**Supplementary File S1. Technical Appendix for manuscript:**

**Evaluation of an ensemble-based distance statistic for clustering MLST datasets using epidemiologically defined clusters of cyclosporiasis**

Fernanda S. Nascimento1†, Joel Barratt1,2†\*, Katelyn Houghton1,2, Mateusz Plucinski3, Julia Kelley3, Shannon Casillas1, Carolyne (Cody) Bennett1,2, Cathy Snider4, Rashmi Tuladhar4, Jenny Zhang4, Brooke Clemons5, Susan Madison-Antenucci5, Alexis Russell6, Elizabeth Cebelinski7, Jisun Haan7, Trisha Robinson7, Michael J. Arrowood8, Eldin Talundzic3, Richard S. Bradbury1, and Yvonne Qvarnstrom1

1 Parasitic Diseases Branch, Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, GA, USA

2 Oak Ridge Institute for Science and Education, Oak ridge, TN, USA

3 Malaria Branch, Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

4Texas Department of State Health Services, TX, USA

5New York State Department of Health, Wadsworth Center Parasitology Laboratory, NY, USA

6New York State Department of Health-Bureau of Communicable Disease Control, NY, USA

7Minnesota Department of Health, MN, USA

8 Waterborne Disease Prevention Branch, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA

\*To whom correspondence should be addressed. Email: jbarratt@cdc.gov Alternative email: joelbarratt43@gmail.com

†Authors contributed equally

# Table of Contents

[Table of Contents ii](#_Toc46323164)

[Supplementary notes on study rationale 1](#_Toc46323165)

[Study rationale 1](#_Toc46323166)

[The Jaccard Index 3](#_Toc46323167)

[Bray-Curtis dissimilarity 3](#_Toc46323168)

[Supplementary Methods 4](#_Toc46323169)

[DNA extractions protocols used at participating NY and MN state laboratories 4](#_Toc46323170)

[Methods for PCR and deep amplicon sequencing used at the CDC laboratory 5](#_Toc46323171)

[Methods for PCR and deep amplicon sequencing used at the MN laboratory 6](#_Toc46323172)

[Methods for PCR and deep amplicon sequencing used at the NY laboratory 7](#_Toc46323173)

[Methods for PCR and deep amplicon sequencing used at the TX laboratory 8](#_Toc46323174)

[Haplotype assignment 9](#_Toc46323175)

[Haplotype assignment for markers 1 to 4 9](#_Toc46323176)

[Haplotype assignment for markers 5, 6, and 8 10](#_Toc46323177)

[Haplotype assignment for marker 7 11](#_Toc46323178)

[Haplotype data sheet generation 12](#_Toc46323179)

[PCR controls, sequencing controls and proficiency testing 12](#_Toc46323180)

[Minimum data requirements for distance calculation 13](#_Toc46323181)

[Table S1. Some characteristics of the eight (14) Cyclospora cayetanensis genotyping loci used in this study 15](#_Toc46323182)

[Table S2. Sequencing success rate for markers tabulated from specimens that met the inclusion criteria 16](#_Toc46323183)

[Barratt’s heuristic algorithm 16](#_Toc46323184)

[Plucinski’s naïve Bayes and determination of epsilon 22](#_Toc46323185)

[Normalization of distances for construction of an ensemble matrix 26](#_Toc46323186)

[Table S3. Explanation of specimen counts and case-patient counts at different points in this study 27](#_Toc46323187)

[Supplementary Results 28](#_Toc46323188)

[Sequencing and amplification success for each sequencing laboratory protocol 28](#_Toc46323189)

[Figure S1. Proportion of markers (M) that were successfully sequenced and assigned a genotype for each participating laboratory. 28](#_Toc46323190)

[Table S4. Number of genotyping markers sequenced (out of 8 markers total) for the fecal specimens obtained from cyclosporiasis case-patients meeting the inclusion criteria 29](#_Toc46323191)

[Proficiency/validation testing 29](#_Toc46323192)

[Figure S2. Visualization of validation (VAL) specimen genotypes as determined by three independent laboratories (CDC, NY, TX), and the genotype of repeat specimens from two patients. 30](#_Toc46323193)

[Examination of discordantly clustered specimens 31](#_Toc46323194)

[Figure S3. Haplotypes (black bars) detected in discordantly clustered specimens compared to the consensus genotype of the Vendor A and Vendor B clusters (blue bars) 32](#_Toc46323195)

[Table S5. Illness onset dates for patients with discordantly clustered specimens 34](#_Toc46323196)

[Consecutive specimens from the same patient 34](#_Toc46323197)

[Supplementary discussion 35](#_Toc46323198)

[Notes on amplification and sequencing success 35](#_Toc46323199)

[Proposed modifications to laboratory protocols based on sequencing and amplification success 36](#_Toc46323200)

[Robustness of clustering despite missing data 37](#_Toc46323201)

[Appendix 38](#_Toc46323202)

# Supplementary notes on study rationale

## Study rationale

Phylogeny does not constitute an appropriate analytic approach for *C. cayetanensis* MLST datasets because multiple haplotypes are often observed at nuclear loci sequenced from a fecal specimen. Indices such as *FST* are also not applicable to this type of data. Calculation of *FST* requires *a priori* knowledge on how the populations should be defined and this is usually based on host or geographical differences. This meta information is often not available in molecular surveillance contexts. Consequently, alternative analytic approaches for this type of MLST data are required.

In this study we use a recently described distance statistic calculated using an ensemble of two machine learning algorithms. These algorithms compare the mixtures of haplotypes observed at various loci for a cohort of specimens and then generate a pairwise distance matrix. While several distance statistics are available for use in similar contexts, these available statistics were not designed for analyzing genetic data specifically, and so fail to consider certain aspects of genetic data that are highly informative. Statistics such as Jaccard distances or Bray–Curtis dissimilarity are sometimes applied to MLST data, though these relatively simple statistics do not consider loci entropy, allelic frequencies, and nuclear versus extranuclear inheritance, among other aspects of MLST data. In contrast, the ensemble-based statistic used here takes these aspects of genetic data into account. The performance of this ensemble has been evaluated once previously, though that study suffered from several limitations relating to the size and quality of the data available at the time, which ultimately prevented a rigorous performance assessment. In light of this, the present study sought to perform the first rigorous performance assessment of this ensemble-based distance statistic.

*Limitations of other distance statistics commonly used for MLST data*

Ecological distance statistics such as Jaccard distances or Bray–Curtis dissimilarity are often applied to MLST data, though fail to consider loci entropy, allelic frequencies, and nuclear versus extranuclear inheritance. Ignoring these aspects of genetic data leaves simpler distance statistics subject to bias, which can lead to inaccurate results. Higher entropy loci are generally more informative when elucidating genetic relationships than lower entropy loci. It therefore seems inappropriate that all loci in a MLST panel should be weighted the same way when distances are calculated. Allelic frequencies are also an important consideration. Observation of a rare allele shared by 3% of a population generally represents good evidence that these individuals share a relationship as opposed to an allele observed in the same group of individuals that also occurs in 90% of the total population. Important differences between the mechanisms of inheritance for nuclear and mitochondrial loci mean that matching haplotypes observed between two specimens at a mitochondrial versus a nuclear locus, require a different interpretation. The way missing data are handled when distance statistics are calculated is also imperative. If a subset of specimens cannot be sequenced at all loci in a MLST panel due to physical limitations (discussed in the main manuscript text), it is inappropriate that these missing data points be treated as a true absence when in fact, data were present for these loci in the physical specimen but could not be observed due to these physical limitations.

## The Jaccard Index

The Jaccard index for specimens *A* and *B* is calculated as follows

The Jaccard index simply represents the number of intersecting () haplotypes for specimens *A* and *B* over the number of haplotypes at the union () of the haplotypes in specimens *A* and *B*, given that vertical bars indicate that the number of haplotypes in the set detected in specimens *A* and *B* should be taken.

## Bray-Curtis dissimilarity

Bray-Curtis dissimilarity for specimens *i* and *j* () is calculated as follows

When assessing the similarity between two distinct geographic sites in terms of their taxon diversity, is typically taken as the sum of the lesser counts of sites *i* and *j* for taxa that are common to both sites only. However, given that MLST data are typically represented as either the presence or absence of an allele as opposed to the number of times an allele was observed in a specimen, all alleles are considered to have been observed once. Therefore, in the context of an MLST dataset, is the number of matching alleles between specimens *i* and *j*. is the number of alleles detected in specimen *i* and is the number of alleles observed in specimen *j*.

