Cryptosporidium in fish: alternative sequencing approaches and analyses at multiple loci to resolve mixed infections

ANDREA PAPARINI[1]*, RONGCHANG YANG[1], LINDA CHEN[1], KAISING TONG[1], SUSAN GIBSON-KUEH[2], ALAN LYMBERY[2] *and* UNA M. RYAN[1]

[1] *Vector and Water-Borne Pathogen Research Group, School of Veterinary & Life Sciences, Murdoch University, WA, Australia*
[2] *Freshwater Fish Group and Fish Health Unit, School of Veterinary and Life Sciences, Murdoch University, Murdoch, Western 6150, Australia*

(*Received 5 April 2017; revised 1 June 2017; accepted 8 June 2017; first published online 14 August 2017*)

SUMMARY

Currently, the systematics, biology and epidemiology of piscine *Cryptosporidium* species are poorly understood. Here, we compared Sanger – and next-generation – sequencing (NGS), of piscine *Cryptosporidium*, at the 18S rRNA and actin genes. The hosts comprised 11 ornamental fish species, spanning four orders and eight families. The objectives were: to (i) confirm the rich genetic diversity of the parasite and the high frequency of mixed infections; and (ii) explore the potential of NGS in the presence of complex genetic mixtures. By Sanger sequencing, four main genotypes were obtained at the actin locus, while for the 18S locus, seven genotypes were identified. At both loci, NGS revealed frequent mixed infections, consisting of one highly dominant variant plus substantially rarer genotypes. Both sequencing methods detected novel *Cryptosporidium* genotypes at both loci, including a novel and highly abundant actin genotype that was identified by both Sanger sequencing and NGS. Importantly, this genotype accounted for 68·9% of all NGS reads from all samples (249 585/362 372). The present study confirms that aquarium fish can harbour a large and unexplored *Cryptosporidium* genetic diversity. Although commonly used in molecular parasitology studies, nested PCR prevents quantitative comparisons and thwarts the advantages of NGS, when this latter approach is used to investigate multiple infections.

Key words: *Cryptosporidium*, fish, next-generation sequencing (NGS), 18S rRNA gene, actin gene, Sanger sequencing, molecular phylogeny, molecular systematics, parasitology, mixed infections.

INTRODUCTION

Currently three species of *Cryptosporidium* are recognized in fish: *Cryptosporidium molnari* (Alvarez-Pellitero and Sitja-Bobadilla, 2002; Palenzuela *et al*. 2010), *C. scophthalmi* (Alvarez-Pellitero *et al*. 2004) and *C. huwi* (Ryan *et al*. 2015). Molecular studies, however, have highlighted an extensive genetic diversity – often unexplored and occurring as mixed infections – within fish-derived *Cryptosporidium* species; a further 12 genotypes have been identified (piscine genotypes 2–8 and 5 un-named genotypes) (Murphy *et al*. 2009; Palenzuela *et al*. 2010; Reid *et al*. 2010; Zanguee *et al*. 2010; Morine *et al*. 2012; Koinari *et al*. 2013; Ryan *et al*. 2015; Yang *et al*. 2015). Elucidating this diversity is important to advance our understanding of parasite–host interactions, host specificity, epidemiology, phylogeny and public health and veterinary implications.

Unless cloning is performed, Sanger sequencing has proven unsuitable to sequence a mixture of amplicons generated by genus-specific primers, co-amplifying multiple genetic variants of *Cryptosporidium*. Unlike the Sanger method, at adequate sequencing depths, next-generation sequencing (NGS), can allow resolving mixtures of amplicons, thanks to the massive parallelization of the sequencing reaction. NGS has already been successfully exploited to characterize the genotypes present in mixed human infections of influenza virus (H1N1) (Ghedin *et al*. 2011) and cytomegalovirus (HCMV) strains (Gorzer *et al*. 2010).

Attempts to reconstruct the evolutionary relationships between *Cryptosporidium* species in fish, using nested PCR and conventional Sanger sequencing, have been hampered by a lack of concordance between the commonly utilized markers 18S rRNA and actin (Yang *et al*. 2015). The inconsistency resulted in conflicting phylogenetic trees at the actin and 18S loci, with the main clades identified by the 18S not reproduced at the actin locus. It is likely that the discrepancy is due to frequent mixed infections and diverse genetic constraints associated with the two phylogenetic markers, which resulted in diverse discriminatory power. This limitation provided the rationale for the present investigation and prompted the adoption of NGS methods.

The purpose of the present study was to test the potential of the Sanger method and NGS (ion semiconductor NGS), for identifying and typing

* Corresponding author: Vector- and Water-Borne Pathogen Research Group, School of Veterinary & Life Sciences, Molecular and Biomedical Sciences, Murdoch University, 90 South Street, Murdoch WA, 6150, Australia. E-mail: A.Paparini@murdoch.edu.au

doi:10.1017/S0031182017001214

Table 1. Taxonomic classification of the ornamental fish species analysed during the present study (if known)

Common name	Order	Family	Species	No. of fish
Neon Tetra	Characiformes	Characidae	*Paracheirodon innesi*	7
Goldfish	Cypriniformes	Cyprinidae	*Carassius auratus*	3
Tiger Barb	Cypriniformes	Cyprinidae	*Puntigrus tetrazona*	1
Guppy	Cyprinodontiformes	Poeciliidae	*Poecilia reticulata*	2
Blue Tang	Perciformes	Acanthuridae	*Paracanthurus hepatus*	1
Oscar	Perciformes	Cichlidae	*Astronotus ocellatus*	1
Yellow-Headed Jawfish	Perciformes	Opistognathidae	*Opistognathus aurifrons*	1
Azure Damsel	Perciformes	Pomacentridae	*Chrysiptera hemicyanea*	2
Orange Clownfish	Perciformes	Pomacentridae	*Amphiprion percula*	2
Red Stripe Angelfish	Perciformes	Pomacentridae	*Centropyge eibli*	1
Peach Anthias	Perciformes	Serranidae	*Pseudanthias dispar*	1
Unknown				1

