American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2023.

Vol. 192, No. 5
https://doi.org/10.1093/aje/kwad006
Advance Access publication:
January 6, 2023

Practice of Epidemiology

Epidemiologic Utility of a Framework for Partition Number Selection When Dissecting Hierarchically Clustered Genetic Data Evaluated on the Intestinal Parasite *Cyclospora cayetanensis*

Joel L. N. Barratt* **and Mateusz M. Plucinski**

* Correspondence to Dr. Joel L. N. Barratt, Parasitic Diseases Branch, Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, 1600 Clifton Road, NE Atlanta, GA 30329 (e-mail: nsk9@cdc.gov).

Initially submitted September 13, 2021; accepted for publication January 4, 2023.

Comparing parasite genotypes to inform parasitic disease outbreak investigations involves computation of genetic distances that are typically analyzed by hierarchical clustering to identify related isolates, indicating a common source. A limitation of hierarchical clustering is that hierarchical clusters are not discrete; they are nested. Consequently, small groups of similar isolates exist within larger groups that get progressively larger as relationships become increasingly distant. Investigators must dissect hierarchical trees at a partition number ensuring grouped isolates belong to the same strain; a process typically performed subjectively, introducing bias into resultant groupings. We describe an unbiased, probabilistic framework for partition number selection that ensures partitions comprise isolates that are statistically likely to belong to the same strain. We computed distances and established a normalized distribution of background distances that we used to demarcate a threshold below which the closeness of relationships is unlikely to be random. Distances are hierarchically clustered and the dendrogram dissected at a partition number where most within-partition distances fall below the threshold. We evaluated this framework by partitioning 1,137 clustered *Cyclospora cayetanensis* genotypes, including 552 isolates epidemiologically linked to various outbreaks. The framework was 91% sensitive and 100% specific in assigning epidemiologically linked isolates to the same partition.

dendrogram; framework; genetic distance; hierarchical clustering; outbreaks; partitioning; threshold; tree dissection

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

Recent advances in genotyping techniques have introduced novel tools for investigation of parasitic disease outbreaks. Molecular epidemiology allows investigators to compare genotypes to differentiate between concurrent outbreaks and link related genotypes, facilitating identification of common sources (1). Originally applied to outbreaks of bacterial foodborne disease (2, 3), genotyping tools for use in epidemiologic settings have now been developed for more complex infectious agents, such as the apicomplexan parasite *Cyclospora cayetanensis*. *C. cayetanensis* is the etiological agent of cyclosporiasis, a food-borne illness characterized by watery diarrhea, nausea, abdominal cramps, and weight loss (4, 5). *C. cayetanensis* is responsible for annual outbreaks in the United States and is currently reportable in 43 US states, the District of Columbia, and New York City (6).

Linking isolates based on genotype to differentiate concurrent outbreaks is not trivial, especially for complex, sexually reproducing organisms like *C. cayetanensis*. A data set of N genotypes sequenced during an outbreak necessitates comparison of $N(N-1)/2$ genotype pairs and determination of which of these $N(N-1)/2$ pairs are genetically linked. Comparison of genotypes first requires computation of genetic distances (Δ) between all possible pairs. A small distance (i.e., $\Delta \approx 0$) is assigned to pairs possessing a close genetic relationship, and a large distance (i.e., $\Delta \approx 1$) is assigned to unrelated pairs. Most distances fall somewhere between 0 and 1, reflecting varying degrees of similarity

among isolates. These distances are then clustered using a hierarchical clustering strategy to generate hierarchical trees. In epidemiologic contexts, a close genetic relationship revealed by hierarchical clustering is considered evidence that isolates were derived from a common source, which may inform downstream epidemiologic investigations when viewed in association with other data (5).

Hierarchical clustering possesses a major limitation. Hierarchical clusters are not discrete groups but instead are nested, meaning that clusters comprising small numbers of similar individuals exist within larger clusters that get progressively larger as relationships among cluster members become increasingly distant (Figure 1). Hierarchically clustered data sets require partitioning to produce discrete groups comprising isolates that cluster closely and are thus genetically linked (7). Here, the onus is on the investigator to decide what constitutes a sufficiently similar genetic match and then select an appropriate partition number when dissecting hierarchical trees to ensure that each partition comprises a set of sufficiently similar isolates (7). A major problem in the field is that tree dissection is often performed arbitrarily, based on the investigators' personal judgment (8, 9).

The number of partitions selected has a profound impact on the isolates that are perceived as being related or unrelated. Too few partitions can lead to wasted resources as epidemiologists investigate "false leads" identified due to grouping of isolates that are not sufficiently similar. Too many partitions can lead to the separation of isolates that are linked to a common source into distinct groupings. Published statistical methods that predict partition number for clustered data, including the elbow method, gap statistic, and the silhouette method (8–10), could be used to eliminate arbitrary decision making when dissecting hierarchical trees. However, these methods were not designed specifically for molecular epidemiologic applications, so systematic methods for partition number selection in epidemiologic contexts are lacking. Consequently, this process has historically been performed with an arbitrary cutoff selection that can be biased by what investigators believe should be grouped (8, 9). Factors such as the genotyping markers selected, mixed-strain infections, and heterozygosity may result in genotypes that differ slightly among closely related pathogen isolates. Consequently, selecting a partition number where isolates in each partition are separated by a distance of $\Delta = 0$ (i.e., requiring identical genotypes) is a poor choice; this ignores that genotyping data sets from parasites like *C. cayetanensis* and *Plasmodium falciparum* require a more nuanced interpretation due to the factors listed above (4, 11).

A framework for statistically testing the relatedness *P. falciparum* genotype pairs in the context of antimalarial drug efficacy trials was recently described (11). For these trials, *P. falciparum* parasites in the blood of an infected patient are genotyped at day 0 (i.e., when the infection is first identified and before treatment) and again if the patient experiences a recurrent *P. falciparum* infection (11). The pair of genotypes is compared to determine if the day 0 genotype is the same as the genotype observed at the time of recurrence. In that study, comparison of only genotype pairs was required, and that framework established a genetic distance threshold for binary classification of pairs as being "the same" or "different".

