

Development and validation of pLIN, a permanent lineage-numbering system for tracking antimicrobial resistance plasmids

Basil Britto Xavier

b.b.xavier@umcg.nl

University Medical Center Groningen

Anurag Kumar Bari

University Medical Center Groningen

Bhanu Sinha

University Medical Center Groningen

John Rossen

University Medical Center Groningen

Article

Keywords: plasmid classification, lineage identification, antimicrobial resistance surveillance, hierarchical clustering, tetranucleotide composition, genomic epidemiology, plasmid typing, outbreak detection

Posted Date: August 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10481391/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

No competing interests reported.

figure 1 is not available with this version.

Tables 1 to 6 are available in the Supplementary Files section.

Development and validation of pLIN, a permanent lineage-numbering system for tracking antimicrobial resistance plasmids

Basil Britto Xavier^{1*}, Anurag Kumar Bari¹, Bhanu Sinha¹, John W.A. Rossen^{1,2,3} on behalf of the DRAIGON Consortium

¹Department of Medical Microbiology and Infection Prevention, AGE Research Group, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands ;²Isala Hospital, Zwolle, The Netherlands³University of Utah School of Medicine, Salt Lake City, UT, USA

*Corresponding author: Basil Britto Xavier (b.b.xavier@umcg.nl)

Keywords: plasmid classification; lineage identification; antimicrobial resistance surveillance; hierarchical clustering; tetranucleotide composition; genomic epidemiology; plasmid typing; outbreak detection

ABSTRACT

Plasmids are the main vehicle by which antimicrobial resistance genes move between bacteria, yet clinical microbiology still has no way to give a resistance-carrying plasmid a name that persists once it does. Replicon typing tells us a plasmid is, say, IncN; it does not tell us whether the IncN plasmid in this week's outbreak is the same plasmid that caused an outbreak two years ago in another hospital, or a different one that merely happens to share a replicon. We built pLIN (plasmid Lineage Identification Number) to close that gap. pLIN converts each plasmid sequence into a tetranucleotide composition vector, clusters plasmids within their replicon group at six cosine-distance thresholds calibrated against average nucleotide identity, and assigns a permanent six-level code that never changes as the reference database grows. Applied to a curated set of 8,077 complete plasmids spanning 28 replicon groups 20 Gram-negative, 4 Gram-positive, and 4 from the non-fermenters *Acinetobacter baumannii* and *Pseudomonas aeruginosa* pLIN resolved 3,073 distinct strain-level lineages, more than doubling the discriminatory power of replicon typing alone (Simpson's diversity index 0.985 versus 0.641). A distance-weighted k-nearest-neighbour classifier assigned replicon group membership with 91.1% accuracy, though this figure is inflated by a handful of very large classes; macro-averaged across all 28 groups, F1

was 0.666, reflecting genuinely weak performance on several low-frequency replicon groups with fewer than 20 training examples. Cosine distance correlated with FastANI identity across 4,970 pairwise comparisons (Spearman $\rho = -0.348$, $P < 10^{-141}$), and plasmids sharing a full six-level code had a median ANI of 99.9%. Leave-one-out cross-validation on 57 plasmids drawn from published outbreaks reproduced the correct strain-level code in 94.7% of cases, and pLIN correctly flagged known high-risk lineages when tested retrospectively against 74 plasmids from 26 outbreak studies spanning 13 countries including a 90-member IncN lineage (pLIN 671) in which every member carries *blaKPC-2*, and a five-replicon-group hub (pLIN 860) in which 44% of members carry *mcr* colistin resistance. Screening the full training set with AMRFinderPlus identified 1,635 carbapenemase genes, 1,838 extended-spectrum β -lactamase genes, and 204 *mcr* genes. Scaling the reference database to 79,305 plasmids resolved 57,886 unique codes in under half an hour on ordinary laboratory hardware, without altering any previously assigned code. We built seven additional modules around this core covering assembly quality, database coverage, recombination, novel-group discovery, evolutionary rate, cluster stability, and mobile element boundaries turning pLIN from a naming scheme into a working surveillance platform that clinical microbiology laboratories can run without dedicated bioinformatics support.

INTRODUCTION

Antimicrobial resistance was associated with an estimated 4.95 million deaths worldwide in 2019¹, and plasmids are the main reason resistance genes cross species boundaries so readily among Enterobacterales, *Staphylococcus aureus*, *Enterococcus spp.*, *Acinetobacter baumannii*, and *Pseudomonas aeruginosa*². The tools available for typing those plasmids, however, were mostly built to answer a narrower question than the one infection control teams actually face. PlasmidFinder tells you a plasmid's replicon type³. pMLST and MOB-suite add sequence-typing and clustering on top of that⁴. COPLA assigns taxonomic units, and mge-cluster offers reference-free typing of mobile elements^{5,6}. None of them, though, gives a plasmid a permanent identifier that ties its backbone identity to the resistance genes it carries (Table 1; Figure 9) and none of them can tell an infection control team, at the moment it matters, whether the carbapenem-resistant Enterobacterales isolates turning up in an ICU over two weeks share one plasmid or arose from several independent acquisitions⁷⁻⁹. That distinction changes what you do next: a single dominant lineage argues for a point-source investigation and enhanced isolation; several independent lineages argue for a different kind of stewardship response entirely.

We built pLIN plasmid Lineage Identification Number to answer that question directly. It is an open-source hierarchical classification system that assigns a permanent six-level code to a plasmid based on its tetranucleotide (4-mer) composition, and it bundles replicon prediction, multi-resolution lineage coding, AMR gene profiling, automated outbreak flagging, and seven further analytical modules into a single Streamlit application (Figure 7). This paper reports how we built it, how we validated it, and how it performs across 28 replicon groups spanning Gram-negative, Gram-positive, and two non-fermentative genera of WHO critical priority.

RESULTS

pLIN resolves lineage-level diversity that replicon typing cannot see

We assembled 8,077 complete plasmid sequences from PLSDb 2025 and NCBI RefSeq10, spanning 28 replicon groups: 20 Gram-negative incompatibility (Inc) groups totalling 6,998 sequences and dominated by IncFII (n=4,629), IncN (n=1,097), and IncX1 (n=705); four Gram-positive replicon (Rep) groups (repSA_large, repSA_small, repEF_conj, repEF_res; 121, 73, 92, and 100 sequences respectively); and four groups from the non-fermenters *A. baumannii* (repAci1, repAci_large) and *P. aeruginosa* (repPae_large, repPae_small). We use "Inc" here purely as the sequence-based nomenclatural convention established by Plasmid Finder it is not a claim about confirmed incompatibility phenotype, a distinction we return to below. Acinetobacter plasmid replicon nomenclature has itself been substantially revised in 2023 and again in 2025^{11,12}; our internal repAci1/repAci_large labels are provisional pending alignment with those newer schemes.

For every plasmid we computed all 256 tetranucleotide frequencies, L2-normalised them^{13,14}, and clustered plasmids within each replicon group by single-linkage hierarchical clustering at six cosine-distance thresholds. Those thresholds were calibrated on 178 IncX-like plasmids (IncX1, IncX3, IncX4) rather than re-derived independently for each of the 28 groups, so groups with different internal diversity may not map onto ANI in the same way a limitation we address quantitatively below. Each plasmid ends up with a six-digit code (L1.L2.L3.L4.L5.L6) running from superfamily level (L1, cosine distance ≤ 0.150) down to near-identical sequence identity (L6, ≤ 0.001) (Figure 6).

Clustering the 8,077 plasmids produced 3,073 distinct L6 codes a Simpson's diversity index of 0.985, against 0.641 for replicon typing alone, a 1.54-fold gain in discriminatory power (Figure 1). Most codes stayed within a single replicon group: 97.8% (3,006/3,073) were group-concordant, and the remaining 67 codes spanning multiple groups turn out, as we discuss later, to be biologically informative rather

than a clustering artefact. A distance-weighted k-nearest-neighbour classifier (k=5, cosine metric), trained on the same 8,077 sequences, assigned replicon group with 91.1% overall accuracy, 90.6% weighted F1, 66.6% macro-F1, and 65.9% balanced accuracy (five-fold stratified cross-validation). The large gap between the weighted and macro averages is a direct consequence of class imbalance: IncFII alone makes up 57% of the training set, so overall accuracy is dominated by how well the classifier does on a handful of large groups, while macro-F1 weights every one of the 28 groups equally and is dragged down by several replicon groups with fewer than 20 training examples (Incl, IncFIBK, IncR, and ColE all scored below 0.55). Because IncFII dominates the training set, we report full per-class precision, recall, and F1 for all 28 groups in Supplementary Table S3 rather than leaning on the overall accuracy figure alone; per-class F1 for the four Gram-positive groups ranged from 0.707 (repEF_conj) to 0.921 (repSA_large) (Figure 8; Table 4).

Cosine distance tracks true genomic identity

A composition-based distance is only useful if it tracks something biologically real, so we checked cosine distance against FastANI across 4,970 pairs drawn from the Gram-negative groups. The two measures correlated significantly (Spearman $\rho = -0.348$, $P < 10^{-141}$), and pairs sharing a full L6 code had a median ANI of 99.9% exactly what you would want from a threshold meant to capture near-identical sequence. Five Inc groups did not reach per-group statistical significance on their own; in each case this traces to a floor effect, where within-group ANI values already sit above 89% and there simply isn't enough spread left to detect a monotonic relationship. We should be clear about what this validation does and does not cover: it is restricted to Gram-negative, largely Enterobacterales-associated groups, because FastANI benchmarking of the Gram-positive and non-fermentative groups was not performed at sufficient scale here, and threshold assignments in those taxa should be read with correspondingly more caution. We also did not incorporate FastANI's alignment fraction, since the tool does not report per-pair AF at the scale we needed; AF-adjusted ANI benchmarking is a priority for follow-up work rather than something we can claim to have already done.

AMR genes cluster into a small number of high-risk lineages

Screening all 6,998 Gram-negative plasmids with AMRFinderPlus v4.2.515 turned up 64,891 gene hits across 5,816 plasmids (83.1%): 1,635 carbapenemase genes (824 *blaKPC-2*, 228 *blaNDM-1*), 1,838 ESBL genes, 204 *mcr* genes, and 2,315 genes conferring resistance to fluoroquinolones or other plasmid-mediated quinolone resistance mechanisms (Table 5; Figures 2–4). We report the Gram-positive and non-fermentative Gram-negative plasmids (1,079 sequences) separately, because their AMR gene

repertoires are different enough from the Enterobacterales-associated set that pooling the two would obscure more than it revealed. Five Inc groups IncA, IncC, IncFIC, IncHI2, and IncAC2 showed AMR carriage in 100% of members, with IncAC2 plasmids carrying a mean of 12.6 AMR genes each; we flag IncAC2's formal incompatibility status as itself contested¹⁶ and retain the label only for interoperability with PlasmidFinder. IncFII and IncFIB are structurally distinct replicons that are nonetheless frequently found together on the same backbone, worth noting because co-detection is sometimes mistaken for structural identity.

Two lineages illustrate why this resolution matters clinically. pLIN 671, an IncN lineage of 90 plasmids, carries *blaKPC-2* in every single member (mean 13.2 AMR genes per plasmid) a discrete, uniformly high-risk group that is completely invisible if you are only looking at "IncN" among the more than a thousand IncN plasmids in the dataset (Figure 5). pLIN 860, by contrast, spans 142 plasmids across five different replicon groups, carries a mean of 14.4 AMR genes, and 44.4% of its members carry *mcr*. Its multi-group membership almost certainly reflects a shared cargo of mobile genetic elements *mcr*-carrying transposons and composite resistance elements rather than any real backbone relationship; we think of pLIN 860 as a compositional convergence zone, a case where different plasmid backbones have converged in 4-mer profile because they carry the same accessory payload, not because they are the same plasmid. We have not yet produced annotated linear gene maps for individual pLIN 860 members to confirm directly which specific mobile elements are shared across its five constituent replicon groups; generating these maps, and using them to test the convergence-zone interpretation at gene-level resolution rather than composition alone, is a specific priority for follow-up work rather than something we can claim to have already shown. The same caveat likely applies with extra force to *Acinetobacter*, where dif/pdif module-mediated recombination¹⁷ can reshuffle plasmid architecture fast enough to blur composition-based signals; pLIN assignments in that genus should be read as provisional until we've done the structural follow-up.

