PathOGiST: A Novel Method for Clustering Pathogen Isolates by Combining Multiple Genotyping Signals

Mohsen Katebi[1](✉), Pedro Feijao[1], Julius Booth[1], Mehrdad Mansouri[1],
Sean La[1], Alex Sweeten[1], Reza Miraskarshahi[1], Matthew Nguyen[1],
Johnathan Wong[1], William Hsiao[2], Cedric Chauve[1], and Leonid Chindelevitch[1]

[1] Simon Fraser University,
8888 University Ave, Burnaby, BC V5A 1S6, Canada
mkatebi@sfu.ca
[2] British Columbia Centre for Disease Control,
655 West 12th Ave, Vancouver, BC V5Z 4R4, Canada

Abstract. In this paper we study the problem of clustering bacterial isolates into epidemiologically related groups from next-generation sequencing data. Existing methods for this problem mainly use a single genotyping signal, and either use a distance-based method with a pre-specified number of clusters, or a phylogenetic tree-based method with a pre-specified threshold. We propose PathOGiST, an algorithmic framework for clustering bacterial isolates by leveraging multiple genotypic signals and calibrated thresholds. PathOGiST uses different genotypic signals, clusters the isolates based on these individual signals with correlation clustering, and combines the clusterings based on the individual signals through consensus clustering. We implemented and tested PathOGiST on three different bacterial pathogens - *Escherichia coli*, *Yersinia pseudotuberculosis*, and *Mycobacterium tuberculosis* - and we conclude by discussing further avenues to explore.

Keywords: Bacterial pathogens · Whole-genome sequencing · Correlation clustering · Microbiology · Public health

1 Introduction

Partitioning the isolates of a bacterial pathogen into epidemiologically related groups is an important challenge in public health microbiology. Specifically, such a partitioning, which we will refer to as a *clustering*, can provide information on particularly transmissible strains (super-spreaders) and identify where an intervention such as active case finding may be particularly beneficial. In combination with additional metadata, such as geography or time of observation, such a clustering can also help identify rapidly growing groups (transmission hotspots),

C. Martín-Vide et al. (Eds.): AlCoB 2020, LNBI 12099, pp. 108–124, 2020.
https://doi.org/10.1007/978-3-030-42266-0_9

narrow down the potential origins of an outbreak (index case), and distinguish between recent and historical transmissions.

The clustering problem can leverage a variety of genotypic signals. Historically, fairly coarse genotypes such as VNTR (variable-number of tandem repeats, i.e. the number of copies of a set of pre-specified repeated regions in a strain) [29], PFGE (pulsed field gel electrophoresis) [15] and MLST (multi-locus sequence type, i.e. the alleles at a small number of pre-specified housekeeping genes) [18] have been the predominant mode of genotyping bacterial pathogens. These low-resolution signals, which we refer to as "fingerprints", could lead to incorrectly clustered strains [1] since unrelated bacterial isolates may happen to share identical fingerprints. With the advent of next-generation sequencing (NGS) [17], new genotypic signals have become available. These include SNP (single-nucleotide polymorphism) profiles [6], which can be identified at the whole-genome scale, and also wgMLST (whole-genome multi-locus sequence type) [19], which contains the alleles at all of the known genes in the organism of interest.

Methodologically, existing approaches fall into one of two categories. Some methods - including those inspired by and used in metagenomics [24] - use a pure distance-based approach, whereby a sequence similarity cutoff threshold is chosen, and any pair of sequences whose similarity exceeds it are considered to be in the same cluster, with a transitive closure operator applied to ensure the result is a valid partition. Alternatively, such methods may simply apply a standard clustering method, such as hierarchical clustering, to the pairwise distance matrix; in this case, the number of clusters is typically specified in advance [5]. Other methods - which tend to be more computationally expensive - leverage a phylogenetic tree reconstructed from the data to define clusters [2,11]. They also typically require a similarity threshold, but may be less sensitive to outlier isolates or to homoplasy, i.e. convergent evolution.

The majority of existing approaches for clustering bacterial isolates use a single genotypic signal, typically one of the higher-resolution ones, in isolation [12]. However, in this paper we argue for the principled combination of both low-resolution as well as high-resolution genotypic signals. The framework we propose here, called PathOGiST, innovates in several key ways. First, it leverages multiple genotypic signals extracted from NGS data. They can be further subdivided according to granularity into coarse and fine signals; the former get penalized only for grouping together isolates with different genotypes, not for splitting isolates with similar genotypes, while the latter get penalized for both of these. Second, it is based on a distance threshold, but does not apply a transitive closure operator to the similarity graph, or require a pre-specified number of clusters. Instead, it makes use of the *correlation clustering* paradigm, which tries to minimize the number of pairs of distant isolates within clusters while minimizing the number of pairs of close isolates between clusters. Third, it can be calibrated to different bacterial pathogens and genotyping signals, although we also provide an automatic threshold detector based on the distribution of pairwise distances between isolates.

Our results demonstrate that, when applied to a selection of three bacterial pathogens with annotated datasets publicly available - *Escherichia coli*, *Yersinia pseudo-tuberculosis*, and *Mycobacterium tuberculosis* - PathOGiST performs with a higher accuracy than recently published existing methods in most cases, both in terms of its ARI (adjusted Rand index) as well as CP (cluster purity). Our paper establishes that the use of calibrated thresholds and multiple genotypic signals can lead to an accurate clustering of bacterial isolates for public health epidemiology.

