Am. J. Trop. Med. Hyg., 104(4), 2021, pp. 1388–1393
doi:10.4269/ajtmh.20-0767

From 18S to 28S rRNA Gene: An Improved Targeted Sarcocystidae PCR Amplification, Species Identification with Long DNA Sequences

Florence C. H. Lee[1]* and Vickneshwaran Muthu[2]*

[1]*Environmental Health Research Centre, Institute for Medical Research (IMR), National Institutes of Health, Ministry of Health Malaysia, Setia Alam, Malaysia;* [2]*Zoonosis Sector, Disease Control Division, Ministry of Health Malaysia, Putrajaya, Malaysia*

Abstract. Sarcocystosis outbreaks in Tioman and Pangkor islands of Malaysia between 2011 and 2014 have raised the need to improve *Sarcocystis* species detection from environmental samples. In-house works found that published primers amplifying the 18S rRNA gene of *Sarcocystis* either could not produce the target from environmental samples or produced *Sarcocystis* DNA sequence that was insufficient for species identification. Using the primer pair of 18S S5 F (published) and 28S R6 R (new), this study improved the PCR amplification of Sarcocystidae to overcome these two difficulties. The PCR product spanned from the 18S to 28S rRNA genes, providing more information for species identification. The long DNA sequence allowed comparison between the "Ident" and "Query Cover" sorting in GenBank identity matching. This revealed the ambiguity in identity matching caused by different lengths of reference DNA sequences, which is seldom discussed in the literature. Using the disparity index test, a measurement of homogeneity in nucleotide substitution pattern, it is shown that the internal transcribed spacer (ITS)1-5.8S-ITS2 and 28S genes are better than the 18S gene in indicating nucleotide variations, implying better potentials for species identification. The example given by the handful of Sarcocystidae long DNA sequences reported herein calls for the need to report DNA sequence from the 18S to the 28S rRNA genes for species identification, especially among emerging pathogens. DNA sequence reporting should include the hypervariable 5.8S and ITS2 regions where applicable, and not be limited to single gene, per the current general trend.

INTRODUCTION

Sarcocystosis has become an emerging parasitological disease affecting tropical countries, especially those in the Southeast Asia.[1,2] Being a waterborne and foodborne zoonotic disease, environmental sampling is crucial in piecing various aspects of this disease to minimize the risk of future outbreaks. However, attempt to detect *Sarcocystis* species from environmental samples has been elusive.[3] Further in-house works (not shown) to amplify *Sarcocystis* in the 18S rRNA gene of environmental samples taken from Tioman Island have been inconclusive. Published PCR methods were used,[4–6] but two critical issues were encountered. First, dominant environmental eukaryotes such as dinoflagellates, *Galactomyces*, and *Sterkiella* were being amplified instead. Second, positive results produced from PCR targeting the 18S rRNA gene could not conclude as to which *Sarcocystis* species were detected in the samples because of similar values of identity percentage to multiple *Sarcocystis* species when Nucleotide BLAST was performed in GenBank. This happened against the backdrop that *Sarcocystis singaporensis*, *Sarcocystis nesbitti*, and *Sarcocystis* sp. *YLL*-2013 were detected based on the 18S rRNA gene in water samples acquired from Tioman Island.[7] It is hence postulated that not all *Sarcocystis* species are distinctive enough to be differentiated with the 18S rRNA gene, constituting the "blind spot" of phylogeny within this genus.[8] These research bottlenecks call for method improvement in the detection of *Sarcocystis* from environmental samples with better species identification outcome. Presented here is the output of this research aim—a suite of methods suitable for Sarcocystidae detection from environmental samples with a *long-range* PCR amplification targeting a product spanning from the 18S to 28S rRNA genes (4 kb in length), followed by cloning and Re-PCR or nested PCR of the cloned PCR product, to provide sufficient DNA concentration for sequential sequencing that covers DNA sequence from the 18S to 28S rRNA genes.