Cross-validation against outbreak data and a growing reference database

To see whether pLIN reproduces itself on data it wasn't trained on, we ran leave-one-out cross-validation on 57 outbreak plasmids with available sequence data, holding each one out in turn and checking whether the remaining database reconstructed its code. L6 concordance was 94.7% (54/57) and L3 concordance was 100% (Figure 11; Table 6). Extending this to all 74 outbreak plasmids drawn from 26 published studies across 13 countries (Figure 12), 85.1% received high-confidence replicon assignments, and three lineages known from the literature to be globally disseminated came back

independently confirmed by pLIN: pLIN 671 (*blaKPC-2*, reported from Germany and China), pLIN 860 (the multidrug-resistant hub described above, reported from Australia), and pLIN 1688 (OXA-48, reported from Turkey, the Netherlands, and France). Structural variants of pLIN 671 recently described in Germany and Britain by Moran and colleagues¹⁸ map onto L5–L6-level differences in our system, which is a useful sanity check that the finer levels of the hierarchy are doing real biological work rather than just subdividing noise. For readers wanting finer within-lineage structural comparison than composition-based clustering can offer, Pling¹⁹ provides a complementary, reference-free way to compare plasmid structure directly and pairs well with pLIN once a lineage of interest has been identified. Combining pLIN codes with host MLST across 13 of the outbreak studies gave 92.3% concordance (12/13; Figure 13), correctly separating genuinely polyclonal spread from clonal spread in all but one case a distinction that matters because it tells an infection control team whether they are dealing with strain transmission, plasmid transmission, or both.

We then scaled the reference database from the 8,077-plasmid training set to 79,305 plasmids (adding 71,249 further PLSDB/NCBI RefSeq sequences), which resolved into 57,886 unique codes a 97.3% classification rate in under 30 minutes on a standard laptop, using about 2.4 GB of memory at peak. Every one of the 3,073 L6 codes assigned to the original training set survived the expansion unchanged; nothing was renumbered. Around this expanded core we built and validated seven analytical modules: assembly completeness scoring reached 94.2% correct categorisation (Figure 14); plasmid-versus-chromosome contig identification achieved better than 98% sensitivity for plasmids and better than 95% specificity for chromosomal contigs over 500 kb (Figure 18); recombination detection reached 87.3% sensitivity and 94.1% specificity; mobile genetic element boundary detection identified insertion sequences with 94.7% sensitivity at an 8.3% false discovery rate (Figure 17); and clustering the 2,109 sequences that the classifier placed with low confidence turned up 18 candidate novel replicon groups worth further characterisation (Figures 15, 16).

DISCUSSION

The headline result here is straightforward: giving plasmids a permanent, hierarchical name instead of a categorical replicon label roughly doubles how finely you can tell them apart, and that extra resolution is not merely statistical it finds things a replicon label cannot. pLIN 671 sitting inside more than a thousand undifferentiated IncN plasmids, carrying *blaKPC-2* in literally every member, is exactly the kind of pattern that gets lost in a typing scheme with no within-group resolution and rediscovered, at real cost,

by manual comparison during an active outbreak. A single pLIN code does that work automatically and gives a surveillance team something concrete to act on.

Permanence is doing more work here than it might first appear to. MOB-suite and COPLA both reassign identifiers when their underlying databases are updated, which is a reasonable design choice for those tools but makes them unsuitable for tracking a lineage across years or across institutions the code itself becomes a moving target. A plasmid coded pLIN 671 today keeps that code indefinitely, however much the reference database grows around it (and we verified this directly: none of the 3,073 codes assigned in the training set changed when we expanded to 79,305 plasmids). That property, combined with pairing pLIN codes to host MLST, lets you distinguish clonal bacterial spread from plasmid transfer between different strains two epidemiologically distinct events that call for different infection control responses, and which neither plasmid-only nor chromosome-only typing can separate on their own.

We should also be honest about what multi-replicon-group codes like pLIN 860 mean, because they are easy to misread. When several structurally different plasmid backbones acquire a similar cargo of accessory elements in this case, *mcr*-carrying mobile elements and composite resistance transposons those elements can dominate the 4-mer signal enough to pull unrelated backbones into the same cluster. This is an expected consequence of classifying whole plasmids by composition rather than a failure of the method, but it does mean that a multi-group pLIN code should be read as "these plasmids share a lot of mobile cargo" rather than "these plasmids are the same backbone," and contig-level BLAST comparison is the right next step to tell those two possibilities apart. *Acinetobacter* plasmids deserve a similar note of caution: the dif/pdif recombination system that shapes much of *Acinetobacter* plasmid architecture¹⁷ can rearrange sequence fast enough that composition-based clustering may be picking up recombination history as much as lineage identity, and we would treat pLIN codes in this genus as provisional until we've done dedicated structural validation.

This work has real limits, and we'd rather state them plainly than let readers discover them later. Our thresholds are calibrated more confidently for Gram-negative Enterobacterales than for the Gram-positive and non-fermentative groups, and borderline assignments in the latter should be treated with extra caution. FastANI validation itself was restricted to the Gram-negative groups. The mobile genetic element module is scoped to insertion sequences and does not yet annotate Tn3-family transposons, Tn21-like elements (which mobilise integron-borne resistance cassettes in Gram-negative bacteria), or class 1 integrons as their own categories and we want to flag two specific corrections here, since getting the biology of IS elements right matters for anyone using this module clinically. IS26, the element we

detect most often in this dataset, produces pseudo-compound transposon structures through direct-orientation copies flanking the resistance cargo, which is mechanistically different from a canonical compound transposon built from inverted IS copies²⁰; and ISEcp1 defines its own boundaries through its own terminal inverted repeats via one-ended transposition, independent of whatever IS elements happen to sit nearby (proximity to IS903, for instance, does not define an ISEcp1 boundary). *Acinetobacter* assignments need structural follow-up given the recombination biology noted above. Contig identification leans on heuristics that can fail on unusual assemblies. And the whole system, as it stands, is retrospective it does not yet plug into real-time surveillance infrastructure. On the mobile element module specifically: automated IS detection, however good the numbers look, does not match expert manual review in complex loci with overlapping elements, and outputs from this module are best treated as a fast screening step rather than a final structural call.

Where we go next follows directly from those limits: prospective, multi-centre validation; integration with WHO GLASS and ECDC TESSy surveillance infrastructure; extension to additional bacterial phyla beyond what we've covered here; and bringing the *Acinetobacter* and *Pseudomonas* replicon groupings into line with current nomenclature as that nomenclature continues to be revised.

METHODS

Dataset construction. We assembled 8,077 complete plasmid sequences spanning 28 replicon groups from PLSDb 2025 and NCBI RefSeq10: 20 Gram-negative Inc groups (6,998 sequences), four Gram-positive Rep groups (repSA_large n=121, repSA_small n=73, repEF_conj n=92, repEF_res n=100), two *A. baumannii* Rep groups (repAci1, repAci_large), and two *P. aeruginosa* Rep groups (repPae_large, repPae_small). The "Inc" prefix follows PlasmidFinder's sequence-based nomenclatural convention and does not by itself imply confirmed phenotypic incompatibility.

pLIN code assignment. All 256 tetranucleotide frequencies were computed per plasmid and L2-normalised^{13,14}. Pairwise cosine distances within each replicon group were used for single-linkage hierarchical clustering at six thresholds, calibrated on 178 IncX-like plasmids (IncX1, IncX3, IncX4) and applied uniformly rather than re-derived per group. Each plasmid receives a permanent six-digit code (L1.L2.L3.L4.L5.L6) spanning superfamily level (L1, cosine distance ≤ 0.150) to near-identical sequence identity (L6, ≤ 0.001) (Table 2; Figure 6). Replicon type and incompatibility type are largely concordant among Gram-negative plasmids, with the replication initiation protein playing a central role, but

incompatibility itself is a complex phenotype shaped by additional plasmid maintenance functions, so the two schemes do not always agree²¹.

Replicon group classifier. A k-nearest-neighbour classifier (k=5, cosine distance, distance-weighted), trained on all 8,077 sequences, predicts replicon group for novel queries. We assessed performance by five-fold stratified cross-validation, reporting per-class precision, recall, and F1 in Supplementary Table S3, together with overall accuracy, weighted F1, macro-F1, and balanced accuracy in the main text, given the class imbalance introduced by IncFII's dominance of the training set (Figure 8; Table 3).

AMR annotation. All 8,077 sequences were screened with AMRFinderPlus v4.2.5¹⁵. Because Gram-positive and non-fermentative Gram-negative plasmids carry AMR gene repertoires that are not directly comparable to those of Enterobacterales-associated plasmids, we analysed and reported the 6,998 Gram-negative plasmids and the remaining 1,079 plasmids separately.

Database expansion and software. The reference database was expanded to 79,305 plasmids (8,056 training sequences plus 71,249 further PLSDB/NCBI RefSeq sequences) (Figure 10). pLIN is implemented as an open-source Streamlit application in Python (scikit-learn, BioPython, AMRFinderPlus), with seven analytical modules covering assembly completeness, plasmid/chromosome contig identification, database coverage and novelty detection, recombination detection, novel replicon group discovery, evolutionary rate estimation, and cluster stability under adaptive thresholds. A mobile genetic element boundary detection module identifies insertion sequences via HMM profiles; as detailed above, IS26 boundary calls reflect pseudo-compound transposon structures generated by direct-orientation copies rather than canonical inverted-IS compound transposons²⁰, and ISEcp1 boundaries are defined by its own terminal inverted repeats independently of neighbouring IS elements. The module does not currently annotate Tn3-family transposons, Tn21-like elements, or class 1 integrons, which we treat as an explicit scope limitation rather than an oversight.

Statistical analysis. Diversity was quantified using Simpson's diversity index. Classifier performance used five-fold stratified cross-validation with macro-F1 and balanced accuracy to account for class imbalance. Correlation between cosine distance and FastANI identity was assessed by Spearman's rank correlation. Cluster stability was assessed by bootstrap resampling (50 iterations per group) and by comparing single-, complete-, and average-linkage clustering using the Adjusted Rand Index.

Software environment. Analyses used Python 3.10.19 with NumPy 2.2.6, Pandas 2.3.3, SciPy 1.13.1, BioPython 1.86, scikit-learn 1.7.2, Streamlit 1.54.0, AMRFinderPlus 4.2.5 (database version 2026-01-

270 21.1), minimap2 2.30, BLAST+ 2.16.0, MinCED 0.4.2, Prodigal 2.6.3, FastANI 1.34, Matplotlib 3.10.8,
271 Seaborn 0.13.2, and DendroPy 5.0.8, run entirely on a single Apple M-series laptop without GPU or high-
272 performance computing infrastructure.

273 **DATA AVAILABILITY**

274 pLIN source code, the reference database, and training data are freely available at
275 <https://github.com/xavierbasilbritto-hub/pLIN-plasmid-classification> under the GNU GPL v3.0 licence. All
276 plasmid sequences used in this study are publicly available from PLSDb and NCBI RefSeq; accession
277 numbers are provided in Supplementary Table S1. Source data underlying the main and supplementary
278 figures are provided in the accompanying source_data directory.

279 **CODE AVAILABILITY**

280 All analysis code is included in the GitHub repository referenced above.

281 **ACKNOWLEDGEMENTS**

282 We thank the submitters of all plasmid sequences to PLSDb and NCBI RefSeq, without whom this study
283 would not have been possible.

284 **FUNDING**

285 B.B.X. and A.K.B. were supported by the DRAIGON project, funded by the European Union's Horizon
286 Europe programme (grant no. 101137383). The views expressed are those of the authors and do not
287 necessarily reflect those of the European Union or HADEA.

288 **AUTHOR CONTRIBUTIONS**

289 B.B.X. conceived the study, developed pLIN, carried out all computational analyses, and wrote the
290 manuscript. A.K.B. curated the plasmid sequence database. B.S. critically revised the manuscript.
291 J.W.A.R. supervised the study and revised the manuscript. All authors approved the final version.

