MICROBIAL GENOMICS

RESEARCH ARTICLE

Robertson *et al.*, *Microbial Genomics* 2020;6

DOI 10.1099/mgen.0.000435

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance

James Robertson*, Kyrylo Bessonov, Justin Schonfeld and John H. E. Nash

Abstract

Bacterial plasmids play a large role in allowing bacteria to adapt to changing environments and can pose a significant risk to human health if they confer virulence and antimicrobial resistance (AMR). Plasmids differ significantly in the taxonomic breadth of host bacteria in which they can successfully replicate, this is commonly referred to as 'host range' and is usually described in qualitative terms of 'narrow' or 'broad'. Understanding the host range potential of plasmids is of great interest due to their ability to disseminate traits such as AMR through bacterial populations and into human pathogens. We developed the MOB-suite to facilitate characterization of plasmids and introduced a whole-sequence-based classification system based on clustering complete plasmid sequences using Mash distances (https://github.com/phac-nml/mob-suite). We updated the MOB-suite database from 12091 to 23671 complete sequences, representing 17779 unique plasmids. With advances in new algorithms for rapidly calculating average nucleotide identity (ANI), we compared clustering characteristics using two different distance measures – Mash and ANI – and three clustering algorithms on the unique set of plasmids. The plasmid nomenclature is designed to group highly similar plasmids together that are unlikely to have multiple representatives within a single cell. Based on our results, we determined that clusters generated using Mash and complete-linkage clustering at a Mash distance of 0.06 resulted in highly homogeneous clusters while maintaining cluster size. The taxonomic distribution of plasmid biomarker sequences for replication and relaxase typing, in combination with MOB-suite whole-sequence-based clusters have been examined in detail for all high-quality publicly available plasmid sequences. We have incorporated prediction of plasmid replication host range into the MOB-suite based on observed distributions of these sequence features in combination with known plasmid hosts from the literature. Host range is reported as the highest taxonomic rank that covers all of the plasmids which share replicon or relaxase biomarkers or belong to the same MOB-suite cluster code. Reporting host range based on these criteria allows for comparisons of host range between studies and provides information for plasmid surveillance.

DATA SUMMARY

(1) Supplementary tables and figures have been deposited in Figshare: https://doi.org/10.6084/m9.figshare.12678281.v1.

(2) Scripts used for cluster analysis have been deposited in GitHub: https://github.com/jrober84/mobclustering.

INTRODUCTION

Plasmids are autonomously replicating mobile genetic elements that can provide advantageous traits which promote the survival of their host cells and are widely distributed in diverse bacterial species [1–4]. Plasmids are highly variable in gene content, replication mechanism and even conformation. Most plasmids are circular, but linear plasmids have been described within Gram-positive and even some Gram-negative bacteria, such as *Salmonella* [1, 3–6]. Plasmid-mediated traits such as antimicrobial resistance (AMR) are of great public-health concern due to the increasing global prevalence of multidrug resistance in bacteria [7]. Plasmid identification is critical to understanding AMR transmission, since plasmids are the primary vectors for AMR dissemination in *Enterobacteriaceae* [2]. Knowing the prevalence of different plasmid 'types' and their associations with different resistance genes can inform our understanding of the epidemiology

Received 20 May 2020; Accepted 03 September 2020; Published 24 September 2020

Author affiliations: [1]National Microbiology Laboratory, Public Health Agency of Canada, Guelph, ON, Canada.

***Correspondence:** James Robertson, james.robertson@canada.ca

Keywords: bacterial genomes; mobile genetic elements; plasmid host range; plasmids; relaxase typing; replicon typing.

Abbreviations: AMR, antimicrobial resistance; ANI, average nucleotide identity; NCBI, National Center for Biotechnology Information; WGS, whole-genome sequencing.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Five supplementary tables and two supplementary figures are available with the online version of this article.

and transmission of plasmid-mediated antibiotic resistance [2–4, 8, 9]. An improved understanding of the distribution of plasmids and their potential hosts is crucial in the development of intervention strategies to prevent antibiotic-resistance gene spread.

Multiple methods for plasmid classification have been developed, but the two primary methods are replicon typing and relaxase typing, which provide largely complementary information [1–3, 8]. Replication is the only critical function that all plasmids must be able to perform, and there are numerous strategies employed by plasmids to propagate themselves [9–11]. Due to the essential requirement of replication for all plasmids, a typing scheme of 18 incompatibility/replicon types was devised based on the molecular basis for replication using laborious mating experiments, where two plasmids were considered to be from the same incompatibility group if they were unable to be stably maintained in the same cell [12]. A high-throughput PCR assay was developed to rapidly classify plasmids into 18 replicon types, and it was later extended to identify 116 types using *in silico* detection with the PlasmidFinder tool [9, 11]. Plasmids frequently contain multiple replication systems, and this complicates tracking plasmids based on replicon typing [1, 8]. MOB-typing is a complementary typing system based on the relaxase sequence from the conjugation apparatus, which consists of six families [1, 4, 8, 13]. This coarse-grain typing system provides broader context for transmissible plasmids with the advantage that it is uncommon for a plasmid to contain multiple relaxases, something that is not true for replicons; however, this approach is not applicable to non-transmissible plasmids, which do not contain relaxase sequences [1, 4, 8, 13]. A PCR-based assay for MOB-typing made it possible to type isolated plasmids in the laboratory, but uptake of MOB-typing for sequenced plasmids was low due to the lack of tools to easily perform automated typing from sequence data until the inclusion of this feature within the MOB-suite [13, 14]. However, there remains a significant drawback to typing plasmids based on marker sequences, in that novel plasmid types require evaluation by experts in order to update repositories of marker sequences.

Plasmid typing information is used to make epidemiological inferences, and the utility of any particular scheme depends on the proportion of plasmids covered by the scheme and its ability to provide accurate information about the evolutionary relationships within and between plasmid types [8, 9, 11]. Both replicon and MOB-typing provide valuable insights into the distribution of plasmids, but neither approach is applicable to all plasmids [1, 8]. There are no universal marker sequences that are present across the diversity of plasmids and, due to their recombinogenic nature, there can be conflicting phylogenetic signals between different sections of the same plasmid [1, 6, 8, 9]. Phylogenetic inference methods depend on the strong vertical inheritance of sequence [15], and the mosaic nature of plasmids [16] presents a challenge to using these types of methods to understand the evolutionary history of plasmids. To address some of the limitations of these typing methodologies, the MOB-suite implements a scalable nomenclature for plasmid typing by estimating genomic distances based on Mash min-hashing [14, 17]. The MOB-suite clustering approach could be considered analogous to an operational taxonomic unit in bacterial diversity studies, where each plasmid is assigned a cluster code based on a defined similarity threshold [14, 18]. The Mash distance threshold used in the MOB-suite was selected empirically to maximize the ability of the tool to accurately reconstruct individual plasmids within a sample [14]. It is possible for an individual plasmid to contain multiple replicon or relaxase biomarker sequences, and the MOB-suite implemented whole-sequence classification approach solves this problem since each plasmid can only be assigned to a single cluster [14].

