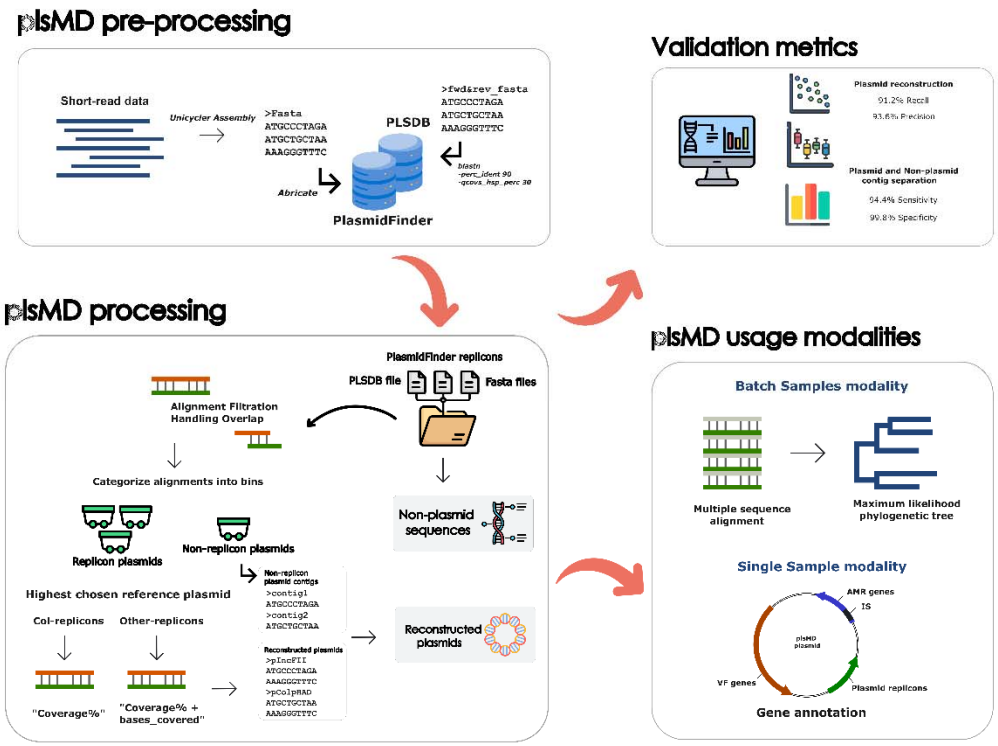
# plsMD: A plasmid reconstruction tool from short-read assemblies

Maryam Lotfi[1†], Deena Jalal[1†] and Ahmed A Sayed[1,2*]

1 Department of Basic Research, Genomics and Metagenomics Program, Children's Cancer Hospital Egypt 57357, Cairo, Egypt
2 Faculty of Science, Department of Biochemistry, Ain Shams University, Cairo, Egypt

† Maryam Lotfi and Deena Jalal contributed equally to this work
* Corresponding author: ahmad.sayed@57357.org




## Graphical abstract

plsMD uses Unicycler assemblies to identify plasmid replicons via PlasmidFinder and aligns sequences to PLSDB. It refines alignments, selects reference plasmids, and reconstructs full plasmid sequences. Circular plasmids without replicons are identified separately. plsMD outputs plasmid and non-plasmid FASTA files. Two workflows are supported: single-sample (plasmid/non-plasmid separation, annotation of AMR, VF, IS, and replicons) and batch-sample (grouping plasmids by replicon, MAFFT alignment, rotation, and phylogenetic tree construction). Validation showed 91.2% plasmid recall, 93.6% precision, 94.6% sensitivity, and 99.8% specificity.

## Abstract

While whole genome sequencing (WGS) has become a cornerstone of antimicrobial resistance (AMR) surveillance, the reconstruction of plasmid sequences from short-read WGS data remains a challenge due to repetitive sequences and assembly fragmentation. Current computational tools for plasmid identification and binning, such as PlasmidFinder, cBAR, PlasmidSPAdes, and Mob-recon, have limitations in reconstructing full plasmid sequences, hindering downstream analyses like phylogenetic studies and AMR gene tracking. To address this gap, we present plsMD, a novel tool designed for full plasmid reconstruction from short-read assemblies. plsMD integrates Unicycler assemblies with replicon and full plasmid sequence databases (PlasmidFinder and PLSDB) to guide plasmid reconstruction through a series of contig manipulations. Using a dataset of 83 samples containing 223 plasmids, plsMD outperformed existing tools, achieving excellent recall and precision of 91.2% and 93.6%, respectively, for plasmid reconstruction, and high sensitivity and specificity of 94.5% and 99.5% for plasmid-chromosomal contig separation. plsMD supports two usage modalities: single-sample analysis for plasmid reconstruction and gene annotation, and multi-sample analysis for phylogenetic investigations of plasmid transmission. This computational tool represents a significant advancement in plasmid analysis, offering a robust solution for utilizing existing short-read WGS data to study plasmid-mediated AMR spread and evolution.

## Keywords

plasmids, mobile genetic elements, DNA sequence analysis, bacterial genomes, plasmid reconstruction

## Key points

- Accurate plasmid reconstruction from short-read assemblies, surpassing existing binning-based tools.
- Replicon-guided approach enables detection of divergent plasmids, enhancing genomic annotation accuracy.
- Supports single- and multi-sample analysis, enabling plasmid transmission and evolutionary studies.

## Introduction

Antimicrobial resistance (AMR) is a critical and growing threat to clinical practice. AMR is primarily disseminated through AMR genes, which are frequently harbored on mobile genetic elements such as plasmids (1). Plasmids are extrachromosomal DNA elements capable of transferring between

organisms via mechanisms such as conjugation and transformation. They also replicate independently using specialized sequences called replicons. These replicons have historically served as key markers for plasmid identification and classification (2). Plasmids that share the same replication mechanism are unable to coexist long-term within the same cell (3). As a result, replicon sequences are used to categorize plasmids into distinct incompatibility groups and, labeled with the "Inc" prefix for incompatibility plasmids and the "Col" prefix for colicin-producing plasmids, which encode bactericidal toxins (3,4). Inc-type plasmids tend to be larger in size with several insertion sequences, whereas colicin plasmids are usually smaller and lack repetitive elements (4,5).

Since 2015, the World Health Organization (WHO) has recommended whole genome sequencing (WGS) as a critical tool for AMR surveillance and pathogen detection (6,7). This has led to its increasing adoption in routine clinical workflows. The vast majority of WGS data is generated using Illumina short-read sequencers, which produce highly accurate reads but face significant challenges during genome assembly (8). In particular, repetitive sequences—commonly found in plasmid DNA, especially near AMR gene cassettes—frequently cause contig breaks during assembly (9). Classical plasmid sequencing approaches rely on the separation of plasmid DNA from chromosomal DNA prior to sequencing. However, these methods struggle to capture large plasmids exceeding 50 kb. To address this, newer methods aim to extract plasmid sequences directly from WGS data (10).

Resolving plasmid sequences from short-read sequencing data has been a persistent challenge (11), yet it remains critically important for understanding AMR gene dissemination and inter-organism gene transfer. Several computational tools have been developed to address this issue. PlasmidFinder relies on a database of replicon sequences to identify and classify contigs as plasmid-derived (10). However, this approach struggles with large plasmids, which often harbor repetitive sequences that cause contig breaks, resulting in only partial plasmid identification (11). cBAR employs a *k*-mer based methodology, using pentamer frequencies to classify contigs as either plasmid or chromosomal, but it does not attempt to bin sequences into distinct plasmid compartments (12). PlasmidSPAdes and Recycler take advantage of assembly graphs derived from De Bruijn graphs (13,14). Recycler searches these graphs for subgraphs that represent distinct plasmids, while PlasmidSPAdes uses depth and circularity characteristics to predict plasmid sequences and group them into bins. Finally, MOB-recon combines in-house databases for complete reference plasmids and plasmid-associated sequences, such as relaxases, to guide the separation of plasmid and chromosomal contigs (15).

