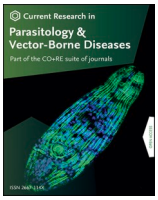
Current Research in Parasitology & Vector-Borne Diseases 4 (2023) 100145

Contents lists available at ScienceDirect

Current Research in Parasitology & Vector-Borne Diseases

journal homepage: www.sciencedirect.com/journal/current-research-in-parasitology-and-vector-borne-diseases



Novel insights on the genetic population structure of human-infecting *Cyclospora* spp. and evidence for rapid subtype selection among isolates from the USA

David K. Jacobson *, Anna C. Peterson, Yvonne Qvarnstrom, Joel L.N. Barratt

Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Cyclospora
Population structure
Genetics
Epidemiology
Genotyping
Cyclosporiasis

ABSTRACT

Human-infecting *Cyclospora* was recently characterized as three species, two of which (*C. cayetanensis* and *C. ashfordi*) are currently responsible for all known human infections in the USA, yet much remains unknown about the genetic structure within these two species. Here, we investigate *Cyclospora* genotyping data from 2018 through 2022 to ascertain if there are temporal patterns in the genetic structure of *Cyclospora* parasites that cause infections in US residents from year to year. First, we investigate three levels of genetic characterization: species, subpopulation, and strain, to elucidate annual trends in *Cyclospora* infections. Next, we determine if shifts in genetic diversity can be linked to any of the eight loci used in our *Cyclospora* genotyping approach. We observed fluctuations in the abundance of *Cyclospora* types at the species and subpopulation levels, but no significant temporal trends were identified; however, we found recurrent and sporadic strains within both *C. ashfordi* and *C. cayetanensis*. We also uncovered major shifts in the mitochondrial genotypes in both species, where there was a universal increase in abundance of a specific mitochondrial genotype that was relatively abundant in 2018 but reached near fixation (was observed in over 96% of isolates) in *C. ashfordi* by 2022. Similarly, this allele jumped from 29% to 82% relative abundance of isolates belonging to *C. cayetanensis*. Overall, our analysis uncovers previously unknown temporal-genetic patterns in US *Cyclospora* types from 2018 through 2022 and is an important step to presenting a clearer picture of the factors influencing cyclosporiasis outbreaks in the USA.

1. Introduction

Sensitive molecular approaches effectively support epidemiological investigations of food-borne disease outbreaks of cyclosporiasis, a gastrointestinal disease caused by apicomplexan parasites of the genus *Cyclospora* (Almeria et al., 2019; Barratt et al., 2023). Cyclosporiasis outbreaks in the USA occur overwhelming in the summer months (Casillas et al., 2018; Almeria et al., 2019) and are thought to primarily be caused by contamination of produce that is imported from *Cyclospora*-endemic regions (Strausbaugh and Herwaldt, 2000; Casillas et al., 2018). However, *Cyclospora* spp. have been detected in produce grown in the USA (Mathison and Pritt, 2021). Additionally, humans are the only known host for the species that cause cyclosporiasis (*Cyclospora cayetanensis*, *Cyclospora ashfordi*, and *Cyclospora henanensis*), meaning animals are not believed to introduce the parasite to new geographical regions (Almeria et al., 2019). Infected patients may not experience symptoms for a week, or longer, after ingestion of food or water contaminated by sporulated *Cyclospora* oocysts, which makes it difficult for individuals to recall specific food exposures that may have caused their illness. This means cyclosporiasis epidemiological traceback investigations are often hampered by incomplete knowledge of produce exposures reported by patients (Barratt et al., 2019; Barratt and Plucinski, 2023). Thus, molecular approaches that complement these investigations are invaluable. In the context of cyclosporiasis, molecular tools divide the wider *Cyclospora* population into subtypes, allowing source attribution investigations to focus on closely related parasites that have a high likelihood of being derived from a common source (Barratt and Plucinski, 2023). Since 2018, the United States Centers for Disease Control and Prevention (CDC) has used the CYCLONE bioinformatic workflow to genetically cluster *Cyclospora* typing data generated from eight genotyping markers to complement source attribution investigations during US cyclosporiasis peak-periods in near-real-time (Barratt et al., 2019, 2021, 2022; Barratt and Plucinski, 2023).

* Corresponding author.
E-mail addresses: quh7@cdc.gov, djacobson@cdc.gov (D.K. Jacobson).

https://doi.org/10.1016/j.crpvbd.2023.100145
Received 18 August 2023; Received in revised form 14 September 2023; Accepted 19 September 2023
Available online 26 September 2023

While recent evaluations indicate that CYCLONE can robustly identify epidemiologically meaningful *Cyclospora* subtypes (Barratt et al., 2022; Jacobson and Barratt, 2023) efforts to improve performance in support of source attribution investigations are ongoing. To this end, addressing knowledge gaps relating to the genetic population structure of human-infecting *Cyclospora* species is paramount. Recently, examination of *Cyclospora* typing data generated from US cyclosporiasis patients between 2018 and 2020 indicated that two-thirds of infections were caused by *C. cayetanensis*, while the remainder were attributed to *C. ashfordi*, a novel species only recently described (Barratt et al., 2023). *Cyclospora cayetanensis* and *C. ashfordi* are recently diverged and are therefore closely related (Barratt et al., 2023), yet their distinction was based on evidence for little to no gene flow between them as supported by the presence (or absence) of certain allele combinations at two of eight CYCLONE genotyping markers. A phylogenetic analysis of randomly selected segments (totaling over 1 million bases) of the *Cyclospora* genome also supported the distinction between *C. ashfordi* and *C. cayetanensis* (Barratt et al., 2023). Epidemiologically, a statistically significant relationship was observed between the geographical residence of cyclosporiasis patients, the month of infection, and the causative *Cyclospora* species; *C. ashfordi* more frequently infected US residents in the southern states and later in the summer compared to *C. cayetanensis* which more often caused infections in midwestern states and earlier in the year (Barratt et al., 2023). These taxonomic, geographical, and temporal trends were uncovered when investigating outbreak-associated *Cyclospora* genotypes from 2018 through 2020 as a single aggregate population (Barratt et al., 2023).

A limitation of this earlier aggregate analysis is that it failed to highlight shifts in the population structure that occur from year to year. By examining outbreak-associated subtypes that emerge and/or dissipate on an annual basis we might be able to differentiate between recurrent subtypes (i.e. those causing outbreaks on a regular basis) *versus* those that occur sporadically. If the majority of cyclosporiasis infections are caused by recurrent types, this could indicate cyclosporiasis infections are originating from the same localities year after year, or, alternatively, that there are a few widespread, dominant *Cyclospora* clades across produce growing regions. Observation of many sporadic types may suggest outbreaks are often isolated and not genetically related on an annual basis, which could result from produce imported from diverse geographical areas. Each scenario has valuable takeaways for how epidemiologists approach cyclosporiasis traceback investigations. Additionally, a yearly analysis will facilitate discernment of annual trends at different levels of genetic classification (e.g. alleles, strains, subpopulations, and species), which could yield important insights about *Cyclospora* population dynamics. For instance, selection of specific alleles at a given locus in both *C. ashfordi* and *C. cayetanensis* would indicate broad selective pressure, while trends limited to one species, or even within a subpopulation, might reflect an isolated temporal/geographical phenomenon or even random genetic drift. From an epidemiological and biological standpoint, such knowledge can elucidate how outbreaks occurring in different years relate to one another.

The present study sought to address the paucity of knowledge on *Cyclospora* genetic population dynamics and how this relates to the epidemiology of cyclosporiasis in the USA. We establish a detailed genetic and epidemiological profile of the *Cyclospora* parasites that infected US residents annually from 2018 through 2022. While our study is limited only to clinical cases in the USA and thus does encompass the full breadth of *Cyclospora* genetic diversity (i.e. environmental/produce sampling, endemic region sampling), our detailed description of these clinical *Cyclospora* genotypes at the species, subpopulation, strain, and the allelic levels, reveals valuable insights about the genetic types of *Cyclospora* that cause annual cyclosporiasis outbreaks in US residents.

