection, as also displayed by the rarefaction curves in Figure 4.

Due to the abundance of unique reads produced by the 18S V9 NGS, a nucleotide strand of 51 bp (CCCTG

Table 2 | New *Sarcocystis* species named in this study

Microorganism	Zotu number	Otu number	Appearance frequency[a]	Dominant PCR product	Cloning	Main ecology	480 bp length availability
Sarcocystis mantioni	Zotu.42	Otu.1	19	✓		Mangrove, estuary, river, WF, S	✓
Sarcocystis varanus	Zotu.12	Otu.2	15	✓	✓	Rivers, WF, S	✓
Sarcocystis biawaki	Zotu.58		12	✓	✓	Rivers, WF, S	✓
Sarcocystis hakimi	Zotu.5	Otu.3	8	✓		Rivers	✓
Sarcocystis amiri	Zotu.56	Otu.5	9	✓		Rivers	✓
Sarcocystis malayani	Zotu.13	Otu.8	16			Bay, rivers, estuary, WF	
Sarcocystis J1.11	Zotu.27		15		✓	Bay, rivers, estuary, mangrove, WF	✓
Sarcocystis pahangi	Zotu.14	Otu.7	10			Estuaries	
Sarcocystis rompini	Zotu.8		13	✓		Bay, rivers, estuary	✓

WF, domestic water filter; S, positive control.

[a]Appearance frequency is the number of times *Sarcocystidae* was detected among the 25 sample replicates subjected to 28S rRNA gene NGS.

Table 3 | NGS results and diversity analysis for 10 sample replicates, 18S rRNA gene V9 region

Sample replicate	Sample type	No. of merged reads	No. of filtered reads	No. of mapped reads	% of mapped reads	Richness	Chao1	Simpson	Berger-Parker
2A-2014[a]	Water	310,992	215,003	212,867	68.4	1,337	1,338.7	0.0894	0.20
8B-2014[a]	Water	265,934	244,516	242,727	91.3	1,137	1,139.6	0.0645	0.19
J1-2015[a]	Water	242,267	238,932	238,086	98.3	433	435.5	0.1390	0.27
G1-2015[a]	Water	262,157	246,849	246,041	93.9	693	698.2	0.0766	0.19
3A-2014[b]	Water	286,001	127,601	126,534	44.2	1,140	1,143.3	0.0430	0.16
H1-2015	Water	276,100	214,323	213,454	77.3	1,313	1,316.5	0.1010	0.21
F2-2015	Water	283,710	197,727	189,347	66.7	2,285	2,290.0	0.0094	0.04
WFMax1B[b]	Filter	309,060	205,167	204,230	66.1	1,359	1,361.7	0.0181	0.10
P3A-2014[b]	Soil	289,990	186,028	179,908	62.0	2,302	2,303.3	0.0073	0.05
K-2014	Soil	275,185	127,444	122,953	44.7	1,542	1,545.0	0.0345	0.13

[a]The same technical replicate of the sample was also subjected to NGS of the 28S rRNA gene (refer Table 1).
[b]A different technical replicate of the same sample was also subjected to NGS of the 28S rRNA gene (refer Table 1). K-2014: River 1, L1; H1-2015: River 2, L1; F2-2015: Seawater, L2.

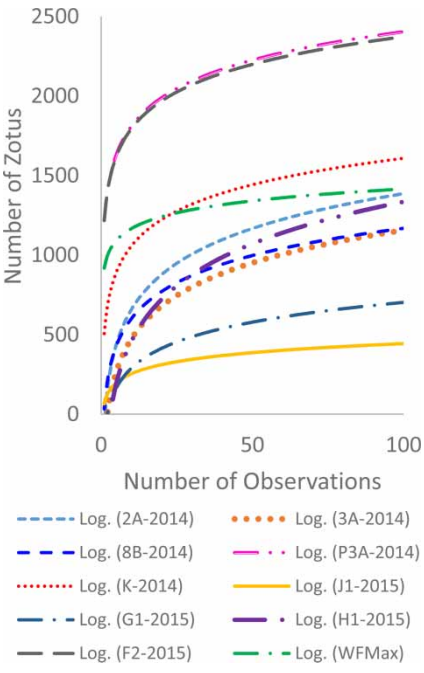

Figure 4 | Rarefaction curves of the 10 sample replicates, 18S V9 hypervariable region. Saturation of detection was achieved overall.

