SPECIAL ISSUE ARTICLE

Suitability of loci for multiple-locus variable-number of tandem-repeats analysis of *Cryptosporidium parvum* for inter-laboratory surveillance and outbreak investigations

RACHEL M. CHALMERS[1,2]*†, GUY ROBINSON[1,2]†, EMILY HOTCHKISS[3]†, CLAIRE ALEXANDER[4], SOPHIE MAY[1], JANICE GILRAY[3], LISA CONNELLY[4] *and* STEPHEN J. HADFIELD[1]

[1] *Cryptosporidium Reference Unit, Public Health Wales Microbiology, Singleton Hospital, Swansea SA2 8QA, UK*
[2] *Swansea University Medical School, Grove Building, Swansea University, Singleton Park, Swansea SA2 8PP, UK*
[3] *Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik, Edinburgh EH26 0PZ, UK*
[4] *Scottish Parasite Diagnostic and Reference Laboratory, Glasgow Royal Infirmary, 10-16 Alexandra Parade, Glasgow G31 2ER, UK*

(*Received 23 October 2015; revised 23 November 2015; accepted 24 November 2015; first published online 2 February 2016*)

SUMMARY

Cryptosporidium parvum is the major cause of livestock and zoonotically-acquired human cryptosporidiosis. The ability to track sources of contamination and routes of transmission by further differentiation of isolates would assist risk assessment and outbreak investigations. Multiple-locus variable-number of tandem-repeats (VNTR) analysis provides a means for rapid characterization by fragment sizing and estimation of copy numbers, but structured, harmonized development has been lacking for *Cryptosporidium* spp. To investigate potential for application in *C. parvum* surveillance and outbreak investigations, we studied nine commonly used VNTR loci (MSA, MSD, MSF, MM5, MM18, MM19, MS9-Mallon, GP60 and TP14) for chromosome distribution, repeat unit length and heterogeneity, and flanking region proximity and conservation. To investigate performance *in vitro*, we compared these loci in 14 *C. parvum* samples by capillary electrophoresis in three laboratories. We found that many loci did not contain simple repeat units but were more complex, hindering calculations of repeat unit copy number for standardized reporting nomenclature. However, sequenced reference DNA enabled reproducible fragment sizing and inter-laboratory allele assignation based on size normalized to that of the sequenced fragments by both single round and nested polymerase chain reactions. Additional *Cryptosporidium* loci need to be identified and validated for robust inter-laboratory surveillance and outbreak investigations.

Key words: *Cryptosporidium parvum*, genotyping, multi-locus, variable-number tandem-repeat, reproducibility, validation, outbreak, surveillance.

INTRODUCTION

Cryptosporidiosis is a gastro-intestinal disease caused by the protozoan *Cryptosporidium*, typically presenting in humans as diarrhoea, abdominal pain, nausea, vomiting and low grade fever (Farthing, 2000). Clinical cases in livestock are mainly in neonates, but older animals can also be significant shedders of oocysts (Pritchard *et al*. 2007; Wells *et al*. 2015). Diagnostic tests identify the genus, with species identification undertaken in specialist and reference or research laboratories (Chalmers and Katzer, 2013). *Cryptosporidium parvum* is one of the major causes of zoonotically-acquired human cryptosporidiosis, and in the UK *C. parvum* accounts for nearly half of all investigated cases of human cryptosporidiosis with an estimated 25% of non-travel-related, sporadic *C. parvum* cases acquired from direct contact with farm animals (Chalmers *et al*. 2011). Other routes of this faecal-oral infection include person-to-person spread, or via a vehicle such as drinking or recreational water, food and fomites (Casemore, 1990). To properly establish the burden of illness from potential exposures and to implement appropriate interventions, the ability to identify sources of contamination and routes of transmission by further differentiation of *C. parvum* isolates is desirable. However, there is currently no standardized genotyping scheme. Sequencing a hyper-variable region of the gene encoding a 60 kDa glycoprotein (GP60) is commonly used, including testing samples from patients and animals during zoonotic outbreak investigations (Chalmers and Giles, 2010). GP60 family IIa is commonly found in cattle and in human cases and outbreaks involving animal contact (Brook *et al*. 2009; Chalmers and Giles, 2010; Chalmers *et al*. 2010; Robertson *et al*. 2014).

* Corresponding author: Cryptosporidium Reference Unit, Public Health Wales Microbiology, Singleton Hospital, Swansea SA2 8QA, UK. Tel. +44 1792 285341. Fax +44 1792 202320. E-mail: rachel.chalmers@wales.nhs.uk
† These authors contributed equally to this paper.

Parasitology (2017), **144**, 37–47.
doi:10.1017/S0031182015001766

Subtype family IId is also commonly found in sheep and goats (Robertson *et al*. 2014) and has been found in human cases in outbreaks linked to open farms and a swimming pool (Cryptosporidium Reference Unit unpublished data). However, multi-locus analyses are more discriminatory (Feng *et al*. 2013), and multi-locus sequence typing (MLST) provides definitive detection of polymorphisms and has been used especially with loci containing variable-number of tandem-repeat (VNTR) units (Gatei *et al*. 2006; Xiao and Ryan, 2008; Widmer and Cacciò, 2015). However, MLST is expensive and time consuming. During outbreak investigations, rapid characterization of multiple isolates may be required to supplement epidemiological and environmental investigations, and for surveillance large numbers may need to be analysed. Multiple-locus VNTR analysis (MLVA) by slab gel or capillary electrophoretic (CE) sizing of amplified DNA fragments may provide a tool to enable initial characterization of outbreak isolates and linkage of cases with each other or suspected sources of contamination or infection. In one comparative study, fragment sizing *C. parvum* loci by CE provided better typability, discriminatory power, ease of use, and was more straightforward than sequencing repeat regions (Díaz *et al*. 2012). Additionally, the presence of multiple genotypes in a sample is likely to be identified more readily than by Sanger sequencing. Although one study has provided direct statistical comparison of fragment sizing and sequencing of four loci and showed that both laboratory methods and data analyses influenced the inferences on the population structure of *C. parvum* (Widmer and Cacciò, 2015), the choice of loci and their underlying characteristics will undoubtedly affect the outcome of such analyses.

