**Supplementary File S1. Technical Appendix for manuscript:**

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

Joel L. N. Barratt1,2,†\*, Subin Park1,2,†, Fernanda S. Nascimento1, Jessica Hofstetter1,2, Mateusz Plucinski4, Shannon Casillas1, Richard S. Bradbury1, Michael J. Arrowood3, Yvonne Qvarnstrom1, Eldin Talundzic4

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

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

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

†Authors contributed equally

\*To whom correspondence should be addressed. Email: jbarratt@cdc.gov Alternative email: joelbarratt43@gmail.com

# Table of Contents

[Table of Contents ii](#_Toc3221401)

[Description of the Heuristic Algorithm 1](#_Toc3221402)

[Description of the Bayesian Algorithm 5](#_Toc3221403)

[Concordance of Genetic Clustering and Epidemiological Data (*P*-value calculation) 7](#_Toc3221404)

[Figure S1. Example R script for *P-*value calculation. 7](#_Toc3221405)

[Genome assemblies and SNP calling 8](#_Toc3221406)

[Table S1. Identities and QUAST statistics for each of the 11 *Cyclospora cayetanensis* genomes 9](#_Toc3221407)

# Description of the Heuristic Algorithm

A heuristic algorithm was used to calculate pairwise distances with arbitrary units that approximate the genetic relationship between a pair of specimens. Distances are calculated for each locus individually and the sum of these distances is taken as the total distance ($Δ\_{T}$) between a pair of specimens.

For nuclear (*Nu*) loci, the raw distance (Δ*Raw*) between a pair is determined by:

$$Δ\_{Nu\\_Raw}=w\left[y=0\right]+2j\left[y>0\right]+\sum\_{i=j}^{2j}-i\left[z=i\right]$$

Where *y* is the number of haplotypes that match for a given pair and square brackets indicate statements of propositional logic, where the value is 1 if true and 0 if false.

The value of *w* is determined by:

$$w=x\left[n>1\right]+4\left[n=1, x=2\right]+\left(1+x\right)\left[n=1,x>2\right]$$

Where *x* is the sum of the number of *Nu* haplotypes detected in the first and second specimens of a pair and *n* is the number of haplotypes detected in the specimen of the pair with the smallest number of haplotypes, or the number of haplotypes in each specimen if they possess the same number of haplotypes.

The variable *j* is determined by:

$$j=2[m = 1] + m[m > 1]$$

Where *m* is the number of *Nu* haplotypes detected in the specimen with the fewest number *Nu* haplotypes considering all specimens in the dataset. Consequently, the variable *j* will always have a value of 2 or larger.

The value of *z* is defined as:

$$z=3\left[\left(\begin{array}{c}2\left[n=1\right]\\+ \left[y=1, x>2\right]\end{array}\right)=3\right]+2\left(\begin{array}{c}y\left[n>1, j\geq y\right]\\+j\left[n>1, y>j\right]\\+j\left[n=1,x=2, y=1\right]\end{array}\right)$$

The final adjusted distance for locus *Nu* (Δ*Nu*) is determined by:

$$Δ\_{Nu}=H\_{Nu}\left(Δ\_{Nu\\_Raw}\left[Δ\_{Nu\\_Raw}>0\right]+P\_{Nu}\left[Δ\_{Nu\\_Raw}=0\right]\right)k$$

Where *H* is the Shannon entropy [16] of locus *Nu* and *PNu* is a frequentist probability, defined as the squared frequency of specimens in the study cohort possessing the same haplotype/s that match for the pair under consideration. For clarification, if a pair possesses two matching haplotypes, *PNu* is the squared frequency of all specimens possessing the same two haplotypes. If a pair possesses a single matching haplotype plus two additional haplotypes at this locus that do not match, *PNu* would be the squared frequency of all specimens in the study cohort possessing that single matching haplotype. The value of *k* is dependent on *y* where if *y* = 0 then *k* = 1, but if *y* > 0 then $k=P\_{Nu}$.

For extra-nuclear (*Ex*) loci, the raw distance (Δ*Raw*) between a pair is determined by:

$$Δ\_{Ex\\_Raw}=2a\left[y=0\right]$$

Where *a* is the sum of the number of *Ex* haplotypes detected in the first and second specimen and *y* is the number of matching *Ex* haplotypes detected for that pair of specimens.

