Paper:

Pérez-Cordón, G., Robinson, G., Nader, J. & Chalmers, R. (2016). Discovery of new variable number tandem repeat loci in multiple Cryptosporidium parvum genomes for the surveillance and investigation of outbreaks of cryptosporidiosis. *Experimental Parasitology*

http://dx.doi.org/10.1016/j.exppara.2016.08.003

Discovery of new variable number tandem repeat loci in multiple *Cryptosporidium parvum* genomes for the surveillance and investigation of outbreaks of cryptosporidiosis

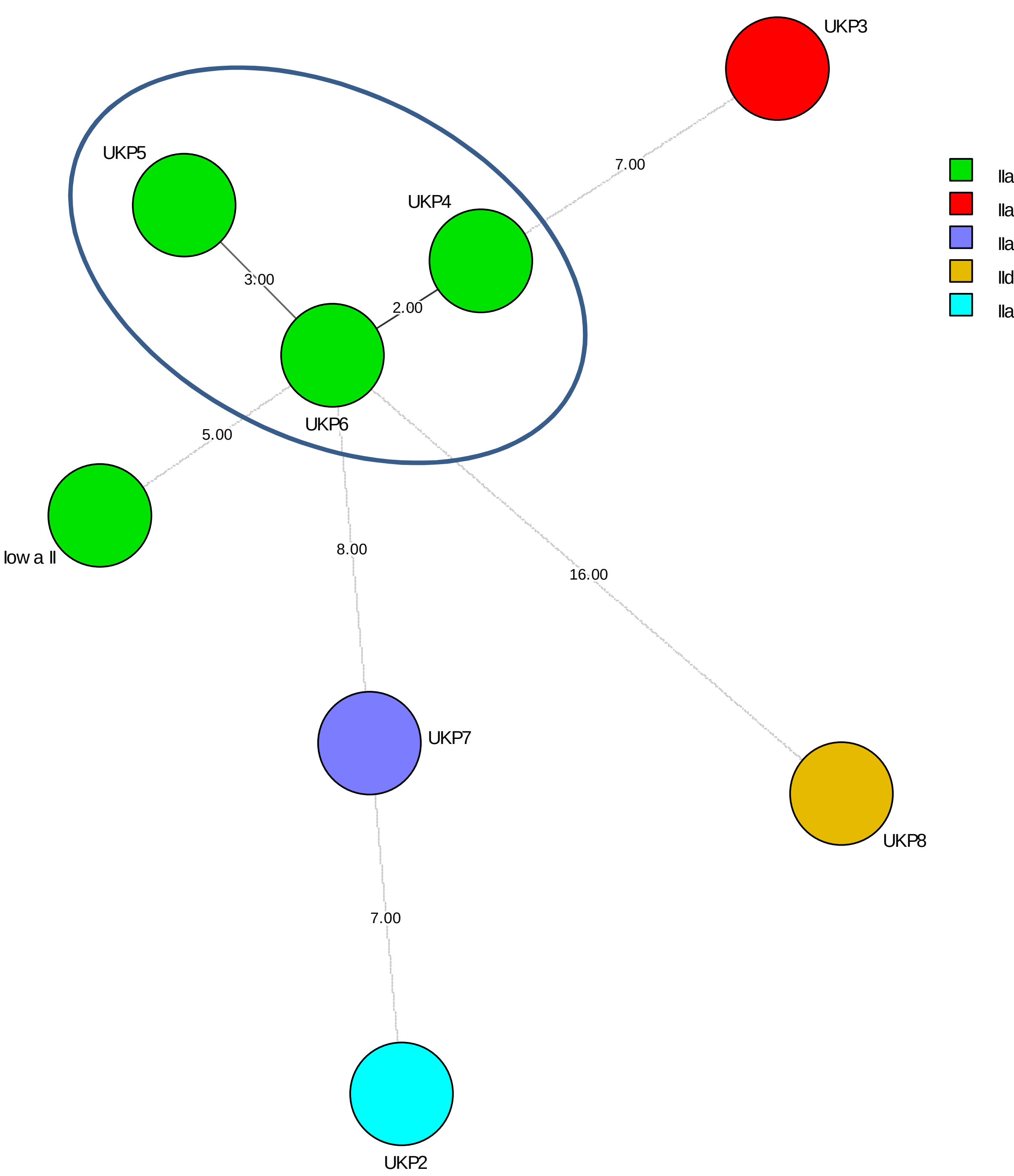
Gregorio Pérez-Cordón, Guy Robinson, Johanna Nader, Rachel M. Chalmers

PII: S0014-4894(16)30161-8
DOI: 10.1016/j.exppara.2016.08.003
Reference: YEXPR 7285

To appear in: *Experimental Parasitology*

Received Date: 26 May 2016
Revised Date: 9 August 2016
Accepted Date: 10 August 2016

Please cite this article as: Pérez-Cordón, G., Robinson, G., Nader, J., Chalmers, R.M., Discovery of new variable number tandem repeat loci in multiple *Cryptosporidium parvum* genomes for the surveillance and investigation of outbreaks of cryptosporidiosis, *Experimental Parasitology* (2016), doi: 10.1016/j.exppara.2016.08.003.


UKP3
7.00
UKP5
UKP4
3.00
2.00
UKP6
5.00
low a ll
8.00
16.00
UKP7
UKP8
7.00
UKP2
🟩 IIa
🟥 IIa
🟦 IIa
🟨 IId
🟦 IIa

Gregorio Pérez-Cordón[a], Guy Robinson[a,b], Johanna Nader[c] and Rachel M Chalmers[a,b]*

[a] Cryptosporidium Reference Unit, Public Health Wales Microbiology, Singleton Hospital, Swansea, SA2 8QA, UK

[b] Swansea University Medical School, Grove Building, Swansea University, Singleton Park, Swansea, SA2 8PP, UK

[c] Norwich Medical School, University of East Anglia, Norwich, UK

*Corresponding author: Tel.: + 44 (0) 1792 285341; Fax: + 44 (0) 1792 202320

E-mail address: Rachel.Chalmers@wales.nhs.uk

(cryptosporidiosis) in humans and animals. The ability to investigate sources of contamination and routes of transmission by characterisation and comparison of isolates in a cost- and time-efficient manner will help surveillance and epidemiological investigations, but as yet there is no standardised multi-locus typing scheme. To systematically identify variable number tandem repeat (VNTR) loci, which have been shown to provide differentiation in moderately conserved species, we interrogated the reference *C. parvum* Iowa II genome and seven other *C. parvum* genomes using a tandem repeat finder software. We identified 28 loci that met criteria defined previously for robust typing schemes for inter-laboratory surveillance, that had potential for generating PCR amplicons analysable on most fragment sizing platforms: repeats ≥ 6 bp, occurring in tandem in a single repeat region, and providing a total amplicon size of < 300 bp including 50 bp for the location of the forward and reverse primers. The qualifying loci will be further investigated *in vitro* for consideration as preferred loci in the development of a robust VNTR scheme.

Keywords

Cryptosporidium parvum, variable number tandem repeats, multilocus

parasite *Cryptosporidium*. The parasite is transmitted via the faecal-oral route through the ingestion of oocysts, either by direct contact with infected hosts or in contaminated food or water, which may lead to the emergence of large scale outbreaks (Ortega and Cama, 2008; Chalmers, 2012). Among the 26 or so species that have been described to date, *Cryptosporidium hominis* is the most common anthroponotic species and *Cryptosporidium parvum* is the most common zoonotic species infecting humans and a wide range of animals, placing an economic and welfare burden on livestock farming as well as public health (Xiao, 2010; Shirley et al., 2012). Subtyping of isolates is of utmost importance to investigate sources of contamination and routes of transmission and in doing so, identify appropriate interventions.