# Supplementary Methods

## DNA extractions protocols used at participating NY and MN state laboratories

In the Parasitology department at the Wadsworth Center, New York (NY), 0.5 mL aliquots of stool were transferred to a 2 mL tube, and these were washed twice by adding 1 mL of PBS followed by vortexing and then centrifuging at 6800 x *g* for 2 minutes. After discarding the supernatant, 0.5 mL of the resulting pellet was re-suspended in NucliSens lysis buffer (BioMerieux, NC, USA), and then transferred to a Matrix E tube (MP Biomedicals, OH, USA). The samples were vortexed again and incubated at 75°C for 15 minutes. The tubes were then placed onto a FastPrep instrument (MP Biomedicals) and homogenized for 45 seconds at a speed setting of 6.0, followed by centrifugation at 16000 x *g* for 5 minutes. Finally, the supernatants were transferred to a 2 mL tube, where 10 µL of a plasmid internal PCR control, and 20 µL of proteinase K were added. This solution was re-extracted with the QIAcube instrument (Qiagen, Hilden, Germany) using the DNA mini protocol with an elution volume of 50 µL. Extracts were stored at 4°C.

At the Minnesota (MN) Department of Health Infectious Disease Laboratory, aliquots of 0.25-1 mL of stool were transferred into a 2 mL tube. The samples were washed twice by adding 1 mL of water, followed by thorough vortexing and then centrifugation at 15,000 x *g* for 3 minutes. After discarding the supernatant, the roughly 0.25 mL size stool pellets were re-suspended in water and delipidized with Citrus Clearing Solvent. The resulting pellets were washed with water and then resuspended in DNAzol (MRC, Inc., OH, USA), transferred to a Matrix C tube (MP Biomedicals) and boiled for 15 minutes. The tubes were placed on a FastPrep instrument (MP Biomedicals) and homogenized for 1 minute at speed setting 5.5 followed by centrifugation at ≥16000 x *g* for 3 minutes. Clarified supernatants were ethanol precipitated, the pellets were washed twice and resuspended in 100 µL of water. The resulting solutions were purified further using the Qiagen QIAamp PCR Purification kit with a final elution volume of 50 µL.

## Methods for PCR and deep amplicon sequencing used at the CDC laboratory

At the CDCs PDB laboratory PCRs for markers 1 to 7 were prepared in a total volume of 25 µL to concentrations of 400 nM for each primer, single strength NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs, MA, USA) and 1 µL of DNA template. The PCR condition for targets one to five were 98°C for 2 min, 35 cycles of 98°C for 15 sec, 67°C for 15 sec, 65°C for 15 sec, and a final extension at 65°C for 5 min. The PCR conditions for target six were 98°C for 2 min, 35 cycles of 98°C for 15 sec, 67°C for 30 sec, 72°C for 30 sec, and a final extension at 72°C for 5 min. For target seven the reaction conditions were 98°C for 2 min, 35 cycles of 98°C for 15 sec, 66°C for 15 sec, 72°C for 30 sec, and a final extension at 72°C for 5 min. The PCR for target eight was performed in a 25 µL reaction containing Platinum PCR Super Mix High Fidelity (Invitrogen, CA, USA), 200 nM of each primer, and 1 µL of DNA template. The PCR conditions were 94°C for 2 min, 35 cycles of 94°C for 15 sec, 55°C for 30 sec, 68°C for 40 sec, and a final extension at 68°C for 5 min. A Verity and GeneAmp 9700 PCR System (Applied Biosystems, CA, USA) thermal cycler was used for all amplifications. The amplicons were subjected to agarose gel electrophoresis on 1.5% agarose gel stained with GelRed (Biotium, CA, USA) and visualized under UV light in a U:Genius3 (Syngene, MD, USA). Amplification and sequencing were attempted on all fecal specimens provided to CDC (e.g. exclusion of specimens based on a real time Ct value was not applied). Sequencing was also attempted on amplicons that were negative by gel electrophoresis.

Due to differences in PCR sensitivity (i.e. amplification of marker 8 was particularly efficient), one volume of amplicon from targets one to seven, and half a volume of amplicon from target eight was pooled for each specimen. Purification and normalization of pools was performed using a SequalPrep Normalization Kit (Thermo Fisher Scientific, MA, USA) with an elution volume of 20 µL. The normalized pool of amplicons was subjected to library preparation using Nextera XT DNA Library Prep Kit (Illumina, CA, USA). After amplicon pooling and library preparation, DNA concentration and DNA molecule size were assessed using a QubitdsDNA HS Assay Kit (Invitrogen) and a High Sensitivity D1000 ScreenTape on the 2200 TapeStation (Agilent, CA, USA), respectively. The pooled amplicon libraries were diluted to 10 -15 pM and sequenced on the MiSeq Platform using either the MiSeq Reagent Kits V2 (500) or Nano V2 (500) (Illumina).

## Methods for PCR and deep amplicon sequencing used at the MN laboratory

Thirty-two stool specimens containing *Cyclospora cayetanensis* as confirmed at a clinical lab (using either the BioFire FilmArray platform or microscopic examination of modified acid-fast stained fecal smears), were subjected to amplification of the eight genotyping markers, using the amplification and reaction conditions described in the CDC laboratory protocol. The thermal cyclers used included either a DNA Engine, iCycler or C1000 (Bio-Rad, CA, USA). The amplicons were individually purified using a Monarch purification kit (New England Biolabs) and amplicons where non-specific bands were apparent following gel electrophoresis, had the band of interest excised from the gel and purified using a MinElute Gel Extraction Kit (Qiagen). The purified amplicons were quantified, diluted, and pooled per specimen and subjected to the Nextera XT DNA Library Prep method described above (in the CDC protocol) which was modified from a previously described protocol[[1]](#footnote-1).

## Methods for PCR and deep amplicon sequencing used at the NY laboratory

Ninety-one stool specimens positive for *C. cayetanensis* based on a routinely used real-time PCR and with a Ct value < 35, were subjected to genotyping. The PCR conditions were performed as described above for targets one to seven (in the CDC method), with the modification that 2 µL of template DNA was used instead of 1 µL. Target eight was amplified using HotStarTaq Master Mix (Qiagen), 400 nM of each primer, and 2 µL of DNA in a 25 µL reaction. The PCR conditions consisted of 95°C for 15 min, 35 cycles of 94°C for 30 sec, 55°C for 30 sec, 72°C for 45 sec, and final extension at 72°C for 10 min. This was performed on a SimpliAmp Thermal Cycler (Thermo Fisher Scientific). The amplicons were visualized on 1.5% agarose gel stained with ethidium bromide (Thermo Fisher Scientific) and visualized under UV light using Bio-Rad Universal Hood II Gel Doc System (Bio-Rad, CA, US). Sequencing was also attempted on amplicons that were negative by gel electrophoresis.

All the amplicons were purified using AMPure XP beads clean up (Beckman Coulter Life Sciences, IN, USA). Six microliters of each amplicon was transferred to a 96 well plate, and 10.8 µL of AMPure XP beads were added to each sample, capped and mixed by vortexing briefly. After five minutes of incubation at room temperature (RT) the plate was placed on a magnetic stand for two minutes and supernatant was removed and discarded without disturbing the beads. The beads were washed twice with 200 µL of 80% ethanol. After ethanol was removed, the plate was taken off the magnetic stand and 40 µL of nuclease-free water was added to each well. Wells were capped and the plate was briefly vortexed. After two minutes of incubation at RT the plate was placed back on the magnet stand and 30 µL of eluted amplicon was transferred to a new plate. Ten microliters of each amplicon from a single sample was pooled into a single tube. The pool was quantified using Qubit (Invitrogen), diluted, and subjected to library preparation as previously described (the CDC method above). The pooled libraries were diluted to 7-8 pM and loaded onto MiSeq using a MiSeq Nano, Micro or standard flow cells (depending on need).

## Methods for PCR and deep amplicon sequencing used at the TX laboratory

PCR was performed using 200 nM of primers for targets one to four and eight, in a 25µL reaction using Platinum PCR Super Mix High Fidelity (Invitrogen) and 1 µL of DNA. The PCR conditions for targets one to four were 94°C for 2 min followed by 35 cycles of 94°C for 15 sec, 55°C for 15 sec, 68°C for 15 sec, and a final extension of 68°C for 5 min. For target eight, 40 seconds of extension was used following the same conditions for targets one to four. Targets five to seven PCRs were performed as previously described (CDC). The thermal cyclers used were a Verity (Applied Biosystems) and Master Cycler Nexus Gradient (Eppendorf, Germany).

The amplicons were visualized on 1.5% agarose gel stained with Ethidium Bromide (Invitrogen) and visualized under UV light using Biorad Universal Hood II Gel Doc System (Bio-Rad). All the amplicons were pooled, per specimen, and then purified using Monarch (New England Biolabs) on a final elution of 20 µL. After quantification using a Qubit, and dilution to 0.2 ng/µL, the pooled amplicons were subjected to library prep using Nextera XT (Illumina) following manufacturer’s instructions. The pooled libraries were again quantified using a Qubit, and 10-20 pM were loaded onto a MiSeq Nano V2 (500) kit, and sequenced on a MiSeq Platform (Illumina). Sequencing was attempted on amplicons that were negative by gel electrophoresis.

## Haplotype assignment

To detect *C. cayetanensis* haplotypes in our Illumina sequence data, the trimmed, filtered, and merged Illumina reads (see manuscript text for details) were mapped to all reference sequences (see Appendix 1) with high stringency, such that only haplotypes that truly exist in the specimen obtain sufficient coverage to be considered present in that specimen. However, as each marker possesses different characteristics in terms of their length, SNP composition, and repeat composition, the Geneious workflows used to assign haplotypes varied between the markers. Each of these workflows is described below.