2 Methods

The goal of our approach is to cluster pathogen isolates from whole-genome sequencing data by using different genotyping approaches, alone and in combination. Each cluster should ideally represent a set of isolates related by an epidemiological transmission chain. We assume that we are given as input several matrices recording the pairwise distances between the isolates, one per genotyping signal. The algorithm proceeds in two stages. We first compute a clustering of the isolates for each distance matrix, and then compute a consensus of these separate clusterings. For the first step, we rely on *correlation clustering* [3], which we describe in Sect. 2.1. For the second step, we use a modified approach to the consensus clustering problem [4], also based on a correlation clustering formulation, described in Sect. 2.2. The whole process is illustrated in Fig. 1.

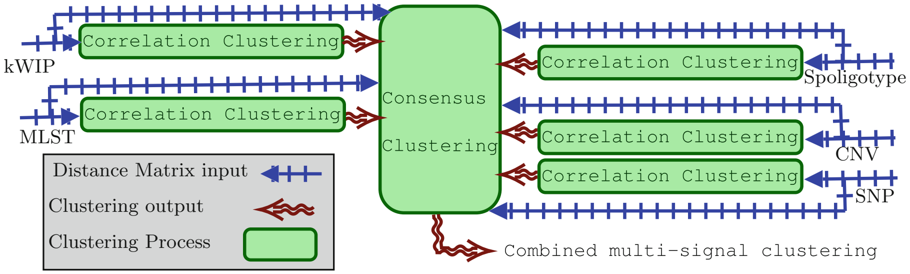

Fig. 1. PathOGiST starts by computing clusters based on single distance signals using correlation clustering. Then we run consensus clustering on the outputs of the correlation clustering.

2.1 Correlation Clustering

Let G be an undirected complete weighted graph with vertices V and edges E. Let $W : E \rightarrow \mathbb{R}$ be the edge-weighting function, which is positive for edges connecting vertices (representing isolates) that are similar and negative for those connecting dissimilar vertices. Correlation clustering aims to partition the vertices into disjoint clusters $C_1, C_2, \ldots, C_N$ where $N \leq n$. Let I be the set of edges whose endpoints lie in the same cluster and let $J = E - I$ be the set of edges whose endpoints lie in different clusters. The goal of the *minimum correlation clustering problem* is to find a clustering that minimizes the total weight of the edges in I with negative weight minus the total weight of the edges in J with positive weight:

$$\underset{C_1, C_2, \ldots, C_N}{\arg\min} \sum_{\substack{e \in I \\ W(e) < 0}} W(e) - \sum_{\substack{e \in J \\ W(e) > 0}} W(e)$$

In this work, we perform the construction of the weighted graph G from a distance matrix. Given a distance matrix D on the input elements (graph vertices) such that d_{ij} is the distance between elements i and j, we define $s_{ij} = T - d_{ij}$, where T is a *distance threshold*, intuitively meaning that if $d_{ij} < T$, i and j are considered similar, while $d_{ij} > T$ means that i and j are considered dissimilar. We use s_{ij} as the weight of the edge between vertices i and j in G.

By defining binary variables x_{ij} such that $x_{ij} = 0$ if i and j are in the same cluster and $x_{ij} = 1$ otherwise, we can write the minimum correlation clustering objective function as

$$f(x) = \sum_{s_{ij} > 0} s_{ij} x_{ij} - \sum_{s_{ij} < 0} s_{ij}(1 - x_{ij}) = \sum s_{ij} x_{ij} - \sum_{s_{ij} < 0} s_{ij}.$$

Since the second term is constant, the minimum correlation clustering problem can be solved optimally with the following Integer Linear Program (ILP):

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \sum s_{ij} x_{ij} \\ \text{s.t.} \quad & x_{ik} \leq x_{ij} + x_{jk} \quad \text{for all } i, j, k \\ & x_{ij} \in \{0, 1\} \quad \text{for all } i, j \end{aligned} \tag{1}$$

Here, the inequality constraints (which we call the "triangle inequality" constraints) together with the binary constraints ensure that the assignment is transitive [3]. Indeed, if $x_{ij} = 0$ and $x_{jk} = 0$, they enforce that $x_{ik} = 0$.

C4: A Fast Parallel Heuristic for the Correlation Clustering Problem. Solving the ILP in Eq. (1) can be time consuming and can require a large amount of memory due to the quadratic number of variables and cubic number of constraints. For this reason, we additionally implemented the faster C4 algorithm, a parallel algorithm that guarantees a 3-approximation ratio of the optimal objective function of correlation clustering in the special case of metric distances (i.e. when the s_{ij} satisfy the triangle inequality $s_{ik} \leq s_{ij} + s_{jk}$) [26].

Our results show that this algorithm is remarkably fast and quite accurate on the input graphs we tested. However, it is non-deterministic, as it depends on the initial permutation of the vertices. With this in mind, we run C4 multiple times with random initial permutations and compute the objective value for each solution. Then, among those solutions, we choose the one that minimizes the objective function. Our experiments show that in practice, this works very well and most of the time is able to find the optimum or near-optimum solution.

Solving the Minimum Correlation Clustering Problem Exactly. In order to solve the minimum correlation clustering problem exactly, while coping with the cubic number of linear constraints, we employed two approaches.

First, recalling that we often get a near-optimum solution from C4, we use it as a warm start to the problem by supplying it to the ILP solver.

