Parasitology Research (2023) 122:2641–2650
https://doi.org/10.1007/s00436-023-07963-8

RESEARCH

Evaluation of nuclear and mitochondrial phylogenetics for the subtyping of *Cyclospora cayetanensis*

Jean P. González-Gómez[1] · Luis F. Lozano-Aguirre[2] · José A. Medrano-Félix[3] · Cristobal Chaidez[1] · Charles P. Gerba[4] · Walter Q. Betancourt[4] · Nohelia Castro-del Campo[1]

Received: 7 July 2023 / Accepted: 30 August 2023 / Published online: 7 September 2023

Abstract

Cyclospora cayetanensis is an enteric coccidian parasite responsible for gastrointestinal disease transmitted through contaminated food and water. It has been documented in several countries, mostly with low-socioeconomic levels, although major outbreaks have hit developed countries. Detection methods based on oocyst morphology, staining, and molecular testing have been developed. However, the current MLST panel offers an opportunity for enhancement, as amplification of all molecular markers remains unfeasible in the majority of samples. This study aims to address this challenge by evaluating two approaches for analyzing the genetic diversity of *C. cayetanensis* and identifying reliable markers for subtyping: core homologous genes and mitochondrial genome analysis. A pangenome was constructed using 36 complete genomes of *C. cayetanensis*, and a haplotype network and phylogenetic analysis were conducted using 33 mitochondrial genomes. Through the analysis of the pangenome, 47 potential markers were identified, emphasizing the need for more sequence data to achieve comprehensive characterization. Additionally, the analysis of mitochondrial genomes revealed 19 single-nucleotide variations that can serve as characteristic markers for subtyping this parasite. These findings not only contribute to the selection of molecular markers for *C. cayetanensis* subtyping, but they also drive the knowledge toward the potential development of a comprehensive genotyping method for this parasite.

Keywords *Cyclospora cayetanensis* · Pangenome · Mitochondrial genome · Molecular markers

Introduction

Cyclospora cayetanensis is an emerging intracellular protozoan that causes gastrointestinal infection in humans known as cyclosporiasis. This disease exhibits clinical signs such as watery diarrhea, nausea, vomiting, bowel movements, stomach cramps, fever, and weight loss (Shields and Olson 2003). In recent years, this microorganism has been globally recognized as the causative agent of various diarrheal outbreaks, associated to fresh fruits and vegetables and contaminated water as transmission vehicles (CDC 2020). The Centers for Disease Control and Prevention (CDC) in the USA has documented a yearly occurrence of nine cyclosporiasis multistate foodborne outbreaks between 2013 and 2020. Epidemiological investigations have consistently associated these outbreaks with the consumption of contaminated fresh produce imported from Mexico (CDC 2020). However, not all cases have been linked to a specific food vehicle or geographic origin due to the lack of a reliable typing method for *C. cayetanensis* that complements epidemiological investigations.

Section Editor: Lihua Xiao

✉ Nohelia Castro-del Campo
ncastro@ciad.mx

[1] Laboratorio Nacional para la Investigación en Inocuidad Alimentaria (LANIIA), Centro de Investigación en Alimentación y Desarrollo, A.C. (CIAD), Carretera a Eldorado km 5.5, Campo El Diez, 80110 Culiacán, Sinaloa, México

[2] Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, AP565-A, 62210 Cuernavaca, Morelos, México

[3] Investigadoras e Investigadores por México-Centro de Investigación en Alimentación y Desarrollo, A.C. (CIAD), Laboratorio Nacional Para la Investigación en Inocuidad Alimentaria (LANIIA), Carretera a El dorado km 5.5, Campo El Diez, 80110 Culiacán, Sinaloa, Mexico

[4] Department of Environmental Science, Water & Energy Sustainable Technology (WEST) Center, University of Arizona, 2959 W, Calle Agua Nueva, Tucson, AZ 85745, USA

The extensive study of this microorganism poses a challenge due to various factors, like its difficulty in propagation in cell culture and/or animal models (Eberhard et al. 2000), and even its growth in humans has been unsuccessful for use as a study model (Alfano-Sobsey et al. 2004). Consequently, a limited number of samples of *C. cayetanensis* have been obtained for analysis, as it is restricted to its isolation from stool samples of infected individuals, which involves a laborious process of purification for the isolation of oocysts (Qvarnstrom et al. 2018). Despite the experimental limitations, novel alternatives have emerged for the study of these troublesome microorganisms; for example, new methods based on sequencing and computational biology have allowed for the analysis of microbial groups that cannot be cultured in the laboratory, such as *Cyclospora*, and making their nuclear and extranuclear genome sequences available in Genbank or ENA databases (Qvarnstrom et al. 2015; Qvarnstrom et al. 2018; Tang et al. 2015). Although routine genotyping through complete genome sequencing for this eukaryotic microorganism proves challenging, a set of eight molecular markers for targeted amplicon deep sequencing has been devised for MLST analysis (Nascimento et al. 2020; Barratt et al. 2021). However, the percentage of successfully sequenced eight markers ranged from 11 to 34%, with some showing a sequencing success rate of 39 to 50% (Yanta et al. 2022; Nascimento et al. 2020). Therefore, the pursuit of novel molecular markers is a valuable endeavor.

Currently, two primary methods have been proposed for subtyping *C. cayetanensis* isolates using molecular markers: core homologous genes and mitochondrial genome analysis (Houghton et al. 2020; Nascimento et al. 2019). In both approaches, these markers should possess characteristics such as genetic variability, specificity to the target species, conservation within the genome, reproducibility, and ease of detection (Grover and Sharma 2016). Core homologous genes are obtained by pangenome analysis, which can provide high-resolution data for inferring relationships among closely related isolates (Contreras-Moreira and Vinuesa 2013; Tettelin et al. 2008). However, this approach has the disadvantage of difficulty in analyzing eukaryotic sequences, although algorithms have been gradually established that allow for their study (Contreras-Moreira et al. 2017). On the other hand, mitochondrial genome analysis involves the retrieval of the maternally inherited mitochondrial genome of isolates, which is typically more conserved than the nuclear genome (Cinar et al. 2016; Tang et al. 2015). This approach can be particularly useful for inferring relationships between distantly related isolates or for studying the evolutionary history of a population. Additionally, molecular diagnostic tools targeting mitochondrial sequences have the advantage of being more numerous and therefore more easily detected than those targeting nuclear genes (Tang et al. 2015).