The 18S, 28S, internal transcribed spacer (ITS), cytochrome c oxidase, and the actin gene have been used for species identification of eukaryotes such as toxoplasma, dinoflagellate, and copepod.[9–16] These developments are needed, but they occur when the current understanding of factors affecting the identification power for taxa is still lacking.[12] The debate as to which hypervariable region in the 18S or 16S genes, or which other genes are better for species identification, will continue for the time being.[12,17,18] Meanwhile, as a consequence of this dominant trend of reporting DNA sequence of single genes, sequence information of a species in public domain is often segmented, causing ambiguity in species identification, given the different lengths of reference DNA sequences. Under this context, we take the encompassing approach of broad sweeping some of those familiar genes, which actually appear side by side. The long-range PCR method presented here detects the partial 18S, complete ITS1-5.8S-ITS2, and partial 28S rRNA genes *in a single reaction*. A suite of primers designed in silico or published for Sarcocystidae is provided along to enable sequential sequencing (primer walking) that yields linked DNA sequences about 4 kb of nucleotides in length.

Longer DNA sequences would theoretically provide more information to improve species identification outcome. The handful of linked Sarcocystidae DNA sequences produced by this study enables a simple exploration on this particular aspect of species identification. Nucleotide BLAST was performed on the linked DNA sequences, and comparison is made between the results sorted with the "Query Cover" and "Ident" functions. In determining the species, the previous one emphasizes on sequence length, whereas the latter on

* Address correspondence to Florence C. H. Lee, Environmental Health Research Centre, Institute for Medical Research (IMR), Ministry of Health Malaysia, C6-L2-19, NIH Complex, No.1, Jalan Setia Murni U13/52, Setia Alam, 40170 Shah Alam, Malaysia, E-mail: florencelee@moh.gov.my or Vickneshwaran Muthu, Zoonosis Sector, Disease Control Division, Ministry of Health Malaysia, Level 3, Block E10, Complex E, Putrajaya 62590, Malaysia, E-mail: drmvwaran@moh.gov.my.

percentage of identity matching. On the other hand, the linked DNA sequences also enable us to revisit the subject of comparing genes for species identification, in a different light—using the disparity index test; we compared all rRNA gene segments of the five unknown Sarcocystidae sequences, dividing the linked DNA sequences in shared conserve regions into three segments, namely, the partial 18S gene, the complete ITS1-5.8S-ITS2 genes and the partial 28S gene. The disparity index test detects significant differences in substitution pattern between pairwise sequences,[19] hence reflecting the usefulness of the gene segments for species identification.

MATERIALS AND METHODS

PCR amplification of genomic DNA from water and soil samples. This study used the genomic DNA acquired from the previous study of Lee,[20] whereby grab water samples and grab soil samples were collected from Tioman, a recreational island off the east coast of Peninsular Malaysia, in October 2014 and August 2015. A total of 28 environmental samples were tested as Sarcocystidae positive by the new primer pair of 28S R7F-28S R8 Deg R that targeted the 28S rRNA gene, of which 21 were grab water samples and seven were grab soil samples. The samples were subjected to genomic DNA extraction using a PowerWater DNA Isolation Kit, DNeasy PowerSoil DNA Isolation Kit, or PowerMax Soil DNA Isolation Kit (Mo Bio, Carlsbad, California). The extraction followed the manufacturer's instruction, whereby heating step for the alternate lysis method was also performed. However, sample tubes were vortexed for 20 minutes, instead of 5 minutes. In this study, long-range PCR on the extracted DNA was performed using either the ProFlex or Veriti PCR system (Applied Biosystems, Waltham, Massachusetts) with the content of 25 μL of Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs, Ipswich, Massachusetts), 2.5 μL (10 μM) each of 18S S5 F forward (Fischer and Odening[6]) and 28S R6 R reverse (new) primer, 3 μL of genomic DNA, and 17 μL of nuclease-free water (Qiagen, Hilden, Germany). Initial denaturation of PCR was run at 98°C for 3 minutes, followed by 40 cycles of 98°C for 10 seconds, 70°C for 45 seconds, and 72°C for 4 minutes; final elongation at 72°C for 7 minutes; and hold at 4°C. Primers are listed in Supplemental Data Table S1. *Sarcocystis* species references used for primer design are compiled in Supplemental Data S2. The presence of PCR products was screened with a FlashGel DNA Cassette (1.2% agarose, Lonza, Basel, Switzerland) and the presumptively positive ones sent for Sanger sequencing using primer 18S S5F (Genomics, Taipei, Taiwan).

