Deliverable D-JRP-PARADISE-WP3.2
Report on the identification of markers for multi-locus sequence typing of *Cryptosporidium parvum* and *Giardia duodenalis*

Workpackage 3 of JRP19-ET1.1-PARADISE

Responsible Partners: ISS (27), RKI (11), SVA (41), UoS (23)

GENERAL INFORMATION

European Joint Programme full title	**Promoting One Health in Europe through joint actions on foodborne zoonoses, antimicrobial resistance and emerging microbiological hazards**
European Joint Programme acronym	**One Health EJP**
Funding	**This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.**
Grant Agreement	**Grant agreement n° 773830**
Start Date	**01/01/2018**
Duration	**60 Months**

DOCUMENT MANAGEMENT

JIP/JRP deliverable	Deliverable D-JRP-PARADISE-WP3.2
Project Acronym	JRP19-ET1.1-PARADISE
Authors	Simone M. Cacciò (ISS); Christian Klotz (RKI); Karin Troell (SVA); Martha Betson (UoS)
Other contributors	PARADISE consortium
Due month of the report	M42
Actual submission month	M45
Type *R: Document, report* *DEC: Websites, patent filings, videos, etc.; OTHER*	R **Save date**: September 8th, 2021
Dissemination level *PU: Public (default)* *CO: confidential, only for members of the consortium (including the Commission Services)*	PU **This is the default setting.** If this project deliverable should be confidential, please add justification here (may be assessed by PMT): ....
Dissemination *Author's suggestion to inform the following possible interested parties.*	OHEJP WP 1 ☐ OHEJP WP 2 ☐ OHEJP WP 3 ☒ OHEJP WP 4 ☐ OHEJP WP 5 ☒ OHEJP WP 6 ☐ OHEJP WP 7 ☐ Project Management Team ☐ Communication Team ☐ Scientific Steering Board ☐ National Stakeholders/Program Owners Committee ☐ EFSA ☒ ECDC ☒ Other international stakeholder(s): European Union Reference Laboratory for Parasites, National EU Reference Laboratories for Parasites Social Media: ..... **Other recipient(s):** .....

This activity is part of the European Joint Programme One Health EJP.
This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.

D-JRP-PARADISE-WP3.2

REPORT ON THE IDENTIFICATION OF MARKERS FOR MULTI-LOCUS SEQUENCE TYPING OF *CRYPTOSPORIDIUM PARVUM* AND *GIARDIA DUODENALIS*

This is a public deliverable of One Health EJP Joint Research Project:
JRP19-ET1.1-PARADISE – PARAsite Detection, ISolation and Evaluation
(https://onehealthejp.eu/jrp-paradise/)

Work Package:
JRP- PARADISE–WP3 Design, implementation and validation of MLST schemes

Task:
JRP-PARADISE-PARADISE-WP3-T2 Development of MLST schemes for *C. parvum* and *G. duodenalis*

Project Leader: Simone M. Cacciò, ISS; Deputy Project Leader: Karin Troell, SVA
WP Leader: Karin Troell, SVA; Deputy WP Leader: Christian Klotz, RKI
Task Leader: Martha Betson, UoS; Deputy Task Leader: Simone M. Cacciò, ISS

Contacts: Martha Betson, m.betson@surrey.ac.uk; Simone M. Cacciò, simone.caccio@iss.it

1. Background

The PARADISE project aims to deliver informative typing schemes and innovative detection strategies applicable to food matrices for the food-borne parasites *Cryptosporidium* and *Giardia*. Using genomics and metagenomics, the project is generating much needed data to enrich our understanding of the epidemiology and genomics of these organisms, and provide the basis for development and testing of improved strain-typing schemes. In parallel, strategies (nanobodies, aptamers, use of hybridization probes) to enrich for the target pathogens in different matrices are under development and validation. These new methodologies will form the basis for integrated approaches to control *Cryptosporidium* and *Giardia* in the European food chain.

