



Swansea University  
Prifysgol Abertawe



## Cronfa - Swansea University Open Access Repository

---

This is an author produced version of a paper published in :

*Experimental Parasitology*

Cronfa URL for this paper:

<http://cronfa.swan.ac.uk/Record/cronfa29586>

---

### 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>

---

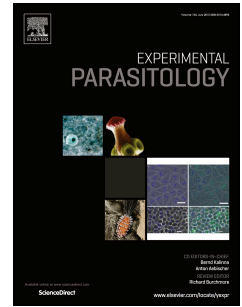
This article is brought to you by Swansea University. Any person downloading material is agreeing to abide by the terms of the repository licence. Authors are personally responsible for adhering to publisher restrictions or conditions. When uploading content they are required to comply with their publisher agreement and the SHERPA RoMEO database to judge whether or not it is copyright safe to add this version of the paper to this repository.

<http://www.swansea.ac.uk/iss/researchsupport/cronfa-support/>

# Accepted Manuscript

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](https://doi.org/10.1016/j.exppara.2016.08.003)

Reference: YEXPR 7285

To appear in: *Experimental Parasitology*

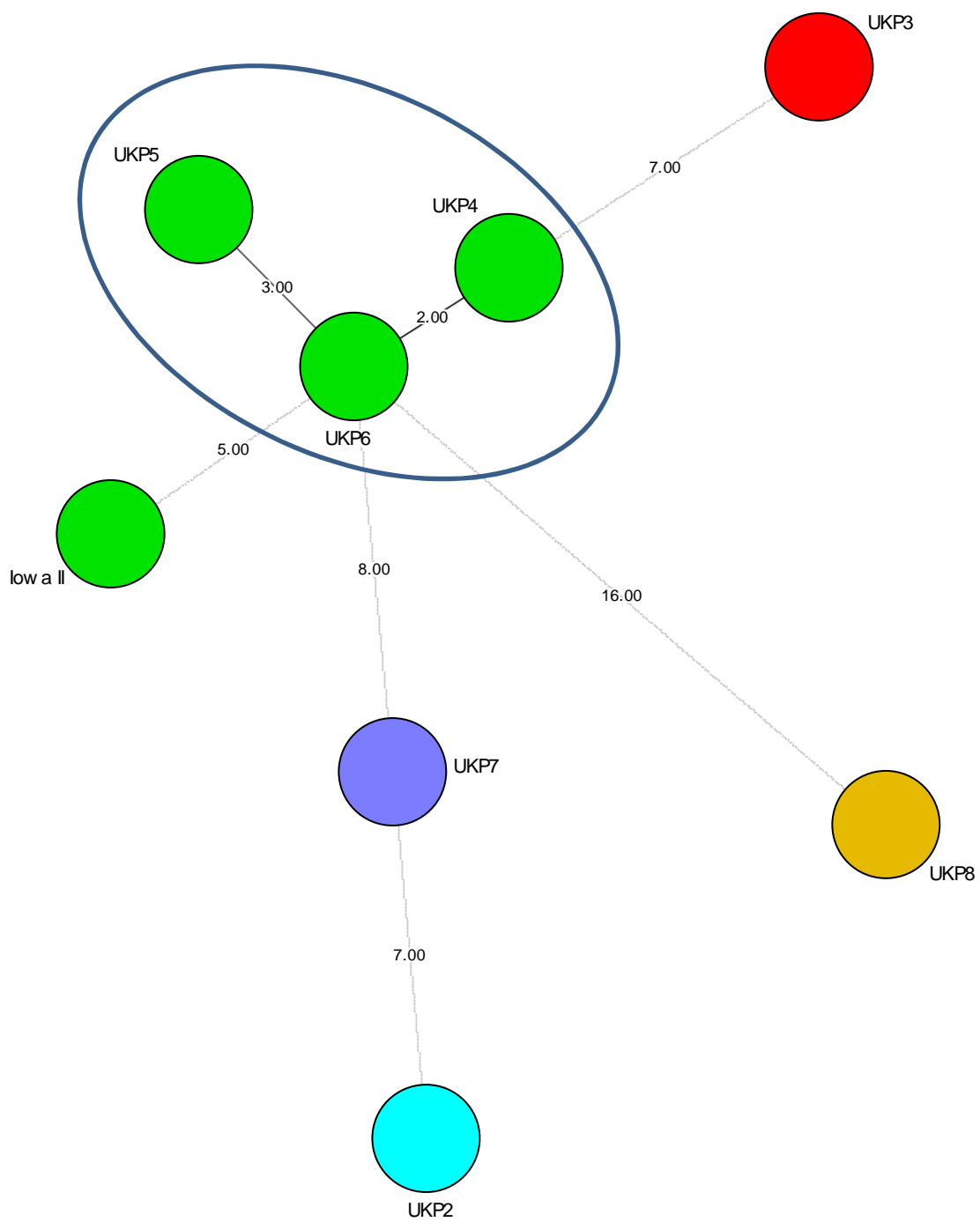
Received Date: 26 May 2016

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.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



3

4 Gregorio Pérez-Cordón<sup>a</sup>, Guy Robinson<sup>a,b</sup>, Johanna Nader<sup>c</sup> and Rachel M Chalmers<sup>a,b\*</sup>

5

6 <sup>a</sup> Cryptosporidium Reference Unit, Public Health Wales Microbiology, Singleton Hospital,

7 Swansea, SA2 8QA, UK

8 <sup>b</sup> Swansea University Medical School, Grove Building, Swansea University, Singleton Park,