Examples of the utility of MLVA of *C. parvum* have been documented previously but few investigations have used the same sets of loci, primers, analytical platforms, or allele nomenclature, hindering both comparison of allelic profiles and performance (Robinson and Chalmers, 2012). One meta-analysis of three sets of data generated using different analytical platforms used the assumption that fragment sizes generated were comparable across platforms (Caccio *et al*. 2015). If MLVA is to be applied as a rapid tool to support outbreak investigations and have meaningful application across both human and animal health surveillance internationally, then there needs to be structured development to enable harmonized application in different laboratories using different analytical platforms and running conditions, accounting for the potential influence of sequence composition and DNA conformation (Pasqualotto *et al*. 2007). Nadon *et al*. (2013) have identified, through consensus agreement, processes for the development of MLVA for bacterial surveillance and outbreak investigations, which should also be applicable to polyclonal samples such as *Cryptosporidium* spp. oocysts. These steps include: selection and naming of loci, assay design and validation, the need for calibration sets of samples, and standardized allele nomenclature (Nadon *et al*. 2013). Specifically pertaining to the selection of loci, Nadon acknowledged that, while there is an inverse relationship between repeat unit length and detected variation, repeat units <5 bp may be hard to differentiate in capillary electrophoresis. However, 3 bp differences have been reported to be differentiated using platforms such as ABI 3730 (Life Technologies) (Hotchkiss *et al*. 2015) and the QIAxcel (Qiagen) (Drumo *et al*. 2012; Caccio *et al*. 2015). Additionally, it was advised that insertions and deletions should be absent in repeat units, that only those loci with 100% conserved flanking sequences should be used, and that primers should be placed as close as possible to the repeat unit (Nadon *et al*. 2013).

To investigate the suitability of selected loci for the potential application of MLVA to *C. parvum* surveillance and outbreak investigations, we undertook *in silico* and *in vitro* studies. Since human *C. parvum* outbreak investigations frequently involve animal sampling, this included inter-laboratory sample exchange between laboratories involved in both human and animal health investigations.

MATERIALS AND METHOD

Loci and their attributes

Cryptosporidium parvum VNTR loci containing repeat units >2 bp, identified previously as being the potentially most useful (Robinson and Chalmers, 2012) or used in previous studies (Caccio *et al*. 2015; Hotchkiss *et al*. 2015), were selected: MSA, MSD, MSF, MM18, MM19, MS9-Mallon (hereafter referred to as MS9), GP60 and TP14.

To evaluate whether the loci met the standards for inter-laboratory surveillance and outbreak investigation proposed by Nadon *et al*. 2013, sequences were selected to represent a broad range of alleles and aligned using BioEdit 7·0·9 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). These sequences were selected from our own archives and the National Center for Biotechnology Information's GenBank database (MM5: KP172504, KP172505, KP265906-KP265911; MM18: KP172508; MM19: KP172512-KP172515, KP265912, KP265914-KP265926; GP60: AB242224-AB242227, AB242229, AF403166-AF403168, AY149610, AY149612, AY149614-AY149616, AY382675, AY738185-AY738186, AY738188-AY738189, AY738191, AY738193-AY738195, AY873780-AY873782, DQ192502, DQ192508, DQ630514-DQ630516, DQ630519, DQ648531-DQ648537, DQ648541, DQ648544, EU140508, EU164810-EU164811; TP14: KM222505-KM222508). Individual sequences were checked for completeness

https://doi.org/10.1017/S0031182015001766 Published online by Cambridge University Press

(for the purpose of this study the primer sequences shown in Table 1 were retained) and quality (no ambiguous bases or suspected anomalies). The true fragment size of each allele was identified and the following attributes tabulated and assessed for suitability (Nadon *et al*. 2013): chromosome location, repeat unit length, repeat unit heterogeneity of DNA and amino acid sequences, flanking region conservation and proximity to repeat unit.

Reproducibility of MLVA

To investigate the impact of the attributes of the loci and to pilot test the reproducibility of MLVA, providing a proof of concept for future inter-laboratory investigations, the nine loci were used *in vitro* in our three laboratories. These have remits either for investigation of human cryptosporidiosis and suspected animal sources (Cryptosporidium Reference Unit, CRU and Scottish Parasite Diagnostic and Reference Laboratory, SPDRL) or livestock cryptosporidiosis (Moredun Research Institute, MRI). A set of 14 DNA samples, extracted from the national collection of *Cryptosporidium* oocysts at the CRU as described previously (Chalmers *et al*. 2009, 2011), was confirmed as containing *C. parvum* DNA by real-time polymerase chain reaction (PCR) of the Lib13 gene (Hadfield *et al*. 2011) and GP60 subtypes were identified by sequencing (Alves *et al*. 2003; Sulaiman *et al*. 2005). Isolates were selected to represent a range of GP60 subtypes. DNA was distributed by post. In house PCRs were used to amplify fragments corresponding to the variable regions of each locus as described below. The primer sets are described in Table 1. DNA from isolates representing a range of sequenced reference alleles was included in each PCR and sizing reaction.

At the CRU, all nine loci were investigated with previously validated single round PCRs (CRU unpublished data) using 1 μL template, except MM19 using 5 μL, in final reaction volumes of 20 μL containing 2·5 mM $MgCl_2$, 200 μM dNTPs, 500 μg mL^{-1} non-acetylated bovine serum albumin and 1 unit of Hotstar DNA Taq polymerase in 1× PCR buffer. Primer concentrations were 500 nM for MSA, MSD, MSF, MS9 and MM5, 300 nM for MM18, TP14 and GP60, and 200 nM for MM19. An addition of 2 μL Q solution was included for MM18, TP14 and GP60. Standard PCR cycling conditions were 40 cycles of denaturation at 95 °C for 30 s, annealing at 55 °C except MM18 at 63 °C and MM19 at 61 °C for 30 s and extension at 72 °C for 60 s followed by a final extension at 72 °C for 10 min. Fragment sizing of PCR products, diluted 1 in 10 in QX dilution buffer, was by capillary electrophoresis in a temperature-controlled room (25 °C using a QIAxcel on programme OH700 with a 15 bp/600 bp QX DNA Alignment Marker and a 25–500 bp QX Size Marker (Qiagen, Crawley, UK).