The objectives of PARADISE WP3 are:

- To select new markers *in silico* for robust high resolution genotyping of *C. parvum* and *Giardia duodenalis* using whole genome data generated in WP2-T1, unpublished data from within the consortium and publically available data
- To develop multilocus sequence typing (MLST) schemes for *C. parvum* and *G. duodenalis*
- To conduct an inter-laboratory comparison of MLST schemes

The procedures undertaken for *in silico* detection of variable regions suitable for subsequent selection of markers for *C. parvum* and *G. duodenalis* to address are described in deliverable:

> Deliverable D-JRP-PARADISE-WP3.1 Report on the *in silico* selection of highly polymorphic sequences in *C. parvum* and *G. duodenalis* genomes
> https://zenodo.org/record/4452772

This deliverable D-JRP-PARADISE-WP3.2 reports on work carried out to address the second objective of PARADISE WP3, namely development of MLST schemes for *C. parvum* and *G. duodenalis,* and includes a plan for future work towards this objective.

2. *Cryptosporidium parvum*

2.1 Novel MLST markers

Primers were designed to amplify highly polymorphic regions in the *C. parvum* genome which were identified as described in D-JRP-PARADISE-WP3.1. A total of 28 primer pairs, 18 designed at SVA and 10 designed at ISS, were then tested in the laboratory. Primers were initially tested at SVA and ISS on DNA extracted from purified *C. parvum* oocysts and products were submitted for Sanger sequencing in both directions. Eighteen primer pairs were subsequently tested on five DNA samples extracted from *C. parvum*-positive faecal samples containing different oocyst concentrations based on a *C. parvum* specific qPCR. Table 1 summarises the results for each primer pair. Overall, 21 primer pairs showed strong and consistent amplification of a single band with no non-specific amplification. Clean, interpretable Sanger sequences were obtained from the PCR products amplified by these primers.

Table 1. Summary of characteristics and performance of 28 primer pairs designed for typing of *C. parvum* by MLST

Marker	Chromosome	PCR product size (bp)	Amplification of DNA from purified *C. parvum*	Amplification of DNA from *C. parvum*-positive faecal samples
1 (CPATCC_0035770)	1	489	Yes/Yes	Yes (Cq – 38.21)*
2 (CPATCC_0035880)	1	535	Yes/Yes	Yes (Cq – 35.51)*
3 (CPATCC_0039030)	1	469	Yes/Yes	Yes (Cq – 33.3)*
4 (CPATCC_0028230)	2	498	Yes/Yes	Yes (Cq – 33.06)*
5 (CPATCC_0029910)	2	576	Yes/Yes	Yes (Cq – 35.51)*
6 (CPATCC_0031960)	3	595	Yes/Yes	Yes (Cq – 33.3)*
7 (CPATCC_0035350)	3	493	Yes/Yes	Yes (Cq – 35.51)*
8 (CPATCC_0018920)	4	442	Yes/Yes	Yes (Cq – 35.51)*
9 (CPATCC_0021750)	4	509	Yes/Yes	Yes (Cq – 33.3)*

10 (CPATCC_0024650)	5	536	Yes/Yes	Yes (Cq – 38.21)*
11 (CPATCC_0011480)	6	568	Yes/Yes	Yes (Cq – 38.21)*
12 (CPATCC_0011530)	6	514	Yes/Yes	Yes (Cq – 38.21)*
13 (CPATCC_0012400)	6	512	Yes/Yes	Yes (Cq – 33.3)*
14 (CPATCC_0007000)	7	568	Yes/Yes	Yes (Cq – 38.21)*
15 (CPATCC_0010930)	7	600	Yes/Yes	Yes (Cq – 33.06)*
16 (CPATCC_0001010)	8	560	Yes/Yes	Yes (Cq – 33.06)*
17 (CPATCC_0001650)	8	439	Yes/ Weak, multiple bands	Yes (Cq – 33.3)*
18 (CPATCC_0002930)	8	564	Yes/ Weak, multiple bands	Yes (Cq – 38.21)*
21 (CP1051)	4	618	Yes/Yes	N/D
22 (CP611)	2	678	Yes/No	N/D
23 (CP907)	3	655	Yes/Yes	N/D
24 (CP1516)	6	430	Yes/Yes, weak	N/D
25 (CP686)	3	362	Not distinct/No	N/D
26 (CP318)	2	261	Yes/Yes	N/D
27 (CP1860)	6	242	Yes/Yes	N/D
28 (CP2089)	7	346	Not distinct/No	N/D
29 (CP2337)	8	475	Yes/Yes	N/D
30 (CP1495)	5	539	Yes/Yes	N/D