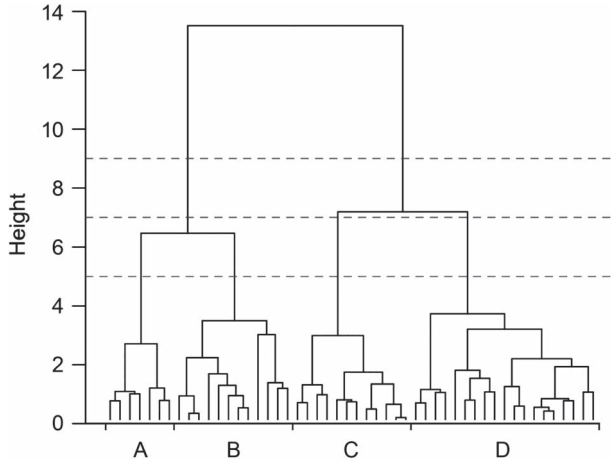

Figure 1. Problem definition: The impact of partition number selection on partition memberships. This figure displays a hierarchical tree with 3 possible dissection heights, indicated by dashed lines. Hierarchical clusters are nested, meaning that clusters comprising small numbers of similar individuals exist within larger clusters that get progressively larger as genetic relationships become increasingly distant. Hierarchically clustered data sets therefore require partitioning by dissecting hierarchical trees at a level that results in groups comprising isolates that probably belong to the same strain. The level at which a hierarchical tree is dissected affects the perceived genetic relationships among clustered isolates. Dissecting this tree at a height of 5 arbitrary units results in 4 partitions, where A, B, C, and D are separated. Alternatively, dissecting the tree at 7 arbitrary units results in 3 partitions, where A and B are assigned to the same partition while C and D remain separated. Dissecting this tree at 9 units results in 2 partitions, one containing A and B and the other containing C and D. When genotyping for epidemiologic purposes, partitions must comprise isolates that are sufficiently similar (i.e., separated by a distance below an appropriate Δ threshold) to warrant epidemiologic follow-up. However, selecting an appropriate number of partitions is not straightforward, is often performed empirically, and may be subject to bias.

Here, we adapt the framework previously used to classify *P. falciparum* genotype pairs to facilitate the unbiased selection of an appropriate partition number (k) when dissecting hierarchical trees. Briefly, a distance matrix is computed for all genotypes using an appropriate genetic distance statistic. This set of distances is then normalized to generate a comparatively unbiased distribution of distances. A threshold distance is computed from the empirical distribution of normalized distances by selecting the distance falling at the empirical lower 5th percentile, to demarcate a threshold distance (11). Finally, A hierarchical tree is generated and dissected at the smallest number of partitions where at least 99.5% of all within-partition distances fall below this threshold.

We evaluated this procedure on 1,137 *C. cayetanensis* genotypes, including 552 with epidemiologic links that were identified during previous cyclosporiasis outbreak investigations (5, 12–15). Genetic distances were calculated as

previously described (4, 5), and a partition number was selected using our framework. By comparing the partitions identified here to available epidemiologic data, we assessed the ability of this framework to detect epidemiologically meaningful partitions. Our approach was highly sensitive and specific at assigning epidemiologically linked isolates to the same partition and therefore provides a robust, systematic method for selecting an appropriate number of partitions from a set of hierarchically clustered genetic distances.

METHODS

Data collection

This study utilized publicly available *C. cayetanensis* genotyping data generated by the US Centers for Disease Control and Prevention (CDC), the Public Health Agency of Canada, and participating US state public health departments as part of genotyping performed during the US cyclosporiasis peak periods of 2018, 2019, and 2020 (5, 12–16). Cyclosporiasis outbreak clusters were defined using epidemiologic methods described in those studies (5, 12–15). These clusters are listed in Table 1. The use of clinical specimens in those studies (5, 12–16) was reviewed by the CDC and was conducted consistent with applicable federal law and CDC policy (Center for Global Health Human Research Protection Office determination number 2018–123). Illumina (San Diego, California) sequence data for all *C. cayetanensis* isolates is accessible under National Center for Biotechnology Information BioProject Number PRJNA578931. Genotypes were determined from this data using methods previously described (5, 15, 16).

Distance calculation and framework for partition number selection

A pairwise distance matrix (M_1) was computed from the genotypes after their conversion to haplotype data sheet (HDS) format (Web Table 1, available at https://doi.org/10.1093/aje/kwad006) using Barratt's heuristic definition of genetic distance, which has been applied previously to genotypes of *C. cayetanensis* (4, 5, 15, 16), *P. falciparum* (11), and nematodes of the genus *Strongyloides* (17). M_1 was hierarchically clustered using Ward's method (5, 15). The resulting hierarchical tree was dissected at the maximum number of partitions resulting in all partitions containing more than 1 isolate. Next, the number of isolates within the partition containing the smallest number of isolates (n) was identified. Subsequently, n isolates were randomly selected from every partition, producing a set (L) of isolates. A second pairwise distance matrix (M_2) was generated containing the distance between every combination of isolate pairs in the set L by extracting the distances for the same pairs from M_1 and inserting these values into M_2. Matrix M_2 therefore comprised a distribution of distances where each strain was represented n times. A threshold distance was computed from the empirical distribution of the distances in M_2 by selecting the distance falling at the empirical lower 5th percentile ($\Delta_{0.05}$), as previously described (11), after exclusion of self-to-self distances. Selecting a genetic distance threshold at the empirical lower 5th percentile is based on the assumption that the distribution of distances in M_2 resembles a unimodal, normal probability distribution. Grounded in this logic, for specimens separated by a genetic distance below the $\Delta_{0.05}$ threshold, the likelihood that this level of genetic similarity is random (i.e., is not due to recent genetic kinship) would be less than 5%, corresponding to a statistical test with a false positive rate (α) of 5%.