## Haplotype assignment for markers 1 to 4

A curated list of haplotypes was generated after accumulation of data from cyclosporiasis cases over the past 5 years, including from locally acquired cases and international travelers. The sequence of each known haplotype for each marker is provided in Appendix 1 of this document. Amplicons of markers 1 to 4 are between 175 and 246 bases in length, with SNPs that differentiate each haplotype falling within a short span of 80 bases. Merged reads spanned the entire length of the amplicon. To assign a haplotype for these markers, a custom Geneious workflow merged reads and mapped these reads to sequences in the reference database requiring a minimum overlap of 100 bases with 100%. Mismatches and gaps were not tolerated though a single ambiguous base was allowed. A minimum mapping depth of 10 was required for every base along the reference sequence before a given haplotype was considered present in the specimen.

## Haplotype assignment for markers 5, 6, and 8

Before defining haplotypes for markers 5, 6 and 8, these markers were split into discrete sections that were each treated as separate loci. Amplicons of these markers are 650, 469 and 674 bases long, respectively and the Nextera XT DNA Library Prep Kit includes a tagmentation step that results in breakage at random points along the DNA molecules. Given the random nature of this step, and because the length of amplicons for markers 5, 6 and 8 exceed the length of a merged pair of Illumina reads, accurately reconstructing full-length haplotypes for these markers is challenging; when multiple SNPs exist at the ends of an amplicon, these fragmented reads provide little information on which of these distal SNPs are linked. The presence of multiple haplotypes in a single amplicon is especially problematic due to PCR-induced chimeras which are a common problem for genotyping. These artefacts are extremely challenging to distinguish from potentially real haplotypes and can lead to misinterpretation of results. Instead of attempting to reconstruct haplotypes at their full length while excluding all possible PCR-induced chimeras, these markers were divided into sections that were treated as separate loci. Marker 5 was divided into three variable regions of less than 110 bases (sections A, B and C), and the haplotypes within these regions were defined separately. This same procedure was performed for marker 6, which was divided into four sections (A, B, C and D) of less than 105 bases. Marker 8 was divided into two sections; ‘Left’ and ‘Right’, that capture 129 and 107 bases from the respective ends of the amplicon (the variable sites for this marker).

Splitting of marker 7 into segments was deemed unnecessary because haplotypes of this marker are largely based on repeat length, and the short length of this repeat ensures that its sequence typically fits within a single read; this informs us of the true length of the repeat. Splitting of markers 1 to 4 was deemed less critical also, because the amplicons are relatively short such that a merged pair of Illumina reads would span all polymorphic sites. Therefore, construction of full-length haplotypes was less problematic. Additionally, these markers are less diverse than markers 5, 6 and 8; markers 1, 2, and 4 possess two haplotypes, and marker 3 possesses three haplotypes. Due to their relatively low diversity, many specimens in this study were homozygous at these loci (in contrast to markers 5, 6, and 8), allowing us to confidently distinguish true haplotypes from PCR-induced artefacts (see Appendix 1 of this document for all haplotypes).

## Haplotype assignment for marker 7

Assignment of haplotypes for marker 7 required a different Geneious workflow to the other markers, as haplotype lengths varied from between 109 and 214 bases due to the existence of three distinct 15-mer repeat sequences that occur at different frequencies and in different combinations within each haplotype. A haplotype of marker 7 was only called if at least two merged reads extended across the entire length of the repeat region, mapping with zero gaps and 96% identity. While this is still stringent, this identity threshold is relaxed compared to the other workflows because the haplotype lengths for this marker are inconsistent (Appendix 1 of this document). To ensure that reads spanning the entire repeat could map to longer haplotypes, we relaxed this identity threshold so that reads spanning the entire repeat were not discarded due to the presence of one or two errors. Additionally, each base of the reference had to obtain coverage of at least 2 reads for haplotypes of this marker to be assigned to a specimen. The minimum mapping requirement of 2 reads is also relaxed relative to the other workflows, and was optimized to accommodate differences in repeat length. Because the tagmentation step favors high coverage towards the middle of an amplicon but low coverage at the ends, for the longer repeats we sometimes observed very low coverage at the ends of the references, but excellent coverage towards the center. Lower coverage at the ends of the references was tolerated because these regions include only the priming sites (which provide no typing value), while coverage towards the center of these longer references was always much higher. These mapping requirements would ensure that haplotypes of the correct repeat length and repeat-type composition would be identified.

## Haplotype data sheet generation

A custom BASH script was used to examine the haplotypes listed in each text file and export this information to a haplotype data sheet (shown in Supplementary File S2, Tab A). This data sheet is a table providing all the haplotype information obtained for each specimen. Each row begins with a specimen name and each column heading represents a known haplotype. An “X” is placed in the appropriate cell at the intersection of a row containing a given specimen name, and a given column representing a haplotype detected within that specimen. The resulting ‘haplotype data sheet’ is saved as a text file or csv file and this file is used as the input for our ensemble procedure following the directions provided here: <https://github.com/Joel-Barratt/Eukaryotyping>.

## PCR controls, sequencing controls and proficiency testing

PCRs using water instead of DNA template were included with every PCR run, and at CDC, DNA extracted from stool specimens negative for *C. cayetanensis* were used as negative controls. State public health laboratories provided samples that were negative for *C. cayetanensis* but positive for other gut pathogens (*Cryptosporidium* sp., *Salmonella* sp. and pathogenic *E. coli*) to CDC as blinded negatives. These specimens were extracted as additional controls for PCR. Three laboratories (CDC, NY and TX) participated in a proficiency test in which six stool specimens containing *C. cayetanensis* were divided into three aliquots, and one aliquot of each specimen was subjected to DNA extraction, PCR, and sequencing at the three different laboratories. Sequential specimens from two patients were also genotyped as an additional assessment of the robustness of these methods. Specimens C\_WI091\_18 and C\_WI120\_18 were collected on the same day from one cyclosporiasis patient, while C\_WI053\_18 and C\_WI157\_18 were collected 14 days apart from a second cyclosporiasis patient.

## Minimum data requirements for distance calculation

An advantage of our ensemble-based statistic over simpler statistics is that it accommodates specimens with data missing at some loci by not treating these missing data points as truly absent. One of the algorithms underpinning the ensemble enters an imputation routine at the end of a computation that estimates a rational distance for missing loci, which is more appropriate than treating missing loci as being absent from a sample. However, there are limits on the minimum number of available markers required. For example, it is unreasonable to expect that an accurate distance would be calculated if data are available for only one of eight markers. Therefore, in this study, only specimens that met one or both of the two following inclusion criteria were retained for downstream analysis: (1) specimens must have sequence data available for any three of markers 5, 6, 7 and 8 (see Table 1 – main manuscript) plus at least one additional marker, and/or (2) specimens must have sequence data available for at least any five markers. Specimens that failed to meet at least one of these criteria were excluded from the analysis.

These criteria were based on the first description of the ensemble procedure, which cautioned against analyzing specimens where too few markers were successfully sequenced. The two criteria used here ensure that at least half of the markers were available to support accurate clustering downstream. The specific requirement that any three of markers of 5, 6, 7 and 8 must be successfully sequenced is related to the relatively high Shannon entropy (*H*) of these markers, as Shannon entropy can be used to define the amount of information provided by each marker. The Shannon entropy of a marker can be determined by

where in the context of this study, the range of to extends across all known haplotypes of the marker , and is the probability that haplotype will be detected when you sequence marker from any given specimen. is a frequentist probability determined by the number of specimens possessing haplotype of marker , divided by the number of times a sequence was obtained for marker . For our purposes, if three haplotypes were detected in each of 20 specimens, two haplotypes were detected in one specimen, and one haplotype was detected in another specimen, the denominator for calculating would be 63. A of base 2 means the units of entropy will be expressed in bits while a of base 10 results in units of entropy expressed in bans. The entropy of our typing loci expressed in units of bans after the splitting of markers 5, 6, and 8, is provided Table S1 (below).

Criterion 2 for retaining specimens in this analysis (any 5 markers) considers that the markers with the highest entropy coincidently also had the highest sequencing success rates (Table S2 - below). Therefore, the likelihood of having a specimen with data for only one of markers 5, 6, 7 or 8, and four of the remaining ‘low entropy’ markers (1 to 4) would be very low. The present study confirmed this experimentally when only 1 of 648 specimens that were retained for analysis obtained data for low entropy markers 1 to 4 and a single high entropy locus (marker 7).

