Genotyping Genetically Heterogeneous *Cyclospora cayetanensis* Infections to Complement Epidemiological Case Linkage

Joel L. N. Barratt[1,2,†,*], **Subin Park**[1,2,†], **Fernanda S. Nascimento**[1], **Jessica Hofstetter**[1,2], **Mateusz Plucinski**[4], **Shannon Casillas**[1], **Richard S. Bradbury**[1], **Michael J. Arrowood**[3], **Yvonne Qvarnstrom**[1], **Eldin Talundzic**[4]

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

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

[3]Waterborne Disease Prevention Branch, National Center for Enteric and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA

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

Abstract

Sexually reproducing pathogens such as *Cyclospora cayetanensis* often produce genetically heterogeneous infections where the number of unique sequence types detected at any given locus varies depending on which locus is sequenced. The genotypes assigned to these infections quickly become complex when additional loci are analyzed. This genetic heterogeneity confounds the utility of traditional sequence-typing and phylogenetic approaches for aiding epidemiological

trace-back, and requires new methods to address this complexity. Here, we describe an ensemble of two similarity-based classification algorithms, including a Bayesian and heuristic component that infer the relatedness of *C. cayetanensis* infections. The ensemble requires a set of haplotypes as input and assigns arbitrary distances to specimen pairs reflecting their most likely relationships. The approach was applied to data generated from a test cohort of 88 human fecal specimens containing *C. cayetanensis*, including 30 from patients whose infections were associated with epidemiologically defined outbreak clusters of cyclosporiasis. The ensemble assigned specimens to plausible clusters of genetically related infections despite their complex haplotype composition. These relationships were corroborated by a significant number of epidemiological linkages ($P<0.0001$) suggesting the ensemble's utility for aiding epidemiological trace-back investigations of cyclosporiasis.

Keywords

Cyclospora cayetanensis; bioinformatics; epidemiology; Bayesian; heuristic; ensemble; algorithm

INTRODUCTION

Cyclosporiasis is a foodborne diarrheal disease caused by the human intestinal pathogen *Cyclospora cayetanensis*. Although the parasite is endemic to tropical and sub-tropical regions, seasonal outbreaks of cyclosporiasis often occur in developed temperate countries such as the United States (U.S.) and Canada. In the past, outbreaks occurring in the U.S. and Canada have been associated with imported fresh produce items including berries, herbs, vegetables and salads, imported from endemic countries (Abanyie *et al.*, 2015; Ortega & Sanchez, 2010; Sim *et al.*, 2017; Whitfield *et al.*, 2017). Difficulties are often encountered in identifying vehicles of *C. cayetanensis* infection in the U.S., as many cases are geographically dispersed and seemingly sporadic. Cases that result from a single imported food source may lack a common source of exposure (e.g., a single restaurant or event), and thus not be readily epidemiologically linked. Additionally, detecting the parasite on or in implicated food items has rarely been successful. Due to these challenges, the development of molecular typing tools for *C. cayetanensis* is a major priority for U.S. public health agencies (Abanyie *et al.*, 2015; Hall *et al.*, 2011).

Whole genome sequencing is a routine practice for supporting epidemiological investigations of outbreaks caused by bacterial pathogens in some public health settings (Brown *et al.*, 2018; Kwong *et al.*, 2016; Quainoo *et al.*, 2017) (e.g. PulseNet (CDC, 2017c)). These investigations involve examination of whole genome SNP profiles that distinguish related and unrelated isolates, facilitating linkage to a common source or exposure event (Kanagarajah *et al.*, 2018). This procedure is made feasible for routine applications by the comparatively small genome size of many bacterial pathogens i.e. from ~3–6 megabases (MB), and the ability to culture them directly from clinical specimens enabling enrichment of their genetic material. By comparison, the *C. cayetanensis* genome is approximately 10-fold larger (~44 MB (Qvarnstrom *et al.*, 2015)), and methods for its cultivation have not been described. Fewer than 10^5 parasites are typically present in diagnostic human fecal specimens yielding only picogram amounts of *C. cayetanensis* DNA

(Nascimento *et al.*, 2016). Therefore, parasite enrichment is required prior to DNA extraction and genome sequencing. Humans are the only known host of *C. cayetanensis* and previous attempts to develop animal models of cyclosporiasis have failed (Eberhard *et al.*, 2000). As a result, animal bioassay is not a viable approach for *C. cayetanensis* enrichment even if the high costs and time associated with such an undertaking were of no consequence in a routine setting. Due to these obstacles, multi-locus sequence typing (MLST) is currently the most feasible option for *C. cayetanensis* genotyping. However, given the genetic heterogeneity of *C. cayetanensis* infections, analysis of a MLST dataset generated for this pathogen requires the use of specialized analysis methods or indices that can accommodate this complexity.

While there is an increasing need for robust *C. cayetanensis* typing tools, the genetic complexity of *C. cayetanensis* infections has been a challenge for their development. Stool specimens from cyclosporiasis patients may contain one, two or more sequence types for a given locus depending on the locus sequenced. These datasets become increasingly complex as additional loci are examined and the number of permutations increases. This confounds phylogenetic inference as a means of assisting epidemiological case-linkage. The field of molecular epidemiology establishes that pathogens linked to a common source are usually genetically related. While this appears to be the case for *C. cayetanensis*, due to sexual re-assortment of this pathogens' genome within the gastrointestinal tract of its human host, no two infections are completely alike. This same phenomenon is not encountered in asexually reproducing prokaryotes and viruses that typically present as clonal populations in outbreak scenarios (Tibayrenc & Ayala, 2012), a feature exquisitely suited to trace-back and case-linkage investigations aided by phylogenetic approaches (Romero-Severson *et al.*, 2016).

In the field of population genetics, fixation indices (F statistics) such as F_{ST} are commonly used to define the genetic structure of sexually reproducing populations (Nagylaki, 1998). This index is based largely on haplotype frequencies and is considered a useful indicator of population differentiation. F_{ST} has been applied to population studies of various apicomplexan parasites (Feng *et al.*, 2013; Lumkul *et al.*, 2018; Muwanika *et al.*, 2016) and is used as a measure of distance, enabling comparison of populations in a pairwise fashion where the larger the F_{ST} value, the more genetically isolated the two populations are. In the context of epidemiological case linkage the F_{ST} statistic is not always applicable as it requires that the populations undergoing comparison are already defined by a tangible characteristic such as their geographical distribution. While it is not inconceivable that an index based on haplotype frequency might possess utility for assisting epidemiological case-linkage, information on how the test populations are to be defined is often absent or unclear in the context of U.S. cyclosporiasis outbreaks. As such, indices like F_{ST} cannot be used directly. Another drawback of using fixation indices in this context is that their accuracy relies on the allele frequencies in a population. F_{ST} assesses whether two populations possess an expected amount of heterozygosity or if there is a reduction in heterozygosity based on the Hardy–Weinberg expectation (Crow, 1988). Because a single *C. cayetanensis* infection represents a sexually reproducing population of its own, and given that these infections may involve one, two, three or more sequence types at various loci, the true frequency of alleles (or sequence types) cannot be determined from a standard MLST

dataset, yet these frequencies are required for calculation of true Hardy-Weinberg proportions.