9 Swansea, SA2 8PP, UK

10 <sup>c</sup> Norwich Medical School, University of East Anglia, Norwich, UK

11

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

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

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

29

### 30 **Keywords**

31 *Cryptosporidium parvum*, variable number tandem repeats, multilocus

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

44  
45 The life cycle of *Cryptosporidium* involves both asexual and sexual reproduction and genetic  
46 recombination has been demonstrated experimentally in *C. parvum* (Feng et al., 2002).  
47 Therefore, it is feasible to suppose that recombination between different genotypes occurs  
48 in nature giving rise not only to new genotypes but also to heterogeneous populations,  
49 although the scale of occurrence within hosts is not known as many genotyping methods  
50 lack sensitivity for their detection (Grinberg and Widmer, 2016). *Cryptosporidium parvum*  
51 genotypes have traditionally been identified based on sequence analysis of the gp60 gene,  
52 in which variable numbers of tandem serine codons as well as downstream polymorphisms  
53 differentiate subtypes (Strong et al., 2000).

54

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

73

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

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



105

## 106 **2. Methods**

### 107 2.1. Identification of variable VNTR loci and their attributes

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

121

122 The repeat size, sequence and copy number, gene name and chromosome location, GC  
123 content and conservation of the sequences flanking the repeat units of the remaining repeat  
124 regions was recorded. The corresponding loci within seven other *C. parvum* whole genomes  
125 (UKP2 through to UKP8; Table 1) published previously (Hadfield et al., 2015) and obtained  
126 from the umbrella BioProject PRJNA215218 on the NCBI database

129 were edited to include only the VNTR and immediate flanking regions, the orientation  
130 checked and the validity of coding sequences and reading frames identified in the *C. parvum*  
131 Iowa II reference genome on CryptoDB. The true repeat units were identified by checking  
132 that repeats in coding regions were represented by whole codons in the correct interval  
133 from the methionine start codon. Motifs similar to the true repeat that consistently flanked  
134 the VNTR units without variation were not included in the definition of the repeat region; an  
135 example is shown in Figure 1. Only those loci that displayed variations in the number of  
136 repeats in the eight aligned isolates were included the final selection.

137

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

144

## 145 2.2. Literature and database search

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

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

162

### 163 2.3. Bioinformatic analyses

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

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

179

### 180 **3. Results**

#### 181 3.1. Identification of variable VNTR loci and their attributes

182 A total of 2284 tandem repeat loci were identified initially in the *C. parvum* Iowa II reference  
183 genome, but after rejecting 2074 loci with repeats of < 6 bp or showing < 90 % similarity  
184 among the copies of the repeat, and 182 loci that showed no variation in copy number  
185 within the other seven genomes, 28 remained for further examination (Table 2).

186 Interrogating the UKP8 genome, 2016 loci were identified initially, but after applying our  
187 selection criteria and removing duplicates identified initially in the IOWA II genome, eight  
188 additional loci remained. However, those eight were also rejected as they showed no  
189 variation in copy number within the other genomes investigated (Table 2).

190

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

201

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

205

206 The majority of VNTR sequences in the qualifying loci were non-polymorphic (18/28),  
207 especially those found in chromosome 2 where there was no sequence variation within the  
208 six repeat units. Twenty five of the 28 qualifying loci were coding, and the most common  
209 repeat unit length was 6 bp and the longest was 27 bp (cgd6\_4290\_9811) (Table 3). The  
210 three non-coding loci (one on chromosome 5 and the two on chromosome 8) were 6, 13  
211 and 18 bp in length. With the intention of developing *in vitro* assays and designing PCR  
212 primers, we looked at the GC % content as well as the conservation of the sequences  
213 flanking the repeat region. In all cases, the GC content was  $\leq 50\%$  and all but 5 qualifying  
214 loci showed 100 % conservation of flanking sequences upstream and downstream of the  
215 repeat region (Table 3).

216

217 Of the 28 qualifying loci, 19 were found in all eight genomes interrogated. The non-detects  
218 occurred mostly as singles (six loci) but three loci (cgd4\_3940\_298, cgd4\_1340\_1688, and  
219 cgd5\_4490\_2941) were not detected in two, three and four genomes respectively (Table 4).

220 The number of alleles identified for each qualifying locus in the eight genomes investigated  
221 varied between two (21 loci), three (4 loci), four (one locus), five (one locus) and seven (one  
222 locus, cgd8\_NC\_4440\_505) (Table 4). Of the 19 loci found in all eight genomes, eight

225 Ila genomes but could not separate IId (Table 4).

226

### 227 3.2. Literature and database search

228 Of the 55 VNTR loci reviewed by Robinson and Chalmers (2012), 18 were  $\geq 6$  bp, but only  
229 MSF (Tanriverdi and Widmer, 2006) was selected by our criteria for further examination  
230 (cgd5\_10\_310, Table 3). The remaining 17 were not included as six showed  $< 90$  % similarity  
231 among the repeat copies and in 11 the variation was distributed throughout the repeat  
232 region. One locus overlooked previously, MSC6-5 (Xiao and Ryan, 2008), was also selected  
233 through our process (cgd6\_4290\_9811, Table 3).

234

235 A total of 35 new publications were identified using the search terms defined in PubMed  
236 within the time period considered, of which 19 were considered relevant. Only three of  
237 these reported “new” loci. Herges et al. (2012) described the GRH locus, detected with the  
238 same TRF software that we used, identified in our study as cgd1\_470\_1429 (Table 3) with  
239 the repeat re-defined based on the correct reading frame. The two others (Ramo et al.,  
240 2016a and 2016b) included four previously un-described VNTR loci. We found all four in our  
241 initial screening of the *C. parvum* Iowa II genome, and two qualified in our analysis (Table 3),  
242 although again we defined the repeat sequences differently, based on their DNA codons in  
243 the correct open reading frame. Additionally, one was translated from the antisense strand  
244 (Table 3). Two were rejected (cgd2\_3850 and cgd6\_5400) as they presented  $< 90$  %  
245 similarity throughout the repeat regions.

