The Journal of Infectious Diseases

BRIEF REPORT

# Genotyping *Cyclospora cayetanensis* From Multiple Outbreak Clusters With An Emphasis on a Cluster Linked to Bagged Salad Mix—United States, 2020

**Joel Barratt,[1,a] Lauren Ahart,[1,a] Marion Rice,[1] Katelyn Houghton,[1] Travis Richins,[1] Vitaliano Cama,[1] Michael Arrowood,[2] and Yvonne Qvarnstrom,[1] and Anne Straily[1]**

[1]Parasitic Diseases Branch, Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, Georgia, USA,[2]Waterborne Disease Prevention Branch, Division of Foodborne, Waterborne, and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Cyclosporiasis is a diarrheal illness caused by the foodborne parasite *Cyclospora cayetanensis*. Annually reported cases have been increasing in the United States prompting development of genotyping tools to aid cluster detection. A recently developed *Cyclospora* genotyping system based on 8 genetic markers was applied to clinical samples collected during the cyclosporiasis peak period of 2020, facilitating assessment of its epidemiologic utility. While the system performed well and helped inform epidemiologic investigations, inclusion of additional markers to improve cluster detection was supported. Consequently, investigations have commenced to identify additional markers to enhance performance.

**Keywords.** *Cyclospora cayetanensis*; cyclosporiasis; outbreaks; genotyping; epidemiology.

Cyclosporiasis is a diarrheal illness caused by the parasite *Cyclospora cayetanensis*. Cases of foodborne cyclosporiasis in the United States typically occur during spring and summer, and numbers of annually reported cases have been increasing [1]. Cyclosporiasis clusters are challenging to identify and investigate, partly because validated tools to genetically link cases are lacking. A promising *Cyclospora* genotyping tool developed at the Centers for Disease Control and Prevention (CDC) uses targeted deep amplicon sequencing of 2 mitochondrial and 6 nuclear targets; analysis of resulting sequences is completed using an ensemble of algorithms to resolve clusters [2]. This tool was evaluated over multiple years by retrospective comparison of genetic clusters to epidemiologically defined clusters, supporting the tools excellent sensitivity and specificity [2, 3]. Evaluation of the CDC's algorithms on genotyping data from other parasites also yielded promising results [4].

During the 2020 cyclosporiasis season, this genotyping tool was used by the Parasitic Diseases Branch at the CDC in near real time to complement epidemiologic cluster investigations. This facilitated detection of multiple genetic clusters and a further appraisal of this method on a novel data set. Our results supported the excellent epidemiologic utility of the method, though opportunities for further development were identified.

## METHODS

In parallel with epidemiologic investigations, *Cyclospora* in stool specimens submitted by state health departments (SHDs) were genotyped at the CDC. In addition, sequence data from 2 SHDs (Texas and New York State) and the Public Health Agency of Canada (PHAC), were sent to the CDC for analysis. Sequencing and identification of genetic links were performed as described elsewhere [2]. Briefly, amplicons were Illumina sequenced and resulting data were analyzed using a workflow comprising various bioinformatic modules that (1) characterize each specimens' genotype, (2) compute genetic distances using the CDC ensemble, and (3) establish a genetic distance threshold, where genotypes separated by distances below this threshold are considered genetically linked [3].

Genetically linked specimens collected within the same 2-week period were considered a temporally supported genetic cluster and were sequentially assigned temporal-genetic cluster (TGC) codes (designated 2020_xxx). A "sliding window" strategy for fine-tuning TGCs was also applied, where specimens not collected within the 2-week period initially defined for a TGC but linked genetically to specimens assigned to that TGC and collected within 7 days before or after its defined period, were also assigned to that TGC. The earliest and latest collection dates for that TGC were then adjusted (if required) according to collection dates of newly added specimens.

This process was repeated until genetically linked specimens were no longer detected 7 days before or after the latest dates. The initial 2-week window is based on the shelf life of refrigerated produce (the main vehicle of cyclosporiasis), which is approximately 2 weeks. The sliding window accounts for variation in the time patients take to submit fecal specimens for laboratory testing after illness onset. TGCs were shared weekly with cyclosporiasis surveillance epidemiologists, who reviewed available case report data linked to genotyping results and determined associations with known outbreaks and/or identified

Received 14 July 2021; editorial decision 23 September 2021; accepted 24 September 2021; published online 4 October 2021.