CCCTTTGTACACACCGCCCGTCGCTCCTACCGATTG AGTGTTCCGG) was used to narrow down the search for reads related to *Sarcocystis*. The length of this strand was selected based on the eight in-house unknown *Sarcocystidae* sequences. Among the 132,734 unique sequences, 15 sequences matched with this 51 bp 'screening' nucleotide strand, of which only two are *Sarcocystis* species (Supplementary Material S6, available online). They are Uniq.V9.21756, which is 99% similar to *S. nesbitti* (HF544323.1) with 98% query cover, and Uniq.V9.39999, that is 91% similar to *Sarcocystis fulicae* with 100% query cover. Uniq.V9.21756 reported 105 reads in WFMax1B and 3 reads in P3A-2014; Uniq.V9.39999 had 34 reads in WFMax1B. These 15 sequences in the V9 region (160 bp), and the sequences of eight in-house *Sarcocystidae* species in the similar region are reported in Supplementary Material S7 (available online).

DISCUSSION

Comparison of NGS results between 28S R7–R8 and 18S V9 region

The difference is clear between the numbers of *Sarcocystis* species detected based on sequences of the 28S rRNA gene as compared to the sequences of V9 region of the 18S rRNA gene. That is, 62 Zotus results in potentially 50 *Sarcocystis* species (not including non-*Sarcocystis* Sarcocystidae and Zotus without search result by Megablast, Supplementary Material S3), versus 6,995 Zotus leading to the identification of only two *Sarcocystis* species. This shows surpassing sensitivity and specificity of the 28S R7F–28S R8 Deg R primer pair in targeted *Sarcocystis* amplification compared to the 18S V9 region. The primer pair of 528F and 706R that

amplifies the 18S rRNA gene V4 hypervariable region was also assessed *in silico*, and was found to have a mismatch to *S. singaporensis*, an important 'cluster' of *Sarcocystis* in this study, hence disregarded. Amplifying hypervariable regions in the 18S gene for *Sarcocystis* detection was not a priority over the concerns of specificity. During the early phase of method exploration with the acquired samples, amplification of the 18S gene using published primers (Fischer & Odening 1998; Yang *et al.* 2001) yielded sequence fragments that were similar among multiple *Sarcocystis* species, according to a public database (data not shown).

Finding the most prevalent *Sarcocystis* species on the Tioman Island

The 98% coverage and 99% similarity of Uniq.V9.21756 to *S. nesbitti* HF544323.1 happened because the reference microorganism fell short of four nucleotides compared to the query sequence and has one nucleotide difference over the 156 bp of comparison. Importantly, Uniq.V9.21756 was identical to the parallel V9 region of WF2B, one of the in-house unknown *Sarcocystidae* species with nucleotide sequence ranging from the 18S to the 28S rRNA gene. This means Uniq.V9.21756, very likely, is WF2B. The corresponding region of WF2B in the 28S rRNA gene (R7–R8 region) was closely related to *Sarcocystis mantioni* (Zotu 42) and Zotu 50 (Figure 3, Supplementary Material S2). *S. mantioni* is the most prevalent sequence found in this study. In other words, notwithstanding the single nucleotide differences between WF2B and *S. mantioni* in the 28S R7–R8 region, and between Uniq.V9.21756 and *S. nesbitti* (HF544323.1) in the 18S V9 region, the most prevalent *Sarcocystis* species from the Tioman Island, is one that is highly similar to *Sarcocystis nesbitti*, as determined by NGS of two different identification genes (18S and 28S). This is the best possible postulation, as HF544323.1 is the longest publicly available sequence for *S. nesbitti* for the 18S gene. Sequences of other genes for this species have not been reported.

Sarcocystis in seawater?