PCR product cloning. The 4-kb-long PCR products that showed average to good band intensity from PCR using the primer set 18S S5 F - 28S R6 R were electrophoresized, excised (Supplemental Data Figure 1), purified, and cloned. The cloning process is necessary to get single DNA templates of sufficient concentration for downstream application. This is because given the high microbial diversity of environmental samples, PCR tend to produce mixed DNA templates from different species belonging to the same family/genus, which may together appear as "a single band" in PCR product gel electrophoresis. Cloning of the excised and purified 4-kb PCR products was conducted using either the NEB PCR Cloning Kit with NEB 10-Beta chemically competent cells (New England Biolabs) following the manual instruction, or the pGC Blue Cloning & Amplification Kits (Lucigen, Middleton, Wisconsin) with the E. cloni 10G Supreme Electrocompetent Cells. For the latter, transformation was conducted with the Gene Pulser Xcell electroporation System (Bio-Rad, Hercules, California). The transformed cells were further incubated overnight at room temperature after the recommended incubation at 250 revolutions per minute for 1 hour at 37°C. All transformed cells were grown in Lennox Luria broth (Pronadisa, Madrid, Spain) until log phase. The cloned plasmids were harvested with Monarch Plasmid Miniprep (New England Biolabs) and their presence screened with FlashGel DNA Cassette (1.2% agarose, Lonza).

Re-PCR or nested PCR amplification of cloned plasmids to increase DNA concentration for sequential sequencing that produces linked DNA sequences. Some cloned plasmids showed nonspecific bands for unknown reasons, or weak band intensities (Supplemental Data Figure 2). Hence, Re-PCR using primer set 18S S5 F-28S R6 R, or nested PCR with primer set 18S 3L Deg F^4 - 28S R5 Deg R (new) was performed (Supplemental Data Figures 3 and 4) to increase the DNA concentration of the plasmids as the subsequent step of sequential sequencing requires high concentration of DNA material. The total PCR reaction content was adjusted to 30 μL, containing 5 μL of plasmid DNA. For Re-PCR, the condition was 98°C for 3 minutes, followed by 37 cycles of amplification at 98°C for 10 seconds, 70°C for 45 seconds, and 72°C for 2 minutes; final elongation at 72°C for 5 minutes; and hold at 4°C. Condition for the nested PCR was 98°C for 3 minutes, 40 cycles of amplification at 98°C for 10 seconds, 63°C for 45 seconds, and 72°C for 2.5 minutes; final elongation at 72°C for 5 minutes; and hold at 4°C. Alternatively, the nested PCR was performed with the purified primary PCR product (Monarch PCR & DNA Cleanup Kit, New England Biolabs) to yield the dominant PCR product directly without going through cloning. All Re-PCR or nested PCR products were electrophoresized and the 4-kb bands excised. The excised bands were then subjected to sequential sequencing using a list of primers in the order as presented in Supplemental Data Table S1. Two additional primers, 18S/ITS1 F (new) and 18S 1H F,[4] were used for further sequencing when the DNA had longer than usual sequence in the neighboring genes of ITS1, 5.8S, and ITS2. Nucleotides from the adjacent sequencing sections of a re-amplified cloned 4-kb PCR product were aligned using the Clustal Omega tool in EMBL-EBI (United Kingdom) and MUSCLE Clustal Alignment function in MEGA 7 software (United Kingdom), which were then linked up to produce a single concatenated sequence based on data cleanup criteria presented in Supplemental Data S3.

Sorting Nucleotide BLAST results with the functions "Query Cover" and "Ident"; and disparity index test. The linked DNA sequences of the samples were aligned using the Nucleotide BLAST tool in the GenBank database (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The results were sorted with two functions: by "Query Cover" and by "Ident," which follows the highest query coverage percentage and identity matching percentage, respectively. Identity of the top three microorganisms shown through these two different sorting methods was compared. Meanwhile, to perform the test of homogeneity of substitution patterns, that is, the disparity index test, the linked DNA sequences were trimmed to start with an identical strand of nucleotides at the earliest available position in the 18S gene, and ends similarly at the farthest available position at the 28S gene. Two nucleotide strands that mark the end of 18S gene and the beginning of 28S gene were also