Cryptosporidium species in fish, at both the 18S and actin loci. The main objectives were: to (i) confirm the rich genetic diversity of piscine *Cryptosporidium* spp. and the high frequency of mixed infections in ornamental fish, and (ii) to explore the potential of NGS as a molecular typing tool, in the presence of complex genetic mixtures. We anticipate that improving the sampling effort of fish-derived *Cryptosporidium* species and genotypes, and resolving the current technical limitations is important to support future phylogenetic studies and improve the robustness of the current molecular systematics. A greater understanding of piscine *Cryptosporidium* infections will also clarify the broader implications bore by ornamental fish release in natural environments.

MATERIALS AND METHODS

Samples and PCR amplification

From various fish hosts, a total of 23 fish genomic DNA preparations were obtained (Table 1) (NCBI BioProject ID: 326557; Accession No. PRJNA326557). Fishes were sourced from a commercial aquarium shop in Perth, Western Australia, and euthanized using an ice slurry upon arrival at the laboratory. The ethics committee approved the study under Murdoch University animal ethics permits W2325/10 and RW2618/13. Dissections and DNA extractions from ~25 mg of gastric and intestinal tissues of each fish were carried out as previously published, using the PowerSoil DNA Isolation Kit (Qiagen, Hilden Germany, formerly MO BIO, Carlsbad, CA, USA) (Yang *et al.* 2015). Tissues were scraped using sterile scalpel blades and surgical instruments. This procedure increases the chances of collecting oocysts that are embedded in the fish tissues (i.e. infecting), rather than simply contained within the stomach and gut. Extraction blanks, consisting of mock extractions using DNA-free reagents and consumables, were also included as controls. An earlier investigation (Yang *et al.* 2015) analysed 55 piscine samples that included 21 samples from the present study, but used an alternative set of 18S primers, and different phylogenetic reconstructions, based on Sanger sequencing only. NGS was not performed in the previous study (Yang *et al.* 2015).

Samples and controls were initially amplified at the 18S locus using the *Cryptosporidium*-specific primer pair 18SiF AGTGACAAGAAATAACAA TACAGG and 18SiR CCTGCTTTAAGCACTC TAATTTTC (~292 bp) (Morgan *et al.* 1997). However, this single-round PCR approach failed to either amplify any product or to produce sufficient template for NGS. As a result, a nested PCR was used to amplify samples at the 18S locus, using the primary primer pair 18SiCF2 GACATATCATTCAAG TTTCTGACC and 18SiCR2 CTGAAGGAGTAA GGAACAACC (~763 bp), located externally to the 18SiF/iR amplicon (Ryan *et al.* 2003), followed by the internal primer pair 18SiF and 18SiR (Morgan *et al.* 1997).

The following amplification conditions were used for 18SiCF2/18SiCR2 PCR: 94 °C – 5 min; then 45 cycles of: 94 °C – 30 s, 58 °C – 30 s, 72 °C – 30 s; followed by 72 °C – 7 min and 4 °C – hold. The following amplification conditions were used for 18SiF/18SiR PCR: 94 °C – 5 min; then 40 cycles of: 94 °C – 30 s, 58 °C – 20 s, 72 °C – 30 s; followed by 72 °C – 7 min and 4 °C – hold.

A hemi-nested PCR was used for the actin locus using the piscine-specific primers ActinallF1 GT-AAATATACAGGCAGTT and reverse primer ActinallR1 GGTTGGAACAATGCTTC (~392 bp) as previously described (Koinari *et al.* 2013). For the actin primary PCR, the conditions used were: 94 °C – 5 min; then 45 cycles of: 94 °C – 30 s, 46 °C – 30 s, 72 °C – 30 s; followed by 72 °C – 7 min and 4 °C – hold. For the actin secondary PCR, a fragment of ~278 bp was amplified using 1 μL of primary PCR product with forward primer ActinallF2 CCTCAT-GCTATAATGAG and reverse primer ActinallR1 (Koinari *et al.* 2013). The conditions used for the actin secondary PCR were identical to those for the primary PCR.

No template controls (NTC), consisting of DNA-free molecular grade water, were used during each PCR run. Sample preparation and amplification

https://doi.org/10.1017/S0031182017001214 Published online by Cambridge University Press

areas were physically separated to prevent contamination of test samples by PCR products. PCRs were conducted in a G-Storm thermal cycler (G-Storm, Somerton, Somerset, UK). Each reaction (25 μL) included: 0·4 mM of each primer, 0·8 mM dNTP mix, 2·0 mM $MgCl_2$, and 0·625 U Kapa Taq (Kapa Biosystems, Wilmington, MA, USA).

Sanger sequencing and phylogenetic analysis

PCR products were run on a 1% agarose gel containing SYBR Safe Gel Stain (Thermo Fisher Scientific, Waltham, MA, USA), and visualized with a dark reader trans-illuminator (Clare Chemical Research, Dolores, CO, USA). For Sanger sequencing, secondary amplicons were purified using an in-house filter tip-based method as previously described (Yang *et al*. 2013). Amplicons were sequenced in both directions using an ABI Prism Dye Terminator cycle sequencing kit (Applied Biosystems, Foster City, CA, USA) and a 3730 × 1 DNA Analyzer (Applied Biosystems, USA), according to the manufacturer's instructions. Chromatograms obtained by Sanger sequencing were visualized and checked individually using Finch TV Version 1.4.0 (http://www.geospiza.com). Sanger sequences were submitted to GenBank and are available under accession numbers KX527727 to KX527747 (18S rRNA) and KX453766 to KX453782 (actin). In Geneious pro 8.1.8 (Kearse *et al*. 2012), paired forward and reverse sequences were merged to generate consensus sequences, allowing no mismatches within the overlapping region. These were identified based on phylogenetic reconstructions (at the two loci), including matching hits from GenBank retrieved by BLAST searches (option megablast).

