Supplementary Methods Text for **­Genomic analysis of *Clostridioides difficile* in two regions of the United States reveals a diversity of strains and limited transmission**

**Determination of SNP cutoffs for potential transmission events**

To determine the number of core SNPs generated by this pipeline that indicate likely transmission between individuals (high risk pairs), we evaluated the core SNP distance between pairs across the whole set and the sharing of a common healthcare location within 12 weeks of CDI diagnosis. We found that there is a clear inverse relationship between the SNP distance between pairs and potential overlap at the same healthcare facility.

To perform this analysis we divided 750 pairs of samples that ≤ 10core SNPs into three groups. **Group 1** consists of 38 pairs of samples that have 0-1 SNPs . Of them, 11 (28.9%) were from individuals who had contact with the same healthcare facility. **Group 2** consists of 212 pairs of samples with 2- ≤ 5 SNPs. Of them, 27 (12.7%) had contact with the same facility. **Group 3** consists of 451 pairs of samples with SNP distance of > 5 but ≤ 10 . Out of them 49 (9.8%) had exposure to the same facility. A Chi-squared test showed that the observed inverse relationship between the SNP distance and the percentage of pairs demonstrating contact with a common facility is statistical significant (p=0.0015). As a reference, among 84,325 pairs of samples with > 10 different SNPs, 8,394 (less than 10%) were found to have contact with the same facility. This percentage is comparable with percentage of contact of the pairs to a common facility in Groups 2 and 3, suggesting that transmission is highly unlikely between strains with more than 1 core SNP between them.

Next, we fitted a logistic regression model between the SNP distances and the percentage of pairs demonstrating contact with a common facility using all three subgroups (those pairs of samples with SNP distances less or equal 10). Those data with more than 10 different SNPs were not included, in part to avoid highly unbalanced sample size in the model and in part because it is highly unlikely that two stains with >10 SNP differences could be biologically related. This inverse relationship was statistically significant (p=0.0016 based on likelihood ratio test) and visualized in Fig. 1.

****

Figure 1: Inverse relationship between the SNP distances and the proportions of samples collected from individuals who had contact with a common facility. Vertical bars represent the observed proportion of pairs with contact to a common facility. Samples with SNP distances less or equal 10 were used to fit a logistic regression model. The fitted proportions are visualized by the broken curve.

In addition to the above analysis, a small number (N=5) of cases had repeat specimens collected within 13-84 days that appeared to be the same strain based on identical MLST and SNP distance <10. All of these cases had 0-1 SNPs across the core genome and 2-7 SNPs distance across the whole genome. Using this information along with the common location data above we set the core SNP threshold at 0-1 SNPs.

**Principal Coordinates Analysis (PCoA)**

## Principal Coordinates Analysis (PCoA) was used to explore and visualize isolates based on the SNPs (Hamming distance) between all isolates. The loadings to the first two principal components (PCs) of each isolate were represented as the x- and y-axis coordinates, so they could be visualized on 2D figures. MClust was applied to these loadings to classify them into several data-driven clusters, where the optimal number of clusters was determined by the Bayesian Information Criterion (BIC). A Chi-square test was used to determine the associations between state, epidemiologic classification, MLST and these clusters.

## Permutation-based Tests

Due to the discrete nature of the sequencing data, Hamming distance between two samples are integers that are not normally distributed. Therefore, permutation-based hypothesis tests are more reliable than their parametric counterparts for those distances. Let $D$ be the $(n×n)$-dimensional pairwise Hamming distance matrix between all samples, and assume that there is a categorical variable $X$ with $K$-levels (e.g., the State variable with two levels, NY and MN) to be associated with $D$. Let $D\_{k}$ be the submatrix of $D$ computed from samples with $X=k$, for $k=1,2,…K$; $\overbar{D}\_{k}$, $\overbar{D}$ be the sample means of $D\_{k}$ and $D$, respectively; and $n\_{k}$ be the number of samples such that $X=k$. We define the following two test statistics to be used in permutation-based tests.

$$t\_{1}=\sqrt{\sum\_{k=1}^{K}\left(\overbar{D\_{k}}- \overbar{D}\right)^{2}} and t\_{2}=\frac{\sum\_{i=1}^{K}n\_{k}^{2}\overbar{D}\_{k}}{\sum\_{k=1}^{K}n\_{k}^{2}} . $$

By construction, $t\_{1}$ is the between-group standard deviation of mean distance and $t\_{2}$ is the grand mean within-group distance. A large value of $t\_{1}$ implies high level of heterogeneity of samples among all $K$ groups, therefore, it can be used to test the following hypothesis:

$$H\_{0}:ED\_{1}=ED\_{2}=…=ED\_{K}, versus H\_{1}:ED\_{k}\ne ED\_{k^{'}}, for some k^{'}\ne k.$$