Given the stochasticity introduced when randomly selecting isolates in set L, this step (generating sets of L and selection of $\Delta_{0.05}$) was repeated 1,000 times and an average of the 1,000 values of $\Delta_{0.05}$ was taken to demarcate a threshold distance ($\overline{\Delta}_{0.05}$). Finally, M_1 was clustered using Ward's method, and the resulting hierarchical tree was dissected at the smallest partition number (k) where at least 99.5% of all within-partition distances fell below $\overline{\Delta}_{0.05}$. This involved generation of 199 hierarchical trees from M_1 using Ward's method and dissecting each tree at a single value of k ranging from 2 to 200. For each dissected tree, the genetic distances between isolates assigned to the same partition was extracted from M_1, producing a single list of within-partition distances for each value of k. Starting at the list of distances obtained for $k = 2$ and moving up, each list was examined to identify the smallest value of k where at least 99.5% of distances fell below $\overline{\Delta}_{0.05}$. This value of k was considered optimal, and isolates assigned to each partition for this k-value were noted.

Assessment of framework performance by comparison to epidemiologic data

We calculated performance metrics including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy using epidemiologic links as a reference for expected clustering outcomes, as previously described (5, 15). This data set included 552 isolates linked to one of 18 epidemiologic clusters identified during US cyclosporiasis peak periods of 2018, 2019, and 2020 (Table 1). Each metric was weighted by the ratio of the number of isolates in an epidemiologic cluster to the total number possessing epidemiologic links (n = 552), so that epidemiologic clusters with more genotyped specimens provided a greater contribution to the metric. We calculated the discriminatory power of our framework using Simpson's index of diversity (D) as described elsewhere (2). The value of D was determined by:

$$D = 1 - \left(\frac{1}{N\,(N-1)} \times \sum_{J=1}^{S} n_j \left(n_j - 1\right) \right),$$

where N is the number of isolates (n = 1,137), S is the number of partitions, and n_j represents the number of isolates within the jth partition. Simpson's index assesses a method's ability to distinguish between unrelated strains sampled randomly from a given species (2), where values of D between 0.9 and 1.0 generally indicate good discriminatory power. We

Table 1. Summary of Results Obtained When Comparing Partition Memberships With Epidemiologic Data for 552 Isolates With Epidemiologic Links, United States, 2018 Through 2020[a]

Epidemiologic Cluster	Outbreak Year	Total No. of Isolates From Cluster	Mode Partition Number[b]	TP	TN	FP	FN	Sensitivity: $\frac{TP}{TP+FN}$	Specificity: $\frac{TN}{TN+FP}$	PPV: $\frac{TP}{TP+FP}$	NPV: $\frac{TN}{FN+TN}$	Accuracy: $\frac{(TP+TN)}{(TP+TN+FP+FN)}$
Distributor A, type 1	2019	13	30	13	539	0	0	100.0	100.0	100.0	100.0	100.0
Distributor A, type 17	2019	43	1	39	509	0	4	90.7	100.0	100.0	99.2	99.3
Distributor A, type 18	2019	14	29	9	537	1	5	64.3	99.8	90.0	99.1	98.9
Distributor A, type 3[c]	2019	18	7	18	530	0	0	100.0	100.0	100.0	100.0	100.0
Prepackaged salad mix 2020_001	2020	140	6	127	410	2	13	90.7	99.5	98.4	96.9	97.3
Prepackaged salad mix 2020_003	2020	79	14	75	473	0	4	94.9	100.0	100.0	99.2	99.3
Prepackaged salad 002[d]	2020	8	4	8	525	0	0	100.0	100.0	100.0	100.0	100.0
Restaurant A[d]	2019	13	4	13	525	0	0	100.0	100.0	100.0	100.0	100.0
Restaurant B	2019	13	13	11	539	0	2	84.6	100.0	100.0	99.6	99.6
Restaurant C	2019	6	5	6	546	0	0	100.0	100.0	100.0	100.0	100.0
Restaurant C (herb 2) associated cluster	2018	2	19	2	550	0	0	100.0	100.0	100.0	100.0	100.0
Restaurant D	2019	13	11	13	539	0	0	100.0	100.0	100.0	100.0	100.0
Salad chain A, 2020_025[c]	2020	4	7	4	530	0	0	100.0	100.0	100.0	100.0	100.0
Supplier X, restaurants A and B (herb 1–associated cluster)[d]	2018	6	4	6	525	0	0	100.0	100.0	100.0	100.0	100.0
Temporospatial cluster A	2018	8	46	8	544	0	0	100.0	100.0	100.0	100.0	100.0
TN/GA/VA Mexican-style restaurant/cilantro subcluster	2020	10	3	10	542	0	0	100.0	100.0	100.0	100.0	100.0
Vendor A	2018	99	17	88	453	0	11	88.9	100.0	100.0	97.6	98.0
Vendor B	2018	63	32	51	489	0	12	81.0	100.0	100.0	97.6	97.8
Adjusted metrics[e]								90.8	99.9	99.4	98.3	98.5
Averages								94.2	100.0	99.4	99.4	99.5

Abbreviations: FN, false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive.

[a] Partition memberships used to generate this table are available in Web Table 3. TP: number of specimens assigned to the correct partition number. TN: number of specimens with different epidemiologic linkage not assigned to the same partition. FP: number of specimens with different epidemiologic linkage assigned to the same partition. Note that there are some exceptions to this classification as described below. FN: number of specimens from this epidemiologic cluster not assigned to the correct partition. Values of sensitivity, specificity, PPV, NPV, and accuracy are represented as percentages.