The final adjusted distance for locus *Ex* (Δ*Ex*) is determined by:

$$Δ\_{Ex}=H\left(Δ\_{Ex\\_Raw}\left[Δ\_{Ex\\_Raw}>0\right]+P\_{Ex }\left[Δ\_{Ex\\_Raw}=0\right]\right)k$$

Where *H* is the Shannon entropy of locus *Ex*, and the value of *k* is dependent on *y* where if *y* = 0 then *k* = 1 but, if *y* > 0 then *k*= *PEx* (the same as for Nu loci). The value of *PEx* is a frequentist probability calculated in the same fashion described for *PNu*.

The total distance (Δ*T*) between a pair of specimens given a set of haplotypes generated for two nuclear (*Nu*) loci and one mitochondrial locus (which is an *Ex* locus) is determined by:

$$Δ\_{T}=Δ\_{Nu1}+Δ\_{Nu2}+Δ\_{Ex}$$

In cases where genotypes are incomplete due to amplification or sequencing failures, a mean distance $\left(∆\_{μ}\right)$ is calculated in place of $∆\_{Nu}$ or $∆\_{Ex}$. Depending on the circumstance, Δ*µ* is calculated in one of five ways (i –v):

1. Consider a dataset consisting of three loci (A, B and C). When a specimen with data missing for locus A (e.g. Specimen 1) is compared to a specimen with data available for locus A (e.g. Specimen 2), $∆\_{μi}$ is calculated in place of a true distance. The value of $∆\_{μi}$ is taken as the mean distance between Specimen 2 at locus A and all specimens in the cohort possessing an identical B and C genotype to Specimen 1, excluding specimens that have data missing for locus A. If a specimen (e.g. Specimen 3) only has data available for locus C and is being compared to Specimen 2, then $∆\_{μi}$ is calculated by taking the mean distance between Specimen 2 and all other specimens in the cohort that are identical to Specimen 3 at only locus C. This calculation is performed for both locus A and B in this case. However, for specimens with multiple loci missing we suggest that any links identified be considered with caution.
2. If Specimen 1 possesses a locus B and C genotype that is different to any other specimen in the cohort (i.e. Specimen 1 is unique), $∆\_{μii}$ is used. The value of $∆\_{μii} $is determined by taking the mean of all distances for locus A from the entire cohort of specimens, excluding specimens with data missing at locus A and excluding self-to-self distances. If a specimen only has data available for locus C and its sequence is unique, then $∆\_{μii}$ is taken as the average pairwise distances for all specimens (determined separately for locus A and B), excluding those missing data for locus A and/or B, and excluding self-to-self distances.
3. When comparing two specimens that both have data missing at locus A (e.g., Specimen 1 and Specimen 4), $∆\_{μiii}$ is used. The value of $∆\_{μiii}$ is determined by taking the mean of all $∆\_{μi}$ values calculated for Specimen 1 at locus A, only for specimens that are identical to Specimen 4 at locus B and C (or identical at one of these if B or C is also absent). Alternatively, the value of $∆\_{μiii}$ can be determined by taking the mean of values of $∆\_{μi}$ calculated for Specimen 4 at locus A, only for specimens that are identical to Specimen 1 at locus B and C (or identical at one of these if B or C is also absent). The value of $∆\_{μiii}$ should be the same regardless. If Specimen 1 and/or Specimen 4 possess a unique B and C genotype, then $∆\_{μii}$ is considered in the calculation of $∆\_{μiii}$.
4. When calculating a self-to-self distance for Specimen 1 (i.e., comparing Specimen 1 to itself) with locus A missing, $∆\_{μiv}$ is used. The value of $∆\_{μiv}$ is taken as the mean of all self-to-self distances for specimens with a B and C genotype identical to Specimen 1 (or identical to one of these if B or C is also absent in Specimen 1), excluding specimens with missing data for locus A.
5. If Specimen 1, which is missing data for locus A, has a B and C genotype that is unique and a self-distance is to be calculated, $∆\_{μv}$ is used. This is taken as the mean of all self-to-self distances only considering specimens that have sequence data available for locus A. If data is only available for locus C for example and its sequence is unique, then $∆\_{μv}$ is calculated in this way for locus A and B.