Bioinformatic approaches that support the analysis of complex *C. cayetanensis* MLST datasets must account for these limitations, yet provide a robust estimation of the most plausible genetic relationships. Additionally, given the unique challenges discussed above, development of novel methods for analysis of this data is warranted. We describe an ensemble of two similarity-based classification algorithms including a Bayesian and heuristic component, for inferring genetic relationships between *C. cayetanensis* infections. As input, the ensemble requires a set of haplotypes from each specimen enriched by PCR and then Sanger sequenced. The algorithms' design includes aspects of Mendelian genetics, information theory, frequentist probabilities, and Bayesian probabilities, and it generates a set of pairwise distances that reflect the most plausible relationships between the specimens in the study cohort. These distances are then clustered for downstream interpretation.

To assess the performance of our approach we selected three *C. cayetanensis* genotyping markers, including two nuclear loci and one mitochondrial locus and Sanger sequenced them from 88 human fecal specimens containing *C. cayetanensis*. This dataset included 30 specimens from people whose infections had been epidemiologically linked to U.S. outbreaks of cyclosporiasis. Specimens from geographically and temporally diverse cases of cyclosporiasis were also analyzed, including one specimen from China that served as an outlier for delineating clusters. Manual examination of the haplotype composition of specimens that were linked using the ensemble confirmed that the distances generated result in the assignment of plausible genetic linkages. On comparing the links assigned using epidemiological methods to those assigned using our method, the concordance between the two approaches was statistically significant, suggesting the utility of this approach for assisting epidemiological trace-back and case-linkage investigations of cyclosporiasis.

MATERIALS AND METHODS

Epidemiologic investigations and classification

Fecal specimens were categorized as either associated with outbreaks or as seemingly sporadic cases of cyclosporiasis identified in the U.S. or in other countries. The categorization of specimens as associated with a U.S. outbreak was based on linkages established using current epidemiological approaches including, but not limited to, investigation of temporospatial links, shared exposures to stores or restaurants, distribution supply chains, and food production facilities (CDC, 2017a; CDC, 2017b). Outbreaks were defined as at least two epidemiologically linked cases (e.g., a cluster of cyclosporiasis cases linked to a restaurant, supplier, store and/or event). A temporospatial cluster was defined as cases occurring in the same geographical area (e.g., in the same community or town) where illness onset dates fell within approximately 15 days of each other. Epidemiologic evidence for linking cases in persons with common exposures (e.g., restaurant, grocery store, and/or social events) is typically stronger than for temporospatial clusters. An international travel-associated case was defined as a case in a person who spent at least one day during their pertinent incubation period (i.e., 14 days before illness onset) outside of the U.S.A.

DNA Extraction from Feces and Genome Sequencing

Fecal specimens were washed free of preservative where necessary, first by centrifugation in a Sorvall Legend Micro 17R centrifuge at maximum speed (13300 RPM) and then decanting the supernatant. Approximately 500μL of PBS was added to each fecal specimen with agitation, until a thick slurry was formed. The Universal Nucleic Acid Extraction (UNEX) method described by Shields *et al.* (Shields *et al.*, 2013) was followed with the adjustments described by Qvarnstrom *et al.* (Qvarnstrom *et al.*, 2017). As an internal control for this workflow, repeat fecal specimens (N=3) from each of two cyclosporiasis patients (patients A and B), were extracted alongside batches of other specimens for downstream typing. Extractions were performed on 88 human fecal specimens. To identify genotyping markers, 11 of these 88 human fecal specimens were selected for genome sequencing using methods previously described (Qvarnstrom *et al.*, 2018). These 11 specimens came from patients representing temporally and geographically diverse cases of cyclosporiasis (Table 1), and were selected for whole genome sequencing (WGS) due to the presence of relatively high numbers of oocysts (estimated by microscopy), which is a prerequisite for successful WGS. Extraction of DNA from purified *C. cayetanensis* oocysts and preparation for WGS using the Ovation Ultralow Library System (Nugen Technologies Inc.) was performed as previously described (Nascimento *et al.*, 2016; Qvarnstrom *et al.*, 2018). The resulting libraries were sequenced using the MiSeq Reagent Kit v2 (500 cycles) on the Illumina MiSeq system.

Marker Selection and Polymerase Chain Reaction

A detailed description of the workflow designed for selecting suitable MLST markers is shown in Figure 1. Once candidate markers were identified, the possibility that multiple copies of these loci (i.e., multiple paralogues) exist within the haploid *C. cayetanensis* genome was excluded by examination of the draft genomes generated here using BLAST and discontiguous BLAST searches, with the candidate MLST loci used as query sequences. Ruling out multiple paralogues in the haploid genome would ensure that if mixed Sanger chromatograms were encountered, they would indicate the presence of multiple haploid genomes in the specimen that possess different haplotypes (i.e., true heterozygosity) as opposed to the presence of multiple paralogues that differ slightly in sequence occurring in the same haploid genome. PCR primers were designed to amplify three SNP-dense loci identified using this workflow, which included two nuclear (Nu) loci and one mitochondrial (Mt) locus (Table 1). The Mt locus was amplified using reagents provided in the HotStarTaq Master Mix Kit (Qiagen). Amplification was performed with an initial melt at 95°C for 15 min, followed by 35 cycles of 94°C for 30 sec, annealing at 55°C for 30 sec; and extension at 72°C for 45 sec, concluding with a final extension step of 72°C for 10 min and holding at 4°C. Each nuclear locus was amplified using reagents accompanying the Q5 High-Fidelity DNA Polymerase (New England BioLabs). For the 360i2 locus, amplification was performed with an initial melt of 98°C for 2 min, followed by 35 cycles of 98°C for 15 sec; 67°C for 30 sec and 72°C for 45 sec. This was followed by a final extension step of 72°C for 5 min and holding at 4°C. For the 378 locus, amplification was performed with an initial melt of 98°C for 2 min, followed by 35 cycles of 98°C for 15 sec; 71°C for 30 sec and 65°C for 45 sec. This was followed by a final extension step of 65°C for 5 min and holding at 4°C. All reactions contained 10 picomoles of forward and reverse primer and 1 μL of template

DNA, in a total reaction volume of 25 μL. Amplicons were subjected to agarose gel electrophoresis on 1.5% gels, followed by staining with ethidium bromide for visualization under UV light.