Two bootstrapped phylogenetic reconstructions of the Sanger sequencing results (actin and 18S rRNA genes) were carried out in Geneious pro 8.1.8 (Kearse *et al*. 2012), using the plugins MAFFT v.7.017 (Katoh and Standley, 2013) and FastTree v. 2.15 (Price *et al*. 2010) (options: GTR model; optimized Gamma20 likelihood). The tree topology was compared with previous analyses obtained with longer sequences at the same two loci (Yang *et al*. 2015). For these two analyses, the genetic distance was calculated as the percentage of base differences per site from between sequences (p-dist option).

All ambiguous positions were removed for each sequence pair (pairwise deletion option). The analysis at the 18S rRNA locus involved 40 nucleotide sequences (256 positions in the final 18S rRNA dataset). The analysis at the actin locus involved 28 nucleotide sequences (197 positions in the final actin dataset). Codon positions included were 1st + 2nd + 3rd + non-coding. Evolutionary analyses were conducted in MEGA6 (Tamura *et al*. 2013).

A third analysis was conducted at the actin locus only, to specifically resolve the phylogenetic position of the operational taxonomic units (OTUs) obtained by NGS that were classified as 'unidentified' during the taxonomy assignment step implemented. This analysis was conducted using Geneious pro 8.1.8 (alignment, global trimming, basic manipulations and edits) (Kearse *et al*. 2012) and MEGA6 (calculation of genetic distance, selection of the best nucleotide substitution model and tree building) (Tamura *et al*. 2013). For this third analysis (actin locus), the bootstrapped phylogenetic reconstruction used the Maximum Likelihood method based on the Tamura three-parameter model. A discrete γ distribution was used to model evolutionary rate differences among sites [five categories (+G, parameter = 2·2201)]. The analysis involved 74 nucleotide sequences. Codon positions included were 1st + 2nd + 3rd + non-coding. There were a total of 209 positions in the final dataset.

Ion semiconductor NGS

The same combination of primer pairs used for Sanger sequencing was also used for NGS. The template preparation workflow was performed according to the manufacturer's recommendations on the One-Touch 2/One-Touch ES system (Thermo Fisher Scientific, Waltham, MA, USA, formerly Life Technologies, Camarillo, CA, USA). Sequencing was performed on an Ion Torrent PGM (Thermo Fisher Scientific, Waltham, MA, USA, formerly Life Technologies, Camarillo, CA, USA) using the Ion PGM Sequencing 400 Kit and 316-V2 semiconductor chips, following the manufacturer's recommendations. Fusion primers (IDT, Coralville, IA, USA) were based on the secondary primers 18SiF and 18SiR (Morgan *et al*. 1997) and ActinallF2 and ActinallR1, and included unique sample-specific barcodes (MID tags) and P1 and A adaptors. All PCR amplicons were double purified using the Agencourt AMPure XP Bead PCR purification protocol (Beckman Coulter Genomics, Brea, CA, USA), pooled in roughly equimolar ratios after Qubit fluorometric quantitations (Thermo Fisher Scientific, Waltham, MA USA) and sequenced.

Data deconvolution and bioinformatics analysis

From the raw sequencing output file (3 512 884 sequences), *Cryptosporidium*-specific amplicons (reads) were extracted, bioinformatically, parsing the sequence for the 18S rRNA and actin primers, with Geneious Pro 8.1.8 (Biomatters Ltd, Auckland, NZ) (Kearse *et al*. 2012). The reads were then de-multiplexed under stringent conditions into sample batches, using the unique sample-specific barcodes (MID tags), as previously described (Paparini *et al*. 2015). Only sequences exhibiting full-length and exact matches to the flanking regions were processed further. MID tags,

https://doi.org/10.1017/S0031182017001214 Published online by Cambridge University Press

Table 2. Sanger sequencing-based taxonomic identifications of *Cryptosporidium*-specific amplicons obtained from 23 fish samples, at the 18S rRNA and actin loci