selected based on the GenBank *Sarcocystis* species sequences used for primer design (Supplemental Data S2). These four strands of nucleotides (double underlined in Supplemental Data S4) were used to divide the linked DNA into partial 18S gene, complete ITS1 to ITS2 genes (including ITS1, 5.8S, and ITS2), and partial 28S gene. The trimmed and segmented sequences of the samples were first aligned in MEGA 7 software. The disparity index test was then performed using the "use all sites" option. A test value of ≤ 0.05 means that the pairwise samples have significantly different substitution patterns.

RESULTS AND DISCUSSION

Six of the 28 Sarcocystidae-positive environmental samples that produced average to good band intensity from PCR using the primer set 18S S5 F-28S R6 R were subjected to cloning or nested PCR. This yielded five linked DNA sequences: WFMax1B from a nested PCR of the purified primary PCR product; J1.11 from Lucigen cloning (Nested-PCR); and C2.5, C2.1, and C1.9 from NEB cloning (Re-PCR, refer Supplemental Data Figure 3). These sequence data have been submitted to GenBank with the numbers of MH590233, MH590230, MH590232, MH590231, and MH567098, and their multiple sequence alignments are shown in Supplemental Data S4. Nucleotides of each adjacent segment in a linked DNA sequence overlapped for at least 250 bp (Supplemental Data S5).

Discrepancies of identity when Nucleotide BLAST results were sorted differently using the "Query Cover" and "Ident" functions. Table 1 presents the potential identities of these five linked DNA sequences as sorted by the descending values of "Query Cover" and "Ident" functions, respectively, in GenBank. The top three identities of the sequences were found to "change" with the two functions. For sequence C2.5, C2.1, and C1.9, the identities even changed to a different genus. For sequence J1.11, although the top matching species remained the same, the percentage of identity matching actually "dropped" from 99% to 96% when sorting was switched from the function "Ident" to "Query Cover." These discrepancies happened because there are not many reported long sequences in the database that contain more than one gene to be used for the sequence alignments. By contrast, the "Ident" sorting produced apparently high percentage of identity matching through alignments to the 18S gene, which are yet to be "challenged" by longer sequences that contain more hypervariable regions from other marker genes. In the case of the *Sarcocystis* genus, some long sequences that span from partial 18S to 28S are available in GenBank, but they are vastly outnumbered by shorter sequences of single genes. Longer DNA sequences would theoretically provide more information for species identification and overcome phylogeny "blind spot" to enable phylogeny analysis with assurance. This can be achieved by amplifying the whole mitochondrial genome,[21] or the ribosomal RNA genes from the 18S to the 28S rRNA genes, as presented here. Nucleotide BLAST results should always be sorted using both the "Query Cover" and "Ident" functions to check for any changes in matching identity.

Disparity index test as a tool to measure identity difference between sequences, showing that the complete ITS1-5.8S-ITS2 and partial 28S rRNA genes are better than the 18S gene in species identification among the rRNA genes. Table 2 shows the disparity index test values of the five linked DNA sequences of WFMax1B, J1.11, C1.9, C2.1, and C2.5 when they were compared pairwise at their respective segments of partial 18S, complete ITS1-5.8S-ITS2, and partial 28S gene. Of the 10 compared pairs of sequences, none showed significantly different substitution pattern in the partial 18S gene ($P > 0.05$). In comparison, the complete ITS1-5.8S-ITS2 genes had seven, and the partial 28S gene had four pairs of sequences, showing significantly different substitution pattern ($P < 0.05$). The ITS1 gene was reportedly more useful than cox1, 18S, and 28S (in the D2/D3 region) in differentiating closely related *Sarcocystis* species found in birds and carnivores.[22] This is in accordance with the disparity index test results of this study whereby three pairs of sequences were found to have significantly different nucleotide substitution patterns only in the ITS1-5.8S-

TABLE 1
Potential identities of samples based on the top three rankings sorted by the "Query Cover" and "Ident" functions in GenBank