Impact Statement

Bacterial plasmids play a large role in allowing bacteria to adapt to changing environments and can pose a significant risk to human health if they confer virulence and antimicrobial resistance. Plasmid typing provides insights into the host-range distribution of plasmids in populations, as well as performing epidemiological surveillance. The two primary methods for typing are not universally applicable to all plasmids and provide limited resolution. With the advent of large-scale sequencing of complete plasmid sequences, it is possible to utilize the entire sequence of the plasmid in comparisons rather than relying solely on a small number of marker genes. The MOB-suite implemented a whole-sequence-based typing system that provides cluster codes for reconstruction and tracking of plasmids, but with advances in rapid genome comparison methods, we identified refinements that could be made to the approach to maximize concordance with existing typing methods. We have leveraged this cluster information along with existing biomarker-based typing and literature evidence to provide predictions of the taxonomic range in which a plasmid could replicate. This information will be valuable for building risk-based models on plasmid transmission.

Recently, new high-throughput methods for determining genetic distances for whole genomes that leverage the MinHash technique have been developed [17]. Mash was the first implementation of this approach applied to genomics data and can rapidly estimate the Jaccard index of similarity between genomes [17]. Mash is also the distance measure used by MOB-suite to develop and assign plasmids to MOB-cluster codes [14]. Average nucleotide identity (ANI) is another commonly used approach for estimating relatedness between organisms at the genomic level [19]. However, ANI suffered from poor scalability due to its reliance on alignment-based methods until fastANI was developed with the MinHash technique to rapidly determine ANI between genomes [19]. Mash and ANI are tightly correlated when comparing similar genomes within the range of 90–100% ANI, but this correlation becomes weaker with increasing divergence [17, 19]. ANI is calculated based on the sequences that are present in both genomes [17, 19], and so requires post-processing of results to determine whether a sufficient amount of the genome was considered in the comparison. Mash measures the shared number of k-mer sketches in relation to the entire unique set of k-mers in the two genomes being compared, and so the distance considers both shared and unique sequence data [19]. Due to the high sequence variability of plasmids, it is not known whether Mash or ANI distance measures are more reliable for delineating plasmid groups of epidemiological and biological relevance.

Plasmid host range, in the context of the present work, describes the hosts in which a plasmid can replicate successfully. Plasmid host range is a complex trait and depends on the molecular basis for plasmid replication and maintenance, as well as capacity for horizontal transfer [20–23]. The taxonomic breadth in which a given plasmid can be transferred and successfully established varies wildly between plasmids, but is qualitatively categorized as either ‘broad’ or ‘narrow’ [20–23]. The terms broad and narrow are used to describe plasmids with compatible hosts across very different taxonomic ranges, a broad host range plasmid could be known to be compatible with bacteria from the same order (e.g. *Enterobacteriales*) or even with bacteria from separate kingdoms. As a result, the qualitative labels broad and narrow by themselves cannot be used to infer the possible bacterial hosts for a given plasmid.

Understanding the host-range potential of plasmids is of great interest due to the ability of plasmids to disseminate traits such as AMR through bacterial populations and into human pathogens [2, 20]. Numerous laboratory experiments have been performed using representatives from diverse sets of taxa and provide valuable insights into the potential host range [20–23]; however, testing all of the potential hosts for a plasmid under the variety of experimental conditions required, in order to assay host range exhaustively, is an intractable challenge. Sequence-based approaches for assessing replication host range are an attractive alternative, and it has been previously demonstrated that there are measurable differences in genomic signatures between narrow and broad host range plasmids compared to their corresponding host chromosome [24]. This genomic signature-based approach will identify taxa where the chromosome and plasmid signatures are similar as potential hosts in which the plasmid has evolved [24]. However, this approach likely reflects only the evolutionary hosts of a plasmid and may not reflect all the potential hosts in which a plasmid could currently replicate.

Routine analysis of human pathogens using whole-genome sequencing (WGS) generates large volumes of sequence data that can be exploited to examine the distribution of plasmids in pathogens of interest [25]. Plasmid host distributions have been examined using complete plasmid and relaxase sequences as queries in more than 449 000 WGS samples [25]. This empirical observation of complete plasmid and marker sequences is an attractive approach to predicting host range, since it leverages existing data generated for other applications. For epidemiological applications, an understanding of which hosts a plasmid can successfully transfer into and replicate in will inform outbreak and surveillance investigations of potential risks, such as emergence of highly antimicrobial-resistant human pathogens or potential outbreak onset. However, estimates of host range based on data within the National Center for Biotechnology Information (NCBI) database will be biased towards culturable bacterial species of health concern, research interest or industrial applications [26].

Here, we have evaluated Mash- and ANI-based distance measures for delineating plasmid groups. We have also refined clustering within the MOB-suite to provide a plasmid nomenclature giving insight into the distribution and transmission dynamics of plasmids at epidemiological relevant timescales. If the distance threshold used to generate the clusters is set too low, then a single cluster will potentially contain multiple distinct plasmids from the same host cell (violating the plasmid incompatibility requirement), limiting the usefulness of the cluster as a biologically relevant identifier. However, if the threshold is set too stringent, then a draft plasmid may end up in a separate cluster from its complete sequence. Within this range, we determined an optimal threshold that provides the best trade-offs between cluster size and concordance with traditional typing. Replicon and MOB-typing schemes will continue to be useful, since they provide meaningful contextual and functional information about the plasmid clusters. MOB-suite’s clustering approach is an attractive complementary typing approach, especially for plasmids that are not typeable using the existing approaches.

METHODS

Expanding MOB-suite's internal database of high-quality plasmids

The MOB-suite v. 1 database contains 12091 complete plasmids, and due to the increased number of plasmid sequences made available since the original publication, we expanded the database using new data from the NCBI utilizing the same approach described in the supplementary materials of the MOB-suite paper [14]. The NCBI Entrez nucleotide database was queried in November 2019 with the query ‘plasmid’ AND ‘complete sequence’ AND ‘bacteria [organism]’. The results were then filtered for sequences between 1500 to 400000 bp in length, with ‘plasmid’ as the genetic compartment and limited to the INSDC (International Nucleotide Sequence Database Collaboration). This yielded an initial set of 33875 sequences that were then typed using MOB-typer v. 2.1.0 (Table S1, available with the online version of this article). Records were excluded due to the presence of any of the following terms in the title or description: gene, cds, protein, transposon, insertion, protein, region, operon, pseudogene, integrase, transposase, integron, partial, shotgun. The remaining set of 23280 sequences was then merged with the MOB-suite v.1 database plasmids, which were then de-duplicated by clustering plasmids that had a Mash v. 2.2.2 [17] distance of 0 and selecting a single representative for subsequent analyses (Table S2). A priority was given to those plasmids that were part of the initial construction of the MOB-suite clusters, which resulted in a total of 17779 records. The database exhibits a strong bias towards plasmids from *Enterobacteriaceae* (35%) as shown in a Krona plot of the plasmid dataset taxonomic composition (Fig. S1). This has the consequence that the threshold optimization may not be fully representative of the underrepresented taxonomic groups.

Identification of initial Mash threshold range

Clustering thresholds used by MOB-suite were designed to be specific enough as to not result in multiple representatives of the same cluster arising from a single host cell, and broad enough to include draft and closed versions of the same plasmid in the same cluster [14]. The first constraint being necessary to remain consistent with the definition of an Inc group and the second necessary to give the tool the flexibility needed to work with draft genomic data. We determined an effective threshold range empirically using the closed genome set previously employed to benchmark MOB-suite [14]. Mash distances are sensitive to k-mer and sketch sizes, and so we examined the effects of these parameters on generated clusters over a range of thresholds (0 to 0.1). We tested pairwise combinations of k-mer sizes 12–21 and sketch sizes 400, 1000 and 2000, and calculated the adjusted rand index of cluster concordance for all combinations within a single cluster threshold. Violin plots of the results are available in Fig. S2. For our dataset, we found the worst concordance of 0.79 at a distance threshold of 0 when comparing the smallest k-mer and sketch (k=12, sketch=400) against the largest (k=21, sketch=2000). Altering sketch and k-mer sizes has an effect on the resulting clusters, and we chose to use the default settings for Mash (k=21, sketch=1000), since the developers identified high concordance with Mash and ANI using these parameters [17].