The permutation based *p*-value is computed in this way: (a) compute $t\_{1}^{(0)}$, the $t\_{1}$-statistic computed from $D$ without permutation; (b) randomly permute samples in $D$ for 5,000 times and compute $t\_{1}^{(j)}$ from those permuted distance matrices; and (c) the p-value is calculated by

$$p=\frac{\# of t\_{1}^{\left(j\right)}> t\_{1}^{(0)}}{\# of permutaions}$$

For $t\_{2}$, the null hypothesis is that the mean of within-group distances equals to that of between-group distances; the alternative hypothesis is that the mean within-group distance is strictly smaller than that of between-group distances. Therefore, the *smaller* $t\_{2}$ is, the more evidence we have to reject the null hypothesis. The p-value is calculated by

$$\frac{\# of t\_{2}^{\left(j\right)}< t\_{2}^{(0)}}{\# of permutaions}$$

We applied these two permutation-based tests to the following data:

1. Associating sites with the Hamming distance matrix computed from the whole and core genome. The results showed that neither the mean nor the variance of the distribution of Hamming distance differed between the two sites. See Figures 1 and 2 for more details.
2. Associating epidemiological classifications with the Hamming distance matrix computed from the whole and core genome. The results showed that neither the mean nor the variance of the distribution of Hamming distance differed between CA and HCA. See Figures 3 and 4 for more details.

Of note, we also applied these two tests to study the association between the MLST and Hamming distance matrices as a sanity test. As expected, the observed between-group standard deviation ($t\_{1}^{(0)}$) was significantly larger than its permutation-based counterparts, and the observed mean within-group distance ($t\_{2}^{(0)}$) was significantly smaller than its permutation-based counterparts. Overall, we conclude that samples within each MLST type were much more closely related than two randomly selected (permuted) samples. This expected outcome confirmed that the two permutation-based tests were able to characterize the associations between pre-defined groups and the Hamming distance matrices. These results were provided in Figures 5 and 6 for references.



**Figure 1**. Between-group standard deviation of mean distance ($t\_{1}$) and grand mean within-group distance ($t\_{2}$) based on the association between **sites** and the Hamming distance matrix computed from the **whole genome**. The histograms illustrates the empirical null distribution of $t\_{1}$ and $t\_{2}$ based on 5,000 permutations. The red vertical lines represent the observed between-group standard deviation of mean distance ($t\_{1}^{(0)}=781.5$, $p=0.4142$) and grand mean within-group distance ($t\_{2}^{(0)}=10725.4$, $p=0.772$).

**

**Figure 2**. Between-group standard deviation of mean distance ($t\_{1}$) and grand mean within-group distance ($t\_{2}$) based on the association between **sites** and the Hamming distance matrix computed from the **core genome**. The histograms illustrates the empirical null distribution of $t\_{1}$ and $t\_{2}$ based on 5,000 permutations. The red vertical lines represent the observed between-group standard deviation of mean distance ($t\_{1}^{(0)}=418.7$, $p=0.293$) and grand mean within-group distance ($t\_{2}^{(0)}=5532.1$, $p=0.8316$).

**

**Figure 3**. Between-group standard deviation of mean distance ($t\_{1}$) and grand mean within-group distance ($t\_{2}$) based on the association between **epidemiological classifications** and the Hamming distance matrix computed from the **whole genome**. The histograms illustrates the empirical null distribution of $t\_{1}$ and $t\_{2}$ based on 5,000 permutations. The red vertical lines represent the observed between-group standard deviation of mean distance ($t\_{1}^{(0)}=1169.3$, $p=0.1574$) and grand mean within-group distance ($t\_{2}^{(0)}=10693.9$, $p=0.7028$).

**

**Figure 4**. Between-group standard deviation of mean distance ($t\_{1}$) and grand mean within-group distance ($t\_{2}$) based on the association between **epidemiological classifications** and the Hamming distance matrix computed from the **core genome**. The histograms illustrates the empirical null distribution of $t\_{1}$ and $t\_{2}$ based on 5,000 permutations. The red vertical lines represent the observed between-group standard deviation of mean distance ($t\_{1}^{(0)}=480.8$, $p=0.1696$) and grand mean within-group distance ($t\_{2}^{(0)}=5510.9$, $p=0.6112$).

**

**Figure 5**. Between-group standard deviation of mean distance ($t\_{1}$) and grand mean within-group distance ($t\_{2}$) based on the association between **MLST profiles** and the Hamming distance matrix computed from the **whole genome**. The histograms illustrates the empirical null distribution of $t\_{1}$ and $t\_{2}$ based on 5,000 permutations. The observed between-group standard deviation of mean distance ($t\_{1}^{(0)}=29184.8$, $p=0$) and grand mean within-group distance ($t\_{2}^{(0)}=7991.4$, $p=0$) are both out of the range therefore not shown in the above figure.