[b] The most common partition number to which specimens with this epidemiologic linkage were assigned using a *k*-value of 46. Note that partition numbers are arbitrary.

[c] Epidemiologic clusters with a mode partition number of 7 were not considered as FP classifications for one another as these outbreaks occurred in different years and were seemingly caused by the same strain.

[d] Epidemiologic clusters with a mode partition number of 4 were not considered FP classifications for one another as the outbreaks occurred in different years and were seemingly caused by the same strain.

[e] Metrics adjusted by ratio of genotyped specimens in an epidemiologic cluster versus sum of all TP and FN isolates.

therefore considered this an indicator of whether the predicted k-values provide useful strain discrimination.

Application of other partitioning methods to the present data set

We applied R (R Foundation for Statistical Computing, Vienna, Austria) implementations of the gap statistic (clusGap function), elbow method (fviz_nbclust function), and the silhouette method (fviz_nbclust function) to M_1 to compare the results obtained using those methods to results obtained using our framework. Briefly, M_1 was clustered using Ward's method and the resulting hierarchical tree was dissected at k-values ranging from 2 to 100, where each method was applied to the resultant partition memberships to identify the optimal k-value. For the clusGap function 500 bootstrap replicates were applied.

RESULTS

Partition number selection using our framework

Matrix M_1 was calculated from the 1,137 genotypes (Web Table 2). The empirical distribution of distances in M_1 (excluding self-to-self distances) was multimodal, displaying several peaks including a major peak at $\Delta \approx 0$ (Figure 2A). Alternatively, iterations of M_2 more closely resembled a normal probability distribution without a major density peak near $\Delta \approx 0$, as observed in M_1 (Figure 2B). A value of 0.19 was computed for $\overline{\Delta}_{0.05}$ resulting in an optimal partition number of $k = 46$ (Figure 3), representing the number of partitions where at least 99.5% of all within-partition distances fell below $\overline{\Delta}_{0.05}$. At 46 partitions, 99.62% of within-partition distances fell below $\overline{\Delta}_{0.05}$ where 49,497 within-partition distances (excluding self-to-self distances) were observed, and 189 (0.38%) of these were outlying distances greater than or equal to 0.19 (Figure 4). We investigated the impact of manually selecting a partition number of 45, and observed that 99.44% of within-partition distances fell below $\overline{\Delta}_{0.05}$. At $k = 44$, 99.42% of within-partition distances fell below $\overline{\Delta}_{0.05}$. We next investigated the impact of the $\overline{\Delta}_{0.05}$ threshold on k by manually selecting numerous $\overline{\Delta}_{0.05}$ values and dissecting the clustered matrix M_1 at a partition number where at least 99.5% of within-partition distances fell below each one. We plotted the result, revealing a long-tailed distribution (Figure 5).

Performance of approach based on comparison to epidemiologic data

After partitioning the clustered matrix M_1 using a k-value of 46, the 552 isolates possessing epidemiologic links were classified as either a true positive, true negative, false positive, or false negative, as per the definitions in Table 1. Weighted metrics calculated for sensitivity, specificity, PPV, NPV, and accuracy, based on these definitions, were 90.8%, 99.9%, 99.4%, 98.3%, and 98.5%, respectively (Table 1). The discriminatory power (Simpson's index) calculated for 46 partitions was 0.92. The partition memberships used to compute these values can be found in Web Table 3.

Application of other partitioning methods to the present data set

The elbow method supported k-values ranging from 6 to 9, while the silhouette method and gap statistic supported 11 and 100 partitions, respectively. We clustered M_1 using Ward's method and dissected the tree into 9, 11, and 100 partitions for examination of the resultant partition memberships (Web Tables 4–6). For values of $k = 11$ and $k = 9$, isolates linked to unrelated outbreaks were inappropriately grouped. For $k = 100$, numerous epidemiologically linked isolates were inappropriately separated among several partitions.

DISCUSSION

Our results support the idea that the principles previously used to classify *P. falciparum* genotype pairs as belonging to the same strain or different strains (11) are applicable in the context of selecting a number of partitions in a hierarchical tree. The original framework for classifying pairs (11) and the modified framework described here each aim to establish a genetic distance threshold for identifying isolates possessing a high likelihood of belonging to the same strain. However, computation of this threshold in the context of hierarchically clustered data sets requires additional steps that control for bias and overcome obstacles associated with identifying genetically similar groups as opposed to pairs.

Establishing an appropriate background distance distribution is required for demarcation of the empirical lower 5th percentile (i.e., $\Delta_{0.05}$), and our threshold demarcation approach assumes this distribution resembles a normal distribution. Naturally, genetic distances computed from isolates causing large outbreaks will be biased toward $\Delta = 0$ and toward distances between highly represented strains. These biased outbreak data sets are unlikely to produce a normal distribution and may not be appropriate for derivation of $\Delta_{0.05}$ because they do not represent what is expected when sampling naturally occurring populations in a randomized, representative way. Indeed, 645,816 distances were computed for 1,137 genotypes, and 21,212 (3.3%) of these represent isolate pairs linked to the same outbreak that likely belong to the same strain. Most of these 21,212 distances account for the sharp peak near zero in the M_1 density distribution, reflecting biased sampling toward same-strain isolates. Generally, random sampling of naturally occurring populations would be unlikely to yield such a large number of same-strain isolates, so a distance distribution resembling the normalized M_2 distribution is more likely. Consequently, the M_2 distribution is more appropriate for computing $\Delta_{0.05}$ as it reflects what would more likely be encountered randomly in nature. In the context of *P. falciparum* therapeutic efficacy trials, sampling bias is a smaller issue as the isolates encountered more likely represent the diversity of strains circulating within the geographic vicinity of the study area (11). To overcome sampling bias here, we subsampled M_1