2. Materials and methods

2.1. Genotypes

This study utilized publicly available Illumina sequence data generated from all clinical *Cyclospora* isolates sequenced as part of routine US outbreak surveillance performed from 2018 through 2022 ($n = 5258$). This initial sample set is the largest clinical *Cyclospora* dataset available and includes all samples sequenced from patients whose illness was linked to cyclosporiasis outbreaks occurring from 2018 through 2022, as well as all sporadic cases identified during this period, where a link to a specific outbreak could not be identified. Thus, this dataset is not limited to only a select few, widespread cyclosporiasis outbreaks. One limitation of this dataset is that it does not include sampling from environmental sources or *Cyclospora*-endemic regions. Raw data were accessed under NCBI BioProject Number PRJNA578931 and the genotype of each isolate was elucidated using the CYCLONE bioinformatic workflow which is described in detail elsewhere (Barratt et al., 2021). After genotype elucidation, the 5258 resultant genotypes were filtered to exclude isolates with a sequence available for fewer than 7 of the 8 CYCLONE markers. If an isolate was genotyped multiple times, only the most complete genotype was retained. After applying these filtering criteria, 2841 genotyped isolates remained for downstream analysis (Supplementary file S1, Table A). The breakdown of the 2841 isolates according to year detected is as follows: 2018: $n = 398$ (14.0% of total); 2019: $n = 585$ (20.6%); 2020: $n = 587$ (20.7%); 2021: $n = 583$ (20.5%); and 2022: $n = 688$ (24.2%). In total, 452 (15.9%) isolates belonged to an epidemiological cluster and the remainder were from sporadic cases with no known endemiological link (Supplementary file S1, Table A). Isolates originated from 29 US states plus the District of Columbia (Supplementary file S1, Table A), with the majority of isolates coming from New York ($n = 741$, 26.1%), Texas ($n = 531$, 18.7%), and Florida ($n = 390$, 13.7%).

2.2. Genetic distance computation

A pairwise genetic distance matrix was computed from the remaining 2841 genotypes using Barratt's heuristic definition of genetic distance (https://github.com/Joel-Barratt/Eukaryotpying-Python) (Jacobson et al., 2022). Note that this genetic distance computation method is utilized as part of the CYCLONE method as it includes imputation steps that can predict missing values when a partial genotype (e.g. possessing 7 of our 8 CYCLONE markers) is encountered (Nascimento et al., 2020; Jacobson et al., 2022). The resultant genetic distance matrix was clustered using Wards clustering method which is available with the *cluster* R package (Maechler, 2018). The hierarchical tree was then rendered using the *ggtree* package in R (Yu et al., 2017).

2.3. Species assignment

As per Barratt et al. (2023) the 360i2 locus is the key marker from among the CYCLONE 8-marker genotyping panel for distinguishing *C. cayetanensis* and *C. ashfordi*. *Cyclospora* genotypes meeting the previous filtering step ($n = 2841$) were designated as either *C. cayetanensis*, *C. ashfordi*, mixed species (i.e. genotype possesses a mix of *C. cayetanensis* and *C. ashfordi* alleles), or unknown species (i.e. no 360i2 sequence available) as per Barratt et al. (2023).

2.4. Defining major subpopulations

The cutreeHybrid function in the *dynamicTreeCut* R package (Langfelder et al., 2008) was previously used to detect several *Cyclospora* populations below the species level designations described above (i.e. subpopulations) (Jacobson et al., 2023). The cutreeHybrid approach is advantageous because hierarchical trees are split dynamically (i.e. not at a constant height across the tree) by accounting for the shape of

branching, as well as dissimilarity between isolates (Langfelder et al., 2008). This approach can yield biologically informative clusters (Bailey et al., 2016; Rosato et al., 2018) and here, we applied the default parameters of cutreeHybrid to the hierarchically clustered genetic distance matrix of 2841 genotypes to identify *Cyclospora* subpopulations from 2018 through 2022.

2.5. Defining strains via the CYCLONE approach

For the purposes of the present study, a *Cyclospora* 'strain' was defined as the set of genotypes assigned to a given partition that was identified using a recently described statistical framework for dissecting hierarchical trees (Barratt and Plucinski, 2023). Briefly, a stringency parameter is selected by the user to determine the percent of genotypes within a partition that must fall under an empirically calculated genetic distance. A stringency setting of 96.5%, which is optimized for the CYCLONE genotyping process and yields epidemiologically meaningful partitions, see (Jacobson and Barratt, 2023), was used to identify strains from the hierarchically clustered genetic distance matrix of 2841 genotypes.

2.6. Mitochondrial genotype abundance

Currently, two mitochondrial (Mt) loci are captured by the CYCLONE process: the 'MSR' locus comprising part of the SSU rDNA gene (Barratt et al., 2021) and the mitochondrial junction sequence (Nascimento et al., 2019). CYCLONE bioinformatically splits the MSR locus into 6 segments (parts A through F) and variation (single nucleotide polymorphisms - SNPs) have thus far only been detected at parts A and F. For this analysis, the 2841-genotype dataset was filtered to retain isolates possessing a single haplotype for the Mt Junction, Mt_MSR Part A, and Mt_MSR Part F (i.e. genotypes with either multiple alleles at any of these markers, or no alleles at any of these markers, were excluded). After these filters were applied, 2003 genotypes remained for subsequent Mt genotype analysis.

Next, genotypes were binarily classified into those possessing either short or long Mt Junction haplotypes. The mitochondrial genome of *Cyclospora* has a linear concatemeric structure, where one genome copy is linked to another *via* a repetitive junction region (Cinar et al., 2015; Gopinath et al., 2018). This junction region contains a variable combination of three different 15 base pair (bp) repeats (Gopinath et al., 2018; Nascimento et al., 2019). Historically, *Cyclospora* samples from the USA have between one and six repeats (Nascimento et al., 2019; Barratt et al., 2023), while a sample from China (strain CHN_HEN01; an isolate of *Cyclospora henanensis*) had zero repeats (Barratt et al., 2023). In our analysis, isolates possessing two or fewer 15 bp repeats were categorized as *Short* and isolates with three or more 15 bp repeats were categorized as *Long*.

Combining the Mt_Junction length with the MSR A and MSR F haplotypes yields the full mitochondrial genotype as defined by the CYCLONE process. A total of 10 unique mitochondrial genotype combinations (i.e. of haplotypes observed at MSR Part A, and MSR Part F, and the junction length category) were identified *via* this process. Next, mitochondrial genotype diversity within each year was evaluated using Hill numbers, an established framework for analyzing diversity in a biological dataset (Hill, 1973; Chao et al., 2014; Alberdi and Gilbert, 2019). Briefly, Hill numbers assist in quantifying the diversity profile of a community at different orders of the parameter q: at $q = 0$, the returned Hill number value represents the number of unique types (i.e. richness), as q approaches 1, the Hill number value is the Shannon index, and at $q = 2$, the calculated value is the Simpson index (Hill, 1973; Chao et al., 2014; Alberdi and Gilbert, 2019). At each order q, higher values indicate greater diversity, and in this study, q assesses the diversity of mitochondrial types found in each year. Hill numbers were calculated using the *vegan* package in R (Dixon, 2003).

2.7. Identification of temporal trends

We calculated the proportion of genotypes sequenced from 2018 through 2022, falling into a given classification category at three levels of our genetic hierarchy (i.e. the percentage of isolates falling into any of the two species categories, the six subpopulation categories, and possessing any of the ten mitochondrial genotype combinations identified previously). Proportional data were used since there was an unequal number of specimens analyzed in each year. We used the Mann-Kendall (MK) test statistic to test for the presence of a monotonic trend in proportional abundance of species, subpopulations, and mitochondrial genotypes between 2018 and 2022. The non-parametric MK test was chosen because proportional abundance data is not normally distributed, and available data were limited to five years. MK's test statistic (*tau*) was calculated in R using mk.test from the *trend* package (Pohlert, 2016); we report *tau*, *z*-score, and *P*-value for each test performed. One-sided *P*-values are reported for categories where a monotonic downward/upward trend is apparent after plotting the values, while two-sided *P*-values are reported where no monotonic trend is immediately obvious.