At the MRI all nine loci, and at the SPDRL eight loci (GP60 was not used), were investigated with validated nested PCRs using 1 μL DNA or primary product diluted 1:100 as template in final reaction volumes of 20 μL as described previously (Hotchkiss *et al*. 2015). Standard PCR cycling conditions were 30 cycles of 95 °C for 50 s, 50 °C for 50 s and 65 °C for 60 s. Fragment sizing of FAM-labelled (Eurofins Genomics, UK) PCR products was undertaken using capillary electrophoresis on two different analytical platforms: MRI used the ABI 3730 (Applied Biosystems; University of Dundee) with the Genescan ROX500 size standard (Applied Biosystems), and SPDRL used the ABI 3500XL with the GeneScan 600 LIZ size standard. Trace files were analysed at the MRI using STRand (http://www.vgl.ucdavis.edu/informatics/strand) and at the SPDRL using GeneMapper Software 5 (Applied Biosystems).

In all three laboratories, the peak sizes were compared and matched with those of the sequenced reference amplicons to enable an adjusted fragment size to be recorded, representing the true fragment size of the sequenced reference standard. Any samples that could not be aligned to a reference standard were sequenced to confirm the presence of a new allele. Sequences generated and/or newly used in this study were deposited in GenBank under accession numbers KT922174 to KT922224.

Reproducibility of allele assignment based on fragment sizing

Alleles were compared between laboratories and primer sets in two ways: first, using the adjusted fragment sizes, but this did not permit ready comparison where different primers were used for four of the nine loci: MM19, MS9, TP14 and GP60 (Table 1); second, the adjusted fragment sizes were normalized by deducting from the larger products the difference between the larger and shorter sequenced products, as this was found to be consistent for the reference alleles.

Standardized allele nomenclature

To determine if a standardized allele nomenclature could be generated that would circumvent the need for standardized primer sets, the copy number of repeats was calculated from the adjusted fragment size minus the off set size divided by the repeat size. For complex loci with more than one repeat region it was assumed that the fragment was generated by the same combination of repeat unit copy numbers as the reference sequence for that allele. Thus, for the first repeat one to nine copies were designated 01 to 09, and 10 or more copies by the two digit integer and likewise for the second repeat, so that an allele containing two copies of

https://doi.org/10.1017/S0031182015001766 Published online by Cambridge University Press

Table 1. Polymerase chain reaction primers used to amplify variable-number of tandem-repeat loci in *Cryptosporidium parvum*

Locus	Laboratory	Forward primers (F)	Reverse primers (R)	Distance between primer and repeat unit F : R (bp)
MSA	MRI & SPDRL 1°	TGCACTGTATTTCAACCCCA	GAACTAGGCTCGGGTTCAGA	31 : 15
	MRI & SPDRL 2°; CRU	TAGGCTCGGGTTCAGA	GACTGTCACAAAAGTTAATCC	
MSD	MRI & SPDRL 1°	AGGTCAGTAAGCTCAGTTCGT	CAGTCTTGAATTTGGTTTCTGCA	6 : 68
	MRI & SPDRL 2°; CRU	CATCTCAAGAAATTCAGTCTC	CTCCTTTTGCTCCAGC	
MSF	MRI & SPDRL 1°	TCTTATCTTTGTCCTTCTTGCCC	AGCTGAGAAGGGCAAGAAGA	60 : 1
	MRI & SPDRL 2°; CRU	AGAAGAAAGCCAAGAAGGGT	TCGGCCTCCTCTACAG	
MM5	MRI & SPDRL 1°	TCACAAGTTACCCCTTCTGATGCTG	TCCACCTCCGGATTGGTTGTG	97 : 32
	MRI & SPDRL 2°; CRU	CCTGGACTTGGATTTGGACTTACACC	GGAGAAGATAAGCTAGCCGAATCT	
MM18	MRI & SPDRL 1°	GTTCAGCTGATACGGGTTTGCAACA	CATCACCATCTCCTCCGCCAGA	85 : 51
	MRI & SPDRL 2°; CRU	CTTTCTGGAGGGTTTGTTCCTCC	CTTCCTGATGATCCAGGCCAAGC	
MM19	CRU	GATTCTGCAACTTTGAATTCAGTA	CCAACCCCGAATTCATTTCCAAC	0 : 150
	MRI & SPDRL 1°	TGGTTTTAGCTAAGGAAGCGATAG	CTGCTGCTGCTGTTGCTTTA	
	MRI & SPDRL 2°	GATTCTGTCAACTTTGAATTCAG	CCAACCCCGAATTCATTTCCAAC	
TP14	CRU	GTTCACAGCCAACAGT	CATTTTGATTTTGGGAGT	44 : 31
	MRI & SPDRL 1°	GAGAAGGAGCAATGGGAGCA	TCCTCCTTTTTGCCCTTGAA	41 : 107
	MRI & SPDRL 2°	CTAACGTTCACAGCCAACAGTACC	CAATAAAGACCATTATTACCACC	
MS9	CRU	ACCTGGAGTGTGATTTGG	GTTCTTGTTCAAAGTCA	−6[a] : 1
	MRI & SPDRL 1°	TTAGTCGACCTCTTCAACAGTTGG	CAGAAT TGGAATCATTTTCTGAAT	48 : 199
	MRI & SPDRL 2°	GGACTAGAAATAGAGCTTTGGCTGG	GTCTGAGACAGAATCTAGGATCTAC	
GP60	CRU[b]	GCCGTTCCACTCAGAGGAAC	CCACATTACAAATGAAGTGCCGC	24–25 : 199–319
		GCCGTTCCACTCAGAGGCA	CCACATTATAAATGAAGTACCACATTC	
			CCACATAACAAATGAAGTACCGCA	
	MRI 1°	ATAGTCTCCGCTGTATTC	GAGATATATCTTGGTGCG	25 : 265–391
	MRI 2°	GCCGTTCCACTCAGAGGAAC	GGTGCATAGACGATAGTGTA	

[a] Forward primer overlaps first repeat.
[b] A primer cocktail (equal concentrations) was used to allow for polymorphisms in *C. parvum* primer sites.

the first repeat and three of the second repeat would be named 0203.

Sensitivity

The number of alleles identified using single round PCRs was compared with those assigned using nested PCR.

RESULTS

Loci and their attributes

Comparison of the attributes of MLVA loci revealed variable performance for the nine *C. parvum* loci (Table 2). The loci were not distributed across all eight *C. parvum* chromosomes; one was on each of chromosomes one and three, there were two loci on each of chromosomes five and six, and three were on chromosome eight (Table 2).