This set consisted of 133 closed genomes with 377 associated plasmids, which were sequenced using PacBio and Illumina technologies and previously used to validate the MOB-suite plasmid reconstruction [14] (Table S3). We determined the minimum Mash v. 2.2.2 distance between all pairs of plasmid sequences in a genome to identify the maximum threshold that could be used before genomes would contain multiple representative plasmids from the same cluster. Since the Mash distances consider all of the k-mers present in both plasmid sequences, it was used to construct the initial clustering boundaries. We discounted the use of ANI for this particular analysis since it will only consider what is shared and, thus, is not appropriate for comparing sequences that have nothing in common. We determined the lowest threshold possible that would not assign a draft version of a plasmid to a different cluster by comparing the Mash distances between draft assembly versions of the plasmids and their completed sequence. The Illumina data was assembled using Unicycler v. 0.4.4 with the default parameters and the resulting assemblies were assigned to the reference using BLASTN v. 2.6.0 [27] with the following options: -max_hsps 1 -num_alignments 1 -perc_identity 50 -qcov_hsp_perc 50. A separate FASTA file was constructed for the chromosome and each plasmid, and the Jaccard distance was estimated using Mash v. 2.2.2 between the draft and completed versions of each molecule with the default parameters.

Selecting an optimal threshold for clustering

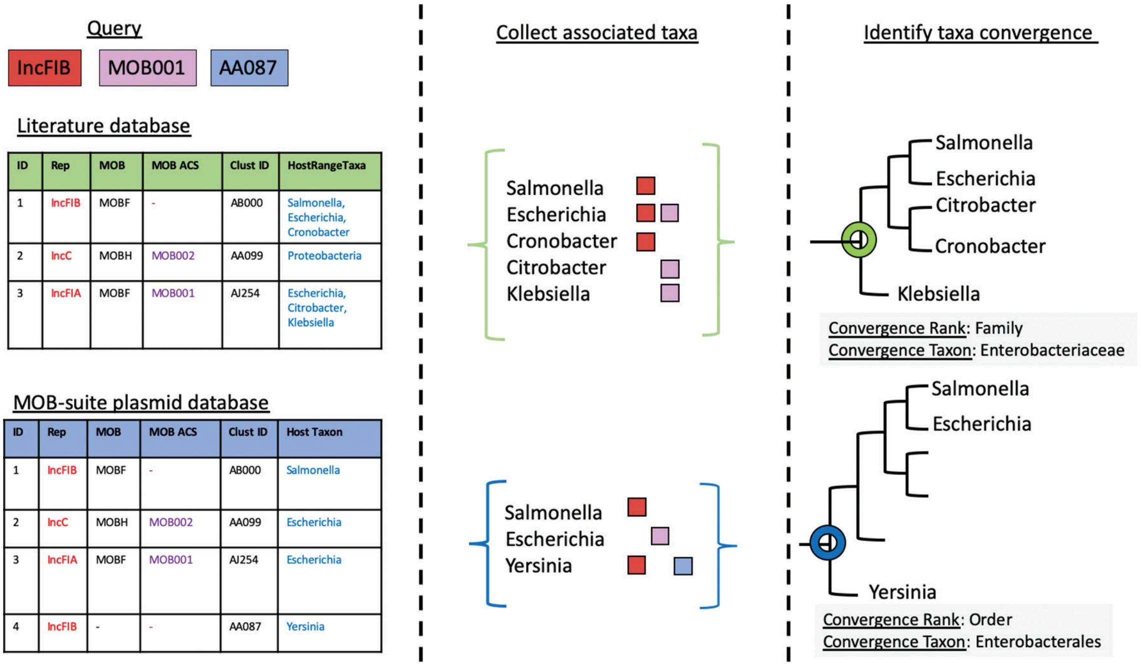
Pairwise ANI was calculated over the set of 17779 putative plasmid sequences using fastANI v. 1.3 with the following parameters: -k 21, -t 32, --fragLen 500, --minFrag 1 [19]. These fastANI parameters were selected so that both large and small plasmids could be analysed together, and such that the k-mer length used between Mash and fastANI would be the same when comparing these two algorithms. ANI results were then filtered using a Python script to set ANI to zero when there was less than 50% overlap between sequences being compared (https://github.com/jrober84/mobclustering). In order to compare Mash and ANI distances, the obtained ANI values were converted to Jaccard distances by the following equation (1 – ANI), as was done elsewhere [17]. Mash v. 2.2.2 was used to calculate the distance between sequences using the default parameters. Clustering on each of the distance measures was performed across thresholds ranging from 0 to 0.1, with an increment of 0.01 between steps. Three different linkage methods were used to cluster the plasmids at each threshold: complete, single and average (implemented via SciPy v. 1.4.1). The cluster membership information was overlaid with the replicon and relaxase information.

Each clustering was evaluated by computing the properties of its component clusters. The component clusters were evaluated for size and purity with respect to either replicon or relaxase types. Purity was captured in two ways: using the Shannon entropy (Equation 1) to measure the diversity of each cluster, along with counts of the number of replicon/relaxase types present in each cluster (https://github.com/jrober84/mobclustering).

The Shannon entropy is computed as $S = -\sum_i P_i \log P_i$, where P_i is the probability of each replicon or relaxase type occurring in a cluster (Equation 1).

For an individual clustering, at a given threshold, the distribution of cluster sizes, entropies and contained types are often not normally distributed. This means that simply averaging the properties (i.e. cluster size) will yield a deceptive answer. Consider a clustering of 10 isolates that resulted in four clusters: one cluster of size 7 and three clusters of size 1. The mean cluster size for this clustering is 2.5, which gives the impression that the clustering consists of a number of smaller clusters when most of the isolates are in one large cluster. In our approach, we weighted cluster size by the number of isolates contained in the cluster. An isolate that was the member of a cluster with a size of 7 would be given a score of 7 for computation of the mean cluster size. This would result a mean of 5.2 instead of 2.5, which is a better representation of the state of the average isolate.

In order to determine an optimal threshold for clustering plasmids, we considered the sum of each of the means of each of the three properties (Equation 2). We normalized cluster



Literature database

ID	Rep	MOB	MOB ACS	Clust ID	HostRangeTaxa
1	IncFIB	MOBF	-	AB000	Salmonella, Escherichia, Cronobacter
2	IncC	MOBH	MOB002	AA099	Proteobacteria
3	IncFIA	MOBF	MOB001	AJ254	Escherichia, Citrobacter, Klebsiella

MOB-suite plasmid database

ID	Rep	MOB	MOB ACS	Clust ID	Host Taxon
1	IncFIB	MOBF	-	AB000	Salmonella
2	IncC	MOBH	MOB002	AA099	Escherichia
3	IncFIA	MOBF	MOB001	AJ254	Escherichia
4	IncFIB	-	-	AA087	Yersinia

Fig. 1. A simplified example of the host-range prediction feature implemented within MOB-typer. Host-range prediction uses replicon type, relaxase biomarker accession number and MOB-cluster to individually query a literature database of publications associated with plasmids and the MOB-suite plasmid database. The taxonomy associated with each of the records is aggregated and placed into a taxonomic hierarchy. The hierarchy is then processed to identify the point of taxonomic convergence, the lowest taxonomic rank that is parent to all of the taxa involved. Both the literature host range and the plasmid database convergence ranks are reported to the user.

size and number of types within a cluster to be within the range 0 to 1 (Equation 3). Since we were interested in minimizing the Shannon entropy and the number of types within a cluster, we took (1 – property) for each of those properties.