Here, we performed analyses of the nuclear pangenome and mitochondrial genome of *C. cayetanensis* from different geographic regions, in order to study the genomic diversity of the protozoan and determine whether strains from different origins are genetically related. The objective was to identify promising markers, both nuclear and mitochondrial, that can enable the subtyping of this parasite and assist in the epidemiological surveillance of outbreaks.

Materials and methods

Cyclospora cayetanensis nuclear and mitochondrial genomes

For the pangenome analysis, we retrieved all the genomic sequences available in the NCBI database for genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/) of the *C. cayetanensis* species from different geographic regions, with a total of 36 nuclear genomic sequences (Table S1). Validation of their taxonomic classification and completeness was performed using average nucleotide identity and coverage with the ANI program (Richter and Rossello-Mora 2009), selecting those with a percentage of identity and coverage greater than 96% concerning the *C. cayetanensis* CcayRef3 reference genome (GCF_002999335.1). Genome annotation was also performed using the AUGUSTUS program (Stanke et al. 2008) trained with the annotated reference genome (Hoff and Stanke 2013), except for another previously annotated genome in NCBI. In addition, retrieval of the complete mitochondrial genomes available in Genbank (https://www.ncbi.nlm.nih.gov/genbank/) was performed, and, in order to include sequences linked to Mexico and China, six mitochondrial genomes were extracted from previously obtained and annotated complete genomes (Table S2).

Pangenome analysis

To determine the pangenome, a database was created using sequence files of nucleotide and protein sequences from genomes retrieved from NCBI. The selection of homologous genes based on sequence similarity was performed using two methods, OMCL (OrthoMCL) and BDBH (Bidirectional Best Hits), employing the software packages GET_HOMOLOGUES (Contreras-Moreira and Vinuesa 2013) and GET_HOMOLOGUES-EST (Contreras-Moreira et al. 2017). Subsequently, the unique copy genes and the intersection of clusters were selected using the compare_clusters.pl script, and the size graphs of both the core genome and pangenome were generated using the plot_pancore_matrix.pl script, with the parameters -f core_Tettelin and -f pan, respectively. Furthermore, an analysis of the accessory genome was conducted to search for unique genes related to geographic origin, using the compare_clusters.pl script with the parameters -m and -T, specifying subsets A and B for comparison.

Phylogenomic analysis

The phylogenomic relationship was examined using the genes from the core genome that showed the greatest potential as biomarkers according to criteria such as phylogenetic informativeness, conserved regions, variation, coverage, and reproducibility. To select these genes, we utilized the GET_PHYLOMARKERS pipeline (Vinuesa et al. 2018) with the BDBH- and OMCL-generated clusters, using the -R1 execution mode for phylogenetics. This program utilizes PhiPack for recombination analysis, the kdtree test to identify trees significantly deviating from the sample's overall tree distribution, using topology and branch length analysis, and FastTree for assessing phylogenetic signal content through Shimodaria-Hasegawa-like likelihood ratio test branch support values. This process generates a directory of genes classified as the best molecular markers for inferring evolutionary relationships. These concatenated markers were aligned using MUSCLE v3.8 (Edgar 2004), and the phylogenomic relationship was inferred using PhyML v3.3 (Guindon et al. 2010) through maximum likelihood, with the GTR substitution model and SH-like support for branches. The tree was visualized and edited using the online tool iTOL v5 (Letunic and Bork 2021).

Haplotype network and phylogenetic analysis of the mitochondrial genome

The alignment of the 33 downloaded nucleotide sequences for haplotype networks and phylogenetic analysis based on the mitochondrial genome was performed using MUSCLE v3.8 software. The resulting alignment was saved in NEXUS format, along with the necessary geographic origin data for haplotype association. To generate the haplotype network, the PopART v1.7 software (Leigh et al. 2015) was employed, utilizing the median-joining algorithm on the NEXUS file. For the construction of the phylogenetic tree, the MEGA 7 software (Kumar et al. 2016) was used. The tree was built based on the alignment generated by MUSCLE, employing the neighbor-joining method with 500 bootstrap replicates. The visualization and editing of the phylogenetic tree were performed using iTOL v5.

Results and discussion

Pangenome analysis

To visualize the incorporation of novel genes or the reduction of core genes during the inclusion of *C. cayetanensis* genomes in the analysis, graphical representations of the core genome and pangenome were generated. The primary objective of constructing these graphs was to discern the observed patterns and ascertain whether the pangenome exhibits an open or closed nature. This information holds significance as it enables us to deduce whether the current number of sequenced genomes suffices for comprehensive species characterization or if additional sequencing is required to faithfully depict its complete gene repertoire.

The count of shared genes was extrapolated through the application of exponential models established by Tettelin et al. (2005) to assess the dimensions of the core genome and pangenome (Fig. 1). As anticipated, the count of shared genes diminished progressively with the addition of each new genomic sequence. Extrapolation of the curve revealed that the core genome (Fig. 1b) reaches a minimum of 2966 genes when considering 36 genomes. Conversely, the extrapolation of the pangenome demonstrates a linear increment in the number of genes as each new genomic sequence is integrated. The extrapolated growth rate of the pangenome size, based on the 36 genomes, averages at 205.8 genes. This implies that, on average, an additional 205.8 specific genes are introduced for each newly sequenced genome. Furthermore, the graphical representation indicates that the pangenome surpasses 12,000 genes. According to this pattern, it can be inferred that the pangenome of *C. cayetanensis* follows an open configuration (as opposed to a closed configuration that would eventually reach an asymptote). Consequently, the influx of new genes will experience a substantial increase upon the incorporation of additional *C. cayetanensis* genome sequences. Hence, the inclusion of a greater number genome sequences is necessary to achieve a comprehensive and representative characterization of the species' gene repertoire.

Moreover, the *C. cayetanensis* genomes employed in this study are derived from fecal isolates, a process that lacks an axenic culture step for DNA extraction, rendering the obtained genetic material potentially contaminated. This decontamination process primarily relies on bioinformatic tools post-sequencing (Barratt et al. 2019; Qvarnstrom et al. 2018). Additionally, a significant subset of the available *C. cayetanensis* genomes originates from isolates preserved over extended periods, evident from the biosample information detailing collection and submission dates. As a consequence, the absence of certain genes among isolates might be attributable to sample degradation over time or insufficient DNA availability during the initial extraction process. These factors introduce inherent variability that could impact the comparative analysis of the pangenome. Despite the meticulous methodology employed in our study, it is crucial to recognize that these challenges in sequence quality might contribute to the observed variations in gene presence or absence across diverse isolates. This recognition underscores the need for cautious interpretation of our pangenome findings and underscores the call for future research to prioritize the acquisition of high-quality, freshly isolated genomes to ensure more accurate and reliable comparative genomic analyses.