DNA sequence analysis and alignment identified that all of the loci were within open reading frames and the repeat units encoded various amino acid residues (Table 2). Translation to the amino acid sequences and their subsequent alignment simplified identification of the true start and end points of the repeat units, and revealed that additional repeat units were present in six loci: consistently in MSA, MSD, MM9 and TP14 and more rarely in MM18 and GP60, the latter being well documented in GP60 family IIa (Table 2 and examples in Fig. 1). Heterogeneity of the DNA sequences within the repeat units was identified commonly, sometimes affecting the amino acids (MM18, MM19, first region in MSA, second region in MSD, first region in TP14) and sometimes not (GP60, MM5, two regions within MS9) (Table 2). Furthermore, insertions were found interspersed between copies of the repeat in MM18, interrupting the tandem nature of the repeats and changing the fragment size non-uniformly (Fig. 1). Only MSF contained a single repeat region with a homogenous repeat unit (Fig. 1).

The primer sets used varied in their proximity to the repeat unit (Table 1), but most generated amplicons <400 bp with the exception of the MRI/SPDRL primers for MS9 and the largest MM19 and GP60 alleles (Table 3). The regions flanking the repeat units were generally well conserved, with the major exception of GP60 (Table 2). In GP60, the region downstream of the repeat unit is highly polymorphic and allows for differentiation of isolates of the same species into allelic families based on sequence data (Strong *et al*, 2000). For example, the downstream regions of families IIa and IId, are only 70% similar. In addition, at MM19 rare insertions were identified downstream of the repeat unit in two sequences found on GenBank: KP265923 which has a 6 bp [AG] insert and KP265925 which has a 36 bp insert [TGAGIEAGVGIG].

Reproducibility of allele assignment based on fragment sizing

Although this pilot study was too small for robust analysis of the relationship between real and measured fragment sizes, one observed trend was that the measured fragments at the MRI were more often larger than the sequenced size, and those from the SPDRL and CRU were more often smaller. Additionally, the size difference appeared to be more consistent at those loci with a generally lower GC content (MSD, MS9, MM5, GP60 and TP14), whereas for MM19 and MSF size differences tended to increase with fragment size and for MM18 and MSA there was no discernable relationship (data not shown). However, for most loci assigning the correct allele was straightforward although for loci with short repeat units (3 bp in MM5, GP60 and TP14), the concentration of the PCR amplicon could affect the ability to align the test samples to the sequenced standards, especially on the QIAxcel. For alleles to be correctly assigned, it was essential that sequenced reference standards were included in the PCR and analysis.

The use of normalized fragment sizes permitted naming regardless of whether the same or different primer sets were used (Table 3). Allele assignation by the three laboratories was concordant with the exception of MS9 where interpretable results were not obtained from one laboratory (Table 4).

The primary purpose of investigating this set of 14 samples was to investigate whether the attributes identified *in silico* affected the reproducibility of allele assignation, but we also found that samples with the same GP60 sequenced allele were readily differentiated by the combination of loci investigated. The three GP60 IIdA17G1 samples differed from each other at three, six and five other loci, and the three IIdA18G1 samples differed at six, five, and three other loci (Table 4). Of the three GP60 family IIa samples, IIaA16G2R1 and IIaA17G1R1 could not be differentiated by 8 of the 9 loci and no amplicons were generated using MM18 for the IIaA17G1R1 sample. The IIaA16G3R1 sample could be differentiated using MM5, MM18, MM19 and TP14. In GP60 family IId, only TP14 was mono-allelic, with multiple alleles identified for the other loci (Table 4).

Standardized allele nomenclature based on copy number of repeats

The calculation of the copy number of repeats was readily applied to the adjusted fragment sizes of MSF, MM5 and MM19 which are simple loci containing single repeat units (Tables 2 and 3). However, application of this nomenclature in the complex loci MSA, MSD, MS9, MM18, TP14 and GP60 with multiple repeat units (Tables 2

https://doi.org/10.1017/S0031182015001766 Published online by Cambridge University Press

Table 2. Attributes of *Cryptosporidium parvum* loci used in this study

Locus	Chr.	Repeat unit	Nucleotide sequence(s)	Length	Amino acid sequence	Conservation of the sequences flanking the repeat unit
MSA	1	1	TT(G/A)(G/A)GCTCA(T/G)(G/T)CTCA(A/G)GTTTAGG (C/T)TCATTCTCAGGT	36 bp	L(G/S)S(C/F/G)S(S/G) LGSFSG	Mainly yes
		2	TCAGGC	6 bp	SG	
MSD	3	1	ACCCAAATCAAGAGGATCCCAGTCAAGAAG	30 bp	TQIKRIPVKK	Yes
		2	(A/C/T)(C/T)CAATCCCAAGAAA	15 bp	(L/T/S)NPKK	
MSF[a]	5		GCTCAGGAAGGA	12 bp	AQEG	Yes
MM5	6		TC(T/C/A)	3 bp	S	Mainly yes
MM18	8		GAAGCAGGA(G/C)CAGGACCAGGACCA (and occasionally an additional GGACCA)	24 bp (rarely 30 bp)	EAG(P/A)GPGP and occasionally an additional GP at end	Mainly yes, although the repeat unit is interspersed with heterologous sequences
MM19	8		GGA(G/T)C(A/T)	6 bp	G(A/S)	Mainly yes
TP14	8	1	CA(A/G/T)	3 bp	Q/H	Mainly yes
		2	CAACACAAT	9 bp	QHN	
MS9	5	1	AT(T/C)TGG	6 bp	IW	Yes
		2	ATCTGGACTTGGAGTATG	18 bp	IWTWSM	
		3	TGGATC	6 bp	WI	
		4	TTGATCTTGATCTGGACCTGGATCTGGAT(C/T)	30 bp	LILIWTWIWI with an LI motif between	
GP60	6	1	TC(A/G/T)	3 bp	S	No. The region downstream of the repeat unit is polymorphic and determines the allelic family.
		2	the R repeat ACATCA in family IIa	6 bp	TS	

[a] The nucleotide sequence for MSF was originally published in reverse orientation (Tanriverdi *et al*. 2006).