At the strain level, the MK approach was not used for temporal trends due to the small size of the majority of the clusters. Rather, we assessed temporal trends by defining recurrent, sporadic, and indeterminant pattern strains across the 5-year period. A strain was considered recurrent if there were 30 or more genotypes assigned to the strain every year; this value was chosen so that the yearly genotype prevalence in recurrent strains would exceed 1% of total specimens analyzed (i.e. 1% of 2841 is 28.41). A sporadic strain was defined where $> 75\%$ of all genotypes assigned to this strain were observed in a single year. Sporadic strains are not required to contain genotyped isolates from each year analyzed. Strains that fit neither of these criteria were classified as indeterminant pattern strains.

2.8. Alternative bead-to-sample ratio evaluation

Minor modifications have been made to the *Cyclospora* laboratory workflow over time, including changes to the library preparation method, as commercially available kit protocols have changed. This includes changes to the bead-based DNA purification/cleanup protocols recommended by Illumina, which is the sequencing chemistry used to sequence all *Cyclospora* genotypes. We selected 11 samples from CDC's archive of *Cyclospora*-positive samples to evaluate how purification bead ratios used in the clean-up step of library preparation impacts the length of the Mt_Junction marker recovered. These 11 samples were processed in duplicate at a purification bead:library ratio of 0.8× , 1.0× , and 1.8× . These ratios were chosen because the 1.8× ratio of Ampure Beads (Beckman Coulter, Brea, California, USA) was used on samples processed at CDC between 2018 and 2020 in accordance with the Illumina Nextera XT kit (Illumina, San Diego, California, USA). In 2021 CDC switched to the Illumina DNA Prep kit which uses a different bead chemistry and a 1.0× ratio was used in 2021 and 2022 after optimization for recovering *Cyclospora* amplicons using the Illumina DNA Prep Kit. Finally, we ran these 11 samples with a 0.8× ratio to further test how purification bead:library ratio impacts Mt_Junction marker detection. We compared the Mt_Junction haplotype detected for each replicate of each bead:library ratio to the Mt_Junction haplotype detected in the sample when it was originally processed at CDC (all of the original processing was done with a 1.0× bead:library ratio). Other than the modifications to bead:library ratio described above, the library preparation and sequencing methods used here are the same as those described previously (Barratt et al., 2021). Illumina MiSeq data from this experiment were deposited under BioProject PRJNA578931.

3. Results

3.1. Species level trends

Following the approach outlined in *Section 2.3*, 2692 of the 2841 genotypes were definitively assigned to a species while the remainder were assigned to the categories of "unknown species" ($n = 91$) or "mixed species genotype" ($n = 57$). *Cyclospora cayetanensis* was the most abundant species of *Cyclospora* infecting US residents annually; however, there was some fluctuation in the proportions of *Cyclospora* species in the samples that were sequenced between 2018 and 2022 (Fig. 1A, $tau = -0.20$, $z = 0.24$, two-sided P-value $= 0.81$). *Cyclospora cayetanensis* was responsible for approximately 75% of cyclosporiasis infections we sampled in the USA in both 2018 and 2021, while in 2019, 2020, and 2022 between 50% and 60% of infections were caused by *C. cayetanensis*. The percent of specimens with an undetermined species assignment (i.e. mixed or unknown) ranged from a high of 8.7% (4.3% mixed; 4.4% unknown) of isolates in 2019 to a low of 3.3% (1.9% mixed; 1.5% unknown) of isolates in 2022 (Fig. 1A; Supplementary file S1, Tables A and B). Interestingly, the percent of isolates with an unknown species assignment significantly decreased from 2018 through 2022 ($tau = -1.0$, $z = -2.21$, one-sided P-value $= 0.01$), which is likely related to the introduction of qPCR screening of samples prior to Illumina sequencing as part of the CYCLONE process. This was introduced in an attempt to exclude samples with a low parasite load (i.e. a high Cycle Threshold value) that may fail to yield a complete genotype.

3.2. Subpopulation-level trends

Cyclospora cayetanensis and *C. ashfordi* were split into six different subpopulations using cutreeHybrid: subpopulations A, B, D, E belonged to *C. cayetanensis* and subpopulations C and F belonged to *C. ashfordi* (Figs. 1B and 2). Overall, there were no significant temporal trends identified for any subpopulation (Supplementary file S1, Table B); however, we did observe some fluctuations in the subpopulation abundance in *C. cayetanensis* from year to year (Fig. 1B). For example, each of the four *C. cayetannsis* subpopulations was the most abundant subpopulation for that species at least once across the five years we analyzed. On the other hand, Subpopulation C was found to be the most abundant subpopulation for *C. ashfordi* every year. Across the full dataset, there were two subpopulations (A and C) with relatively high abundance each year (i.e. $>$ 10% annual abundance); yet neither of these subpopulations, nor any other subpopulation, ever exceeded more than 40% of infections (Fig. 1).

3.3. Strain-level trends

Thirty strains were detected using a previously described statistical framework for dissecting hierarchical trees (Jacobson and Barratt, 2023). The five most prevalent strains each belonged to distinct subpopulations (Fig. 3, Supplementary file S1, Table A) and each of these strains accounted for greater than 45% of all isolates within the respective subpopulation. However, Subpopulation F did not follow this description, as the three strains (Strain 14, Strain 15, Strain 23) each made up roughly one third of the total isolates within Subpopulation F (Supplementary file S1, Table A). Using the recurrent/sporadic/indeterminant categorization scheme described in *Section 2.7*, two strains qualified as recurring, while another six strains were classified as sporadic, and the remaining 22 strains were neither recurring nor sporadic (Fig. 3). Unsurprisingly, the two most prevalent strains were both recurring, with Strain 3 ($n = 549$ of 2841) belonging to Subpopulation C, and Strain 1 ($n = 330$ of 2841) being a member of Subpopulation A. There were distinct epidemiological outbreaks from multiple years found in both Strain 1 and Strain 3 (Supplementary file S1, Table C). Sporadically occurring strains of *C. cayetanensis* and *C. ashfordi* were typically observed in association with a single previously described epidemiological outbreak in each strain. For *C. cayetanensis*, the sporadic Strain 30 was epidemiologically linked to the "2018 Vendor A" outbreak (Nascimento et al., 2020), Strain 13 was linked to the "2021 Lettuce 1" outbreak (Ahart et al., 2023), Strain 9 was linked to the "2018 Vendor B" outbreak (Nascimento et al., 2020) (Supplementary file S1, Table C). In *C. ashfordi*, the sporadic Strain 27 was largely related to the "2019 Restaurant B" outbreak (Barratt et al., 2021) and Strain 23 was linked to the "2019 Distributor A Type 18" outbreak (Barratt et al., 2021).

3.4. Trends relating to mitochondrial genotype

In total, we observed two Mt_Junction haplotypes (*Short*, *Long*), three Mt_MSR Part A haplotypes (*A1*, *A2*, *A3*) and two Mt_MSR Part F haplotypes (*F1*, *F2*) (Supplementary file S2). The Mt_Junction and Mt_MSR markers showed clear patterns in allelic segregation across the 2003 genotypes in the Mt analysis dataset. First, strictly within the Mt_MSR locus, haplotype *A2* was never found in combination with haplotype *F2*; similarly, we only observed four occurrences where haplotype *A3* was found in the same genotype as haplotype *F1* (Table 1). After incorporating the Mt_MSR A/F alleles with their respective Mt_Junction length classifications, we observed additional Mt genotypes that were rare: *Short* + haplotype *A1* + *F1* ($n = 8$, 0.40% of isolates), *Long* + *A1* + *F2* ($n = 12$, 0.60% of isolates), *Short* + *A2* + *F1* ($n = 1$, 0.05% of isolates). This left 5 mitochondrial allelic combinations that were commonly observed in the USA between 2018 and 2022 (Table 1).