While some of these tools perform well in plasmid binning and separating chromosomal contigs, none are designed to reconstruct full plasmid sequences. This limitation hinders key downstream analyses, such as phylogenetic studies, tracking plasmid evolution, and understanding the mechanisms underlying AMR gene spread. Full plasmid reconstruction is essential for identifying structural variations, mapping plasmid gene content, and characterizing plasmid-mediated AMR dissemination.

In this study, we present plsMD (https://github.com/Genomics-and-Metagenomics-unit-57357/plsMD), a tool designed for full plasmid reconstruction beyond contig binning from short-read sequence assemblies. This tool utilizes Unicycler assemblies together with established replicon and full plasmid sequence databases to guide plasmid reconstruction. Through a series of contig sequence manipulations, we reconstruct full plasmid sequences. Using a dataset of 83 samples with 223 plasmids, we benchmarked plsMD against two plasmid binning tools, MOB-recon and PlasmidSPAdes, achieving superior accuracy and precision of 91.2% and 93.6%, respectively. For the separation of plasmid and chromosomal contigs, plsMD demonstrated sensitivity and specificity of 94.7% and 99.8%, respectively. We propose two usage modalities: single-sample analysis, which allows plasmid reconstruction and separation of chromosomal contigs, followed by gene annotation using relevant bioinformatics tools; and multi-sample analysis, which enables the investigation of multiple samples by grouping reconstructed plasmids with the same replicons and performing phylogenetic analysis to gain insights into plasmid relatedness and transmission between organisms.

## Materials and Methods

### Annotation and Preprocessing

The input for plsMD consists of assembled shotgun sequencing data processed using Unicycler (16). To account for both DNA strand orientations, all resulting contigs are reversed and merged with their forward counterparts into a combined FASTA file. plsMD integrates replicon and full plasmid sequences databases, PlasmidFinder and PLSDB respectively (10,17,18). First, it runs Abricate (https://github.com/tseemann/abricate) on the Unicycler assembly using PlasmidFinder database to identify plasmid-associated replicons and the contigs they are harbored on. To capture alignments on both forward and reverse orientations, the combined FASTA is aligned to PLSDB database using BLASTn (19), parameters are optimized for plasmid detection (-perc_identity 90 -qcov_hsp_perc 30). The PLSDB output is modulated to include; the reference plasmid (*sseqid*), the length of the *sseqid* (*slen*), coordinates of the *sseqid* alignments (*sstart* and *send*), the contig aligned (*qseqid*), coordinates of the

*qseqid* alignments (*qstart* and *qend*), alignment metrics (*pident*, *evalue*, *bitscore*) and the percentage of alignment relevant to length of contig (*qcovs*). Plasmid names in the abricate output files are modified, and a suffix is appended to account for multiple occurrences of the same plasmid replicon.

## Alignment Refinement

### Filtration

To ensure final plasmid assembly in the reference plasmid orientation, plsMD filters out alignments where *sstart* is greater than *send* position, as the proper orientation will be inherently covered by the reverse-contigs generated during preprocessing. Contigs that are identified as circular in the original unicycler assembly (circular = true in header), are marked for circularity. For each *sseqid* alignment, if it includes circular contigs along with other contigs, they are filtered out entirely and this *sseqid* is excluded from use as a reference plasmid.

### Handling overlap

To ensure non-redundant alignments and proper reconstruction of resulting plasmids, plsMD identifies and handles nested and overlapping contigs rigorously. Nested contig alignments, which are entirely contained within larger contigs (relative to the subject), are excluded from plasmid reconstruction, since they provide no additional information beyond the larger contig that contains them. Nested contigs are also flagged for removal from non-plasmid contigs. Partially overlapping contigs are handled by calculating the overlap between contigs using *sseqid* alignment coordinates and the number of overlapped bases to be used for trimming from contigs during reconstruction. Additionally, *send* and *qend* are updated for proper downstream calculations. For repeated contigs with overlapping *qseqid* coordinates, the alignments are retained only if *qcovs* is more than 80.

## Reference Plasmid identification for reconstruction

### Coverage Calculations

Coverage calculation determines the percentage of the *sseqid* covered by the query alignment. This involves merging overlapping *sseqid* alignment intervals and calculating the total length of the covered region (*bases_covered*). The coverage percentage is then calculated using the formula:

$$coverage_percentage = \frac{bases_covered}{slen} * 100$$

### Reference plasmid choice

Bins are created for each replicon in a sample, with only plasmid alignments containing a specific replicon contig binned together for reference plasmid selection. To choose the highest scoring reference plasmid for each reconstructed plasmid, plsMD uses two different scoring approaches depending on the replicon type: ***(i) Col replicons;*** the reference plasmid (*sseqid*) is chosen based on the highest coverage percentage as Col-replicon plasmids are typically smaller and have few or no insertion sequences; ***(ii) Inc/Other/Col-hybrid replicons;*** reference plasmid choice is based on a combined scoring of both coverage percentage and bases covered, to allow larger plasmid capturing as these plasmids commonly contain repetitive elements. To prioritize plasmids with high coverage, a progressive coverage filter is applied, starting at 80% and decreasing in 10% increments if no plasmids meet the threshold, with no lower limit, ensuring the detection of all possible replicon plasmids.

### Redundant plasmid removal

All chosen reference plasmids across all replicons, with their corresponding alignments are concatenated together and checked for redundancy as follows; (i) If the same reference plasmid (*sseqid*) is chosen across different replicon bins, only one instance is included in the final reconstructed plasmid file for each sample. (ii) For hybrid Col-Inc plasmids, the chosen plasmid in the Col replicon bin is excluded, as the selected Inc plasmid would be more comprehensive in its selection of reference plasmid.

## plsMD output sequences

### Plasmid sequences

#### (i) Reconstructed Replicon plasmids

Reconstruction of the plasmid is done by extracting the contig IDs (*qseqids*) for the chosen plasmid from the combined FASTA files. Overlapping regions in the contigs are trimmed, and the contigs are merged into a single sequence for each plasmid. Naming of the reconstructed plasmid is based on the created replicon bins.

#### (ii) Non-replicon plasmids

plsMD relies on the presence of replicons in plasmid sequences. However, to enhance inclusivity, non-replicon-containing plasmids can still be identified as plasmids under specific conditions. A plasmid is considered valid if its reference plasmid alignments (*sseqids*) consist of a single circular

contig with 100% coverage_percentage of the *sseqid* and *qcovs* = 100. These contigs are directly included in the plasmid contig sequences without further refinement.

### Non-plasmid sequences

Contigs included in plasmid sequences, regardless of orientation (forward or reverse), are excluded from assembled FASTA files to generate non-plasmid contigs. Additionally, nested contigs that align to reference plasmids but were excluded from plasmid assembly due to redundancy are also removed from the non-plasmid contigs.

## Usage Modalities

### Single Sample modality – Annotation

The separation of plasmid and non-plasmid sequences enables the annotation of various genetic elements, including AMR genes, virulence factor (VF) genes, insertion sequences, and plasmid replicons. Annotation is performed using Abricate and AMRFinder (20) for AMR gene identification, VFDB (21) for virulence factor genes, ISfinder (22) for insertion sequences, and PlasmidFinder (10) for plasmid replicons.

### Batch modality – Phylogenetic Analysis

For the purpose of investigating phylogenetic relationships between identified plasmids in different samples, reconstructed plasmids from different samples with same replicons (in the same replicon bin) are grouped together in one multi-fasta file. Prior to multiple sequence alignment, plasmid sequences must be rotated to a common start point to resolve inconsistencies caused by arbitrary starting positions in circular plasmid sequences. This is achieved by identifying a shared reference common sequence, adjusting plasmid sequences to start from this point, and reversing them if necessary. Once rotated, the sequences are aligned using MAFFT (23), which automatically determines the optimal alignment strategy. Finally, a maximum likelihood phylogenetic tree is generated using IQ-TREE (24).