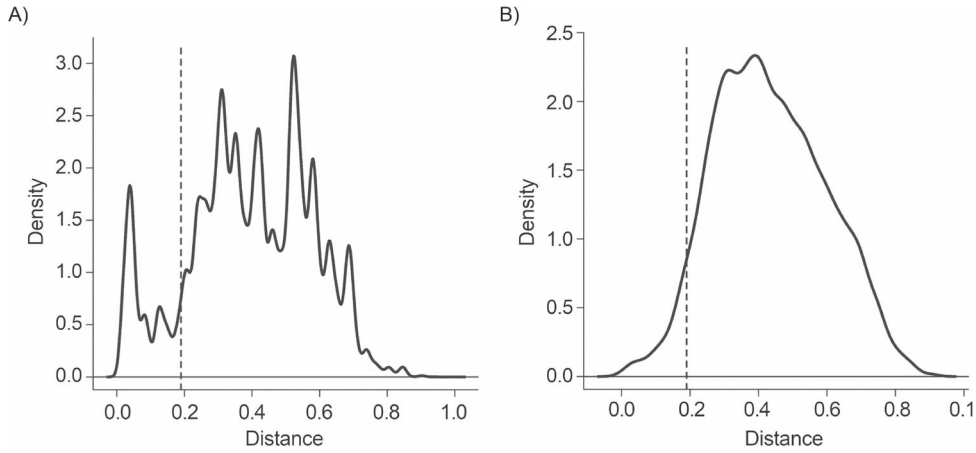

Figure 2. Density plot of the empirical distribution of distances in M_1 (A) and of the empirical distribution of distances identified for a single iteration of M_2 (B). A) The plot represents the distribution of distances observed for matrix M_1, with the threshold of $\overline{\Delta}_{0.05} = 0.19$ indicated with a dashed line. The frequency of each strain/type in the data set used to compute M_1 is biased toward those causing large outbreaks. Therefore, a large peak in density is observed near zero—at approximately $\Delta = 0.04$ in this case—reflecting the large number of distances computed for isolate pairs that belong to the same strain (i.e., bias toward $\Delta \approx 0$). As a consequence of this bias toward $\Delta \approx 0$, the distance observed at the empirical lower 5th percentile of matrix M_1 is approximately the same distance observed at the first high-density peak in the distribution. For this reason, it is inappropriate to compute a threshold distance from M_1. Instead, a threshold is computed from 1,000 iterations of the subsampled matrix M_2 where each iteration comprises a relatively unbiased distribution of background distances that are representative of the study population. Using the empirical distribution of each of 1,000 iterations of subsampled matrix M_2, a value of $\overline{\Delta}_{0.05} = 0.19$ was computed. B) The plot represents the empirical distance distribution of a single iteration of M_2 with the threshold of 0.19 indicated with a dashed line. This distribution is almost unimodal and almost symmetrical, resembling a normal probability distribution, with a markedly reduced bias toward distances approaching zero relative to M_1. Isolates separated by a distance below $\overline{\Delta}_{0.05} = 0.19$ are likely to belong to the same strain (i.e., possess recent genetic kinship) because the likelihood of randomly observing a genetic distance below this threshold for a pair of unrelated isolates is low, corresponding to a statistical test with a false positive rate (α) of 5%.

to generate 1,000 iterations of M_2 wherein each strain was effectively represented at an equal frequency. We then used M_2—possessing a distance distribution resembling a normal distribution—to derive $\Delta_{0.05}$. This eliminated some inherent sampling bias by demonstrably reducing bias toward $\Delta = 0$ at the empirical lower 5th percentile, providing a distance distribution that might have been encountered if a randomized sampling strategy had been employed.

A large genetically diverse study population will help ensure that iterations of M_2 possess a distance distribution resembling a normal distribution. This will depend on marker variation and the number of isolates sampled. Generally, an ideal data set will include multiple markers with some possessing several haplotypes. Defining a set of criteria to describe an ideal data set is difficult, although investigators should use their best judgment. Examining the distance distribution for multiple iterations of M_2 is recommended to establish whether they resemble a normal, unimodal probability distribution. This represents an indicator of whether a data set is amenable to this analysis, understanding that biological data are variable and may never yield a perfectly symmetrical distribution. The assumption that the distance distributions in iterations of M_2 resemble a normal distribution should be met, or $\overline{\Delta}_{0.05}$ may be biased toward zero. In this study, M_2 did not yield a perfectly symmetrical distribution, but it resembled a normal distribution. We acknowledge however, that this may not always hold true. Consider a genotyping data set comprising isolates from 2 related species sequenced at several loci. Some loci may possess alleles shared between the species while others may possess species-specific alleles. The species-specific alleles would divide the population, resulting in a bimodal distance distribution; the leftmost peak would represent distances between strains of the same species, while the rightmost peak would include distances between strains of different species. Our framework could still be applied to such a data set, although only distances within the leftmost peak should be used to establish a distribution for threshold demarcation.

Two apicomplexan isolates of the same strain do not necessarily possess identical genotypes due to factors including mixed-strain infections and heterozygosity. While genotypes derived from the same strain could be identical (separated by $\Delta = 0$), they are more appropriately defined as being separated by a distance approaching or equal to zero. Consequently, a threshold distance of zero is a poor choice. Indeed, when a manually selected threshold of $\overline{\Delta}_{0.05} = 0.04$ was evaluated in this study, 363 partitions resulted. Our framework selected a threshold representing a good trade-off between the high discriminatory power afforded at lower $\overline{\Delta}_{0.05}$ values versus the cost of setting $\overline{\Delta}_{0.05}$ too low, which may inappropriately separate isolates of the same strain.