The life cycle of *Cryptosporidium* involves both asexual and sexual reproduction and genetic recombination has been demonstrated experimentally in *C. parvum* (Feng et al., 2002). Therefore, it is feasible to suppose that recombination between different genotypes occurs in nature giving rise not only to new genotypes but also to heterogeneous populations, although the scale of occurrence within hosts is not known as many genotyping methods lack sensitivity for their detection (Grinberg and Widmer, 2016). *Cryptosporidium parvum* genotypes have traditionally been identified based on sequence analysis of the gp60 gene, in which variable numbers of tandem serine codons as well as downstream polymorphisms differentiate subtypes (Strong et al., 2000).

characterization of outbreak isolates and infer linkage (Hotchkiss et al., 2015; Chalmers et al., 2016). However, VNTRs have been used in many combinations on different analytical platforms in a limited number of studies for genotyping *C. parvum* and investigating population structure and transmission and there is as yet no standardised multilocus subtyping scheme (Robinson and Chalmers, 2012). Some of the currently used VNTR loci are either poorly suited to fragment sizing (Chalmers et al., 2016) or have been found to be monoallelic in some populations (Hotchkiss et al., 2015). For international surveillance and outbreak investigations, a robust, multilocus VNTR scheme, incorporating suitable loci for the different analytical platforms that might be used in different laboratories, would provide a portable tool. Criteria and processes for the selection of markers have been described for bacterial pathogens (Nadon et al., 2013). However, many of the VNTR loci used for fragment sizing analyses of *C. parvum* have been identified as sub-optimal. Either they are very short repeat units producing similar sized fragments that are prone to amplification errors due to slippage and that are hard to differentiate on many analytical platforms, or are complex and non-tandem in occurrence, and there is a need for the identification of new loci (Robinson and Chalmers, 2012; Chalmers et al., 2016).

Traditionally, options for identifying new candidate VNTRs include: screening thousands of clones in genomic libraries through colony hybridization with repeat-containing probes such as RAPD-based to avoid library construction and screening, primer extension-based methods for the production of libraries enriched in microsatellite loci, and selective hybridization (Zane el al., 2002). In recent years, with the continued improvement of next

E. coli, this is now the standard typing method (Dallman et al. 2015) and for others it provides a means for identifying new markers. Interrogating whole genome sequences provides an efficient, simplified method of identifying new VNTR regions (Lim et al., 2012; Zapala et al., 2012). However, whole genome sequencing of *Cryptosporidium* spp. has lagged behind that of other pathogens, such as those that are culturable, present in greater abundance, or in less complex samples than faeces. Until recently, only three *Cryptosporidium* genomes were available, one each of *C. parvum*, *C. hominis* and *Cryptosporidium muris* (Abrahamsen et al., 2004; Xu et al., 2004; http://cryptodb.org). However, through the use of appropriate faecal sample selection, oocyst purification by flotation and immunomagnetic separation, followed by bleach treatment to degrade exogenous nucleic acid, new *Cryptosporidium* whole genome sequences have been generated from clinical samples, increasing the number of sequences available (Hadfield et al., 2015). Genomes can be mined rapidly and efficiently using bioinformatics tools, expanding the potential for the identification of new diagnostic and genotyping markers. For *Cryptosporidium*, several studies have used software programs to mine the previously limited number of genomes to identify VNTR loci in *C. parvum*, *C. hominis* and *C. muris* and used them to multilocus genotype isolates by sequencing or fragment sizing (Tanriverdi and Widmer, 2006; Feng et al., 2011; Herges et al., 2012; Li et al., 2013; Ramo et al. 2016a). Additionally, *Cryptosporidium* genome mining of newly produced genomes has been used in the identification of unique gp60 sequences within the genome of the emerging pathogen *Cryptosporidium ubiquitum* (Li et al., 2014). Here we describe the mining of multiple *C. parvum* genomes for the identification of VNTR loci and the verification *in silico* of their

2. Methods

2.1. Identification of variable VNTR loci and their attributes

To identify robust MLVA candidate loci for inter-laboratory surveillance and outbreak investigations, selection criteria were first defined on the basis of a previous *in vitro* evaluation study (Chalmers et al., 2016) and published guidance (Nadon et al., 2013): repeats ≥ 6 bp, occurring in tandem in a single repeat region, and providing a total amplicon size of < 300 bp including 50 bp for the location of the forward and reverse primers which would give fragments suitable for sizing on most platforms. The *C. parvum* Iowa II reference genome (Table 1; Puiu et al., 2004) was retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov) and interrogated for qualifying loci meeting our selection criteria using Tandem Repeat Finder (TRF) software (version 4.07b, Boston University) (Benson., 1999) using the default settings. The output table of identified tandem repeats was transferred to a spreadsheet (Excel 2007, Microsoft) and repeats of < 6 bp rejected. Repeats with < 90% sequence similarity among the copies were also rejected and those with ≥ 90%, with the variation limited to only the ends of the region, examined further.

The repeat size, sequence and copy number, gene name and chromosome location, GC content and conservation of the sequences flanking the repeat units of the remaining repeat regions was recorded. The corresponding loci within seven other *C. parvum* whole genomes (UKP2 through to UKP8; Table 1) published previously (Hadfield et al., 2015) and obtained from the umbrella BioProject PRJNA215218 on the NCBI database were edited to include only the VNTR and immediate flanking regions, the orientation checked and the validity of coding sequences and reading frames identified in the *C. parvum* Iowa II reference genome on CryptoDB. The true repeat units were identified by checking that repeats in coding regions were represented by whole codons in the correct interval from the methionine start codon. Motifs similar to the true repeat that consistently flanked the VNTR units without variation were not included in the definition of the repeat region; an example is shown in Figure 1. Only those loci that displayed variations in the number of repeats in the eight aligned isolates were included the final selection.

The number of true repeat units was determined for each locus in each genome, and any additional features of interest that could influence the further selection of qualifying loci for PCR development were noted. To investigate whether any potential tandem repeats were present as only single copies in the Iowa II reference genome, the process was repeated using the genome of *C. parvum* UKP8 (selected as it is a different gp60 family compared to the other seven and therefore more likely to vary from Iowa; Table 1) as the reference.

2.2. Literature and database search

To validate our identification procedure, we looked in the TRF output spreadsheet for the loci reviewed previously by Robinson and Chalmers (2012) and those arising from a new literature search using the terms Cryptosporidium AND parvum AND (VNTR OR tandem OR microsat* OR minisat* OR multiloc* OR multi-loc*) undertaken in PubMed for the time period 1st November 2011 to 20th May 2016.

http://EuPathDB.org using the BLAST search tool to see if they were present in the reference genomes of *C. hominis* and *C. muris*, which may be desirable if a common subtyping approach is required for both *C. parvum* and *C. hominis* for example. Likewise, the genomes of genera within the other taxa available on the database (Amoebozoa, Apicomplexa, Chromerida, Diplomonadida, Fungi, Kinetoplastida, Oomycetes, and Trichomonadida) were also checked as homology in potential primer sequences would compromise the specificity of any assay based on these loci. The repeat regions and 50 bp of the flanking sequences from each of the identified *C. parvum* loci were used as the query sequence using default parameters.