	Top three rankings of potential identities in GenBank							
	Sorted with descending order of "Query Cover"				Sorted with descending order of "Ident"			
Sample†	Query cover (%)	Identity to sample (%)	Microorganism*	Accession number	Query Cover (%)	Identity to sample (%)	Microorganism*	Accession number
WFMax1B (3,902 bp)	67	91	*S. zamani*	KU244528.1	57	93	*S. singaporensis*	KU341123.1
	67	89	*S. zuoi*	KU341120.1	32	92	*S. sp. ex Morelia viridis*	KC201640.1
	57	93	*S. singaporensis*	KU341123.1	62	90	*S. zuoi*	KU341121.1
J1.11 (3,979 bp)	88	96	*S. singaporensis*	KU341123.1	32	99	*S. singaporensis*	KY513624.1
	74	96	*S. zamani*	KU244528.1	74	96	*S. zamani*	KU244528.1
	73	95	*S. zuoi*	KU341120.1	63	96	*S. zuoi*	KU341121.1
C2.5 (4,003 bp)	84	87	*B. besnoiti*	DQ227420.1	33	99	*G. szekelyi*	GU479656.1
	66	94	*T. gondii*	X75453.1	32	98	*G. koertingi*	GU479647.1
	40	94	*H. hammondi*	AH008381.2	31	98	*G. pannonica*	GU479642.1
C2.1 (4,164 bp)	81	88	*B. besnoiti*	DQ227420.1	31	98	*G. janae*	GU479644.1
	75	94	*T. gondii*	X75453.1	30	98	*G. pannonica*	GU479642.1
	36	94	*I. belli*	DQ060661.2	30	98	*G. koertingi*	GU479647.1
C1.9 (4,053 bp)	80	88	*B. besnoiti*	DQ227420.1	28	98	*G. pannonica*	GU479642.1
	73	94	*T. gondii*	X75453.1	28	97	*G. janae*	GU479644.1
	34	94	*I. belli*	DQ060661.2	28	97	*G. koertingi*	GU479647.1

* Genus names: *B* = *Besnoitia*; *G* = *Goussia*; *H* = *Hammondia*; *I* = *Isospora*; *S* = *Sarcocystis*; *T* = *Toxoplasma*.
† Sequence of WFMax1B was produced from nested PCR of purified primary PCR; J1.11 from Lucigen cloning (nested PCR); C2.5, C2.1, and C1.9 from NEB cloning (Re-PCR).

TABLE 2

Disparity index test (test of homogeneity of nucleotide substitution patterns) between pairwise sequences for partial 18S gene, complete ITS1-5.8S-ITS2 genes, and partial 28S gene

Disparity index test value between pairwise sequences				
Pairwise sequences		Genes segments*		
Sequence 1	Sequence 2	Partial 18S (1,270 bp)	Complete ITS1-5.8S-ITS2 (1,457 bp)	Partial 28S (1,294 bp)
WFMax1B	J1.11	0.22	1.00	0.16
WFMax1B	C2.5	1.00	**0.00**	0.22
J1.11	C2.5	1.00	0.15	0.40
WFMax1B	C2.1	1.00	**0.00**	0.17
J1.11	C2.1	1.00	**0.00**	**0.05**
C2.5	C2.1	0.38	**0.01**	**0.03**
WFMax1B	C1.9	1.00	**0.00**	0.24
J1.11	C1.9	1.00	**0.00**	**0.03**
C2.5	C1.9	0.30	**0.01**	**0.02**
C2.1	C1.9	1.00	1.00	0.20

* Number of nucleotides reported follows the "use all sites" function in gaps/missing data treatment of MEGA7. Bold test value ≤ 0.05 indicates a significantly different nucleotide substitution pattern.

ITS2 genes, and not in the 18S or 28S genes (Table 2). In moving toward the third generation sequencing characterized by long read lengths,[23] the disparity index test can serve as a useful tool to deal with the legacy of short read lengths by measuring identity difference between sequences.