246

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

253

### 254 3.3. Bioinformatic analyses

255 All eight isolates were differentiated *in silico* by MLVA using all 28 loci (Figure 2). In fact the  
256 minimum number of loci required to differentiate all eight isolates was two  
257 (cgd8\_NC\_4440\_506 and any one of eight others, the most discriminatory being  
258 cgd4\_2350\_796, Table 4).

259

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

264

## 265 4. Discussion

266 The clinical and economic impact of cryptosporidiosis demands the development of  
267 strategies for improved surveillance and control including the ability to investigate, through  
268 genotyping, sources of contamination and routes of transmission in a fast and reliable way.  
269 The availability of seven new *C. parvum* genomes (Hadfield et al., 2015), in addition to the  
270 reference Iowa II genome, allowed us to perform an *in silico* analysis of new potential VNTR

273 sequences (Zane, 2002), and was fruitful in our analysis; of the 28 qualifying loci identified,  
274 23 were new and just 5 had been identified previously.

275

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

280

281 Although the accuracy of NGS may be challenged by homopolymers, one study reported the  
282 acceptable identification of short tandem repeats, present in the yeast *Saccharomyces*  
283 *cerevisiae* in copy numbers of a similar order of magnitude to those in our whole genome  
284 sequences (Zavodna et al., 2014). The depth of coverage of the genome sequencing was  
285 identified as being important, but cannot alone resolve assembly gaps caused by repetitive  
286 regions with lengths that approach or exceed those of the short NGS reads (Sims et al.,  
287 2014). The required average mapped depth to allow reliable calling of SNPs and small indels  
288 across 95 % of the genome has reduced from 50x to 35x due to improvements in sequencing  
289 chemistry reducing GC bias and yielding a more uniform coverage (Sims et al., 2014). The  
290 overall range of coverage of the *C. parvum* genomes in our study ranged from 26.86x to  
291 192.48x (mean 113.52x ) for the UKP genomes (Hadfield et al., 2015) and 13x for Iowa II  
292 (Abrahamsen et al., 2004) (Table 1). We therefore considered that using the genome  
293 sequences not only allowed us to locate and describe the VNTRs, but also compare the  
294 outputs and outcomes phylogenetically.



297 in coding regions. It is likely that selection pressure for sequence conservation drove the  
298 occurrence of homogeneous repeats mainly in coding regions (Madesis et al., 2013). The use  
299 of only perfect non-polymorphic repeats for MLVA was recommended by Nadon et al.  
300 (2013), but to identify these it was necessary to loosen the parameters to include those  
301 repeats with  $\geq 90\%$  similarity before manually determining the true repeat, as flanking  
302 sequences similar to the repeat would sometimes stop the software from returning some of  
303 the results when set to 100%. For example, TCA TCA TCT would not return if set to 100%,  
304 because the TCT unit would be counted as part of the repeat, even if it was consistently  
305 present and non-variable. While this undoubtedly resulted in the loss of a number of VNTR  
306 loci that may indeed be useful, the objective was not to identify all of the tandem repeats  
307 present, but to identify new suitable candidates that could be examined further to develop  
308 a robust typing scheme. The 90% cut-off was not pre-determined, but selected arbitrarily  
309 based on the number of initial results that it returned (210 before assessing the spread of  
310 variation throughout the region and discrimination with the other *C. parvum* isolates).  
311 Additionally, we made the assumption that loci with the highest similarity between repeat  
312 copies would be more robust in a typing method.

313

314 Most repeat units were short (6 bp) but some longer ones were identified, up to 27 bp, but  
315 there didn't appear to be any major significance associated with the length of the repeat  
316 and potential for discrimination, although this is probably due to only 2 or 3 alleles being  
317 found at most loci (25/28). The most discriminatory locus was the 6 bp repeat  
318 cgd8\_NC\_4440\_506 that separated the 8 genomes into 7 different alleles, but the second

321 number of alleles within the maximum fragment size requirements for a multi-platform  
322 scheme. For example, when the two most discriminatory loci are compared in the Iowa II  
323 genome, cgd8\_NC\_4440\_506 had 30 copies of the repeat opposed to the 13 copies of  
324 cgd4\_3450\_4336, but the latter is at the top end of the preferred size range (< 300 bp  
325 including 50 bp flanking regions for primer annealing) because each copy is 15 bp. The  
326 advantage however, with longer repeats is the easier separation of alleles based on  
327 fragment size as variation in the sizing is less likely to overlap with the next allele size.  
328

329 The distribution of qualifying loci was across all chromosomes, with the number per  
330 chromosome ranging from one (chromosome 3) to six (chromosomes 2 and 4) (Table 3). The  
331 selection of loci for a scheme based on the diversity and spread across different  
332 chromosomes is particularly important in *Cryptosporidium* due to the potential for  
333 recombination during the sexual stage of the life cycle (Widmer & Sullivan, 2012). While a  
334 spread of loci across chromosomes is required, a representative from each chromosome is  
335 not necessary, because the aim would be to identify a multilocus method providing good  
336 resolution but with the smallest number of markers (Widmer & Sullivan, 2012).  
337

338 Of the 28 qualifying loci, nine were not detected in one or more of the whole genomes  
339 (Table 4). There could be a few explanations for this including, mismatches in the sequence  
340 inhibiting the identification of the target sequence, poor coverage of the genome at that  
341 particular locus or a true absence of the repeat in that isolate. The locus that had the most  
342 non-detects (cgd5\_4490\_2941) only identified alleles in half of the genomes. However, each

345 was not detected in half of the genomes and whether following primer design to specifically  
346 target it can the VNTR be detected in all isolates.