## Results

## Dataset and tools

To evaluate plsMD, we compared it with two other commonly used plasmid binning tools, Mob-recon (15) and PlasmidSPAdes (13). We used the dataset compiled by Robertson and Nash (15), which

originally consisted of 103 closed genomes and 377 associated plasmids sequenced using PacBio and Illumina. To exclude the samples with potential discrepancies between the uploaded complete genomes and uploaded Illumina FASTQ files, we removed those with differing replicon genes identified in uploaded complete genomes and Unicycler assembly. This resulted in 83 samples and 223 associated plasmids being included in our validation.

## Evaluation metrics

Illumina short-reads were assembled using unicycler (v0.5.0) with default parameters. Reference genomes and plasmids for each sample were used to build BLAST databases (*plasmid_db* and *chr_db*) to use for downstream alignments with BLASTn using cut off (-perc_identity 90).

### Plasmid Reconstruction Assessment: Recall and Precision

For each sample, each reconstructed plasmid was separately aligned against its corresponding *plasmid_db*. For each alignment, *bases_covered* and *coverage_percentage* were calculated as previously described. The total number of bases in each reconstructed plasmid FASTA file was determined (*sum_qlen*), where plsMD produces a single FASTA per reconstructed plasmid, whereas Mob-recon and PlasmidSPAdes generate multi-FASTA contig bins. The number of bases aligned in each reconstructed plasmid FASTA relative to each sseqid in the *plasmid_db* was computed (*sum_qseqid_aligned*), and *q_perc_cov* was calculated as:

$$q_perc_cov = \frac{sum_qseqid_aligned}{sum_qlen} * 100$$

Pairs of qseqid and sseqid alignments were extracted along with their associated metrics: *bases_covered*, *coverage_percentage*, *sum_qseqid_aligned*, and *q_perc_cov*.

To evaluate plasmid reconstruction accuracy, the Hungarian algorithm was implemented to find the optimal matching between query and subject sequences, ensuring that each qseqid and sseqid were matched only once. The algorithm maximized the total *coverage_percentage* (recall) by assigning each query to its best-matching subject, constructing a weight matrix based on *coverage_percentage* values. Query sequences without a match were assigned *q_perc_cov* = 0, while subject sequences without a query match were assigned *coverage_percentage* = 0.

The recall of each plasmid was represented by the *coverage_percentage* chosen by the matching algorithm, while the precision of each reconstructed plasmid was determined by the

*q_perc_cov* assigned by the algorithm. Average recall for each tool was calculated as the mean *coverage_percentage* across all 223 plasmids, while average precision was computed as the mean *q_perc_cov* across all reconstructed plasmids for each tool.

**Plasmid Contig Classification: Sensitivity and Specificity**

For each sample, all reconstructed plasmids and non-plasmid contigs were aligned against the corresponding *plasmid_db* and *chr_db*. The number of bases aligned for qseqids when aligned to either database was computed (*sum_qseqid_aligned*). The *sum_qseqid_aligned* from aligning reconstructed plasmids against *plasmid_db* represented True Positives (TP), while alignment against *chr_db* represented False Positives (FP). Similarly, *sum_qseqid_aligned* from aligning non-plasmid sequences against *plasmid_db* represented False Negatives (FN), while alignment against *chr_db* represented True Negatives (TN). Sensitivity and Specificity were calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN} * 100$$

$$Specificity = \frac{TN}{TN + FP} * 100$$

## Validation Results

We compared the performance of plsMD with the two established plasmid binning tools; MOB-recon and PlasmidSPAdes. Tools primarily designed for separating plasmid sequences from metagenomic data, such as cBAR and Recycler, were excluded. Other tools were also omitted as they consistently underperformed in plasmid reconstruction and plasmid-contig separation compared to MOB-recon (11,25). Recall (the percentage of the reference plasmid covered by the reconstructed plasmid) and precision (the percentage of the reconstructed plasmid covered by the reference) were calculated for each tool and summarized in Supplementary Table 2 and Figure 2a. Out of the 223 plasmids to be reconstructed, plsMD achieved the highest overall recall (91.22%), outperforming MOB-recon (87.39%) and PlasmidSPAdes, which had a significantly lower recall of 55.05%. However, plsMD failed to reconstruct 12 plasmids, whereas MOB-recon failed to reconstruct only 7. For precision, plsMD again outperformed the other tools, achieving 93.59%, compared to 86.68% for MOB-recon and 48.53% for PlasmidSPAdes. Additionally, plsMD produced only 8 extra plasmids, demonstrating superior accuracy, whereas MOB-recon generated 25 extra plasmids, and PlasmidSPAdes produced 70. Plasmid recall by plsMD remained consistently high across all plasmid size categories—small (<10 kbp), medium

(10–50 kbp), and large (>50 kbp). In contrast, MOB-recon exhibited reduced recall for larger plasmids (Figure 2b). The replicon types of different plasmids are color-coded to highlight size differences, with Col-replicon plasmids generally being smaller and Other-type replicon plasmids tending to be larger. Despite the attempt to reconstruct non-replicon plasmids, plsMD only effectively captures non-replicon plasmids with smaller sizes (Figure 2b).

For assessing contig classification to plasmid and non-plasmids compartments, we calculated sensitivity (percentage of correctly classified plasmid contigs) and specificity (percentage of correctly classified non-plasmid contigs) for each of the 83 samples. PlasmidSPAdes was excluded from this analysis because, although it separates plasmid compartments, it does not generate non-plasmid contigs, preventing the assessment of true and false negatives. plsMD and MOB-recon performed similarly in both sensitivity and specificity; 94.74% and 99.79% for plsMD, and 96.50% and 99.81% for MOB-recon (Supplementary Table 2 and Figure 2c).

## Example usage

The strength of plsMD lies in its ability to fully reconstruct and assemble plasmid sequences—an approach not attempted by any other available tool. As an example, we demonstrate the batch mode of plsMD for analyzing the phylogenetic relationships of plasmids carrying the IncFII(pAR0022) replicon among the 83 validation samples. This replicon was selected for phylogenetic analysis due to its high prevalence, with 27 plasmids identified across 26 samples. A common sequence was identified to standardize the starting point for rotation. This was followed by multiple sequence alignment using MAFFT v7.525, maximum likelihood phylogenetic tree construction with IQ-TREE v2.3.6, and visualization in iTOL version 7. The phylogenetic trees generated for both the complete reference plasmids harboring IncFII(pAR0022) replicon as well as those generated from plsMD are visualized in figure 3a. The high similarity between these trees, with minimal changes in clustered plasmids, highlights the accuracy of plsMD's reconstructions. Furthermore, our analysis demonstrates the ability of plsMD to differentiate plasmids from distinct organisms and identify highly similar plasmids, suggesting potential inter-organism plasmid transfer.

Additionally, full plasmid reconstruction enables precise genomic annotation, preserving the exact order of genomic features (Figure 3b). This allows for the identification of mobile genetic elements as well as AMR and VF genes capable of co-transferring, further emphasizing the advantages of using plsMD for plasmid analysis.