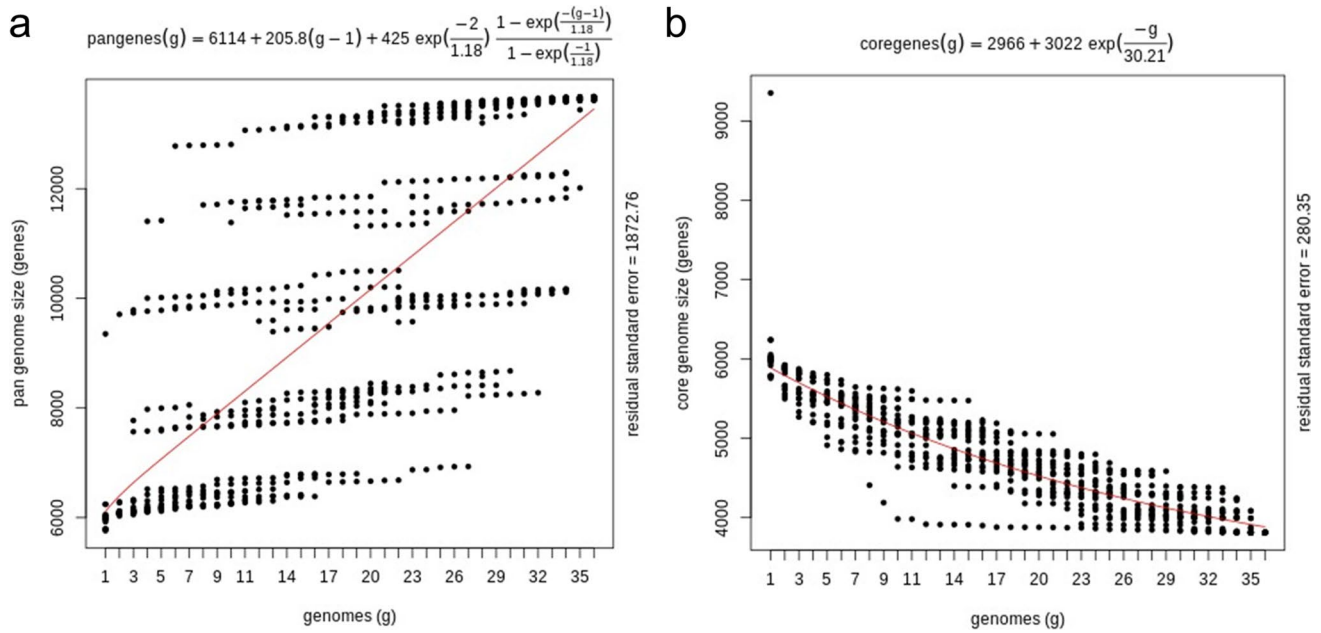

Fig. 1 Graphic of the statistical estimation of the pangenome (**a**) and core genome (**b**) size of *Cyclospora cayetanensis* based on the exponential function fit

Phylogenomic analysis

To identify candidate homologous and single-copy genes as molecular markers, we utilized the results generated by the GET_HOMOLOGUES pipeline. A total of 3349 genes were identified using the OMCL method, while 3740 genes were identified using the BDBH method, with an overlap of 2297 genes detected by both methods. Subsequently, the GET_PHYLOMARKERS pipeline was employed for recombination analysis and phylogenetic estimation, resulting in the identification of 547 markers that passed the Phi recombination test. Among these markers, only 47 passed the kdetree and phylogenetic signal tests and were considered optimal for phylogenetic inference (Table S3).

The phylogenomic tree constructed using these 47 markers revealed clusters of sequences based on their geographical origin, including genomic sequences from China, Nepal, and Indonesia (Fig. 2). The findings from the Chinese sequences align with the results reported by Li et al. (2017), who observed that Chinese isolates formed a closely knit cluster compared to those from other countries (USA, Nepal, and Peru) in their evolutionary analysis using five microsatellite loci as MLST markers. Additionally, it is noteworthy that two sequences from an outbreak in New York, USA, in 2001 (sequences Cc_2893425_USA:NY and Cc_2893435_USA:NY) are clustered with the sequences from Nepal, for which no year of sample collection is given. This suggests a close relationship between these sequences and implies that the *Cyclospora* oocysts responsible for the New York outbreak may have originated from Nepal.

Similarly, the associations among sequences related to Mexico were examined. The genomic sequences from Texas outbreaks in 2013 and 2014, which were linked to the consumption of cilantro from Puebla, Mexico (Cc_3057635_USA:TX_Cilantro, Cc_2893365_USA:TX_Cilantro, Cc_2019455_USA:TX2014, and Cc_2893485_USA:TX2014), as well as two cases involving US travelers who visited Mexico and acquired the infection (Cc_2893335_USA:TX_Vmex and Cc_2893465_USA:TX_Vmex), and another sequence originating from Mexico (Cc_3945065_Mexico) according to Qvarnstrom et al. (2018), were analyzed. These seven sequences were distributed throughout the phylogenetic tree without exhibiting a distinct profile or clustering among them. However, it is worth noting that the genomic sequences from the 2013 Texas outbreak associated with cilantro from Puebla are closely related to each other but distinct from the sequences of the 2014 Texas outbreak, which are also linked to contaminated cilantro from the same geographical area. Additionally, the sequences from the 2014 Texas outbreak are closely related to one of the cases involving US travelers, as well as a genomic sequence of *C. cayetanensis* from Guatemala. The discrepancies in these findings may arise from the fact that these sequences do not originate from a common geographical source. The investigations of these cases have relied on speculative interpretations based on statements provided by affected individuals, rather than