Our requirement that 99.5% of within-partition distances fall below $\overline{\Delta}_{0.05}$ is noteworthy, and was introduced to

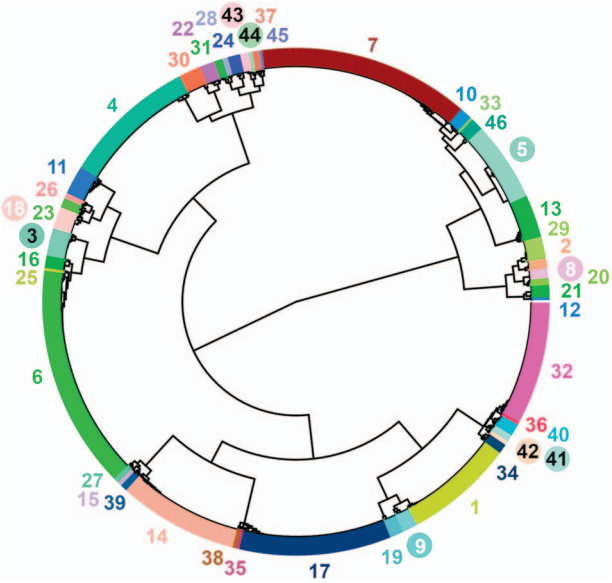

Figure 3. Hierarchical cluster dendrogram generated from 1,137 *Cyclospora* genotypes. This figure shows a hierarchical cluster dendrogram (a tree), generated from matrix M_1, that was computed from 1,137 *C. cayetanensis* genotypes. Matrix M_1 was clustered using Ward's method to produce the tree shown. The tree has been dissected into 46 partitions; the partition number where at least 99.5% of all within-partition distances fell below $\overline{\Delta}_{0.05} = 0.19$. Each of the 46 partitions is shaded in a unique color; colors differentiate the boundary between partitions and have no other meaning.

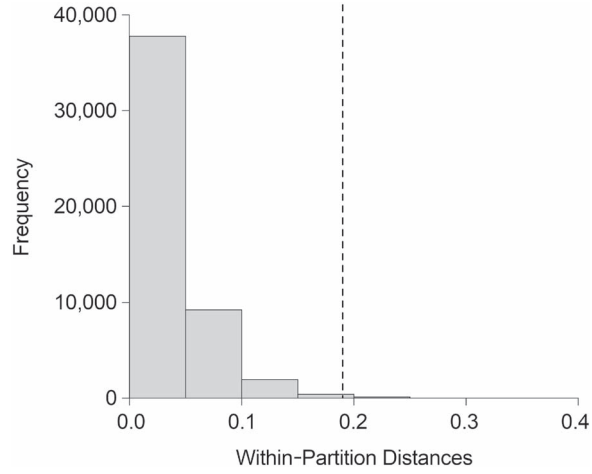

Figure 4. Frequency histogram of within-partition distances determined for matrix M_1 using a partition number (k) of 46. Frequency histogram showing distances between *Cyclospora cayetanensis* genotypes assigned to the same genetic partition at k = 46. At 46 partitions, 49,497 within-partition distances (excluding self-to-self distances) were observed. Of these, 49,308 (99.62%) fell below the $\overline{\Delta}_{0.05}$ threshold of 0.19, indicated by a dashed line, and 189 of these distances (0.38%) were greater than or equal to 0.19. The within-partition distances in this histogram are divided into 8 bins with the highest frequency observed for distances between $\Delta = 0.0$ and $\Delta = 0.05$. All within-partition distances fell below $\Delta = 0.4$. The number of distances represented by each bin, from left to right is as follows: 37,768, 9,219, 1,944, 409, 121, 25, 10, and 1.

account for outlying distances. Suppose an unusual isolate is separated from other isolates within the same partition by a genetic distance slightly above $\overline{\Delta}_{0.05}$, yet is also separated from additional isolates in this partition by distances below $\overline{\Delta}_{0.05}$. Assuming all isolates within this partition (excluding the unusual one) are separated by distances below $\overline{\Delta}_{0.05}$, the hierarchical clustering technique will have made an appropriate decision in placing the unusual isolate at a position in the tree culminating in its assignment to this partition. It may be impossible to select a $\overline{\Delta}_{0.05}$ threshold that excludes "unusual" outlying isolates while avoiding separation of appropriately grouped isolates. To accommodate such anomalies, requiring that close to 100% (e.g., 99.5%) of within-partition distances fall below $\overline{\Delta}_{0.05}$ is an appropriate trade-off. Specificity may be increased slightly by increasing this "stringency" setting to above 99.5% (e.g., to 99.7%) while keeping it less than 100%, and investigators may experiment with a setting that maximizes the k-value without affecting sensitivity too greatly.

We evaluated performance using epidemiologic data to assign labels to isolates that represent expected outcomes. Values calculated for sensitivity, specificity, PPV, NPV, and accuracy were 90.8%, 99.9%, 99.4%, 98.4%, and 98.5%, respectively. For 11 of 18 epidemiologic clusters, scores of 100% were obtained for every metric. A score below 80% for any metric was observed only for "Distributor A, type 18," where 64.3% sensitivity was observed. Simpson's Index of Diversity produced a value of $D = 0.92$, supporting the idea that 46 partitions provide good discriminatory power. The lower sensitivity observed for "Distributor A, type 18" arose because isolates possessing this linkage were assigned to 3 partitions: 9 of 14 isolates to partition 29, 1 to partition 20, and 4 to partition 2. Examining the genotype of these 14 isolates (Web Table 1) showed that some were uncharacteristically mixed, possessing 3–4 haplotypes at markers where 1 or 2 are typically observed. This suggests exposure to multiple *C. cayetanensis* strains. Notably, epidemiologic labels may be imperfect as they rely on case-patient accounts of what they ate weeks to months prior to illness, making it difficult to identify mixed exposures. Regardless, sensitivity was generally high and experience from real outbreak scenarios has shown that our framework can support epidemiologic investigations as outbreaks emerge, in addition to detecting re-emergent strains. For instance, the *C. cayetanensis* strain represented by partition 4 caused unrelated outbreaks in 3 consecutive years: Supplier X in 2018 (5), Restaurant A in 2019 (15), and prepackaged salad 002 in 2020 (16) (Table 1).