## Discussion

With the rapid advancements in long-read sequencing technologies and the development of newer sequencing techniques, the future of plasmid identification and sequencing is promising. However, for the vast amount of existing short-read sequencing data, current tools have yet to achieve the precision and reliability needed to accurately extract plasmid sequences. Plasmid identification and binning tools generally rely on three main approaches. Assembly graph inspection methods, such as PlasmidSPAdes and Recycler, search for sub-graphs within assembly graphs. However, these tools typically score low on precision and recall and often fail to classify large plasmids with highly repetitive sequences due to the fragmented and ambiguous assembly graphs they produce (11). Reference-based approaches, including MOB-recon and PlasmidFinder, depend on precompiled databases for plasmid identification. While these methods can be effective, they inherently struggle to detect novel plasmids, limiting their broader applicability. Algorithm-based methods, such as cBAR and PlasmidHunter, use sequence-specific features (e.g., GC content or *k*-mer frequencies) to identify plasmids. These approaches perform well in metagenomic contexts (12,26) but generally do not attempt plasmid binning, as they cannot fully resolve sequences from different plasmids based solely on sequence characteristics. Hybrid approaches, such as plasBin-flow, attempt to combine these methods to improve performance (25). While assembly graph-based methods struggle with classification accuracy and algorithm-based approaches excel in metagenomic applications but lack binning capabilities, reference-based strategies hold promise if applied effectively—despite their limitation in identifying previously uncharacterized plasmids. plsMD, while being a reference-based tool, allows the reconstruction of plasmids around replicon genes, even with low similarity to reference plasmids. This allows the identification of highly divergent plasmids effectively.

The performance of plsMD is inherently influenced by the quality of sequencing reads, assembly accuracy, and the comprehensiveness of the reference database (PLSDB). Despite these dependencies, plsMD achieves superior recall and precision compared to existing tools, with few instances of non-reconstructed and extra plasmids. Unlike MOB-recon, which relies on plasmid-associated genes such as relaxases for plasmid construction (15), plsMD primarily utilizes replicons—a more established and plasmid-specific feature—as the foundation for plasmid reconstruction. This approach, combined with its ability to avoid over-reliance on circularity (which can occasionally result from spurious assembly graph artifacts, particularly in the presence of long repetitive sequences (16,27), contributes to plsMD's significantly higher precision and lower number of extra plasmids. Furthermore, to accurately capture

plasmids of varying sizes, plsMD does not treat all plasmids equally. We recognize that the key factor affecting the success of Illumina-based assemblies is the presence of repetitive sequences, which frequently cause contig breaks. By distinguishing between low-repetitive sequence plasmids, such as Col replicon plasmids, and those with highly repetitive sequences, plsMD enables the complete and accurate assembly of plasmids across a wide size range.

Overall, plsMD uses established databases and tools in a streamlined workflow to address a critical and challenging problem—without relying on complex algorithms or methodologies. While plasmid binning may be sufficient in a range of application, by providing fully assembled and polished plasmid sequences, plsMD opens up opportunities for a broad range of analyses and insights that otherwise would not be possible.

**Funding information**

This project was funded by the Children's Cancer Hospital Egypt CCHE 57357.

**Conflicts of interest**

The authors declare that there are no conflicts of interest.

**Data Availability**

plsMD is implemented in Bash and Python3, which can be downloaded from the following repositories: https://doi.org/10.6084/m9.figshare.28465241.v1 and https://github.com/Genomics-and-Metagenomics-Unit-57357/plsMD/.

## Figures and Figure Legends

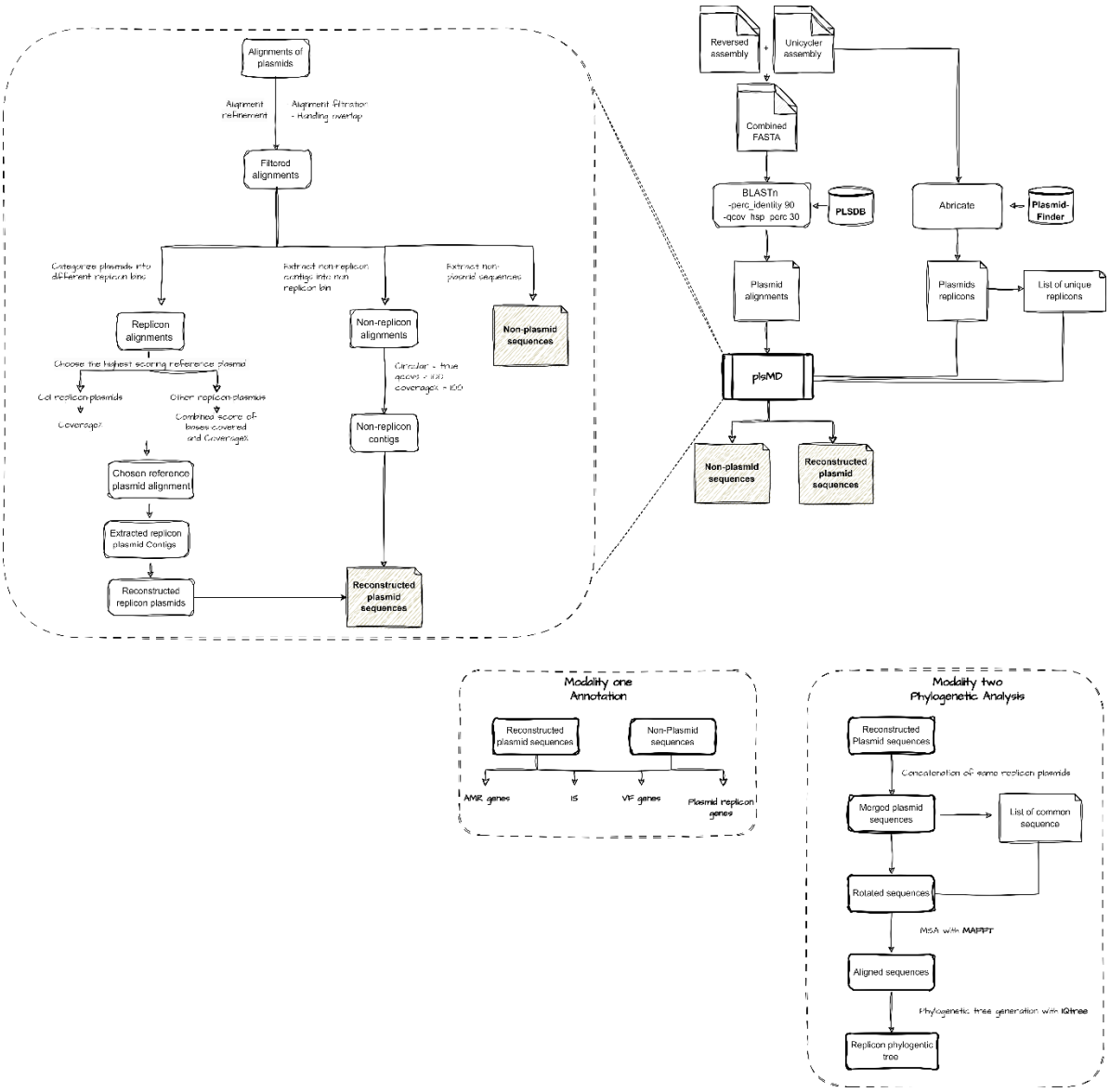



**Figure 1: Flowchart of the plsMD workflow.** Unicycler assemblies serve as the input for plsMD, which integrates PlasmidFinder and PLSDB databases to identify plasmid replicons in a sample and align sequences to the reference database, respectively. This is followed by alignment refinements, overlap handling, and the selection of a reference plasmid based on the replicon present. The chosen reference plasmid then guides the reconstruction of a full plasmid sequence. Circular, non-replicon-containing plasmids are identified separately if they align to PLSDB with qcovs = 100 and 100% coverage. plsMD outputs two FASTA files: one for plasmid sequences and another for non-plasmid sequences. The tool can be used in two workflows: single-sample and batch-sample. In the single-sample workflow, reconstructed plasmid sequences and non-plasmid contigs are separated, followed by annotation of AMR genes, virulence factor (VF) genes, insertion sequences (ISs), and plasmid replicons. In the batch-sample workflow, plasmids with the same replicon across multiple samples are grouped, aligned using MAFFT, and rotated to a common sequence. Finally, a phylogenetic tree is constructed.