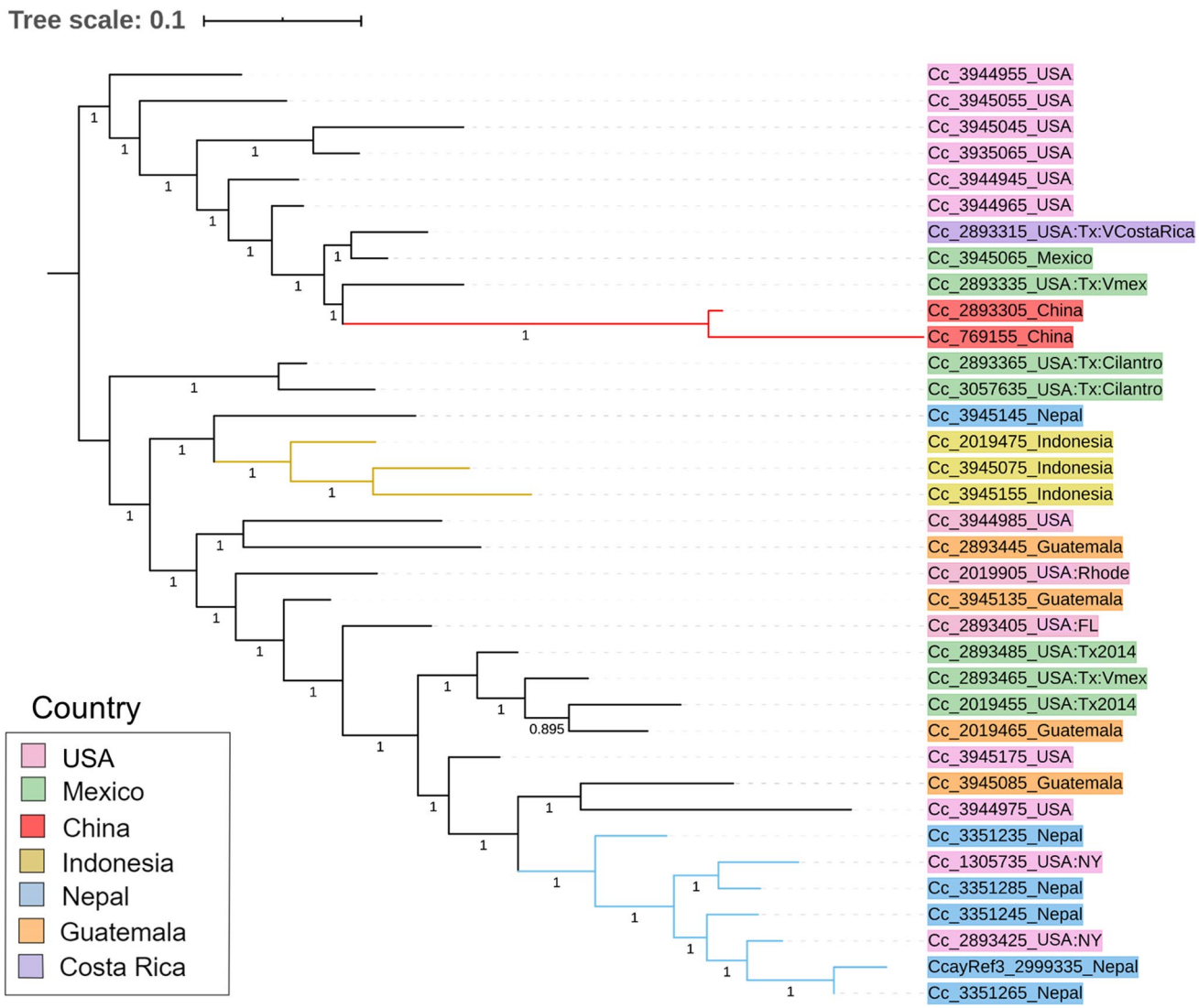

Fig. 2 Phylogenetic tree of the 36 *Cyclospora cayetanensis* isolates based on maximum likelihood using 47 markers obtained by the GET_PHYLOMARKERS pipeline. Numbers on the branches indicate the level of support according to the SH-like supports statistic and the root was determined using the midpoint method

solely tracing the source of pathogen contamination. The absence of subtyping methods hinders the ability to ascertain whether the individual acquired the pathogen from an alternative source or location, or if contamination occurred during food export processes, for example. Another possibility is that the *C. cayetanensis* sequences from Mexico exhibit genetic diversity, indicating the presence of multiple strains within the country. However, confirming the latter hypothesis is challenging given the limited availability of genomes with documented origins and subtyping markers. This underscores the importance of incorporating sequences with well-established isolation origins, particularly from Mexico as a whole, as it will enhance the selection of nuclear markers and facilitate accurate inferences regarding the phylogenetic relationships among *C. cayetanensis* isolates.

In the selection of molecular markers, the GET_HOMOLOGUES and GET_PHYLOMARKERS pipelines yielded promising results by identifying 47 high-quality markers suitable for robust genomic phylogenetic estimation (Table S3). Houghton et al. (2020) recently proposed a workflow for identifying molecular markers from the nuclear genome of *C. cayetanensis*, which resulted in the identification of 485 candidate markers. These findings align with our study, where we identified 547 preliminary markers that underwent screening to select the optimal set of 47 markers. This selection process involves analyzing sequence recombination, as recombined sequences have a detrimental impact on phylogenetic inference. Furthermore, topology verification was performed through tree estimations, excluding atypical sequences, and evaluating their phylogenetic signal

based on branch support values (Vinuesa et al. 2018). The resulting phylogenetic tree (Fig. 2) exhibited distinct clustering patterns based on geographical origin, indicating that incorporating additional genomic sequences into this workflow could yield similar and more precise results. Additionally, the selected optimal markers have the potential to be employed in multilocus typing and complement the existing panel of markers used in previous investigations. For example, Barratt et al. (2019) utilized three markers for MLST typing, consisting of two derived from the nuclear genome and one from the mitochondrial genome. This approach achieved concordance in clustering with epidemiological data for half of the analyzed outbreaks.

The most recent and widely used panel developed by the CDC for investigating outbreak epidemiology comprises eight molecular markers that have demonstrated satisfactory sensitivity and specificity in clustering samples based on their relatedness. However, among the specimens analyzed by Nascimento et al. (2020), only 229 out of 666 and by Yanta et al. (2022), only 21 out of 187 successfully exhibited sequencing for the eight markers. This outcome may arise from physical limitations, including the quantity of the provided sample or the extent of parasite presence within it, which can notably vary. Noteworthy, an 11.1% sequencing success rate was observed for some of these markers. Unfortunately, neither of the aforementioned studies can be directly compared to ours, as none of the six nuclear markers encompassed within the CDC panel was selected through our rigorous high-quality molecular marker selection process. The primary limitation of our work lies in the valuable assessment of the proposed molecular markers, while access to clinical isolates for robust validation remains restricted to a few laboratories.