Fish ID	Fish species	18S rRNA (GenBank match) (% p-distance)	Actin (GenBank match) (% p-distance)	Agreement between loci
KS109	Neon Tetra	*Cryptosporidium huwi* (HM989835) (0·4)	*C. huwi* (AY524772) (0·5)	Match
KS123	Neon Tetra	*C. huwi* (HM989835) (0·8)	NA	
KS27	Oscar	Piscine genotype 2 (FJ769050) (0·4)	NA	
KS33	Neon Tetra	*C. huwi* (HM989835) (0·0)	NA	
KS35	Neon Tetra	*C. huwi* (HM989835) (0·4)	*C. huwi* (AY524772) (0·5)	Match
KS36	Neon Tetra	*C. huwi* (HM989835) (0·0)	*C. huwi* (AY524772) (0·0)	Match
KS37	Neon Tetra	*C. huwi* (HM989835) (0·0)	*C. huwi* (AY524772) (0·5)	Match
KS39	Neon Tetra	*C. huwi* (HM989835) (0·0)	NA	
KS43	Guppy	*C. huwi* (HM989835) (0·0)	*C. huwi* (AY524772) (0·5)	Match
KS46	Guppy	*C. huwi* (HM989835) (0·4)	*C. huwi* (AY524772) (0·5)	Match
KS52	Tiger Barb	*C. huwi* (HM989835) (0·4)	*C. huwi* (AY524772) (0·5)	Match
LC01	Orange Clownfish	Novel genotype LC01 (**KX527738**)	*Cryptosporidium molnari*-like (KR610337) (1·0)	Mismatch
LC04	Unknown	NA	NA	
LC06	Yellow-Headed Jawfish	*C. molnari*-like (KR610356) (0·0)	*C. molnari*-like (KR610337) (0·5)	Match
LC09	Blue Tang	Piscine genotype 5 (HM989837) (1·3)	Piscine genotype 5 (KR610339) (0·5)	Match
LC12	Peach Anthias	*C. molnari*-like (KR610356) (0·9)	*C. molnari*-like (KR610337) (0·5)	Match
LC16	Goldfish	*C. molnari*-like (KR610356) (0·0)	*C. molnari*-like (KR610337) (0·0)	Match
LC38	Goldfish	*C. molnari*-like (KR610356) (0·0)	*C. molnari*-like (KR610337) (0·0)	Match
LC47	Azure Damsel	*C. molnari*-like (KR610356) (0·9)	*C. molnari*-like (KR610337) (0·0)	Match
LC48	Orange Clownfish	Novel genotype LC48 (**KX527745**)	Novel genotype LC48 (**KX453780**)	Match
LC50	Goldfish	NA	NA	
LC51	Azure Damsel	Novel genotype LC51 (**KX527746**)	*C. molnari*-like (KR610337) (0·5)	Mismatch
LC73	Red Stripe Angelfish	*C. molnari*-like (KR610356) (0·9)	*C. molnari*-like (KR610337) (0·0)	Match

Identification of *Cryptosporidium* sequences obtained during the present study is based on phylogenetic reconstructions and % genetic distance (p-dist) from the closest GenBank matches (given in brackets). Novel genotypes obtained during the present study that had no close matches in GenBank were submitted to GenBank with a new accession number (in bold). Novel genotypes showed minimum genetic distances ranging from 2·5% (18S rRNA) to 13·2% (actin) from the closest matches already available in GenBank.
NA, no amplification.

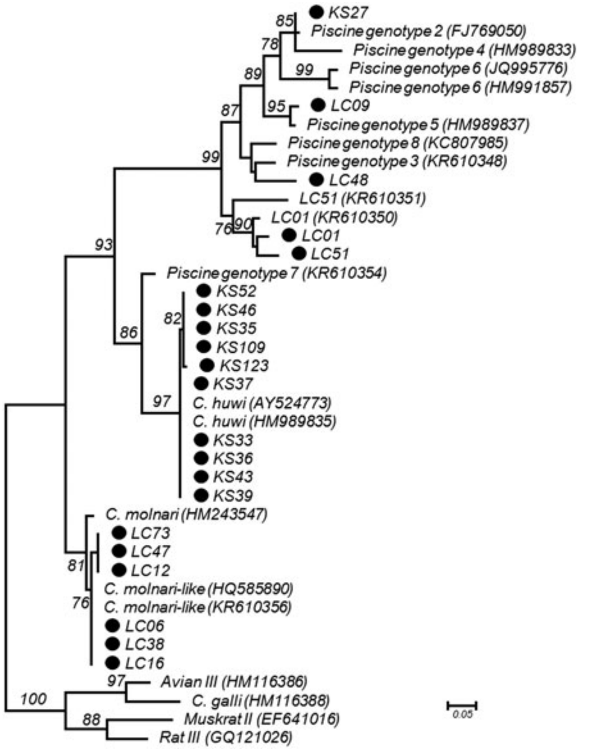

Fig. 1. Phylogenetic tree showing the position of the 18S genotypes identified by Sanger sequencing.

sequencing adapters and primers were trimmed. NGS data are available under NCBI BioProject ID: 326557 (Accession No. PRJNA326557). The reads were quality-filtered using USEARCH 8.1 (Edgar, 2010) (-fastq_filter; maxee settings = 2·0); ⩾90·1% of the reads passed the filtering step. At this stage, the quality-filtered reads were 906 267 and 934 411, for actin and 18S rRNA, respectively. Singletons were removed (on a per-run basis), prior to clustering into OTUs, at 97% similarity. High-throughput OTU clustering was performed by the UPARSE-OTU algorithm (http://drive5.com/uparse/) (Edgar, 2013). The cluster_otus command in USEARCH was used; ⩾99·0% of the reads were assigned to OTUs and the chimeras proportion was ⩽1·0%. OTUs were checked with the UCHIME algorithm (-uchime_denovo) (Edgar *et al*. 2011) to ensure OTUs were not the result of chimeric reads, and chimeras were discarded. High-stringency taxonomic assignments were performed in QIIME v. 1.9.1 (Caporaso *et al*. 2010), using the BLAST algorithm (−*e* 0·0001) and two databases curated and edited ‘in-house’. These included *Cryptosporidium* 18S rRNA and actin sequences, obtained from GenBank and/or during previous studies from our group. The databases for the 18S and actin loci consisted of 68 (length range: 243–272 bp) and 64 (length range: 197–1103 bp) overlapping *Cryptosporidium* sequences from multiple vertebrate hosts, respectively. Data representation, multivariate analyses and diversity core analyses were also carried out in QIIME (Caporaso *et al*. 2010).