347

348 The eight *C. parvum* genomes investigated comprised two *C. parvum* gp60 families (IIa and  
349 IIId) which are prevalent in both humans and animals worldwide (Wang et al., 2014). While  
350 the gp60 marker does provide relatively good discrimination between isolates of *C. parvum*,  
351 it, along with other single loci, does not serve as a surrogate for other loci or multilocus  
352 genotypes (Widmer and Lee, 2010). However, as these data were readily available for each  
353 of our genomes, the gp60 genotype of each isolate could provide some initial indication to  
354 differences between isolates for comparative purposes with the newly identified alleles.  
355 Eight of our loci could only differentiate the two gp60 families IIa and IIId (Table 4), whereas  
356 the remaining candidate VNTR loci allowed for some intra-gp60 family discrimination (e.g.  
357 cgd1\_3060\_604 with two alleles or cgd8\_NC\_4440\_505 with seven alleles) sometimes in  
358 addition to family discrimination. Although within-host populations of *Cryptosporidium* are  
359 likely to be genetically diverse (Grinberg and Widmer, 2016), MLVA has the potential to  
360 identify these mixed populations. The genome sequences interrogated in our study were  
361 reported to show no evidence of being mixed species (Hadfield et al., 2015), but from the  
362 sequence data alone we cannot be certain that there are no mixed populations of *C. parvum*  
363 genotypes present.

364

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

369 Hotchkiss et al. 2015). Consideration of the hosts likely to be investigated is important; in  
370 Spain and the UK, gp60 family IIa is more common in cattle (Quilez et al., 2008a; Hotchkiss  
371 et al., 2015) and IIc in sheep and goats (Quilez et al., 2008b), so loci such as cgd5\_10\_310  
372 (MSF, Tanriverdi and Widmer, 2006) would be less informative in cattle isolates compared  
373 to sheep and goats (Hotchkiss et al., 2015). Indeed, in a recent study by Ramo et al. (2016b)  
374 two VNTR loci that were previously used for intra-species typing in cattle and showed to be  
375 poorly discriminatory (Ramo et al., 2016a) were among the most informative for typing in  
376 sheep. Due to the prior selection of samples for whole genome sequencing (Hadfield et al.,  
377 2015), only one of the genomes analysed in our study was IIc, whereas the other seven  
378 genomes were IIa, which may have resulted in selection bias towards loci that are more  
379 variable in IIa. Further testing *in vitro* of a larger, varied panel of isolates is required to  
380 provide more detailed information about the discriminatory capabilities of the qualifying  
381 loci. For example, Herges et al. (2012) identified 10 different GRH (syn. cgd1\_470\_1429 in  
382 this study) alleles in 254 *C. parvum* isolates from humans and cattle, second in  
383 discrimination only to gp60 with 22 alleles. There remains a need for more *Cryptosporidium*  
384 whole genomes to be published ideally from different sources to increase the amount of  
385 potential variation and allow us to make less biased comparisons. The number of available  
386 genomes is increasing (Andersson et al., 2015; Hadfield et al., 2015; Guo et al., 2015) and  
387 mining their data will help further in the development of efficient MLVA schemes.

388

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

393 (2013), linkage equilibrium was observed in the gp60 subtype IIaA15G2R1 group but not in  
394 the non-IIaA15G2R1 group, indicating the possible presence of genetic recombination and  
395 maybe explaining the variation at other loci within the IIaA15G2R1 gp60 genotype.  
396 However, in our study the *cgd2\_3300\_1504* and *cgd4\_2350\_796* loci only differed between  
397 the three isolates in UKP5 and it is a possibility that this variation could be due to  
398 inaccuracies in the UKP5 sequence assembly as the depth of coverage was only 26.86x.  
399 Another potential for inaccuracy in two of the loci (*cgd4\_2350\_796* and  
400 *cgd8\_NC\_4440\_505*) is that the repeats are approaching the size of the raw NGS reads,  
401 which as described above can make it hard to resolve assembly gaps in these regions (Sims  
402 et al., 2014). Testing these three isolates with carefully designed PCR assays at these loci  
403 would help resolve whether these isolates are indeed different from each other. It is also  
404 possible that in outbreaks, especially ones where there must have been a high degree of  
405 contamination to cause geographically widespread illness in >300 confirmed cases (McKerr  
406 et al., 2015), mixed populations of oocysts may have caused the infections resulting in  
407 differing allelic profiles. Alternatively, the cases may not have been linked by a common  
408 exposure to the source of the outbreak and may have been background, unlinked cases. A  
409 comparison by MST with all 28 loci and phylogenetic analysis of the whole genome both  
410 suggested that although slightly different, the outbreak samples cluster together separately  
411 from the Iowa II isolate (Figures 2 and 3). This suggests that with careful selection of loci,  
412 MLVA may serve as a surrogate to whole genome analysis when studying relationships  
413 between epidemiological relevant isolates with clear cost-saving benefits. In addition to  
414 cost, whole genome sequencing of *Cryptosporidium* is also hindered by the non-culturable

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

418

## 419 **Conclusions**

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

430

## 431 **Acknowledgments**

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

435

## 436 **References**

437 Abrahamsen, M.S., Templeton, T.J., Enomoto, S., Abrahante, J.E., Zhu, G., Lancto, C.A., Deng,  
438 M., Liu, C., Widmer, G., Tzipori, S., Buck, G.A., Xu, P., Bankier, A.T., Dear, P.H., Konfortov, .A.,