2.3. Bioinformatic analyses

To compare the eight *C. parvum* isolates at all selected loci, a Minimal Spanning Tree (MST) was produced using Bionumerics 7.6 (Applied Maths). To determine the potential for the MLVA approach to be used as a surrogate for whole genome comparison of closely related isolates, the MST was compared with phlyogenetic analysis of four isolates with the same gp60 subtype, UKP4, 5, 6 and Iowa II, conducted on the FASTA files from the NCBI Bioprojects (Table 1; Hadfield et al., 2015) using MEGA version 6 (Molecular Evolutionary Genetics Analysis; Tamura et al. 2013) and aligned using the integrated ClustalW multiple sequence alignment program. Isolates UKP4, 5 and 6 were from cryptosporidiosis cases diagnosed during a widespread foodborne outbreak in the UK in 2012 (McKerr et al., 2015). The ~9.08 Mb whole genome alignment was subsequently examined manually to ensure sequence integrity and consensus across the four isolates. Phylogenetic reconstruction of

Likelihood model and uniform rates among sites. Confidence of the phylogenetic tree was assessed using 1000 bootstrap replications.

3. Results

3.1. Identification of variable VNTR loci and their attributes

A total of 2284 tandem repeat loci were identified initially in the *C. parvum* Iowa II reference genome, but after rejecting 2074 loci with repeats of < 6 bp or showing < 90 % similarity among the copies of the repeat, and 182 loci that showed no variation in copy number within the other seven genomes, 28 remained for further examination (Table 2). Interrogating the UKP8 genome, 2016 loci were identified initially, but after applying our selection criteria and removing duplicates identified initially in the IOWA II genome, eight additional loci remained. However, those eight were also rejected as they showed no variation in copy number within the other genomes investigated (Table 2).

The repeat size, sequence and copy number, gene name and chromosome location, GC content and conservation of the sequences flanking the repeat units of the remaining edited and validated repeat regions are shown in Tables 3 and 4. Of the 28 qualifying VNTR loci, 16 met all of the guidance criteria published by Nadon et al. (2013) while 12 had some variation. For two loci this was in the flanking region only, for seven it was towards the ends of the VNTR region and for three loci it was in both the flanking region and towards the ends of the VNTR region (Table 3). The variability in the flanking regions was not predicted to hinder assay design or affect fragment sizing because it was due to substitutions and not

The 28 qualifying loci were found across all eight *C. parvum* chromosomes (Tables 2, 3 and 4). Chromosomes 2 and 4 had the most qualifying loci, with six loci each; chromosome 3 had the least with only a single qualifying locus.

The majority of VNTR sequences in the qualifying loci were non-polymorphic (18/28), especially those found in chromosome 2 where there was no sequence variation within the six repeat units. Twenty five of the 28 qualifying loci were coding, and the most common repeat unit length was 6 bp and the longest was 27 bp (cgd6_4290_9811) (Table 3). The three non-coding loci (one on chromosome 5 and the two on chromosome 8) were 6, 13 and 18 bp in length. With the intention of developing *in vitro* assays and designing PCR primers, we looked at the GC % content as well as the conservation of the sequences flanking the repeat region. In all cases, the GC content was ≤ 50 % and all but 5 qualifying loci showed 100 % conservation of flanking sequences upstream and downstream of the repeat region (Table 3).

Of the 28 qualifying loci, 19 were found in all eight genomes interrogated. The non-detects occurred mostly as singles (six loci) but three loci (cgd4_3940_298, cgd4_1340_1688, and cgd5_4490_2941) were not detected in two, three and four genomes respectively (Table 4). The number of alleles identified for each qualifying locus in the eight genomes investigated varied between two (21 loci), three (4 loci), four (one locus), five (one locus) and seven (one locus, cgd8_NC_ 4440_505) (Table 4). Of the 19 loci found in all eight genomes, eight

IIa genomes but could not separate IId (Table 4).

3.2. Literature and database search

Of the 55 VNTR loci reviewed by Robinson and Chalmers (2012), 18 were ≥ 6 bp, but only MSF (Tanriverdi and Widmer, 2006) was selected by our criteria for further examination (cgd5_10_310, Table 3). The remaining 17 were not included as six showed < 90 % similarity among the repeat copies and in 11 the variation was distributed throughout the repeat region. One locus overlooked previously, MSC6-5 (Xiao and Ryan, 2008), was also selected through our process (cgd6_4290_9811, Table 3).

A total of 35 new publications were identified using the search terms defined in PubMed within the time period considered, of which 19 were considered relevant. Only three of these reported “new” loci. Herges et al. (2012) described the GRH locus, detected with the same TRF software that we used, identified in our study as cgd1_470_1429 (Table 3) with the repeat re-defined based on the correct reading frame. The two others (Ramo et al., 2016a and 2016b) included four previously un-described VNTR loci. We found all four in our initial screening of the *C. parvum* Iowa II genome, and two qualified in our analysis (Table 3), although again we defined the repeat sequences differently, based on their DNA codons in the correct open reading frame. Additionally, one was translated from the antisense strand (Table 3). Two were rejected (cgd2_3850 and cgd6_5400) as they presented < 90 % similarity throughout the repeat regions.

confirmed as present only in *C. parvum.* None of the loci were indicated to be present in *C. muris.* The BLAST results against other taxa on EuPathDB only returned results showing low similarity, or close matches over very short sequence spans suggesting that non-specific amplification would be avoided by careful primer design.

3.3. Bioinformatic analyses

All eight isolates were differentiated *in silico* by MLVA using all 28 loci (Figure 2). In fact the minimum number of loci required to differentiate all eight isolates was two (cgd8_NC_4440_506 and any one of eight others, the most discriminatory being cgd4_2350_796, Table 4).

Both MLVA and whole genome comparison of UKP4, 5 and 6 and Iowa II showed similar outcomes: while each individual isolate could be identified separately, the UKP4, 5 and 6 clustered closely together when compared to the other genomes (Figure 2) and when compared to Iowa II (Figure 3).

4. Discussion

The clinical and economic impact of cryptosporidiosis demands the development of strategies for improved surveillance and control including the ability to investigate, through genotyping, sources of contamination and routes of transmission in a fast and reliable way. The availability of seven new *C. parvum* genomes (Hadfield et al., 2015), in addition to the reference Iowa II genome, allowed us to perform an *in silico* analysis of new potential VNTR sequences (Zane, 2002), and was fruitful in our analysis; of the 28 qualifying loci identified, 23 were new and just 5 had been identified previously.

Cryptosporidium genome mining for VNTR loci has been restricted in the past because of the limited number of genomes available, and required subsequent laboratory experiments to predict their discriminatory potential (Tanriverdi and Widmer, 2006, Feng et al., 2011, Herges et al., 2012, Li et al., 2013, Ramo et al., 2016a).

Although the accuracy of NGS may be challenged by homopolymers, one study reported the acceptable identification of short tandem repeats, present in the yeast *Saccharomyces cerevisiae* in copy numbers of a similar order of magnitude to those in our whole genome sequences (Zavodna et al., 2014). The depth of coverage of the genome sequencing was identified as being important, but cannot alone resolve assembly gaps caused by repetitive regions with lengths that approach or exceed those of the short NGS reads (Sims et al., 2014). The required average mapped depth to allow reliable calling of SNPs and small indels across 95 % of the genome has reduced from 50x to 35x due to improvements in sequencing chemistry reducing GC bias and yielding a more uniform coverage (Sims et al., 2014). The overall range of coverage of the *C. parvum* genomes in our study ranged from 26.86x to 192.48x (mean 113.52x) for the UKP genomes (Hadfield et al., 2015) and 13x for Iowa II (Abrahamsen et al., 2004) (Table 1). We therefore considered that using the genome sequences not only allowed us to locate and describe the VNTRs, but also compare the outputs and outcomes phylogenetically.

in coding regions. It is likely that selection pressure for sequence conservation drove the occurrence of homogeneous repeats mainly in coding regions (Madesis et al., 2013). The use of only perfect non-polymorphic repeats for MLVA was recommended by Nadon et al. (2013), but to identify these it was necessary to loosen the parameters to include those repeats with ≥ 90 % similarity before manually determining the true repeat, as flanking sequences similar to the repeat would sometimes stop the software from returning some of the results when set to 100 %. For example, TCA TCA TCT would not return if set to 100 %, because the TCT unit would be counted as part of the repeat, even if it was consistently present and non-variable. While this undoubtedly resulted in the loss of a number of VNTR loci that may indeed be useful, the objective was not to identify all of the tandem repeats present, but to identify new suitable candidates that could be examined further to develop a robust typing scheme. The 90% cut-off was not pre-determined, but selected arbitrarily based on the number of initial results that it returned (210 before assessing the spread of variation throughout the region and discrimination with the other *C. parvum* isolates). Additionally, we made the assumption that loci with the highest similarity between repeat copies would be more robust in a typing method.