Furthermore, our study demonstrates congruence with the recent research on *C. cayetanensis* phylogenetics by Leonard et al. (2023). Both studies utilized the available whole-genome sequences from NCBI to examine the genetic relationships of the parasite. Notably, the consistent clustering patterns observed in chromosomal core genes phylogenetics correspond closely with the 47 molecular markers outlined in this study. This convergence of findings underlines the dependability of molecular markers in characterizing *C. cayetanensis* isolates and further bolsters the effectiveness of targeted amplicon sequencing, which has proven its efficacy in genotyping challenging samples, including fecal specimens and fresh produce samples inoculated with low oocyst levels (Leonard et al. 2023).

In line with this, the inclusion of additional nuclear markers to the existing eight-marker MLST scheme has been proposed to enhance cluster stability, especially as specimens are acquired over time (Barratt et al. 2022). This recommendation aligns with the evolving nature of genetic studies, emphasizing the need for expanding marker panels to ensure accurate and meaningful clustering outcomes. Moreover, the target amplicon sequencing approach presents a notable advantage over the eight-loci MLST method, as it encompasses a greater number of informative markers. This increased marker density contributes to enhance confidence in sample placement within a cluster, even when sequence data for certain markers are incomplete.

Haplotype network and phylogenetic analysis of the mitochondrial genome

Haplotype construction was performed by aligning 33 mitochondrial genome sequences obtained from samples collected in Nepal, China, Texas, Montana, Massachusetts, South Carolina, Maine, and Ohio. A total of 19 variable sites were identified, leading to the determination of 13 distinct haplotypes across these eight geographical regions (Fig. 3). These haplotypes, separated by 1 to 9 mutations, exhibit segregation according to their respective geographic locations. Notably, haplotypes 2 and 4, associated with the isolate from China, differ from other haplotypes by 9 mutations. Conversely, haplotype 5 groups together sequences from Nepal and the USA. Therefore, similar to the phylogenetic analysis of the nuclear genome, clustering haplotypes based on the mitochondrial genome is insufficient for definitively establishing the geographic origin of the isolates. However, it serves as a valuable tool for outbreak traceability due to certain advantages offered by utilizing the mitochondrial genome for haplotype determination instead of the nuclear genome. Barratt et al. (2019) reported that, in contrast to the nuclear genome locus, which displayed multiple haplotypes, most specimens analyzed exhibited a single haplotype at the mitochondrial locus. This distinction arises from the fact that an apicomplexan oocyst can be heterozygous, harboring up to two alleles for a given locus, thereby increasing the number of detected sequence types in the nuclear genome of *C. cayetanensis*. Nevertheless, to achieve conclusive results regarding the predictive power of geographic origin, a greater number of samples from diverse geographical regions is required. Unfortunately, studies on this parasite often face challenges stemming from the lack of animal models and cell culture systems (Alfano-Sobsey et al. 2004; Eberhard et al. 2000).

Phylogenetic analysis was performed on mitochondrial genomes, unveiling two primary clusters: one exclusive to sequences from China and another encompassing the remaining sequences (Fig. 4). Within the second cluster, the four sequences from isolates in Nepal, one from Texas, one from Montana, and one from Ohio

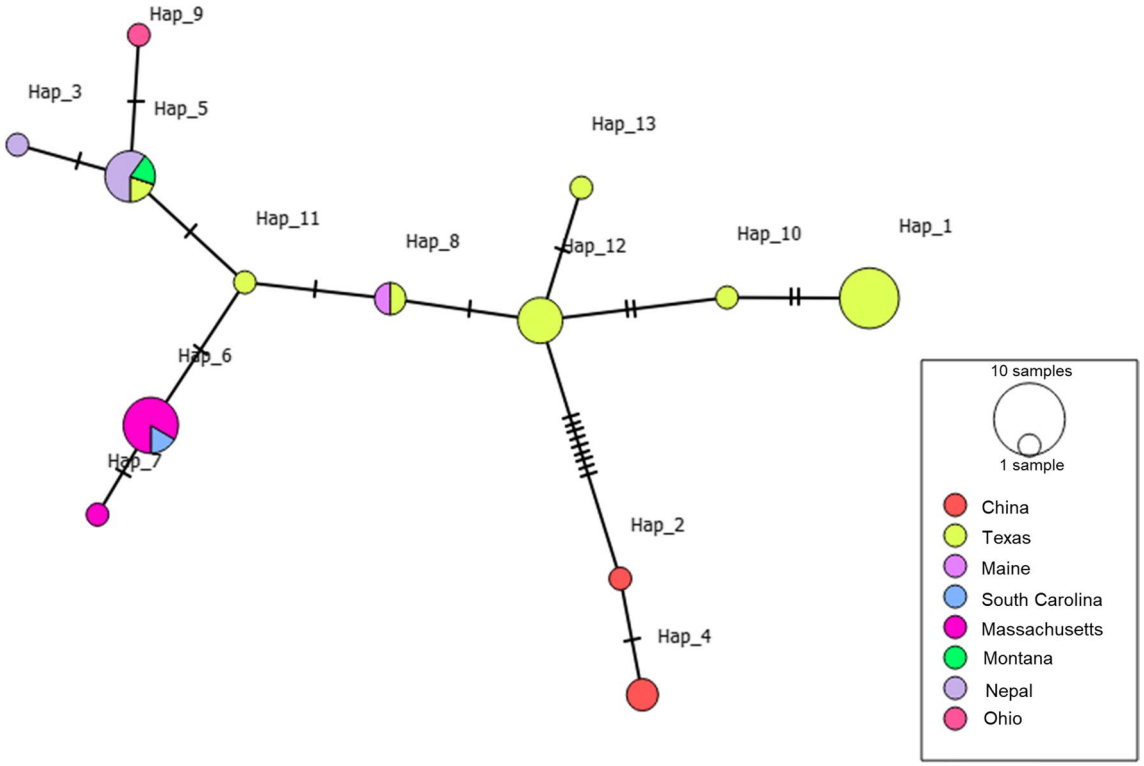

Fig. 3 Median-joining haplotype network of the *Cyclospora cayetanensis* mitochondrial genome from eight geographic regions. Each mark indicates the number of different mutations between haplotypes