Sanger Sequencing

The PCR products were purified using a Monarch® PCR & DNA Cleanup Kit (New England Biolabs, Ipswich, MA) and sequenced using the BigDye Terminator V3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA). The DyeEx 2.0 Spin Kit (Qiagen, Hilden, Germany) was used to remove the unincorporated dyes (Qiagen, Hilden, Germany). All PCR products were sequenced in both directions on an ABI PRISM® 3100 Genetic Analyzer (AppliedBiosystems, Foster City, CA). The ABI files were imported into Geneious v10 (Kearse *et al.*, 2012) for quality trimming, generation of consensus sequences, and sequence alignments. Heterozygous sequences were identified with the aid of the Geneious Heterozygote Plugin (Version: 1.5.1), using a peak similarity threshold of 25% followed by confirmation by manual inspection. Finished sequences were aligned and exported in .fasta format for downstream analysis. Heterozygous sequences were unphased using the PHASE algorithm in DnaSP v6.10.01, with default parameters (Rozas *et al.*, 2017). Sequences containing more than two haplotypes were identified when automatic unphasing indicated that certain specimens shared no genetic relationship with any other specimen in the dataset. Because this was considered highly unlikely, their chromatograms and consensus sequences were re-inspected and unphased manually where necessary, to identify the underlying haplotypes.

Algorithms and clustering

Detailed descriptions of the Bayesian and heuristic similarity-based classification algorithms underpinning the ensemble are provided in Supplemental File S1. The heuristic algorithm is of novel design while the Bayesian algorithm was modified from a previously described algorithm for comparison of multi-clonal *P. falciparum* infections (Plucinski *et al.*, 2015). Pairwise distances were calculated using both algorithms, the values were normalized, and the mean of these two datasets was taken. The distances were clustered using agglomerative nesting (AGNES), executed in the R package 'cluster', version 2.0.6 (Maechler, 2011). AGNES was performed using Manhattan distances and the Ward clustering method (Strauss & von Maltiz, 2017) with all other parameters set to default. The 'as.hclust' function was used to convert the 'agnes' object into a 'hclust' object, and the 'as.phylo' function (part of the 'ape' R package) was used to convert this into a 'phylo' object. This 'phylo' object was visualized as a circular cluster dendrogram using the R package 'ggtree' (Yu *et al.*, 2018). A specimen from China was used as an outlier for delineating clusters as it possessed a mitochondrial haplotype not detected in any other specimen as well as a rare nuclear haplotype observed only in one other specimen (Supplemental File S2). To assess the concordance between the epidemiologically assigned links and those assigned using the ensemble, the probability that any identified associations arose by random chance was determined. Briefly, the probability (*P*) value was calculated using a binomial distribution, the null probability that any two random specimens in the dataset will be linked using our method and, the total number of specimens with epidemiologically defined links (Supplemental File S1).

RESULTS

Polymerase Chain Reaction and Sequence Analysis

Three SNP-rich loci were selected; one Mt locus designated as MSR and two Nu loci designated as 360i2 and 378, with amplicon sizes less than 1 kilobase pair (kb) and capturing between four and twenty SNPs (Table 1). The amplification and sequencing success rate for each of the three loci was greater than 90%. The success rate of sequencing all three loci for a given specimen was 74/88 (84%). The Mt MSR locus was the least diverse with five haplotypes detected. The Nu 378 locus was the most diverse and possessed the most intra-specimen heterozygosity, with ten haplotypes identified. Seven haplotypes were detected for the Nu 360i2 locus. The existence of several homozygous specimens for the Nu 360i2 and Nu 378 loci facilitated accurate unphasing of heterozygous specimens into two distinct haplotypes in most cases. Five specimens possessed a mixture of three confirmed Nu 378 haplotypes, one specimen contained a mixture of four Nu 378 haplotypes, and one specimen contained two Mt MSR haplotypes (Supplementary File S2). The sequencing results supported that the nuclear loci examined exist as one copy in the haploid *C. cayetanensis* genome as BLAST searches against these genomes had originally suggested. For example, at the 360i2 locus, specimens were identified that were homozygous for haplotypes 1, 2, 3, 4, 6 and 7, while nearly all other specimens had different combinations of these. Approximately 20% (17 out of 87) of specimens sequenced at the 360i2 locus were homozygous, supporting that the heterozygosity observed in some chromatograms was due to the presence of different haploid genome copies in the specimen possessing different haplotypes. For the 378 locus, homozygous specimens were found for haplotypes 2, 3, 4 and 7 representing approximately 8% of the specimens (6 out of 80) that were sequenced at this locus. This data supports our original BLAST analysis of the *C. cayetanensis* genomes generated here, indicating that the 360i2 and 378 loci exist as one copy in each haploid genome.

Assignment of specimens to clusters using the ensemble

The 88 specimens genotyped in this study were divided into 16 genetic clusters using the ensemble (Figure 2), with a cutoff assigned at the node where the Chinese specimen split from its nearest neighbor (two specimens from Texas). To assess the performance of the ensemble in terms of the plausibility of its cluster assignments the 88 genotypes were represented as a barcode (Figure 3). Manual examination of the barcodes within each cluster indicated that the ensemble assigned rational distances based on the haplotype composition of each specimen (Figure 3). Some of the 88 specimens had an incomplete genotype due to amplification and or sequencing failures (Table 1, Figure 2). However, the algorithms underlying the ensemble were designed to accommodate incomplete datasets allowing all specimens to be assigned to a cluster. For instance, multiple attempts at PCR amplification of the Nu 378 locus failed for one of three repeat specimens from patient B yet this incompletely typed specimen was still linked to the other specimens from this patient using our method. The parasites within repeat fecal specimens from patients A and B were consistently assigned to the same respective clusters (Figure 2), suggesting that cross contamination between specimens was absent.