**

**Figure 6**. Between-group standard deviation of mean distance ($t\_{1}$) and grand mean within-group distance ($t\_{2}$) based on the association between **MLST profiles** and the Hamming distance matrix computed from the **core genome**. The histograms illustrates the empirical null distribution of $t\_{1}$ and $t\_{2}$ based on 5,000 permutations. The observed between-group standard deviation of mean distance ($t\_{1}^{(0)}=16420.1$, $p=0$) and grand mean within-group distance ($t\_{2}^{(0)}=4496.6$, $p=0$) are both out of the range therefore not shown in the above figure.

## Principal Coordinates Analyses and Cluster Analyses

We conducted some additional data explorations based on principal coordinates analyses (PCoA) and model-based cluster analyses. Specifically, we applied PCoA to the pairwise Hamming distance matrices for both core and whole genome, and calculated the first two principal components (PCs) in each case. Loadings to those two PCs for each sample are visualized in the following Figures.

**

**Figure 7**. Visualizing the site effects in PCoA plots. Upper panel uses the whole genome and the bottom panel uses the core genome. In both cases, x-axis and y-axis represent the first and second PCs, respectively. Each sample is represented by its corresponding loadings to those two leading PCs and colored by sites.

**

**Figure 8**. Visualizing the site effects in PCoA plots. Upper panel uses the whole genome and the bottom panel uses the core genome. In both cases, x-axis and y-axis represent the first and second PCs, respectively. Each sample is represented by its corresponding loadings to those two leading PCs and colored by epidemiological classifications.

From the above two figures, we conclude that there are no apparent clustering patterns that may be associative with either sites nor epidemiological classifications.

Next, we applied MClust to the loadings to identify data-driven clusters, and associated these clusters with sites, epidemiological classifications, and MLST profiles (Fraley & Raftery, 2003). Specifically, the Bayesian Information Criterion (BIC) was used to select the optimal number of clusters and the types of models. For analysis based on the whole genome, the optimal BIC was achieved with eight clusters and model “EEV”. For analysis based on the core genome, optimal BIC was achieved with eight clusters and model “VEV”.

Using Pearson Chi-squared test, we found that data-driven clusters were not significant associated with either sites (whole genome: p=0.848; core genome: p=0.6089) or epidemiological classifications (whole genome: p=0.7004; core genome: p=0.6637). These results are consistent with the PCoA plots.

On the other hand, the associations between data-driven clusters and MLST profiles were highly significant. The following two tables summarizes the distribution of MLST in those eight data-driven clusters. One notable finding is that many similar MLST profiles were grouped into the same data-driven clusters, suggesting that these MLST types may be merged in practice. For example, using the whole genome, Mclust classified ST2, ST3, ST6, ST8, ST14, ST42, ST46, ST53, ST58, and ST110, as one cluster (Cluster 2); and it grouped ST43 and ST54 into another cluster (Cluster 3).

**Table S1:** clusters vs MLST (**whole genome**). The overall association is highly significant ($p<0.0001$) based on Pearson Chi-square test. Highlighted are MLST types grouped by the MClust procedure.

|  |  |
| --- | --- |
| Clusters | MLST Types |
| 1 | 11 | 110 | 14 | 2 | 3 | 41 | 42 | 43 | 46 | 53 | 54 | 58 | 6 | 8 | others |
| 1 | 32 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 |
| 2 | 0 | 0 | 15 | 15 | 34 | 13 | 0 | 39 | 0 | 13 | 18 | 0 | 13 | 12 | 33 | 77 |
| 3 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 10 | 2 | 0 | 12 | 0 | 0 | 0 | 12 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| 5 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |

**Table S2:** clusters vs MLST (**core genome**). The overall association is highly significant ($p<0.0001$) based on Pearson Chi-square test. Highlighted are MLST types grouped by the MClust procedure

|  |  |
| --- | --- |
| Clusters | MLST Types |
| 1 | 11 | 110 | 14 | 2 | 3 | 41 | 42 | 43 | 46 | 53 | 54 | 58 | 6 | 8 | others |
| 1 | 32 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 |
| 2 | 0 | 0 | 16 | 15 | 34 | 5 | 0 | 39 | 10 | 0 | 18 | 0 | 13 | 12 | 33 | 51 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |
| 4 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| 6 | 0 | 1 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 15 | 0 | 12 | 0 | 0 | 0 | 29 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |

Fraley, C., & Raftery, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *Journal of Classification, 20*(2), 263-286.