Abbrevations: Fw = forward sequence, Rv = reverse sequence,- = no readable sequence generated, double = double peaks, mc = ?, N/D = not determined; * Cq = the sample with lowest amount of DNA (i.e. the highest Cq from a *C. parvum* specific qPCR) that can be amplified by the primer pair.

To investigate the utility of the markers amplified by the primer sets in distinguishing between *C. parvum* isolates, and to narrow down the number of markers for further testing, an *in silico* approach was undertaken. Nineteen markers were selected based on good results in initial laboratory testing and coverage of the eight *C. parvum* chromosomes. Whole genome sequence data (either publicly available or generated as part of PARADISE WP2) from 137 *C. parvum* isolates from human and animals in 13 European countries were interrogated to obtain sequences at each marker locus. Sequences for each locus were aligned, SNPs were identified and the number of different genotypes at each locus was determined. Ultimately, each isolate was assigned a 19-digit code indicating which genotype was found at each locus. Using all 19 markers, it was possible to distinguish between most isolates except for 6 sets (generally pairs) of isolates from the same country, which may represent outbreaks/shared infection origin. Analysis is ongoing to reduce the number of markers to eight while ensuring coverage of all chromosomes and maintaining the level of discrimation between isolates.

Table 2. *In silico* discrimination of *C. parvum* isolates using 19 MLST markers extracted from whole genome sequences

Country	Number of isolates analysed	Number of unique genotypes	Number of shared genotypes
United Kingdom	29	29	0
Italy	27	26	1
Finland	17	13	1
Germany	16	16	0
Hungary	11	8	1
Denmark	9	8	1
France	7	7	0
Slovenia	6	6	0
Sweden	5	4	1
Norway	4	4	0
Poland	2	2	0

Portugal	1	1	0
China	1	1	0
Egypt	1	1	0

2.2. VNTR markers

As an alternative approach to *C. parvum* typing, seven unpublished variable number tandem repeat (VNTR) markers, developed at the *Cryptosporidium* Reference Unit (CRU) in the UK (CRU is an associate partner of the Paradise project, Responsible Dr. Rachel M. Chalmers) were tested on DNA extracted from 48 *C. parvum* positive faecal samples from humans in the Netherlands and cattle in Sweden. All samples were successfully amplified at the seven marker loci and PCR products were submitted for fragment size analysis. Fragment sizes were determined using Genemapper and translated into single and mutlilocus genotype codes. The matrix of multilocus genotypes was then used to infer the clustering of isolates using a principal component analysis (see Figure 1).