Other methods for partition number selection over- or underestimated the number of partitions, which might have confounded epidemiologic investigations had they been used in practice. For k = 100 (gap statistic) many linked isolates were separated, while k = 9 (elbow) and k = 11 (silhouette) assigned unrelated outbreak isolates to the same partition. Partition 3 defined using k = 11 (Web Table 4) contained 2 strains detected in 2020, one linked to prepackaged salads

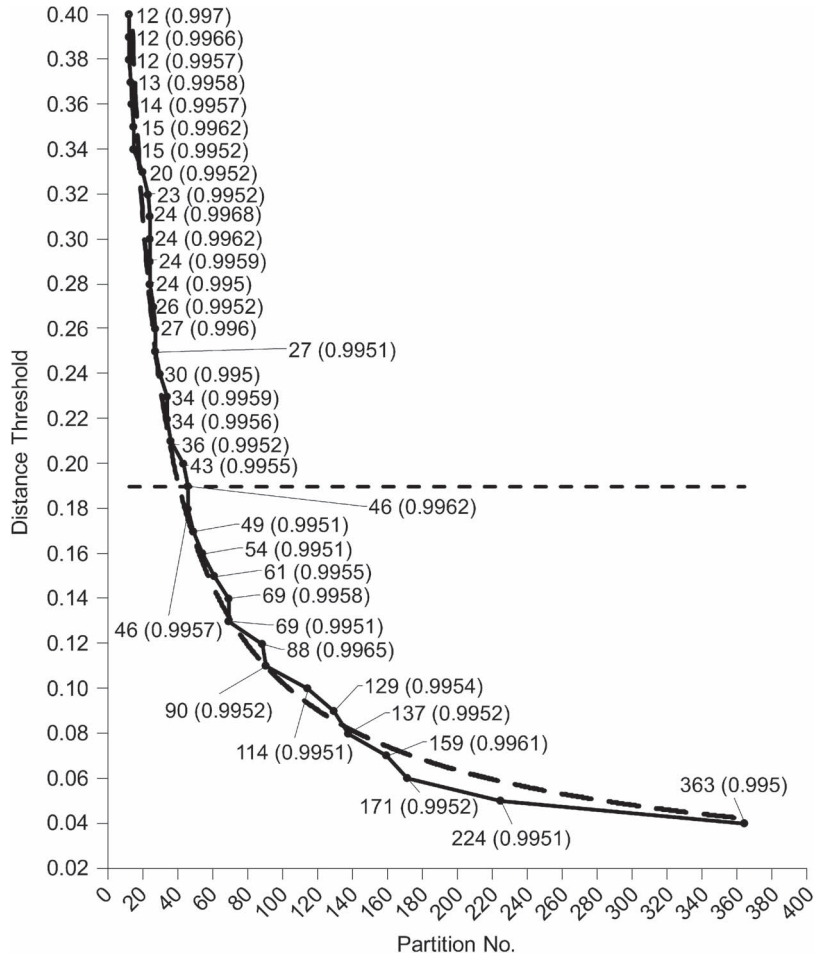

Figure 5. Impact of distance threshold on the number of partitions. This figure shows a line graph plotting a series of manually selected genetic distance threshold values (*y*-axis) against the number of partitions that result from those manually selected values (*x*-axis) and highlights how genetic distance threshold values above and below $\overline{\Delta}_{0.05} = 0.19$ (horizontal small-dashed line) affect the number of partitions predicted. Briefly, matrix M_1 was clustered using Ward's method and the resulting hierarchical tree was dissected at a partition number where at least 99.5% of within-partition distances fell below each of a range of manually selected threshold demarcation values (*y*-axis) (values from $\Delta = 0.04$ to $\Delta = 0.40$). The predicted partition number (*x*-axis) obtained for each manually selected distance threshold value was plotted, resulting in a long-tailed distribution. Each point is labeled with the partition number (the *x* coordinate) and the percentage of within-partition distances falling below the respective distance threshold for this value of *x* (in brackets). A large-dashed trendline with the equation $y = 2.4232x^{(-0.687)}$ is also shown.

and the other to cilantro served at Mexican-style restaurants. Our framework ($k = 46$) correctly assigned isolates linked to these outbreaks to different partitions. Importantly, these other statistical methods were not specifically designed for molecular epidemiologic purposes while our framework was, which likely accounts for its greater efficacy here.

The genetic distance statistic employed may also affect our frameworks' efficacy. While the framework is nonparametric and could be used in conjunction with any distance statistic, investigators should consider the advantages and disadvantages of available distance statistics before making a selection. For example, Barratt's heuristic accommodates isolates where sequence data are absent for some markers by attempting to impute missing values (5, 15). This is an advantage on one hand, as discarding specimens with data available for most markers (but not all) can be avoided. However, the imputation step becomes increasingly inaccurate as the number of missing markers increases for an isolate (17). Another limitation of Barratt's heuristic is that haplotypes are defined across a span of nucleotide bases (15, 17). Therefore, distances are not calculated by considering each variant base, meaning that Barratt's heuristic may lack resolution compared to phylogenetic substitution models. As with the choice of distance statistic, efficacy may depend on the hierarchical clustering method employed. Several hierarchical clustering methods exist, and investigators should evaluate each before making a selection.