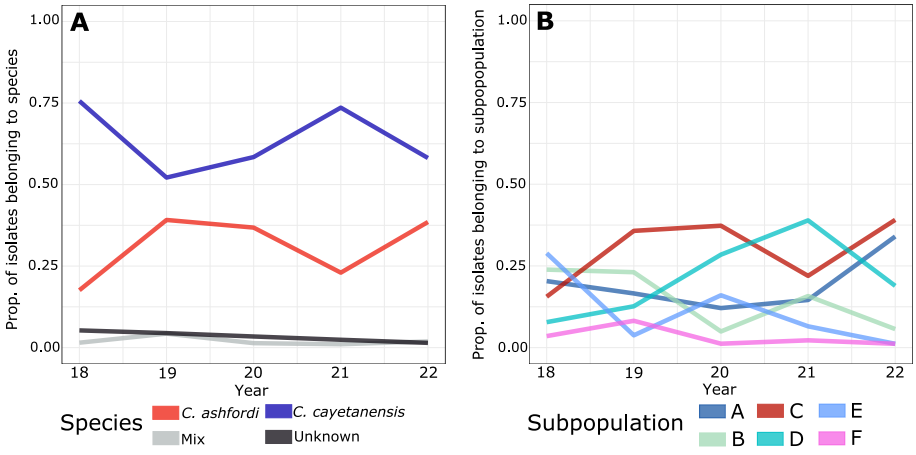


Fig. 1. Trends in *Cyclospora* species and subpopulation abundance from 2018 through 2022. Yearly proportional abundance of species (**A**) and subpopulations (**B**). In **B**, subpopulations belonging to *C. cayetanensis* (A, B, D, E) have a blue hue, while subpopulations belonging to *C. ashfordi* (C, F) have a red hue.

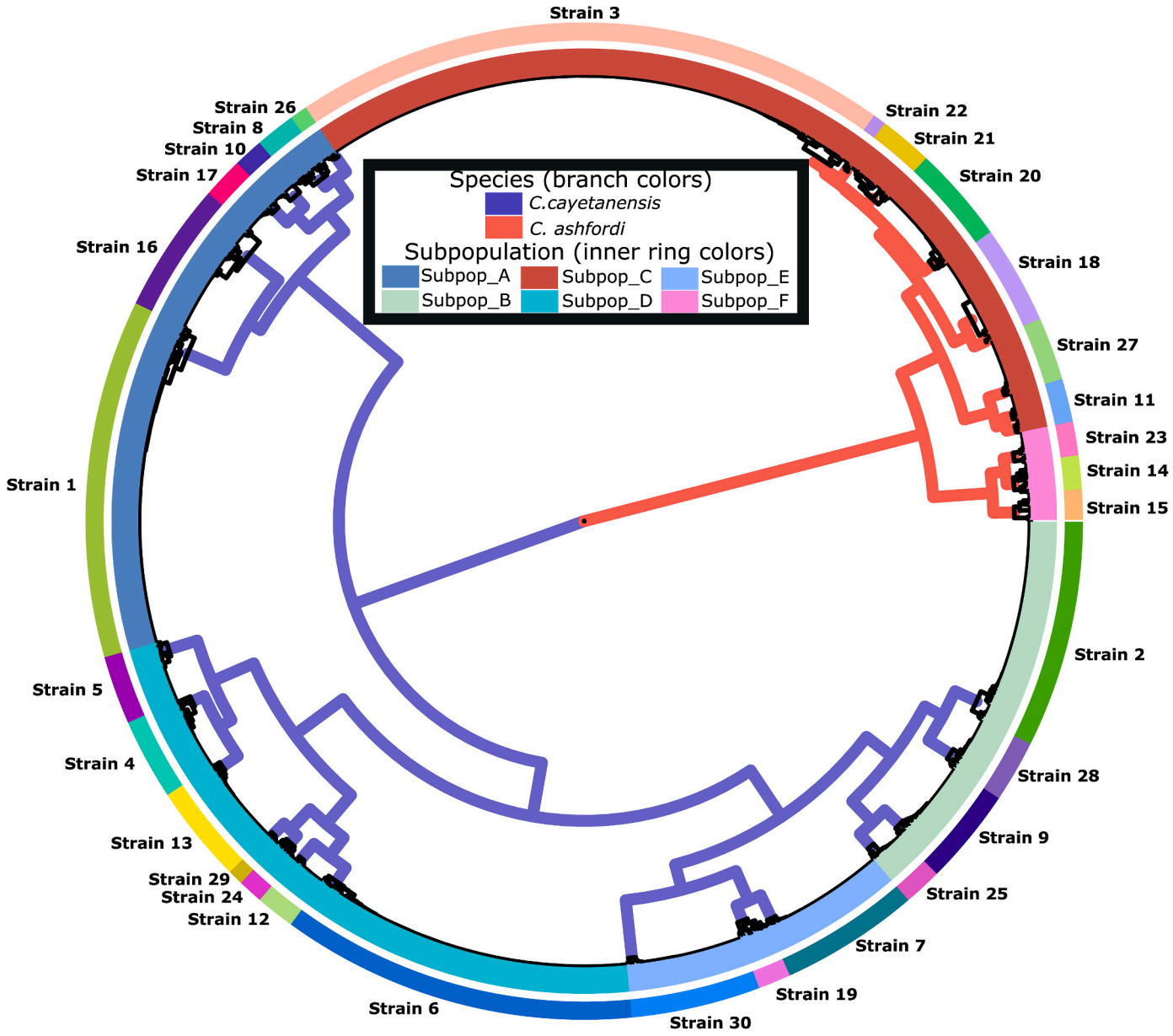


Fig. 2. Hierarchically clustered tree of 2841 *Cyclospora* genotypes during 2018–2022. The colored branches represent species and the inner ring represents subpopulations. Strains are colored by the outer ring and labeled in text.

A statistically significant trend was observed with respect to combinations of MSR genotype and junction sequence length (*tau* = 0.8, *z*-score = 1.71, one-sided *P*-value = 0.04), favoring an increase in the frequency of the *Short* + *F2* + (*A1* or *A3* haplotype) genotype over time, which steadily increased from 38% of all genotypes in 2018 to 88.2% by 2022 (Fig. 4, Supplementary file S1, Table B). The trend was observed in both species but with slightly different dynamics. In *C. cayetanensis*, this allele rose from relatively low abundance in 2018 (29.1% of isolates) to high abundance in 2022 (82.3% of isolates). The mitochondrial *Short* + *F2* + (*A1* or *A3* haplotype) genotype was already found in over 76% of *C. ashfordi* isolates in 2018 yet it spread to near fixation by 2022 (96.7% of isolates) (Fig. 4). Overall, Hill numbers indicate that *C. cayetanensis* possessed a more diverse makeup of mitochondrial alleles (Fig. 5): there were a higher number of unique mitochondrial alleles observed in *C. cayetanensis* (at $q = 0$, Hill number equates to richness, or number of unique types) and those alleles were more evenly distributed across the population (at $q = 1$, Hill number equates to Shannon index; at $q = 2$, Hill number equates to Simpson index), supporting the finding that *C. ashfordi* mitochondrial diversity was limited to a few, highly abundant alleles (Fig. 5). Looking at proportional abundance, the *Short* + *A1* + *F2* genotype was present in greater than 85% of all *C. ashfordi* isolates each year following 2018. Contrastingly, no individual *C. cayetanensis* mitochondrial genotype exceeded 62% relative abundance in any year (Fig. 4).