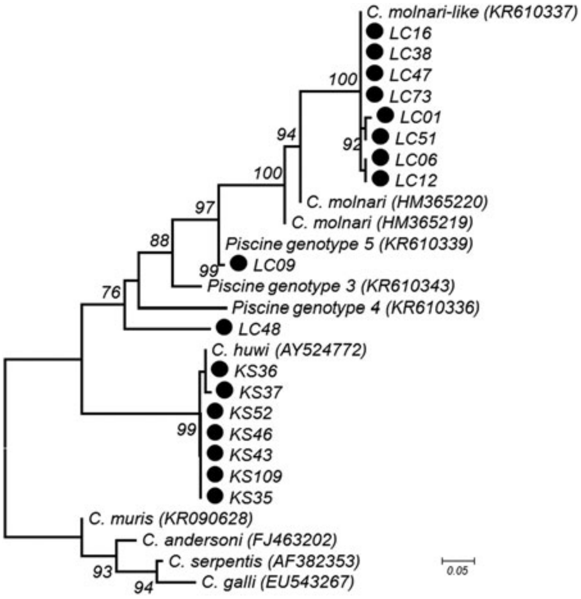

Fig. 2. Phylogenetic tree showing the position of the actin genotypes identified by Sanger sequencing.

RESULTS

Sanger sequencing of the 18S rRNA and actin loci

Sanger sequencing at the 18S locus was successful for 21 of the 23 fishes tested (two samples, LC04 and LC50, did not amplify) (Table 2). The following species/genotypes were identified: piscine genotype 2 in one sample (KS27); *C. huwi* in 10 samples (KS33, KS35, KS36, KS37, KS39, KS43, KS46, KS52, KS109 and KS123); *C. molnari*-like genotype in six samples (LC06, LC12, LC16, LC38, LC47 and LC73); piscine genotype 5 in one sample (LC09); three samples were more unique variants with genetic distances 2·5–6·4% from the closest genotype already present in GenBank (LC01, LC48 and LC51) (Fig. 1; Table 2 and Supplementary Table S1).

At the actin locus 17, Sanger sequences were obtained (Table 2). Of these, *C. huwi* was identified in seven samples (KS35, KS36, KS37, KS43, KS46, KS52 and KS109), *C. molnari*-like genotype in eight samples (LC01, LC06, LC12, LC16, LC38, LC47, LC51 and LC73), piscine genotype 5 in one sample (LC09) and a novel genotype in one sample (LC48) with >13% genetic distance from piscine genotype 3 (KR610343) (Fig. 2; Table 2 and Supplementary Table S2).

Out of the 23 samples analysed, Sanger sequencing worked at both loci for 17 fish samples. Four samples provided only 18S rRNA sequences, and two more samples (LC04 and LC50) provided no sequences. When Sanger sequencing worked at both loci ($n = 17$), in 15 cases, there was agreement between loci (Table 2). As previously noted (Yang *et al*. 2015), a 18S rRNA gene sequence classified as *C. molnari* in GenBank (accession number HQ585890) does in fact belong to the *C. molnari*-

Table 3. Taxonomic identification of *Cryptosporidium*-specific amplicons obtained from 23 fish samples, at the 18S rRNA and actin loci

	NGS 18S rRNA (GenBank match)											NGS actin (GenBank match)			
Fish ID	PG 2 (FJ769050)	Avian III (HM116386)	*C. galli* (HM116388)	LC51 (KR610351)	Muskrat I (EF641016)	LC01 (KR610350)	PG 5 (HM989837)	Rat III (GQ121026)	Novel genotype LC48 (KX527745)	*Cryptosporidium molnari*-like (KR610356)	*Cryptosporidium huwi* (AY524773)	*C. molnari*-like (KR610337)	PG 3 (KR610343)	*C. huwi* (AY524772)	Novel genotype LC48 (KX453780)
KS109											7434			19 786	
KS123											8085			14 986	
KS27	4										24 904				19
KS33											26 430				21 157
KS35											9801			17 421	
KS36											7 586			2	20 978
KS37											23 686			2	21 452
KS39											22 787			21 274	
KS43										1	25 973			10 018	
KS46											20 805			12 182	
KS52										1	10 128			16 001	
LC01				25		718									
LC04		5	5		107			2041	2	2					31 995
LC06									1	33 154	1				27 464
LC09							1			15 088		407	701		
LC12										29 756				1	30 172
LC16										265	23 898				28 970
LC38										33 220			2		15 507
LC47										29 407		4			6
LC48															23 297
LC50										14 738	1				42
LC51								1	3790		9				28 522
LC73							759			1	1 179				4
Total	4	5	5	25	107	718	760	2042	3793	155 633	212 707	411	703	111 673	249 585

Numbers represent the number of reads obtained by NGS, from each fish. Identification of *Cryptosporidium* sequences are based on high-stringency taxonomic assignments performed using the BLAST algorithm (−*e* 0·0001) and two databases curated and edited 'in-house'. These included *Cryptosporidium* 18S rRNA and actin sequences, obtained from GenBank and/or during previous studies from our group. The databases for the 18S and actin loci consisted of 68 (length range: 243–272 bp) and 64 (length range: 197–1103 bp) overlapping *Cryptosporidium* sequences from multiple vertebrate hosts, respectively.

PG, piscine genotype.

like genotype (e.g. KR610356) (Fig. 1), which is genetically distinct from *C. molnari*. For instance, the genetic distance between the *C. molnari*-like genotype and *C. molnari* was 1·3% and ⩽7·1%, at the 18S rRNA and actin genes, respectively (Supplementary Tables S1 and S2).

Ion semiconductor NGS of 18S rRNA and actin loci

The raw output of the sequencing run consisted of 1 036 903 and 952 243 reads, for 18S rRNA and actin, respectively (934 411 and 906 267 reads, after quality filtering). Only three sequences were obtained from the controls (three 18S rRNA sequences from the NTC). After chimera removal, 68 and 76 OTUs were identified, for 18S rRNA and actin, respectively.