Second, rather than creating the ILP with all the constraints right away, we iteratively add the constraints as follows. According to Eq. (1), for every index triple (i, j, k) the ILP has 3 constraints on the decision variables x_{ij}, x_{jk}, x_{ik}:

$$x_{ik} \leq x_{ij} + x_{jk}, \;\; x_{ij} \leq x_{ik} + x_{jk}, \;\; x_{jk} \leq x_{ik} + x_{ij}.$$

To provide the intuition for our second heuristic, assume that all three similarities between elements i, j, k are positive. This implies that the elements i, j, and k are similar to each other, so are more likely to belong to the same cluster. In this case, the three variables x_{ik}, x_{ij}, and x_{jk} will likely be assigned the value 0 and satisfy the inequalities. On the other hand, all three similarities being negative implies that elements i, j, and k are likely to be in different clusters, which would set these variables to 1 and again satisfy the inequalities.

Taking this into account, we use an approach inspired by constraint generation [8], and start by only including constraints induced by element triples whose set of similarities contain both positive and negative edges and solve this trimmed-down ILP. We then check all the excluded constraints in the solution to see whether any of them is violated. If none is violated, then the current solution is also an optimum solution for the original ILP and we are done. Otherwise, we add all the constraints that are not satisfied by the current solution to the ILP and solve the modified ILP again. We repeat this process until no violated constraint remains.

In most experiments (225 out of 235 experiments), we observe that no violated constraints have been found. Almost all of the other cases only required one extra iteration to find a solution that satisfied all the constraints. The average number of iterations was 1.102.

2.2 Consensus Clustering

Given a set of clusterings and a measure of distance between clusterings, the *consensus clustering problem* aims to find a clustering minimizing the total distance to all input clusterings. A simple distance between two clusterings π_1 and π_2 is the number of elements clustered differently in π_1 and π_2, that is, the number of pairs of elements co-clustered in π_1 but not co-clustered in π_2, or vice versa.

Representing a clustering x by a quadratic number of binary variables ($x_{ij} = 0$ if and only if i, j are co-clustered), the distance between x and a clustering π is given by the formula

$$d(x, \pi) = \sum_{\pi_{ij}=1} w_{ij}(1 - x_{ij}) + \sum_{\pi_{ij}=0} w_{ij} x_{ij} = \sum_{i,j} s_{ij} x_{ij} + \sum_{x_{ij}=1} w_{ij}, \qquad (2)$$

where a weight w_{ij} is assigned to each pair of elements i and j penalizing any clustering decision in x that differs from π, and we define $s_{ij} := (-1)^{\pi_{ij}} w_{ij}$.

Notice that solving the minimum consensus problem for a given set of clusterings $\pi^{(1)}, \dots, \pi^{(n)}$ is equivalent to solving a minimum correlation clustering problem with the matrix S defined as

$$s_{ij} = \sum_{\left\{k \mid \pi_{ij}^{(k)}=0\right\}} w_{ij}^{(k)} - \sum_{\left\{k \mid \pi_{ij}^{(k)}=1\right\}} w_{ij}^{(k)} = \sum_{k=1}^{n} (-1)^{\pi_{ij}^{(k)}} w_{ij}^{(k)} \tag{3}$$

Consensus Clustering with Different Granularities. An important feature of our problem is that the different genotyping signals we consider might not cluster the isolates with the same granularity. For example, it was shown in [22] that when clustering *Mycobacterium tuberculosis* isolates using SNPs, MLST, CNVs and spolygotypes, the latter two genotyping signals result in coarser clusters than the former two. For this reason, we assume that the input clusterings can be of different granularities. In this setting, we want to avoid penalizing the differences between a finer clustering π and a coarser clustering π', and we introduce the following asymmetric distance: $d(\pi, \pi') = |\pi - \pi'|$. In this case, assuming π is the coarser clustering and π' the finer one, we penalize only those pairs that are co-clustered in π but not in π'.

Then, given the clusterings $\pi^{(1)}, \dots, \pi^{(m)}$ and a subset F of these clusterings, representing the clusterings with the finer resolution, the *finest consensus clustering* problem is to find a clustering x that minimizes the total distance between x and all input clusterings, where

$$d(x, \pi) = \begin{cases} \sum\limits_{\pi_{ij}=1} w_{ij}(1 - x_{ij}) + \sum\limits_{\pi_{ij}=0} w_{ij} x_{ij}, & \text{if } \pi \in F \\ \sum\limits_{\pi_{ij}=0} w_{ij} x_{ij}, & \text{otherwise} \end{cases} \tag{4}$$

We can then reformulate this problem as a minimum correlation clustering again, with matrix S defined by

$$s_{ij} = \sum_{\left\{k \mid \pi_{ij}^{(k)}=0\right\}} w_{ij}^{(k)} - \sum_{\left\{k \mid \pi_{ij}^{(k)}=1, \pi^{(k)} \in F\right\}} w_{ij}^{(k)} \tag{5}$$

Selecting Appropriate Weights for Consensus Clustering. There might be many meaningful ways of defining the weights $w_{ij}^{(k)}$ used in the previous equations. If we assume that a clustering π was inferred based on a distance matrix D, normalized such that $0 \le d_{ij} \le 1$, we can define w_{ij} as

$$w_{ij} = \begin{cases} d_{ij}, & \text{if } \pi_{ij} = 1 \\ 1 - d_{ij}, & \text{otherwise} \end{cases} \tag{6}$$