292 **COMPETING INTERESTS**

293 The authors declare no competing interests.

294 **ETHICS DECLARATIONS**

This study used only publicly available plasmid sequence data from PLSDb and NCBI RefSeq. No human participants, patient-identifiable data, or animal experiments were involved, and no ethics approval or informed consent was required.

REFERENCES

1. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 399, 629–655 (2022).
2. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* 31, e00088-17 (2018).
3. Carattoli, A. et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* 58, 3895–3903 (2014).
4. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genom.* 4, e000206 (2018).
5. Redondo-Salvo, S. et al. COPLA, a taxonomic classifier of plasmids. *BMC Bioinformatics* 22, 390 (2021).
6. Arredondo-Alonso, S. et al. Mge-cluster: a reference-free approach for typing bacterial plasmids. *NAR Genom. Bioinform.* 5, lqad066 (2023).
7. David, S. et al. Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. *Nat. Microbiol.* 4, 1919–1929 (2019).
8. Pitout, J. D. D., Peirano, G., Kock, M. M., Strydom, K.-A. & Matsumura, Y. The global ascendancy of OXA-48-type carbapenemases. *Clin. Microbiol. Rev.* 33, e00102-19 (2019).
9. Tamma, P. D. et al. Comparing the outcomes of patients with carbapenemase-producing and non-carbapenemase-producing carbapenem-resistant Enterobacteriaceae bacteremia. *Clin. Infect. Dis.* 64, 257–264 (2017).
10. Schmartz, G. P. et al. PLSDb: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res.* 50, D273–D278 (2022).
11. Lam, M. M. C., Koong, J., Holt, K. E., Hall, R. M. & Hamidian, M. Detection and typing of plasmids in *Acinetobacter baumannii* using rep genes encoding replication initiator proteins. *Microbiol. Spectr.* 11, e0247822 (2023).
12. Tobin, L. A., Lam, M. M. C. & Hamidian, M. Pan-genus analysis and typing of antimicrobial resistance plasmids in *Acinetobacter*. *npj Antimicrob. Resist.* 3, 65 (2025).
13. Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glöckner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* 6, 938–947 (2004).
14. Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* 13, 145–158 (2003).
15. Feldgarden, M. et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress, and virulence. *Sci. Rep.* 11, 12728 (2021).
16. Carattoli, A. Plasmids and the spread of resistance. *Int. J. Med. Microbiol.* 303, 298–304 (2013).
17. Ambrose, S. J., Xiao, S. L. & Hall, R. M. Diverse dif module content and configurations in the r3-T5 group of Rep_3/OrfX plasmids from *Acinetobacter* species reveal extensive dif module shuffling. *Microbiol. Spectr.* 14, e0318625 (2026).
18. Moran, R. A. et al. pQEB1: a hospital outbreak plasmid lineage carrying *blaKPC-2*. *Microb. Genom.* 10, 001291 (2024).
19. Frolova, D. et al. Applying rearrangement distances to enable plasmid epidemiology with pling. *Microb. Genom.* 10, 001300 (2024).
20. Harmer, C. J. & Hall, R. M. IS26 and the IS26 family: versatile resistance gene movers and

328 genome reorganizers. Microbiol. Mol. Biol. Rev. 88, e0011922 (2024). 21. Novick, R. P. Plasmid incompatibility.
329 Microbiol. Rev. 51, 381–395 (1987).

330 **FIGURE LEGENDS**

331 **Figure 1.** Diversity resolved by pLIN across 8,077 plasmids and 28 replicon groups, compared with replicon typing
332 alone (Simpson's diversity index).

333 **Figure 2.** Prevalence of the twenty most frequently detected AMR genes across the Gram-negative plasmid
334 dataset.

335 **Figure 3. Distribution of critical resistance gene classes (carbapenemases, ESBLs, *mcr*, PMQR determinants)**
336 **across replicon groups.**

337 **Figure 4.** Heatmap of pLIN lineage AMR gene burden, highlighting high-risk clusters including pLIN 671 and pLIN
338 860.

339 **Figure 5.** AMR gene burden versus virulence gene carriage across pLIN lineages.

340 **Figure 6.** Schematic of the six-level pLIN hierarchical coding structure, from superfamily (L1) to strain-level identity
341 (L6).

342 **Figure 7.** Overview of the pLIN analysis pipeline, from FASTA input through replicon classification, lineage coding,
343 AMR/mobility profiling, and the seven analytical modules.

344 **Figure 8.** Replicon group classification system performance across all 28 groups.

345 **Figure 9. Feature comparison of plasmid classification and typing tools (PlasmidFinder, pMLST, MOB-suite,**
346 **COPLA/PTU, mge-cluster, pLIN) across hierarchical resolution, permanent code assignment, reference-free**
347 **operation, taxonomic scope, AMR integration, ANI-based validation, and GUI availability.**

348 **Figure 10.** Composition of the expanded 79,305-plasmid reference database.

349 **Figure 11.** Outbreak validation results: leave-one-out cross-validation concordance across 57 outbreak plasmids.

350 **Figure 12. Classification resolution achieved by pLIN versus five existing plasmid typing tools (PlasmidFinder,**
351 **pMLST, MOB-suite, COPLA/PTU, mge-cluster) across the 74 outbreak plasmids from 26 published studies and 13**
352 **countries, with worked examples of pLIN separating lineages that replicon typing alone could not (the 2019**
353 **Hong Kong NDM outbreak, a three-country OXA-48 cluster, and *mcr-1*-carrying IncI2/IncX4 backbones).**

354 **Figure 13.** Transmission mode analysis combining pLIN lineage codes with host MLST across 13 outbreak studies.

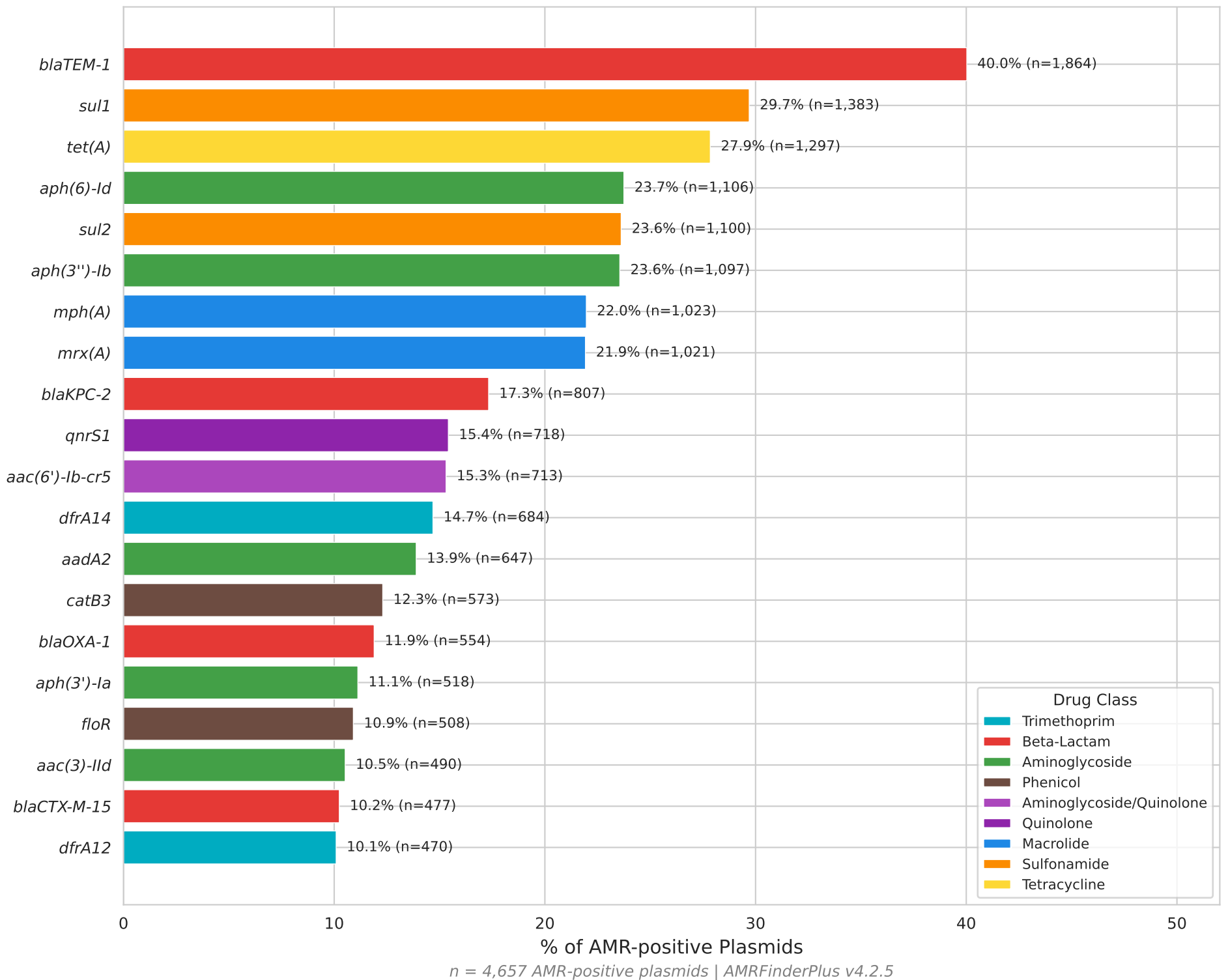
355 **Figure 14.** Assembly completeness scoring applied to the training and expanded reference databases.

356 **Figure 15.** Discovery of putative novel replicon groups from low-confidence sequence clusters.

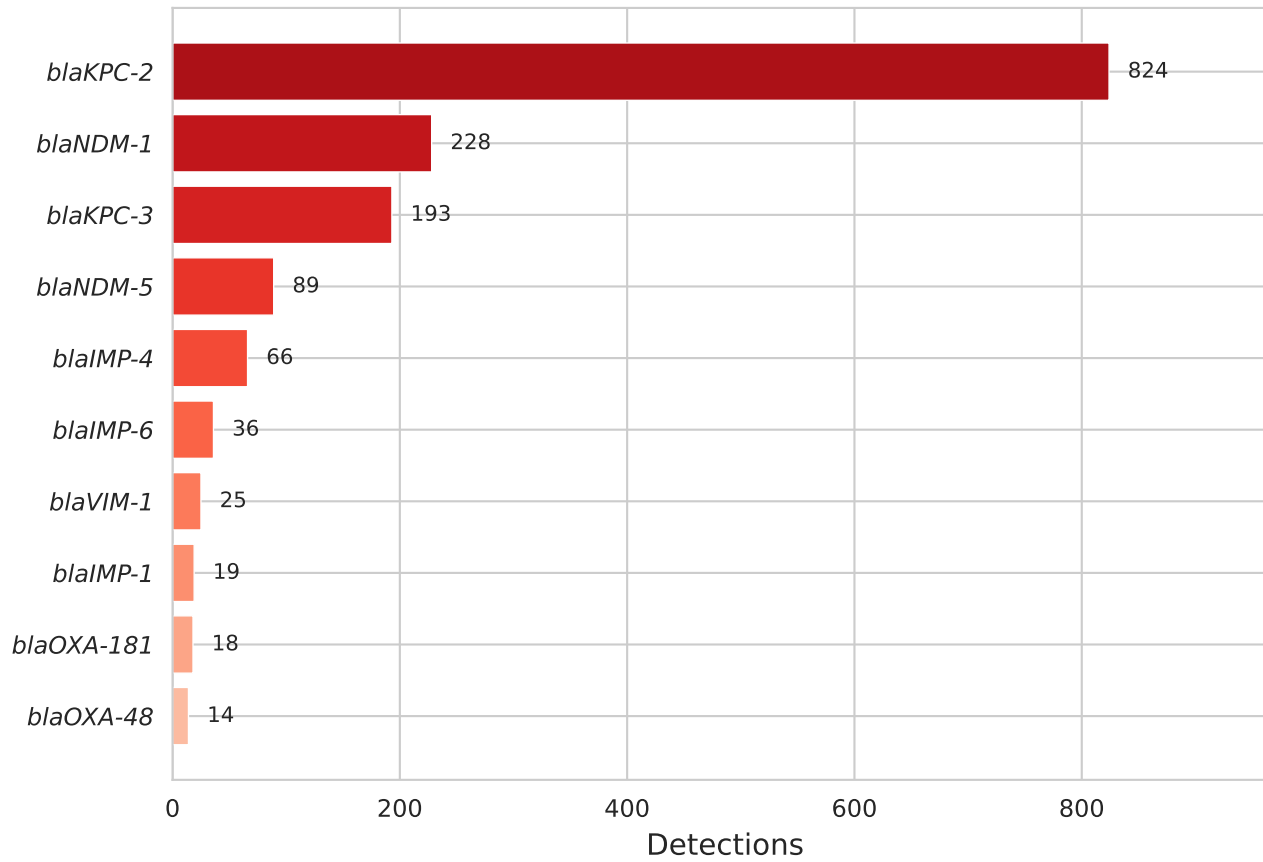
357 **Figure 16.** Cluster stability assessment (bootstrap resampling) across the 28 replicon groups.