Score(t, m) = (1 – AvgEntropy(t, m)) + AvgClusterSize(t, m) + (1 – AvgNumberTypes(t, m)) (Equation 2).

$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$, where i is the cluster size or number of types for an individual sequence (Equation 3).

The AvgEntropy is the weighted mean of the Shannon entropies of each cluster for a given threshold (t) and clustering method (m). The AvgClusterSize and AvgNumberTypes are the weighted means of the cluster sizes and number of replicon/relaxase types in each cluster, respectively.

Taxonomic analysis

Host range was determined by using the taxonomic hierarchy associated with the NCBI records. Plasmids were grouped by replicon type, relaxase ID and MOB-cluster code, and the taxonomic point of convergence was determined as the taxon that was parent to all of the different taxa associated with a given type. For example, the replicon type 'Col(YF27601)' was associated with three records assigned to two species: *Yersinia frederiksenii* and *Yersinia kristensenii*. The taxonomic convergence for this replicon would be at the genus level. Individual query relaxase sequence IDs were used in place of relaxase types, since the majority of the types resolved at the level of the NCBI taxon rank of superkingdom. We are using the NCBI taxonomy for comparisons, but in traditional taxonomy 'bacteria' has the rank of kingdom.

Building the literature database

There is a wealth of knowledge on plasmid biology contained within the literature, which provides valuable insights into the potential host ranges of the plasmids within the MOB-suite database. Therefore, we utilized a literature mining approach to select publications with information about plasmid host range. We performed a search of publications associated with plasmid accession numbers in our reference database and identified 64 unique publications that had information about plasmid host range. Host-range information was extracted manually from each article and added to the literature database to provide users with relevant publication details.

Host-range module workflow description

MOB-suite leverages the typing information provided by MOB-typer to query a literature database of reported host ranges and the internal MOB-suite database of closed plasmids based on replicon type, relaxase accession number and MOB-cluster ID. The taxa associated with each of the three queries are aggregated together and the host range is reported as the taxon and rank that contains all of the associated taxa. This process is illustrated in Fig. 1 with a simplified example

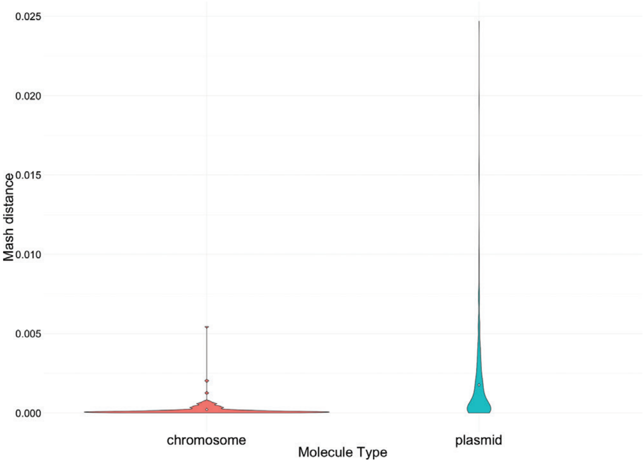


Fig. 2. Violin plot of the Mash distances between complete and draft versions of plasmid and chromosomes.

where MOB-typer has identified a plasmid as belonging to the IncFIB replicon type, possessing a MOB001 relaxase and assigned to MOB-cluster AA087. Both the literature and plasmid databases contain additional information, but only the information relevant to identifying the host range is shown in the diagram. Within the literature database, it is possible to have multiple taxa associated with a single record, since these experiments would test the capability of that plasmid to be transferred into multiple hosts. The reference plasmid database, however, will only contain a single taxon, the taxon associated with the record in the NCBI database.

In each case, the list of taxa is then aggregated together, and using the NCBI taxonomic hierarchy, each taxon is placed on a tree and the point of convergence for the set of taxa is identified as the closest parent taxonomic rank that contains all of the listed taxa. When a plasmid is only observed within a single species, the prediction is moved up to the genus level to account for some ambiguities in species assignments. In the case of the literature database, the taxonomic rank that contains all of the taxa is at the family level and the specific taxon is *Enterobacteriaceae* (Fig. 1). In the case of the plasmid database, the point of convergence for the set of taxa is *Enterobacterales* at the order level (Fig. 1). MOB-typer will provide the reported host range based on the literature, as well as an observed host range based on the MOB-suite plasmid database. The reported and observed host ranges are combined, and the highest rank is reported as the predicted host range.

Assembly, reconstruction and typing of plasmids

Illumina reads and metadata were downloaded from NCBI BioProject PRJNA285421 and assembled using Unicycler v. 0.4.4, plasmids were reconstructed using MOB-recon and typing information was obtained using MOB-typer v. 3.0.0.

RESULTS

Analysis of closed genomes to identify initial threshold ranges for MOB-cluster development

A strong motivator for the development of a nomenclature within the MOB-suite is to develop groupings where it is unlikely to have two plasmids of the same type within the same cell, so that they could be used for reconstruction of individual plasmids from draft assemblies. In order to determine the window of valid thresholds for these purposes, we utilized the closed benchmarking genomes from the MOB-suite paper [14]. Since our goal was to have an approach that could be applied to draft and incomplete plasmids, we examined the distribution of Mash distances between the deposited closed sequences and their corresponding draft assemblies. A violin plot of the distances observed for plasmids and chromosomes within the 133 genomes is presented in Fig. 2. The full list of strains and genomes used for these comparisons is available in Table S3. The highest distance observed between pairs of draft and complete chromosome sequences was 0.005, which was considerably lower than the Mash distance of 0.025 observed as the maximum distance between pairs of draft and complete plasmids. Given that plasmids are much smaller than chromosomes in this set, it follows that even small amounts of missing sequence data would have a much larger impact on the distances obtained. As it is the maximum observed Mash distance between a draft plasmid and its complete sequence counterpart, we chose 0.025 as the lower bound for our threshold.

We examined the pairwise intra-genomic distance for all sequences (plasmid and chromosome) within each complete genome to determine the distribution of minimal Mash distances observed within a single genome. We observed that there were few cases where the minimum Mash distance approached zero, with most intra-genomic distances being 0.1 and above (data not shown). We could not use the absolute minimum Mash distance observed (0) as an upper bound, as it would allow for no meaningful clustering threshold to be chosen, since it is lower than our chosen lower bound threshold of 0.025. Since we are comparing the performance of Mash and ANI, we selected the upper bound to be 0.1, as it has been established previously that they are highly comparable over the range of 0–0.1.