441 10.1126/science.1094786

442

443 Andersson, S., Sikora, P., Karlberg, M.L., Winiiecka-Krusnell, J., Alm, E., Beser, J., Arrighi, R.B.

444 2015. It's a dirty job--A robust method for the purification and de novo genome assembly of

445 *Cryptosporidium* from clinical material. J. Microbiol. Methods. 113, 10-2. doi:

446 10.1016/j.mimet.2015.03.018

447

448 Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic

449 Acids Res. 27, 573-80. doi: 10.1093/nar/27.2.573

450

451 Chalmers, R.M. 2012. Waterborne outbreaks of cryptosporidiosis. Ann. Ist. Super Sanita. 48,

452 429-46. doi: 10.4415/ANN\_12\_04\_10

453

454 Chalmers, R.M., Robinson, G., Hotchkiss, E., Alexander, C., May, S., Gilray, J., Connelly, L.,

455 Hadfield, S.J. 2016. Suitability of loci for multiple-locus variable-number of tandem-repeats

456 analysis of *Cryptosporidium parvum* for inter-laboratory surveillance and outbreak

457 investigations. Parasitology 2, 1-11. doi: 10.1017/S0031182015001766

458

459 Dallman, T.J., Byrne, L., Ashton, P.M., Cowley, L.A., Perry, N.T., Adak, G., Petrovska, L., Ellis

460 R.J., Elson, R., Underwood, A., Green, J., Hanage, W.P., Jenkins, C., Grant, K., Wain, J. 2015.

461 Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia*

462 *coli* O157. Clin. Infect. Dis. 61, 305-12. doi: 10.1093/cid/civ318.

465 recombination in the opportunistic pathogen *Cryptosporidium parvum*. Mol. Biochem.  
466 Parasitol. 119, 55-62. doi: 10.1016/S0166-6851(01)00393-0  
467  
468 Feng, Y., Yang, W., Ryan, U., Zhang, L., Kvac, M., Koudela, B., Modry, D., Li, N, Fayer, R, Xiao,  
469 L. 2011. Development of a multilocus sequence tool for typing *Cryptosporidium muris* and  
470 *Cryptosporidium andersoni*. J. Clin. Microbiol. 49, 34-41. doi: 10.1128/JCM.01329-10  
471  
472 Feng, Y., Torres, E., Li, N., Wang, L., Bowman, D., Xiao, L. 2013. Population genetic  
473 characterisation of dominant *Cryptosporidium parvum* subtype IIaA15G2R1. Int. J. Parasitol.  
474 43, 1141-7. doi: 10.1016/j.ijpara.2013.09.002  
475  
476 Grinberg., A., Widmer., G. 2016. *Cryptosporidium* within-host genetic diversity: systematic  
477 bibliographical search and narrative overview. Int. J. Parasitol. Published online ahead of  
478 print. doi: 10.1016/j.ijpara.2016.03.002  
479  
480 Guo, Y., Li, N., Lysén, C., Frace, M., Tang, K., Sammons, S., Roellig, D.M., Feng, Y., Xiao, L.  
481 2015. Isolation and enrichment of *Cryptosporidium* DNA and verification of DNA purity for  
482 whole-genome sequencing. J. Clin. Microbiol. 53, 641-7. doi: 10.1128/JCM.02962-14  
483  
484 Hadfield, S.J., Pachebat, J.A., Swain, M.T., Robinson, G., Cameron, S.J., Alexander, J.,  
485 Hegarty, M.J., Elwin, K., Chalmers, R.M. 2015. Generation of whole genome sequences of



488

489 Herges, G.R, Widmer, G., Clark, M.E., Khan, E., Giddings, C.W., Brewer, M., McEvoy, J.M.  
490 2012. Evidence that *Cryptosporidium parvum* populations are panmictic and unstructured in  
491 the Upper Midwest of the United States. *Appl. Environ. Microbiol.* 78: 8096-101. doi:  
492 10.1128/AEM.02105-12

493

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

498

499 Li, N., Xiao, L., Cama, V.A., Ortega, Y., Gilman, R.H., Guo, M., Feng, Y. 2013. Genetic  
500 recombination and *Cryptosporidium hominis* virulent subtype IbA10G2. *Emerg. Infect. Dis.*  
501 19, 1573-82. doi: 10.3201/eid1910.121361

502

503 Li, N., Xiao, L., Alderisio, K., Elwin, K., Cebelinski, E., Chalmers, R., Santin, M., Fayer, R., Kvac,  
504 M., Ryan, U., Sak, B., Stanko, M., Guo, Y., Wang, L., Zhang, L., Cai, J., Roellig, D., Feng, Y.  
505 2014. Subtyping *Cryptosporidium ubiquitum*, a zoonotic pathogen emerging in humans.  
506 *Emerg. Infect. Dis.* 20, 217-24. doi: 10.3201/eid2002.121797

507

511

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

514

515 McKerr, C., Adak, G.K., Nichols, G., Gorton, R., Chalmers, .RM., Kafatos, G., Cosford, P.,  
516 Charlett, A., Reacher, M., Pollock, K.G., Alexander, C.L., Morton, S. 2015. An Outbreak of  
517 *Cryptosporidium parvum* across England & Scotland associated with consumption of fresh  
518 pre-cut salad leaves, May 2012. PLoS One. 10, e0125955. doi:  
519 10.1371/journal.pone.0125955

520

521 Nadon, C.A., Trees, E., Ng, L.K., Møller Nielsen, E., Reimer, A., Maxwell, N., Kubota, K.A.,  
522 Gerner-Smidt, P. 2013. MLVA Harmonization Working Group. Development and application  
523 of MLVA methods as a tool for inter-laboratory surveillance. Euro Surveill. 18, 20565. doi:  
524 10.2807/1560-7917.es2013.18.35.20565

525

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

528

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

531

534 1613-1620. doi: 10.1017/s0031182008005088

535

536 Quílez, J., Torres, E., Chalmers, R.M., Hadfield, S.J., Del Cacho, E., Sánchez-Acedo, C. 2008b.

537 *Cryptosporidium* genotypes and subtypes in lambs and goat kids in Spain. Appl. Environ.