Hypervariable regions covered by the Sarcocystidae PCR primer sets, and sensitivity with environmental samples. The DNA sequence produced from the primer set of 18S S5 F-28S R6 R includes nucleotides from the 18S V3 hypervariable regions onward.[16] The 18S V1 and V2 hypervariable regions are not covered by this primer set. The partial 28S gene segment amplified by this primer set also contains the D2 and D3 regions of the 28S gene.[24] The A2F-KL3 primer set used by Gjerde et al.,[22] is the closest comparison to S5F-28S R6R. Given the results of this study and public database, A2F is estimated to be located in a hypervariable region further downstream from S5F in the 18S gene, whereas KL3 is situated between 28S R5 Deg R and 28S R6R. The A2F-KL3 primer set was designed to target *Sarcocystis* species harbored by birds. By contrast, both the S5F and 18S 3L Deg F forward primers (Supplemental Data Table S1 and Supplemental Data S4) are in the same conserved region. Therefore, these primer sets should be chosen depending on application purposes, that is, whether a broad detection of Sarcocystidae species is needed, or a more specific host category is targeted.

Although primer set 18S S5 F-28S R6 R was shown to have the specificity needed to detect Sarcocystidae in environmental samples against other eukaryotes, it however suffered the sensitivity issue common in long-range PCR amplification, as other workers have experienced.[21] Of the 28 environmental samples previously reported as harboring Sarcocystidae using the 28S R7F-28S R8 Deg R primer set that delivered about 480 bp of PCR products,[20] only six samples yielded PCR products of sufficient band intensity when tested again with the primer set of 18S S5 F-28S R6 R. The genomic DNA concentration of these samples ranged between 2.6 and 103.8 ng/μL, and has been kept at –20°C for 17 months before testing. The application of long-range PCR with environmental samples is rather new and is compounded with the additional challenges of inherent DNA degradation, coexistence of closely related species/genera, and low DNA concentration due to dilution effect in the environment, among others. Nonetheless, Deiner et al.[25] had successfully performed long-range PCR with filtered water samples from streams, showing that the application is still plausible with environmental samples. Their success is most likely delivered by the combination of timely PCR (minimal DNA storage time) and the use of bovine serum albumin during PCR reaction.

Other alternatively tested primers can produce longer DNA sequences for broad detection of environmental eukaryotes. More primer sets have initially been designed and tested to produce linked DNA sequences that would comprise all complete rRNA genes. However, only the S5F-28S R6 R and 18S 3L Deg F-28S R5 Deg R (Supplemental Data Table S1) succeeded in amplifying Sarcocystidae out of the environmental samples (Supplemental Data Figure 1), with the compromise of the S5F-28S R6 R pair losing about 380 bp in the beginning of the 18S rRNA gene and about 1,350 bp toward the end of the 28S rRNA gene. Two alternative sets of primers that were tested concurrently with the environmental sample were found to be interesting, namely, the ERIB1 (5′- ACCTGGTTGATCCTGCCAG - 3′)[5] and 28S R1R (new, 5′ – TAGGGACAGTGGGAATCTCG - 3′); and 1LF (5′- CCATGCATGTCTAAGTATAAGC - 3′)[4] and 28S R5 Deg R. Instead of detecting Sarcocystidae in the environmental samples, they detected other eukaryotes from the genus of *Cyclotella*, *Yarrowia*, *Galactomyces*, *Clavispora*, *Picochlorum*, *Nannochloris*, *Tetraselmis*, *Chlorococcum*, *Euplotes*, *Sterkiella*, *Spumella*, and *Rhodotorula* (results not shown). These "unintentionally" detected eukaryotes suggest the potential use of these two primer sets (ERIB1-28S R1R and 1LF-28S R5 Deg R) in ecological or other studies concerning those eukaryotes. The ERIB1-28S R1R pair notably would amplify DNA sequence spanning from the beginning of the 18S gene until about 2,500 bp to the 28S rRNA gene, which is generally 3,200 bp long. The 1LF-28S R5 Deg R may then be used as nested PCR primers, if needed. Although not suitable for targeted Sarcocystidae amplification with environmental samples, these alternative primer sets might still detect Sarcocystidae in samples with comparably less microorganism diversity, for example, samples retrieved from diseased animals suspected of suffering sarcocystosis.