Benchmarking the performance of Mash- and ANI-based distances for clustering complete plasmid sequences

Using our de-duplicated dataset of 17779 closed NCBI plasmids, we determined what would be an optimal threshold for MOB-suite cluster codes with respect to size of clusters, along with adherence to existing replicon- and relaxase-based typing schemes. Concordance with traditional typing was examined using the mean Shannon entropy along with the mean number of types found within a cluster. These two measures are interrelated, but we chose to examine both due to the fact that our dataset is biased to contain a small number of highly abundant types and this can artificially decrease

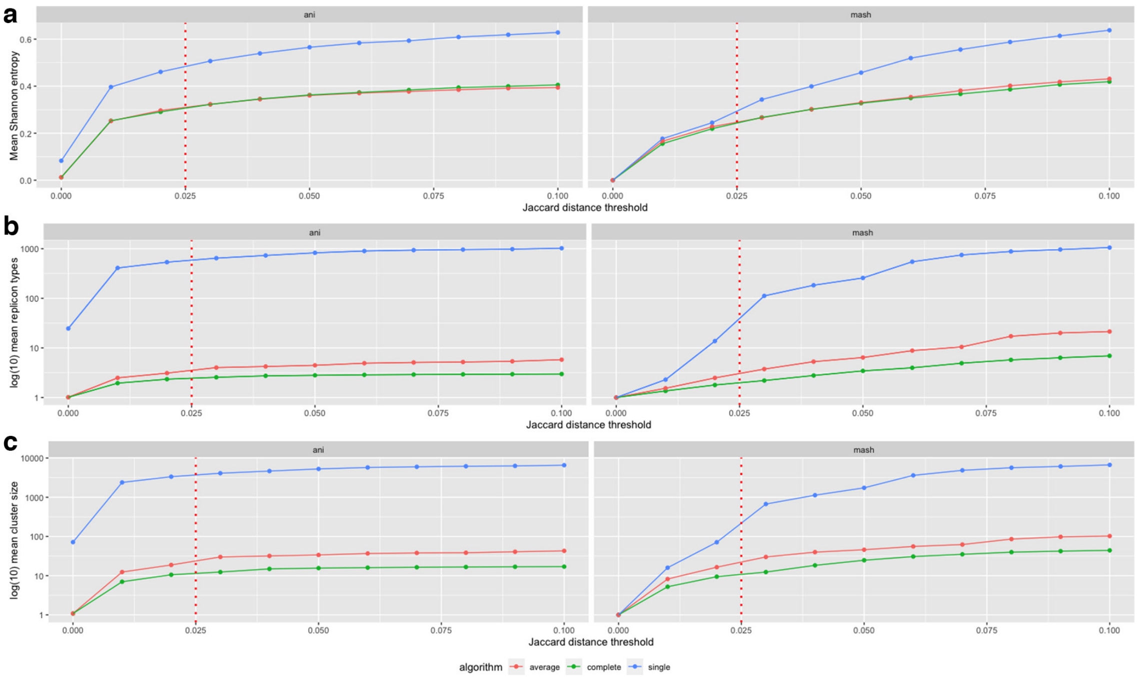


Fig. 3. Replicon-typed plasmids were clustered using either ANI- or Mash-based distances using complete-, single- or average-linkage algorithms. The mean Shannon entropy and mean number of types is based on the number of replicon types present in each of the clusters. The lower bound derived from the closed genome analysis is highlighted by the vertical red dotted line.

entropy. Due to the popularity of ANI for species delineation and the development of a fast approach for calculating ANI, we benchmarked plasmid clustering using both ANI and Mash distance measures with three different clustering algorithms to determine which method would optimally partition the data. The clusters were designed to minimize variability in typing information within a cluster to maximize concordance with existing methods, while maximizing the number of members within a cluster. We examined replicon and relaxase typing information independently, since each method is applicable to a different number of plasmids.

Of the 17779 putative plasmid sequences, 12802 records could be classified using replicon typing and these were used to examine the clustering dynamics over the selected Jaccard distance thresholds (Fig. 3). Single-linkage clustering produced results that were highly divergent from those obtained by complete- and average-linkage clustering across all of the measured attributes when either Mash or ANI distances were used (Fig. 3). Single-linkage clustering produced larger clusters compared to the other algorithms, but also produced clusters with a much higher entropy and containing a much higher number of replicon types (Fig. 3). Complete-linkage clustering displayed a much more conservative behaviour by producing smaller clusters with lower entropy and fewer replicon types (Fig. 3). Average-linkage clustering displayed an intermediate behaviour between the other two methods, but the results much more closely resembled those obtained with the complete-linkage algorithm (Fig. 3).

The same analyses were repeated on the 9640 plasmids that could be classified by relaxase-typing, and similar results were observed for the three algorithms (Fig. 4). The plots for the plasmids classified by relaxase typing mirror what was seen for those classified by replicon typing for each of the three algorithms tested and for both distance measures used (Fig. 4). However, the magnitude of variation for Shannon entropy and number of types is much smaller given that there are only six defined classes of relaxases versus the 1770 replicon types defined within the MOB-suite replicon database. For both replicons and relaxases, the slope for Mash was higher than for ANI (Figs 3 and 4), which is likely due to our requirement that for sequences to be comparable via ANI, there needs to be at least a 50% overlap.

Identification of the optimal distance measure, threshold and clustering algorithm for MOB-cluster

Selecting the optimal distance threshold for partitioning our data is a multi-criteria optimization problem where we are attempting to maximize both cluster purity and cluster size. As described in Methods, we developed a scoring function for each distance threshold based on three factors that were weighted equally: cluster size, number of types and entropy. The measures for both the cluster size and for the number of types per cluster were scaled so that they were between 0 and 1, with the number of types reversed since the best score for this feature should minimize the number of types in a cluster. The scores for both Mash and ANI using each of the

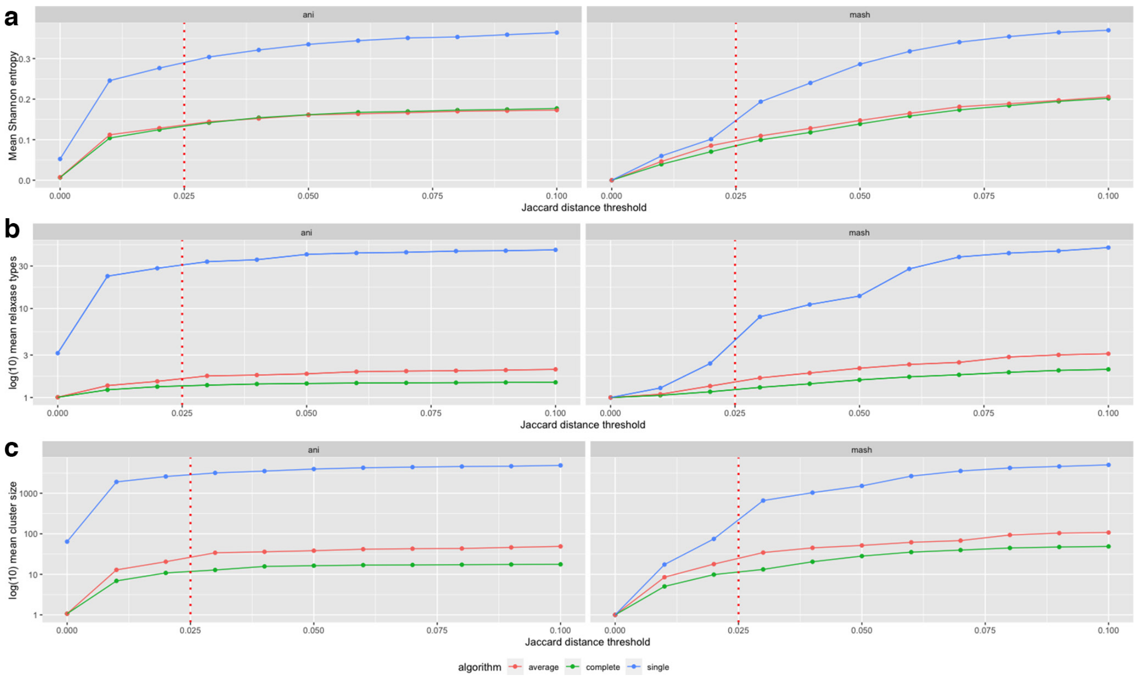


Fig. 4. Plasmids that were typed according to the existing relaxase accession numbers were clustered using either ANI- or MASH-based distances using complete-, single- or average-linkage algorithms. The mean Shannon entropy and mean number of types is based on the number of relaxase accession numbers present in each of the clusters. The lower bound derived from the closed genome analysis is highlighted by the vertical red dotted line.