Our framework was efficacious when applied to the present data set. However, we strongly suggest that investigators wishing to apply it to data sets involving different marker combinations, or data sets from other organisms, evaluate its performance by comparing resultant partition

memberships to expected outcomes based on externally validated metadata (i.e., epidemiologic, geographic, host data, or other metadata that support expected outcomes). Adjusting the stringency setting (e.g., reducing it to 99%) to produce partitions that better support expected outcomes (i.e., framework calibration) may be employed to maximize performance. For example, if we wished to apply our framework to genotypes of human immunodeficiency virus (HIV)—which has an extremely high mutation rate (18)—a lower stringency setting (e.g., 95% or 97%) might be appropriate as additional outliers could be expected. To calibrate this setting for HIV, genotyping of numerous isolates should be performed, including a subset collected from among close contacts such as sexual partners or people who share needles during intravenous drug use (19). The framework could then be applied to genetic distances computed for this data set, across a range of possible stringency values, where the optimal value would be the one that most frequently resulted in isolates sharing close contacts being assigned to the same partition. Our framework has the advantage that it can be calibrated for different applications, or to fit the biology of specific pathogens. Once properly calibrated, one can reasonably expect it to perform similarly when applied to data sets possessing unknown links, assuming calibration was performed on data from the same species and using the same markers.

To conclude, we describe a framework for selecting an appropriate number of partitions in an unbiased way when dissecting hierarchical trees, and we apply it to an extant *C. cayetanensis* genotyping data set. By examining the epidemiologic linkage of *C. cayetanensis* isolates assigned to each partition, we establish that our framework is highly sensitive and specific. While this evaluation was performed on a *C. cayetanensis* genotyping data set, we expect our framework will be broadly applicable to other pathogens.

ACKNOWLEDGMENTS

Author affiliations: Parasitic Diseases Branch, Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, Atlanta, Georgia, United States (Joel L. N. Barratt); Malaria Branch, Division of Parasitic Diseases and Malaria, Centers for Disease Control and Prevention, Atlanta, Georgia, United States (Mateusz M. Plucinski); and US President's Malaria Initiative, Centers for Disease Control and Prevention, Atlanta, Georgia, United States (Mateusz M. Plucinski).

The authors acknowledge the US Centers for Disease Control and Prevention, Division of Parasitic Diseases and Malaria, for support of this work. M.M.P. was funded by the US President's Malaria Initiative.

Data are available in the Web Material accompanying this publication and can be accessed under National Center for Biotechnology Information BioProject No. PRJNA578931.

We thank the Parasitic Diseases Branch and the Malaria Branch at the US Centers for Disease Control and Prevention for their support of this work.

The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the US Centers for Disease Control and Prevention.

Conflict of interest: none declared.

REFERENCES

1. Riley LW. *Molecular Epidemiology of Infectious Diseases*. 1st ed. Washington, DC: ASM Press; 2004.
2. van Belkum A, Tassios PT, Dijkshoorn L, et al. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clin Microbiol Infect.* 2007; 13(suppl 3):1–46.
3. Tolar B, Joseph LA, Schroeder MN, et al. An overview of PulseNet USA databases. *Foodborne Pathog Dis.* 2019;16(7): 457–462.
4. Barratt JLN, Park S, Nascimento FS, et al. Genotyping genetically heterogeneous *Cyclospora cayetanensis* infections to complement epidemiological case linkage. *Parasitology.* 2019;146(10):1275–1283.
5. Nascimento FS, Barratt J, Houghton K, et al. Evaluation of an ensemble-based distance statistic for clustering MLST datasets using epidemiologically defined clusters of cyclosporiasis. *Epidemiol Infect.* 2020;148:e172.
6. Centers for Disease Control and Prevention. Parasites—cyclosporiasis (*Cyclospora* infection): surveillance & outbreak response. https://www.cdc.gov/parasites/cyclosporiasis/surveillance.html. Accessed July 1, 2021.
7. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics.* 2008;24(5):719–720.
8. Zambelli AE. A data-driven approach to estimating the number of clusters in hierarchical clustering [version 1; peer review: 2 approved, 1 approved with reservations]. *F1000Res.* 2016;5(ISCB Comm J):2809.
9. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Series B Stat Methodology.* 2001;63(2):411–423.
10. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* 1987;20:53–65.
11. Plucinski MM, Barratt JLN. Nonparametric binary classification to distinguish closely related versus unrelated *P. falciparum* parasites. *Am J Trop Med Hyg.* 2021;104(5): 1830–1835.
12. Centers for Disease Control and Prevention. Domestically acquired cases of cyclosporiasis—United States, May–August 2018. https://www.cdc.gov/parasites/cyclosporiasis/outbreaks/2018/c-082318/index.html. Accessed July 1, 2021.
13. Centers for Disease Control and Prevention. Domestically acquired cases of cyclosporiasis—United States, May–August 2019. https://www.cdc.gov/parasites/cyclosporiasis/outbreaks/2019/a-050119/index.html. Accessed July 1, 2021.
14. Centers for Disease Control and Prevention. Domestically acquired cases of cyclosporiasis—United States, May–August 2020. https://www.cdc.gov/parasites/cyclosporiasis/outbreaks/2020/seasonal/index.html. Accessed July 1, 2021.
15. Barratt J, Houghton K, Richins T, et al. Investigation of US *Cyclospora cayetanensis* outbreaks in 2019 and evaluation of an improved *Cyclospora* genotyping system against 2019 cyclosporiasis outbreak clusters. *Epidemiol Infect.* 2021; 149:e214.

16. Barratt J, Ahart L, Rice M, et al. Genotyping *Cyclospora cayetanensis* from multiple outbreak clusters with an emphasis on a cluster linked to bagged salad mix—United States, 2020. *J Infect Dis.* 2021;225(12):2176–2180.
17. Barratt JLN, Sapp SGH. Machine learning-based analyses support the existence of species complexes for *Strongyloides fuelleborni* and *Strongyloides stercoralis*. *Parasitology.* 2020;147(11):1184–1195.
18. Bbosa N, Kaleebu P, Ssemwanga D. HIV subtype diversity worldwide. *Curr Opin HIV AIDS.* 2019;14(3):153–160.
19. Wilson DP, Zhang L. Characteristics of HIV epidemics driven by men who have sex with men and people who inject drugs. *Curr Opin HIV AIDS.* 2011;6(2):94–101.