### Table S1. Some characteristics of the eight (14) Cyclospora cayetanensis genotyping loci used in this study

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CDS-1** | **CDS-2** | **CDS-3** | **CDS-4** | **HC378 PART A** | **HC378 PART B** | **HC378 PART C** | **HC360i2 PART A** | **HC360i2 PART B** | **HC360i2 PART C** | **HC360i2 PART D** | **Mt Junction** | **MSR Left** | **MSR Right** |
| Number of known haplotypes for each locus | 2 | 2 | 3 | 2 | 6 | 2 | 4 | 2 | 5 | 4 | 2 | 15 | 4 | 3 |
| Median number of haplotypes per specimen | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| Mode number of haplotypes per specimen | 1 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |
| Max. number of haplotypes detected at this locus for this dataset\* | 2 | 2 | 2 | 2 | 5 | 2 | 3 | 2 | 4 | 4 | 2 | 3 | 3 | 2 |
| Entropy (*H*) \*\* | 0.269 | 0.262 | 0.25 | 0.278 | 0.712 | 0.201 | 0.482 | 0.22 | 0.479 | 0.521 | 0.219 | 0.737 | 0.272 | 0.303 |

Note: While eight genotyping markers were included in our panel, 14 values for *H* shown here to reflect the entropy of loci after three of them (MSR, HC360i2 and HC378) were divided into multiple segments.

\* Within a single specimen.

\*\* Expressed in units of *bans*.

### Table S2. Sequencing success rate for markers tabulated from specimens that met the inclusion criteria

|  |  |  |
| --- | --- | --- |
| **Marker (#)** | **Number of specimens with this marker sequenced** | **Percentage** |
| CDS-1 (1) | 378 | 56.8% |
| CDS-2 (2) | 556 | 83.5% |
| CDS-3 (3) | 553 | 83.0% |
| CDS-4 (4) | 354 | 53.2% |
| HC378 (5) | 640 | 96.1% |
| HC360i2 (6) | 625 | 93.8% |
| Mt-Junction (7) | 574 | 86.2% |
| MSR (8) | 652 | 97.9% |

Note: These values do not differentiate between data generated at the different laboratories and considers only specimens that met the minimum MLST data requirements (648 from 2018 plus the 18 validation specimens).

## Barratt’s heuristic algorithm

The heuristic algorithm first calculates the value of for each locus when a pair of specimens is compared. The value of is the result of a heuristic based on set theory and propositional logic that approximates the degree of genetic separation at locus for the pair. Determination of differs for nuclear (*Nu*) loci and extranuclear (*Ex*) loci because they have different mechanisms of inheritance. Extranuclear loci include those encoded in organellar genomes that are not subject to nuclear inheritance, such as loci encoded in mitochondrial (Mt) and plastid genomes. In this study, markers 7 and 8 are encoded in the Mt genome and are therefore *Ex* loci. Calculation of when comparing specimen *A* to specimen *B* at locus *x*, is performed as follows

For simplicity, square brackets [ ] are used here to indicate statements of propositional logic where the value is 1 if true and 0 if false. Vertical bars indicate the number of elements (haplotypes) in a set. For example, indicates the number of haplotypes in the set of haplotypes *D* observed at locus *x* in specimen *A*. For *Nu* loci, the raw distance () between specimens *A* and *B* at locus *x* is determined by

where

Curly braces { } represent a set-building annotation. For determination of *j* the annotation creates a set representing the haplotypes at locus *x* for all () specimens in the dataset with haplotypes available at this locus (e.g. where ), *min* indicates selection of the smallest value. The value of *z* is determined by

For both *Nu* and *Ex* markers, the value of is weighted by the entropy of locus and a frequentist probability resulting in a final distance at locus for the pair determined by

Consequently, when and

When and

When and

Where is the probability that any two specimens would possess the same haplotypes observed at the intersection of locus for specimens *A* and *B*, as determined by

where builds a set comprised of the sets of haplotypes *D* at locus *x* from all () specimens for which all haplotypes at the intersection of and are elements. Finally, the total distance () for specimens *A* and *B* is determined by

Where is the final distance for isolates *A* and *B* at locus *x*, and *N* is the number of loci in the genotyping panel. In cases where data are missing for a specimen at any locus in a genotyping panel (e.g. when ) due to amplification or sequencing failures, the algorithm commences an imputation routine that calculates a mean distance in place of . Depending on the circumstance is calculated in one of five ways (*i* –*v*):

1. When comparing specimen *A* with data for locus *x* (e.g. ) to specimen *B* with data missing at locus *x* (e.g. ) the value of is determined by

Where is a set comprised of all values of generated when comparing specimen *A* to all specimens (). Curly braces indicate generation of a set containing the values where the specimens used to generate these values meet the criterion and a second criterion

which represents the values of the set for all () specimens compared to *A* possessing data that are identical to the data available for specimen *B* at their intersecting loci (*X*) (e.g. where contain identical haplotypes), given that loci with no data are considered null and absent from sets of *X*. For example, when there are eight markers in a genotyping panel, and specimen *C* has data for all except marker 7, then . If for specimen *B* data is only available at markers 1, 6, 7 and 8, then , therefore .

1. When comparing specimens *A* and *B* where specimen *B* has no data available for locus *x* (e.g. where ), data is available for specimen *A* (e.g. where ), but specimens *B* is unique (e.g. where , given that is a set containing data *D* for all specimens at all loci *X* excluding *x*, the value of is determined by

Where is the set of all () values of calculated for this dataset and indicates exclusion () of all values of generated by self-to-self (*self*) comparisons ().

1. When comparing specimen *A* and *B* at locus *x*, and , the value of is determined in one of four ways (a to d). Options (a) and (b) will yield an identical result when both specimens *A* and *B* are not unique at their respective sets of available loci (*X*):

b.

is a set comprised of all values of generated when comparing specimen *A* to all other specimens. If *A* is unique and *B* is not, calculation of (option a) is used. If *B* is unique and *A* is not, calculation of (option b) is used. In the event that specimens *A* and *B* are both unique the following (options c or d) are used:

c.

d.

is a set comprised of all available values of generated when comparing specimen *A* to other specimens. is a set comprised of all available values of generated when comparing specimen *B* to other specimens. Options (c) and (d) yield identical results.

1. When performing a self-to-self comparison of specimen *A* at locus *x* where , and specimen *A* is not unique at all haplotypes contained in the set X, the value of is determined by
2. Finally, when performing a self-to-self comparison of specimen *A* at locus *x* where , and where , the value of is determined by

## Plucinski’s naïve Bayes and determination of epsilon

Plucinski’s naïve Bayes assumes that a single specimen possesses a genotype derived from the mixing of two populations (or two individuals – either interpretation is valid). When two genotyped specimens are compared, Plucinski’s naïve Bayes first calculates the probability that the pair would possess the observed haplotypes conditional on each isolate sharing one of its two ancestral populations or, , determined by

where is the observed data for isolate *i*, is the number of ancestral populations shared by isolate *i* and *j*, is the *m*th observed allele at locus *k* for isolate *i*. is the set of successfully sequenced loci for specimen *i* (e.g. where *m* > 0), is the number of alleles at locus *k* for isolate *i*, and is a function that returns the empiric frequency of allele *x* at locus *k.* Regarding the upper limit of Pi (), is the intersection of the sets of loci *K* with alleles available (only loci where *m* > 0) for specimens *i* and *j*, given that loci with no available data (e.g. where *m* = 0) are considered null and absent from sets of *K*. For example, for a genotyping panel of 8 markers where specimen *i* has data for all but marker 7, and specimen *j* has data for markers 1, 6, 7 and 8, such that , then. Vertical bars indicate the number of elements in a set*.* The function is defined as follows

where *S* represents the complete set of specimens included in the analysis, the annotation builds a set containing all specimens possessing data at locus *k*, and *x* represents an allele contained within a given set of . Square brackets indicate statements of propositional logic where the value is 1 if true and zero if false.

Next, the algorithm calculates the probability that the genotypes observed in isolates *i* and *j* are conditional on the isolates sharing the same ancestral populations (i.e. both isolates have two ancestral populations in common) or .

where is the number of all possible pairs of alleles at locus *k* for isolate *i* and is the *x*th set of all possible pairs of alleles for isolate *i* at locus *k.* The final distance between specimens *i* and *j* is then determined by

For simplicity, when calculating we can let

such that

When calculating we can let

such that

Before commencing a computation using Plucinski’s naïve Bayes, the value of epsilon () must be determined by the user (see: <https://github.com/Joel-Barratt/Eukaryotyping>). As the final likelihood is taken as the product of the locus-specific likelihoods *U* generated for all loci from to , if a value of zero is determined for any value of or this results in a net value of zero for *P*, which increases . The value of is used to avoid assigning a zero value to *P*, and is applied only in the event that . The value of can be simply described as the approximated rate that haplotypes are missing or absent for a dataset, and when calculating an alternative value for (in place of zero), is used as follows

Alternatively, when calculating an alternative value for , is used as follows

where curly braces indicate a set and is the set of haplotypes observed at locus *k* for specimen *i*. Vertical bars indicate that the number of elements (haplotypes) in a set should be taken, is the same function defined previously (above) that calculates the empiric frequency of each allele in the set . The *min* annotation indicates that the minimum value in a set should be taken.