538 Microbiol. 74, 6026-31. doi: 10.1128/aem.00606-08

539

540 Ramo, A., Quílez, J., Monteagudo, L., Del Cacho, E., Sánchez-Acedo, C., 2016a. Intra-Species

541 Diversity and Panmictic Structure of *Cryptosporidium parvum* Populations in Cattle Farms in

542 Northern Spain. PLoS One. 11, e0148811. doi: 10.1371/journal.pone.0148811

543

544 Ramo, A., Monteagudo, L.V., Del Cacho, E., Sánchez-Acedo, C., Quílez, J. 2016b. Intra-

545 Species Genetic Diversity and Clonal Structure of *Cryptosporidium parvum* in Sheep Farms in

546 a Confined Geographical Area in Northeastern Spain. PLoS One. 11, e0155336. doi:

547 10.1371/journal.pone.0155336.

548

549 Robinson, G., Chalmers, R.M. 2012. Assessment of polymorphic genetic markers for multi-

550 locus typing of *Cryptosporidium parvum* and *Cryptosporidium hominis*. Exp. Parasitol. 132,

551 200-15. doi: 10.1016/j.exppara.2012.06.016

552

553 Shirley, D.A., Moonah, S.N., Kotloff, K.L. 2012. Burden of disease from cryptosporidiosis.

554 Curr. Opin. Infect. Dis. 25, 555-63. doi: 10.1097/qco.0b013e328357e569

555

558 10.1038/nrg3642.

559

560 Strong, W.B., Gut, J., Nelson, R.G. 2000. Cloning and sequence analysis of a highly  
561 polymorphic *Cryptosporidium parvum* gene encoding a 60-kilodalton glycoprotein and  
562 characterization of its 15- and 45-kilodalton zoite surface antigen products. *Infect. Immun.*  
563 68, 4117-34. doi: 10.1128/iai.68.7.4117-4134.2000

564

565 Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: Molecular  
566 Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–9. doi:  
567 10.1093/molbev/mst197

568

569 Tanriverdi, S., Widmer, G. 2006. Differential evolution of repetitive sequences in  
570 *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Infect. Genet. Evol.* 6, 113-22. doi:  
571 10.1016/j.meegid.2005.02.002

572

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

577

580 6639-44. doi: 10.1128/aem.01268-10

581

582 Widmer, G., Sullivan, S. 2012. Genomics and population biology of *Cryptosporidium* species.

583 Parasite Immunol. 34, 61-71. doi: 10.1111/j.1365-3024.2011.01301.x

584

585 Xiao, L. 2010. Molecular epidemiology of cryptosporidiosis: an update. Exp. Parasitol. 124,

586 80-89. doi: 10.1079/9781845933913.0051

587

588 Xiao, L., Ryan, U.M. 2008. Molecular Epidemiology. In: Fayer, R., Xiao, L. (ed)

589 *Cryptosporidium* and Cryptosporidiosis. CRC Press, London, pp 119-171.

590

591 Xu, P., Widmer, G., Wang, Y., Ozaki, L.S., Alves, J.M., Serrano, M.G., Puiu, D., Manque, P.,

592 Akiyoshi, D., Mackey, A.J., Pearson, W.R., Dear, P.H., Bankier, A.T., Peterson, D.L.,

593 Abrahamsen, M.S., Kapur, V., Tzipori, S., Buck, G.A. 2004. The genome of *Cryptosporidium*

594 *hominis*. Nature 31, 1107-12. doi: 10.1038/nature02977

595

596 Zane, L., Bargelloni, L., Patarnello, T. 2002. Strategies for microsatellite isolation: a review.

597 Mol. Ecol. 11, 1-16. doi: 10.1046/j.0962-1083.2001.01418.x

598

599 Zalapa, J.E., Cuevas, H., Zhu, H., Steffan, S., Senalik, D., Zeldin, E., McCown, B., Harbut, R. and

600 Simon, P. 2012. Using next-generation sequencing approaches to isolate simple sequence

601 repeat (SSR) loci in the plant sciences. Am. J Bot. 99, 193-208. doi: 10.3732/ajb.1100394

604 challenges of sequencing short tandem repeats using next-generation sequencing platforms.