The cloning of sample I-2015 (using technical replicate I2-2015), which was a seawater sample from L1, produced three different sequences. One of it, D5.3, was identical to the sequencing results of primary PCR product used for cloning (results not shown), meaning that it is most likely the dominant species in the sample. The sequence of D5.3, which appeared twice among the clones, is reported as the last entry in Supplementary Material S2. D5.3 is 91% similar to *Sarcocysits zamani* KU244528.1, at 49% of query coverage. However, subsequent attempts to amplify sample I-2015 for NGS, and medium-range PCR targeting a longer PCR product with the same sample, were all not successful. This regrettably rendered the interesting quest for how much *Sarcocystis* species could survive in seawater, largely unanswered. Nevertheless, the initially successful PCR and the subsequent cloning results provided sufficient credibility to the reported sequence of C5.3 (467 bp) for future reference. This also forms the basis of the likelihood that *Sarcocystis* may survive in seawater, hence the possibility of contracting sarcocystosis through seawater recreational activities.

Technical replicate variability

The use of NGS for quantitation has been cautioned due to replicate variability, whereby average Otu overlaps of only 17.2% and 8.2% were reported, respectively, among groups of two and three technical replicates (Zhou *et al.* 2011). Similarly, Zotus variability (28S R7–R8 region) is evident among the five technical replicates of the water filter, namely WF-1B_S5, WF-1A_S13, WF-2A_S14, WF-Sup_S21, and D.max.mar (Table 1, Supplementary Material S5). If replicate WF-Sup_S21 (supernatant) that was processed differently was left out, then the remaining four replicates shared 15.9% of Zotus. In comparison, the duplicate pair of J1-2015 and J2-2015 shared 71.1% of Zotus, and the pair of S1-2015 and S2-2015 shared 61.3% of Zotus. Referring back to the replicates of the water filter, the Chao1 (a measure of alpha diversity) values of WF-1B, WF-1A, WF-2A, and D.max.mar were 32.1, 17.5, 17.0 and 6.0, respectively (Table 1). The Chao1 values of J1-2015, J2-2015, S1-2015 and S2-2015 were 33.1, 31.5, 18.0 and 15.0, respectively. These data seem to imply that in the context of technical replicates, a higher percentage of Zotu overlaps might be expected if the alpha diversity metrics (measures of species richness in a sampling site) of the replicates are close in values. Following this simple presumptive 'rule', sample WF-1A and WF-2A that had close Chao1 values

were found to be sharing 63.9% of their Zotus, as compared to the overall 15.9% among the four replicates. To circumvent the issue of technical replicate variability, methods such as removing singletons followed by sequence abundance-weighted OTU overlaps analysis were proposed (Wen *et al.* 2017). The 28S rRNA gene NGS results of this study suggest that before the inevitable sequence abundance-based 'correction' methods, the alpha diversity metrics among the technical replicates, such as Chao1 values, should be noted to give a preliminary idea as to whether a closely similar Zotu or Otu content can be expected. It is likely that technical replicates with very different alpha diversity values would have less shared Zotu/Otu/unique sequences, and hence, require deeper sequencing, if possible.

PCR of environmental samples

Hino *et al.* (2016) have previously shown that the V9 region of the 18S gene was able to detect various parasites from rat faeces. The microbial diversity of faeces is less compared to environmental samples. If pathogens are detected, their concentrations would most likely be significant, since they are already manifested clinically. PCR amplification workflow tested with clinical samples such as faeces, muscles or blood, is often assumed to be also useful for other applications. However, with environmental samples, this study, for example, has shown that the opposite could be true of the 18S V9 region. Similarly, PCR amplification of microalgae (dinoflagellates and diatoms) was reportedly prone to 'contamination' by fungus (Guo *et al.* 2016). The microbial diversity of environmental samples should, therefore, not be underestimated. In other words, PCR workflow optimised with clinical samples should not be assumed to be also useful for environmental samples – it needs to be tested with environmental samples itself.