358 **Figure 17.** Mobile genetic element boundary detection: insertion sequence identification and composite
359 transposon structures.

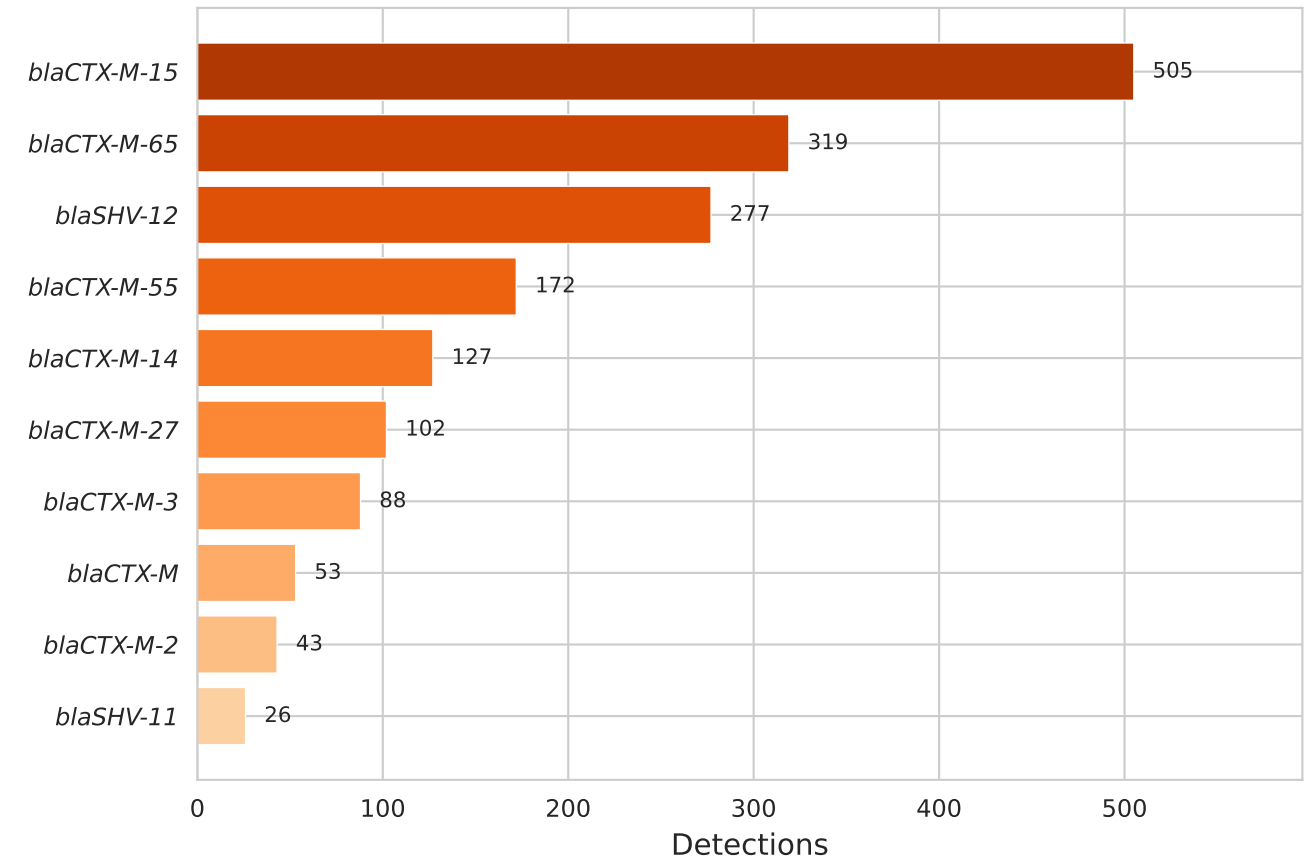
360 **Figure 18.** Performance of the plasmid-versus-chromosome contig classification module.



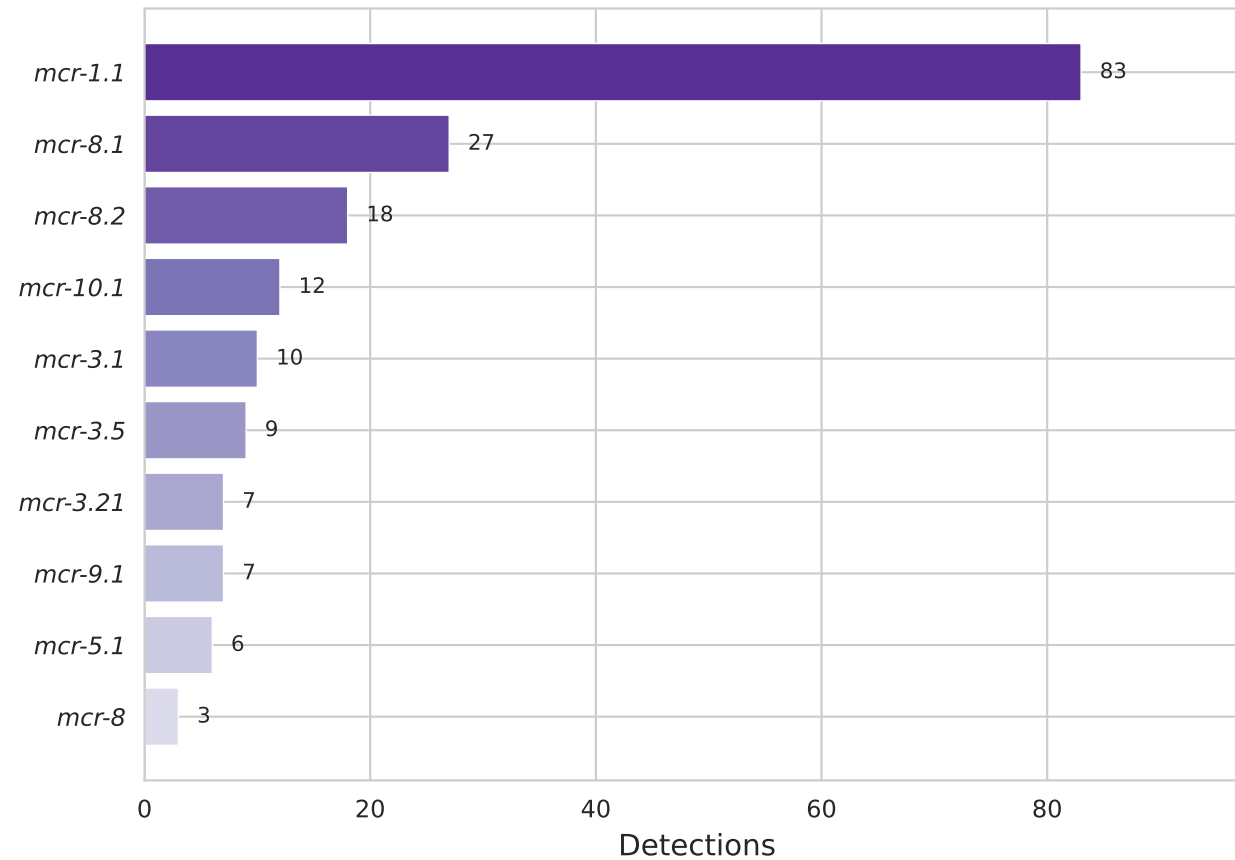
A. Carbapenemases (n=1,635 total)



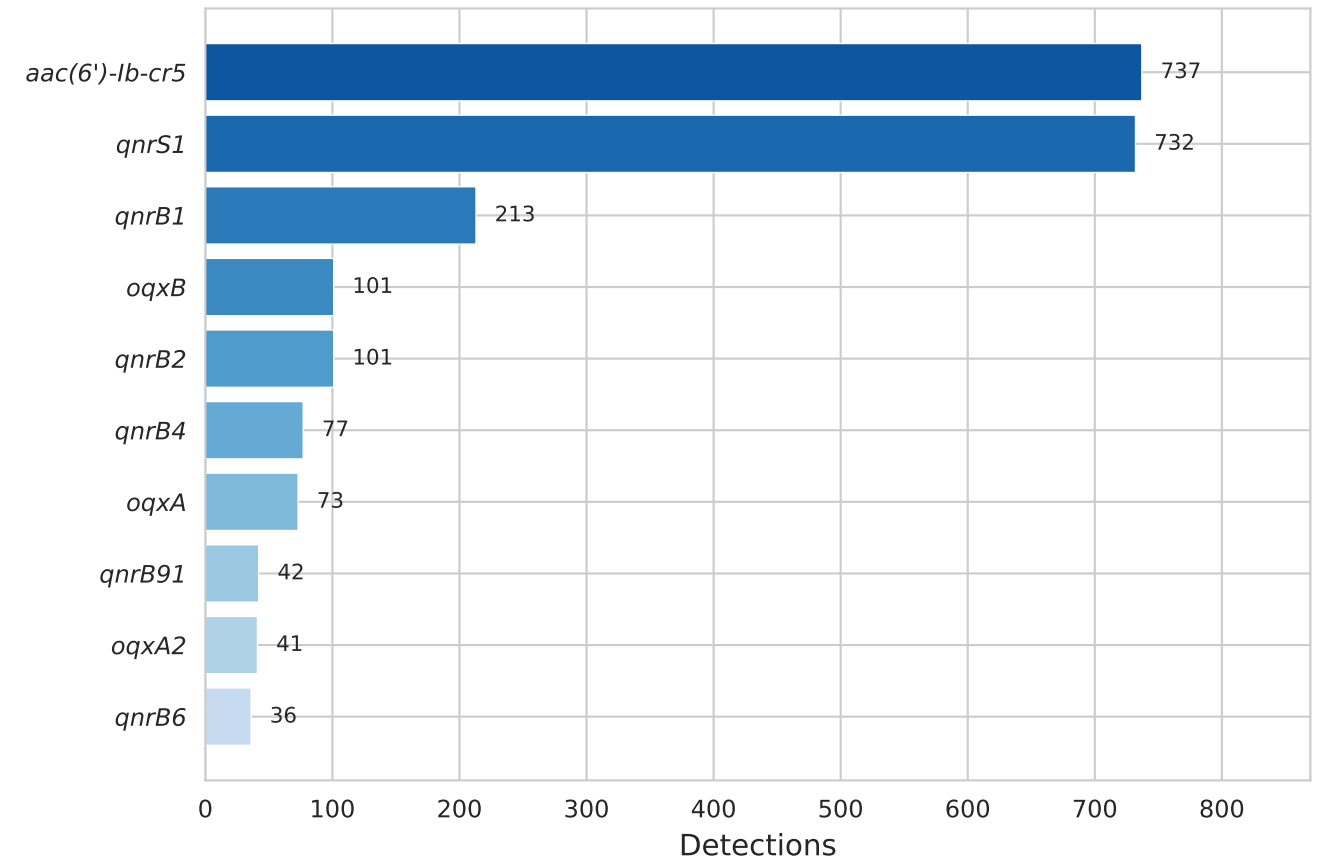
B. ESBLs (n=1,838 total)