605 PLoS One 9, e113862. doi: 10.1371/journal.pone.011386

ACCEPTED MANUSCRIPT

Iowa II	Standard isolate from infected calf	IlaA15G2R1	PRJNA15586	13x <sup>*</sup>
UKP2	Male child, case from north east England in 2012	IlaA19G1R2	PRJNA253836	51.80x <sup>**</sup>
UKP3	Female child from north Wales linked to an outbreak involving lamb contact at school in 2013	IlaA18G2R1	PRJNA253840	166.42x <sup>**</sup>
UKP4	Adult cases from north east England diagnosed during a	IlaA15G2R1	PRJNA253843	192.48x <sup>**</sup>
UKP5	widespread foodborne outbreak in	IlaA15G2R1	PRJNA253845	26.86x <sup>**</sup>
UKP6	2012 (McKerr et al., 2015)	IlaA15G2R1	PRJNA253846	104.83x <sup>**</sup>
UKP7	Male child from north west England linked to an outbreak at an open farm in 2013	IlaA17G1R1	PRJNA253847	77.85x <sup>**</sup>
UKP8	Female adult case from the Midlands of England linked to an outbreak at an open farm in 2013	IIdA22G1	PRJNA253848	174.39x <sup>**</sup>

---

\* 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 lowa II ; UKP8	regions meeting selection criteria * in the lowa 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

\*  $\geq 6$  bp,  $\geq 90\%$  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, loci are identified by chromosome, gene number and location of the repeat region in bp from the start of the gene, and for non-coding regions, the chromosome followed by the label NC, the upstream gene number and location of the repeat region in bp from the start of the gene.**

VNTR locus name	Corrected nucleotide sequence 5' to 3'	Length	Coding / Non-coding	% GC content (not including repeat)	Chromosome
<b>Chromosome 1</b>					
cgd1_470_1429	TC(T/G)GAT <sup>a</sup>	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 <sup>b</sup>	6	Coding	37	1
<b>Chromosome 2</b>					
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	TCCAATTCTGCT	12	Coding	32.7	1
cgd2_3690_5176	GAAAAGGAGGAGAAAGAG	18	Coding	27.3	1
<b>Chromosome 3</b>					
cgd3_3620_1036	AAAGA(C/T)	6	Coding	24.4	1
<b>Chromosome 4</b>					
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
<b>Chromosome 5</b>					
cgd5_10_310	GCTCAGGAAGGA <sup>c</sup>	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
<b>Chromosome 6</b>					
cgd6_530_1561	ACAGGAACA	9	Coding	28.6	1
cgd6_3930_1823	CAGCTCCTC	9	Coding	36.5	U
cgd6_3940_688	ATGCCA <sup>d</sup>	6	Coding	50	U
cgd6_4290_9811	(TCT*/TCC) <sup>e</sup> TCTTCTTCCTCCTCT(TCTTCTTCC/ TCCTCCTCT**)	27	Coding	35.2	1
<b>Chromosome 7</b>					
cgd7_420_4750	(G/A/C)AA(C/G)AA	6	Coding	25.7	1
cgd7_1010_9527	TTGGACAGGGGTGTGGAG	18	Coding	29.7	1
<b>Chromosome 8</b>					
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 indicates that the locus could not be identified. The variation could be due to either mismatches in the sequence inhibiting the identification of the target sequence or polymorphisms at that locus**

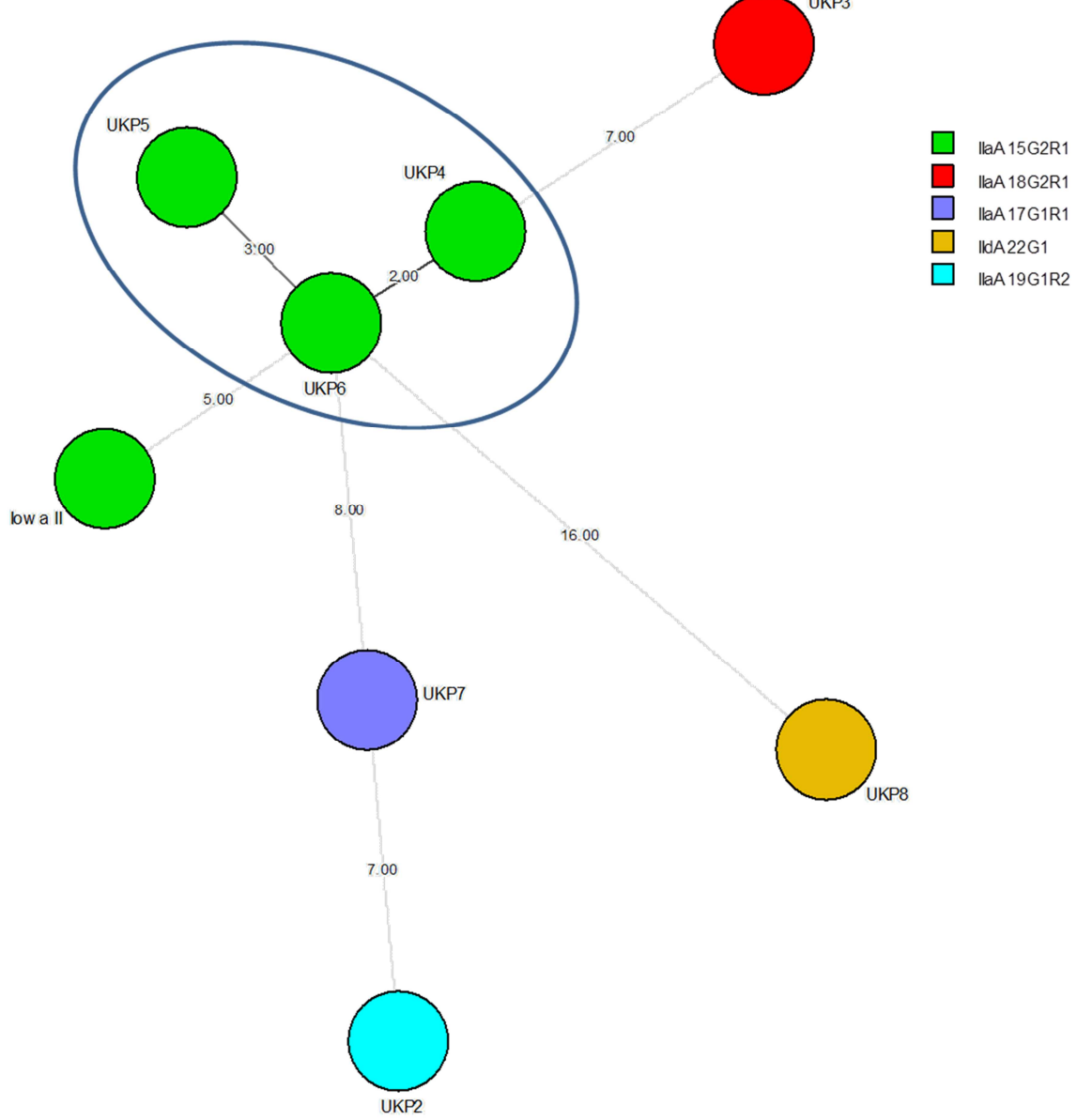
Locus	Number of repeats (gp60 allele)						
	Iowa II (IIaA15G2R1)	UKP2 (IIaA19G1R2)	UKP3 (IIaA18G2R1)	UKP4 (IIaA15G2R1)	UKP5 (IIaA15G2R1)	UKP6 (IIaA15G2R1)	UKP7 (IIaA17G1R1)
cgd1_470_1429	7	4	4	4	4	4	4
cgd1_3060_604	5	5	6	5	5	5	5
cgd1_3170_4182	3	3	3	3	3	3	NF
cgd1_3670_5956	5	5	5	5	5	5	5
cgd2_430_451	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