three algorithms are presented in Fig. 5. Clustering analyses on replicon- and relaxase-typed plasmids are presented separately, since the sets of plasmids typed by the two methods are different. The highest score was achieved for complete clustering using Mash for both replicon and relaxase typing at a threshold of 0.01. However, based on the experiments using the closed genomes, this falls outside the threshold where draft and complete genomes would be separated (Fig. 5). Using our bounding thresholds from the closed genome experiments, a threshold of 0.06 produced the highest scoring clustering of the data for both replicon and relaxase typing when using Mash as the distance measure and complete-linkage as the clustering method (Fig. 5). On the whole, ANI-based clusterings had a lower score than Mash-based clustering over the range of 0.025–0.1 for both replicon- and relaxase-typed plasmids.

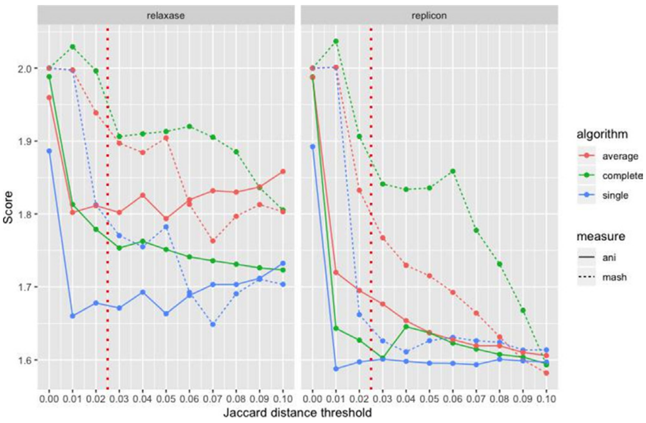


Fig. 5. Performance scores across different distance thresholds of either Mash and ANI distance measures using three different clustering approaches: complete-, single- and average-linkage. Performance scores are the result of combining mean cluster size along with cluster entropy and the mean number of types (replicon or relaxase) within a cluster (Equation 3). The lower bound of the clustering threshold determined by the closed genomes experiments is signified by the vertical red dotted line.

Stability of cluster codes over epidemiological time scales

In order to be useful taxonomic units, the derived clusters will also need to be relatively stable over the course of epidemiologically relevant time scales. To assess the use of the cluster codes in an epidemiological context, we re-analysed data from a study that looked at carriage of *Salmonella enterica* subsp. *enterica* serovar Typhimurium over a period of up to 279 days within 11 patients to see whether the plasmid clusters were stable. MOB-recon reconstructed the plasmid content for each of the samples and each of the resulting plasmids were typed with MOB-typer (Table S4). Patient A possessed two plasmids (Col156 and IncFIB,IncFII) at time point one, and for the other two time points only a single plasmid was detected (Table S4). The IncFIB,IncFII group for all three time points was assigned to MOB-cluster AB460. Patients

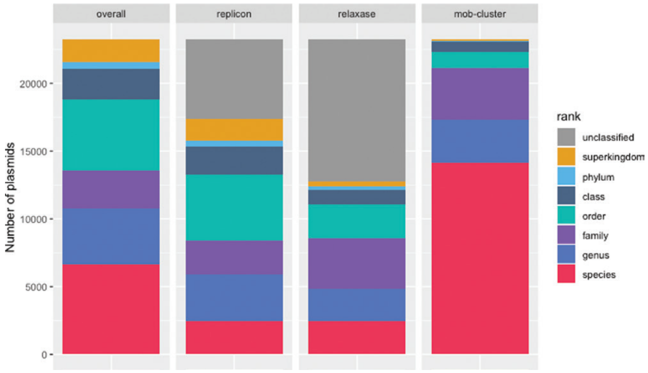


Fig. 6. Stacked bar chart of the highest point of taxonomic convergence for plasmids based on replicon types, relaxase accession numbers and MOB-clusters. An overall convergence was determined using all of the features applicable to a given plasmid and picking the highest convergence point achieved.

B–F and H–K each possessed a single plasmid within all of their tested time points, which was also the IncFIB,IncFII plasmid assigned to MOB-cluster AB460 (Table S4). Patient G presents an interesting case of plasmid flux within the patient, with samples taken at four time points post-infection: 0, 9, 24 and 206 days. The sequence data from the first three time points reconstructed the same three MOB-clusters, AA162, AB037 and AC082, with MOB-cluster AC082 consisting of a ~25 kb IncQ1 plasmid (Table S4). By the fourth time point, only MOB-cluster AC082 was detected in the sample, which indicates that the previous plasmids were not stable over prolonged carriage within this patient. It is worth noting that MOB-cluster AC082 (IncQ1) plasmids lack conjugative transfer genes, but MOB-cluster AA162 contains a MOBP relaxase and is predicted to be mobilizable, which potentially could have mobilized MOB-cluster AC082 plasmid in trans.

Identification of plasmid host ranges based on molecular typing and MOB-clusters

We examined the taxonomic relationships between replicon types, relaxase accession numbers and MOB-clusters based on the host taxonomy of the 23280 closed NCBI plasmids. The observed taxonomic range of each feature (replicon type, relaxase type, MOB-cluster ID) was examined individually, as well as an overall range based on the combined set of features Fig. 6. The number of plasmids that could be typed with each of the methods differed greatly with 5903 (25%) and 10523 (45%) untyped plasmids for replicon and relaxase typing, respectively, but 0 (0%) for MOB-clusters (Fig. 6). As a consequence of the high specificity required for MOB-clusters, 60% of plasmids were identified within a single species compared to the 14 and 19% observed for replicon- and relaxase-typed plasmids (Fig. 6).

We observed that across all three clustering methods (replicon, relaxase and MOB-cluster), the number of plasmids whose host ranges converge at the phylum level is lower than those whose host ranges include multiple phyla. This suggests that plasmids that are capable of replicating in highly diverse taxa in a single phylum are also likely to be able to replicate within other phyla (Fig. 6). We took the highest taxonomic convergence point for replicon, relaxase and MOB-clusters, and observed that 28% of plasmids are found in a single species. Since most plasmids are unlikely to be specialized to replicate in only one host species, this may reflect plasmids with limited mobility or sparse sampling. Based on the dataset, a total of 22% of plasmids are observed to have a host range that converges at the order level.