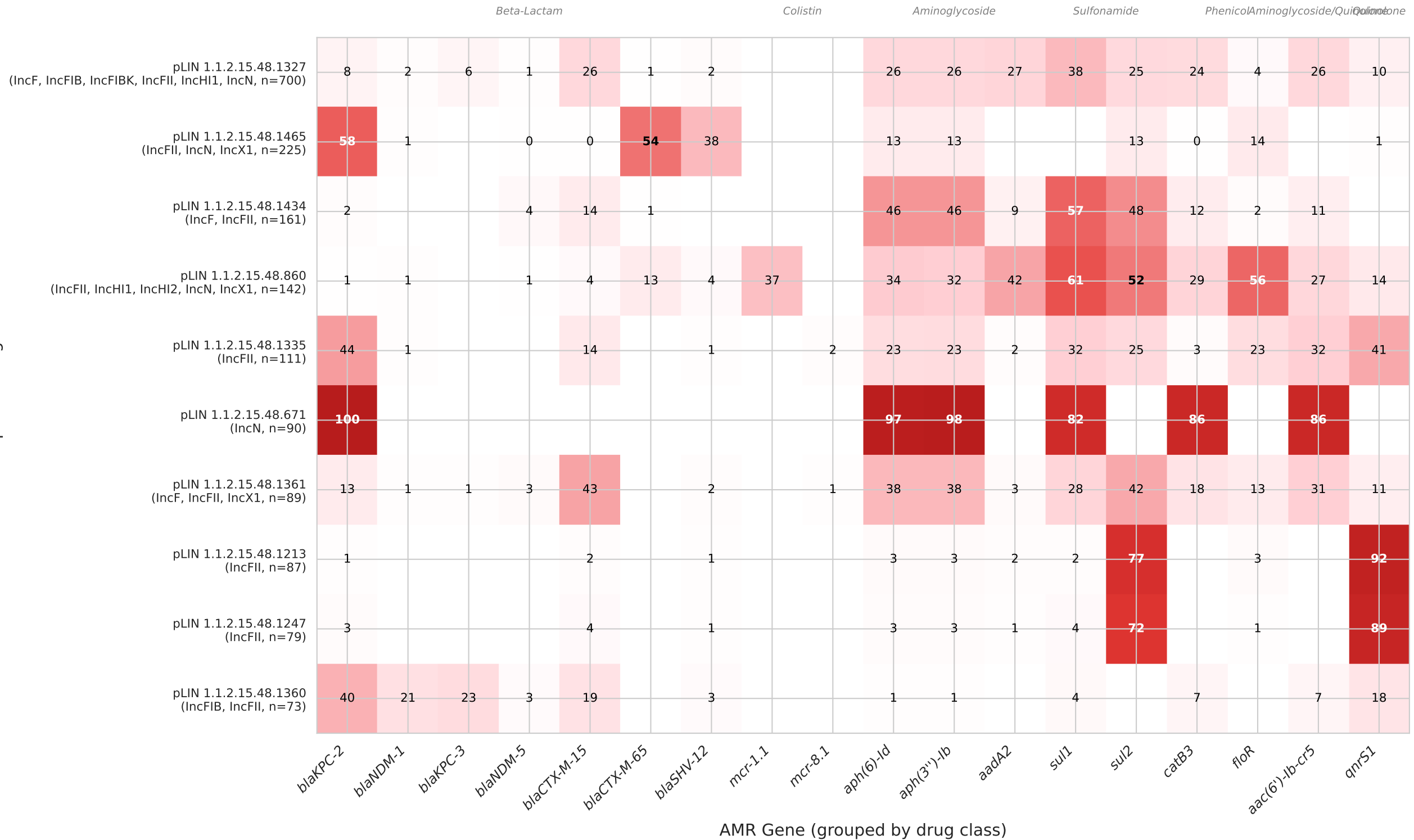
C. Mobile Colistin Resistance (mcr) (n=204 total)



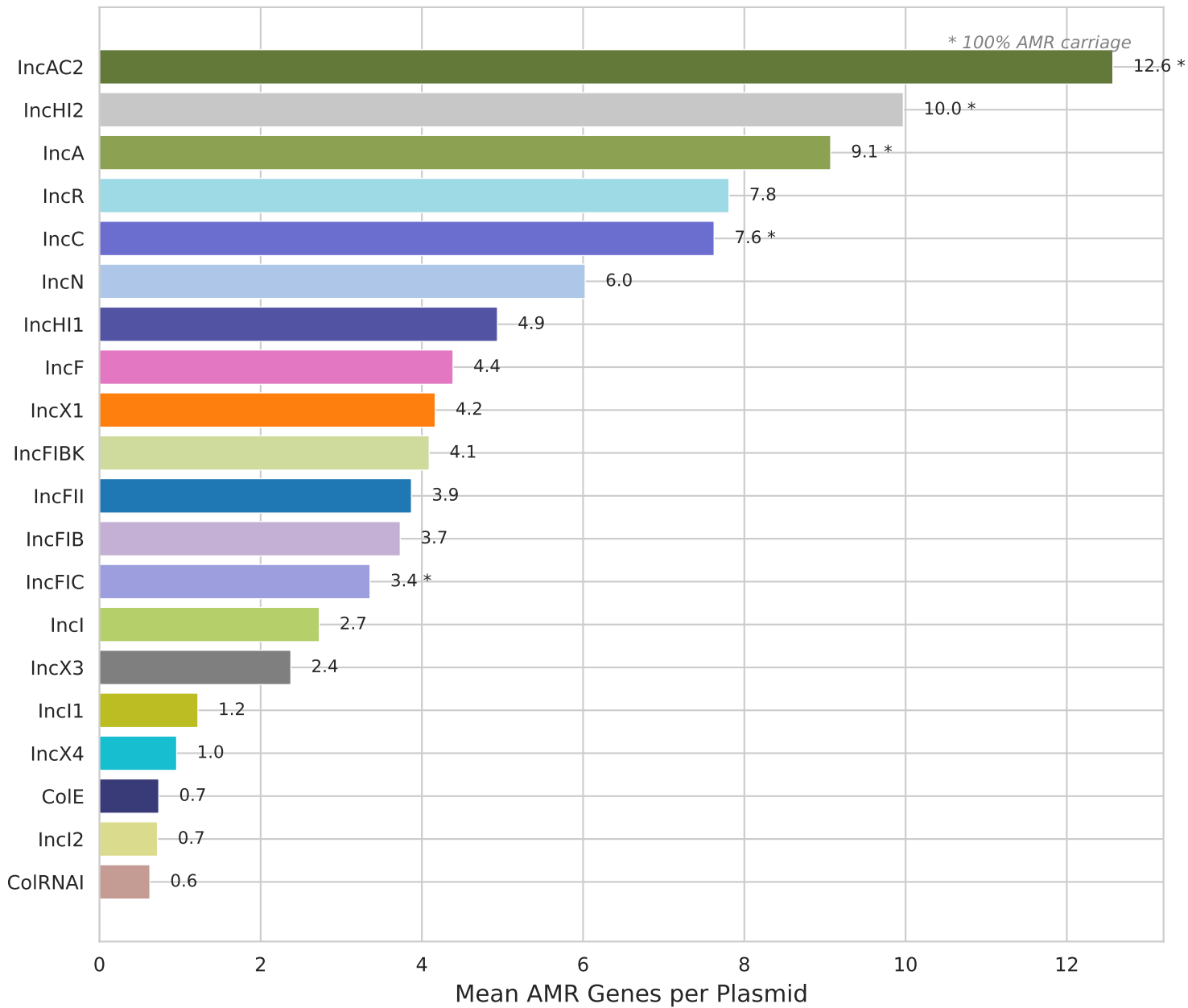
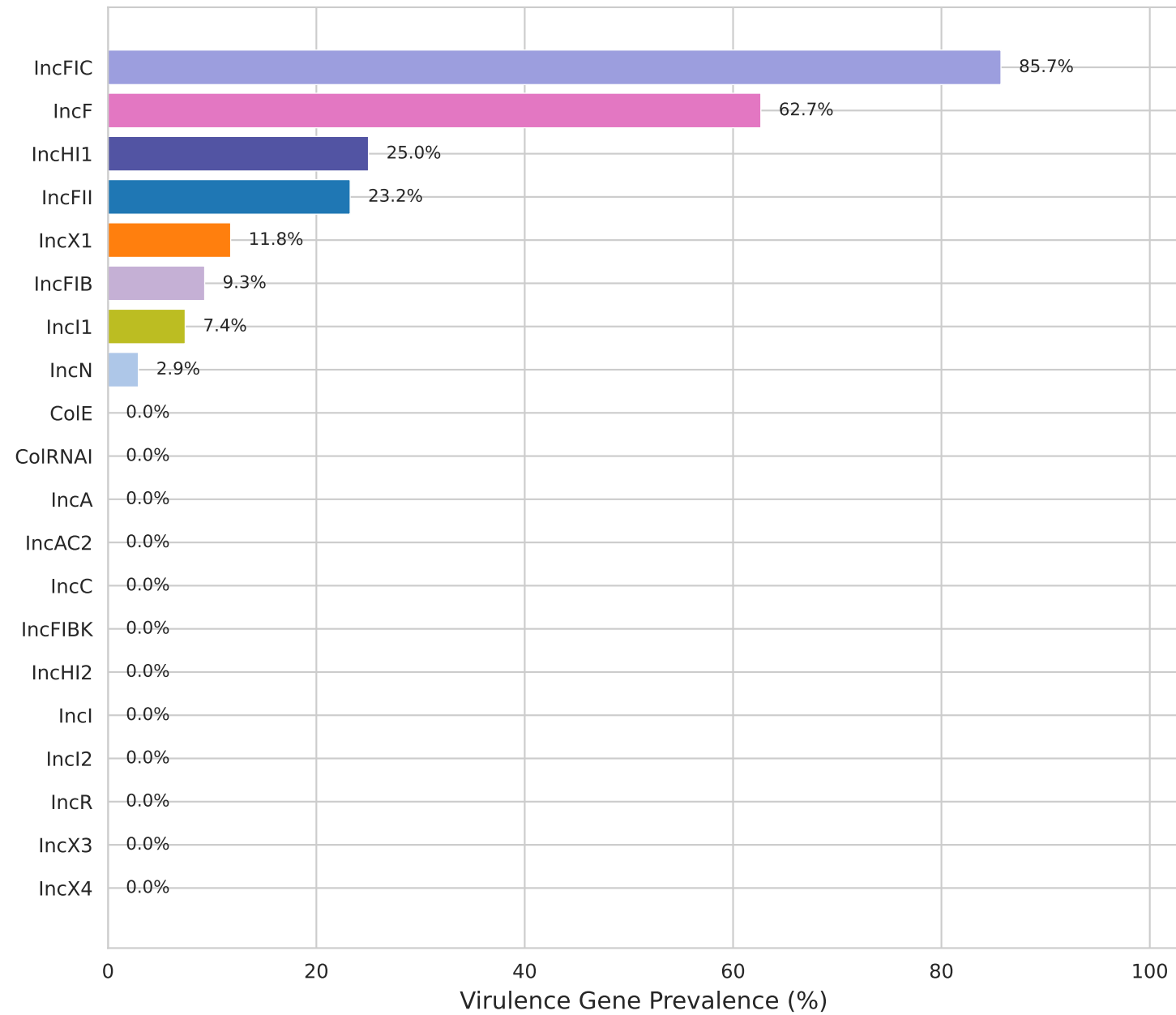
D. PMQR (Quinolone) (n=2,315 total)