[a]J. B. and L. A. contributed equally to this work.

Correspondence: Joel Barratt, US Centers for Disease Control and Prevention, Roybal Campus, 1600 Clifton Rd, H23-9, Atlanta, GA 30333, USA (jbarratt@cdc.gov).

**The Journal of Infectious Diseases® 2022;225:2176–80**


https://doi.org/10.1093/infdis/jiab495

frequently reported food items among cases within a TGC. This activity was reviewed by the CDC and was conducted consistent with applicable federal law and CDC policy (Center for Global Health Human Research Protection Office determination 2018-123).

## RESULTS

Between May and August 2020, a total of 1441 laboratory-confirmed cyclosporiasis cases were reported by 34 states and New York City. On 15 June 2020, the CDC was first notified of a multistate outbreak potentially linked to bagged salads containing iceberg lettuce, carrots, and red cabbage sold at several grocery stores in the midwestern United States. By the end of 2020, 705 laboratory-confirmed cases had been linked to the bagged salad mix. Of the 736 cases not linked to the bagged salads, some were linked to either other produce items, restaurants, or events or could not be linked to a specific vehicle. A total of 1019 specimens from the 1441 case patients were submitted to the CDC for genotyping (964 by SHDs and 55 by PHAC; numbers include sequence data submitted to the CDC by PHAC and by the SDHs from New York State and Texas, which perform sequencing at their own facilities and send data to the CDC for analysis); sequence data of sufficient quality for genotyping were obtained from 816 specimens.

By July, 3 TGCs had emerged. TGCs 2020_001 and 2020_003 appeared to be related to the bagged salads based on epidemiologic data available for some specimens in those TGCs at that time. Samples in TGC 003 had nuclear genotypes in common with specimens in TGC 001 but had a distinct mitochondrial genotype. Onset dates for cases in TGCs 001 and 003 overlapped considerably (Figure 1A), and several specimens assigned to either 001 or 003 demonstrated mixed 001 and 003 mitochondrial genotypes, further supporting that these clusters were attributable to 1 source. When genotyping for 2020 concluded, 163 specimens had been assigned to TGC 001 and 112 to TGC 003; of these 275 specimens, 189 (69%) were from cases epidemiologically linked to bagged salad mix. This confirmed the relationship between bagged salads and TGCs 001 and 003 initially suspected in July. Epidemiologic linkage was not possible for the remaining 86 specimens because of missing, inconclusive, or conflicting data. The genetic relationship between TGCs 001 and 003 was detected solely because of nuclear similarities and mitochondrial differences were sufficient to separate TGCs 001 and 003. Regardless, the distinction between these TGCs is small, as the specimens belonged to the same larger genetic cluster (Figure 1B).

In 2020, a total 26 TGCs were identified (2020_001 to 2020_026). However, the algorithms used to assess genetic relatedness compute distances reflecting the likelihood of isolate pairs being unrelated using data available at the time. Thus, new genotyping data added throughout an investigative period may lead to emergence or dissolution of clusters, in addition to reassignment of isolates to clusters representing improved genetic matches. Consequently, TGCs 002, 004, 006, 008, 010, 011, 012, 014, 015, 016, 020, and 021 dissolved over time, leaving 14 stable TGCs by the end of 2020 (Table 1).

In addition to TGCs 001 and 003 (associated with bagged salads), epidemiologic links were detected among specimens in 8 of the 14 final TGCs. All 10 case patients associated with an epidemiologic cluster linked to restaurants in 3 states (ultimately attributed to cilantro), submitted specimens assigned to TGC 2020_009 (Table 1). Similarly, specimens from 4 case patients with a common exposure at a restaurant chain in New York City (New York City salad chain A) were assigned to 2020_025. Alternatively, epidemiologically linked cases were distributed across multiple TGCs for some outbreaks. Isolates from case patients associated with "grocery store chain A salads/romaine lettuce" were assigned to 3 TGCs: 2020_018 (n = 2), 2020_017 (n = 7), and 2020_026 (n = 3). For 3 TGCs (2020_005, 2020_007, 2020_017), no common exposure was identified, as some specimens within these TGCs were associated with different exposures to each other. Among all specimens assigned to TGCs 2020_013, 2020_019, 2020_022, and 2020_024 (n = 74; Table 1), no epidemiologic links were found. Interestingly, 15 of 20 specimens assigned to 2020_023 were from case patients epidemiologically linked to bagged salads (Table 1).