3.5. Modifications to bead purification protocols

The clear increase in abundance of short Mt Junction haplotypes in the past few years led us to investigate if a change in laboratory methods (i.e. modifications to the bead purification steps in our Illumina sequencing protocol) could partially explain this trend. The ratio of purification beads to sample volume impacts the size of DNA fragments filtered out during library preparation (Verrow et al., 2019) and as discussed *Section 2.8*, this ratio was changed between 2020 and 2021. We re-processed a subset of samples ($n = 11$) with different bead ratios and our results demonstrated that bead ratio had no impact on the

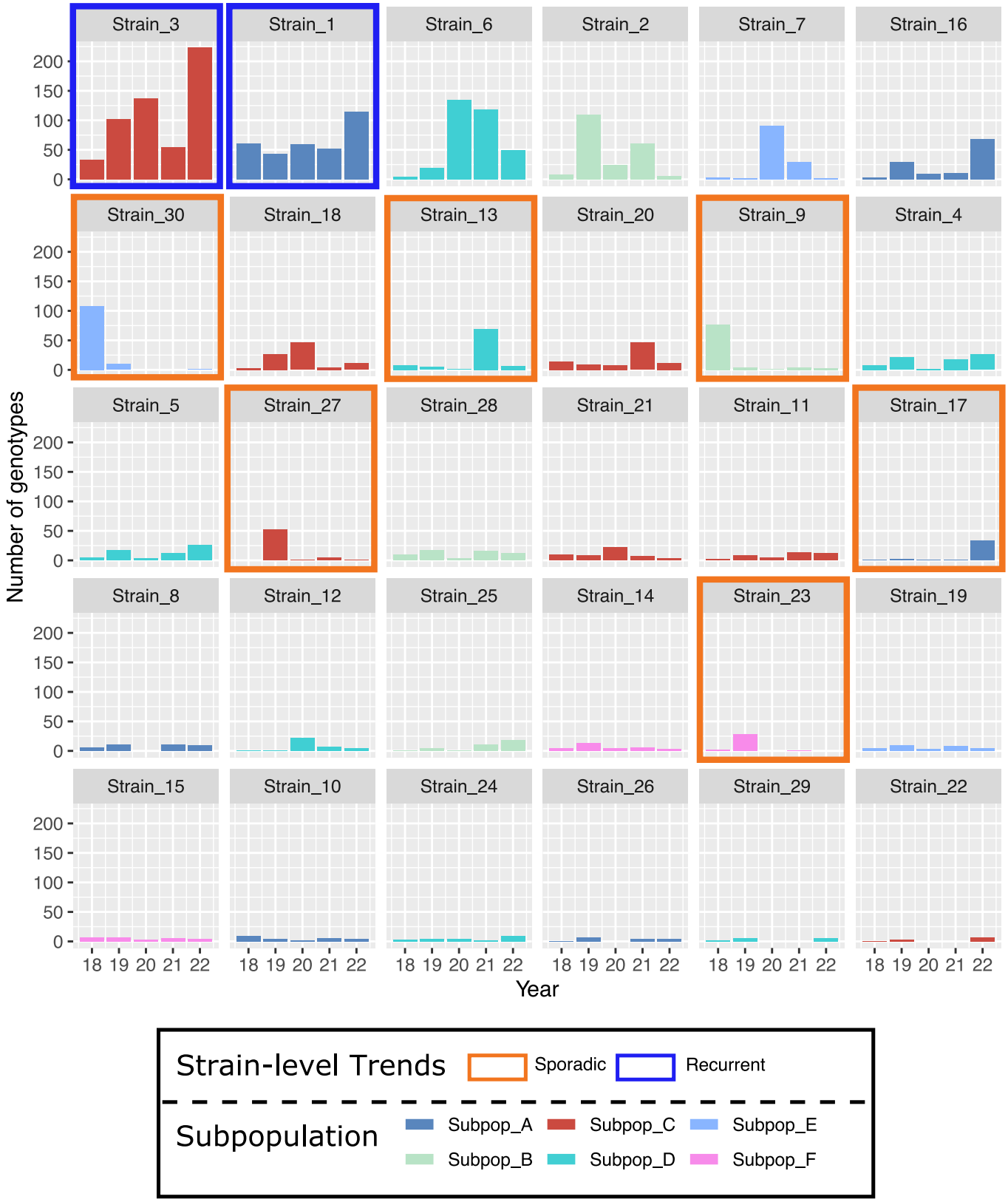


Fig. 3. Trends in *Cyclospora* strain abundance from 2018 through 2022. Bars represent the raw abundance of genotypes in each strain per year. All 30 strains are listed in order of total abundance across the five years analyzed (e.g. Strain 3 has the most genotypes and Strain 22 has the fewest genotypes) and the color of the bars represents which subpopulation the strain sits within. Colored bars with a blue hue belong to *C. cayetanensis*, and colored bars with a red hue belong to *C. ashfordi.* Strains outlined in blue boxes are recurrent strains, strains outlined in orange boxes are sporadic, and strains not outlined have an indeterminant temporal pattern.

Table 1
Frequency of each mitochondrial genotype combination identified.

	Mt MSR haplotype combinations					
	A1/F1	A1/F2	A2/F1	A2/F2	A3/F1	A3/F2
Short junction	8 (0.40%)	**1100 (54.92%)**	1 (0.05%)	0 (0%)	2 (0.10%)	**329 (16.43%)**
Long junction	**275 (13.73%)**	12 (0.60%)	**236 (11.78%)**	0 (0%)	2 (0.10%)	**38 (1.90%)**

Note: Five relatively common allelic combinations (present in greater than 0.6% of all *Cyclospora* isolates) are in bold. The number of isolates is indicated in each cell, where the percent of all isolates possessing this genotype is shown in parentheses.