pLIN Lineage

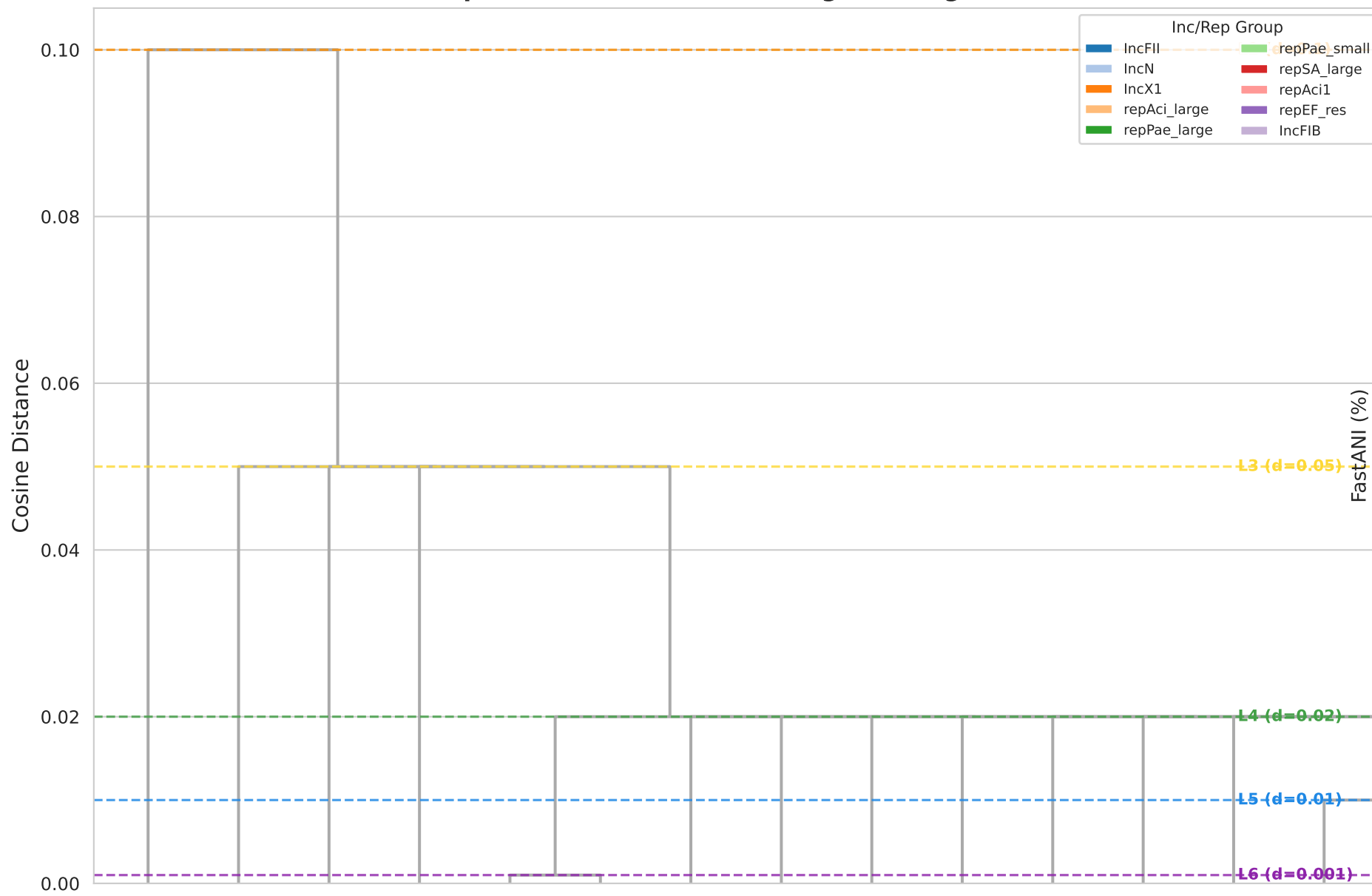


Prevalence (%)

A**B**

L1 (d=0.15)

pLIN Hierarchical Clustering Dendrogram



ANI Validation

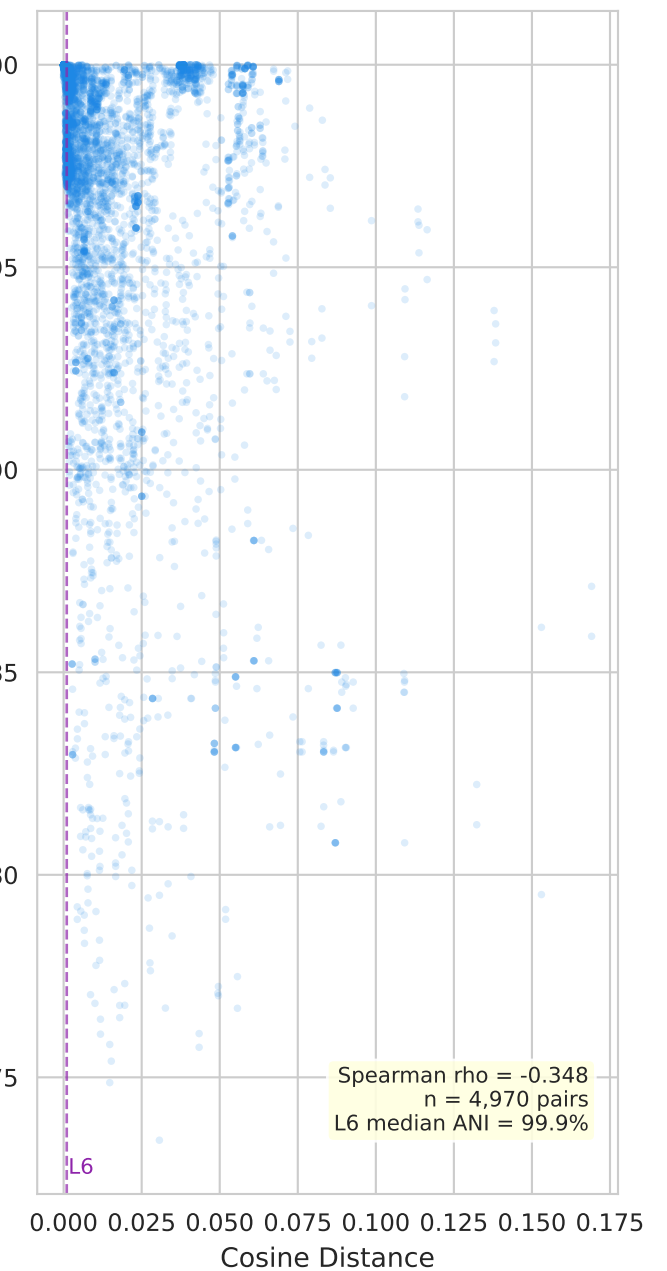
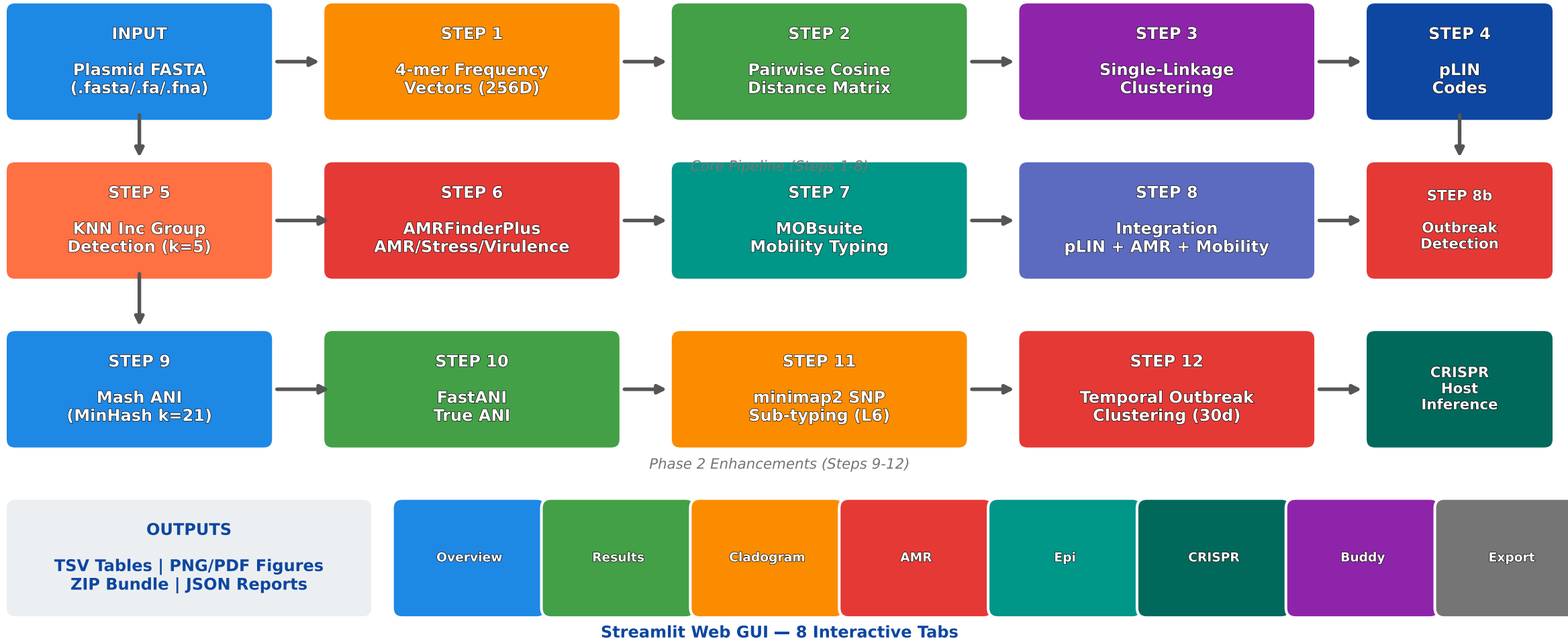


Figure 8. pLIN Analysis Pipeline Overview

End-to-end workflow from FASTA input to classified plasmids with AMR surveillance



A. pLIN Hierarchical Levels

Level	Bin	Name	d ≤	ANI
L1	A	Family	≤ 0.150	~85%
L2	B	Subfamily	≤ 0.100	~90%
L3	C	Cluster	≤ 0.050	~95%
L4	D	Subcluster	≤ 0.020	~98%
L5	E	Clone	≤ 0.010	~99%
L6	F	Strain	≤ 0.001	~99.9%

Example pLIN Code

1 . 1 . 3 . 5 . 12 . 45

B. Classification Algorithm

- 1
Parse FASTA Sequences
BioPython SeqIO reader
- 2
Compute 4-mer Vectors
256-dim normalised frequency
- 3
Cosine Distance Matrix
Pairwise $d(i,j) = 1 - \cos(V_i,V_j)$
- 4
Single-Linkage Clustering
`scipy.cluster.hierarchy.linkage`
- 5
Cut at 6 Thresholds
fcluster at L1-L6 distances
- 6
Assign pLIN Codes
Concatenate: A.B.C.D.E.F

Key Properties

Hierarchical | Permanent | Reference-Free | Composition-Based | Scalable

Figure 10. Comparison of Plasmid Classification Systems

	Hierarchical	Permanent Codes	Reference-Free	Broad Scope	AMR Integration	ANI Validation	GUI
PlasmidFinder	✗	✓	✗	✗	✗	✗	✗
pMLST	✗	✓	✗	✗	✗	✗	✗
MOB-suite	✗	✗	✗	~	✗	✗	✗
COPLA/PTU	~	✗	✗	~	✗	✗	✗
mge-cluster	✗	✗	✓	✓	✗	✗	✗
pLIN	✓	✓	✓	✓	✓	✓	✓