The reasoning behind this definition is that if $\pi_{ij} = 1$ (i, j are not co-clustered in π), then the distance d_{ij} should be large, therefore it is a good penalty for co-clustering i, j in x. On the other hand, if $\pi_{ij} = 0$, d_{ij} can be expected to be small, which means that $1 - d_{ij}$ is a better candidate for the penalty of choosing $x_{ij} = 1$. The distance between two clusterings (Eq. (2)) can then be written as

$$d(x, \pi) = \sum_{\pi_{ij}=1} d_{ij}(1 - x_{ij}) + \sum_{\pi_{ij}=0} (1 - d_{ij})x_{ij} \tag{7}$$

and Eq. (3) becomes

$$s_{ij} = \sum_{\{k|\pi_{ij}^{(k)}=0\}} \left(1 - d_{ij}^{(k)}\right) - \sum_{\{k|\pi_{ij}^{(k)}=1\}} d_{ij}^{(k)} = \Pi_{ij} - D_{ij} \tag{8}$$

where $\Pi_{ij} = \left|\left\{k|\pi_{ij}^{(k)} = 0\right\}\right|$ and $D = \sum_{k=1}^{n} d_{ij}^{(k)}$.

We can naturally combine the weighting with the different granularities within a single formulation. In summary, the finest consensus clustering problem with weights can be formulated as a minimum correlation clustering problem, and thus solved by the algorithms described in Sect. 2.1.

2.3 Evaluation

To evaluate our methods for clustering, we compute two measures between our clustering and a ground truth clustering: Adjusted Rand Index (ARI) and Cluster Purity (CP).

The adjusted Rand index is a measure that computes how similar the clusters are to the ground truth. It is the corrected-for-chance version of the Rand index which is the percentage of correctly clustered elements. It can be computed using the following formula:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where n_{ij}, a_i, b_j are values, row sums, and column sums from the contingency table [13].

Cluster Purity is another measure of similarity between two data clusterings. To compute, assign each cluster to the most common ground truth cluster in it. Then, count the number of correctly assigned data points and divide by the total number of data points. Formally:

$$CP(C, G) = \frac{1}{N} \sum_k \max_j |c_k \cap g_j|$$

where N is the number of data points, $C = \{c_1, c_2, \ldots, c_K\}$ is the set of clusters and $G = \{g_1, g_2, \ldots, g_J\}$ is the set of ground truth clusters [20].

3 Results

3.1 Datasets and Genotyping Methods

We used three published datasets for three pathogens, *Escherichia coli* [14], *Mycobacterium tuberculosis* [10], *Yersinia pseudotuberculosis* [30], and a simulated dataset taken from [16]. Several genotyping signals were extracted from the WGS data: multilocus sequence typing (MLST) using MentaLiST pipeline [7], single nucleotide polymorphisms (SNP) using Snippy [28], copy number variants (CNV) using Prince [21], k-mer weighted inner products (kWIP) using kWIP [23], and spacer oligonucleotide typing (Spoligotyping) using SpoTyping [31] (Table 1).

Table 1. Datasets and genotyping summary

Dataset	Number of isolates	Genotyping signals				
		SNP	MLST	kWIP	CNV	SpoTyping
E. coli	1509	✓	✓	✓	✗	✗
M. tuberculosis (MTB)	1377	✓	✓	✓	✓	✓
Y. pseudotuberculosis (Yp)	163	✓	✓	✓	✗	✗
Simulated Data (SD)	96	✓	✓	✓	✗	✗

For each genotyping signal, in order to apply our correlation clustering algorithm, we needed to determine a threshold T to decide which pairs of isolates should be considered similar. To do so, we consider the pairwise distance distribution for each signal, choosing a threshold range that covers the first valley in the distribution, under the assumption that the first peak likely indicates distances between isolates belonging to the same cluster. The resulting threshold ranges and steps are described in Appendix (Table 4).

In our experiments, for each sample, each signal and each threshold, we ran our two algorithms for solving the minimum correlation clustering problem, the C4 approximation algorithm (with multiple runs) and the exact ILP using delayed constraint generation. Then we ran the consensus clustering algorithm, again using both methods.

3.2 Single Signal Genotyping

The *E. coli* dataset contains 1509 isolates collected from across England and spans an 11-year period. The distance distribution for each genotyping signal (MLST, SNP, kWIP) is shown in the Appendix (Fig. 2).

The second dataset contains 163 isolates of *Y. pseudotuberculosis* mostly collected from New Zealand [30]. We applied the same genotyping methods as for the *E. coli* dataset to this one. The results are presented in Appendix (Fig. 3). For *E. coli* and *Y. pseudotuberculosis*, we consider the MLST groups determined in their respective studies [14,30] as the ground truth, and use them to calculate the ARI and CP values. For the simulated data, since the authors suggest [16] BAPS for building the true phylogenetic tree and clustering, we used the RhierBAPS results as the ground truth for this dataset instead of MLST.

The isolates of the *M. tuberculosis* dataset were obtained from pediatric patients in British Columbia, Canada, and were collected between 2005 and 2014. We used a subset of 1377 isolates, all of which underwent WGS. In addition to SNP, MLST and kWIP information, we considered two additional genotyping signals, CNV and Spolygotyping. For *M. tuberculosis*, due to the lack of MLST groups, we use the strain's lineage, a proxy for its geographic origin [27]. For this dataset, the ARI would not be informative since a lineage is a coarse grouping largely uninformative of the underlying epidemiology, and should be split into multiple clusters. Thus, we only calculate and report the CP in this case. The results for this dataset are illustrated in Appendix (Fig. 4).