Concordance of Genetic Clustering and Epidemiological Data

Nineteen instances occurred where any two specimens linked using the ensemble were also linked using epidemiological methods. The probability that these 19 concordant links arose by random chance was extremely low (P<0.0001). Additionally, given that repeat specimens from patients A and B shared a common source (i.e., they caused the same infections), the ensemble assigned 25 concordant linkages and the probability that these associations arose randomly was also low (raw P = 3.9×10^{-13}). When the links assigned using the Bayesian and heuristic algorithms underpinning the ensemble were examined individually (repeat specimens from patients A and B considered), their concordance with the epidemiology was also unlikely to have arisen randomly (heuristic P=2.3×10^{-10}, Bayesian P=2.7×10^{-8}). However, the P-value obtained using the ensemble was the lowest.

For four of the eight epidemiologically defined outbreak clusters the ensemble-assigned links were fully corroborated by the epidemiological data. This included four cyclosporiasis cases linked to a local business in Texas (Figure 2, red). Sequencing failed at the 378 locus for one of these specimens yet the ensemble still linked this specimen to the other three. Four cyclosporiasis cases were linked to the Maine temporospatial cluster (Figure 2, green) and the ensemble-assigned links were also fully concordant. The cyclosporiasis patients associated with this epidemiological cluster came from neighboring counties of residence and their onset of illness fell within 13 days. These patients had each purchased the same produce item at one of two grocery store chains. The two cyclosporiasis patients linked to Texas restaurant-associated cluster 1 (Figure 2, gold) were also linked by the ensemble. The two patients fell ill one day apart and both had dined at same restaurant. The ensemble also linked two cyclosporiasis cases from South Carolina associated with a temporospatial outbreak cluster (Figure 2, gray). These patients shared the same county of residence and their illness onset dates fell within 7 days of each other. In all of these cases, manual examination of the specimens respective barcodes (Figure 3) supported that the links assigned were rational based on the haplotype composition of the specimens.

For two of the eight epidemiologically-defined outbreaks of cyclosporiasis, the ensemble assigned links that were partially concordant. Only three of seven cases associated with the Michigan event cluster were linked by the ensemble (Figure 2, blue). The foods consumed at this event were prepared in the same kitchen, though came from diverse sources. Patients were exposed via hotel restaurant dining, catered events and/or room service. Consequently, a single food vehicle could not be identified. Despite the lack of concordance with the epidemiology, the haplotype composition of these specimens indicates that the ensemble assigned rational links for all seven cases (Figure 3). For five cases linked to the Wisconsin restaurant-associated cluster (Figure 2, orange), the ensemble assigned links that were partially concordant with the epidemiology. These patients had dined at the same restaurant chain and their onsets of illness fell within nine days of each other. However, three of the five samples associated with this epidemiologically defined cluster had partial genotypes assigned to them due to sequencing failures.

For the specimens in two of the eight epidemiologically defined clusters, the ensemble detected no links. This includes the four specimens associated with the Texas restaurant 2 cluster (Figure 2, purple), where the infections were assigned to four different groups by the

ensemble. The patients associated with this outbreak had dined at the same restaurant chain though in different cities, and their illness onset dates were up to 23 days apart. For the two cyclosporiasis patients that had their infections linked to the Mexican Event associated cluster (Figure 2, brown), the ensemble assigned their infections to different groups. These cyclosporiasis patients had spent one overlapping day at the same event. Despite the lack of concordance with the epidemiological data in some cases, the haplotype composition of the specimens allocated to each cluster suggests that the ensemble had nonetheless assigned plausible linkages (Figure 3).

DISCUSSION

Previous attempts to develop molecular typing tools for *C. cayetanensis* have proved difficult due to challenges in ascertaining cases, collecting specimens, and difficulties associated with PCR amplification and sequencing of target loci (Guo *et al.*, 2016; Li *et al.*, 2017). Like all apicomplexan parasites, *C. cayetanensis* possesses three discrete genomes: one nuclear, one mitochondrial, and a third contained within a vestigial plastid homologous to the chloroplast, called the apicoplast (Cinar *et al.*, 2016; McFadden & Yeh, 2017). Adding to this complexity, the genome sequence of *C. cayetanensis* is relatively large (44 MB), so whole genome sequencing is not feasible for routine genotyping. Therefore, MLST represents the most viable solution for typing *C. cayetanensis*. However, as a consequence of the difficulties discussed above, no independently (i.e., epidemiologically) validated genotyping approaches have been developed for aiding cyclosporiasis outbreak investigations.

The use of phylogenetic inference to aid linkage of cyclosporiasis cases is confounded by the heterogeneous nature of infections which is likely attributable to the parasites sexual reproductive cycle. During this cycle, fusion of a macrogamete (1N) and a microgamete (1N) in the host gut produces a zygote that develops into an oocyst. Oocysts are shed in the hosts' feces and only become infectious after sporulation; a process that takes approximately 2 weeks (Smith *et al.*, 1997). Fully sporulated oocysts possess two sporocysts, each containing two infective sporozoites that are haploid (1N). Sporozoites in an individual sporocyst are thought to be genetically identical based on studies of *Eimeria* (Shirley & Harvey, 1996), while the sporocysts in a single apicomplexan oocyst can be genetically distinct (Mzilahowa *et al.*, 2007). Consequently, one apicomplexan oocyst can be heterozygous, possessing up to two alleles for any given locus so the number of sequence types detected between *C. cayetanensis* infections often varies depending on the locus examined (Supplementary File S2). In this study, approximately 20% of specimens were homozygous at the 360i2 locus, while approximately 8% were homozygous at the 378 locus. Not surprisingly, specimens homozygous at one nuclear locus were often heterozygous at the other and some specimens possessed three or four 378 haplotypes indicating infections of mixed genotypes, which required manual unphasing. For the mitochondrial locus, most specimens possessed a single haplotype with only one specimen possessing two haplotypes.

Examination of the haplotype composition of specimens within each ensemble-assigned cluster indicated that the groupings were plausible (Figure 3). Additionally, the epidemiological data corroborated a significant number of ensemble-assigned links and the

concordance between these methods was unlikely to have arisen by chance. We therefore suggest that the approach shows great promise as an aid for epidemiologic investigation of cyclosporiasis outbreaks.