Most repeat units were short (6 bp) but some longer ones were identified, up to 27 bp, but there didn’t appear to be any major significance associated with the length of the repeat and potential for discrimination, although this is probably due to only 2 or 3 alleles being found at most loci (25/28). The most discriminatory locus was the 6 bp repeat cgd8_NC_4440_506 that separated the 8 genomes into 7 different alleles, but the second number of alleles within the maximum fragment size requirements for a multi-platform scheme. For example, when the two most discriminatory loci are compared in the Iowa II genome, cgd8_NC_4440_506 had 30 copies of the repeat opposed to the 13 copies of cgd4_3450_4336, but the latter is at the top end of the preferred size range (< 300 bp including 50 bp flanking regions for primer annealing) because each copy is 15 bp. The advantage however, with longer repeats is the easier separation of alleles based on fragment size as variation in the sizing is less likely to overlap with the next allele size.

The distribution of qualifying loci was across all chromosomes, with the number per chromosome ranging from one (chromosome 3) to six (chromosomes 2 and 4) (Table 3). The selection of loci for a scheme based on the diversity and spread across different chromosomes is particularly important in *Cryptosporidium* due to the potential for recombination during the sexual stage of the life cycle (Widmer & Sullivan, 2012). While a spread of loci across chromosomes is required, a representative from each chromosome is not necessary, because the aim would be to identify a multilocus method providing good resolution but with the smallest number of markers (Widmer & Sullivan, 2012).

Of the 28 qualifying loci, nine were not detected in one or more of the whole genomes (Table 4). There could be a few explanations for this including, mismatches in the sequence inhibiting the identification of the target sequence, poor coverage of the genome at that particular locus or a true absence of the repeat in that isolate. The locus that had the most non-detects (cgd5_4490_2941) only identified alleles in half of the genomes. However, each was not detected in half of the genomes and whether following primer design to specifically target it can the VNTR be detected in all isolates.

The eight *C. parvum* genomes investigated comprised two *C. parvum* gp60 families (IIa and IId) which are prevalent in both humans and animals worldwide (Wang et al., 2014). While the gp60 marker does provide relatively good discrimination between isolates of *C. parvum*, it, along with other single loci, does not serve as a surrogate for other loci or multilocus genotypes (Widmer and Lee, 2010). However, as these data were readily available for each of our genomes, the gp60 genotype of each isolate could provide some initial indication to differences between isolates for comparative purposes with the newly identified alleles. Eight of our loci could only differentiate the two gp60 families IIa and IId (Table 4), whereas the remaining candidate VNTR loci allowed for some intra-gp60 family discrimination (e.g. cgd1_3060_604 with two alleles or cgd8_NC_ 4440_505 with seven alleles) sometimes in addition to family discrimination. Although within-host populations of *Cryptosporidium* are likely to be genetically diverse (Grinberg and Widmer, 2016), MLVA has the potential to identify these mixed populations. The genome sequences interrogated in our study were reported to show no evidence of being mixed species (Hadfield et al., 2015), but from the sequence data alone we cannot be certain that there are no mixed populations of *C. parvum* genotypes present.

The apparent discriminatory power of VNTR loci has been shown previously to differ between gp60 families. For example, in two studies the VNTR locus MSF (Tanriverdi and

Hotchkiss et al. 2015). Consideration of the hosts likely to be investigated is important; in Spain and the UK, gp60 family IIa is more common in cattle (Quilez et al., 2008a; Hotchkiss et al., 2015) and IId in sheep and goats (Quilez et al., 2008b), so loci such as cgd5_10_310 (MSF, Tanriverdi and Widmer, 2006) would be less informative in cattle isolates compared to sheep and goats (Hotchkiss et al., 2015). Indeed, in a recent study by Ramo et al. (2016b) two VNTR loci that were previously used for intra-species typing in cattle and showed to be poorly discriminatory (Ramo et al., 2016a) were among the most informative for typing in sheep. Due to the prior selection of samples for whole genome sequencing (Hadfield et al., 2015), only one of the genomes analysed in our study was IId, whereas the other seven genomes were IIa , which may have resulted in selection bias towards loci that are more variable in IIa. Further testing *in vitro* of a larger, varied panel of isolates is required to provide more detailed information about the discriminatory capabilities of the qualifying loci. For example, Herges et al. (2012) identified 10 different GRH (syn. cgd1_470_1429 in this study) alleles in 254 *C. parvum* isolates from humans and cattle, second in discrimination only to gp60 with 22 alleles. There remains a need for more *Cryptosporidium* whole genomes to be published ideally from different sources to increase the amount of potential variation and allow us to make less biased comparisons. The number of available genomes is increasing (Andersson et al., 2015; Hadfield et al., 2015; Guo et al., 2015) and mining their data will help further in the development of efficient MLVA schemes.

Comparison of three isolates (UKP4, 5 and 6) from cryptosporidiosis cases who lived in the North East of England and were diagnosed during a large foodborne outbreak in 2012

(2013), linkage equilibrium was observed in the gp60 subtype IIaA15G2R1 group but not in the non-IIaA15G2R1 group, indicating the possible presence of genetic recombination and maybe explaining the variation at other loci within the IIaA15G2R1 gp60 genotype. However, in our study the cgd2_3300_1504 and cgd4_2350_796 loci only differed between the three isolates in UKP5 and it is a possibility that this variation could be due to inaccuracies in the UKP5 sequence assembly as the depth of coverage was only 26.86x. Another potential for inaccuracy in two of the loci (cgd4_2350_796 and cgd8_NC_4440_505) is that the repeats are approaching the size of the raw NGS reads, which as described above can make it hard to resolve assembly gaps in these regions (Sims et al., 2014). Testing these three isolates with carefully designed PCR assays at these loci would help resolve whether these isolates are indeed different from each other. It is also possible that in outbreaks, especially ones where there must have been a high degree of contamination to cause geographically widespread illness in >300 confirmed cases (McKerr et al., 2015), mixed populations of oocysts may have caused the infections resulting in differing allelic profiles. Alternatively, the cases may not have been linked by a common exposure to the source of the outbreak and may have been background, unlinked cases. A comparison by MST with all 28 loci and phylogenetic analysis of the whole genome both suggested that although slightly different, the outbreak samples cluster together separately from the Iowa II isolate (Figures 2 and 3). This suggests that with careful selection of loci, MLVA may serve as a surrogate to whole genome analysis when studying relationships between epidemiological relevant isolates with clear cost-saving benefits. In addition to cost, whole genome sequencing of *Cryptosporidium* is also hindered by the non-culturable

WGS to be applicable (Hadfield et al., 2015).