Identity of the five linked DNA sequences of this study. There is no definite answer to the identity of the five linked DNA sequences reported here, given the lack of long DNA sequence references covering most, if not all, of the rRNA genes of *Sarcocystis*—by species per se, in public database. Nonetheless, after considering the disparity index test results alongside the length differences in the ITS1-5.8S-ITS2 gene segments and the overall outcome of the Nucleotide BLAST results sorted by both the "Query Cover" and "Ident" functions, it is postulated that WFMax1B and J1.11 are currently unreported species belonging to the *Sarcocystis* genus, whereas C2.5, C2.1, and C1.9 are other members of the Sarcocystidae family. J1.11 is the linked long rRNA sequence of *Sarcocystis J1.11*, whereby the 480-bp nucleotide of the latter in the 28S rRNA gene has been reported previously.[20] Sequence WFMax1B is hereby named as *Sarcocystis thatcheiani*.

Future outlook. Besides improving Sarcocystidae detection in response to sarcocystosis outbreaks, this study also addresses the overlooked issue of species identification using single genes, and the application of long-range PCR with environmental samples. The disparity index test results of long eukaryote DNA sequences amplified from environmental samples clearly demonstrate the need to shift the trend of species identification from single genes to long DNA sequence covering the adjacent rRNA genes of 18S, ITS1, 5.8S, ITS2, and 28S among eukaryotes. With the current limitation of short read lengths in next-generation sequencing, at least, the ITS and 28S genes are better choices for species identification than the 18S gene in the context of ribosomal RNA genes. This recommendation is coined on the ground of rRNA genes across eukaryotes, notwithstanding the use of other genes outside the rRNA genes. The "establishment" of the 18S (or 16S for prokaryotes) in species identification is founded by the historical trending of its usage, when the usefulness of other rRNA genes has not really been tested and compared. Long DNA sequence is needed to reconcile the currently fragmented sequence information of the ribosomal RNA genes. This research direction is especially relevant among emerging diseases so that the DNA database of these microorganisms can start and grow right, avoiding the pitfalls of fragmented DNA sequence information common among other microorganisms.

Received June 30, 2020. Accepted for publication October 26, 2020.

Published online February 22, 2021.

Note: Supplemental data, table, and figures appear at www.ajtmh.org.

Acknowledgments: We would like to thank the director general of Health Malaysia for permission to publish this article. Gratitude is extended to personnel from the Tioman *Sarcocystis* Investigation Team, the Environmental Health Research Centre and Molecular Pathology Unit of the Institute for Medical Research, the Zoonosis Sector, Disease Control Division of the Ministry of Health Malaysia, and the National Zoo of Malaysia for their field, sample, and laboratory assistance.

Financial support: This work was funded by the Ministry of Health Malaysia (project code NIH/IMR/15-011 and registration number NMRR-15-2005-27199).

Authors' addresses: Florence C. H. Lee, Environmental Health Research Centre, Institute for Medical Research (IMR), Ministry of Health Malaysia, C6-L2-19, NIH Complex, No.1, Jalan Setia Murni U13/52, Setia Alam, 40170 Shah Alam, Malaysia, E-mail: florencelee@moh.gov.my. Vickneshwaran Muthu, Zoonosis Sector, Disease Control Division, Ministry of Health Malaysia, Putrajaya, Malaysia, E-mail: drmvwaran@moh.gov.my.