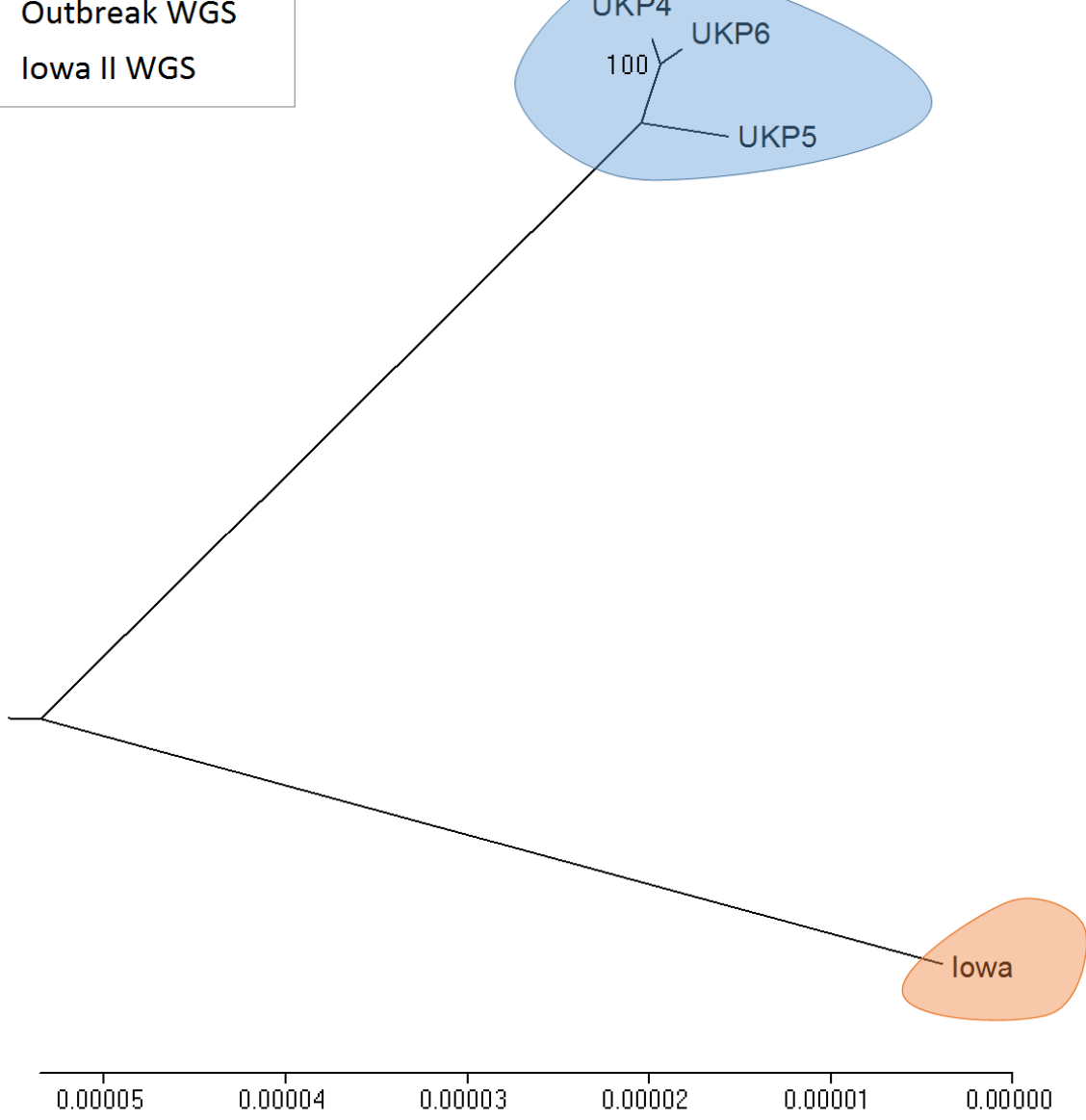
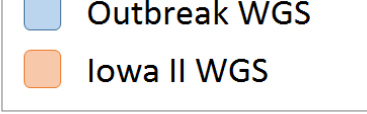
---

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

---







ACU

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

ACCEPTED MANUSCRIPT