We can observe that for all the pathogens and genotyping signals we consider, there is a relatively clear threshold that falls within the chosen range which results in high accuracy clusters, with ARI and CP values above 0.8, and often close to 1. The only exceptions concern the *M. tuberculosis* dataset with SNPs, MLST and kWIP, although the CP statistics is notoriously less robust than the ARI. Moreover, most of the time, around the best thresholds, the clustering obtained with the exact ILP method results in accuracy measures that are either very close to those of the C4 method or slightly better.

3.3 Comparison of the C4 and Exact ILP Methods

The ILP generally gives more accurate results and is a deterministic method. However, its running time and memory usage depend a lot on the size of the dataset. For example, while it is able to cluster the smaller *Y. pseudotuberculosis* dataset in less than a minute, it takes more than three hours to find clusters for larger datasets at some threshold values. On the other hand, the C4 heusritic is significantly faster and requires much less memory even on the larger datasets, as shown in Table 2. However, it is not deterministic, and random restarts may give slightly different, incompatible results. To evaluate the C4 heuristic performance, we compared the objective values of solutions found by C4 and by the exact ILP. In most cases, C4 performs very well and finds a solution whose objective value is close to the optimal (Table 2).

Furthermore, the objective value of CPLEX is affected by the tolerance parameter; when the gap between the lower and upper bound is less than a certain fraction ϵ, set to 10^{-6} by default, the optimization is stopped. In this case, we see that because the magnitude of the objective function is fairly large, it is possible for the C4 method to obtain a better objective function than CPLEX.

Table 2. Average running time (in seconds) and memory footprint (in gigabytes); ILP and C4 objective value comparison.

Dataset	Time (s)		Memory (GB)		Objective value		
	C4	ILP	C4	ILP	ILP	C4: mean	C4: std
E. coli	698	7282	0.22	193.72	-1.9068×10^{10}	-1.9074×10^{10}	3.8817×10^{5}
M. tuberculosis	572	10437	0.20	298.87	-2.9445×10^{8}	-2.8844×10^{8}	2.4238×10^{3}
Y. pseudotuberculosis	14	15	0.13	0.81	-4.1601×10^{7}	-4.1594×10^{7}	1.3238×10^{3}

3.4 Comparison with Existing Clustering Methods

The results from PathOGiST were compared to those generated by two recent methods developed for clustering WGS datasets, Phydelity [11] and TreeCluster [2]; both of them are based on phylogenetic trees. To infer a phylogeny for our datasets, we first calculated a pair-wise distance matrix using Mash [25], then we ran the popular and widely used BIONJ [9] variant of the neighbor joining algorithm on the distance matrix. After we inferred phylogenetic trees, we ran Phydelity and TreeCluster with their default settings. In order to pick a single threshold for each genotyping signal-pathogen combination in PathOGiST, we chose the threshold resulting in the best ARI (CP for *M. tuberculosis*) among all the options. These thresholds are set as the default thresholds for these genotyping signal-pathogen combinations, but can be overridden by the user. Table 5 (Appendix) shows the chosen optimal threshold for each dataset and genotyping signal.

Table 3. ARI (Adjusted Rand Index) and CP (Cluster Purity) computed for different methods and genotyping signals

Method	*E. coli*		*Y. pseudotuberculosis*		*M. tuberculosis*		*Simulated Data*	
	ARI	CP	ARI	CP	ARI	CP	ARI	CP
Phydelity	0.76	0.93	0.23	0.94	-	0.92	0.238	0.645
TreeCluster	0.08	0.96	0.01	0.90	-	0.74	0.940	0.562
PathOGiST								
ILP: SNP	**0.92**	**1.0**	**0.96**	**0.98**	-	0.56	0.970	0.979
ILP: MLST	0.90	0.95	0.94	0.94	-	0.95	0.969	0.968
ILP: kWIP	0.90	**1.0**	**0.96**	0.94	-	0.57	**0.973**	**0.989**
ILP: CNV	-	-	-	-	-	**1.0**	-	-
ILP: SpoTyping	-	-	-	-	-	0.92	-	-
ILP: Consensus	0.91	0.85	**0.96**	0.97	-	0.57	**0.973**	**0.989**
C4: SNP	**0.92**	**1.0**	**0.96**	**0.98**	-	0.57	**0.973**	**0.989**
C4: MLST	0.90	0.95	0.94	0.94	-	0.95	0.969	0.968
C4: kWIP	0.90	0.99	**0.96**	0.94	-	0.60	**0.973**	**0.989**
C4: CNV	-	-	-	-	-	**1.0**	-	-
C4: SpoTyping	-	-	-	-	-	0.92	-	-
C4: Consensus	0.91	0.86	**0.96**	0.97	-	0.47	**0.973**	**0.989**

Having clustering outputs of the single signal correlation clustering algorithm with chosen default thresholds, we ran consensus clustering for each pathogen with all their available genotyping signals. We considered SNP clustering as the finest because it provides a higher resolution signal comparing to other genotyping signals. The results are described in Table 3. The main observation is that in all cases, but *M. tuberculosis*, the consensus clustering ARI is close to the best ARI obtained by a single genotyping signal, showing that our approach indeed removes the need to chose a single signal for clustering.

4 Conclusion

In this paper we described PathOGiST, an algorithmic framework for clustering bacterial isolates. One of our key contributions is to introduce the paradigms of correlation clustering and consensus clustering for the analysis of bacterial pathogens, together with two implementations - one exact and one heuristic - of correlation clustering algorithms, tailored to the problem at hand. Our experimental results suggest that our approach allows to compute a very accurate, often close to optimal, clustering without having to determine an optimal genotyping signal.