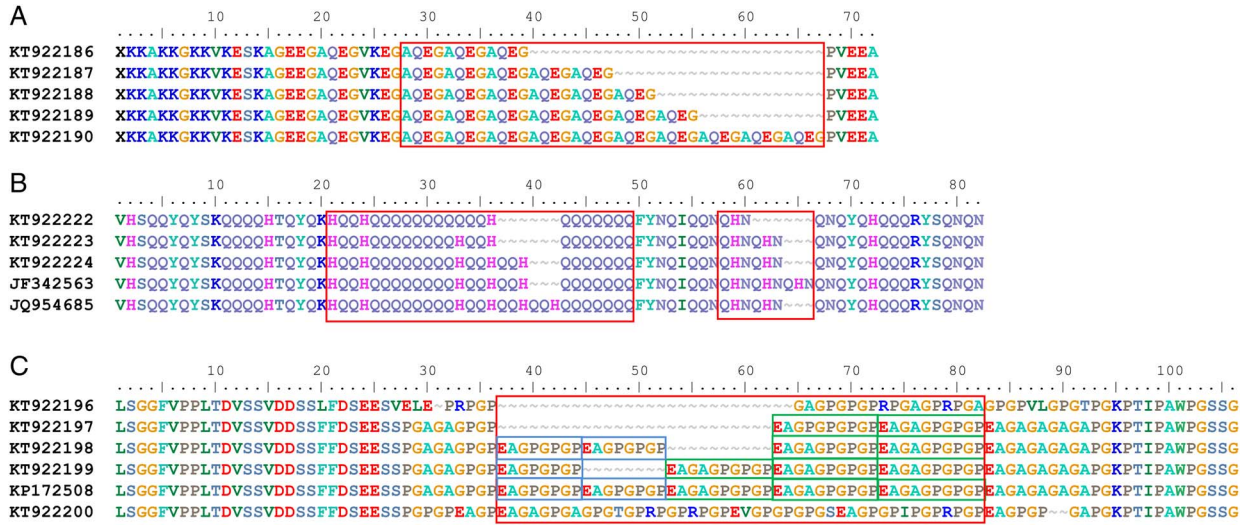

Fig. 1. Examples of amino acid sequence alignments of three variable-number of tandem-repeats (VNTR) loci, MSF, TP14 and MM18, in *Cryptosporidium parvum*. Repeat regions are shown within the coloured boxes. (A) MSF contains a single, homologous repeat unit encoding AQEG so each allele has a fragment size that differs by a multiple of 12 bp. (B) TP14 contains two repeat units, encoding Q/H and QHN. The first four alleles are differentiated by variable numbers of both repeat units generating different fragment sizes. The last two alleles have the same fragments size but different number of repeats in each region. (C) MM18 contains a single repeat unit that has 8 (blue boxes) and/or 10 (green boxes) amino acid motifs. There are also rare alleles (KT922196 and KT922200) that appear to have additional amino acids within the repeat region disrupting its tandem nature.

and 3), was based on the assumption that the copy numbers of the different repeat units in the samples was the same as those in the sequenced reference alleles, which we consider misleading.

Sensitivity

Single round PCRs enabled full allelic profiles to be generated for 12 of the 14 samples, and only 5 alleles overall were not assigned, three in one sample and two in another (Table 4). However, one of these samples was not fully profiled by nested PCR either. Overall, nested PCRs provided only four more data points in the entire sample set compared with single round PCR (Table 4). Laboratory workflow was simplified by single round PCR.

DISCUSSION

We have investigated nine of the ten top ranking *C. parvum* loci identified on the basis of prior MLVA performance for variability (Robinson and Chalmers, 2012), that have been used in previous studies (Caccio *et al*. 2015; Hotchkiss *et al*. 2015), by assessing their attributes *in silico* in terms of proposed guidelines (Nadon *et al*. 2013) and *in vitro* through sample exchange. *In silico* analyses revealed that not all these loci met the proposed guideline criteria and may not be ideal MLVA choices for inter-laboratory surveillance and outbreak investigations. However, despite some of the apparent shortcomings, the *in vitro* study demonstrated that reproducible allele assignation was possible for all these loci in a meaningful way. This was achieved through the use of sequenced reference standards and normalization of fragment sizes, requiring inter-laboratory communication to define a baseline allowing for the use of different PCR protocols. Nested PCRs yielded only very slightly more information than single round PCRs; the latter provides greatly improved workflow in emergency response.

The five attributes used to assess the VNTR loci were: chromosome location; repeat units ⩾5 base pairs; no insertions and deletions in the repeat units; perfect homogenous repeats should be preferred; and only loci with 100% conserved flanking sequences should be used (Nadon *et al*. 2013). First, the loci were found not to be distributed across all eight chromosomes; when selecting MLVA loci for epidemiological investigations, distribution across chromosomes is desirable as it ensures they are sufficiently distant to exclude physical linkage (Widmer and Sullivan, 2012). However, if more than eight markers are needed for high-resolution genotyping some clustering would be inevitable. The inclusion of linked loci can be valuable in population genetics, for example in studies of linkage disequilibrium. Secondly, seven of the nine loci contained repeat units that were longer than 5 bp. Although the capillary electrophoresis platforms used in this study were capable of differentiating 3 bp, which concurs with previous studies (Drumo *et al*. 2012; Caccio *et al*. 2015; Hotchkiss *et al*. 2015), this was through judicious use of sequenced reference standards representing a range of alleles and maintaining optimal running conditions

https://doi.org/10.1017/S0031182015001766 Published online by Cambridge University Press

Table 3. Allele nomenclature for *Cryptosporidium parvum* variable-number of tandem-repeats derived from adjusting fragment sizes to those of reference sequences, and normalizing alleles to shorter fragments

Locus	Adjusted fragment size where alternative primers were used	Size adjustment for normalization	Normalized allele	Copy number of repeats
MSA	See normalized allele	Not applied	197	0205[a]
			227	0304[a]
			233	0305[a]
			239	0306[a]
			245	0307[a]
			257	0309[a]
MSD	See normalized allele	Not applied	246	0104[a]
			261	0203[a]
			276	0204[a]
MSF	See normalized allele	Not applied	133	3[b]
			157	5[b]
			169	6[b]
			181	7[b]
			217	10[b]
MM5	See normalized allele	Not applied	233	18[b]
			248	23[b]
			260	27[b]
MM18	See normalized allele	Not applied	212	May not be applicable because of the different sized repeats within the same region and the occasional sequences that do not demonstrate tandem repeats and are interspersed with random bases
			242	
			290	
			296	
			314	
MM19	222	−1	221	4[b]
	288		287	15[b]
	294		293	16[b]
	324		323	21[b]
	330		329	22[b]
	336		335	23[b]
	366		365	28[b]
	390		389	32[b]
	408		407	35[b]
MS9	432	−267	165	06020201[a]
	438		171	06020500[a]
	444		177	08020201[a]
	450		183	04020701[a]
	462		195	06010502[a]
GP60	377	−62	315	Cannot be applied as fragment size is also influenced by variation in the downstream sequence (Sulaiman *et al*. 2005)
	383		321	
	386		324	
	398		336	
	401		339	
	404		342	
	407		345	
TP14	301	−81	220	2302[a]
	310		229	2602[a]

[a] Observed distribution in the sequenced reference standard.
[b] Calculated from the fragment size minus the offset size divided by the repeat size.

especially for the QIAxcel (CRU unpublished data). The practicalities of assigning 3 bp alleles was more challenging than for longer repeats, and the precision of analysis of MM5 has been reported previously to be impaired (Hotchkiss *et al*. 2015). For a robust, standardized scheme ⩾5 bp would be more desirable.