We compared the host ranges predicted through our analyses of the NCBI closed plasmids to what has been established in the literature for some well-known replicon types. IncP plasmids are a well-characterized family for which plasmids have been experimentally shown to replicate within a broad range of Gram-negative bacteria and even some Gram-positives [20, 21, 23, 23, 27]. Within the collection of plasmids described here, the taxonomic convergence in our dataset for the IncP replicon is consistent with the literature, since it converges at the superkingdom rank with representatives from both *Actinobacteria* and *Proteobacteria* (Table S5). However, a previous study demonstrated that IncP plasmids can replicate in *Firmicutes* as well [21], which highlights the ability of literature surveys to provide insight into plasmid host ranges not captured by taxonomically annotated sequence data.

Plasmids belonging to IncQ1 are known to replicate in a broad range of hosts [20, 24, 28], and the closed plasmid dataset shows that the host taxonomy for these plasmids converges at the phylum level of *Proteobacteria*. Conversely, IncI-complex plasmids are thought to generally have a narrow replication host range [20, 24], and recently the typing of this group has been updated to contain four groups: IncI1/B/O, IncI-gamma/K1, IncI2 and IncK2/Z [29]. The host range for Inc1 was found to be slightly broader than IncI2 with ranges of *Enterobacterales* and *Enterobacteriacea*, respectively (Table S5). This likely represents sampling biases in the dataset rather than a genuine difference in replication host range.

DISCUSSION

Plasmids play a fundamental role in enabling bacteria to survive and exploit new niches, which can have large consequences for human health in the case of virulence and AMR [1, 2, 9]. Plasmid typing information is critical to enable epidemiological surveillance of plasmids and inform investigations into the transmission pathways of plasmids [8, 9]. Replicon typing has served as the primary system of plasmid classification for decades [1, 8, 9, 11], and there is a wealth of historical knowledge about the biology and prevalence of plasmids based on that nomenclature. A broad nomenclature based on relaxase typing has also been used to classify plasmids, and it has proven valuable in terms of grouping plasmids into an evolutionary context [1, 4, 13, 26]. Due to the recombinogenic nature of plasmids [16] and the associated diversity, marker-based classification systems struggle to encompass the full breadth of the plasmid population. With advances in genome sequencing, it is now possible to develop an automated sequence-based nomenclature that can partition plasmids into cluster codes which maximize compatibility with the original marker-based typing methods [14].

It is important to determine which distance measure generates the most useful clusters based on well-defined criteria. Clusters for MOB-suite were originally designed for reconstruction of plasmids from draft assemblies: one driving criterion for selecting the appropriate cluster boundaries is that it would be highly unlikely for two plasmids belonging to the same cluster to co-exist in the same cell [14]. This criterion mirrors the original underlying assumption of replicon typing, that two plasmids of the same type could not stably coexist in the same cell [11, 12]. The original clustering approach in the MOB-suite used Mash-based distances [14], but with the availability of fastANI [19] it was important to determine which distance (Mash or ANI) generated clusters that maximized concordance with relaxase and replicon typing while minimizing the number of singletons. Mash-based clusters empirically generated the highest scoring clusters using the complete-linkage algorithm at a Mash distance of 0.06 (Fig. 5). The original MOB-suite clustering approach utilized the single-linkage algorithm and a Mash distance of 0.05, and when compared against the new clusters produced an adjusted rand index score of 0.81, indicating agreement between the approaches, but there is definite variation in cluster membership. Analysis of cluster size and purity showed that the complete-linkage algorithm in Fig. 5 generated more homogenous clusters and this is a desirable attribute . Complete-linkage clustering does not suffer from the 'chaining-effect' that is present in average- and single-linkage methods [30]. In the case of plasmids, this effect can prove highly problematic, since there is a high degree of mosaicism between plasmids and transposable elements are readily exchanged between otherwise dissimilar plasmids [16].

For a typing methodology to have any merit, it must be repeatable and robust to minor differences in the plasmid content due to technical issues with sequencing technology, as well as sensitive to genuine biological variability. Plasmids can be highly variable in their gene content, but relatively conserved backbones are present within plasmid groups [6, 11, 16, 26, 31–33]. As demonstrated through the longitudinal analysis of patients chronically infected with *Salmonella enterica* Typhimurium [34], the MOB-suite cluster codes can facilitate plasmid tracking due to their stability within the same patient at multiple time points (Table S4). Analysis of the patient data also highlighted the ability of the MOB-suite to identify instances of plasmid flux within a single patient. Patient G had three plasmids over the course of the first month of testing, but by the end of 206 days only a single plasmid remained. The 25 kb IncQ1 plasmid was assigned to the same cluster (AC082) over the course of testing, which demonstrates that despite minor fluctuations in sequence content and draft assembly, the plasmid could reliably be assigned to the same cluster. Some plasmids exhibit extensive plasticity with complex genomic content changes over short time scales [35], and this approach may fail to assign such plasmids to the same cluster if the proportion of the changed sequence content is high. Analysis of Illumina short-read assemblies using the MOB-suite to reconstruct plasmid content is a useful tool for hypothesis generation. However, for applications involving complete mobile elements associated with transposable elements, long reads may be required in order to gain a full understanding of the relationship of those elements with different plasmid backbones [35].

Plasmid host range is a complex phenotype and has only been described in qualitative terms, which makes it difficult to compare results between studies since the broad classification is used to describe plasmids with different magnitudes of taxonomic distributions [20–23]. To address this issue, we propose that plasmid host range should be codified based on the taxonomic hierarchy of the organisms in which plasmids can successfully establish themselves. Broad-host-range plasmids such as IncP would now be described in terms of having a multi-phyla host range of *Actinobacteria*, *Firmicutes* and *Proteobacteria* based on the observed taxonomy of the NCBI plasmids, as well as experiments from the literature [21, 23]. This specificity would readily allow for comparisons between different studies and could be updated as new experimental evidence becomes available. Previously, sequence similarity searches of the publicly deposited WGS data for plasmid sequences have been used to identify host range [25], and here we leveraged the taxonomy of deposited plasmids in the NCBI database to estimate the replication host range of different plasmids based on observation of replicon, relaxase marker sequences and MOB-cluster codes in different taxa. We have added this feature into the MOB-suite for users to gain insight into the observed distribution of their plasmid of interest in the public data and combine that with detailed experimental knowledge where available. This information should be useful for hypothesis generation of plasmid host range, but its ability to be predictive of biological reality should be taken with caution, which is why we have paired it with experimental validation where it is available. As demonstrated here, MOB-cluster codes are conservative in their estimates of host range (Fig. 6); however, they are universally applicable to all sequenced plasmids, because they do not rely on any *a priori* marker scheme. The majority of the MOB-cluster codes are observed within a single species (Fig. 6), and so they are likely only applicable to epidemiological tracking rather than investigations into deeper evolutionary relationships.

MOB-suite v. 3.0.0 has been updated to include the improvements to the MOB-clustering algorithm along with the host range prediction. We have updated the MOB-suite to utilize a clustering threshold of 0.06 for primary cluster designation for plasmid reconstruction, host-range prediction and broader surveillance, along with a nested secondary cluster designation at 0.025 to recognize near duplicate plasmids. We selected the 0.025 threshold for near duplicates since any lower would result in draft plasmids potentially assigned to a difference cluster than a complete version of the plasmid. The current work represents a major change to the MOB-cluster approach and so a complete recalculation of the database was necessary, but the software supports incremental addition of new sequences to the database without changing existing cluster designations. The GitHub repository also contains the cluster designations and metadata for all plasmids that passed quality control (QC) from the NCBI, and will be updated as new public plasmid sequences become available. MOB-cluster codes are now represented in a five-character fixed-length accession code format similar to a GenBank accession number, i.e. AA001, since it permits writing very large numbers in a fixed-length code. A fixed-length accession code also enables validation of codes since a truncated code will be recognized as invalid.