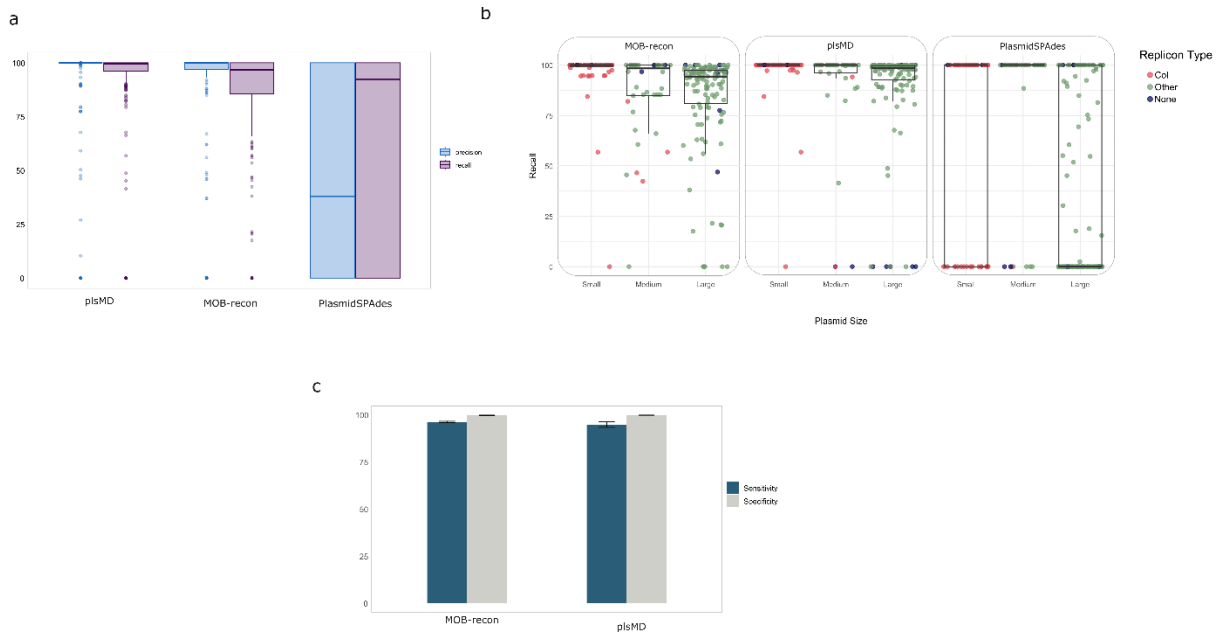



**Figure 2: Performance of plsMD and other tools in plasmid reconstruction and contig classification.** (a) Box plot showing recall and precision for plsMD, MOB-recon, and PlasmidSPAdes across all 223 plasmids in the dataset. Recall is defined as the percentage of the reference plasmid sequence covered by the reconstructed plasmid, while precision is the percentage of the reconstructed plasmid sequence aligning to a reference plasmid. (b) Box plot displaying recall for plsMD, MOB-recon, and PlasmidSPAdes across different plasmid size categories—small (<10 kbp), medium (10–50 kbp), and large (>50 kbp). Plasmid replicon types are color-coded to highlight performance variations across different replicon types. (c) Bar plot illustrating the sensitivity and specificity of plsMD and MOB-recon in classifying plasmid and non-plasmid contigs. Sensitivity is calculated as the percentage of correctly identified plasmid bases relative to the total bases classified as plasmid sequences. Specificity is calculated as the percentage of correctly identified non-plasmid bases relative to the total bases classified as non-plasmid sequences.

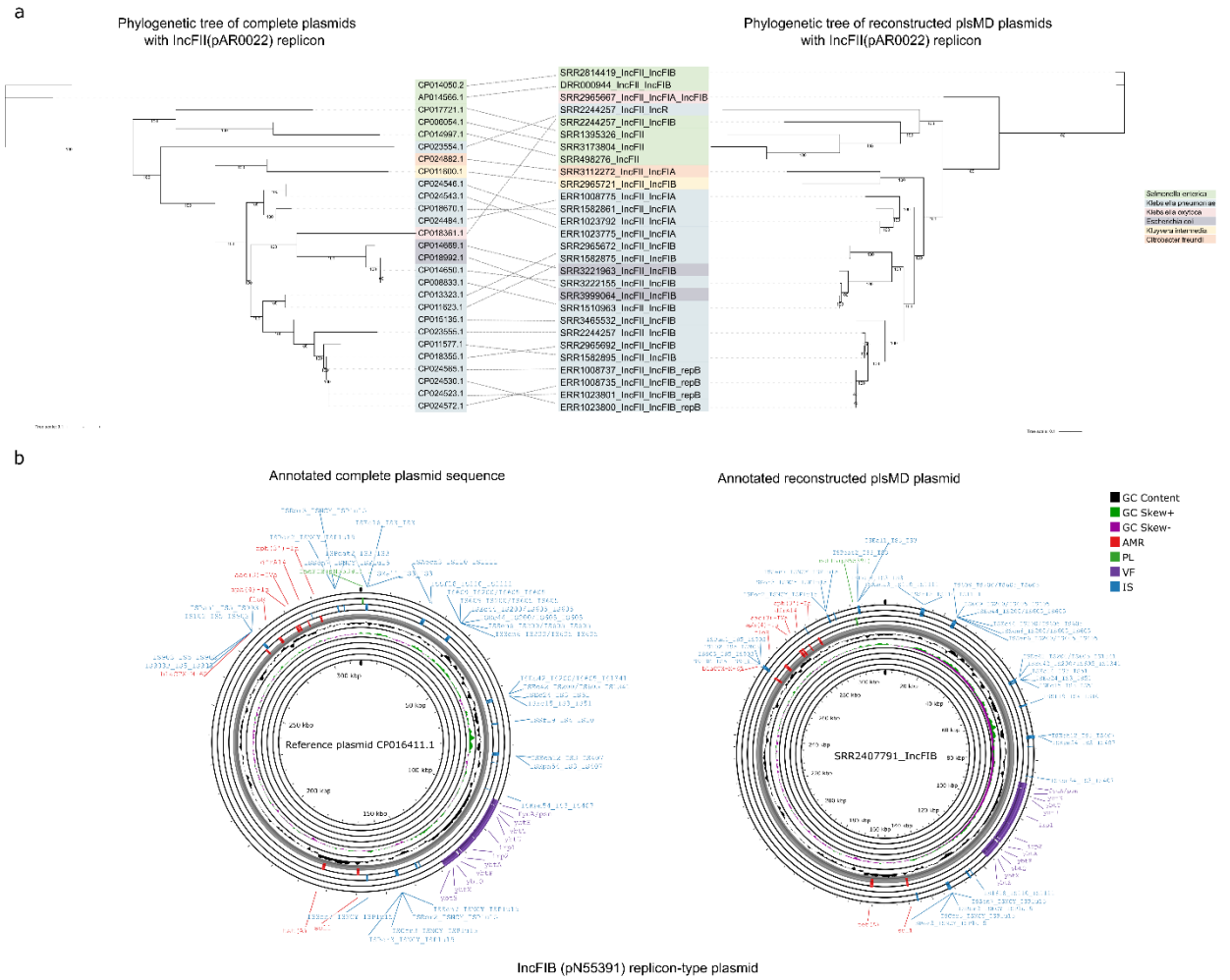



**Figure 3: Example Demonstrating the Advantages of Full Plasmid Reconstruction with plsMD.** (a) Maximum likelihood phylogenetic tree of 27 closed-sequence plasmids harboring the IncFII(pARR002) replicon, shown alongside a tree of 28 reconstructed plasmids harboring the same replicon, generated using plsMD. Phylogenetic trees were constructed using iTOL (29). Plasmid origins are color-coded based on the host organism type, highlighting the ability to differentiate plasmids from different sources. (b) Comparison of a fully closed plasmid harboring the IncFIB(pN55391) replicon with its plsMD-reconstructed counterpart from the same sample, illustrating their similarities. Both sequences are annotated for plasmid replicons, antimicrobial resistance (AMR) genes, virulence factor (VF) genes, and insertion sequences (ISs).