Conclusions

The strategy we followed for this study enabled the identification of 28 VNTR loci that may be suitable for the development of a robust MLVA scheme. This study not only mined a *C. parvum* reference genome (Iowa II) to identify VNTR loci, but also utilised seven additional *C. parvum* genomes to determine the potential for intra-isolate discrimination. The potential for these loci to discriminate isolates was demonstrated by comparing alleles, MST and UPGMA. For an efficient MLVA scheme the number and selection of loci should be ideally reduced to a minimum number of discriminatory loci to maintain cost and time efficiency for epidemiological investigations. The next step will be subjecting selected loci to *in vitro* testing to assess their typability and discriminatory power by capillary electrophoretic sizing of amplified DNA from both related and unrelated isolates.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme ([FP7/2007-2013] [FP7/2007-2011]) under Grant agreement no: 311846.

References

Abrahamsen, M.S., Templeton, T.J., Enomoto, S., Abrahante, J.E., Zhu, G., Lancto, C.A., Deng, M., Liu, C., Widmer, G., Tzipori, S., Buck, G.A., Xu, P., Bankier, A.T., Dear, PH., Konfortov, .A., 10.1126/science.1094786

Andersson, S., Sikora, P., Karlberg, M.L., Winiecka-Krusnell, J., Alm, E., Beser, J., 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. J. Microbiol. Methods. 113, 10-2. doi: 10.1016/j.mimet.2015.03.018

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573-80. doi: 10.1093/nar/27.2.573

Chalmers, R.M. 2012. Waterborne outbreaks of cryptosporidiosis. Ann. Ist. Super Sanita. 48, 429-46. doi: 10.4415/ANN_12_04_10

Chalmers, R.M., Robinson, G., Hotchkiss, E., Alexander, C., May, S., Gilray, J., Connelly, L., Hadfield, S.J. 2016. Suitability of loci for multiple-locus variable-number of tandem-repeats analysis of *Cryptosporidium parvum* for inter-laboratory surveillance and outbreak investigations. Parasitology 2, 1-11. doi: 10.1017/S0031182015001766

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., Wain, J. 2015. Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. Clin. Infect. Dis. 61, 305-12. doi: 10.1093/cid/civ318.

recombination in the opportunistic pathogen *Cryptosporidium parvum*. Mol. Biochem. Parasitol. 119, 55-62. doi: 10.1016/S0166-6851(01)00393-0

Feng, Y., Yang, W., Ryan, U., Zhang, L., Kvác, M., Koudela, B., Modry, D., Li, N, Fayer, R, Xiao, L. 2011. Development of a multilocus sequence tool for typing *Cryptosporidium muris* and *Cryptosporidium andersoni*. J. Clin. Microbiol. 49, 34-41. doi: 10.1128/JCM.01329-10

Feng, Y., Torres, E., Li, N., Wang, L., Bowman, D., Xiao, L. 2013. Population genetic characterisation of dominant *Cryptosporidium parvum* subtype IIaA15G2R1. Int. J. Parasitol. 43, 1141-7. doi: 10.1016/j.ijpara.2013.09.002

Grinberg., A., Widmer., G. 2016. *Cryptosporidium* within-host genetic diversity: systematic bibliographical search and narrative overview. Int. J. Parasitol. Published online ahead of print. doi: 10.1016/j.ijpara.2016.03.002

Guo, Y., Li, N., Lysén, C., Frace, M., Tang, K., Sammons, S., Roellig, D.M., Feng, Y., Xiao, L. 2015. Isolation and enrichment of *Cryptosporidium* DNA and verification of DNA purity for whole-genome sequencing. J. Clin. Microbiol. 53, 641-7. doi: 10.1128/JCM.02962-14

Hadfield, S.J., Pachebat, J.A., Swain, M.T., Robinson, G., Cameron, S.J., Alexander, J., Hegarty, M.J., Elwin, K., Chalmers, R.M. 2015. Generation of whole genome sequences of

Herges, G.R, Widmer, G., Clark, M.E., Khan, E., Giddings, C.W., Brewer, M., McEvoy, J.M. 2012. Evidence that *Cryptosporidium parvum* populations are panmictic and unstructured in the Upper Midwest of the United States. Appl. Environ. Microbiol. 78: 8096-101. doi: 10.1128/AEM.02105-12

Hotchkiss, E.J., Gilray, J.A., Brennan, M.L., Christley, R.M., Morrison, L.J., Jonsson, N.N., Innes, E.A., Katzer, F. 2015. Development of a framework for genotyping bovine-derived *Cryptosporidium parvum*, using a multilocus fragment typing tool. Parasit. Vectors. 8, 500. doi: 10.1186/s13071-015-1107-8

Li, N., Xiao, L., Cama, V.A., Ortega, Y., Gilman, R.H., Guo, M., Feng, Y. 2013. Genetic recombination and *Cryptosporidium hominis* virulent subtype IbA10G2. Emerg. Infect. Dis. 19, 1573-82. doi: 10.3201/eid1910.121361

Li, N., Xiao, L., Alderisio, K., Elwin, K., Cebelinski, E., Chalmers, R., Santin, M., Fayer, R., Kvac, M., Ryan, U., Sak, B., Stanko, M., Guo, Y., Wang, L., Zhang, L., Cai, J., Roellig, D., Feng, Y. 2014. Subtyping *Cryptosporidium ubiquitum*, a zoonotic pathogen emerging in humans. Emerg. Infect. Dis. 20, 217-24. doi: 10.3201/eid2002.121797

10.1093/bib/bbs023

Madesis, P., Ganopoulos, I., Tsaftaris, A. 2013. Microsatellites: Evolution and Contribution, Methods Mol. Biol. 1006, 1-13. doi: 10.1007/978-1-62703-389-3_1

McKerr, C., Adak, G.K., Nichols, G., Gorton, R., Chalmers, .RM., Kafatos, G., Cosford, P., Charlett, A., Reacher, M., Pollock, K.G., Alexander, C.L., Morton, S. 2015. An Outbreak of *Cryptosporidium parvum* across England & Scotland associated with consumption of fresh pre-cut salad leaves, May 2012. PLoS One. 10, e0125955. doi: 10.1371/journal.pone.0125955

Nadon, C.A., Trees, E., Ng, L.K., Møller Nielsen, E., Reimer, A., Maxwell, N., Kubota, K.A., Gerner-Smidt, P. 2013. MLVA Harmonization Working Group. Development and application of MLVA methods as a tool for inter-laboratory surveillance. Euro Surveill. 18, 20565. doi: 10.2807/1560-7917.es2013.18.35.20565

Ortega, Y.R., Cama, V.A. 2008. Foodborne transmission. In: Fayer, R., Xiao, L. (Eds), *Cryptosporidium* and Cryptosporidiosis. CRC Press, Bocan Raton, pp. 289-304.

Puiu, D., Enomoto, S., Buck, G.A., Abrahamsen, M.S., Kissinger, J.C. 2004. CryptoDB: the *Cryptosporidium* genome resource. Nucleic Acids Res. 32, 329-31. doi: 10.1093/nar/gkh050

1613-1620. doi: 10.1017/s0031182008005088

Quílez, J., Torres, E., Chalmers, R.M., Hadfield, S.J., Del Cacho, E., Sánchez-Acedo, C. 2008b. *Cryptosporidium* genotypes and subtypes in lambs and goat kids in Spain. Appl. Environ. Microbiol. 74, 6026-31. doi: 10.1128/aem.00606-08

Ramo, A., Quílez, J., Monteagudo, L., Del Cacho, E., Sánchez-Acedo, C., 2016a. Intra-Species Diversity and Panmictic Structure of *Cryptosporidium parvum* Populations in Cattle Farms in Northern Spain. PLoS One. 11, e0148811. doi: 10.1371/journal.pone.0148811

Ramo, A., Monteagudo, L.V., Del Cacho, E., Sánchez-Acedo, C., Quílez, J. 2016b. Intra-Species Genetic Diversity and Clonal Structure of *Cryptosporidium parvum* in Sheep Farms in a Confined Geographical Area in Northeastern Spain. PLoS One. 11, e0155336. doi: 10.1371/journal.pone.0155336.