## DISCUSSION

Comparison of genotyping results to analogous epidemiologic clusters demonstrated that multiple outbreaks would have been detected had genetic linkages been used to inform epidemiologic investigations. This is exemplified by TGCs 001 and 003, for which examination of epidemiologic questionnaires among associated case patients revealed a link to bagged salads. Similarly, it is likely that cilantro would have been identified as a source of exposure for TGC 009. These results corroborate previous studies where the CDC's method provided data to further inform epidemiologic investigations [2, 3]. This work also highlighted opportunities for refinement of that method. For example, the detection of multiple TGCs comprising genetically related isolates could suggest that 8 markers provide insufficient resolution (Figure 1B). The possibility that additional or alternative markers could identify differences not currently captured is worth exploring.

The temporal criteria used to identify emerging clusters deserves further consideration. While illness onset dates are preferred for defining temporal relationships, specimen collection dates were used as a proxy because illness onset dates do not become available to the CDC for days to weeks after laboratory confirmation of a cyclosporiasis case (ie, after patients are interviewed by their SHD and data are transmitted to the CDC). Collection dates are available immediately on specimen receipt, but using them in this context has limitations. For example, the length of time individuals tolerate symptoms before seeking

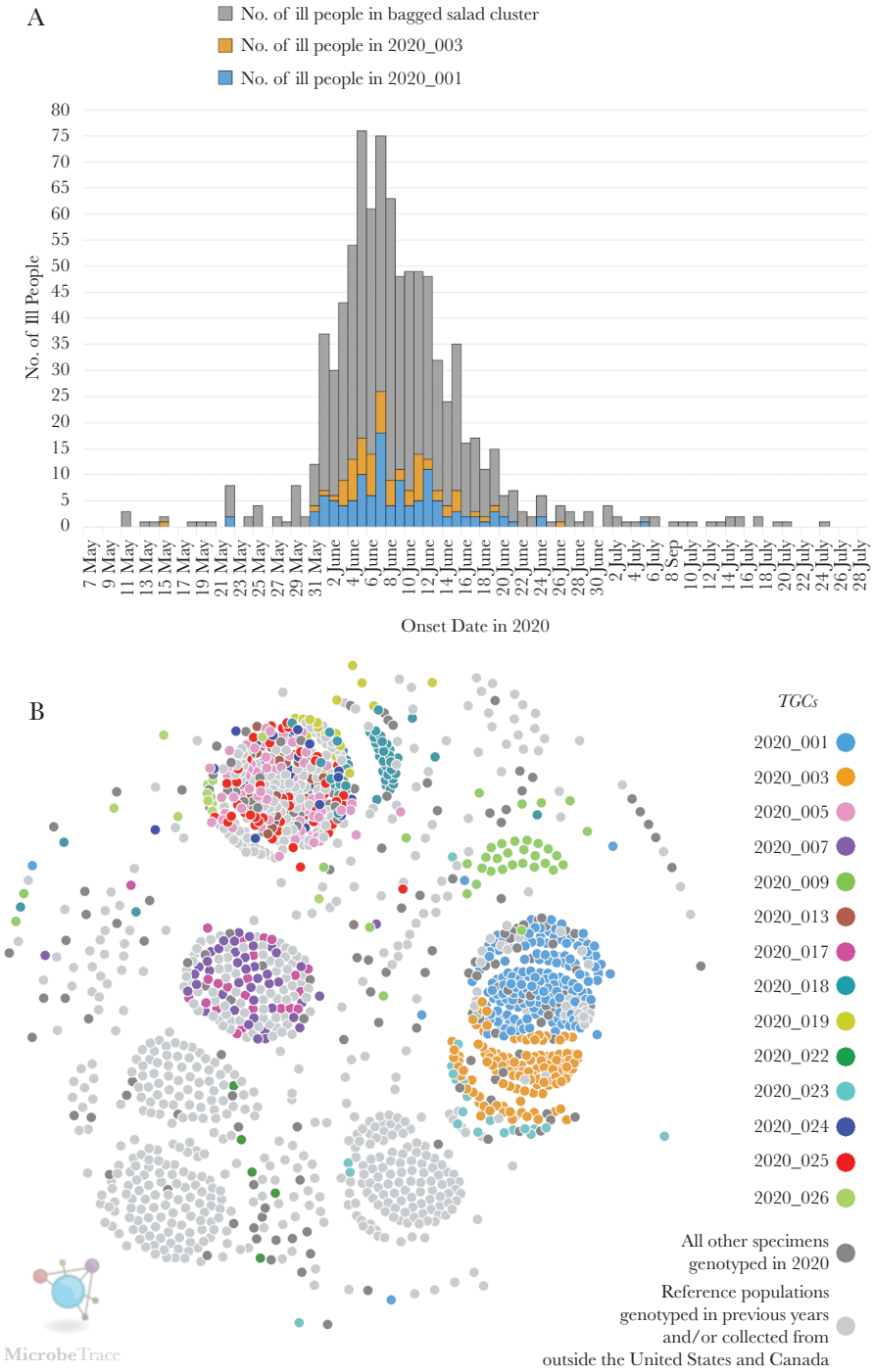


**Figure 1.** Epidemiologic curve of cases linked to the bagged salad outbreak investigation (n = 705) by the end of 2020, by illness onset date (*A*), and MicrobeTrace network generated from a distance matrix containing genotypes analyzed between 24 June and 9 November 2020 (*B*). *A,* Cases linked by genotyping to clusters 2020_001 (n = 116) and 2020_003 (n = 73) and that were also epidemiologically linked to the bagged salad outbreak are highlighted in blue and orange, respectively. Gray represents cases linked to the bagged salad mix outbreak without a specimen provided for genotyping (n = 516). *B,* The matrix used to generate the network shown was calculated using the Barratt-Plucinski ensemble of algorithms, as described elsewhere [2, 3]. Links were filtered in MicrobeTrace [5] to link specimens separated by a distance of <0.15 for visualization purposes. Color coding reflects specimens belonging to each temporal-genetic cluster (TGC), determined from the same distance matrix using Ward's clustering method, as described elsewhere [2]. Some outlying specimens (dots distal from their same-colored partners in 2-dimensional space) appear in this network, reflecting minor discordance between Ward's clustering method and clusters detected using MicrobeTrace. Specimens assigned to clusters 001 (*blue*) and 003 (*orange*) are partitioned into 2 hemispheres of the same larger cluster, based on their mitochondrial genotype. This network includes 778 genotyped specimens submitted to the Centers for Disease Control and Prevention (CDC) by the state health department and 38 submitted by the Public Health Agency of Canada by 9 November 2020, for which sequence data of sufficient quality were obtained (*all colors except light gray*, n = 816), analyzed alongside specimens submitted in previous years and/or those collected from outside the United States or Canada (*light gray*, n = 710). Notably, some genetic clusters (dots positioned closely together in space) are marked with different colors indicating membership in different TGCs despite possessing a similar genotype. This is a consequence of using specimen collection dates to identify outbreaks that may have been epidemiologically unrelated yet were genetically similar based on the CDC's current 8-marker genotyping system.