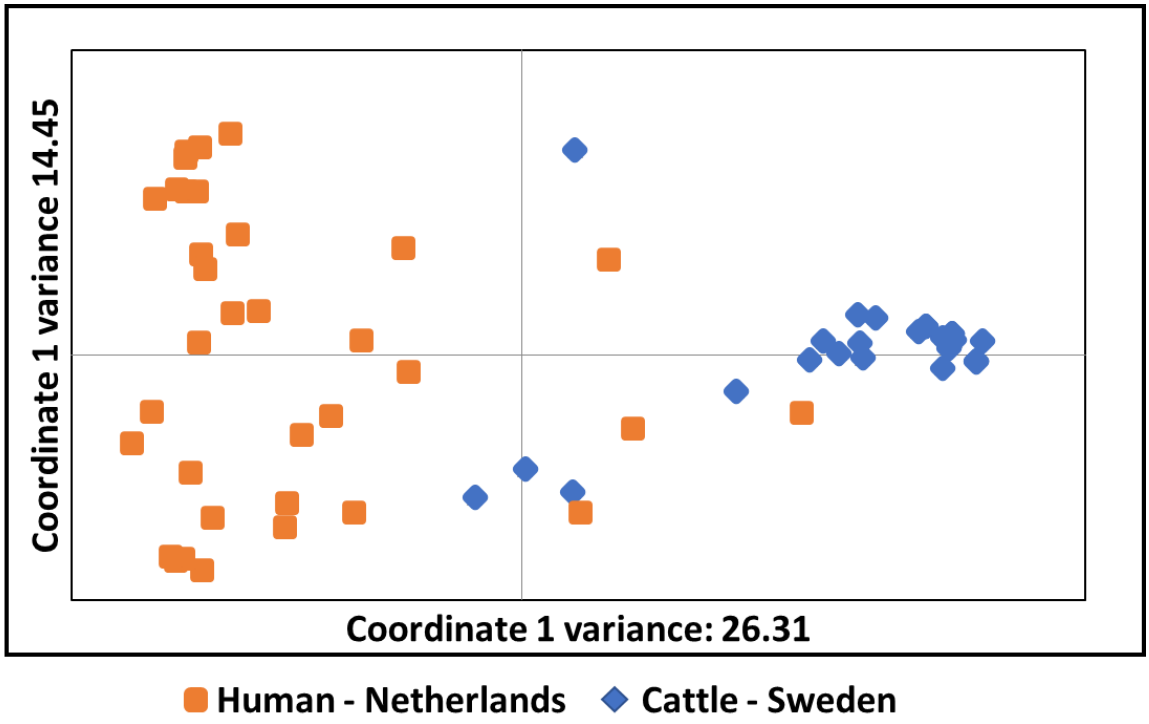

Figure 1. Genetic distances between *C. parvum* isolates based on seven VNTR markers. Principal component analysis was used to assess clustering among *C. parvum* isolates (n=48), based on fragment sizes of seven VNTR markers. Human-derived isolates from the Netherlands are shown in orange and cattle-derived isolates from Sweden in blue.

3. Giardia duodenalis

3.1 Assemblage B markers

Of the 42 highly polymorphic regions in the *G. duodenalis* assemblage B genome identified in WP3 task 1 (as described in D-JRP-PARADISE-WP3.1), 20 were selected and nested primer sets were designed to amplify

the regions. In initial testing, the 40 selected primer pairs amplified PCR products of expected sizes using genomic DNA from 5 different assemblage B isolates (from axenic cultures) as templates. Bidirectional Sanger sequencing of nested PCR products confirmed expected sequences. In further testing on DNA from 10 *Giardia* assemblage B-positive cyst/faecal samples, 8 out of the 10 samples could be amplified by the primers. Sequences for each isolate at each marker locus were aligned and SNPs were identified. Table 3 summarises the results for each nested primer set.

Table 3. Summary of characteristics and performance of 20 primer pairs designed for typing of *G.duodenalis* assemblage B by MLST

Marker	Chromo-some localization (based on WB6 reference)	Outer PCR product size (bp)	Nested PCR product size (bp)	PCR positive samples (*N*=13)	Mean ambiguities per isolate	Mean SNPs per isolate	No. of polymorph-ic sites in consensus
1	1	578	538	13	3.2	15	50
2*	4	593	553	13	3.9	13	43
3	2	700	657	13	6.4	20	79
4	5	573	533	13	2.6	8	32
7*	3	577	562	13	3.2	12	49
8*	2	570	529	12	5.5	15	50
9	3	689	650	11	3.5	13	36
10*	5	598	557	12	6.5	18	56
11	4	629	582	12	4.5	15	43
13	5	661	622	8	3.8	16	37
14*	1	536	498	10	3.1	12	37
15	5	802	762	12	5.7	19	65
16	2	793	752	12	4	19	47
17	4	697	654	11	6.7	19	66
18	5	634	593	11	6.4	19	75
19	5	547	506	12	2.9	10	26
20*	5	656	616	13	3	15	42
21	5	551	512	11	2.2	8	28
22	3	565	523	9	2.3	10	29
27	1	543	503	10	2.3	7	17

* Markers selected for further testing.