The value of was estimated using consensus genotypes predicted from specimens associated with the Vendor A and Vendor B outbreaks, given that these represent the largest epidemiologically-defined clusters; more than 100 genotyped specimens met the minimum inclusion criteria for clustering from each of these outbreaks. To establish a consensus for these two outbreak clusters, if a haplotype was detected in at least 25% of all specimens that met the minimum data requirements for inclusion in our clustering analysis, that haplotype was included as part of the consensus genotype for that cluster (Supplementary File S2, Tabs C and D).

To roughly approximate the rate at which markers were not detected () using our TADS methodology we examine the genotype of all specimens that were linked epidemiologically to the Vendor A or Vendor B outbreaks regardless of whether they met the minimum data requirements or not. We genotyped specimens from 317 case-patients linked to these two outbreaks collectively but only 242 of these were retained for analysis (see Table S2 – below). For specimens genotyped from both the Vendor A and Vendor B outbreaks, 18 haplotypes were included in their respective consensus genotypes. Therefore, if all 18 haplotypes had been detected for each of these 317 cases, the expected frequency of haplotypes detected would be 5,706 markers (i.e. 317 x 18). However, the true number of haplotypes detected was 3,953 of the expected 5,706. Based on this result, the value of was determined as follows.

In this case, the value of is 0.3072.

## Normalization of distances for construction of an ensemble matrix

Each algorithm generates its own pairwise distance matrix and these two matrices are normalized to generate the single ensemble matrix, which is the only output generated using the scripts deposited in the aforementioned GitHub. Normalization of the Bayesian and heuristic matrices is performed first by normalizing each matrix to a maximum value of 1 and then mapping the distribution of distances generated by the Bayesian algorithm to the empiric distribution of distances generated by the heuristic algorithm. The mean of these mapped pairs of distances is used to generate the final ensemble matrix (Supplementary File S2, Tab E).

### Table S3. Explanation of specimen counts and case-patient counts at different points in this study

|  |  |  |
| --- | --- | --- |
| **Definition** | **Number** | **Notes and explanations** |
| Total number of **specimens** from outbreak cases for which genotyping was attempted | 927 | This includes specimens sent to the four participating laboratories for genotyping: CDC (n=686), TX (n=118), NY (n=91) and MN (n=32). This total excludes the 18 validation specimens |
| Number of **case-patients** represented by the 927 specimens tested | 925 | Two case patients provided two specimens each. Therefore, while 927 stool specimens were tested in total, 925 case-patients provided specimens. Specimens from these two case-patients were sent to CDC for genotyping. Therefore, genotyping was attempted at CDC for 686 specimens representing 684 case-patients. |
| Number of **specimens** where at least one marker was successfully sequenced | 869 | Of the 927 specimens for which genotyping was attempted, typing failed at every marker for 58 of these specimens. The remaining 869 specimens where typing was successful for at least one marker, are listed in Supplementary File S2, Tab A. Note in that spreadsheet, that the 18 validation specimens are shown. This results in a total of 887 specimens for which at least 1 marker was successfully sequenced. |
| Number of **specimens** that were retained for classification based on the criteria for inclusion | 648 | As described in the methods, these are the specimens that were retained for the ML analysis. This includes specimens that (1) have sequence data available for any three of markers 5, 6, 7 and 8 (Table 1) plus one additional marker, and/or (2) have sequence data available for any 5 markers. There were 648 specimens collected during 2018 that met these criteria. This number excludes validation specimens. |
| Number of **specimens** collected from previous years that were sequenced for validation purposes only | 18 | This includes six specimens collected during previous years (2014 and 2015) that were divided into three aliquots for genotyping at the CDC, NY and TX laboratories to ensure consistency of the genotyping procedure. Therefore, a total of 666 specimens (648 + 18 = 666) were retained for genotyping. The features of these 666 specimens are shown in Supplementary File S2, Tab B. |
| Total number of **specimens** retained for analysis using our ensemble approach | 666 | This includes 648 specimens from 2018 in addition to the 18 validation specimens. |
| Total number of **laboratory-confirmed cases** associated with the Vendor A and Vendor B outbreaks. | 761 | There were 511 laboratory-confirmed cases linked to Vendor A and 250 laboratory confirmed cases Vendor B. However, the CDC and the State laboratories participating in this study did not receive all of these 761 specimens for genotyping. In total, 242 specimens from these two outbreaks were typed collectively. For the Vendor A outbreak 116 specimens were genotyped and for the Vendor B outbreak 126 specimens were genotyped (Table 2). |
| Number of genotyped **specimens** from this study possessing epidemiologic links | 264 | Excludes validation specimens. Refer to Table 2 for further details. |
| Number of genotyped **specimens** from this study with unknown epidemiologic linkage | 384 | Of these 384 specimens, 22 were assigned to genetic cluster 4 containing the majority of specimens linked to Vendor B, and 44 were assigned to genetic cluster 7 containing the majority of specimens linked to Vendor A. The remaining 318 were distributed amongst the other genetic clusters. Note that 648 specimens from 2018 were retained for genotyping (264 + 384 = 648). |
| Number of **specimens** associated with the Vendor A and Vendor B outbreaks before exclusion of specimens failing to meet the minimum inclusion criteria. | 317 | This value is only mentioned in these supplementary materials. It is the number of specimens associated with the Vendor A and Vendor B outbreaks that had at least 1 marker successfully sequenced. Only 242 of these 317 met the criteria for downstream analysis. These 317 specimens were considered when calculating the value of epsilon. |

CDC = Centers for Disease Control and Prevention, NY = New York, TX = Texa

# Supplementary Results

## Sequencing and amplification success for each sequencing laboratory protocol

At the CDC laboratory where genotyping was attempted on 686 stool specimens (see Table S3 – above), PCR amplification success rates varied between the markers. The nuclear markers (markers 1 to 6) were amplified at rates between 78% and 83% while the mitochondrial targets (markers 7 and 8) amplified for approximately 90% of specimens. Sequencing success rates were lower than the success rates obtained for PCR amplification, ranging from between 48% and 68% for nuclear markers and from 76% to 84% of the mitochondrial markers (Figure S1, Table S2).



### Figure S1. Proportion of markers (M) that were successfully sequenced and assigned a genotype for each participating laboratory.

### Table S4. Number of genotyping markers sequenced (out of 8 markers total) for the fecal specimens obtained from cyclosporiasis case-patients meeting the inclusion criteria

|  |  |  |
| --- | --- | --- |
| **Number of markers successfully sequenced** | **Number of specimens** | **Percentage** |
| 8 | 229 | 34.4% |
| 7 | 140 | 21.0% |
| 6 | 122 | 18.3% |
| 5 | 88 | 13.2% |
| 4 | 87 | 13.1% |
| TOTAL | 666 | 100% |

Note: These values include validation specimens (n = 666)

At the MN laboratory where 32 specimens were tested, there was 100% concordance between amplification and sequencing success. Amplification of marker 6 was the least efficient (91%) and the highest rate of amplification success was for target 8 (100%), noting that the MN laboratory only attempted sequencing for specimens that were positive by agarose gel electrophoresis. In the TX laboratory, a lower amplification success rate was observed for targets 1 to 4, though the sequencing success rate was approximately 90% for all markers except for marker 6, which was slightly lower at 83%. For specimens processed in the NY laboratory, amplification success rates were also below the sequencing success rates. Targets 1-4 and 7 were successfully sequenced for more than 80% of specimens, and markers 5, 6 and 8 for more than 70% of specimens (Figure S1 - above).

## Proficiency/validation testing

The six validation specimens genotyped at the TX, NY and CDC laboratories shared similar haplotypes though were not always identical, and typing at some markers sometimes failed due to insufficient sequencing depth (Figure S2 - below). Despite this, each repeat specimen was assigned to the same genetic cluster as its partners based on the 10 cluster model. For nine negative stool specimens, sequencing was attempted at each marker despite the absence of a visible PCR product following gel electrophoresis. The resulting sequence data contained no detectable *C. cayetanensis* sequences. Similarly, for specimens provided to CDC by state laboratories that were positive for other gastrointestinal pathogens (i.e. for *Cryptosporidium* sp., *Salmonella* sp. and pathogenic *E. coli*), no *C. cayetanensis* sequences were obtained.



### Figure S2. Visualization of validation (VAL) specimen genotypes as determined by three independent laboratories (CDC, NY, TX), and the genotype of repeat specimens from two patients.

The horizontal position of black bars indicates detection of a specific haplotype. Background colors are a visual aid to help differentiate the various markers and hold no other meaning. Haplotypes detected for validation specimens sequenced in each laboratory (CDC, TX, NY) were similar, though with minor differences. Despite these differences, genetic cluster assignments were concordant for each validation specimen. Two repeat specimens (C\_WI091\_18 and C\_WI120\_18, shaded gray) were collected on the same day, while the other repeat specimens (C\_WI053\_18 and C\_WI157\_18, shaded blue) were collected 14 days apart. C-1 to C-4 refers to markers CDS-1 to CDS-4.