# Description of the Bayesian Algorithm

A Bayesian approach was used to calculate the posterior probability of a pair of specimens sharing 0, 1, or 2 common ancestors. The likelihood of observing the data assuming two isolates share 0 ancestors, signifying full independence, can be calculated directly as the product of the frequencies of the observed alleles at each locus.

$$P(D\_{i},D\_{j}|A\_{ij}=0)=\prod\_{k=1}^{N}\left[\prod\_{m=1}^{n\_{i}^{k}}f\_{k}(D\_{i,m}^{k})\prod\_{m=1}^{n\_{j}^{k}}f\_{k}(D\_{j,m}^{k})\right]$$

Where $D\_{i}$ is the observed data for isolate *i*, $A\_{ij}$ is the number of ancestors shared by isolate *i* and *j*, $D\_{i,m}^{k}$ is the *m*th observed allele at locus *k* for isolate *i*, *N* is the total number of loci, $n\_{i}^{k}$ is the number of alleles at locus *k* for isolate *i*, and $f\_{k}(x)$ is a function that returns the empiric frequency of allele *x* at locus *k.*

The probability of observing the alleles conditional on the isolates sharing one common ancestor is:

$$P(D\_{i},D\_{j}|A\_{ij}=1)=\prod\_{k=1}^{N}\left[\frac{1}{n\_{i}^{k}n\_{j}^{k}}\sum\_{x=1}^{n\_{i}^{k}}\sum\_{y=1}^{n\_{j}^{k}}1\_{\{D\_{i,x}^{k}=D\_{j,y}^{k}\}}\prod\_{m\ne x}^{n\_{i}^{k}}f\_{k}(D\_{i,m}^{k})\prod\_{m=1}^{n\_{j}^{k}}f\_{k}(D\_{j,m}^{k})\right]$$

Finally, the likelihood of observing the allele data assuming that both isolates share the exact same strain is:

$$P(D\_{i},D\_{j}|A\_{ij}=2)=\prod\_{k=1}^{N}\left[\frac{1}{n\_{i}^{(2)k}n\_{j}^{(2)k}}\sum\_{x=1}^{n\_{i}^{(2)k}}\sum\_{y=1}^{n\_{j}^{\left(2\right)k}}1\_{\{D\_{i,x}^{(2)k}=D\_{j,y}^{(2)k}\}}\prod\_{m\notin x}^{n\_{i}^{k}}f\_{k}(D\_{i,m}^{k})\prod\_{m=1}^{n\_{j}^{k}}f\_{k}(D\_{j,m}^{k})\right]$$

Where $n\_{i}^{(2)k}$ is the number of all possible pairs of alleles at locus *k* for isolate *i* and $D\_{i,x}^{(2)k}$ is the *x*th set of all possible pairs of alleles for isolate *i* at locus *k.*

The posterior probability of each hypothesis can be calculated from Bayes' formula:

$$P(A\_{ij}=x|D\_{i},D\_{j})=\frac{P(D\_{i},D\_{j}|A\_{ij}=x)P(A\\_ij=x)}{\sum\_{y=0}^{2}P(D\_{i},D\_{j}|A\_{ij}=y)}$$

An uninformative prior was chosen such that $P\left(A\_{ij}=0\right)=P\left(A\_{ij}=1\right)=P\left(A\_{ij}=2\right)=1/3$.

The pairwise distance ($Δ\_{ij}$) between any pair of specimens was defined as a function of the expected number of common ancestors, calculated from the estimated posterior probabilities:

$$Δ\_{ij}=2-\left[0×P\left(A\_{ij}=0|D\_{i},D\_{j}\right)+1×P\left(A\_{ij}=1|D\_{i},D\_{j}\right)+2×P\left(A\_{ij}=2|D\_{i},D\_{j}\right)\right]$$

# Concordance of Genetic Clustering and Epidemiological Data (*P*-value calculation)

The script below was used to calculate the probability that the concordance between the ensemble-assigned links and the epidemiologically assigned links arose due to random chance rather than a true association.

## Figure S1. Example R script for *P-*value calculation.

#ENSEMBLE P-VALUE CALCULATION

#Size of each genetic cluster – 16 clusters total

geneticclustersizes= c(1,2,1,4,1,3,2,9,9,14,13,5,4,10,6,4)

#probability that any two random specimens will be in the same genetic cluster by chance

null\_p = sum(sapply(1:length(geneticclustersizes), function (x) geneticclustersizes[x]/sum(geneticclustersizes)\*(geneticclustersizes[x]-1)/(sum(geneticclustersizes)-1)))

#sizes of each epidemiologically-defined clusters

epiclustersizes = c(4,7,2,3,3,4,2,4,2,5)