## References


1. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. Clin Microbiol Rev [Internet]. 2018 Oct 1 [cited 2025 Feb 9];31(4):e00088-17. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6148190/

2. Helinski DR. A Brief History of Plasmids. EcoSal Plus [Internet]. 2022 Dec 15 [cited 2025 Feb 9];10(1):eESP-0028-2021. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10729939/

3. Novick RP. Plasmid incompatibility. Microbiol Rev [Internet]. 1987 Dec [cited 2025 Feb 9];51(4):381–95. Available from: https://pubmed.ncbi.nlm.nih.gov/3325793/

4. Cascales E, Buchanan SK, Duché D, Kleanthous C, Lloubès R, Postle K, et al. Colicin Biology. Microbiol Mol Biol Rev [Internet]. 2007 Mar [cited 2025 Feb 9];71(1):158. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC1847374/

5. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B, Guerra B, et al. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. J Antimicrob Chemother [Internet]. 2018 May 1 [cited 2025 Feb 9];73(5):1121–37. Available from: https://dx.doi.org/10.1093/jac/dkx488

6. GLASS Whole-genome sequencing for surveillance of antimicrobial resistance Global Antimicrobial Resistance and Use Surveillance System (GLASS).

7. Baker KS, Jauneikaite E, Hopkins KL, Lo SW, Sánchez-Busó L, Getino M, et al. Genomics for public health and international surveillance of antimicrobial resistance. The Lancet Microbe [Internet]. 2023 Dec 1 [cited 2025 Feb 9];4(12):e1047–55. Available from: http://www.thelancet.com/article/S2666524723002835/fulltext

8. El-Metwally S, Hamza T, Zakaria M, Helmy M. Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. PLoS Comput Biol [Internet]. 2013 [cited 2025 Feb 9];9(12):e1003345. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3861042/

9. Abramova A, Karkman A, Bengtsson-Palme J. Metagenomic assemblies tend to break around antibiotic resistance genes. BMC Genomics 2024 251 [Internet]. 2024 Oct 14 [cited 2025 Feb 9];25(1):1–17. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-024-10876-0

10. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, et al. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58(7):3895–903.

11. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microb Genomics [Internet]. 2017 Oct 1 [cited 2025 Feb 6];3(10):1–8. Available from: https://pubmed.ncbi.nlm.nih.gov/29177087/

12. Zhou F, Xu Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics [Internet]. 2010 Aug 2 [cited 2025 Feb 9];26(16):2051–2. Available from: https://pubmed.ncbi.nlm.nih.gov/20538725/

13. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. plasmidSPAdes: assembling

plasmids from whole genome sequencing data. Bioinformatics [Internet]. 2016 Nov 15 [cited 2025 Feb 9];32(22):3380–7. Available from: https://dx.doi.org/10.1093/bioinformatics/btw493

14. Rozov R, Kav AB, Bogumil D, Shterzer N, Halperin E, Mizrahi I, et al. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. Bioinformatics [Internet]. 2016 Feb 15 [cited 2025 Feb 9];33(4):475. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC5408804/

15. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb genomics [Internet]. 2018 Aug 1 [cited 2025 Feb 6];4(8). Available from: https://pubmed.ncbi.nlm.nih.gov/30052170/

16. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13(6):1–22.

17. Galata V, Fehlmann T, Backes C, Keller A. PLSDB: a resource of complete bacterial plasmids. Nucleic Acids Res [Internet]. 2019 [cited 2024 Sep 1];47(D1):D195–202. Available from: https://pubmed.ncbi.nlm.nih.gov/30380090/

18. Schmartz GP, Hartung A, Hirsch P, Kern F, Fehlmann T, Müller R, et al. PLSDB: advancing a comprehensive database of bacterial plasmids. Nucleic Acids Res [Internet]. 2022 Jan 7 [cited 2024 Sep 1];50(D1):D273–8. Available from: https://dx.doi.org/10.1093/nar/gkab1111

19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct;215(3):403–10.

20. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep [Internet]. 2021;11(1):1–9. Available from: https://doi.org/10.1038/s41598-021-91456-0

21. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: Hierarchical and refined dataset for big data analysis - 10 years on. Nucleic Acids Res. 2016;44(D1):D694–7.

22. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006 Jan;34(Database issue):D32-6.

23. Katoh K, Misawa K, Kuma K ichi, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul;30(14):3059–66.

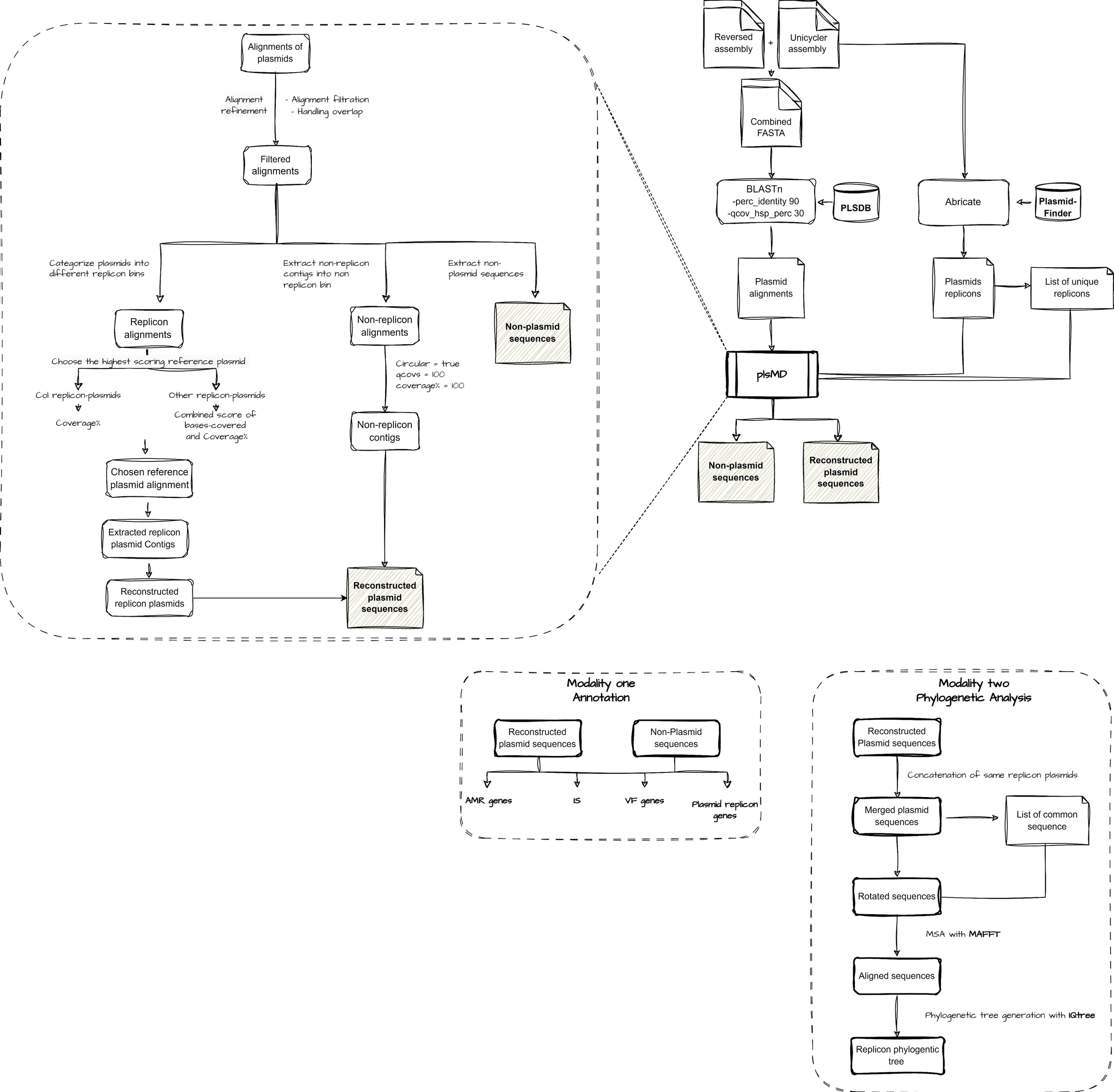
24. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020 May;37(5):1530–4.