## Examination of discordantly clustered specimens

Overall, 15 discordant cluster assignments were observed, where the ensemble distances generated culminated in the assignment of specimens to a genetic cluster that did not correspond with that specimen’s epidemiologic linkage. To identify potential causes of these false positive and/or false negative (discordant) clustering results (see definitions in Table 2, main manuscript text), the genotype of discordantly clustered specimens was manually compared to the consensus genotype of their respective epidemiologically-defined clusters (Supplementary File S2, Tabs C and D) (Figure S3 - below). Illness onset dates for the associated case-patients who submitted these discordantly clustered specimens were also examined (Table S5 - below). After examining the sequencing results manually, we concluded that the genotype of 11 of the 15 discordantly clustered specimens was not consistent with the consensus genotype for their respective epidemiologic clusters. Illness onset dates for five of these 15 case-patients occurred earlier or later than the peak illness dates for their respective epidemiologic clusters. For three of these 15 discordantly clustered specimens, both the genotype and illness onset dates did not support their epidemiologic linkage (Table S5 - below).



### Figure S3. Haplotypes (black bars) detected in discordantly clustered specimens compared to the consensus genotype of the Vendor A and Vendor B clusters (blue bars)

The horizontal position of black bars indicates detection of a specific haplotype for each specimen. Gray bars indicate expected haplotypes based on the consensus genotypes observed for each epidemiologic (Epi) cluster (shown in dark blue). Results (Res) shown in red are discordant classifications where the genotype observed seems unrelated to the consensus genotype for the specific epidemiologic cluster. Results (Res) in black indicate samples with genotypes closely related to genotypes from samples epidemiologically assigned to the Vendor B (VB) outbreak despite not being assigned to genetic cluster (Cl) 4. Ten false negative (FN) classifications occurred due to exclusion of specimens from genetic cluster 7 (associated with Vendor A), yet their genotype was not consistent with the consensus genotype of the Vendor A (VA) outbreak (upper panel). Five false negative classifications occurred due to exclusion of specimens from genetic cluster 4 (associated with Vendor B), yet the genotype of one of these (C\_IA008\_18) was not consistent with VB (lower panel). One false positive classification was identified (lower panel), where a specimen linked to the VA outbreak (associated with genetic cluster 7) was assigned to genetic cluster 4 (associated with VB). However, the genotype of that false positive (C\_IA043\_18) suggests its assignment to cluster 4 might be plausible (lower panel). Background colors are a visual aid to help differentiate the markers and hold no other meaning. C-1 to C-4 refers to markers CDS-1 to CDS-4.

If we manually examine some of the discordantly clustered genotypes shown in Figure S3, it becomes apparent that for some of these specimens, the ensemble may have led to an appropriate cluster assignment despite the specimens associated epidemiologic linkage. For example, Specimen C\_IA008\_18 was epidemiologically assigned to the Vendor B outbreak but shared only five out of its 18 haplotypes with the Vendor B consensus genotype (Figure S3). Consequently, C\_IA008\_18 was assigned to genetic cluster 3 while the majority of Vendor B-associated specimens were placed in genetic cluster 4. Similarly, specimen C\_IL030\_18 was epidemiologically associated with the Vendor A outbreak and the vast majority of specimens associated with this outbreak were assigned to genetic cluster 7. However, C\_IL030\_18 differs from the Vendor A consensus by 12 haplotypes and was placed in genetic cluster 1. Considering that the respective consensus genotypes of the *C. cayetanensis* responsible for the Vendor A and Vendor B outbreaks differ by only four haplotypes (Figure S3, blue barcodes), and noting that specimens like C\_IA008\_18 and C\_IL030\_18 represent false negative results, we propose that our performance assessment (based on this epidemiologic gold standard) probably represents a conservative ‘lower boundary’ of the true performance of our ensemble-based distance statistic. As this study was retrospective, we did not readjust the epidemiologic linkages based on our genotyping results. However, these observations highlight the value this approach could provide if used to complement epidemiologic investigations in real time.

### Table S5. Illness onset dates for patients with discordantly clustered specimens

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Specimen name** | **Illness onset date** | **Genetic cluster** | **Epi-linkage** | **Outbreak period for this epidemiologic cluster α** | **Classification** | **Onset consistent with epi-link? α** | **Genotype consistent with epi-link?** **β** |
| C\_IL143\_18 | June 10 | 10 | VA | June 20 to July 6 | False neg. | No | No |
| C\_IL144\_18 | July 1 | 9 | VA | June 20 to July 6 | False neg. | Yes | No |
| C\_WI241\_18 | July 15 | 6 | VA | June 20 to July 6 | False neg. | No | No |
| C\_IL070\_18 | June 25 | 6 | VA | June 20 to July 6 | False neg. | Yes | No |
| C\_IA038\_18 | June 29 | 6 | VA | June 20 to July 6 | False neg. | Yes | No |
| C\_IA047\_18 | June 29 | 6 | VA | June 20 to July 6 | False neg. | Yes | No |
| C\_IA043\_18\* | July 2 | 4 | VA | June 20 to July 6 | False neg. | Yes | No |
| C\_IL145\_18 | July 1 | 2 | VA | June 20 to July 6 | False neg. | Yes | No |
| C\_IA046\_18 | June 26 | 2 | VA | June 20 to July 6 | False neg. | Yes | No |
| C\_IL030\_18 | June 19 | 1 | VA | June 20 to July 6 | False neg. γ | No | No |
| C\_WI012\_18 | May 16 | 10 | VB | May 20 to June 7 | False neg. | No | Yes |
| C\_WI157\_18 | May 30 | 6 | VB | May 20 to June 7 | False neg. | Yes | Yes |
| C\_WI180\_18 | June 10 | 6 | VB | May 20 to June 7 | False neg. | No | Yes |
| C\_WI117\_18 | May 30 | 6 | VB | May 20 to June 7 | False neg. | Yes | Yes |
| C\_IA043\_18\* | July 2 | 4 | VA | June 20 to July 6 | False pos. | Yes | No |
| C\_IA008\_18 | May 30 | 3 | VB | May 20 to June 7 | False neg. | Yes | No |

α Refer to Figure 3 and Supplementary File S2, Tab B for dates

β Refer to Figure 5 for genotypes of these specimens.

γ Borderline result. Falls one day before the beginning of this outbreak.

\* This specimen is repeated in this table because it represents both a false positive and false negative classification based on the definitions in Table 2 – main manuscript text.

Note: Specimens where the onset of illness date and the genotype are both inconsistent with this epi-cluster

## Consecutive specimens from the same patient

In two instances, consecutive specimens were collected from the same patient. For one of these patients, their two specimens were collected on the same day and these were assigned to the same genetic cluster (Figure S2). In the second instance, the case-patients illness was epidemiologically linked to the vendor B outbreak and the specimens were collected 14 days apart (C\_WI157\_18 and C\_WI053\_18). These two specimens were assigned to different genetic clusters; C\_WI053\_18 to cluster 4 (a true positive for Vendor B) and C\_WI157\_18 to cluster 6 (a false negative result). This discordant result may be attributed to the steady reduction in parasite load over time as the infection progressed, reducing our ability to detect all haplotypes as the infection resolved. Regardless, the set of haplotypes observed in C\_WI157\_18 (Figure S2), support that assignment of this specimen to cluster 6 was still valid; for the most part, specimens from clusters 4 and 6 differ by a single haplotype observed at marker 8.

# Supplementary discussion

## Notes on amplification and sequencing success

These results indicate that deep amplicon sequencing is more sensitive than agarose gel electrophoresis for DNA detection, as useable sequence data was often obtained even in the absence of a band following agarose gel electrophoresis. Additionally, the variations in amplicon purification and/or library normalization employed by the different laboratories clearly had an impact on sequencing success, which is a useful observation from the perspective of laboratory protocol optimization to maximize genotyping success. Prior to genotyping at the NY and MN laboratories, specimens were excluded based on real-time PCR cycle threshold, where only specimens obtaining a Ct ≤ 35 (NY) and ≤ 32 (MN) were subjected to genotyping. At the TX and CDC laboratories, genotyping was attempted for every specimen that received a positive diagnosis regardless of the diagnostic method. Retrospective follow-up of genotyped specimens at the TX lab by morphologic examination of the stool and real-time PCR demonstrated that all specimens obtaining a Ct ≥ 34 led to a failed genotyping attempt, regardless of whether *C. cayetanensis* was present following microscopic examination of the stool specimen. At CDC, retrospective real-time PCR performed on all specimens where genotyping was attempted, later demonstrated that 25% of the specimens had a Ct ≥ 35 and 11% of specimens where typing was attempted were completely negative by real-time PCR.

## Proposed modifications to laboratory protocols based on sequencing and amplification success

Firstly, excluding specimens based on a high real-time PCR cycle threshold value was valuable for exclusion of specimens that were unlikely to yield genotyping results. This will ultimately reduce costs as genotyping would not be attempted on specimens that will likely lead to genotyping failures. Next, regardless of the purification procedure used, normalization for DNA concentration of each amplicon individually prior to pooling amplicons from the same specimen is superior to pooling PCR amplicons before normalization. At CDC, an average of 18% of PCR products processed did not produce sufficient sequence data for haplotype assignment. This is because the raw PCR products were pooled prior to normalization, despite the fact that PCR product concentrations varied between markers amplified from the same specimen, probably due to differences in amplification efficiency. This practice was implemented to increase throughput. However, with a binding capacity of ~25 ng for the SequalPrep Kit (used for purification and normalization at CDC), the markers detected at CDC were likely biased towards those with a higher amplification efficiency, leading to a higher abundance in the amplicon pool. This typically favored the mitochondrial markers (7 and 8), which had the greatest amplification success.