In the future, we hope to address several challenges. The first issue is the risk of overfitting, as the calibration of the threshold relies on the correlation clustering results' comparison to a gold standard. However, we also provide an automatic threshold detector. Our results demonstrate that our approach has the potential to provide reliable clusters.

Second, instead of a single output, a multi-scale or hierarchical representation of the clusters may be helpful in order to provide the user with the flexibility of deciding on their own clustering granularity. Moreover, some metadata, such as collection time or geographic location, may be fruitfully incorporated into the clustering approach in order to better inform the resolution of some groups of isolates.

Finally, due to the lack of existing tools for simulating multiple genotyping signals, we considered MLST as our gold standard. This may not represent the correct clustering, but is the best available among the individual genotyping signals.

Despite these challenges, we believe that PathOGiST is a first step in the right direction, and we hope that it will generate an impetus to further explore the problem of clustering bacterial isolates.

A Appendix Tables

Table 4. Ranges and steps for threshold values. For each experiment we ran the PathOGiST with different thresholds iteratively. We increased threshold by the step starting from beginning of the range through its end.

Dataset	SNP		MLST		kWIP		CNV		SpoTyping	
	Range	Step	Range	Step	Range	Step	Range	Step	Range	Step
E. coli	$(0, 43000]$	2150	$(0, 600]$	20	$[0.21, 0.75]$	0.03	-	-	-	-
MTB	$(0, 500]$	25	$(0, 500]$	25	$[0.125, 0.5]$	0.025	$(0, 50]$	2.5	$(0, 13]$	0.65
Yp	$(0, 40000]$	2000	$(0, 600]$	20	$[0.175, 0.7]$	0.025	-	-	-	-
SD	$(0, 8000]$	400	$(0, 400]$	20	$[0.26, 0.4]$	0.02	-	-	-	-

Table 5. Best clustering thresholds per dataset and genotyping signal.

Dataset	SNP	MLST	kWIP	CNV	SpoTyping
E. coli	17200	400	0.66	-	-
M. tuberculosis	500	475	0.5	50	13
Y. pseudotuberculosis	6000	340	0.625	-	-
Simulated Data	1600	220	0.4	-	-

B Appendix Figures

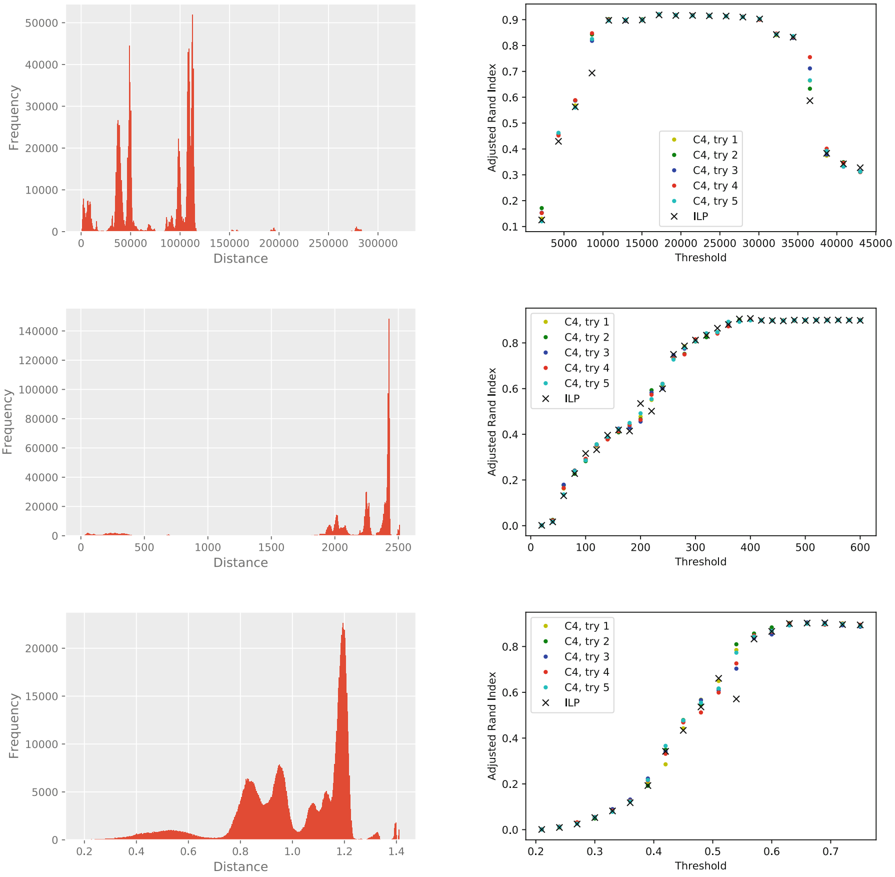

Fig. 2. Distance histograms and ARI results for *E. coli.* From top to bottom: SNP, MLST, kWIP.