However, the underlying algorithms have some limitations and are based on several assumptions that should be considered. For instance, the quality of the links assigned relies on the size of the dataset. Preliminary analyses have shown that because the underlying algorithms rely on frequencies, the approach performs better on larger datasets. It is also important to highlight that the cluster dendrograms generated using this approach are not akin to a genetic distance, but more closely represent the probability of relatedness based on a set of frequencies. A single *C. cayetanensis* infection represents a complex mixture of haplotypes and the precise frequency of these cannot be known from this type of data. Consequently, the probabilities calculated as part of the ensemble are based on the frequency of samples possessing each haplotype rather than the frequency of each haplotype in the population. Therefore, both algorithms substitute the frequency of samples possessing each haplotype in place of a true haplotype frequency, which is less accurate though cannot be avoided using the data generated here. Furthermore, the ensemble may be prone to sampling bias, as the links assigned may be skewed in circumstances where genotypes that are very common in the true population are rare in the study cohort. Such an occurrence could lead to specimens being strongly linked by the ensemble even though their genotype is actually common. In circumstances like this the closeness of a relationship may be over-estimated.

Possibly the most important limitation of this strategy (also a limitation of the laboratory methods) is that only a minor portion of the parasites' genome is sampled, meaning the algorithms attempt to extrapolate genetic relationships from limited data. Our three markers resolved the 88 specimens analyzed into 16 clusters (Figure 2), though improved granularity may be achieved with the inclusion of additional typing markers to our panel. We also propose that the use of deep sequencing technologies in place of Sanger would almost certainly improve accuracy by increasing the number of haplotypes detected within each specimen. Additionally, the similarity-based classification algorithms underpinning the ensemble could be modified to accommodate deep sequencing data, such that the normalized read counts might be used in place of a haplotype frequency.

The similarity-based classification algorithms underlying the ensemble were designed to accommodate partial typing datasets (Supplementary file S1), because amplification and sequencing are not always successful for every specimen. Nonetheless, the amplification and sequencing success rate for this scheme (84%) is superior to previously described approaches (~50%) (Guo *et al.*, 2016; Li *et al.*, 2017) despite most specimens used here being several years old. In any case, the results suggest that the ensemble assigns plausible links in the event of missing data. For example, sequencing of the 378 locus failed for a repeat specimen from Patient B, yet the specimen was still linked closely by the ensemble to the other repeat specimens from this same patient. Similarly, one of four specimens linked to the local business cluster from Texas had missing data for locus 378 (Figure 2, red), yet was still linked to the other specimens from this epidemiological cluster. While retention of samples with missing data is not ideal, this feature of the ensemble reduces data wastage and means that specimens with a partial genotype remain useful.

This approach linked specimens associated with four epidemiologically-defined clusters with full concordance. This included specimens associated with the South Carolina temporospatial cluster, the Maine temporospatial cluster, the Texas local business-associated cluster and the cluster associated with Restaurant 1. The approach linked only three of seven cases from the Michigan event associated cluster (Figure 3, cluster 9), though despite this lack of concordance, the four unlinked specimens (PA1462, MI1436, MI1431, MI1430), are genetically distinct from each other at the loci sequenced (Figure 3). Similarly, for the Restaurant 2 cluster and the cluster associated with the Mexican event, there was no concordance between the epidemiological data and the ensemble-assigned links. However, the haplotype composition of each of these specimens also suggested that they were not closely related (Figure 3). In fact, one specimen from the Mexican retreat associated cluster possessed a 360i2 haplotype not observed in any other specimen in the cohort and was assigned to its own unique cluster by the ensemble (Figure 3, cluster 16). Based on this data, we conclude that the lack of concordance between the epidemiology and the ensemble-assigned links is not related to the performance of the algorithm. We note however, that in the absence of a reference *C. cayetanensis* genotyping method, epidemiological linkages remain the only method to assess the performance of this approach.

Despite its limitations, the ensemble assigned specimens to plausible clusters of genetically related infections based on their complex haplotype composition. These relationships were corroborated by a significant number of epidemiological linkages suggesting the ensemble's utility for aiding epidemiological trace-back and case-linkage investigations.

ACCESSION NUMBERS

The Genbank BioSpecimen numbers for all published *C. cayetanensis* genomes are listed in Table 1. The sequence of each haplotype from the 378, 360i2 and MSR loci are available in GenBank under Accession Numbers MH185772 to MH185793.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

REFERENCES

Abanyie F, Harvey RR, Harris JR, Wiegand RE, Gaul L, Desvignes-Kendrick M, Irvin K, Williams I, Hall RL, Herwaldt B, Gray EB, Qvarnstrom Y, Wise ME, Cantu V, Cantey PT, Bosch S, AJ DAS, Fields A, Bishop H, Wellman A, Beal J, Wilson N, Fiore AE, Tauxe R, Lance S, Slutsker L, Parise M and Multistate Cyclosporiasis Outbreak Investigation, T. (2015). 2013 multistate outbreaks of *Cyclospora cayetanensis* infections associated with fresh produce: focus on the Texas investigations. Epidemiol Infect, 143, 3451–3458. doi: 10.1017/S0950268815000370. [PubMed: 25865140]

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA and Pevzner

PA (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol, 19, 455–477. doi: 10.1089/cmb.2012.0021. [PubMed: 22506599]

Borodovsky M and Lomsadze A (2011). Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinformatics, Chapter 4, Unit 4 6 1–10. doi: 10.1002/0471250953.bi0406s35.

Brown AC, Chen JC, Watkins LKF, Campbell D, Folster JP, Tate H, Wasilenko J, Van Tubbergen C and Friedman CR (2018). CTX-M-65 Extended- Spectrum beta-Lactamase-Producing Salmonella enterica Serotype Infantis, United States(1). Emerg Infect Dis, 24, 2284–2291. doi: 10.3201/eid2412.180500. [PubMed: 30457533]

CDC (2017a). Cyclosporiasis: Outbreak Investigations and Updates. Vol. 2017 Centers for Disease Control and Prevention, Global Health, Division of Parasitic Disease, Atlanta, Georgia, USA.

CDC (2017b). Parasites - Cyclosporiasis (Cyclospora Infection). Centers for Disease Control and Prevention.

CDC (2017c). PulseNet. Centers for Disease Control and Prevention, Atlanta, USA.