To remove sampling depth heterogeneity, α-rarefaction plots were generated at both loci, after rarefying at 2313 and 3081 reads/sample for 18S rRNA and actin, respectively. Flatter rank-abundance curves were observed at both loci for the samples ($n \approx 4$) that, compared with the others, yielded the least reads at either locus. Despite this, all curves appeared to plateau after about 500 reads, based on virtually all metrics available in QIIME (Caporaso *et al.* 2010), suggesting that an adequate sampling depth had been obtained and allowing assessment of the effect of possible sequencing error on accumulation of sequence diversity (Supplementary Figs S1 and S2).

At each locus, NGS was successful for 22 of the 23 fishes: LC01 failed to amplify at the actin locus, and LC48 failed to amplify at the 18S rRNA locus; therefore, 21 fish samples were successfully analysed by NGS at both loci concurrently. Of these 21 samples, seven fishes provided consistent identifications according to the 18S rRNA – and actin – loci (all *C. huwi*).

The total number of reads obtained from 22 samples was 375 799 for the 18S rRNA locus and 362 372 for actin (Table 3). Cluster analyses on 18S NGS data seemed to indicate that co-infections between *C. huwi* and the *C. molnari*-like genotype were uncommon and that all Neon Tetra fishes were infected only by the first species (data not shown).

Multiple genotype/species were often detected by NGS. In reality, most samples harbouring multiple variants were characterized by the presence of only one highly dominant variant plus nine reads or less of one (or more) rarer genotype(s). These latter sequences may represent spurious reads or rarer variants not adequately covered by the sequencing depth implemented during the present study. If rare reads (<9) are ignored, five fishes clearly appeared to harbour mixed infections: two distinct 18S rRNA genotypes/species were obtained by NGS from samples LC01 [LC51 (KR610351) and LC01 (KR610350)], LC04 [Muskrat genotype I (EF641016) and Rat genotype III (GQ121026)], LC16 [(*C. molnari*-like (KR610356) and *C. huwi* (AY524773)] and LC73 [piscine genotype 5 (HM989837) and *C. huwi* (AY524773)]. Sample LC09 harboured two actin variants [LC47 (novel genotype) and piscine genotype 3 (KR610343)].

Identification of a novel actin clade using both Sanger an NGS (LC48)

A novel actin monophyletic clade was identified by both Ion Torrent and Sanger. The clade had 87% bootstrap support value and included 46 of the 76 actin OTUs generated. By Ion Torrent, this group was highly abundant and accounted for 68·9% of all NGS actin reads from all samples (249 585/362 372). During the OTU clustering step (after removal of chimeras), these reads were bioinformatically associated to 46 related OTUs.

The OTUs were successively taxonomically assigned using curated custom databases and the BLAST algorithm with settings of the expected $-e$ value more stringent than the default ($-e = 0{\cdot}0001$ instead of 0·001); assignments were also confirmed by BLAST searches against the complete nr/nt database at NCBI. While OTU 01 (223 bp) was found 248 963 times, the other OTUs were less abundant (⩽220 copies). BLAST searches against a curated custom actin database showed that all novel OTUs matched with the Sanger LC48 actin sequence (KX453780) (Tables 2 and 3; Supplementary Fig. S3).

Within this heterogeneous monophyletic clade, consisting of a group of related genotypes, the average genetic distance was 5·7% with a maximum value of 14·3% (i.e. between OTUs 07, 16, 31, 60 and 72). The minimum genetic distance from the Sanger LC48 actin sequence (KX453780) was 1·1% (e.g. OTUs 01, 09, 12, 38 and 39). The clade including the Sanger LC48 actin sequence (KX453780) and all the NGS OTUs exhibited 13·2 genetic distances from piscine genotype 3 (KR610343) and piscine genotype 4 (KR610336) (Supplementary Fig. S3).

The existence of the novel clade including LC48 and the OTUs obtained by NGS is compelling and strongly supported by the number of highly similar NGS sequences obtained. Therefore, fish LC48 harboured a novel genotype that was detected both by sequencing methods, but NGS captured a greater diversity associated with this group of genetic variants (Supplementary Fig. S3).

Agreement between sequencing platforms at 18S and actin loci

At the 18S rRNA locus, 20/23 fish samples analysed provided both NGS – and Sanger sequencing – results. Of these 20 fishes, there were 16 corresponding identifications between NGS – and Sanger sequencing, including the species/genotypes *C. huwi* (HM989835), *C. molnari*-like (KR610356)

https://doi.org/10.1017/S0031182017001214 Published online by Cambridge University Press

and the LC01 (KX527738). Multiple NGS variants were obtained by NGS from fishes LC01 and LC16; however, these variants also included those detected by Sanger sequencing.

At the actin locus, 15/23 fish samples analysed provided both NGS – and Sanger sequencing – results. Of these 15 fishes, there were five corresponding identifications between NGS and Sanger sequencing: four fishes (KS43, KS46, KS52 and KS 109) were *C. huwi* (GenBank: AY524772) and one fish (LC48) was the same novel genotype (LC48) detected by both sequencing methods.

DISCUSSION

In the present study, 23 *Cryptosporidium*-positive fishes were assessed by molecular interrogation comparing ion semiconductor NGS and Sanger sequencing approaches. The set of samples consisted of 11 common species of ornamental fish, belonging to four orders and eight families (Table 1). Overall, molecular identification suggested a broad host range for *C. huwi* and *C. molnari*-like genotype, which are the two main *Cryptosporidium* spp. detected during the present study. Similarly, different genotypes were also found co-infecting the same host species. These indications, however, require further testing on more species of fish hosts. The present study has also confirmed the genetic distinctness of the *C. molnari*-like genotype and, given the very large genetic distance from *C. molnari* at both loci tested (1·3–7·1%), it is likely a separate species. Further analysis is required to confirm this.