are grouped in the same subcluster, suggesting their shared genotype. This grouping could be attributed to the distinct geographic origin of sequences obtained from oocysts of patients in the USA. Until recently, it was widely held that cyclosporiasis was not endemic to the USA. However, recent discoveries of this parasite in fresh produce and irrigation water have challenged this assumption (Durigan et al. 2022; Gottlieb 2018). Consequently, reported cases could now be attributed to both imported and local products or travel to endemic regions with higher incidences of this disease, expanding the range of potential contamination sources. This pattern is evident in the phylogenetic relationship observed between the Ohio sequence MN260353.1, the Montana sequence MN260352.1, and the mitochondrial sequences of isolates from Nepal, signifying a common origin of these genomes. Additionally, four samples from Texas, including two samples linked to cilantro consumption, are clustered together. Moreover, another six samples from Texas are placed in a separate clade that includes the PDMM01002259.1 sequence, associated with oocysts from a patient with cyclosporiasis who traveled to Mexico. Therefore, it is imperative to increase sequencing efforts and genotype reporting in the most important endemic regions. Relying solely on reports from the USA introduces bias due to incomplete patient histories or uncertainty regarding the site of infection. This approach would also aid in the identification of foodborne outbreaks by associating genotypes from endemic regions with their principal exported products.

Subtyping of *Cyclospora cayetanensis* based on the mitochondrial genome offers a higher probability of success, facilitated by the availability of diverse approaches for sequence acquisition, given the substantial 67 copies per haploid genome of *C. cayetanensis* (Cinar et al. 2022). Recovering the mitochondrial genome from metagenomic data represents a suitable strategy, considering the non-cultivable nature of *C. cayetanensis*, which hinders isolation and replication. Furthermore, positive samples may contain genetic material from other protozoa or bacteria. However, comprehensive DNA sequencing of the samples can potentially enable the retrieval of the mitochondrial genome (Qvarnstrom et al. 2018). Alternatively, specific primers can be employed to amplify the target sequence for subsequent sequencing. A set of four primer pairs has been documented to cover the complete mitochondrial genome (Cinar et al. 2020), which is advantageous considering the dispersion of the 19 single-nucleotide

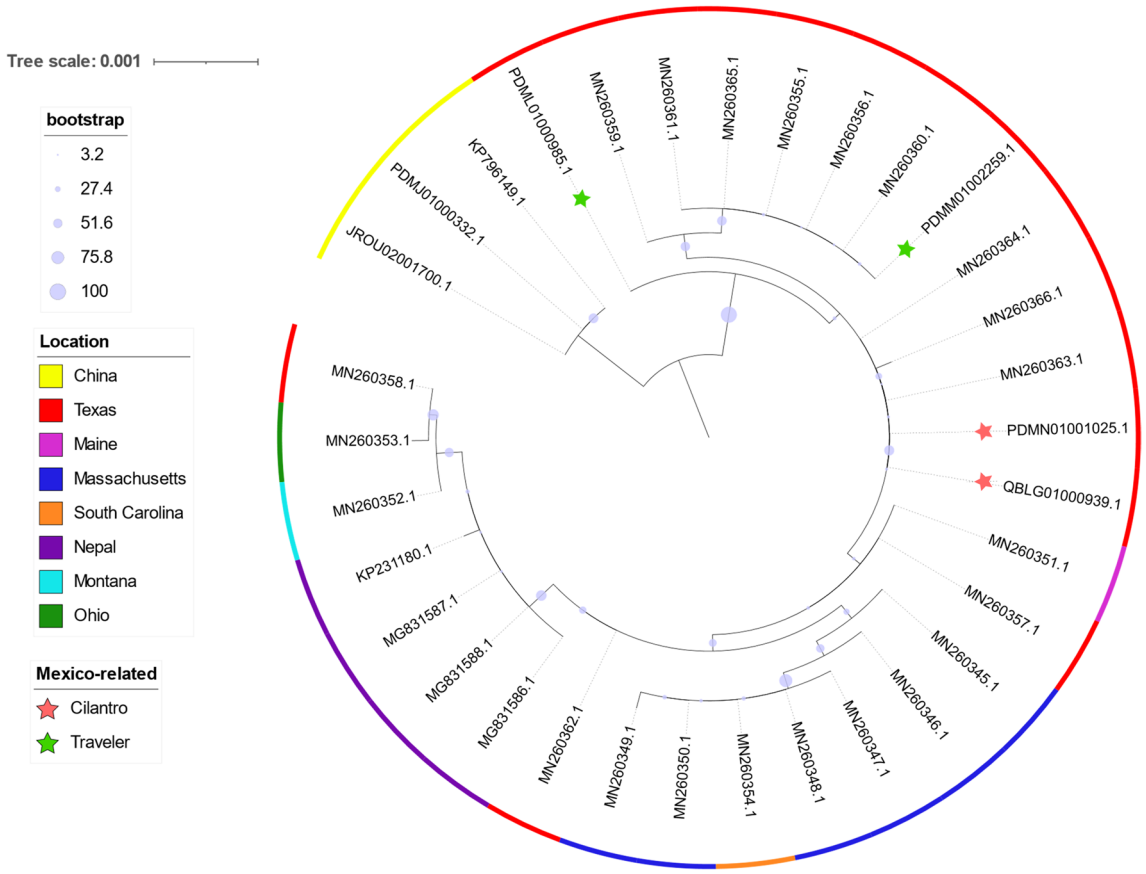

Fig. 4 Phylogenetic analysis of the mitochondrial genome of *Cyclospora cayetanensis* inferred from 33 *Cyclospora cayetanensis* mitochondrial genome sequences using the neighbor-joining method with 500 bootstrap replicates

variations identified across the entire sequence in our analysis. This approach is particularly recommended for sequencing numerous samples, offering inherent benefits in terms of time and cost efficiency.