25. Mane A, Faizrahnemoon M, Vinař T, Brejová B, Chauve C. PlasBin-flow: a flow-based MILP algorithm for plasmid contigs binning. Bioinformatics [Internet]. 2023 Jun 30 [cited 2025 Feb 9];39(Supplement_1):i288–96. Available from: https://dx.doi.org/10.1093/bioinformatics/btad250

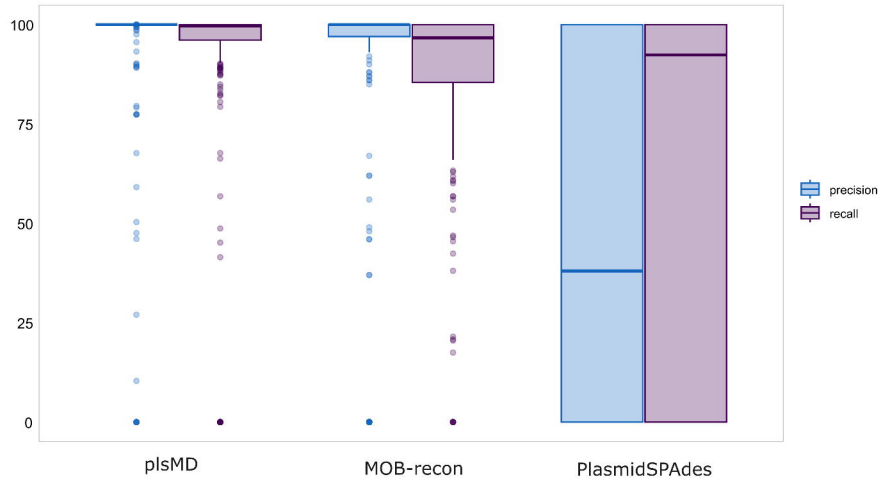
26. Tian R, Zhou J, Imanian B. PlasmidHunter: accurate and fast prediction of plasmid sequences using gene content profile and machine learning. Brief Bioinform [Internet]. 2024 May 23 [cited 2025 Feb 9];25(4). Available from: https://dx.doi.org/10.1093/bib/bbae322

27. Schäfer L, Jehle JA, Kleespies RG, Wennmann JT. A practical guide and Galaxy workflow to avoid

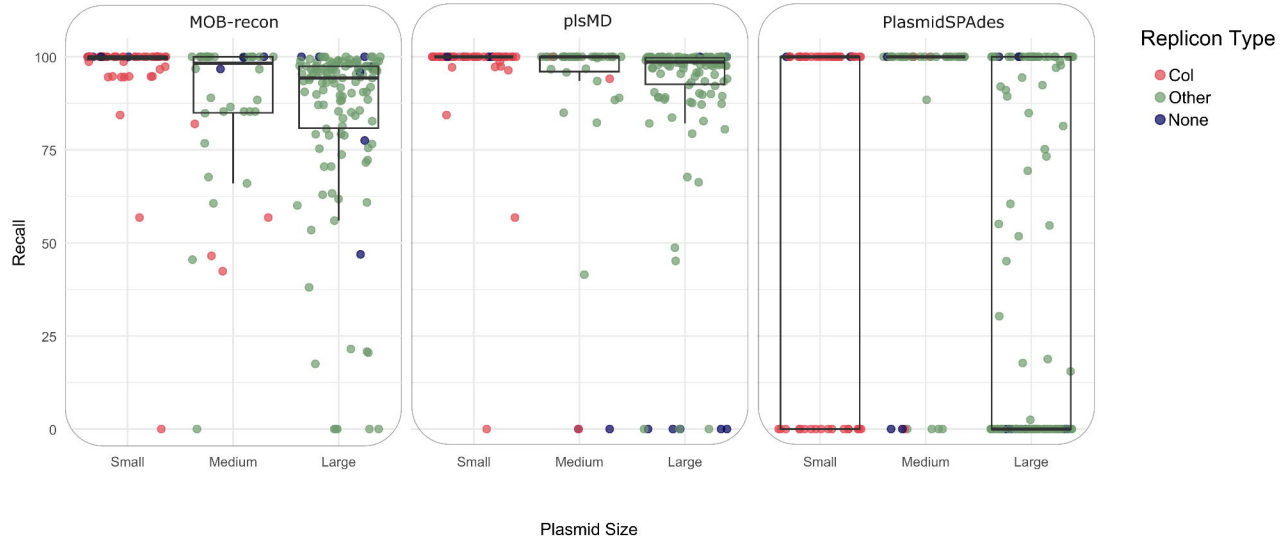
inter-plasmidic repeat collapse and false gene loss in Unicycler's hybrid assemblies. Microb genomics [Internet]. 2024 Jan 1 [cited 2025 Feb 6];10(1). Available from: https://pubmed.ncbi.nlm.nih.gov/38197876/


Alignments of plasmids
Alignment refinement
- Alignment filtration
- Handling overlap
Filtered alignments
Categorize plasmids into different replicon bins
Extract non-replicon contigs into non replicon bin
Extract non-plasmid sequences
Replicon alignments
Non-replicon alignments
Non-plasmid sequences
Choose the highest scoring reference plasmid
Col replicon-plasmids
Coverage%
Other replicon-plasmids
Combined score of bases-covered and Coverage%
Circular = true
qcovs = 100
coverage% = 100
Chosen reference plasmid alignment
Non-replicon contigs
Extracted replicon plasmid Contigs
Reconstructed replicon plasmids
Reconstructed plasmid sequences
Reversed assembly
+
Unicycler assembly
Combined FASTA
BLASTn
-perc_identity 90
-qcov_hsp_perc 30
PLSDB
Abricate
Plasmid-Finder
Plasmid alignments
Plasmids replicons
List of unique replicons
plsMD
Non-plasmid sequences
Reconstructed plasmid sequences
Modality one
Annotation
Reconstructed plasmid sequences
Non-Plasmid sequences
AMR genes
IS
VF genes
Plasmid replicon genes
Modality two
Phylogenetic Analysis
Reconstructed Plasmid sequences
Concatenation of same replicon plasmids
Merged plasmid sequences
List of common sequence
Rotated sequences
MSA with MAFFT
Aligned sequences
Phylogenetic tree generation with IQtree
Replicon phylogentic tree

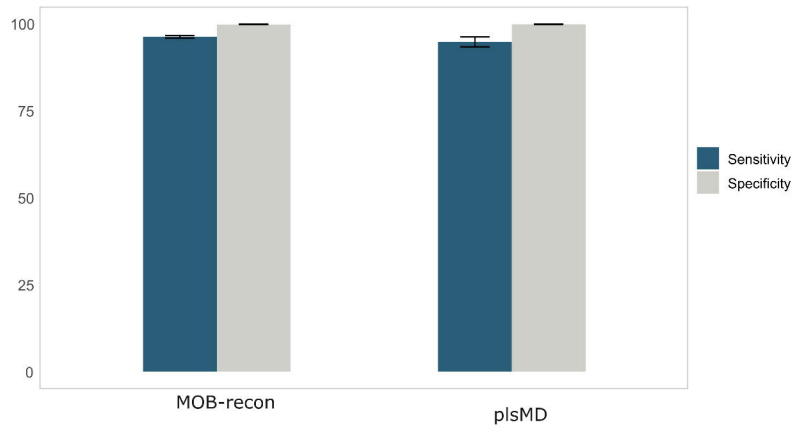
a


100
75
50
25
0
plsMD
MOB-recon
PlasmidSPAdes
precision
recall

b


MOB-recon
plsMD
PlasmidSPAdes
Recall
100
75
50
25
0
Small
Medium
Large
Plasmid Size
Replicon Type
Col
Other
None