CONCLUSIONS

This study presents a new region in the 28S rRNA gene for *Sarcocystidae* identification, which is better than the V9 region of the 18S rRNA gene. The Zotu sequences of the respective samples are not exhaustively discussed herein. Other than the nine named Zotus, some of the other interesting Zotus are Zotu 15, 3, 26, 40, 2, 29, 61, 50, and 59. Their actual identities, as well as the identities of all other Zotus reported in this study, await to be discovered, either as species that have been known using other identification genes like the 18S or ITS (internal transcribed spacer) as such, or as new species that would then be recognised as first detected on the Tioman Island of Malaysia. It is demonstrated here that NGS is a very useful tool in assessing outbreaks whereby a clearer picture of the microbiological entities contributing to the issue is needed, in order to devise effective outbreak control and prevention steps. Information from this study fills the critical environmental perspective gap of *Sarcocystis* research against the backdrop of various *Sarcocystis* species reported from animal hosts.

ACKNOWLEDGEMENTS

The author would like to thank the Director General of Health Malaysia for permission to publish this article. Dr Lu Ping Tan and Dr Zuraifah Asrah Mohamad from the Institute for Medical Research Malaysia are greatly appreciated for their review of the manuscript. Acknowledgement goes to personnel from The Tioman Island *Sarcocystosis* Investigation Team, the Environmental Health Research Centre and Molecular Pathology Unit of the Institute for Medical Research Malaysia, the Disease Control Division of the Ministry of Health Malaysia, and the National Zoo of Malaysia for their field, sample and laboratory assistance. This work was funded by the Ministry of Health Malaysia, with project code NIH/IMR/15-011 and registration number NMRR-15-2005-27199.

REFERENCES

Bannert, B. 1994 Investigations on the host specificity of dihomoxenous sarcosporidia in the intermediate and definitive host. *Journal of Eukaryotic Microbiology* **41** (3), 183–188.

CDC 2012 Notes from the field: acute muscular sarcocystosis among returning travelers – Tioman Island, Malaysia, 2011. *Morbidity and Mortality Weekly Report* **61** (2), 37–38.

Charles, K. J., Shore, J., Sellwood, J., Laverick, M., Hart, A. & Pedley, S. 2009 Assessment of the stability of human viruses

and coliphage in groundwater by PCR and infectivity methods. *Journal of Applied Microbiology* **106** (6), 1827–1837.

Dubey, J. P. 2015 Foodborne and waterborne zoonotic sarcocystosis. *Food and Waterborne Parasitology* **1** (1), 2–11.

Dubey, J. P., Howe, D. K., Furr, M., Saville, W. J., Marsh, A. E., Reed, S. M. & Grigg, M. E. 2015 An update on *Sarcocystis neurona* infections in animals and equine protozoal myeloencephalitis (EPM). *Veterinary Parasitology* **209** (1–2), 1–42.

Edgar, R. C. 2016 UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* doi:10.1101/081257.

Edgar, R. C. 2018 Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34** (14), 2371–2375.

Edgar, R. C. & Flyvbjerg, H. 2015 Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31** (21), 3476–3482.

Esposito, D. H., Freedman, D. O., Neumayr, A. & Parola, P. 2012 Ongoing outbreak of an acute muscular *Sarcocystis*-like illness among travellers returning from Tioman Island, Malaysia, 2011–2012. *Euro Surveillance* **17** (45), 672–674.

Esposito, D. H., Stich, A., Epelboin, L., Malvy, D., Han, P. V., Bottieau, E., da Silva, A., Zanger, P., Slesak, G., van Genderen, P. J., Rosenthal, B. M., Cramer, J. P., Visser, L. G., Munoz, J., Drew, C. P., Goldsmith, C. S., Steiner, F., Wagner, N., Grobusch, M. P., Plier, D. A., Tappe, D., Sotir, M. J., Brown, C., Brunette, G. W., Fayer, R., von Sonnenburg, F., Neumayr, A. & Kozarsky, P. E. 2014 Acute muscular sarcocystosis: an international investigation among ill travelers returning from Tioman Island, Malaysia, 2011–2012. *Clinical Infectious Diseases* **59** (10), 1401–1410.

Fayer, R. 2004 *Sarcocystis* spp. in human infections. *Clinical Microbiology Reviews* **17** (4), 894–902.

Fischer, S. & Odening, K. 1998 Characterization of bovine *Sarcocystis* species by analysis of their 18S ribosomal DNA sequences. *The Journal of Parasitology* **84** (1), 50–54.