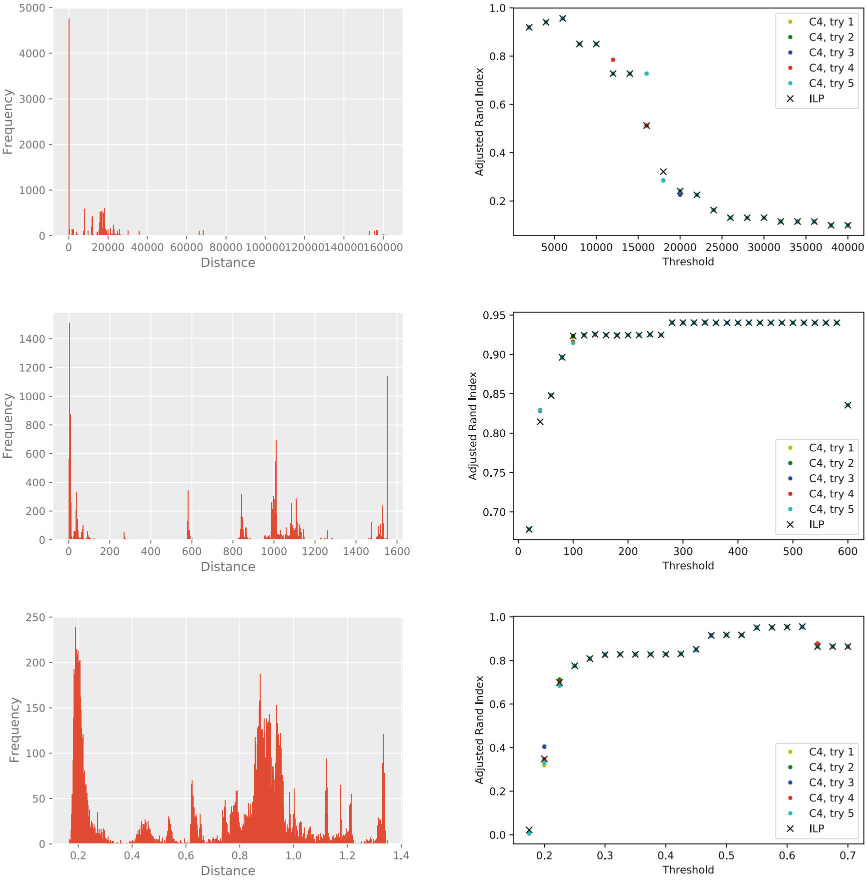

Fig. 3. Distance histograms and ARI results for the *Y. pseudotuberculosis* dataset. From top to bottom: SNP, MLST, kWIP.

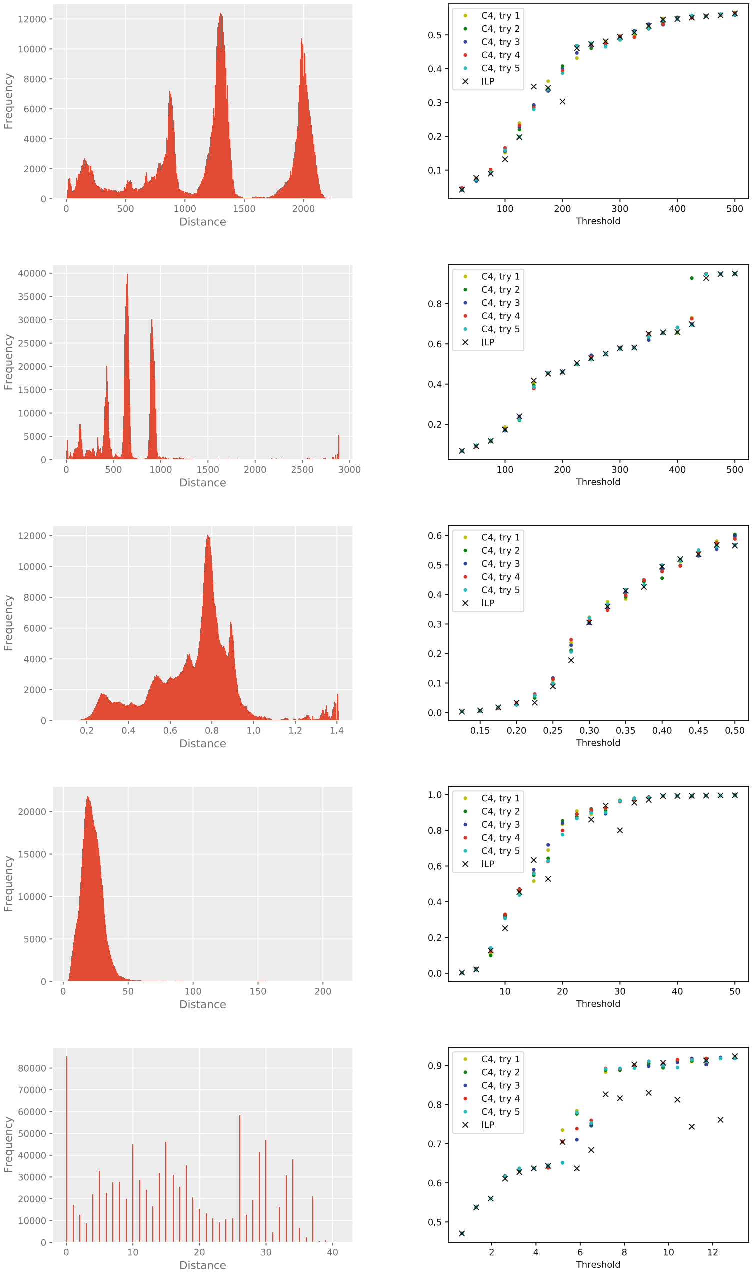

Fig. 4. Distance histograms and CP results for *M. tuberculosis*. From top to bottom: SNP, MLST, kWIP, CNV, Spolygotyping.