REFERENCES

1. Latif B, Muslim A, 2016. Human and animal sarcocystosis in Malaysia: a review. *Asian Pac J Trop Biomed 6:* 982–988.
2. Fayer R, Esposito DH, Dubey JP, 2015. Human infections with Sarcocystis species. *Clin Microbiol Rev 28:* 295–311.
3. Husna Maizura AM, Khebir V, Chong CK, Azman Shah AM, Azri A, Lokman Hakim S, 2012. Surveillance for sarcocystosis in Tioman Island, Malaysia. *Malaysian J Public Heal Med 12:* 39–44.
4. Yang ZQ, Zuo YX, Ding B, Chen XW, Luo J, Zhang YP, 2001. Identification of Sarcocystis hominis-like (Protozoa: Sarcocystidae) cyst in water buffalo (*Bubalus bubalis*) based on 18S rRNA gene sequences. *J Parasitol 87:* 934–937.
5. Barta JR et al., 1997. Phylogenetic relationships among eight Eimeria species infecting domestic fowl inferred using complete small subunit ribosomal DNA sequences. *J Parasitol 83:* 262–271.
6. Fischer S, Odening K, 1998. Characterization of bovine Sarcocystis species by analysis of their 18S ribosomal DNA sequences. *J Parasitol 84:* 50–54.
7. Shahari S, Tengku-Idris TIN, Fong MY, Lau YL, 2016. Molecular evidence of *Sarcocystis nesbitti* in water samples of Tioman Island, Malaysia. *Parasit Vectors*. *9:* 598.
8. Woo PCY, Lau SKP, Teng JLL, Tse H, Yuen KY, 2008. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clin Microbiol Infect 14:* 908–934.
9. Guo L, Sui Z, Zhang S, Liu Y, Du Q, 2014. Preliminary comparison of quantification efficiency between DNA-derived dataset and cell-derived dataset of mixed diatom sample based on rDNA-ITS sequence analysis. *Biochem Syst Ecol 57:* 183–190.
10. Ellis JT, Amoyal G, Ryce C, Harper PAW, Clough KA, Homan WL, Brindley PJ, 1998. Comparison of the large subunit ribosomal DNA of Neospora and Toxoplasma and development of a new genetic marker for their differentiation based on the D2 domain. *Mol Cell Probes 12:* 1–13.
11. Pereira TJ, Baldwin JG, 2016. Contrasting evolutionary patterns of 28S and ITS rRNA genes reveal high intragenomic variation in Cephalenchus (Nematoda): implications for species delimitation. *Mol Phylogenet Evol 98:* 244–260.
12. Tanabe AS, Nagai S, Hida K, Yasuike M, Fujiwara A, Nakamura Y, Takano Y, Katakura S, 2016. Comparative study of the validity of three regions of the 18S-rRNA gene for massively parallel sequencing-based monitoring of the planktonic eukaryote community. *Mol Ecol Resour 16:* 402–414.
13. Fišer Pečnikar Ž, Buzan EV, 2014. 20 years since the introduction of DNA barcoding: from theory to application. *J Appl Genet 55:* 43–52.
14. Ki JS, 2012. Hypervariable regions (V1–V9) of the dinoflagellate 18S rRNA using a large dataset for marker considerations. *J Appl Phycol 24:* 1035–1043.
15. Wu S, Xiong J, Yu Y, 2015. Taxonomic resolutions based on 18S rRNA genes: a case study of subclass Copepoda. *PLoS One 10:* e0131498.
16. Cooper MK, Phalen DN, Donahoe SL, Rose K, Šlapeta J, 2016. The utility of diversity profiling using Illumina 18S rRNA gene amplicon deep sequencing to detect and discriminate *Toxoplasma gondii* among the cyst-forming coccidia. *Vet Parasitol 216:* 38–45.
17. Hadziavdic K, Lekang K, Lanzen A, Jonassen I, Thompson EM, Troedsson C, 2014. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS One 9:* e87624.
18. Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM, 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One 4:* e6372.
19. Kumar S, Gadagkar SR, 2001. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics 158:* 1321–1327.

20. Lee FCH, 2019. Finding *Sarcocystis* spp. on the Tioman Island: 28S rRNA gene next-generation sequencing reveals nine new *Sarcocystis* species. *J Water Health 17:* 416–427.
21. Briscoe AG, Goodacre S, Masta SE, Taylor MI, Arnedo MA, Penney D, Kenny J, Creer S, 2013. Can long-range PCR Be used to amplify genetically divergent mitochondrial genomes for comparative phylogenetics? A case study within spiders (Arthropoda: Araneae). *PLoS One 8:* e62404.
22. Gjerde B, Vikøren T, Hamnes IS, 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. *Int J Parasitol Parasites Wildl 7:* 1–11.
23. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C, 2018. The third revolution in sequencing technology. *Trends Genet 34:* 666–681.
24. Bae CH, Robbins RT, Szalanski AL, 2010. Secondary structure models of D2-D3 expansion segments of 28S rRNA for Hoplolaiminae species. *J Nematol 42:* 218–229.
25. Deiner K, Renshaw MA, Li Y, Olds BP, Lodge DM, Pfrender ME, 2017. Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods Ecol Evol 8:* 1888–1898.