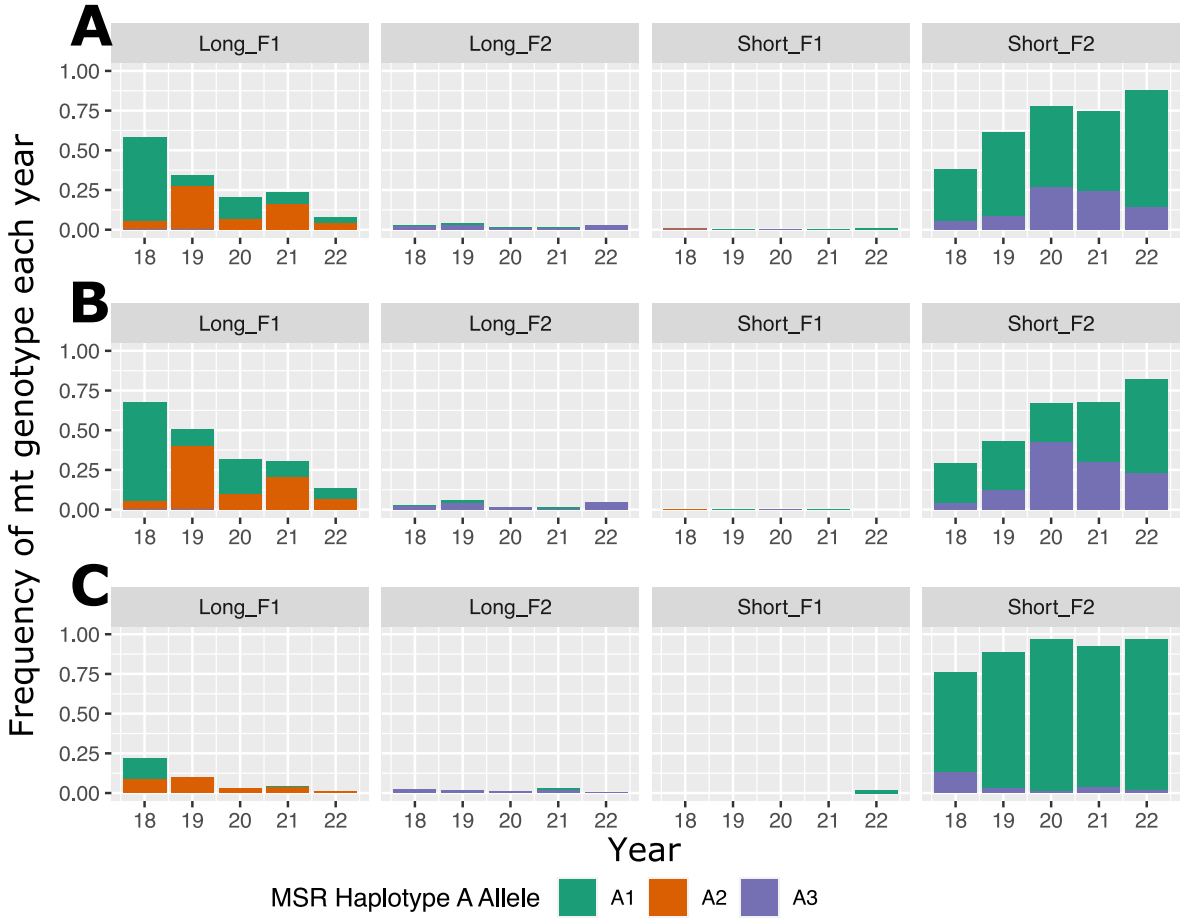


Fig. 4. Mitochondrial (Mt) genotype abundance by *Cyclospora* species. Proportional abundance of mitochondrial genotypes across the filtered Mt analysis dataset (n = 2003). Bar height represents the proportion of genotypes within a given year that have the Mt genotype of interest within the full dataset (**A**), within only *C. cayetanensis* (**B**), or only *C. ashfordi* (**C**) genotypes. Each of the four boxes is a distinct Mt junction length plus MSR Part F combination, while the colors in the bars indicate the MSR Part A haplotype found with the Mt junction + MSR F genotype.

length of Mt Junction haplotype identified in a given sample (Supplementary file S1, Table D). Thus, we concluded that the observed shift towards the vastly abundant *Short + F2* mitochondrial genotype was not observed due to changes in our laboratory protocols.

4. Discussion

Our investigation revealed a complex population genetic structure for *C. cayetanensis* and *C. ashfordi* isolates causing cyclosporiasis outbreaks in the USA from the years 2018–2022. We did not observe any clear directional time related trends in abundance at the species or subpopulation level, which supports relative stability among US outbreak-causing *Cyclospora* varieties at these higher orders of classification (subpopulation and species). This is to be expected as outbreak dynamics may drive a temporary spike in a strain within a species or a particular subpopulation from one year to the other, but the higher-level genetic structure does not drastically shift. Along those lines, we identified two subpopulations (A and C) that were each responsible for greater than 10% of total cyclosporiasis cases every year, further suggesting that we can expect cyclosporiasis cases to be regularly caused by parasites belonging to these two subpopulations. Interestingly, Subpopulation C is one of only two *C. ashfordi* subpopulations identified here and is the more abundant *C. ashfordi* subpopulation in each year we analyzed. On the other hand, there is not a single *C. cayetanensis* subpopulation that accounts for the plurality of infections within this species in more than two years; rather, the four subpopulations each have at least one year where they are the most abundant subpopulation.

Five of the six subpopulations showed greater than 45% of isolates belonging to a single strain within the subpopulation; however, only 2 of 30 strains were identified as recurring, indicating that the *Cyclospora* isolates causing seasonal outbreaks in the USA represent a genetically diverse population when looking below the subpopulation level. The

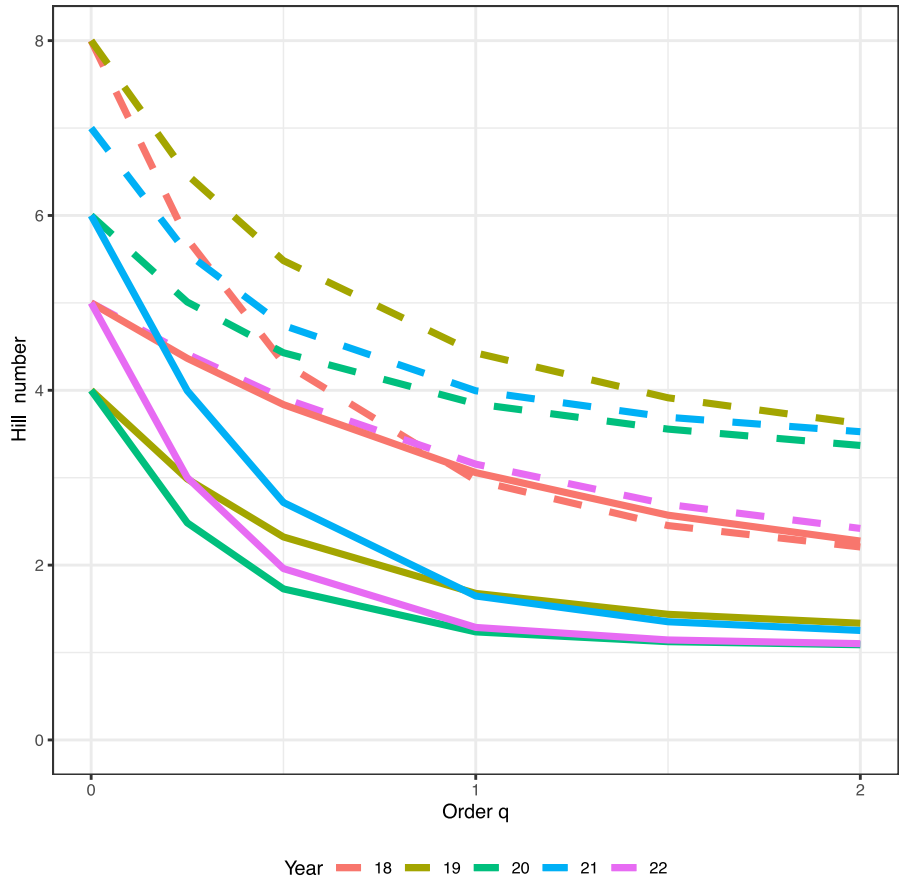


Fig. 5. Hill number diversity profile for mitochondrial genotypes identified each year. The dashed lines indicate *C. cayetanensis* and the solid lines indicate *C. ashfordi*. Line color corresponds to year of infection. The X-axis lists the different diversity indices measured by Hill numbers: at $q = 0$, the returned Hill number value is the richness of mitochondrial alleles found in the year; as q approaches 1, the Hill number value is the Shannon index; and at $q = 2$, the calculated Hill number value is the Simpson index. For each value of q, higher values on the Y-axis indicate a more diverse population.

fact that the majority of strains occur sporadically or have indeterminant temporal patterns suggests that sources of contamination for food vehicles of cyclosporiasis are relatively isolated for a given outbreak. Since *Cyclospora* parasites sexually reproduce within humans and then contaminate produce/water after being shed in human feces, it is likely that *Cyclospora* genetic diversity seen in US cyclosporiasis cases is a subset of the diversity found in human populations from endemic regions. For example, the source for a sporadic genetic type observed in cyclosporiasis cases in the USA may be found in a limited geographical area or from a small group of infected individuals. Alternatively, outbreaks caused by these sporadic strains could be related to the maintenance and circulation of many strains across a wide geographical area at low abundance, where due to the diversity of *Cyclospora* it is unlikely that the same strain would be observed across multiple years. Additionally, the sexual reproductive cycle of *Cyclospora* likely continually adds to the diversity of isolates circulating within humans living in endemic regions, contributing to the trend that many strains tend to be encountered infrequently. The observation of a few, highly prevalent, recurring strains may be related to different, but not mutually exclusive, scenarios where *Cyclospora* can contaminate produce. One explanation could be the presence of a few highly reproductively successful clades that are geographically widespread, and thus various produce items across multiple growing regions can be contaminated with the same strain. Conversely, a strain may be endemic to a relatively small geographical area, but this area accounts for a large proportion of fresh produce consumed in the USA, meaning this strain is found in cyclosporiasis patients on an annual basis. A third consideration is that the recurrent strains consist of different genetic types; however, our current genotyping approach is unable to split these types into discrete groups. Incorporation of additional genotyping markers will help address this question in the future.