#count total number of epidemiological links

nepilinks = sum(sapply(1:length(epiclustersizes), function (x) epiclustersizes[x]\*(epiclustersizes[x]-1)/2)) ##Sum of all epi-links in each cluster (number of links = (n(n-1))/2)

#plot null distribution of expected number of epi links in genetic map

barplot(dbinom(1:40,nepilinks,null\_p),names.arg=1:40)

#observed epi-links that are concordant with the genetic links

observedepilinks = sum(1,6,3,1,1,3,3,1,6)

#p-value

1-pbinom(observedepilinks, nepilinks,null\_p)

# Genome assemblies and SNP calling

Eleven high quality *C. cayetanensis* genomes were successfully sequenced and assembled (Table 1). The NY 2001 genome assembly was of the highest quality and length. The assembled nuclear genome of *C. cayetanensis* consisted of 486 contigs (mean coverage 233.33), comprising 45,493,142 bp, with a N50 of 343,466 bp. The largest contig was 1,703,124 bp. In comparison, TXT1 2015 was of the lowest quality, with a nuclear genome assembly consisting of 2,688 contigs (mean coverage, 3.22), comprising 43,735,863 bp, with a N50 of 20,473 bp. The largest contig was 293,917 bp. The median length of all nuclear genomes was 44,520,949 bp. The complete mitochondrial genome for all isolates was assembled to a size of 6400 bp with a mean coverage of 208.24. A total of 27,109 and 14 informative SNPs were identified using kSNP within the Nu and Mt genomes of *C. cayetanensis*, respectively. Following WGS, four of the 11 sequenced specimens had insufficient DNA remaining for downstream PCR, leaving 88 DNA extracts available for genotyping.

## Table S1. Identities and QUAST statistics for each of the 11 *Cyclospora cayetanensis* genomes

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Genome Alias** | **Specimen Origin** | **GenBank BioSpecimen** | **Specimen ID/Strain** | **# Readsα** | **#contigs (≥1 kb)** | **Total length (bp)** | **Largest contig (bp)** | **GC (%)** | **N50** | **L50** |
| NY 2001 | New York, USA | SAMN7377742 | HCNY16:01 | 92,455,630 | 486 | 44,808,213 | 1,703,114 | 51.87 | 344,083 | 39 |
| TX 2014 | Texas, USA | SAMN7377743 | HCTX48:14β | 4,277,817 | 567 | 44,449,911 | 1,191,273 | 51.90 | 278,301 | 48 |
| FL 2013 | Florida, USA | SAMN7377744 | HCFL47:13 | 5,567,424 | 719 | 45,345,321 | 1,162,148 | 51.87 | 244,493 | 56 |
| VA 2015 | Virginia, USA | SAMN7377745 | HCVA02:15β | 2,319,171 | 2,725 | 32,009,723 | 400,228 | 53.27 | 11,399 | 545 |
| TXC1 2015 | Texas, USA | SAMN7377746 | HCTX204:15β | 5,705,130 | 1,732 | 43,987,511 | 307,457 | 51.93 | 63,489 | 211 |
| TXC2 2015 | Texas, USA | SAMN7377747 | HCTX205:15 | 3,952,855 | 965 | 47,105,559 | 746,765 | 51.68 | 108,591 | 134 |
| TXT1 2015 | Texas, USA | SAMN7377749  | HCTX208:15 | 4,280,160 | 2,688 | 43,735,863 | 293,917 | 51.31 | 20,473 | 624 |
| TXT2 2015 | Texas, USA | SAMN7377750  | HCTX227:15 | 9,791,641 | 626 | 42,776,694 | 1,472,454 | 51.87 | 248,541 | 53 |
| TXT3 2015 | Texas, USA | SAMN7377751  | HCTX230:15 | 3,444,066 | 2,002 | 46,800,813 | 643,039 | 51.86 | 54,290 | 235 |
| CH 2011 | China | SAMN7377752  | CHN\_HEN01 | 7,241,205 | 631 | 44,326,258 | 816,169 | 51.91 | 248,371 | 55 |
| GM 1997 | Guatemala | SAMN7377753  | HCGM11:97β | 3,079,603  | 583 | 44,425,195 | 1,019,316 | 51.90 | 252,491 | 58 |

α Number ofIllumina MiSeq 500 cycle V2 kit reads generated (250 by 250 bp paired end).

β These specimens were not subjected to genotyping due to insufficient DNA following genome sequencing or amplification failures.