Cinar HN, Qvarnstrom Y, Wei-Pridgeon Y, Li W, Nascimento FS, Arrowood MJ, Murphy HR, Jang A, Kim E, Kim R, da Silva A and Gopinath GR (2016). Comparative sequence analysis of *Cyclospora cayetanensis* apicoplast genomes originating from diverse geographical regions. Parasit Vectors, 9, 611. doi: 10.1186/s13071-016-1896-4. [PubMed: 27899155]

Crow JF (1988). Eighty years ago: the beginnings of population genetics. Genetics, 119, 473–476. [PubMed: 3042506]

Eberhard ML, Ortega YR, Hanes DE, Nace EK, Do RQ, Robl MG, Won KY, Gavidia C, Sass NL, Mansfield K, Gozalo A, Griffiths J, Gilman R, Sterling CR and Arrowood MJ (2000). Attempts to establish experimental Cyclospora cayetanensis infection in laboratory animals. J Parasitol, 86, 577–582. doi: 10.1645/0022-3395(2000)086[0577:ATEECC]2.0.CO;2. [PubMed: 10864257]

Feng Y, Torres E, Li N, Wang L, Bowman D and Xiao L (2013). Population genetic characterisation of dominant Cryptosporidium parvum subtype IIaA15G2R1. Int J Parasitol, 43, 1141–1147. doi: 10.1016/j.ijpara.2013.09.002. [PubMed: 24126186]

Gardner SN, Slezak T and Hall BG (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics, 31, 2877–2878. doi: 10.1093/bioinformatics/btv271. [PubMed: 25913206]

Guo Y, Roellig DM, Li N, Tang K, Frace M, Ortega Y, Arrowood MJ, Feng Y, Qvarnstrom Y, Wang L, Moss DM, Zhang L and Xiao L (2016). Multilocus Sequence Typing Tool for *Cyclospora cayetanensis*. Emerg Infect Dis, 22, 1464–1467. doi: 10.3201/eid2208.150696. [PubMed: 27433881]

Gurevich A, Saveliev V, Vyahhi N and Tesler G (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29, 1072–1075. doi: 10.1093/bioinformatics/btt086. [PubMed: 23422339]

Hall RL, Jones JL and Herwaldt BL (2011). Surveillance for laboratory-confirmed sporadic cases of cyclosporiasis--United States, 1997–2008. MMWR Surveill Summ, 60, 1–11.

Kanagarajah S, Waldram A, Dolan G, Jenkins C, Ashton PM, Carrion Martin AI, Davies R, Frost A, Dallman TJ, De Pinna EM, Hawker JI, Grant KA and Elson R (2018). Whole genome sequencing reveals an outbreak of Salmonella Enteritidis associated with reptile feeder mice in the United Kingdom, 2012–2015. Food Microbiol, 71, 32–38. doi: 10.1016/j.fm.2017.04.005. [PubMed: 29366466]

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P and Drummond A (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28, 1647–1649. doi: 10.1093/bioinformatics/bts199. [PubMed: 22543367]

Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T and Howden BP (2016). Prospective Whole-Genome Sequencing Enhances National Surveillance of Listeria monocytogenes. J Clin Microbiol, 54, 333–342. doi: 10.1128/JCM.02344-15. [PubMed: 26607978]

Li J, Chang Y, Shi KE, Wang R, Fu K, Li S, Xu J, Jia L, Guo Z and Zhang L (2017). Multilocus sequence typing and clonal population genetic structure of *Cyclospora cayetanensis* in humans. Parasitology, 1–8. doi: 10.1017/S0031182017001299.

Lumkul L, Sawaswong V, Simpalipan P, Kaewthamasorn M, Harnyuttanakorn P and Pattaradilokrat S (2018). Unraveling Haplotype Diversity of the Apical Membrane Antigen-1 Gene in Plasmodium falciparum Populations in Thailand. Korean J Parasitol, 56, 153–165. doi: 10.3347/kjp. 2018.56.2.153. [PubMed: 29742870]

Maechler M (2011). Package 'cluster'. Vol. 2018 pp. version 2.0.6.

McFadden GI and Yeh E (2017). The apicoplast: now you see it, now you don't. Int J Parasitol, 47, 137–144. doi: 10.1016/j.ijpara.2016.08.005. [PubMed: 27773518]

Muwanika V, Kabi F and Masembe C (2016). Population genetic structure of Theileria parva field isolates from indigenous cattle populations of Uganda. Ticks Tick Borne Dis, 7, 291–297. doi: 10.1016/j.ttbdis.2015.11.004. [PubMed: 26613662]

Mzilahowa T, McCall PJ and Hastings IM (2007). "Sexual" population structure and genetics of the malaria agent *P. falciparum*. PLoS One, 2, e613. doi: 10.1371/journal.pone.0000613. [PubMed: 17637829]

Nagylaki T (1998). Fixation indices in subdivided populations. Genetics, 148, 1325–1332. [PubMed: 9539445]

Nascimento FS, Wei-Pridgeon Y, Arrowood MJ, Moss D, da Silva AJ, Talundzic E and Qvarnstrom Y (2016). Evaluation of library preparation methods for Illumina next generation sequencing of small amounts of DNA from foodborne parasites. J Microbiol Methods, 130, 23–26. doi: 10.1016/j.mimet.2016.08.020. [PubMed: 27553132]

Ortega YR and Sanchez R (2010). Update on *Cyclospora cayetanensis*, a food-borne and waterborne parasite. Clin Microbiol Rev, 23, 218–234. doi: 10.1128/CMR.00026-09. [PubMed: 20065331]

Plucinski MM, Morton L, Bushman M, Dimbu PR and Udhayakumar V (2015). Robust Algorithm for Systematic Classification of Malaria Late Treatment Failures as Recrudescence or Reinfection Using Microsatellite Genotyping. Antimicrobial agents and chemotherapy, 59, 6096–6100. [PubMed: 26195521]

Quainoo S, Coolen JPM, van Hijum S, Huynen MA, Melchers WJG, van Schaik W and Wertheim HFL (2017). Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clin Microbiol Rev, 30, 1015–1063. doi: 10.1128/CMR.00016-17. [PubMed: 28855266]

Qvarnstrom Y, Benedict T, Marcet P, Wiegand R, Herwaldt B and da Silva A (2017). Molecular detection of *Cyclospora cayetanensis* in human stool specimens using UNEX-based DNA extraction and real-time PCR. Parasitology, 8, 1–6. doi: 10.1017/S0031182017001925.

Qvarnstrom Y, Wei-Pridgeon Y, Li W, Nascimento FS, Bishop HS, Herwaldt BL, Moss DM, Nayak V, Srinivasamoorthy G, Sheth M and Arrowood MJ (2015). Draft Genome Sequences from *Cyclospora cayetanensis* Oocysts Purified from a Human Stool Sample. Genome Announc, 3. doi: 10.1128/genomeA.01324-15.