Perhaps the most surprising observation made here is the rapid increase in relative abundance of specific mitochondrial genotypes in both *C. ashfordi* and *C. cayetanensis*. Specifically, the *Short* + *F2* genotype from 2018 through 2022 swiftly grew in proportional representation, including a species-specific trend for *C. ashfordi* where we observed near fixation of the *Short* + *A1* + *F2*, which expanded to > 96% of *C. ashfordi* isolates by 2022. The key difference between *C. cayetanensis* and *C. ashfordi* in terms of the mitochondrial genotype shift is that *A1* and *A3* are both found with *Short* + *F2* in *C. cayetanensis* genotype, while only *A1* is found in *C. ashfordi*. While the finding that the *Short* + *F2* genotype reached over 80% prevalence in both species in a short period (i.e. less than five years) could be the result of random genetic drift, the paralleled rapid increase in relative abundance in both *C. ashfordi* and *C. cayetanensis* suggests an environmental or biological pressure that is selecting for outbreaks caused by parasites carrying this mitochondrial genotype. This is because the two species are not thought to sexually reproduce (Barratt et al., 2023), and thus mirrored random genetic drift resulting in the same rapid increase in abundance of a specific mitochondrial allele seems highly unlikely. The fact that *C. ashfordi* predominantly possesses the MSR *A1* haplotype in conjunction with the *Short* + *F2* combination, while *C. cayetanensis* possesses either *A1* or *A3*, may be due to *C. ashfordi* having a smaller population size and less initial mitochondrial diversity after speciation, particularly considering that

the *Short* + *A1* + *F2* genotype was already found in over three-quarters *C. ashfordi* isolates in 2018 when our analysis began. Although our study only includes genotyping data from a five-year period, due to the lack of extensive genotyping data from before 2018, we believe our observations point to a clear pattern with regard to Mt genetic diversity in *Cyclospora*.

The trend towards fixation of specific mitochondrial genotypes was not replicated in the nuclear markers. However, it must be noted that the six nuclear markers used in CYCLONE cover less than 0.1% of the complete *Cyclospora* genome. While this combination of markers has proven useful for linking strains to outbreaks (Barratt et al., 2021, 2022; Jacobson et al., 2022), it is likely that these markers do not capture a sufficient amount of the *Cyclospora* nuclear genome to detect similar evolutionary trends. Additionally, the nuclear markers are subject to genetic recombination, while the mitochondrial markers are not, which may further complicate the detection of genetic trends within the six nuclear markers analyzed using CYCLONE. With that said, the impact of genetic recombination on *Cyclospora* genetic diversity would be best evaluated with a whole-genome sequencing approach, where both intragenic and intergenic recombination can be assessed.

The function of the Mt junction region is not well understood, and it is difficult to speculate on the biological mechanisms that may be driving selection for the *Short* Junction alleles. It is possible that a shift in Junction repeat lengths is related to genomic flexibility (Wickstead et al., 2003) and there are unknown factors favoring a shorter junction over the past five years. It should be noted, however, that the fixation of mitochondrial alleles may not be biologically or evolutionarily driven. Rather, there may be shifts in produce importation practices that could account for the trends observed. For example, specific *Cyclospora* strains could be isolated to specific growing regions and regulatory/economic, or SARS-CoV-2 pandemic-related reasons could bring about shifts in where produce was imported from (Karov et al., 2009; Huang, 2013; Chenarides et al., 2021). However, this would have required a steady shift in supply chains for most or all US states that experience cyclosporiasis outbreaks, and for various produce items consumed by Americans, which seems unlikely. Furthermore, large-scale market and labor data suggest that North American fresh produce supply chains were resilient through the SARS-CoV-2 pandemic in 2020 (Chenarides et al., 2021), suggesting there were no widespread shifts in produce supply chains that would impact genetic diversity associated with cyclosporiasis cases in the USA.

The absence of any major changes over time at the higher levels of classification (species, subpopulation), compared to notable patterns at the lower classification levels (strain, mitochondrial alleles) supports that those forces impacting *Cyclospora* genetic diversity are likely widespread, whether they be biological/environmental or related to produce importation decisions. For example, the decreasing abundance of *Long* Mt Junction and MSR *F1* haplotypes is more extreme in *C. ashfordi*, but the trend is also observed in *C. cayetanensis*. Likewise, we observe large recurring and year-specific sporadic genetic clusters in both *C. cayetanensis* and *C. ashfordi*, which indicates that the population dynamics appear to be broadly similar between the two species.

It is difficult to ascertain the exact reason for the patterns we observed without successful traceback investigations. Our data suggest that there are intriguing genetic patterns developing related to cyclosporiasis cases in the USA; yet we are not able to draw strong conclusions about the reasons for these patterns without knowing the produce item (s) linked to specific genetic types and whether they are associated with produce grown in the USA or imported produce. The paucity of knowledge on the genetic diversity of *Cyclospora* in regions where cyclosporiasis is endemic further limits our interpretations. Likewise, our dataset only consists of cyclosporiasis outbreak specimens from the USA, which means we are highly likely to be undersampling the true genetic diversity of *Cyclospora* parasites found in people, soil, water, and produce from various geographical regions (Chacin-Bonilla and Santin, 2023). This lack of non-outbreak samples may obfuscate larger patterns of change in *Cyclospora* genetic diversity. Genotyping environmental samples is challenging due to low parasite load in these matrices (Durigan et al., 2020). Regardless, strides have been made towards detecting *Cyclospora* in environmental samples (Durigan et al., 2022), and genotyping data generated from such samples may soon follow. Combining the data generated here with environmental genotypes and improved sampling from humans in endemic regions, in addition to improved traceback information, would likely shed light on the mechanisms by which *Cyclospora* continues to infect thousands of Americans annually. Nevertheless, there are still uncertainties about how long oocysts can remain viable in the environment (Chacín-Bonilla, 2010), and it is important to acknowledge that environmental genotyping may yield data from oocysts that are no longer infectious.

Despite the limitations outlined above, the data presented here may prove useful to epidemiologists. If a future outbreak is found to be associated with a previously sporadic strain with a known epidemiological link, epidemiologists could focus on any exposure similarities from the ongoing outbreak and the previous outbreak. An outbreak from a recurring strain may seem less informative at first glance; however, these data can be useful under the assumption that recurring strains are likely to be from a widespread, reproductively successful parasite. Therefore, what seems like a single outbreak, may in reality be the same genotype causing distinct outbreaks, and epidemiologists can use this information to see if there are any splits in exposures, rather than trying to find one exposure linking all cases together.

5. Conclusions

Addressing temporal trends at each level of genetic substructure is important for painting a complete picture of cyclosporiasis outbreaks in the USA. Multilevel genetic characterization can be used to build highly sensitive profiles of what genetic types are making Americans sick on a yearly basis. While using these data to proactively link genetic clusters to specific growing regions or produce types is not on the immediate horizon, the first step is to build a detailed genetic profile of the types of *Cyclospora* that infect US residents. Moving forward, these genetic profiles could be reviewed in the context of existing epidemiological and environmental information which could help identify associations between specific species/subpopulations/strains/alleles and certain types of produce or food processing systems. Ultimately, continuing to monitor genetic trends in *Cyclospora* using methods similar to those applied here will help to improve cyclosporiasis outbreak and traceback investigations in the longer term, lessening the impact of foodborne disease in the USA.

Funding

This work was supported by the US Centers for Disease Control and Prevention's Advanced Molecular Detection (AMD) Initiative.

Ethical approval

Ethics approval for the use of clinical specimens was reviewed by the CDC Center for Global Health Human Research Protection Office under project determination number 2018–123. The need for patient informed consent was waived because the specimens were de-linked from any personal identifiers prior to submission to CDC.

CRediT authorship contribution statement

David K. Jacobson: Conceptualization, Methodology, Formal analysis, Data curation, Investigation, Visualization, Writing – original draft. **Anna C. Peterson:** Investigation, Data curation, Writing – review & editing. **Yvonne Qvarnstrom:** Project administration, Supervision, Resources, Writing – review & editing. **Joel L.N. Barratt:** Conceptualization, Software, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data supporting the conclusions of this article are included within the article and its supplementary files. FASTQ sequence data for all isolates analysed in this manuscript are available under NCBI BioProject PRJNA578931. Readers may contact the authors for access to the CYCLONE code used in this manuscript's bioinformatic analysis.