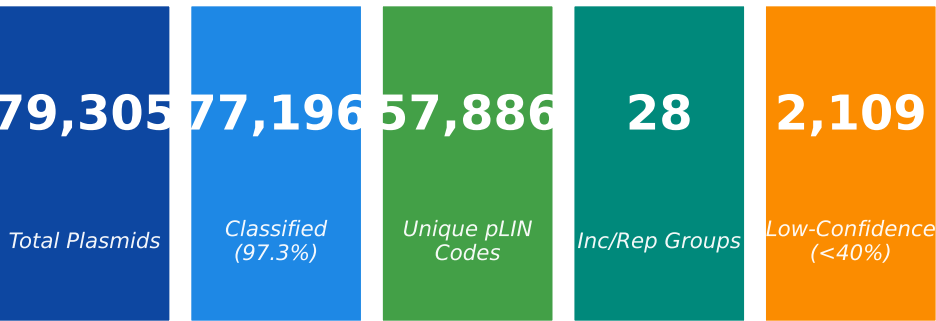
✓ = Full support

~ = Partial support

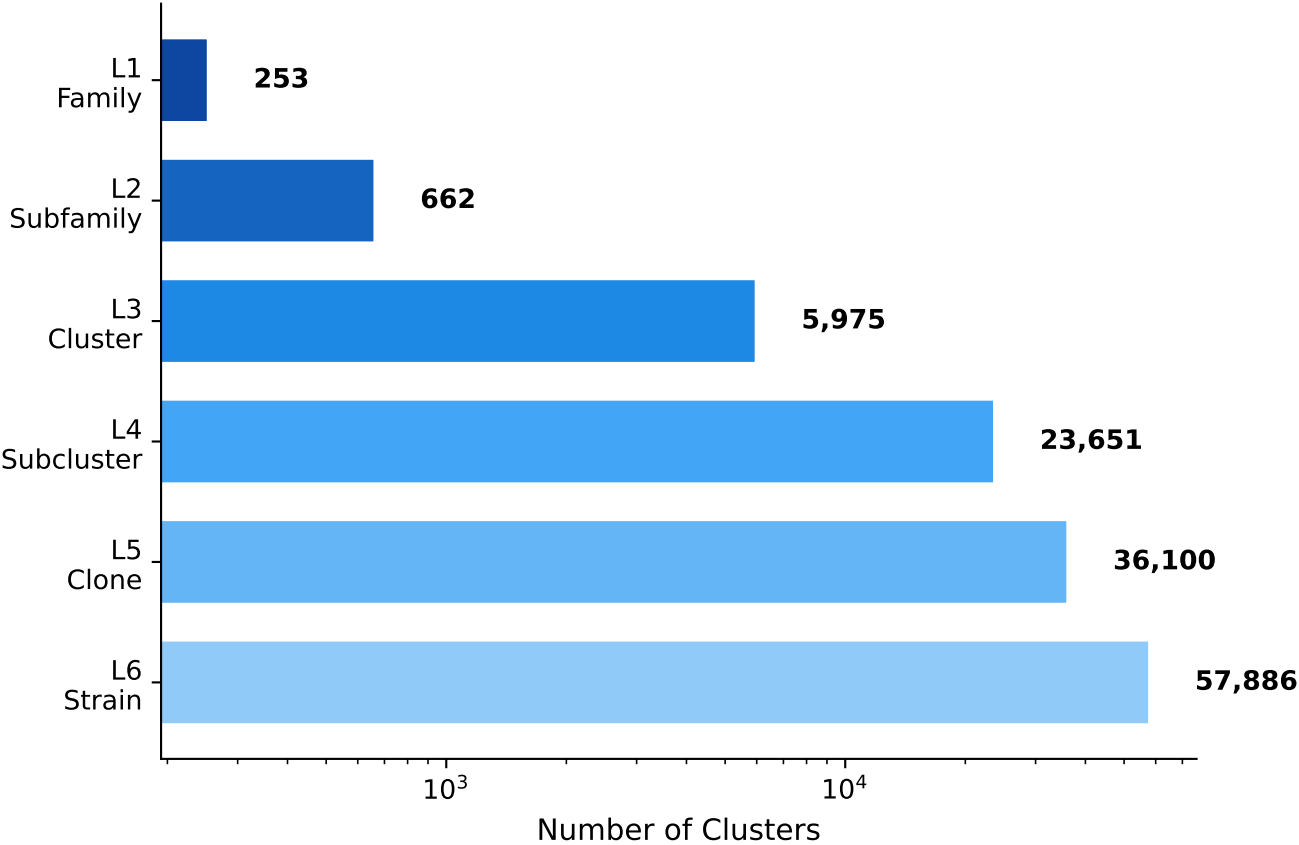
✗ = Not supported

Figure 10. Reference Database — Full pLIN Assignment

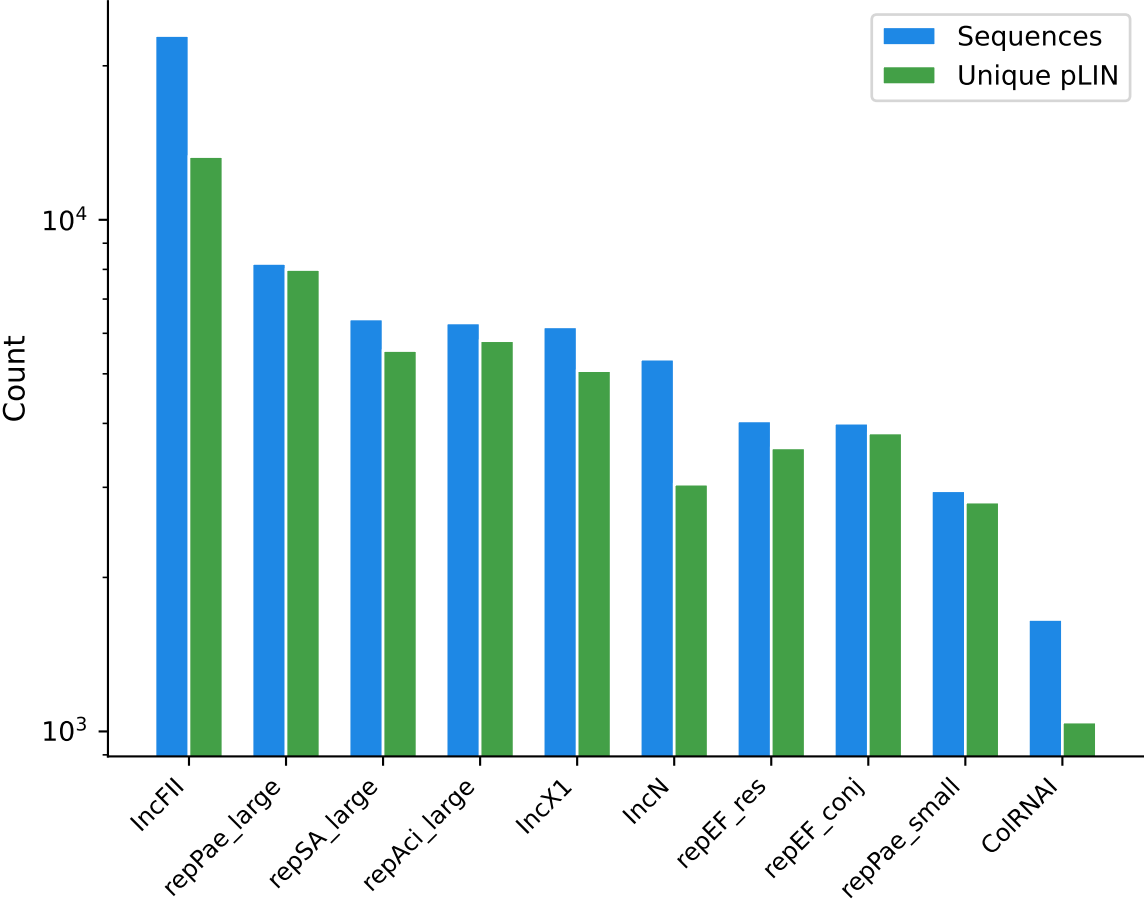
(A) Key Metrics



(B) Hierarchical Resolution



(C) Top 10 Inc/Rep Group Distribution



(D) Summary Statistics

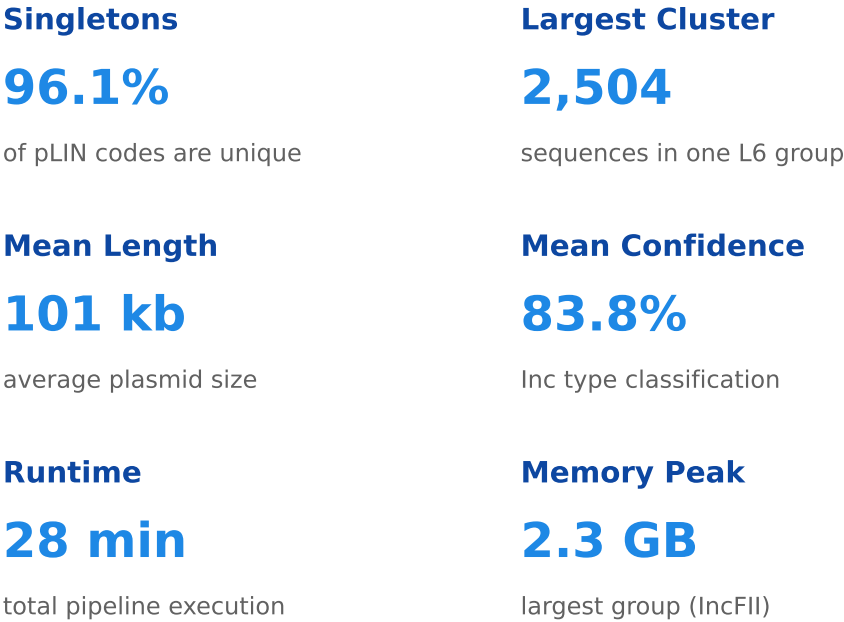
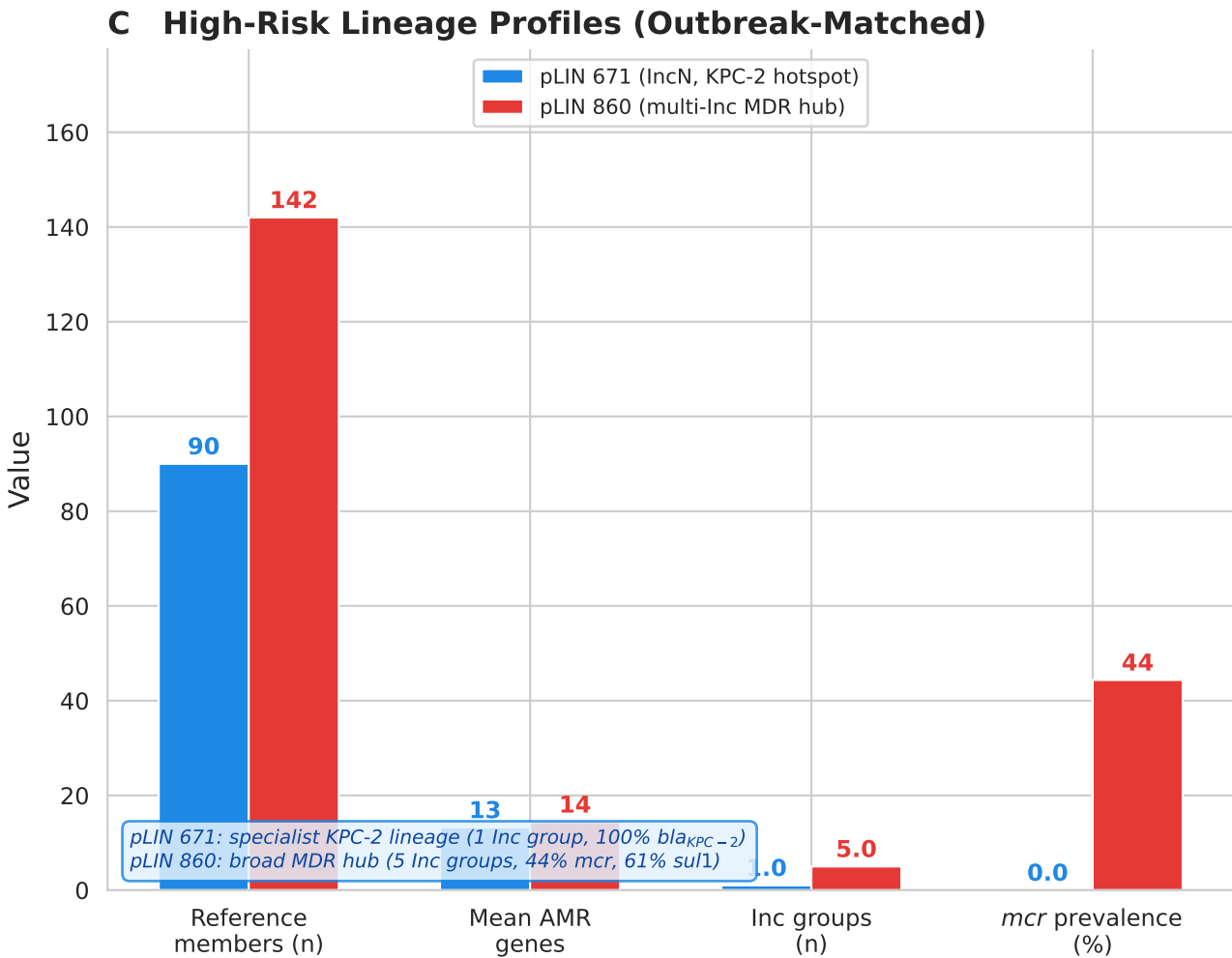
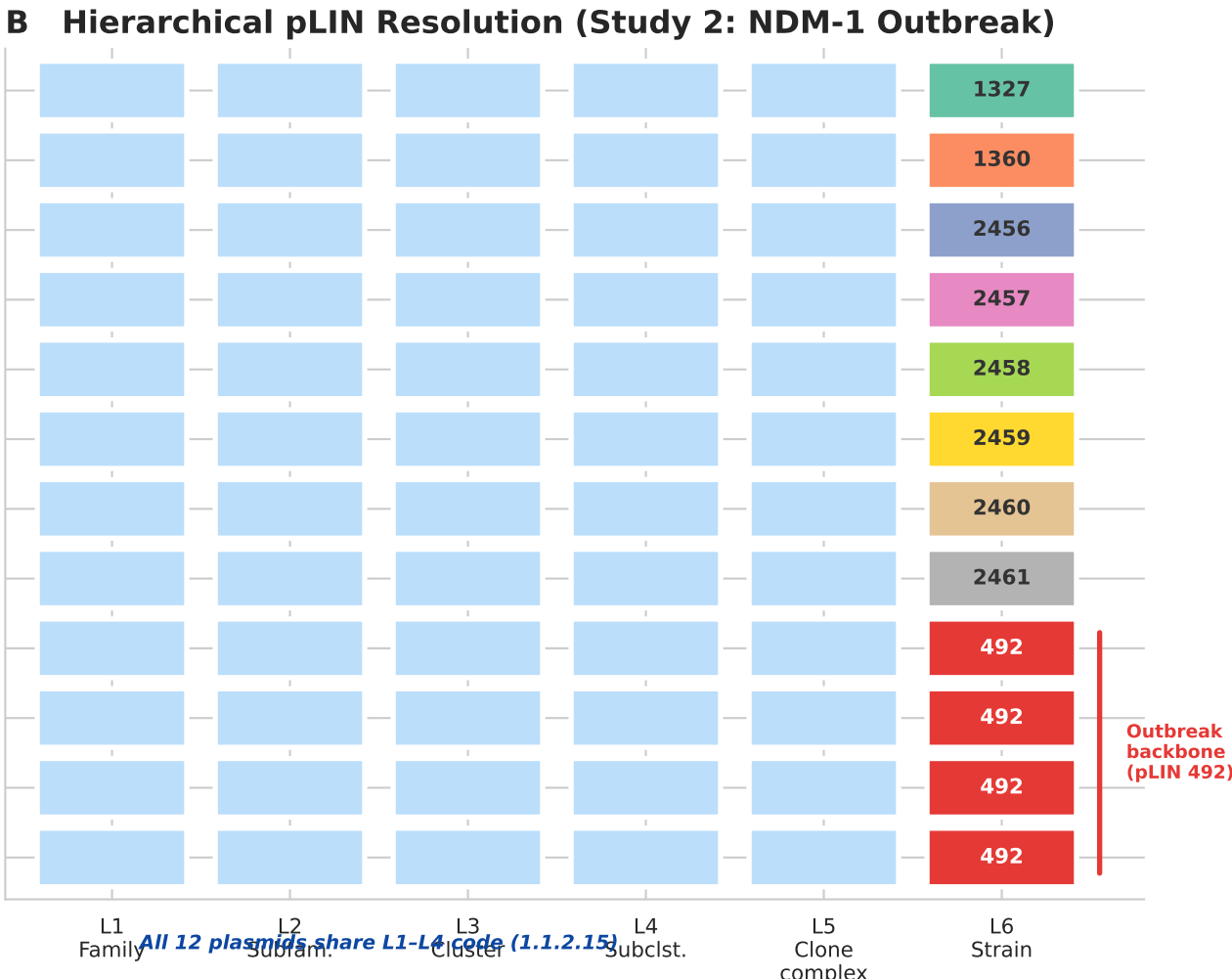
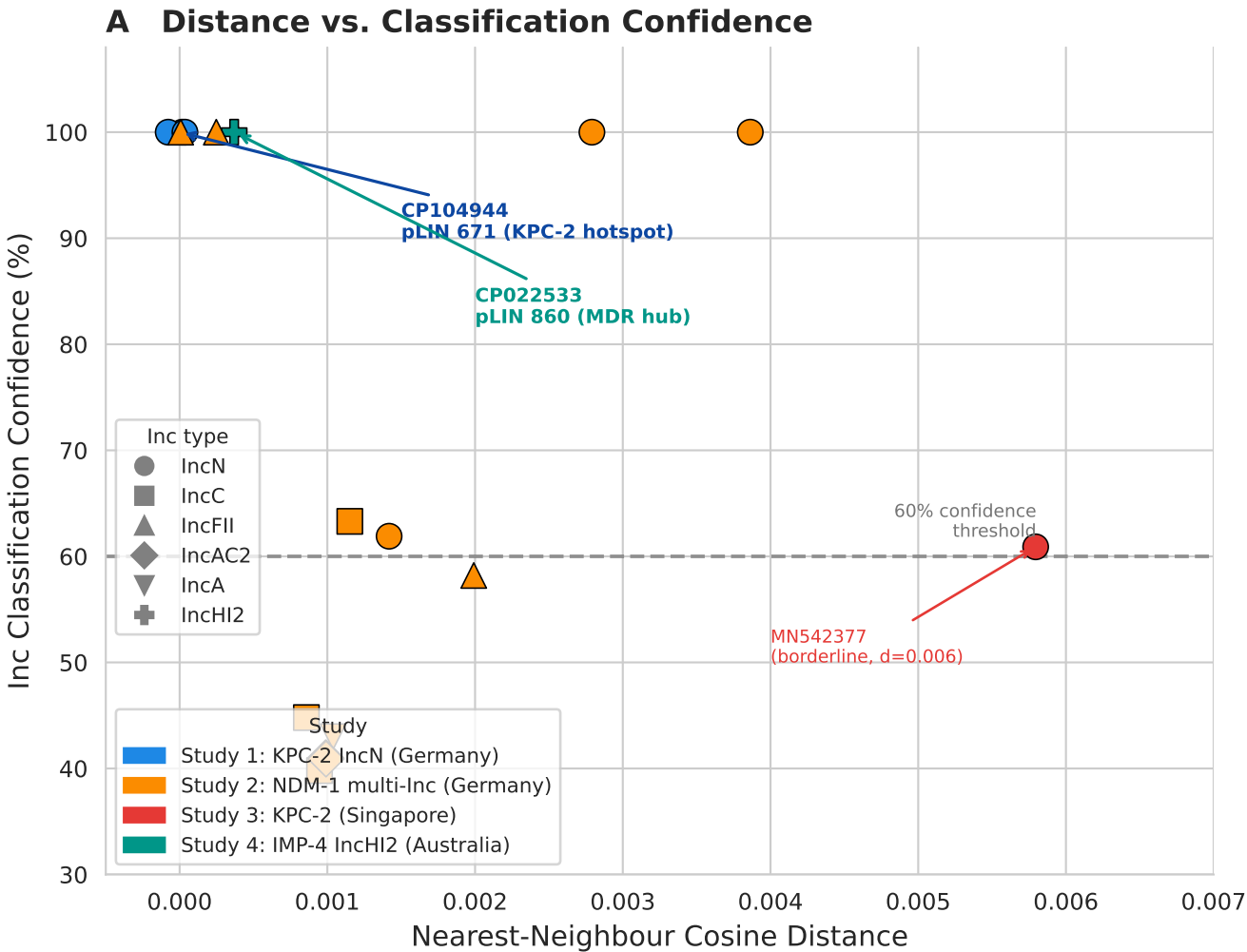


Figure 11. Cross-Validation Against Published Outbreak Datasets



D Validation Summary