In summary, the MOB-suite clustering approach has been designed to maximize concordance with replicon and relaxase clustering, and is applicable to all plasmid sequences since it does not have any reliance on the presence of specific marker sequences. The clusters are usually found within a single species due to the conservative nature of the clustering threshold selected. This aspect is suggestive that the cluster threshold employed is specific enough that it may be detecting species-specific plasmid variants and so can have applications for surveillance of transmission of plasmids within and between species. The MOB-suite has been updated to predict the replication host range of plasmids based on host taxonomy of plasmids deposited into the NCBI database and where possible to provide information from the literature on the host range. The cluster codes can be a useful tool for tracking plasmid transmission within an epidemiological context, and when combined with host ranges may be used to build relative risk models of transmission of plasmids that are of public-health concern due to the presence of AMR or virulence.

Funding information

This work was funded by the Public Health Agency of Canada, and the Government of Canada's Genomics R and D Initiative Phase VI Shared Priority Project Management Plan on Antimicrobial Resistance.

Acknowledgements

We thank Dr Anil Nichani, Dr Genevieve Labbe and Amanda Saunders for critical review of the manuscript, and the Bioinformatics core facility of the National Microbiology Laboratory, Public Health Agency of Canada, for computational infrastructure.

Author contributions

Conceptualization: J. R., J. H. E. N. Data curation: J. R., K. B. Formal analysis: J. R., K. B., J. S. Funding acquisition: J. H. E. N. Investigation: J. R., K. B., J. S. Methodology: J. R., K. B., J. S., J. H. E. N. Project administration: J. H. E. N. Resources: J. H. E. N. Software: J. R., K. B. Supervision: J. H. E. N. Writing – original draft: J. R., J. S. Writing – review and editing: J. R., K. B., J. S., J. H. E. N.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical statement

No ethics nor consent approval was required for the research in this study.

References

1. Shintani M, Sanchez ZK, Kimbara K. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Front Microbiol* 2015;6:242.
2. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B *et al*. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *J Antimicrob Chemother* 2018;73:1121–1137.
3. Couturier M, Bex F, Bergquist PL, Maas WK. Identification and classification of bacterial plasmids. *Microbiol Rev* 1988;52:375–395.
4. Garcillán-Barcia MP, Alvarado A, de la Cruz F. Identification of bacterial plasmids based on mobility and plasmid population biology. *FEMS Microbiol Rev* 2011;35:936–956.
5. Baker S, Hardy J, Sanderson KE, Quail M, Goodhead I *et al*. A novel linear plasmid mediates flagellar variation in *Salmonella* Typhi. *PLoS Pathog* 2007;3:e59.
6. Robertson J, Lin J, Wren-Hedgus A, Arya G, Carrillo C *et al*. Development of a multi-locus typing scheme for an Enterobacteriaceae linear plasmid that mediates inter-species transfer of flagella. *PLoS One* 2019;14:e0218638.
7. Aslam B, Wang W, Arshad MI, Khurshid M, Muzammil S *et al*. Antibiotic resistance: a rundown of a global crisis. *Infect Drug Resist* 2018;11:1645–1658.
8. Orlek A, Stoesser N, Anjum MF, Doumith M, Ellington MJ *et al*. Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. *Front Microbiol* 2017;8:182.
9. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O *et al*. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;58:3895–3903.
10. del Solar G, Giraldo R, Ruiz-Echevarría MJ, Espinosa M, Díaz-Orejas R. Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev* 1998;62:434–464.
11. Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL *et al*. Identification of plasmids by PCR-based replicon typing. *J Microbiol Methods* 2005;63:219–228.
12. Novick RP. Plasmid incompatibility. *Microbiol Rev* 1987;51:381–395.
13. Alvarado A, Garcillán-Barcia MP, de la Cruz F. A degenerate primer MOB typing (DPMT) method to classify gamma-proteobacterial plasmids in clinical and environmental settings. *PLoS One* 2012;7:e40438.
14. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 2018;4:e000206.
15. Philippe H, Douady CJ. Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol* 2003;6:498–505.
16. Pesesky MW, Tilley R, Beck DAC. Mosaic plasmids are abundant and unevenly distributed across prokaryotic taxa. *Plasmid* 2019;102:10–18.
17. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH *et al*. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132.
18. Chen W, Zhang CK, Cheng Y, Zhang S, Zhao H. A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS One* 2013;8:e70837.
19. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9:5114.
20. Jain A, Srivastava P. Broad host range plasmids. *FEMS Microbiol Lett* 2013;348:87–96.
21. Klümper U, Riber L, Dechesne A, Sannazzarro A, Hansen LH *et al*. Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *Isme J* 2015;9:934–945.
22. Datta N, Hedges RW. Host ranges of R factors. *J Gen Microbiol* 1972;70:453–460.

23. Yano H, Rogers LM, Knox MG, Heuer H, Smalla K *et al*. Host range diversification within the IncP-1 plasmid group. *Microbiology* 2013;159:2303–2315.
24. Suzuki H, Yano H, Brown CJ, Top EM. Predicting plasmid promiscuity based on genomic signature. *J Bacteriol* 2010;192:6045–6055.
25. Bradley P, den Bakker HC, Rocha EPC, McVean G, Iqbal Z. Ultrafast search of all deposited bacterial and viral genomic data. *Nat Biotechnol* 2019;37:152–159.
26. Orlek A, Phan H, Sheppard AE, Doumith M, Ellington M *et al*. Ordering the mob: insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid* 2017;91:42–52.
27. Popowska M, Krawczyk-Balska A. Broad-host-range IncP-1 plasmids and their resistance potential. *Front Microbiol* 2013;4:44.
28. Brooks LE, Kaze M, Sistrom M. Where the plasmids roam: large-scale sequence analysis reveals plasmids with large host ranges. *Microb Genom* 2019;5:e000244.
29. Zhang D, Zhao Y, Feng J, Hu L, Jiang X *et al*. Replicon-based typing of IncI-complex plasmids, and comparative genomics analysis of IncIγ/K1 plasmids. *Front Microbiol* 2019;10:48.
30. Murtagh F. A survey of recent advances in hierarchical clustering algorithms. *Comput J* 1983;26:354–359.
31. García-Fernández A, Villa L, Moodley A, Hasman H, Miriagou V *et al*. Multilocus sequence typing of IncN plasmids. *J Antimicrob Chemother* 2011;66:1987–1991.
32. Chen W, Fang T, Zhou X, Zhang D, Shi X *et al*. IncHI2 plasmids are predominant in antibiotic-resistant *Salmonella* isolates. *Front Microbiol* 2016;7:1566.
33. Hancock SJ, Phan M-D, Peters KM, Forde BM, Chong TM *et al*. Identification of IncA/C plasmid replication and maintenance genes and development of a plasmid multilocus sequence typing scheme. *Antimicrob Agents Chemother* 2017;61:e01740-16.
34. Octavia S, Wang Q, Tanaka MM, Sintchenko V, Lan R. Genomic variability of serial human isolates of *Salmonella enterica* serovar Typhimurium associated with prolonged carriage. *J Clin Microbiol* 2015;53:3507–3514.
35. Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A *et al*. Nested Russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene blaKPC. *Antimicrob Agents Chemother* 2016;60:3767–3778.