Gjerde, B., Vikøren, T. & Hamnes, I. S. 2018 Molecular identification of *Sarcocystis halieti* n. sp., *Sarcocystis lari* and *Sarcocystis truncata* in the intestine of a white-tailed sea eagle (*Haliaeetus albicilla*) in Norway. *International Journal for Parasitology: Parasites and Wildlife* **7** (1), 1–11.

Guo, L., Sui, Z. & Liu, Y. 2016 Quantitative analysis of dinoflagellates and diatoms community via Miseq sequencing of actin gene and v9 region of 18S rDNA. *Scientific Reports* **6**, 34709.

Hino, A., Maruyama, H. & Kikuchi, T. 2016 A novel method to assess the biodiversity of parasites using 18S rDNA Illumina sequencing; parasitome analysis method. *Parasitology International* **65** (5 Pt B), 572–575.

Husna Maizura, A., Khebir, V., Chong, C., Azman Shah, A., Azri, A. & Lokman Hakim, S. 2012 Surveillance for sarcocystosis in Tioman Island, Malaysia. *Malaysian Journal of Public Health Medicine* **12** (2), 39–44.

Italiano, C. M., Wong, K. T., AbuBakar, S., Lau, Y. L., Ramli, N., Syed Omar, S. F., Kahar Bador, M. & Tan, C. T. 2014 *Sarcocystis nesbitti* causes acute, relapsing febrile myositis with a high attack rate: description of a large outbreak of muscular sarcocystosis in Pangkor Island, Malaysia, 2012. *PLoS Neglected Tropical Diseases* **8** (5), e2876.

Koudela, B. & Modrý, D. 2000 *Sarcocystis muris* possesses both diheteroxenous and dihomoxenous characters of life cycle. *Journal of Parasitology* **86** (4), 877–879.

Kumar, S., Stecher, G. & Tamura, K. 2016 MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33** (7), 1870–1874.

Latif, B. M., Al-Delemi, J. K., Mohammed, B. S., Al-Bayati, S. M. & Al-Amiry, A. M. 1999 Prevalence of *Sarcocystis* spp. in meat-producing animals in Iraq. *Veterinary Parasitology* **84** (1–2), 85–90.

Lau, Y. L., Chang, P. Y., Tan, C. T., Fong, M. Y., Mahmud, R. & Wong, K. T. 2014 *Sarcocystis nesbitti* infection in human skeletal muscle: possible transmission from snakes. *The American Journal of Tropical Medicine and Hygiene* **90** (2), 361–364.

Letunic, I. & Bork, P. 2011 Interactive Tree of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research* **39** (2), W475–W478.

Nguyen, N. H., Smith, D., Peay, K. & Kennedy, P. 2015 Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist* **205** (4), 1389–1393.

Shahari, S., Tengku-Idris, T. I. N., Fong, M. Y. & Lau, Y. L. 2016 Molecular evidence of *Sarcocystis nesbitti* in water samples of Tioman Island, Malaysia. *Parasites & Vectors* **9**, 598.

Tappe, D., Ernestus, K., Rauthe, S., Schoen, C., Frosch, M., Müller, A. & Stich, A. 2013 Initial patient cluster and first positive biopsy findings in an outbreak of acute muscular *Sarcocystis*-like infection in travelers returning from Tioman Island, Peninsular Malaysia, in 2011. *Journal of Clinical Microbiology* **51** (2), 725–726.

Wen, C., Wu, L., Qin, Y., Van Nostrand, J. D., Ning, D., Sun, B., Xue, K., Liu, F., Deng, Y., Liang, Y. & Zhou, J. 2017 Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. *PLoS ONE* **12** (4), e0176716.

Yang, Z. Q., Zuo, Y. X., Ding, B., Chen, X. W., Luo, J. & Zhang, Y. P. 2001 Identification of *Sarcocystis hominis*-like (Protozoa: Sarcocystidae) cyst in water buffalo (*Bubalus bubalis*) based on 18S rRNA gene sequences. *The Journal of Parasitology* **87** (4), 934–937.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., Xie, J., Van Nostrand, J. D., He, Z. & Yang, Y. 2011 Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME Journal* **5**, 1303.

First received 31 January 2019; accepted in revised form 1 February 2019. Available online 22 March 2019