Acknowledgements

We would like to thank all participating laboratories that submitted samples and/or data that make the *Cyclospora* genotyping project at CDC possible. We would like to thank cyclosporiasis epidemiologists at CDC, namely Lauren Ahart, Marion Rice, and Anne Straily, who link genotyping data with ongoing cyclosporiasis outbreaks in the United States.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crpvbd.2023.100145.

References

Ahart, L., Jacobson, D., Rice, M., Richins, T., Peterson, A., Zheng, Y., et al., 2023. Retrospective evaluation of an integrated molecular-epidemiological approach to cyclosporiasis outbreak investigations - United States, 2021. Epidemiol. Infect. 151, e131.

Alberdi, A., Gilbert, M.T.P., 2019. A guide to the application of Hill numbers to DNA-based diversity analyses. Mol. Ecol. Resources 19, 804–817.

Almeria, S., Cinar, H.N., Dubey, J.P., 2019. *Cyclospora cayetanensis* and cyclosporiasis: An update. Microorganisms 7, 317.

Bailey, P., Chang, D.K., Nones, K., Johns, A.L., Patch, A.-M., Gingras, M.-C., et al., 2016. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 531, 47–52.

Barratt, J., Ahart, L., Rice, M., Houghton, K., Richins, T., Cama, V., et al., 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, 2176–2180.

Barratt, J., Houghton, K., Richins, T., Straily, A., Threlkel, R., Bera, B., et al., 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.

Barratt, J.L., Plucinski, M.M., 2023. Epidemiologic utility of a framework for partition number selection when dissecting hierarchically clustered genetic data evaluated on the intestinal parasite *Cyclospora cayetanensis*. Am. J. Epidemiol. 192, 772–781.

Barratt, J.L.N., Park, S., Nascimento, F.S., Hofstetter, J., Plucinski, M., Casillas, S., et al., 2019. Genotyping genetically heterogeneous *Cyclospora cayetanensis* infections to complement epidemiological case linkage. Parasitology 146, 1275–1283. https://doi.org/10.1017/S0031182019000581.

Barratt, J.L.N., Shen, J., Houghton, K., Richins, T., Sapp, S.G., Cama, V., et al., 2023. *Cyclospora cayetanensis* comprises at least 3 species that cause human cyclosporiasis. Parasitology 150, 269–285.

Casillas, S.M., Bennett, C., Straily, A., 2018. Notes from the field: Multiple cyclosporiasis outbreaks - United States, 2018. Morb. Mortal. Wkly. Rep. 18, 1101–1102.

Chacín-Bonilla, L., 2010. Epidemiology of *Cyclospora cayetanensis*: A review focusing in endemic areas. Acta Trop. 115, 181–193.

Chacin-Bonilla, L., Santin, M., 2023. *Cyclospora cayetanensis* infection in developed countries: Potential endemic foci? Microorganisms 11, 540.

Chao, A., Gotelli, N.J., Hsieh, T., Sander, E.L., Ma, K., Colwell, R.K., et al., 2014. Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. Ecol. Monogr. 84, 45–67.

Chenarides, L., Richards, T.J., Rickard, B., 2021. Covid-19 impact on fruit and vegetable markets: One year later. Can. J. Agroecon. 69, 203–214.

Cinar, H.N., Gopinath, G., Jarvis, K., Murphy, H.R., 2015. The complete mitochondrial genome of the foodborne parasitic pathogen *Cyclospora cayetanensis*. PLoS One 10, e0128645.

Dixon, P., 2003. Vegan, a package of R functions for community ecology. J. Veg. Sci. 14, 927–930.

Durigan, M., Murphy, H.R., da Silva, A.J., 2020. Dead-end ultrafiltration and DNA-based methods for detection of *Cyclospora cayetanensis* in agricultural water. J. Appl. Environ. Microbiol. 86, e01595-01520.

Durigan, M., Patregnani, E., Gopinath, G.R., Ewing-Peeples, L., Lee, C., Murphy, H.R., et al., 2022. Development of a molecular marker based on the mitochondrial genome for detection of *Cyclospora cayetanensis* in food and water samples. Microorganisms 10, 1762.

Gopinath, G., Cinar, H., Murphy, H., Durigan, M., Almeria, M., Tall, B., et al., 2018. A hybrid reference-guided de novo assembly approach for generating *Cyclospora* mitochondrion genomes. Gut Pathog. 10, 1–8.

Hill, M.O., 1973. Diversity and evenness: A unifying notation and its consequences. Ecology 54, 427–432.

Huang, S.W., 2013. Imports contribute to year-round fresh fruit availability. US Department of Agriculture, Economic Research Service. https://www.ers.usda.gov/webdocs/outlooks/37056/41739_fts-356-01.pdf?v=5548.2.

Jacobson, D., Barratt, J., 2023. Optimizing hierarchical tree dissection parameters using historic epidemiologic data as 'Ground Truth'. PLoS One 18, e0282154.

Jacobson, D., Zheng, Y., Plucinski, M.M., Qvarnstrom, Y., Barratt, J.L., 2022. Evaluation of various distance computation methods for construction of haplotype-based phylogenies from large mlst datasets. Mol. Phylogenet. Evol. 177, 107608.

Jacobson, D.K., Low, R., Plucinski, M.M., Barratt, J.L., 2023. An improved framework for detecting discrete epidemiologically-meaningful partitions in hierarchically clustered genetic data. Bioinform. Adv. 2023, vbad118.

Karov, V., Roberts, D., Grant, J.H., Peterson, E.B., 2009. A preliminary empirical assessment of the effect of phytosanitary regulations on US fresh fruit and vegetable imports. In: Agricultural & Applied Economics Association 2009 AAEA & ACCI Joint Annual Meeting. Milwaukee, WI, USA.

Langfelder, P., Zhang, B., Horvath, S., 2008. Defining clusters from a hierarchical cluster tree: The dynamic Tree Cut Package for R. Bioinformatics 24, 719–720. https://doi.org/10.1093/bioinformatics/btm563.

Maechler, M., 2018. Cluster: Cluster analysis basics and extensions. In: R Package, version 2.0.

Mathison, B.A., Pritt, B.S., 2021. Cyclosporiasis - updates on clinical presentation, pathology, clinical diagnosis, and treatment. Microorganisms 9, 1863.

Nascimento, F.S., Barratt, J., Houghton, K., Plucinski, M., Kelley, J., Casillas, S., et al., 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, F.S., Barta, J.R., Whale, J., Hofstetter, J.N., Casillas, S., Barratt, J., et al., 2019. Mitochondrial junction region as genotyping marker for *Cyclospora cayetanensis*. Emerg. Infect. Dis. 25, 1314.

Pohlert, T., 2016. Trend: Non-parametric trend tests and change-point detection. Tech. Rep. 4, 1–18.

Rosato, A., Tenori, L., Cascante, M., De Atauri Carulla, P.R., Martins dos Santos, V.A., Saccenti, E., 2018. From correlation to causation: Analysis of metabolomics data using systems biology approaches. Metabolomics 14, 37.

Strausbaugh, L.J., Herwaldt, B.L., 2000. *Cyclospora cayetanensis*: A review, focusing on the outbreaks of cyclosporiasis in the 1990s. Clin. Infect. Dis. 31, 1040–1057.

Verrow, S., Blair, M., Packard, B., Godfrey, W., 2019. Gel-free size selection using SPRIselect for next generation sequencing.. BeckmanCoulter. https://ls.beckmancoulter.co.jp/files/appli_note/Gel_Free_Using_SPRIselect.pdf.

Wickstead, B., Ersfeld, K., Gull, K., 2003. Repetitive elements in genomes of parasitic protozoa. Microbiol. Mol. Biol. Rev. 67, 360–375.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.Y., 2017. Ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36.