At the TX laboratory, a Monarch DNA Purification Kit (New England Biolabs, Ipswich, MA, USA) was used to purify the eight PCR amplicons after they had been pooled. This approach did not suffer from preferred selectivity of any one amplicon, or loss of any specific amplicon, probably due to the column recovery capacity being up to 5 μg of DNA, far higher than the 25 ng capacity of the SequelPrep kit. At the TX and NY laboratories, quantification and normalization were performed after this purification step, and before library preparation. At the MN laboratory, amplicons were normalized individually prior to pooling, resulting in a more homogenous concentration for each marker. Additionally, in cases where a PCR product was not detected upon gel electrophoresis, usable sequence data was often recovered. As such, for specimens passing the real-time PCR cycle threshold exclusion criterion, agarose gel electrophoresis is probably only necessary for positive and negative PCR controls rather than every amplicon, only to ensure that the reaction has occurred correctly.

## Robustness of clustering despite missing data

Only 34.4% of specimens that met the inclusion criteria for classification obtained a complete genotype, and more than 25% of all specimens were successfully genotyped at only 4 or 5 markers. Among specimens retained for analysis, sequencing success was highest for MSR and lowest for CDS-1 and CDS-4 (Table S2). It is worth noting that of the specimens retained for analysis, 106 of 116 cyclosporiasis case-patients whose illness was epidemiologically linked to the Vendor A outbreak had their specimens correctly assigned to genetic cluster 7, and 76 of these specimens had an incomplete genotype. Furthermore, 17 of those specimens had data available for only 4 of 8 markers yet were still correctly clustered. Similarly, of the specimens retained for analysis, 121 of 126 cyclosporiasis case-patients whose illness was epidemiologically linked to the Vendor B outbreak had their specimen correctly assigned to genetic cluster 4, and 84 of these had an incomplete genotype. As with the Vendor A outbreak, 17 of these had data available for only 4 markers, yet the specimens were correctly clustered.

# Appendix

Appendix 1. Reference sequences of known *Cyclospora cayetanensis* haplotypes used in this study

>360i2\_A\_Hap\_1

GGCCTGCCCCTAAATCTTTAATCGGAACGCTCTAAGAGAGCGTACAGCACTTTGTTCGTGCGCTGCAGCTGTTGCGGGAGAAGCAGGGCTATCTGATGTCA

>360i2\_A\_Hap\_2

GGCCTGCCCCTAAATCTTTAATCGGAACGTTGTAAGAGAGCGTACAGCACTTTGTTCGTGCGCTGCAGCTGTTGCGGGAGAAGCAGCGCTATCTGATGTCA

>360i2\_B\_Hap\_1

TTGAGTTTTGATGAGAAAAGGTGTGTTCCTAGGATTAGGCGTTTTGCATGCAGGAATCAAGCCAGACTCGAAAAAATTGAGGCCCTCTTGGCATACCCTTT

>360i2\_B\_Hap\_2

TTGAGTTTTGATGAGAACAGGCGTGTTCCTGGGATTAGGCGTTTTGCATGCAGAAATCAAGCCAGACTCAGAAAAATTGAGGCTCTCTTGGCATACCCTTT

>360i2\_B\_Hap\_3

TTGAGTTTTGATGAGAACAGGCGTGTTCCTGGGATAAGGCGTTTTGCATGCAGGAATCAAGCCAGACTCAGAAAAATTGAGGCCCTCTTGGCATACCCTTT

>360i2\_B\_Hap\_4

TTGAGTTTTGATGAGAACAGGCGTGTTCCTGGGATTAGGCGTTTTGCATGCAGGAATCAAGCCAGACTCAGAAAAATTGAGGCCCTCTTGGCATACCCTTT

>360i2\_B\_Hap\_5

TTGAGTTTTGATGAGAACAGGTGTGTTCCTGGGATTAGGCGTTTTGCATGCAGAAATCAAGCCAGACTCAGAAAAATTGAGGCTCTCTTGGCATACCCTTT

>360i2\_B\_Hap\_6

TTGAGTTTTGATGAGAACAGGCGTGTTCCTGGGATTAGGCGTTTTGCATGCAGAAATCAAGCAAGACTCAGAAAAATTGAGGCTCTCTTGGCATACCCTTT

>360i2\_C\_Hap\_1

GTGCTCGGATCGTTGCTCTCCTAACTTACTAATCCGAGAAGATGCAATCTGTGGATGGGGCAGAAGACGACGATACATCAGCAAGAAGCCCAGCCGCCGC

>360i2\_C\_Hap\_2

GTGCTCGGATCGTTGCTCTCCTAACTTACTAATCCGAGAAGATGCAATCTGTGGATGGGGCAGAAGACGACGATACGTCAGCAAGAAGCCCAGCCGCCGC

>360i2\_C\_Hap\_3

GTGCTCGGATCGTTGCTCTCCTAACTTACTTATCTGAGAAGATGCAATCCGTGGATGGGGCAGAAGACGACGATACATCAGCAAGAAGCCCAGCCGCCGC

>360i2\_C\_Hap\_4

GTGCTCGGATCGTTGCTCTCCTAACTTACTTATCCGAGAAGATGCAATCTGTGGATGGGGCAGAAGACGACGATACATCAGCAAGAAGCCCAGCCGCCGC

>360i2\_D\_Hap\_1

TGCTAGTGACGCAACATTAGGTGACATTCCCTCCCCACAGGCTACGGATTATGAAAAATACACCAAAATGACCGCGTGCCATTCTCTTTGCTTGAGCAGAACCA

>360i2\_D\_Hap\_2

TGCTAGTGACGCAACATTAGGTGACATTCCCTCCCCACAAGCTACGGATTATGAAAAATACACCAAAATGACTGCGTGCCATTCTCTTTGCTTGAGCGGACCCA

>378\_A\_Hap\_1

ACAAAGTGCAACCAACTCACTCCTACTCTTGCTCCTGCATCTACCACCGAAAAATATGCAGACGCAAGGAACGTTCTATGGACGCACATCGCTCTGCAGAGCCC

>378\_A\_Hap\_2

ATAAAGTGCAACCAACTCACTCCTACTCTTGCTCCTGCATCTACCACCGAAAAATATGCAGACACGAGGAACGTTCTTTGGATGCAAATTCCTCTGCAGAGCCC

>378\_A\_Hap\_3

ACAAAGTGCAACCAACTCACTCCTACTCTTGCTCCTGCATCTACCACCGAAAAATATGCAGACACAAGGAACGTTCTATGGATGCACATCGCTCTGCAGAGCCC

>378\_A\_Hap\_4

ACAAAGTGCAACCAACTCACTCCTACTCTTGCTCCTGCATCTACCACCGAAAAATATGCAGACGCAAGGAACGTTCTATGGATGCAAATTCCTCTGCAGAGCCC

>378\_A\_Hap\_5

ATGAAGTGCAACCAACTCACTCCTACTCTTGCTCCTGCATCTACCACCGAAAAATATGCAGACACGAGGAACGTTCTGTGGATGCAAATTCCTCTGCAGAGCCC

>378\_A\_Hap\_6

ACAAAGTGCAACCAATTCACTCCTACTCTTGCTCCTGCATCTACCACCGGAAAATATGCAGACACGAGGAACGTTCTATGGATGCAAATTCCTCTGCAGAGCCC

>378\_A\_Hap\_7

ATAAAGTGCAACCAACTCACTCCTACTCTTGCTCCTGCATCTACCACCGAAAAATATGCAGACGCAAGGAACGTTCTTTGGATGCAAATTCCTCTGCAGAGCCC

>378\_B\_Hap\_1

GCCTCCCCAATCGCCCTATCCTGTCGAACGACTGACCTCAAAAAGAACTCAAGGCAAAAGATGACCAGCCGCGTCACCTTCTCCTCTGATCCAGACTCTCGCTGCAA

>378\_B\_Hap\_2

GCCTCCCCAATCGCCCTATTTTGTCGAACGACTGACCTCAAAAAGAACTCAAGGCAAAAGATGACCAGCCGCGTCACCTTCTCCTCTGATCCAGACTCTCGCTGCAA

>378\_C\_Hap\_1

GTACAAGGAGTGGACGCATCTTTCGGTCCCTTCCGTGCCAGTAACATGCACCTACATTCCTCAGTTCAGAATACACCTCGGACTCCAACACCTGGTGCGACAATC

>378\_C\_Hap\_2

GTACAAGGAGTGGACGCATCTTTCGGTCTCTTCCGTGCCAGTAACATGCACCTACATTCCTCAGTTCAGAATACACCTCGGACTCCAACACCTGGTACGACAATC

>378\_C\_Hap\_3

GTACAAGGAGTGGACGCATCTTTCGGTCCCTTCCGTGCCAGTAACATGCACCTACATTCCTCAGTTCAGAATACACCTCGGACTCCAACACCTGGTACGACAATC

>378\_C\_Hap\_4

GTACAAGGAGTGGACGCATCTTTCGGTCTCTTCCGTGCCAGTAACATGCACCTACATTCCTCAGTTCAGAATACACCTCGGACTCCAACACCTGGTAAGACAATC

>Cmt109.A\_Junction\_Hap\_1

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt127.A\_Junction\_Hap\_2

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGTCTCTTATACACATCTC

>Cmt154.A\_Junction\_Hap\_3

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTTTAATAGTACTATTATAAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt154.B\_Junction\_Hap\_4