However, the most robust approach for genotyping *C. cayetanensis* isolates involves a combined utilization of nuclear, apicoplast, and mitochondrial markers. Barratt et al. (2022) reported that assessing nuclear, apicoplast, and mitochondrial markers distinguishes three primary lineages of *Cyclospora* that infect humans, potentially indicating distinct species. Among these newly proposed species, the main lineage remains consistent; however, *Cyclospora henanensis*, encompassing the Chinese genomes, showcases differing mitochondrial characteristics compared to *Cyclospora ashfordi*. Our study similarly highlights greater variations in the mitochondrial phylogenetics of the Chinese isolate compared to those from other geographical regions, in contrast to nuclear phylogenetics. Hence, it is crucial to consider both the quantity of markers and the genomic regions utilized for a precise interpretation of phylogenetic studies concerning this parasite.

Conclusions

This research provides valuable insights into the subtyping of *Cyclospora cayetanensis* using nuclear and mitochondrial phylogenetics. The analysis of the nuclear genome revealed the presence of two main lineages across different geographical regions, suggesting genetic differentiation. However, due to the potential heterozygosity of the apicomplexan oocyst, the nuclear genome alone may not be conclusive for establishing the geographic origin of isolates. On the other hand, the analysis of the mitochondrial genome proved to be a more suitable approach for subtyping *C. cayetanensis*. The identification of 19 single-nucleotide variations within the mitochondrial genome provided characteristic markers for distinguishing different genotypes. This method has the advantage of easier acquisition of sequences and the ability to amplify and sequence the mitochondrial DNA using specific primers. These findings not only contribute to the selection of molecular markers for *C. cayetanensis* subtyping, but they also drive the knowledge toward the potential development of a comprehensive genotyping method for this parasite.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00436-023-07963-8.

Acknowledgements We thank Yazmin Pelayo-Ayón, Estefanía Alarcón-Fabela, Célida I. Martínez-Rodríguez, and Miriam Vega-Rodríguez for the technical support.

Author contribution JPG-G: conceptualization, formal analysis, methodology, writing — original draft. LFL-A: formal analysis, methodology, writing — review and editing. JAM-F: conceptualization, methodology, writing — review and editing. CC: conceptualization, funding acquisition, writing — review and editing. CPG: funding acquisition, review and editing. WQB: writing, review and editing. NC-delC: conceptualization, funding acquisition, validation, writing — review and editing.

Funding The research presented in this article was conducted with the financial support of UA-CONACYT Binational Consortium for the Regional Scientific Development and Innovation.

Data availability The sequences utilized in this study are openly accessible and can be retrieved via the GenBank accession numbers provided in Table S1 and Table S2.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Alfano-Sobsey EM, Eberhard ML, Seed JR, Weber DJ, Won KY, Nace EK, Moe CL (2004) Human challenge pilot study with *Cyclospora cayetanensis*. Emerg Infect Dis 10(4):726–728. https://doi.org/10.3201/eid1004.030356

Barratt J, Ahart L, Rice M, Houghton K, Richins T, Cama V, Arrowood M, Qvarnstrom Y, Straily A (2022) Genotyping *Cyclospora cayetanensis* from multiple outbreak clusters with an emphasis on a cluster linked to bagged salad mix-United States, 2020. J Infect Dis 225(12):2176–2180. https://doi.org/10.1093/infdis/jiab495

Barratt J, Houghton K, Richins T, Straily A, Threlkel R, Bera B, Kenneally J, Clemons B, Madison-Antenucci S, Cebelinski E, Whitney BM, Kreil KR, Cama V, Arrowood MJ, Qvarnstrom Y (2021) Investigation of US *Cyclospora cayetanensis* outbreaks in 2019 and evaluation of an improved *Cyclospora* genotyping system against 2019 cyclosporiasis outbreak clusters. Epidemiol Infect 149:e214. https://doi.org/10.1017/S0950268821002090

Barratt JLN, Park S, Nascimento FS, Hofstetter J, Plucinski M, Casillas S, Bradbury RS, Arrowood MJ, Qvarnstrom Y, Talundzic E (2019) Genotyping genetically heterogeneous *Cyclospora cayetanensis* infections to complement epidemiological case linkage. Parasitology 146(10):1275–1283. https://doi.org/10.1017/S0031182019000581

Barratt JLN, Shen J, Houghton K, Richins T, Sapp SGH, Cama V, Arrowood MJ, Straily A, Qvarnstrom Y (2023) *Cyclospora cayetanensis* comprises at least 3 species that cause human cyclosporiasis. Parasitology 150(3):269–285. https://doi.org/10.1017/S003118202200172X

CDC (2020) List of selected multistate foodborne outbreak investigations. Center for Disease Control and Prevention, Atlanta (GA). https://www.cdc.gov/foodsafety/outbreaks/lists/outbreaks-list.html. Accesed 25 May 2023

Cinar HN, Gopinath G, Almería S, Njoroge JM, Murphy HR, da Silva A (2022) Targeted next generation sequencing of *Cyclospora cayetanensis* mitochondrial genomes from seeded fresh produce and other seeded food samples. Heliyon 8(11):e11575. https://doi.org/10.1016/j.heliyon.2022.e11575

Cinar HN, Gopinath G, Murphy HR, Almeria S, Durigan M, Choi D, Jang A, Kim E, Kim R, Choi S, Lee J, Shin Y, Lee J, Qvarnstrom Y, Benedict TK, Bishop HS, da Silva A (2020) Molecular typing of *Cyclospora cayetanensis* in produce and clinical samples using targeted enrichment of complete mitochondrial genomes and next-generation sequencing. Parasit Vectors 13(1):122. https://doi.org/10.1186/s13071-020-3997-3

Cinar HN, Qvarnstrom Y, Wei-Pridgeon Y, Li W, Nascimento FS, Arrowood MJ, Murphy HR, Jang A, Kim E, Kim R, da Silva A, Gopinath GR (2016) Comparative sequence analysis of *Cyclospora cayetanensis* apicoplast genomes originating from diverse geographical regions. Parasit Vectors 9(1):611. https://doi.org/10.1186/s13071-016-1896-4

Contreras-Moreira B, Cantalapiedra CP, Garcia-Pereira MJ, Gordon SP, Vogel JP, Igartua E, Casas AM, Vinuesa P (2017) Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. Front Plant Sci 8:184. https://doi.org/10.3389/fpls.2017.00184

Contreras-Moreira B, Vinuesa P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79(24):7696–7701. https://doi.org/10.1128/AEM.02411-13