**Table 1. Numbers of Specimens Assigned to Each Temporal Genetic Cluster Code and Their Epidemiologic Linkages**

| TGC Code[a] | Specimens Assigned to Code, No.[b] | Epidemiologic Linkage of Specimens Assigned to Code (No. of Specimens) |
|---|---|---|
| 2020_001 | 163 | Bagged salad mix (n = 116); unknown linkage (n = 47; 39/47 due to absence of epidemiologic data) |
| 2020_003 | 112 | Bagged salad mix (n = 73; unknown linkage (n = 39; 29/39 due to absence of epidemiologic data) |
| 2020_005 | 69 | Mail grocery delivery service (n = 3); bagged salad mix (n = 5)[c]; unknown linkage (n = 61; 43/61 due to absence of epidemiologic data) |
| 2020_007 | 43 | NYC brand-name salad product (n = 1); NYC salad chain B restaurant (n = 2); NE chain restaurant (n = 1); unknown linkage (n = 39; 26/39 due to absence of epidemiologic data) |
| 2020_009 | 36 | TN/GA/VA restaurant/cilantro subcluster (n = 10); unknown linkage (n = 26; 23/26 due to absence of epidemiologic data) |
| 2020_013 | 23 | Unknown linkage (n = 23; 14/23 due to absence of epidemiologic data) |
| 2020_017 | 23 | NE chain restaurant (n = 4); grocery store chain A salads/romaine lettuce (n = 7); unknown linkage (n = 12; 6/12 due to absence of epidemiologic data) |
| 2020_018 | 51 | Grocery store chain A salads/romaine lettuce (n = 2); unknown linkage (n = 49; 22/49 due to absence of epidemiologic data) |
| 2020_019 | 18 | Unknown linkage (n = 18; 15/18 due to absence of epidemiologic data) |
| 2020_022 | 5 | Unknown linkage (n = 5; 2/5 due to absence of epidemiologic data) |
| 2020_023 | 22 | Bagged salad mix (n = 15)[c]; unknown linkage (n = 7; 5/7 due to absence of epidemiologic data) |
| 2020_024 | 28 | Unknown linkage (n = 28; 14/28 due to absence of epidemiologic data) |
| 2020_025 | 46 | NYC salad chain A (n = 4); unknown linkage (n = 42; 33/42 due to absence of epidemiologic data) |
| 2020_026 | 12 | Grocery store chain A salads/romaine lettuce (n = 3); unknown linkage (n = 9; 6/9 due to absence of epidemiologic data) |
| No code assigned[d] | 165 | IL Mexican-style restaurant (n = 2); mail grocery delivery service (n = 1); NE chain restaurant (n = 1); bagged salad mix (n = 37)[c]; grocery store chain A salads/romaine lettuce (n = 3); unknown linkage (n = 121; 81/121 due to absence of epidemiologic data) |

Abbreviations: GA, Georgia; IL, Illinois; NE, Nebraska; NYC, New York City; TGC, temporal genetic cluster; TN, Tennessee; VA, Virginia.

[a]Only stable TGCs that remained by the end of the 2020 cyclosporiasis season are listed.

[b]A total of 816 specimens were genotyped (778 US specimens and 38 from the Public Health Agency of Canada).

[c]Other specimens epidemiologically linked to the bagged salads not assigned to TGCs 2020_001 or 2020_003. Specimens in TGC 2020_023 are genetically similar to those assigned to 2020_003 (Figure 1), although they were assigned a separate TGC code. Five specimens associated with bagged salads were assigned to TGC 2020_005, representing a discordance between the epidemiologic data and genotyping, as the genotype of these 5 specimens is dissimilar from that of specimens assigned to 2020_001 and 2020_003. In addition, 37 specimens associated with bagged salad mix were not assigned a cluster code (see the following footnote for an explanation).

[d]Genotyping was successful for these specimens, but they were not assigned to a cluster code because their collection dates did not support their belonging to a larger genetic cluster of cyclosporiasis.

medical attention may differ, as might the speed at which they can secure a medical appointment and receive an appropriately diagnosis. To account for this variability, the 7-day sliding window criteria was applied. This strategy assigned related specimens submitted from 4 June to 19 August to TGC 001 and those submitted from 7 June to 29 August to TGC 003, facilitating identification of these large TGCs.

Two major TGCs described here (001 and 003) were derived from a single source: bagged salads. Specimens assigned to these TGCs had different mitochondrial genotypes, although they shared nuclear genotypes. While the 8-marker system nevertheless linked these TGCs to bagged salads, the distinction between clusters 001 and 003, driven solely by mitochondrial markers, may have been inappropriate. The poor understanding of *Cyclospora* gametogenesis and its impact on mitochondrial inheritance complicates interpretation of this result. However, the genotyping system overcame this and closely clustered TGCs 001 and 003 (Figure 1B). As the mitochondrial genome of *Cyclospora* (approximately 6.3 kilobases) is substantially smaller than its nuclear genome (>40 megabases) [6, 7], having 2 of 8 markers be mitochondrion derived may assign too much weight to mitochondrial loci. Nonetheless, mitochondrial markers are useful [2, 8], so retaining them while including additional nuclear markers seems warranted.

It is noteworthy that 12 of 26 TGCs were unstable. These 12 clusters eventually dissolved, leaving 14 stable TGCs remaining when genotyping concluded in 2020. Instability among some TGCs is a limitation of the CDC's method, although applicable genotyping approaches are limited by the large genome of *Cyclospora* [6]. Whole-genome sequencing is not a feasible *Cyclospora* typing strategy because it is difficult to purify enough parasites from stool samples to obtain sufficient genome coverage. *Cyclospora* cannot currently be cultured [6], so culture-based enrichment of parasites for sequencing is impractical. Consequently, the CDC is restricted to amplicon-based strategies that extrapolate genetic relationships from sequences obtained for a few markers.