Qvarnstrom Y, Wei-Pridgeon Y, Van Roey E, Park S, Srinivasamoorthy G, Nascimento FS, Moss DM, Talundzic E and Arrowood MJ (2018). Purification of Cyclospora cayetanensis oocysts obtained from human stool specimens for whole genome sequencing. Gut Pathog, 10, 45. doi: 10.1186/s13099-018-0272-7. [PubMed: 30337964]

Romero-Severson EO, Bulla I and Leitner T (2016). Phylogenetically resolving epidemiologic linkage. Proc Natl Acad Sci U S A, 113, 2690–2695. doi: 10.1073/pnas.1522930113. [PubMed: 26903617]

Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE and Sanchez-Gracia A (2017). DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. Mol Biol Evol. doi: 10.1093/molbev/msx248.

Schubert M, Lindgreen S and Orlando L (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes, 9, 88. doi: 10.1186/s13104-016-1900-2. [PubMed: 26868221]

Shields JM, Joo J, Kim R and Murphy HR (2013). Assessment of three commercial DNA extraction kits and a laboratory-developed method for detecting *Cryptosporidium* and *Cyclospora* in

raspberry wash, basil wash and pesto. J Microbiol Methods, 92, 51–58. doi: 10.1016/j.mimet. 2012.11.001. [PubMed: 23147278]

Shirley MW and Harvey DA (1996). *Eimeria tenella*: infection with a single sporocyst gives a clonal population. Parasitology, 112 (Pt 6), 523–528. [PubMed: 8684826]

Sim S, Won J, Kim JW, Kim K, Park WY and Yu JR (2017). Simultaneous Molecular Detection of *Cryptosporidium* and *Cyclospora* from Raw Vegetables in Korea. Korean J Parasitol, 55, 137–142. doi: 10.3347/kjp.2017.55.2.137. [PubMed: 28506035]

Smith HV, Paton CA, Mitambo MM and Girdwood RW (1997). Sporulation of *Cyclospora* sp. oocysts. Appl Environ Microbiol, 63, 1631–1632. [PubMed: 9097458]

Strauss T and von Maltitz MJ (2017). Generalising Ward's Method for Use with Manhattan Distances. PLoS One, 12, e0168288. doi: 10.1371/journal.pone.0168288. [PubMed: 28085891]

Tibayrenc M and Ayala FJ (2012). Reproductive clonality of pathogens: a perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. Proc Natl Acad Sci U S A, 109, E3305–3313. doi: 10.1073/pnas.1212452109. [PubMed: 22949662]

Whitfield Y, Johnson K, Hanson H and Huneault D (2017). 2015 Outbreak of Cyclosporiasis Linked to the Consumption of Imported Sugar Snap Peas in Ontario, Canada. J Food Prot, 80, 1666–1669. doi: 10.4315/0362-028X.JFP-17-084. [PubMed: 28876133]

Yu G, Tsan-Yuk Lam T, Zhu H and Guan Y (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. Mol Biol Evol. doi: 10.1093/molbev/msy194.
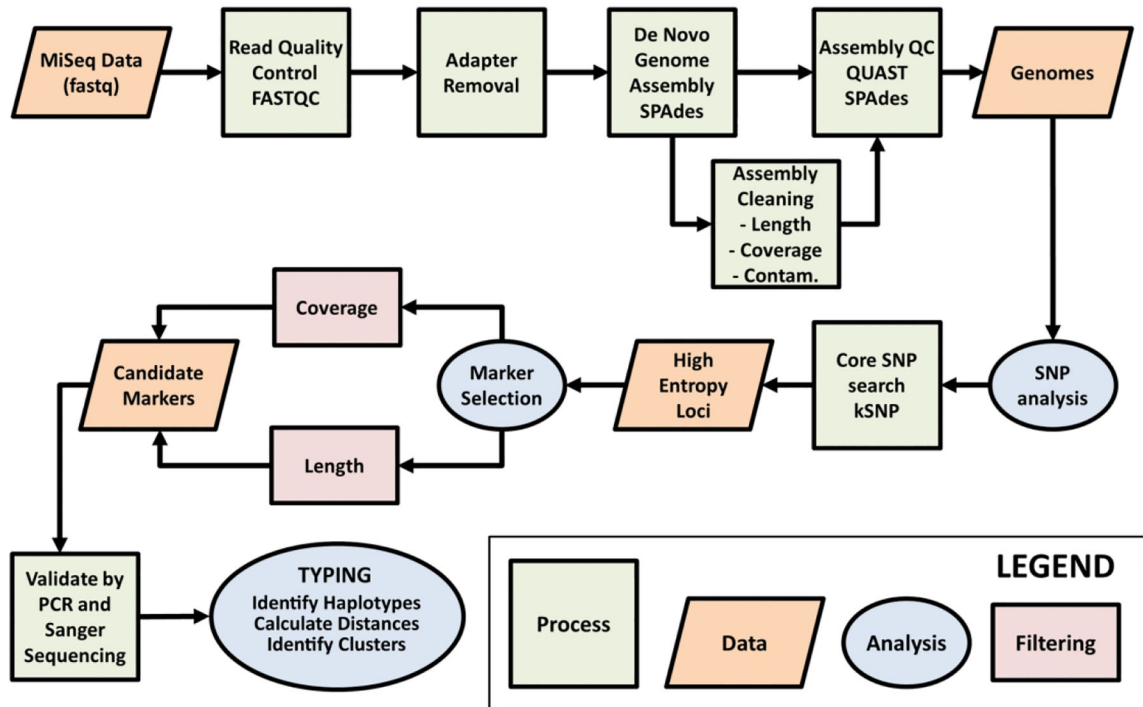
Figure 1. Workflow for selection of *Cyclospora cayetanensis* typing markers