Durigan M, Patregnani E, Gopinath GR, Ewing-Peeples L, Lee C, Murphy HR, Almeria S, Cinar HN, Negrete F, da Silva AJ (2022) Development of a molecular marker based on the mitochondrial genome for detection of *Cyclospora cayetanensis* in food and water samples. Microorganisms 10(9):1762. https://doi.org/10.3390/microorganisms10091762

Eberhard ML, Ortega YR, Hanes DE, Nace EK, Quy Do R, Robl MG, Won KY, Gavidia C, Sass NL, Mansfield K, Gozalo A, Griffiths J, Gilman R, Sterling CR, Arrowood MJ (2000) Attempts to establish experimental *Cyclospora cayetanensis* infection in laboratory animals. J Parasitol 86(3):577–582. https://doi.org/10.1645/0022-3395(2000)086[0577:ATEECC]2.0.CO;2

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797. https://doi.org/10.1093/nar/gkh340

Gottlieb S (2018) The FDA's ongoing efforts to prevent foodborne outbreaks of *Cyclospora*. Available at: https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-fdas-ongoing-efforts-prevent-foodborne-outbreaks Accessed 9 Aug 2023

Grover A, Sharma PC (2016) Development and use of molecular markers: past and present. Crit Rev Biotechnol 36(2):290–302. https://doi.org/10.3109/07388551.2014.959891

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59(3):307–321. https://doi.org/10.1093/sysbio/syq010

Hoff KJ, Stanke M (2013) WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res 41(W1):W123–W128. https://doi.org/10.1093/nar/gkt418

Houghton KA, Lomsadze A, Park S, Nascimento FS, Barratt J, Arrowood MJ, VanRoey E, Talundzic E, Borodovsky M, Qvarnstrom Y (2020) Development of a workflow for identification of nuclear genotyping markers for *Cyclospora cayetanensis*. Parasite 27:24. https://doi.org/10.1051/parasite/2020022

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33(7):1870–1874. https://doi.org/10.1093/molbev/msw054

Leigh JW, Bryant D, Nakagawa S (2015) Popart: full-feature software for haplotype network construction. Methods Ecol Evol 6(9):1110–1116. https://doi.org/10.1111/2041-210x.12410

Leonard SR, Mammel MK, Gharizadeh B, Almeria S, Ma Z, Lipman DJ, Torrence ME, Wang C, Musser SM (2023) Development of a targeted amplicon sequencing method for genotyping *Cyclospora cayetanensis* from fresh produce and clinical samples with enhanced genomic resolution and sensitivity. Front Microbiol 14:1212863. https://doi.org/10.3389/fmicb.2023.1212863

Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49(W1):W293–W296. https://doi.org/10.1093/nar/gkab301

Li J, Chang Y, Shi KE, Wang R, Fu K, Li S, Xu J, Jia L, Guo Z, Zhang L (2017) Multilocus sequence typing and clonal population genetic structure of *Cyclospora cayetanensis* in humans. Parasitology 144(14):1890–1897. https://doi.org/10.1017/S0031182017001299

Nascimento FS, Barratt J, Houghton K, Plucinski M, Kelley J, Casillas S, Bennett C, Snider C, Tuladhar R, Zhang J, Clemons B, Madison-Antenucci S, Russell A, Cebelinski E, Haan J, Robinson T, Arrowood MJ, Talundzic E, Bradbury RS, Qvarnstrom Y (2020) Evaluation of an ensemble-based distance statistic for clustering MLST datasets using epidemiologically defined clusters of cyclosporiasis. Epidemiol Infect 148:e172. https://doi.org/10.1017/S0950268820001697

Nascimento FS, Barta JR, Whale J, Hofstetter JN, Casillas S, Barratt J, Talundzic E, Arrowood MJ, Qvarnstrom Y (2019) Mitochondrial junction region as genotyping marker for *Cyclospora cayetanensis*. Emerg Infect Dis 25(7):1314. https://doi.org/10.3201/eid2507.181447

Qvarnstrom Y, Wei-Pridgeon Y, Li W, Nascimento FS, Bishop HS, Herwaldt BL, Moss DM, Nayak V, Srinivasamoorthy G, Sheth M, Arrowood MJ (2015) Draft genome sequences from *Cyclospora cayetanensis* oocysts purified from a human stool sample. Genome Announc 3(6). https://doi.org/10.1128/genomeA.01324-15

Qvarnstrom Y, Wei-Pridgeon Y, Van Roey E, Park S, Srinivasamoorthy G, Nascimento FS, Moss DM, Talundzic E, Arrowood MJ (2018) Purification of *Cyclospora cayetanensis* oocysts obtained from human stool specimens for whole genome sequencing. Gut Pathog 10:45. https://doi.org/10.1186/s13099-018-0272-7

Richter M, Rossello-Mora R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A 106(45):19126–19131. https://doi.org/10.1073/pnas.0906412106

Shields JM, Olson BH (2003) *Cyclospora cayetanensis*: a review of an emerging parasitic coccidian. Int J Parasitol 33(4):371–391. https://doi.org/10.1016/S0020-7519(02)00268-0

Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24(5):637–644. https://doi.org/10.1093/bioinformatics/btn013

Tang K, Guo Y, Zhang L, Rowe LA, Roellig DM, Frace MA, Li N, Liu S, Feng Y, Xiao L (2015) Genetic similarities between *Cyclospora cayetanensis* and cecum-infecting avian *Eimeria* spp. in apicoplast and mitochondrial genomes. Parasit Vectors 8:358. https://doi.org/10.1186/s13071-015-0966-3

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102(39):13950–13955. https://doi.org/10.1073/pnas.0506758102

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11(5):472–477. https://doi.org/10.1016/j.mib.2008.09.006

Vinuesa P, Ochoa-Sanchez LE, Contreras-Moreira B (2018) GET_PHYLOMARKERS, a software package to select optimal orthologous clusters for phylogenomics and inferring pan-genome phylogenies, used for a critical geno-taxonomic revision of the genus *Stenotrophomonas*. Front Microbiol 9:771. https://doi.org/10.3389/fmicb.2018.00771

Yanta CA, Barta JR, Corbeil A, Menan H, Thivierge K, Needle R, Morshed M, Dixon BR, Wasmuth JD, Guy RA (2022) Genotyping Canadian *Cyclospora cayetanensis* isolates to supplement cyclosporiasis outbreak investigations. Microorganisms 10(2):447. https://doi.org/10.3390/microorganisms10020447