As a partial consequence of sexual reproduction, *Cyclospora* genotypes from different patients infected from a common source are typically similar but not necessarily identical [2, 3, 6]. To address this, the CDC developed algorithms (ie, the Barratt-Plucinski ensemble) that consider the effects of sexual reproduction when assessing genetic relationships [2, 3, 6]. These algorithms assign a numeric distance to all genotype pairs, where smaller distances represent close relationships and larger distances, distant relationships. These distances reflect the most probable relationships, given available data. As new genotypes are identified throughout cyclosporiasis peak periods, better matches may be encountered than previously observed, causing some TGCs to dissolve. Sequencing additional nuclear markers could improve cluster stability by increasing the amount of information available for assessing genetic linkage.

Four TGCs could not be linked to a common source of exposure. This observation highlights an important point: while genotyping can facilitate detection of related clusters, high-quality epidemiologic data is indispensable when identifying common exposures. Therefore, rigorous epidemiologic investigations must continue alongside genotyping which complements but cannot replace epidemiologic investigations. Unfortunately, the COVID-19 pandemic likely limited the ability of public health workers to follow up on cyclosporiasis case investigations, for example, because resources were prioritized for pandemic response or because patients with cyclosporiasis declined to participate. The large amount of missing epidemiologic case data limits our ability to further discern potential links from the 2020 genotyping results.

In conclusion, the CDC's *Cyclospora* genotyping method shows promising epidemiologic utility. Despite the absence of epidemiologic data for many case patients, the use of collection dates as a proxy for illness onset, and the limited number of markers sequenced, a signal was detected for the largest cyclosporiasis cluster from 2020 (bagged salad mix), distinguishing it from smaller clusters. This work also supports the need for additional *C. cayetanensis* nuclear markers, an aspect that the CDC is currently investigating as part of the continuous improvement of this genotyping tool.

## Notes

***Acknowledgments.*** The authors acknowledge the contribution of all participating US state public health departments and laboratories to this work. Special thanks to Christine Yanta, Rebecca Guy, and Tanis Kershaw from the Public Health Agency of Canada for their contributions to this study.

***Financial support.*** This work was supported by the US Centers for Disease Control and Prevention.

***Potential conflicts of interest.*** All authors: No potential conflicts. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

## References

1. Tack DM, Ray L, Griffin PM, et al. Preliminary incidence and trends of Infections with pathogens transmitted commonly through food—foodborne diseases active surveillance network, 10 U.S. Sites, 2016-2019. MMWR **2020**; 69:509–14.
2. Nascimento FS, Barratt J, Houghton K, et al. Evaluation of an ensemble-based distance statistic for clustering MLST datasets using epidemiologically defined clusters of cyclosporiasis. Epidemiol Infect **2020**; 148:e172.
3. Barratt J, Houghton K, Richins T, et al. Investigation of US *Cyclospora cayetanensis* outbreaks in 2019 and evaluation of an improved *Cyclospora* genotyping system against 2019 cyclosporiasis outbreak clusters. Epidemiol Infect **2021**; 149:e214.
4. Barratt JLN, Sapp SGH. Machine learning-based analyses support the existence of species complexes for *Strongyloides fuelleborni* and *Strongyloides stercoralis*. Parasitology **2020**; 147:1184–95.
5. Campbell EM, Boyles A, Shankar A, et al. MicrobeTrace: retooling molecular epidemiology for rapid public health response. PLoS Comput Biol **2021**; 17:e1009300.
6. Barratt JLN, Park S, Nascimento FS, et al. Genotyping genetically heterogeneous *Cyclospora cayetanensis* infections to complement epidemiological case linkage. Parasitology **2019**; 146:1275–83.
7. Qvarnstrom Y, Wei-Pridgeon Y, Li W, et al. Draft genome sequences from *Cyclospora cayetanensis* oocysts purified from a human stool sample. Genome Announc **2015**; 3:e01324-15.
8. Nascimento FS, Barta JR, Whale J, et al. Mitochondrial junction region as genotyping marker for *Cyclospora cayetanensis*. Emerg Infect Dis **2019**; 25:1314–9.