The nine loci were all within open reading frames and all the repeat units coded for amino acids; identifying some repeat units from DNA sequences was open to interpretation, but was clarified by analysis of the amino acid sequences. Sequence variation was identified within the repeat units of eight of the nine loci, the only exception being MSF. This variation has not been reported previously for MSA, MSD, MS9 and TP14 and contrasts with the simple sequence repeats reported previously (summarized by Robinson and Chalmers, 2012). The variation seen in the amino acid sequences of the repeat units in MSA, MSD, TP14, MM18 and MM19 may have a biological effect.

Multiple repeat units were identified in six loci and although recognized previously in GP60

https://doi.org/10.1017/S0031182015001766 Published online by Cambridge University Press

Table 4. Final allelic profiles based on normalized fragment sizes (consensus agreement across all three laboratories unless otherwise stated; MS9 and GP60 were analysed at CRU and MRI only)

		Locus								
ID	GP60 subtype	MSA	MSD	MSF	MM5	MM18	MM19	MS9	GP60	TP14
UKP40	IIaA16G2R1	227	276	157	260	296	293	183	336	220
UKP41	IIaA16G3R1	227	276	157	233	290	287	183	339	229
UKP43	IIaA17G1R1	227	276	157	260[a]	DAMP	293[b]	183	336[c]	220[d]
UKP32	IIdA15G1	233	246	157	248	242	407	165	315	229
UKP45	IIdA17G1	233	276	169	233	242	323	165	321	229
UKP42	IIdA17G1	233	261	169	233	314	365	165	321	229
UKP46	IIdA17G1	239	246	217	233	242	335	171	321	229
UKP47	IIdA18G1	197	246	169	248	242	389	195	324	229
UKP48	IIdA18G1	239	246	181	233	212	365	165	324	229
UKP49	IIdA18G1	197	246	181	248	242	323	177	324	229
UKP50	IIdA22G1	257	246	133	233	242	221	177	336	229
UKP51	IIdA23G1	239	261	181	233	242	329	165	339	229[a]
UKP52	IIdA24G1	245	246	133	233	242	221	165	342	229
UKP53	IIdA25G1	245	246	133	233	242	221	165	345	229

DAMP – did not amplify.
[a] MRI and SPDRL only, CRU DAMP.
[b] SPDRL and CRU only, MRI DAMP.
[c] MRI only, CRU DAMP.
[d] CRU and MRI only, SPDRL DAMP.

(Alves *et al*. 2003; Sulaiman *et al*. 2005) this was identified for the first time in MSA, MSD, MS9, MM18 and TP14. The presence of multiple repeat units did not prevent allele assignation based on adjusted fragment sizes, although the size difference between alleles was not as predictable as for homogenous units. A standardized allele nomenclature based on calculation of the actual copy number of repeats that would also allow for the use of alternative primers (Larsson *et al*. 2009) meant that assumptions were made about the distribution of the copy numbers within those loci that were more complex than originally thought. The practice of allocating the same copy number pattern for the different repeat units as that found in the sequenced reference allele (Nadon *et al*. 2013) would lead to under-reporting of variation in the complex loci, biased by the selection of the reference sequence. For example, we identified that TP14 had two repeat units, the length of the first being 3 bp and the second 9 bp (Table 2; Fig. 1). The two alleles in this study were newly identified and therefore sequence data identified their configuration 2302 and 2602; however, had we found a 238 bp fragment this could have been assigned to reference sequence JF342563 which is configured with 2603, but another sequence, JQ954685, also has the same sized fragment but was configured 2902. We consider that the assumption is not helpful, and this strategy should not be pursued; the issue could be avoided altogether if only simple VNTR loci are used. However, these seem to be in the minority of those currently identified and further work is needed to identify more suitable loci.

The proximity of the (internal) primers to the repeat region partially determined the overall size of the amplicons, which determines the size markers to use and has been shown to affect the performance of the CE machine (Hotchkiss *et al*. 2015). The resolution of the QIAxcel is optimal for fragments <300 bp especially with shorter repeat units (Qiagen). Thus the primers need to be designed taking this into account. Finally, most of the flanking regions were either homogenous or generally conserved, but where they were not, such as in GP60, heterogeneity may pose two problems: the fragment size could be affected not only by the VNTRs but also by variation in the flanking sequence, and some of the primer sites also included polymorphisms that requires a primer cocktail to improve the sensitivity by allowing amplification of a range of variants. This heterogeneity is acknowledged by, and forms a critical part of, GP60 sequence nomenclature (Sulaiman *et al*. 2005) but may affect fragment sizing.

Only MSF met all of the criteria and was the only true simple tandem repeat, providing a good example for identification of future loci. The attributes of the nine loci may go some way to clarify the arguments that have been raised against the use of fragment sizing for genotyping *Cryptosporidium* isolates. In one study, fragment sizing was compared with sequencing amplicons of MM5, MM19, MS9 and GP60 and showed that single locus distance matrices were weakly correlated, but that this correlation was not maintained when the data were combined in multi-locus genotypes (Widmer and Cacciò, 2015). The authors argued that the simplicity of genotyping using amplicon length data is potentially offset by its

https://doi.org/10.1017/S0031182015001766 Published online by Cambridge University Press

limited resolution (Widmer and Cacciò, 2015). However, we propose that the attributes of the loci investigated are critical to this and the comparison needs to be explored further using loci that are better suited to MLVA since the repeat units of MM5, MM19, MS9 and GP60 are all polymorphic and we have demonstrated that MS9 contains four repeat units (Table 2). We agree that the development and adherence to a set of guidelines for locus identification and standardization of genotyping analyses by any method is important.