Robinson, G., Chalmers, R.M. 2012. Assessment of polymorphic genetic markers for multi-locus typing of *Cryptosporidium parvum* and *Cryptosporidium hominis*. Exp. Parasitol. 132, 200-15. doi: 10.1016/j.exppara.2012.06.016

Shirley, D.A., Moonah, S.N., Kotloff, K.L. 2012. Burden of disease from cryptosporidiosis. Curr. Opin. Infect. Dis. 25, 555-63. doi: 10.1097/qco.0b013e328357e569

10.1038/nrg3642.

Strong, W.B., Gut, J., 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. Infect. Immun. 68, 4117-34. doi: 10.1128/iai.68.7.4117-4134.2000

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 30, 2725–9. doi: 10.1093/molbev/mst197

Tanriverdi, S., Widmer, G. 2006. Differential evolution of repetitive sequences in *Cryptosporidium parvum* and *Cryptosporidium hominis*. Infect. Genet. Evol. 6, 113-22. doi: 10.1016/j.meegid.2005.02.002

Wang, R., Zhang, L., Axén, C., Bjorkman, C., Jian, F., Amer, S., Liu, A., Feng, Y., Li, G., Lv, C., Zhao, Z., Qi, M., Dong, H., Wang, H., Sun, Y., Ning, C., Xiao, L. 2014. *Cryptosporidium parvum* IId family: clonal population and dispersal from Western Asia to other geographical regions. Sci. Rep. 4, 4208. doi: 10.1038/srep04208

6639-44. doi: 10.1128/aem.01268-10

Widmer, G., Sullivan, S. 2012. Genomics and population biology of *Cryptosporidium* species. Parasite Immunol. 34, 61-71. doi: 10.1111/j.1365-3024.2011.01301.x

Xiao, L. 2010. Molecular epidemiology of cryptosporidiosis: an update. Exp. Parasitol. 124, 80-89. doi: 10.1079/9781845933913.0051

Xiao, L., Ryan, U.M. 2008. Molecular Epidemiology. In: Fayer, R., Xiao, L. (ed) *Cryptosporidium* and Cryptosporidiosis. CRC Press, London, pp 119-171.

Xu, P., Widmer, G., Wang, Y., Ozaki, L.S., Alves, J.M., Serrano, M.G., Puiu, D., Manque, P., Akiyoshi, D., Mackey, A.J., Pearson, W.R., Dear, P.H., Bankier, A.T., Peterson, D.L., Abrahamsen, M.S., Kapur, V., Tzipori, S., Buck, G.A. 2004. The genome of *Cryptosporidium hominis*. Nature 31, 1107-12. doi: 10.1038/nature02977

Zane, L., Bargelloni, L., Patarnello, T. 2002. Strategies for microsatellite isolation: a review. Mol. Ecol. 11, 1-16. doi: 10.1046/j.0962-1083.2001.01418.x

Zalapa, J.E., Cuevas, H., Zhu, H., Steffan, S., Senalik, D., Zeldin, E., McCown, B., Harbut, R. and Simon, P. 2012. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am. J Bot. 99, 193-208. doi: 10.3732/ajb.1100394 challenges of sequencing short tandem repeats using next-generation sequencing platforms. PLoS One 9, e113862. doi: 10.1371/journal.pone.011386

Iowa II	Standard isolate from infected calf	IIaA15G2R1	PRJNA15586	13x*
UKP2	Male child, case from north east England in 2012	IIaA19G1R2	PRJNA253836	51.80x**
UKP3	Female child from north Wales linked to an outbreak involving lamb contact at school in 2013	IIaA18G2R1	PRJNA253840	166.42x**
UKP4	Adult cases from north east England diagnosed during a widespread foodborne outbreak in 2012 (McKerr et al., 2015)	IIaA15G2R1	PRJNA253843	192.48x**
UKP5		IIaA15G2R1	PRJNA253845	26.86x**
UKP6		IIaA15G2R1	PRJNA253846	104.83x**
UKP7	Male child from north west England linked to an outbreak at an open farm in 2013	IIaA17G1R1	PRJNA253847	77.85x**
UKP8	Female adult case from the Midlands of England linked to an outbreak at an open farm in 2013	IIdA22G1	PRJNA253848	174.39x**

* Random shotgun sequencing (Abrahamsen et al., 2004)

** Illumina sequencing reads mapped to *C. parvum* Iowa II (Hadfield et al., 2015)

	tandem repeat regions found in Iowa II ; UKP8	regions meeting selection criteria* in the Iowa II genome (additional repeats in the UKP8 genome)	repeat regions showing variation in copy number of repeats within eight genomes studied (Table 1)
1	194 ; 192	18 (0)	4
2	276 ; 279	26 (0)	6
3	215 ; 212	29 (1)	1
4	312 ; 202	38 (1)	6
5	351 ; 332	24 (3)	3
6	326 ; 282	19 (0)	4
7	227 ; 142	22 (0)	2
8	383 ; 375	34 (3)	2

* ≥6 bp, ≥ 90% similarity among the copies of the repeat, sequence variation limited to the ends of the repeat region

Table 3. Attributes of the selected VNTR loci identified in *Cryptosporidium parvum.* Within coding regions, lo chromosome, gene number and location of the repeat region in bp from the start of the gene, and for non-co the chromosome followed by the label NC, the upstream gene number and location of the repeat region in b gene.

VNTR locus name	Corrected nucleotide sequence 5' to 3'	Length	Coding / Non-coding	% GC content (not including repeat)	C fl
Chromosome 1					
cgd1_470_1429	TC(T/G)GAT[a]	6	Coding	38.2	1
cgd1_3060_604	TCCTCA	6	Coding	34.6	1
cgd1_3170_4182	TGATTCCAATTC	12	Coding	27.4	1
cgd1_3670_5956	GAGCCT[b]	6	Coding	37	1
Chromosome 2					
cgd2_430_451	TCAAGT	6	Coding	45.5	1
cgd2_3300_1504	CATTCTGGTAGGGGAGGA	18	Coding	31.5	1
cgd2_3320_1621	GAACAGGAGCAT	12	Coding	34.5	1
cgd2_3490_2029	TCATCT	6	Coding	39.1	1
cgd2_3550_1474	TCCACTTCTGCT	12	Coding	32.7	1
cgd2_3690_5176	GAAAAGGAGGAGAAAGAG	18	Coding	27.3	1
Chromosome 3					
cgd3_3620_1036	AAAGA(C/T)	6	Coding	24.4	1
Chromosome 4					
cgd4_1340_1681	GGTACTAAAATTAC(C/T)AATACC	21	Coding	20	1
cgd4_2350_796	CC(T/C)GGTATGGG(T/C)CC(A/G)	15	Coding	40.4	U

cgd4_3450_4336	TCTGAA	6	Coding	41.5	1
cgd4_3630_880	CCAAGTAG(C/G)(A/G)CT	12	Coding	45.5	U
cgd4_3940_298	GAAAGCGATTCTGATAGT	18	Coding	25.4	1
cgd4_3970_1525	ATGCCT	6	Coding	30.6	1
Chromosome 5					
cgd5_10_310	GCTCAGGAAGGA[c]	12	Coding	38.2	1
cgd5_NC_3600_3666	CATCATCACCA(A/T)CATCAC	18	Non-Coding	44.1	1
cgd5_4490_2941	CAGAGC	6	Coding	24.1	1
Chromosome 6					
cgd6_530_1561	ACAGGAACA	9	Coding	28.6	1
cgd6_3930_1823	CAGCTCCTC	9	Coding	36.5	U
cgd6_3940_688	ATGCCA[d]	6	Coding	50	U
cgd6_4290_9811	(TCT*/TCC)[e]TCTTCTTCCTCCTCT(TCTTCTTCC/ TCCTCCTCT**)	27	Coding	35.2	1
Chromosome 7					
cgd7_420_4750	(G/A/C)AA(C/G)AA	6	Coding	25.7	1
cgd7_1010_9527	TTGGACAGGGGTGTGGAG	18	Coding	29.7	1
Chromosome 8					
cgd8_NC_ 4440_505	TGAGC(C/T)	6	Non-Coding	41	U
cgd8_NC_4990_360	GGCGG(G/T)CAATTTT	13	Non-Coding	26	1

* present only in first repeat, ** present only in last repeat

a) Previously presented as TTCTGA (Herges et al., 2012)

b) Previously presented as TGAGCC (Ramo et al., 2016a)

c) Reverse complement of MSF (Tanriverdi and Widmer, 2006)

d) Reverse complement, adjusted repeat previously presented as TTGGCA (Ramo et al., 2016a).