Raw genome sequence data generated on the Illumina MiSeq platform were assessed for quality using FASTQC. AdaptorRemoval v2.1.7 (Schubert *et al.*, 2016) was used to remove adaptor sequences from reads and to merge overlapping paired reads into consensus sequences. SPAades v3.9.0 (Bankevich *et al.*, 2012) was used to *de novo* assemble the reads. During the assembly cleaning process, contigs derived from contaminating (Contam.) prokaryotic human gut flora were removed using BBMap (http://sourceforge.net/projects/bbmap/). The assemblies were assessed for quality using QUAST v4.3 (Gurevich *et al.*, 2013) before and after the cleaning phase. Contigs with 60 times coverage, 3000 base pairs (bp) long and with coding regions identified using GeneMark-ES v4.33 (Borodovsky & Lomsadze, 2011), were retained as part of the core genome. Single nucleotide polymorphisms (SNPs) were detected across the core genome assemblies using kSNP v3.021 (Gardner *et al.*, 2015) and this information was used to identify high-entropy genomic loci. Genomic regions containing high confidence SNPs (i.e., those SNPs within genomic regions of the highest coverage) occurring within SNP-dense regions (i.e., where several informative SNPs exist within a genomic region of less than 1 kilobase pair in size), were identified as candidate typing markers for validation by PCR amplification and Sanger sequencing. The markers with the highest amplification and sequencing success rate were considered ideal candidates for *C. cayetanensis* typing, and were PCR amplified and sequenced from stool specimens provided by a diverse range of patients. The resulting sequences were then subjected to typing.
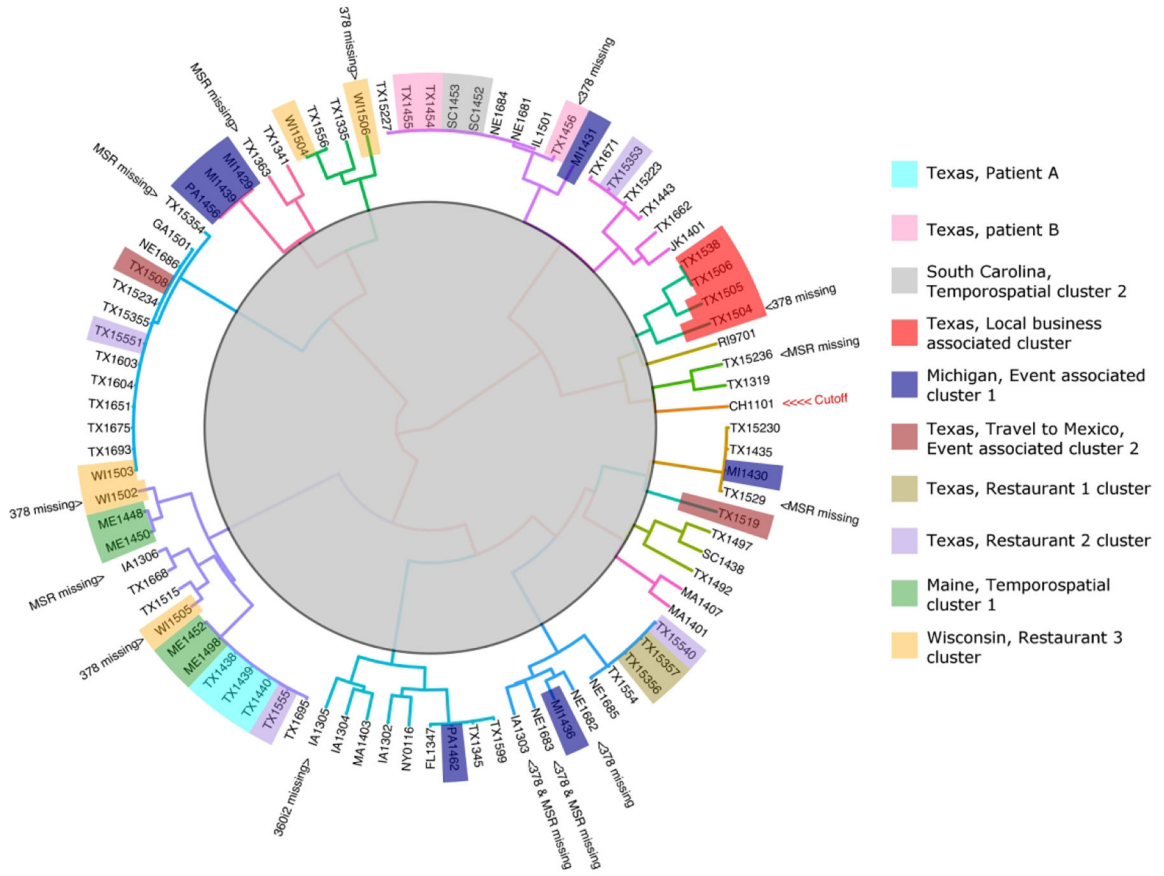
Figure 2. Cluster Dendrogram Generated from the Ensemble Distance Matrix

Our ensemble of two similarity-based classification algorithms resolved the *C. cayetanensis* infections from 88 fecal specimens into sixteen clusters (different branch colors). Clusters were delineated by cutting the tree at the node indicating the separation of the Chinese sample (CHN_HEN01) from its nearest neighbor. The specimen names are shaded in colors according to their epidemiological linkage. Unshaded specimen names represent sporadic or unlinked cases of cyclosporiasis. Specimen identity codes begin with a two letter state abbreviation (except for Jakarta, Indonesia; JK), followed by two numbers indicating the year, and ending a unique identifier assigned to that specimen (2 to 3 digits). The specimen from China (CHN_HEN01) follows a different naming convention as sequence data from this specimen had been submitted to GenBank previously by different investigators (GenBank accession: NW_019211453).

Figure 3. The haplotype composition of each specimen genotyped in this study represented as a barcode

The 88 specimens in the study cohort were assigned to 16 distinct clusters by the ensemble, with cluster assignments shown on the right hand side of each panel. These cluster assignments were made based on the haplotype composition of each sample, with the loci and their respective haplotype numbers shown along the two top rows. Boxes are shaded black if the corresponding haplotype was detected in a specimen. Specimen names are listed in the far left column of each panel. Rows are shaded gray if sequencing was unsuccessful for a given marker. This figure was generated to graphically represent the groupings assigned by the ensemble when presented with a set of complex genotyping data.

Table 1.

Primers designed for PCR enrichment of the selected typing markers

Genome	Locus (Alias)	SNPs[a]	Primer Name	Primer sequence (5′ – 3′)	T_a	Amplicon size (sequence length[β])	Amplification success (%)
Mt	Mt rRNA (MSR)	4	15F	GGACATGCAGTAACCTTTCCG	55°C	686 (573) bp	81/88(92.0)
			688R	AGGAAAGGTTAACCGCTGTCA			
Nu	Nu, undefined (360i2)	20	HC360i2F	CCCATTACGCCGCATAGAGT	67°C	650 (541) bp	87/88(98.9)
			HC360i2R	GCATTGCAAAGCCAGTCAGC			
Nu	Nu, Sec14 family protein (378)	15	HC378F	CCCCTGCCTTGTTCTTGGTGAA	71°C	469 (364) bp	80/88(90.9)
			HC378R	CCGGCGACACAGAGGTACC			

[a]The number of variable sites present after trimming of sequences to equal lengths.

[β]After trimming sequences to equal lengths.