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt154.C\_Junction\_Hap\_5

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt.154.D\_Junction\_Hap\_6

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTTTAATAGTACTATTTTAAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt154.E\_Junction\_Hap\_7

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt169.A\_Junction\_Hap\_8

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt169.B\_Junction\_Hap\_9

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTATAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt169.C\_Junction\_Hap\_10

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTTTAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt184.A\_Junction\_Hap\_11

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTTAAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt184.B\_Junction\_Hap\_12

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt184.C\_Junction\_Hap\_13

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTATAATAGTATTATTTATAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt184.D\_Junction\_Hap\_14

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTATAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt184.E\_Junction\_Hap\_15

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTTTAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt184.F\_Junction\_Hap\_16

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt199.A\_Junction\_Hap\_17

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt199.B\_Junction\_Hap\_18

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTTTAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt199.C\_Junction\_Hap\_19

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTATAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>Cmt214.A\_Junction\_Hap\_20

TACCAAAGCATCCATCTACAGCTGCGGAAACTGTATTTTTATTATTTAATTTTACTATTTTAAATAGTATTATTTATAATAGTATTATTTATAATAGTATTATTTATAATAGTATTATTTTTAATAGTATTATTTTTAATAGTATTATTTTTAATAGTACTATTTTTAATATAAAAATTTAGTACACCTAGCCAACACGATCCGATTGCTTGGG

>CDS1\_Hap\_1

CTCCTTGCTGCTCAGAACGAGGATCACTCTCCAGGCGACACCCTCCTCGCGCTCTGTCCACCGTTTCTCTACGCTTCCCCATGTCTCTGTGCTTGCATGCATCCGTAGATGCATGTCACCTCTAGCCCTTTGTTGCTTGAGGCGCTGCTTATCCGTTGCCACTGCTCCTCTCTTG

>CDS1\_Hap\_2

CTCCTTGCTGCTCAGAACGAGGATCACTCTCCAGGCGCTGCCCTCCTCGCACTGTGTCCACCCTTTCTCTACGCTTCCCCATGTCTCTGTGCTTGCATGCATCCGTAGACGCATGTCACCTCTAGCCCTTTGTTGCTTGAGGCGCTGCTTATCCGTTGCCACTGCTCCTCTCTTG

>CDS2\_Hap\_1

TGCAAACTACTAAGGGCGCACCAGATAGGTGGAGGCGCAGATGGCCTTAAGAGCAAGAATGCCGGACTGGTGATGCTGGCGAGGGGTGATGTTGCCGCGGCGACTGCTGCTAGATCCCCTGCTGCTGCCGCAGCGGCAGCAGCGGAAGTTCTGCGAACCGCCTCCCGTTCGTCCCTCGTGTCTACGCGAAGCGCCAAGAAGTTGCAGCATACAAAGAAGCAGCTTGTCAAGGCTCAAGAGAAGGCG

>CDS2\_Hap\_2

TGCAAACTACTAAGGGCGCACCAGATAGGTGGAGGCGCAGATGGCCTTAAGAGCAAGAATGCCGGACTGGTGATGCTGGCGAGGGGCGATGTTGCCGCGGCGACTGCTGCTAGATCCCCTGCTGCTGCCGCAGCGGCAGCAGCGGAAGTTCTGCGAACCGCCTCCCGTTCGTCCCTCGTGTCTACGCGAAGCGCCAAGAAGTTGCAGCATACAAAGAAGCAGCTTGTCAAGGCTCAAGAGAAGGCG

>CDS3\_Hap\_1

AATCGAATCGGTGCAGTGCTTAGTTGCTTTATCTTTGGCAGTTAGACAGGGAGGGATTCCTACATTTGCAGGTGATGCTGCTACCTCCATGCGCAGCTGCAACACCAGCAACATCTACAAGTGCACCAGCAGAACTAACCCTGTCACGAAAGGCGGCACGAGTGCTGCGCTAGAGGCTTCTCGGTTTGTTCAGCCGCCGCCCCCTCTCACACGTTCAGTC

>CDS3\_Hap\_2

AATCGAATCGGTGCAGTGCTTAGTTGCTTTATCTTTGGCAGTTAGACAGGGAGGGATTCCTACATTTGCAGGTGATGCTGCTACCTCCGTGCGCAGCTGCAACACCAGCAACATCTACAAGTGCACCAGCAGAACTAACCCTGTCACGAAAGGCGGCACGAGTGCTGCGCTAGAGGCTTCTCGGTTTGTTCAGCCGCCGCCCCCTCTCACACGTTCAGTC

>CDS3\_Hap\_3

AATCGAATCGGTGCAGTGCTTAGTTGCTTTATCTTTGGCAGTTAGACAGGGAGGGATTCCTACATTTGCAGGTGATGCTGATACCTCCGTGCGCAGCTGCAACACCAGCAACATCTACAAGTGCACCAGCAGAACTAACCCTGTCACGAAAGGCGGCACGAGTGCTGCGCTAGAGGCTTCTCGGTTTGTTCAGCCGCCGCCCCCTCTCACACGTTCAGTC

>CDS4\_Hap\_1

GTAGATGGGTCCTTGAAGGCTTCCGACGCTTCTGATGCCGCAGCTTCCATTCAGATCGACGACGTCGAGTCATCTGGTGAAGATGACTCGGTGGAGACGCCAGAAATCTCCGCCGCTCCTGGCACGGCCTCAGCATCCGTAGCCACAGCCTCGTCCACTTTCGGTTCCTTAGGCGTCTG

>CDS4\_Hap\_2

GTAGATGGGTCCTTGAAGGCTTCCGACGCTTCTGAGACCGCAGCTTCCATTCAGATCGACGACGTCGAGTCATCTGGTGAAGATGACTCGGTGGAGACGCCAGAAATCTCTGCCGCTCCTGGCACGGCCTCAGCATCCGTAGCCACAGCCTCGTCCACTTTCGGTTCCTTAGGCGTCTG

>MSR\_Left\_Hap\_1

TAACTTCTGGAGTCCCTTTTGGTAAAAAACAGTACAAAGTTGGACTGAGGCTGAGTCTCTATGCCGAATGAGGAGCACTGGATTGGATACCCAGGTACCAGCGCTTCCATTAATAAGAAAGTGAGAAGA

>MSR\_Left\_Hap\_2

TAACTTATGGAGTCCCTTTTGGTAAAAAACAGTACAAAGTTGGACTGAGGCTGAGTCTCTATGCCGAATTAGGAGCACTGGATTGGATACCCAGGTACCAGCGCTTCCATTAATAAGAAAGTGAGAAGA

>MSR\_Left\_Hap\_3

TAACTTATGGAGTCCCTTTTGGTAAAAAACAGTACAAAGTTGGACTGAGGCTGAGTCTCTATGCCGAATGAGGAGCACTGGATTGGATACCCAGGTACCAGCGCTTCCATTAATAAGAAAGTGAGAAGA

>MSR\_Left\_Hap\_4

TAACTTCTGGAGTCCCTTTTGGTAAAAAACAGTACAAAGTTGGACTGAGGCTGAGTCTCTATGCCGAATGAGGAGCACTGGATTGGATACCCAGGTACCAGCGCTTCCATTAATAAGAAAGTGATAAGA

>MSR\_Right\_Hap\_1

CTTAGTAACGATTCCGTCTTCCAGCTGCCAATTTTTATTTCAAATGTTCCACTACCAAATTATATCCATATGTTCCAAATGTTCAGGAGTGTATGTAATAAGTTTTT

>MSR\_Right\_Hap\_2

CTTAGTAACGATTCCGTCTTCCAGCTGCCAATTTTTATTTCAAATGTTCCACTACCAAATTATATCCATCTGTTCCAAATGTTCAGGAGTGTATGTAATAAGTTTTT

>MSR\_Right\_Hap\_3

CTTAGTAACGATTCCGTCTTCCAGCTGCCAATTTATATTTCAAATGTTCCACTACCAAATTATATCCATCTGTTCCAAATGTTCAGGAGTGTATGTAATAAGTTTTT

1. Modified from PulseNet: <https://www.cdc.gov/pulsenet/pathogens/protocols.html> [↑](#footnote-ref-1)