References

1. Alaridah, N., Hallbäck, E.T., Tångrot, J., et al.: Transmission dynamics study of tuberculosis isolates with whole genome sequencing in southern Sweden. Sci. Rep. **9**(1), 4931 (2019)
2. Balaban, M., Moshiri, N., Mai, U., et al.: TreeCluster: clustering biological sequences using phylogenetic trees. bioRxiv (2019). https://doi.org/10.1101/591388
3. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. Mach. Learn. **56**, 89–113 (2004)
4. Bonizzoni, P., Vedova, G.D., Dondi, R., Jiang, T.: On the approximation of correlation clustering and consensus clustering. J. Comput. Syst. Sci. **74**, 671–696 (2008)
5. Cheng, L., Connor, T.R., Sirén, J., et al.: Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol. Biol. Evol. **30**, 1224–1228 (2013)
6. Faison, W.J., et al.: Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human genomes. Genomics **104**(1), 1–7 (2014)
7. Feijao, P., Yao, H.T., Fornika, D., et al.: MentaLiST-a fast MLST caller for large MLST schemes. Microb. Genom. **4** (2018)
8. Dantzig, G., Fulkerson, R., Johnson, S.: Solution of a large-scale traveling salesman problem. Oper. Res. **2**, 393–410 (1954)
9. Gascuel, O.: BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**(7), 685–695 (1997)
10. Guthrie, J.L., Delli Pizzi, A., Roth, D., et al.: Genotyping and whole-genome sequencing to identify tuberculosis transmission to pediatric patients in British Columbia, Canada, 2005–2014. J. Infect. Dis. **40**, 1–9 (2018)
11. Han, A.X., Parker, E., Maurer-Stroh, S., et al.: Inferring putative transmission clusters with Phydelity. bioRxiv (2019). https://doi.org/10.1101/477653
12. Hanage, W.P., Fraser, C., Spratt, B.G.: Sequences, sequence clusters and bacterial species. Philos. Trans. R. Soc. B: Biol. Sci. **361**(1475), 1917–1927 (2006)
13. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)
14. Kallonen, T., Brodrick, H.J., Harris, S.R., et al.: Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. Genome Res. **27**, 1437–1449 (2017)
15. Kaufmann, M.E.: Pulsed-field gel electrophoresis. In: Woodford, N., Johnson, A.P. (eds.) Molecular Bacteriology, pp. 33–50. Springer, Heidelberg (1998). https://doi.org/10.1385/0-89603-498-4:33
16. Lees, J.A., Kendall, M., Parkhill, J., Colijn, C., Bentley, S.D., Harris, S.R.: Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. Wellcome Open Res. **3** (2018)
17. Loman, N.J., Pallen, M.J.: Twenty years of bacterial genome sequencing. Nat. Rev. Microbiol. **13**(12), 787 (2015)
18. Maiden, M.C., Bygraves, J.A., Feil, E., et al.: Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. PNAS **95**(6), 3140–3145 (1998)
19. Maiden, M.C., Van Rensburg, M.J.J., Bray, J.E., et al.: MLST revisited: the gene-by-gene approach to bacterial genomics. Nat. Rev. Microbiol. **11**(10), 728 (2013)

20. Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval. Nat. Lang. Eng. **16**(1), 100–103 (2010)
21. Mansouri, M., Booth, J., Vityaz, M., et al.: PRINCE: accurate approximation of the copy number of tandem repeats. In: WABI 2018, pp. 20:1–20:13 (2018)
22. Meehan, C.J., Moris, P., Kohl, T.A., et al.: The relationship between transmission time and clustering methods in Mycobacterium tuberculosis epidemiology. EBioMedicine **37**, 410–416 (2018)
23. Murray, K.D., Webers, C., Ong, C.S., et al.: kWIP: the k-mer weighted inner product, a de novo estimator of genetic similarity. PLoS Comput. Biol. **13**, 1–17 (2017)
24. Nguyen, N.P., Warnow, T., Pop, M., White, B.: A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. NPJ Biofilms Microbi. **2**, 16004 (2016)
25. Ondov, B.D., Treangen, T.J., Melsted, P., et al.: Mash: fast genome and metagenome distance estimation using minhash. Genome Biol. **17**(1), 132 (2016)
26. Pan, X., Papailiopoulos, D.S., Oymak, S., et al.: Parallel correlation clustering on big graphs. In: NIPS 2015, pp. 82–90 (2015)
27. Reed, M., Pichler, V., McIntosh, F., et al.: Major Mycobacterium tuberculosis lineages associate with patient country of origin. J. Clin. Microbiol. **47**, 1119–1128 (2009)
28. Seemann, T.: Snippy (2015). https://github.com/tseemann/snippy
29. Vergnaud, G., Pourcel, C.: Multiple locus variable number of tandem repeats analysis. In: Caugant, D. (ed.) Molecular Epidemiology of Microorganisms, pp. 141–158. Springer, Heidelberg (2009). https://doi.org/10.1007/978-1-60327-999-4_12
30. Williamson, D.A., Baines, S.L., Carter, G.P., et al.: Genomic insights into a sustained national outbreak of Yersinia pseudotuberculosis. Genome Biol. Evol. **8**, 3806–3814 (2017)
31. Xia, E., Teo, Y.Y., Ong, R.T.H.: SpoTyping: fast and accurate in silico mycobacterium spoligotyping from sequence reads. Genome Med. **8**(1), 19 (2016)