The increasing availability of *C. parvum* whole genome sequences (Andersson *et al*. 2015; Hadfield *et al*. 2015) provides the means to identify new, appropriate loci for a robust MLVA scheme, and this work is underway. In addition, genome sequence data have contributed to our understanding of these loci, for example MSF was originally published in reverse orientation (Tanriverdi *et al*. 2006). For many pathogens, especially culturable bacteria such as Shiga toxin-producing *Escherichia coli* O157, whole genome sequencing has superseded MLVA and other traditional typing methods (Dallman *et al*. 2015). However, for *Cryptosporidium* lengthy processing is required to generate suitable DNA from clinical samples (Hadfield *et al*. 2015) even when whole genome amplification is used (Andersson *et al*. 2015), and routine application for timely *Cryptosporidium* surveillance and outbreak investigations is currently a distant reality.

We undertook a preliminary assessment of the reproducibility of MLVA applied to 14 DNA samples selected to provide a range of GP60 alleles from families IIa and IId. Even in this small study, where some samples with the same GP60 sequences were compared, different allelic profiles were generated concurring with previous findings that single locus analysis underestimates diversity in *C. parvum* (Widmer and Sullivan, 2012). While the use of GP60 sequencing has been useful in characterizing the aetiology of zoonotic *C. parvum* outbreaks (Chalmers and Giles, 2010), a multilocus approach is needed to improve discrimination during outbreak investigations. Previously, in a study focussing on GP60 family IIa, MSA, MSD and MSF were monoallelic which is what we found here (Hotchkiss *et al*. 2015). However, multiple alleles were found at these three loci in family IId, demonstrating that consideration of the host and parasite population is important in marker selection.

Concluding remarks

Although most loci were not ideal for MLVA according to the proposed guideline standards, it was possible to use different capillary electrophoresis platforms and assign reproducible allelic profiles to a set of samples, by using previously sequenced, co-amplified reference standards. If a centrally curated database and archive of all identified alleles were maintained then cloned, reference material could be circulated to participating laboratories. In this way, laboratories could use bespoke protocols and primer sets without compromising allele assignment. MLVA assays for *Cryptosporidium* are still in the development phase and there is no consensus on the number of markers or which they should be. While resolution might be increased by using more markers, the necessity depends on the epidemiological question being asked. From this proof of principle study it is not possible to comment on how many or which markers are desirable or essential. There is a need to re-define loci and a set of rules for selection, application and analysis for inter-laboratory schemes, as well as nomenclature for locus and allele naming. This could be achieved through a consensus meeting and it is proposed that this is enabled by COST Action FA1408: A European Network for Foodborne Parasites (Euro-FBP; www.euro-fbp.eu). Loci for investigation of both *C. hominis* and *C. parvum* should be considered. Full validation studies, supported by calibration samples, are needed to compare MLVA analysis between different laboratories following guidelines for validation of typing schemes (Struelens, 1996; van Belkum *et al*. 2007; Nadon *et al*. 2013) and permitting analysis for typability, discriminatory power, reproducibility and epidemiological concordance. The cost of MLVA could be reduced by multiplexing loci with significantly different expected fragment sizes and different fluorescent labels. Finally, standardized nomenclature needs to be agreed, including consultation with end users including health professionals (Palm *et al*. 2012).

ACKNOWLEDGEMENTS

We are grateful to Frank Katzer, Moredun Research Institute, for helpful comments on the manuscript.

FINANCIAL SUPPORT

The research leading to these results has received funding from the European Union Seventh Framework Programme (RMC, GR and SM [FP7/2007-2013] [FP7/2007-2011] under Grant agreement no: 311846); the Scottish Government (EH and JG) under SPASE workstrand 3.2.3.

REFERENCES

Alves, M., Xiao, L., Sulaiman, I., Lal, A. A., Matos, O. and Antunes, F. (2003). Subgenotype analysis of *Cryptosporidium* isolates from humans, cattle, and zoo ruminants in Portugal. *Journal of Clinical Microbioogy* **41**, 2744–2747.

Andersson, S., Sikora, P., Karlberg, M. L., Winiecka-Krusnell, J., Alm, E., Beser, J. and Arrighi, R. B. (2015). It's a dirty job – A robust method for the purification and de novo genome assembly of *Cryptosporidium* from clinical material. *Journal of Microbiological Methods* **113**, 10–12.

https://doi.org/10.1017/S0031182015001766 Published online by Cambridge University Press

Brook, E. J., Hart, C. A., French, N. P. and Christley, R. M. (2009). Molecular epidemiology of *Cryptosporidium* subtypes in cattle in England. *The Veterinary Journal* **179**, 378–382.

Caccio, S. M., de Waele, V., Widmer, G. (2015). Geographical segregation of *Cryptosporidium parvum* multilocus genotypes in Europe. *Infection, Genetics and Evolution* **31**, 245–249.

Casemore, D. (1990). Epidemiological aspects of human cryptosporidiosis. *Epidemiology and Infection* **104**, 1–28.

Chalmers, R. M. and Giles, M. (2010). Zoonotic cryptosporidiosis. *Journal of Applied Microbiology* **109**, 1487–1497.

Chalmers, R. M. and Katzer, F. (2013). Looking for *Cryptosporidium*: the application of advances in detection and diagnosis. *Trends in Parasitology* **29**, 237–251.

Chalmers, R. M., Elwin, K., Thomas, A. L., Guy, E. C. and Mason, B. (2009). Long-term *Cryptosporidium* typing reveals the aetiology and species-specific epidemiology of human cryptosporidiosis in England and Wales, 2000 to 2003. *Eurosurveillance* **14**, 15.

Chalmers, R. M., Smith, R., Elwin, K., Clifton-Hadley, F. A., Giles, M. (2011). Epidemiology of anthroponotic and zoonotic human cryptosporidiosis in England and Wales, 2004 to 2006. *Epidemiology and Infection* **139**, 700–712.

Dallman, T. J., Byrne, L., Ashton, P. M., Cowley, L.A., Perry, N. T., Adak, G., Petrovska, L., Ellis, R. J., Elson, R., Underwood, A., Green, J., Hanage, W. P., Jenkins, C., Grant, K. and Wain, J. (2015). Whole genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clinical Infectious Diseases* **61**, 305–312.