The present study also assessed the concordance between the molecular identification allowed by alternative sequencing methods. During the present study, 375 799 18S rRNA and 362 372 actin NGS reads were obtained from 23 *Cryptosporidium*-positive fish samples. In line with the comparative focus of the present study and irrespective of the number of reads per sample, all samples providing *Cryptosporidium*-specific sequences, by either Sanger sequencing or Ion Torrent (at least at one locus), were included in the analysis. Thus, fish LC04, which yielded 18S rRNA avian and rodent genotypes not usually associated with piscine hosts (Table 3) and the novel LC48 genotype at the actin locus, was maintained in the study. For the same reason, NGS data in Table 3 are presented without rarefaction, to better appreciate the detection of rarer species by the NGS-based approach, in comparison to the more conventional Sanger sequencing method.

A previous analysis using nested PCR and Sanger sequencing of 55 piscine samples, which included 21 of the 23 samples analysed in the present study, also reported poor concordance between the 18S and actin loci (Yang *et al*. 2015). The present investigation utilized some DNA preparations obtained previously (Yang *et al*. 2015). The genetic characterization of samples were carried out independently, during the two studies and, more importantly, in the analysis of Yang *et al*. (2015), a different 18S nested PCR was used (Silva *et al*. 2013) compared with the current study (Morgan *et al*. 1997; Ryan *et al*. 2003). The primers of Silva *et al*. (2013) produce a longer secondary amplicon (~553 bp) compared with the 292 bp secondary product produced by the primers used in the present study (Morgan *et al*. 1997). The shorter amplicon size was chosen for the present study to make it compatible with the amplicon length recommended by Ion Torrent system. The same actin primers were used for both studies.

Disagreements observed between the 18S and actin loci for both Sanger and Ion Torrent are most likely due to the presence of mixed infections. For samples containing mixed genotypes, the discrepancies in the identifications obtained by different loci is likely due to preferential amplification of one genotype over another. Novel genotype-specific PCR primers may alleviate this problem. Unfortunately, in the present study, we were unable to amplify the samples using single-round PCR, due to low parasite DNA concentrations in the samples. Previous studies have implemented a nested PCR approach to screen fish immunobiomes by NGS (Boutin *et al*. 2012). However, it is generally recognized that nested PCR approaches have an inherent risk of contamination and have previously been shown to exhibit strong amplification biases and/or stochastic variation (Park and Crowley, 2010). By involving two sequential rounds of amplification, nested PCR may not accurately represent the extent of genetic diversity initially present in the sample, because it introduces a bottleneck between the first and second rounds. In molecular parasitology studies however, nested PCR is often critical to obtain enough DNA copies to sequence by the Sanger method. This is an inherent problem with *Cryptosporidium* epidemiology, as environmental water samples, gastric/intestinal tissues from fish or feces from wildlife, frequently contain very low numbers of oocysts and high levels of PCR inhibitors, and therefore nested PCR is often necessary to amplify the parasite DNA. However, the bottleneck effect counteracts the potential of NGS, which, unlike the Sanger method, can be particularly useful in the presence of mixed infections. Previous studies have also reported a strong bias when using a nested PCR approach for Ion Torrent sequencing (Whiteley *et al*. 2012).

Concluding remarks

The present study confirms that ornamental fish can harbour a large variety of novel *Cryptosporidium* spp. and genotypes. This may be due to the enclosed environment these animals are restricted to, which

https://doi.org/10.1017/S0031182017001214 Published online by Cambridge University Press

may favour transmission and higher prevalence of infections. For the same reason, we speculate that *Cryptosporidium* spp. prevalence in aquariums may not reflect that found in other wild piscine species. The pathogenicity and public health significance of these protozoan species, including the novel LC48 genotype identified from the present study, require further studies and improvement of amplification efficiency and amplicon length.

For the present comparative study, primer sets and test conditions matched rigorously between the alternative sequencing methods evaluated. The implemented assays have been extensively validated in previous publications. Nonetheless, future similar comparisons should be extended to DNA extractions from other matrices (tissues, feces, sediments, water, etc.), and other loci, pathogens and single-round amplification protocols.

From a technical point of view, although often essential in molecular parasitology studies, nested PCR can change the relative proportion of the species and genotypes of parasites initially present in the sample and therefore the benefits of NGS, and quantitative comparisons between alternative DNA sequencing methods are hampered by the need to used nested PCR to produce sufficient DNA products for analyses.

SUPPLEMENTARY MATERIAL

The supplementary material for this article can be found at https://doi.org/10.1017/S0031182017001214.

ACKNOWLEDGEMENTS

Authors' gratitude also goes to the West Coast Microbiome Network (http://www.wcmicrobiome.org), Dr David Berryman, Ms Frances Brigg, Mr Alex Gofton and Dr Elvina Lee (Murdoch University) for assistance. Resources of the Pawsey Supercomputing Centre were used during some steps of the data analysis (BLAST searches).

FINANCIAL SUPPORT

Partial financial support for this study was indirectly provided by Murdoch University, the Australian Research Council, Bayer Healthcare, Bayer Australia Ltd, Western Australia Water Corporation, WaterRA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Alvarez-Pellitero, P. and Sitja-Bobadilla, A. (2002). *Cryptosporidium molnari* n. sp. (Apicomplexa: Cryptosporidiidae) infecting two marine fish species, *Sparus aurata L.* and *Dicentrarchus labrax L. International Journal for Parasitology* **32**, 1007–1021.

Alvarez-Pellitero, P., Quiroga, M.I., Sitja-Bobadilla, A., Redondo, M.J., Palenzuela, O., Padros, F., Vazquez, S. and Nieto, J.M. (2004). *Cryptosporidium scophthalmi* n. sp. (Apicomplexa: Cryptosporidiidae) from cultured turbot *Scophthalmus maximus*. Light and electron microscope description and histopathological study. *Diseases of Aquatic Organisms* **62**, 133–145.