Six markers (indicated with asterisks in Table 3), were tentatively selected for further testing based on the following criteria: coverage of all chromosomes, positivity rate of PCR and Sanger sequencing, number of polymorphic sites, avoidance of intergenic regions and sequence length.

An issue identified in the analysis of sequences from the 20 marker loci, which as previously been demonstrated for other markers used to type *G. duodenalis* assemblage B (Woschke et al., 2021), is the presence of "double peaks" in sequencing chromatograms, likely due to allelic sequence heterozygosity (ASH), which is common in assemblage B and complicates the interpretation of typing data. This is represented by "mean ambiguities per isolate" in Table 3. However, in line with the objectives of the work that focuses on the suitability of the new typing scheme for reliable analysis of isolates for epidemipological purposes such as outbreak investigation, the identified markers revealed higher potential discriminatory power than current typing markers. They will therefore be further pursued for establishment of a new typing scheme.

3.2 Assemblage A markers

The availability within the consortium of *G. duodenalis* assemblage A samples from different countries and hosts has provided an opportunity to further test a published MLST scheme involving six markers for

assemblage A (Ankarklev et al. 2018). Twenty samples from humans in Italy (including samples from an outbreak), nine samples from humans in Germany and nine samples from cattle in Poland were typed using the MLST scheme as described (Ankarklev et al. 2018). The resulting data were combined with published data (Ankarklev et al. 2018, Woschke et al. 2021), and the resolution of the scheme assessed by cluster analyses. It was concluded that the MLST scheme for assemblage A has good discriminatory power and a publication describing the results is being prepared.

Next steps

- Select 8 *C. parvum* MLST markers for further testing
- Confirm selection of 6 *G. duodenalis* assemblage B MLST markers
- Conduct laboratory testing of the selected markers on a panel of *C. parvum*, *G. duodenalis* assemblage B samples from humans and animals available at ISS, RKI SVA and UoS (see Table 4) to confirm amplification and determine ability to discriminate between parasite strains
- Develop MLST protocols for *C. parvum* and *G. duodenalis* assemblage B
- Share protocols with PARADISE consortium members for further sample testing
- Conduct an analysis of the population structure of *C. parvum* and of *G. duodenalis* across Europe using the MLST data generated
- Complete analysis of *G. duodenalis* assemblage A typing

Table 4. Samples available for MLST marker testing

Parasite	Source	Country	Sample number
Cryptosporidium	Cattle	Denmark	~10
	Humans/environment/ cattle	Hungary	50
	Humans/animals	Italy	50
	Animals	Latvia	29 Cp; 9 other
	Humans	Netherlands	26 Cp; 49 Ch
	Humans/animals	Norway	30
	Cattle	Poland	29 Cp; 20 other
	Cattle	Portugal	8
	Humans/animals	Sweden	200
	Human/animals	UK	86 Cp; 10 Ch; 4 other
	Cattle/buffalo	USA	50 Cp
Giardia	Humans/environment	Hungary	51
	Humans/animals	Italy	50
	Humans/Animals	Norway	~100
	Cattle/pigs	Poland	37 Ass A, B; 4 Ass E
	Animals	Portugal	6
	Humans/animals	UK	100 Ass A, B
	Human	Travellers returning to Germany)	200 Ass A, B
	Humans/environment	Hungary	51

References

Ankarklev J, Lebbad M, Einarsson E, Franzén O, Ahola H, Troell K, Svärd SG. (2018) A novel high-resolution multilocus sequence typing of *Giardia intestinalis* Assemblage A isolates reveals zoonotic transmission, clonal outbreaks and recombination. Infect Genet Evol. 60:7-16.

Woschke A, Faber M, Stark K, Holtfreter M, Mockenhaupt F, Richter J, Regnath T, Sobottka I, Reiter-Owona I, Diefenbach A, Gosten-Heinrich P, Friesen J, Ignatius R, Aebischer T, Klotz C. (2021) Suitability of current typing procedures to identify epidemiologically linked human *Giardia duodenalis* isolates. PLoS Negl Trop Dis. 2021 Mar 25;15(3):e0009277.