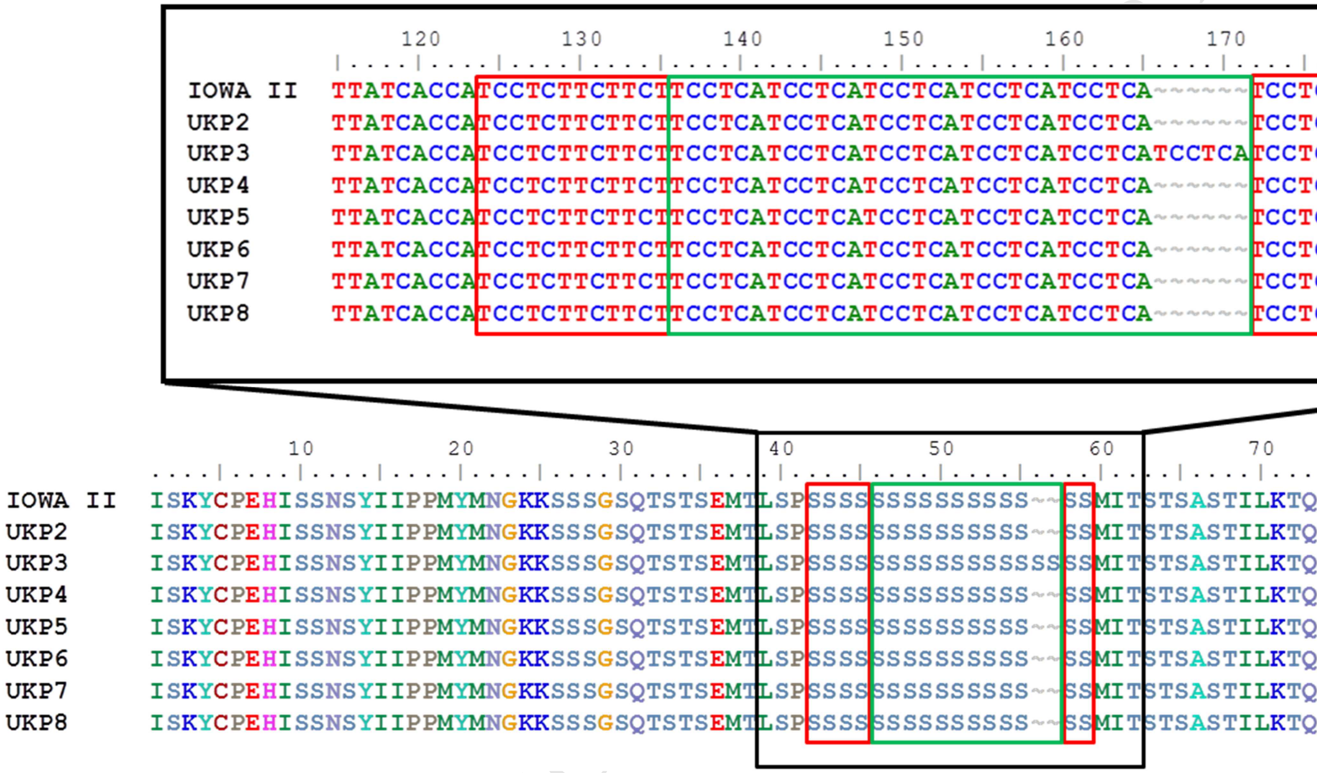
e) Identified previously as MSC6-5 (Xiao and Ryan, 2008)

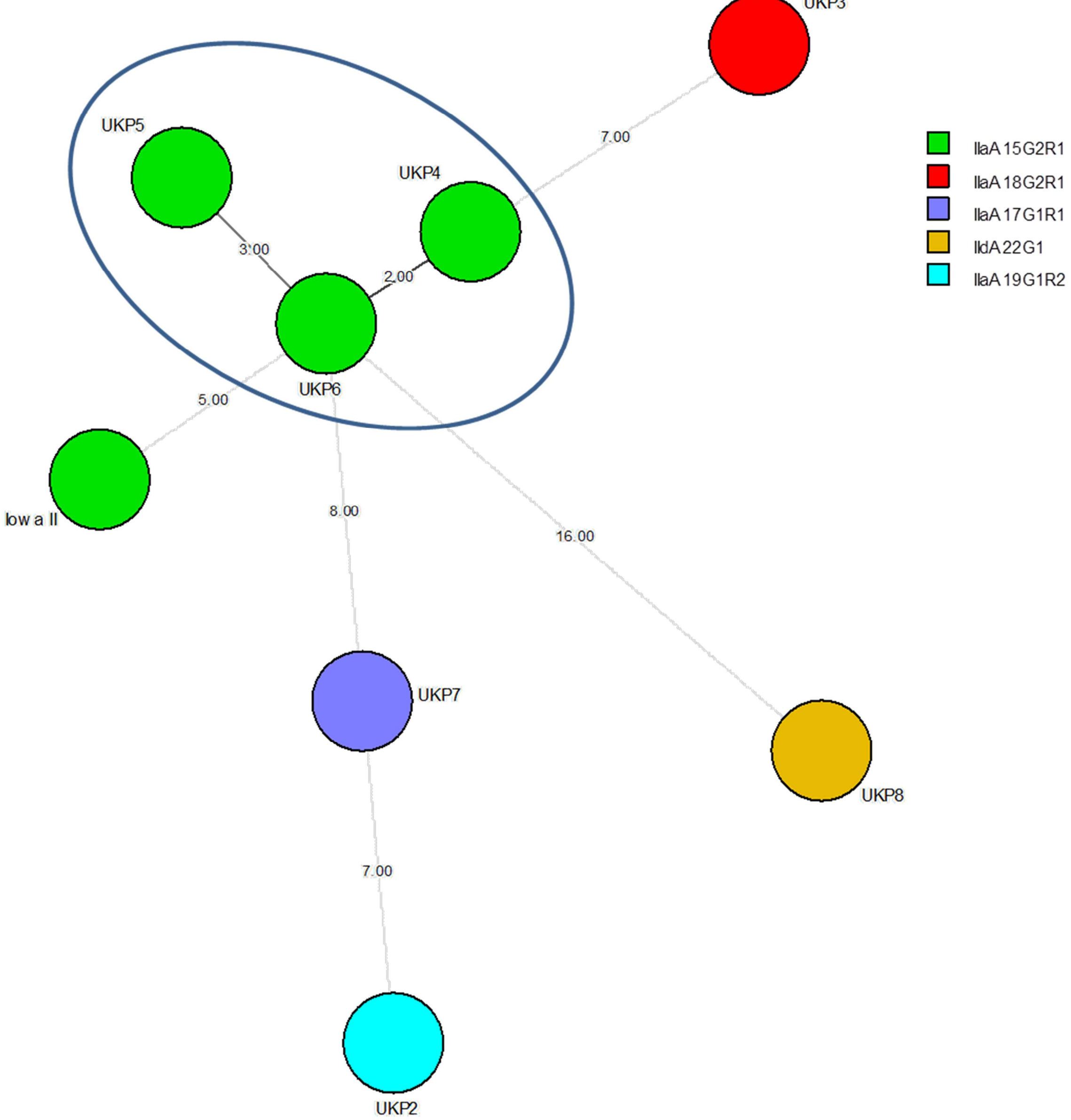
Table 4. Amount of variation within eight *Cryptosporidium parvum* genomes at the qualifying VNTR loci. NF i could be due to either mismatches in the sequence inhibiting the identification of the target sequence or p that locus