Boutin, S., Sevellec, M., Pavey, S.A., Bernatchez, L. and Derome, N. (2012). A fast, highly sensitive double-nested PCR-based method to screen fish immunobiomes. *Molecular Ecology Resources* **12**, 1027–1039.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J. R., Tumbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461.

Edgar, R.C (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**, 996–8.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200.

Ghedin, E., Laplante, J., DePasse, J., Wentworth, D.E., Santos, R.P., Lepow, M.L., Porter, J., Stellrecht, K., Lin, X., Operario, D., Griesemer, S., Fitch, A., Halpin, R.A., Stockwell, T.B., Spiro, D.J., Holmes, E.C. and St George, K. (2011). Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. *Journal of Infectious Diseases* **203**, 168–174.

Gorzer, I., Guelly, C., Trajanoski, S. and Puchhammer-Stockl, E. (2010). Deep sequencing reveals highly complex dynamics of human cytomegalovirus genotypes in transplant patients over time. *Journal of Virology* **84**, 7195–7203.

Katoh, K. and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. and Drummond, A. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649.

Koinari, M., Karl, S., Ng-Hublin, J., Lymbery, A.J. and Ryan, U.M. (2013). Identification of novel and zoonotic *Cryptosporidium* species in fish from Papua New Guinea. *Veterinary Parasitology* **198**, 1–9.

Morgan, U.M., Constantine, C.C., Forbes, D.A. and Thompson, R.C.A. (1997). Differentiation between human and animal isolates of *Cryptosporidium parvum* using rDNA sequencing and direct PCR analysis. *Journal of Parasitology* **83**, 825–830.

Morine, M., Yang, R., Ng, J., Kueh, S., Lymbery, A.J. and Ryan, U.M. (2012). Additional novel *Cryptosporidium* genotypes in ornamental fishes. *Veterinary Parasitology* **190**, 578–582.

Murphy, B.G., Bradway, D., Walsh, T., Sanders, G.E. and Snekvik, K. (2009). Gastric cryptosporidiosis in freshwater angelfish (*Pterophyllum scalare*). *Journal of Veterinary Diagnostic Investigation* **21**, 722–727.

Palenzuela, O., Alvarez-Pellitero, P. and Sitja-Bobadilla, A. (2010). Molecular characterization of *Cryptosporidium molnari* reveals a distinct piscine clade. *Applied and Environmental Microbiology* **76**, 7646–7649.

Paparini, A., Gofton, A., Yang, R.C., White, N., Bunce, M. and Ryan, U.M. (2015). Comparison of Sanger and next generation sequencing performance for genotyping *Cryptosporidium* isolates at the 18S rRNA and actin loci. *Experimental Parasitology* **151**, 21–27.

Park, J.W. and Crowley, D.E. (2010). Nested PCR bias: a case study of *Pseudomonas* spp. in soil microcosms. *Journal of Environmental Monitoring* **12**, 985–988.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010). FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490.

Reid, A., Lymbery, A., Ng, J., Tweedle, S. and Ryan, U. (2010). Identification of novel and zoonotic *Cryptosporidium* species in marine fish. *Veterinary Parasitology* **168**, 190–195.

Ryan, U., Xiao, L., Read, C., Zhou, L., Lal, A.A. and Pavlasek, I. (2003). Identification of novel *Cryptosporidium* genotypes from the Czech Republic. *Applied and Environmental Microbiology* **69**, 4302–4307.

Ryan, U., Paparini, A., Tong, K., Yang, R., Gibson-Kueh, S., O'Hara, A., Lymbery, A. and Xiao, L. (2015). *Cryptosporidium huwi* n. sp. (Apicomplexa: Eimeriidae) from the guppy (*Poecilia reticulata*). *Experimental Parasitology* **150**, 31–35.

Silva, S.O., Richtzenhain, L.J., Barros, I.N., Gomes, A.M., Silva, A.V., Kozerski, N.D., de Araujo Ceranto, J.B., Keid, L.B.

and Soares, R. M. (2013). A new set of primers directed to 18S rRNA gene for molecular identification of *Cryptosporidium* spp. and their performance in the detection and differentiation of oocysts shed by synanthropic rodents. *Experimental Parasitology* **135**, 551–557.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* **30**, 2725–2729.

Whiteley, A. S., Jenkins, S., Waite, I., Kresoje, N., Payne, H., Mullan, B., Allcock, R. and O'Donnell, A. (2012). Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. *Journal of Microbiological Methods* **91**, 80–88.

Yang, R., Murphy, C., Song, Y., Ng-Hublin, J., Estcourt, A., Hijjawi, N., Chalmers, R., Hadfield, S., Bath, A., Gordon, C. and Ryan, U. (2013). Specific and quantitative detection and identification of *Cryptosporidium hominis* and *C. parvum* in clinical and environmental samples. *Experimental Parasitology* **135**, 142–147.

Yang, R., Palermo, C., Chen, L., Edwards, A., Paparini, A., Tong, K., Gibson-Kueh, S., Lymbery, A. and Ryan, U. (2015). Genetic diversity of *Cryptosporidium* in fish at the 18S and actin loci and high levels of mixed infections. *Veterinary Parasitology* **214**, 255–263.

Zanguee, N., Lymbery, J. A., Lau, J., Suzuki, A., Yang, R., Ng, J. and Ryan, U. (2010). Identification of novel *Cryptosporidium* species in aquarium fish. *Veterinary Parasitology* **174**, 43–48.

https://doi.org/10.1017/S0031182017001214 Published online by Cambridge University Press