c


100
75
50
25
0
MOB-recon
plsMD
Sensitivity
Specificity

a

Phylogenetic tree of complete plasmids with IncFII(pAR0022) replicon

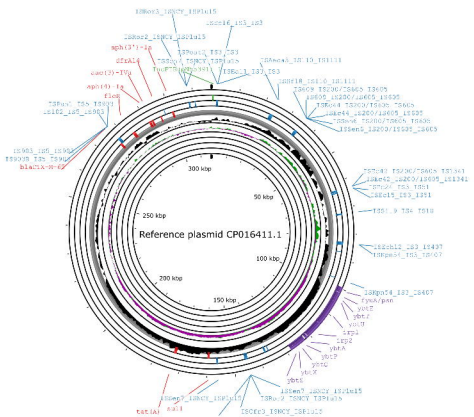
Phylogenetic tree of reconstructed plsMD plasmids with IncFII(pAR0022) replicon

| | |
|---|---|
| CP014050.2 | SRR2814419_IncFII_IncFIB |
| AP014566.1 | DRR000944_IncFII_IncFIB |
| CP017721.1 | SRR2965667_IncFII_IncFIA_IncFIB |
| CP006054.1 | SRR2244257_IncFII_IncFII |
| CP014997.1 | SRR2244257_IncFII_IncFII |
| CP023554.1 | SRR1395326_IncFII |
| CP024882.1 | SRR3173804_IncFII |
| CP011600.1 | SRR498276_IncFII |
| CP024546.1 | SRR3112272_IncFII_IncFIA |
| CP024543.1 | SRR2965721_IncFII_IncFIB |
| CP018670.1 | ERR1008775_IncFII_IncFIA |
| CP024484.1 | SRR1582861_IncFII_IncFIA |
| CP018361.1 | ERR1023792_IncFII_IncFIA |
| CP014669.1 | ERR1023775_IncFII_IncFIA |
| CP018992.1 | SRR2965672_IncFII_IncFIB |
| CP014650.1 | SRR1582875_IncFII_IncFIB |
| CP008833.1 | SRR3221963_IncFII_IncFIB |
| CP013323.1 | SRR3222155_IncFII_IncFIB |
| CP011623.1 | SRR3999064_IncFII_IncFIB |
| CP015135.1 | SRR1510963_IncFII_IncFIB |
| CP023555.1 | SRR3485532_IncFII_IncFIB |
| CP011577.1 | SRR2244257_IncFII_IncFIB |
| CP018355.1 | SRR2965692_IncFII_IncFIB |
| CP024585.1 | SRR1582895_IncFII_IncFIB |
| CP024530.1 | ERR1008737_IncFII_IncFIB_repB |
| CP024523.1 | ERR1008735_IncFII_IncFIB_repB |
| CP024572.1 | ERR1023801_IncFII_IncFIB_repB |
| | ERR1023800_IncFII_IncFIB_repB |

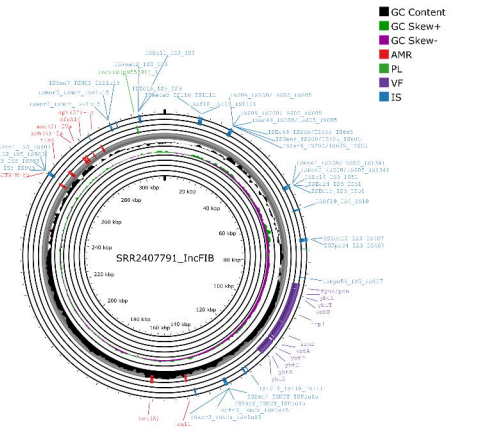
Salmonella enterica
Klebsiella pneumoniae
Klebsiella oxytoca
Escherichia coli
Kluyvera intermedia
Citrobacter freundii

b

Annotated complete plasmid sequence

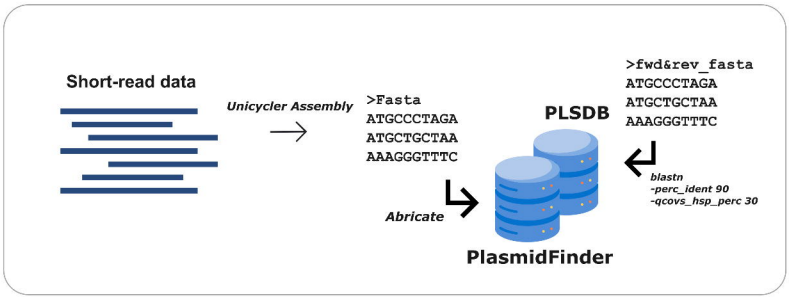



Annotated reconstructed plsMD plasmid

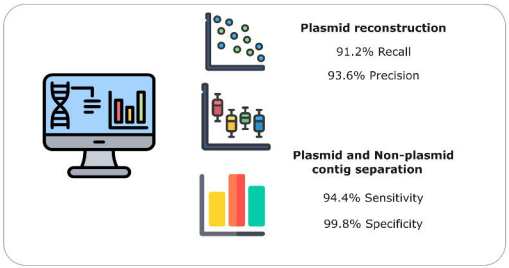



IncFIB (pN55391) replicon-type plasmid

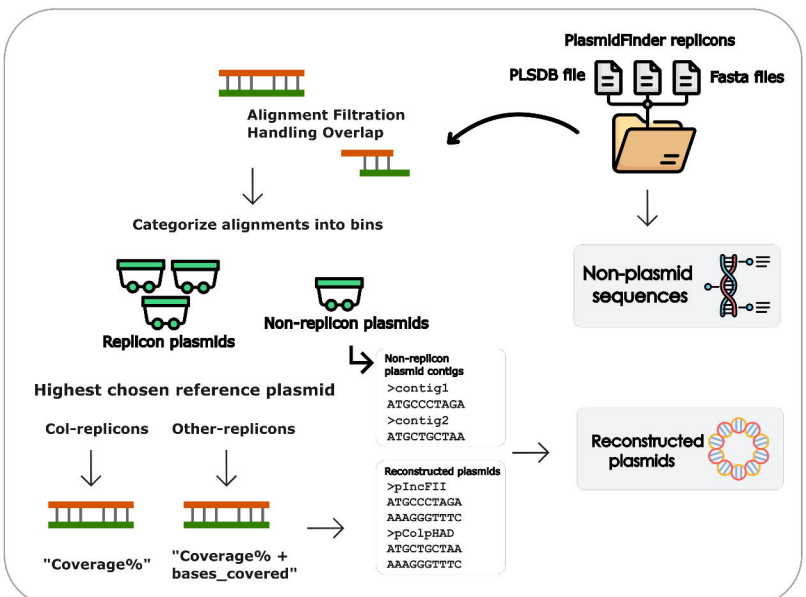
# plsMD pre-processing

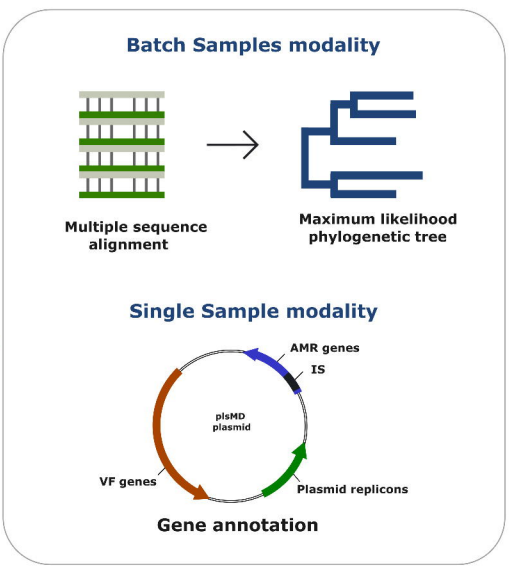



# Validation metrics




# plsMD processing




# plsMD usage modalities