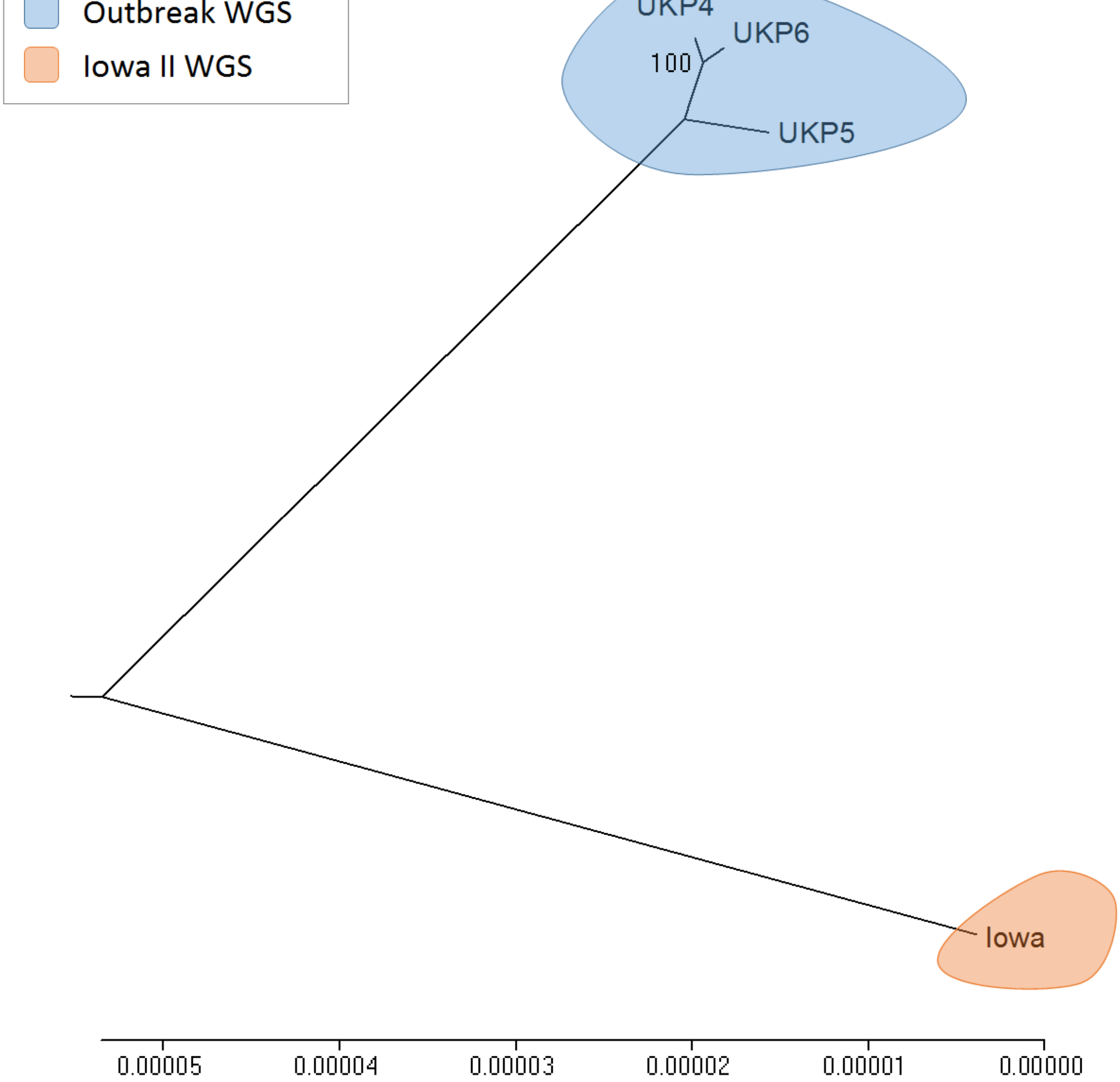
Locus	**Number of repeats (gp60 allele)**						
	Iowa II (IIaA15G2R1)	UKP2 (IIaA19G1R2)	UKP3 (IIaA18G2R1)	UKP4 (IIaA15G2R1)	UKP5 (IIaA15G2R1)	UKP6 (IIaA15G2R1)	UKP7 (IIaA17G1R1)
cgd1_470_1429	7	4	4	4	4	4	4
cgd1_3060_604	5	5	6	5	5	5	5
cgd1_3170_4182	3	3	3	3	3	3	NF
cgd1_3670_5956	5	5	5	5	5	5	5
cgd2_430_451	6	7	6	6	6	6	7
cgd2_3300_1504	3	3	3	3	1	3	3
cgd2_3320_1621	4	4	4	4	4	4	4
cgd2_3490_2029	4	4	4	5	5	5	NF
cgd2_3550_1474	2	1	2	2	2	2	2
cgd2_3690_5176	4	4	4	4	4	4	4
cgd3_3620_1036	7	6	8	8	8	8	6
cgd4_1340_1688	3	3	3	NF	NF	3	NF
cgd4_2350_796	13	6	7	5	9	5	8
cgd4_3450_4336	3	3	3	3	3	3	3
cgd4_3630_880	5	5	5	5	5	5	5

cgd4_3940_298	2	2	2	2	NF	NF	2
cgd4_3970_1525	5	4	4	5	5	5	5
cgd5_10_310	5	5	5	5	5	NF	5
cgd5_NC_3600_3667	2	2	2	2	2	2	2
cgd5_4490_2941	8	NF	NF	NF	6	NF	7
cgd6_530_1561	3	3	3	3	3	3	3
cgd6_3930_1823	2	2	3	2	2	2	3
cgd6_3940_688	11	11	NF	13	11	11	11
cgd6_4290_9811	3	1	3	3	3	3	2
cgd7_420_4750	6	6	6	6	6	6	7
cgd7_1010_9527	2	2	2	2	2	2	2
cgd8_NC_ 4440_506	30	18	16	18	17	14	19
cgd8_NC_4990_361	2	2	2	3	3	3	2

Figure 1. The true cgd1_3060_604 repeat region (green box) occurring in eight Cryptosporidium parvum isola repeats (each translated to two serine repeats), but flanked by similar TCCTCT or TCTTCT repeats (red boxes) repeats) that are not included as part of the repeat region.




UKP3
UKP5
UKP4
7.00
3.00
2.00
UKP6
5.00
Iow a II
8.00
16.00
UKP7
UKP8
7.00
UKP2
IIaA 15G2R1
IIaA 18G2R1
IIaA 17G1R1
IIdA 22G1
IIaA 19G1R2


Outbreak WGS
Iowa II WGS
UKP4
UKP6
100
UKP5
Iowa
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000

- *In silico* analysis of qualifying loci was performed with eight *C. parvum* genomes.
- Multilocus discrimination was high even between closely related isolates.