Díaz, P., Hadfield, S. J., Quílez, J., Soilán, M., López, C., Panadero, R., Díez-Baños, P., Morrondo, P. and Chalmers, R. M. (2012). Assessment of three methods for multilocus fragment typing of *Cryptosporidium parvum* from domestic ruminants in northwest Spain. *Veterinary Parasitology* **186**, 188–195.

Drumo, R., Widmer, G., Morrison, L. J., Tait, A., Grelloni, V., D'Avino, N., Pozio, E. and Caccio, S. M. (2012). Evidence of host associated populations of *Cryptosporidium parvum* in Italy. *Applied and Environmental Microbiology* **78**, 3523–3529.

Farthing, M. J. G. (2000). Clinical aspects of human cryptosporidiosis. In *Cryptosporidiosis and Microsporidiosis*. (ed. Petry, F.), Contributions in Microbiology, vol. 6, pp. 50–74, Karger, Basel.

Feng, Y., Torres, E., Li, N., Wang, L., Bowman, D. and Xiao, L. (2013). Population genetic characterisation of dominant *Cryptosporidium parvum* subtype IIaA15G2R1. *Emerging Infectious Diseases* **16**, 895–896.

Gatei, W., Hart, C. A., Gilman, R. H., Das, P., Cama, V. and Xiao, L. (2006). Development of a multilocus sequence typing tool for Cryptosporidium hominis. *Journal of Eukaryotic Microbiology* **53** (Suppl. 1), S43–S48.

Hadfield, S. J., Robinson, G., Elwin, K. and Chalmers, R. M. (2011). Detection and differentiation of *Cryptosporidium* spp. in human clinical samples by use of real-time PCR. *Journal of Clinical Microbiology* **49**, 918–924.

Hadfield, S. J., Pachebat, J. A., Swain, M. T., Robinson, G., Cameron, S. J., Alexander, J., Hegarty, M. J., Elwin, K. and Chalmers, R. M. (2015). Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples. *BMC Genomics* **16**, 650.

Hotchkiss, E. J., Gilray, J. A., Brennan, M. L., Christley, R. M., Morrison, L. J., Jonsson, N. N., Innes, E. A. and Katzer, F. (2015). Development of a framework for genotyping bovine-derived *Cryptosporidium parvum*, using a multilocus fragment typing tool. *Parasites and Vectors* **8**, 500.

Larsson, J. T., Torpdahl, M., Petersen, R. F., Sørensen, G., Lindstedt, B. A., Nielsen, E. M. (2009). Development of a new nomenclature for Salmonella Typhimurium multilocus variable number of tandem repeats analysis (MLVA). *Euro Surveillance* **14**. pii=19174.

Nadon, C. A., Trees, E., Ng, L. K., Møller Nielsen, E., Reimer, A., Maxwell, N., Kubota, K. A. and Gerner-Smidt, P., the MLVA Harmonization Working Group (2013). Development and application of MLVA methods as a tool for inter-laboratory surveillance. *Euro Surveillance* **18**. pii=20565.

Palm, D., Johansson, K., Ozin, A., Friedrich, A. W., Grundmann, H., Larsson, J. T. and Struelens, M. J. (2012). Molecular epidemiology of human pathogens: how to translate breakthroughs into public health practice, Stockholm, November 2011. *Euro Surveillance* **17**.

Pasqualotto, A. C., Denningm, D. W., Anderson, M. J. (2007). A cautionary tale: lack of consistency in allele sizes between two laboratories for a published multilocus microsatellite typing system. *Journal of Clinical Microbiology* **45**, 522–528.

Pritchard, G. C., Marshall, J. A., Giles, M., Chalmers, R. M. and Marshall, R. M. (2007). *Cryptosporidium parvum* infection in orphan lambs on a farm open to the public. *Veterinary Record* **161**, 11–14.

Robertson, L., Björkman, C., Axén, C. and Fayer, R. (2014). Cryptosporidiosis in Farmed Animals. In *Cryptosporidium: parasite and disease* (ed. Caccio, S. M. and Widmer, G.), pp. 149–236. Springer Wien Heidelberg, New York, Dordrecht, London.

Robinson, G. and Chalmers, R. M. (2012). Assessment of polymorphic genetic markers for multi-locus typing of *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Experimental Parasitology* **132**, 200–215.

Strong, W. B., Gut, J. and Nelson, R. G. (2000). Cloning and sequence analysis of a highly polymorphic *Cryptosporidium parvum* gene encoding a 60-kilodalton glycoprotein and characterization of its 15- and 45-kilodalton zoite surface antigen products. *Infection and Immunity* **68**, 4117–4134.

Struelens, M. J. (1996). Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems. *Clinical Microbiology and Infection* **2**, 2–11.

Sulaiman, I. M., Hira, P. R., Zhou, L., Al-Ali, F. M., Al-Shelahi, F. A., Shweiki, H. M., Iqbal, J., Khalid, N. and Xiao, L. (2005). Unique endemicity of cryptosporidiosis in children in Kuwait. *Journal of Clinical Microbiology* **43**, 2805–2809.

Tanriverdi, S., Markovics, A., Arslan, M. O., Itik, A., Shkap, V. and Widmer, G. (2006). Emergence of distinct genotypes of *Cryptosporidium parvum* in structured host populations. *Applied and Environmental Microbiology* **72**, 2507–2513.

van Belkum, A., Tassios, P. T., Dijkshoorn, L., Haeggman, S., Cookson, B., Fry, N. K., Fussing, V., Green, J., Feil, E., Gerner-Smidt, P., Brisse, S. and Struelens, M. (2007). Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clinical Microbiology and Infection* **13** (Suppl. 3), 1–46.

Wells, B., Shaw, H., Hotchkiss, E., Gilray, J., Ayton, R., Green, J., Katzer, F., Wells, A. and Innes, E. (2015). Prevalence, species identification and genotyping *Cryptosporidium* from livestock and deer in a catchment in the Cairngorms with a history of a contaminated public water supply. *Parasites and Vectors* **8**, 66.

Widmer, G. and Sullivan, S. (2012). Genomics and population biology of *Cryptosporidium* species. *Parasite Immunology* **34**, 61–71.

Widmer, G. and Cacciò, S. M. (2015). A comparison of sequence and length polymorphism for genotyping *Cryptosporidium* isolates. *Parasitology* **142**, 1080–1085.

Xiao, L. and Ryan, U. (2008). Molecular epidemiology. In *Cryptosporidium and Cryptosporidiosis*. (ed. Fayer, R. and Xiao, L.), pp. 119–163. CRC Press, Boca Raton.