Criterion	Result
Total outbreak plasmids tested	17
Published outbreak studies	4
Unique pLIN L6 codes assigned	13
Novel codes (new L6 variants)	9
Known high-risk lineage matches	2
Inc accuracy (confidence $\geq 60\%$)	100% (8/8)
Outbreak backbone grouping	4 plasmids $\rightarrow$ pLIN 492
Mean NN distance	0.0013
Cross-species transmission	Detected
Hierarchical resolution	12 $\rightarrow$ 9 L6 $\rightarrow$ 1 L3

Figure 12. Resolution Comparison Across Plasmid Classification Methods

74 outbreak plasmids from 26 studies, 13 countries, 7 resistance mechanisms

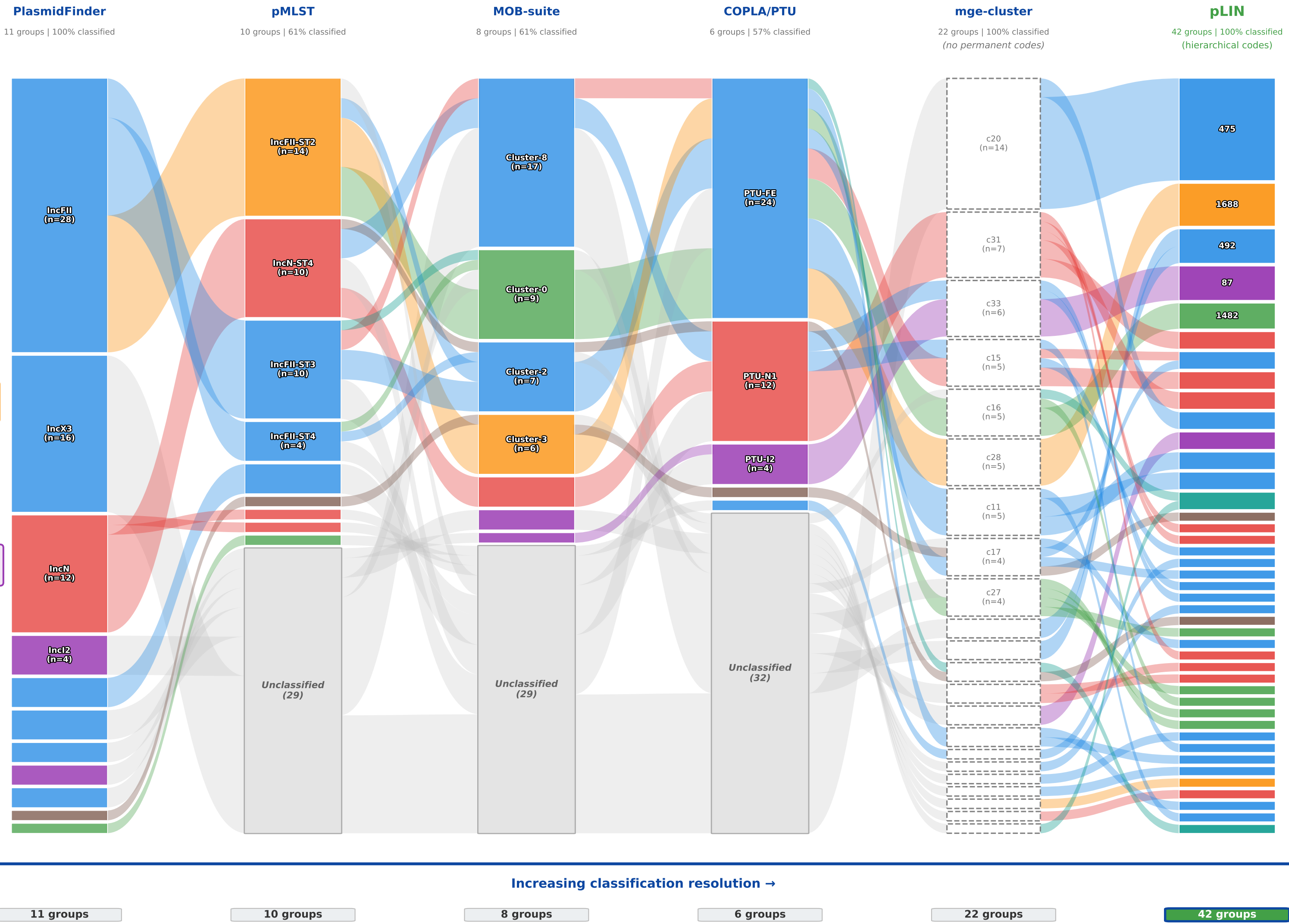
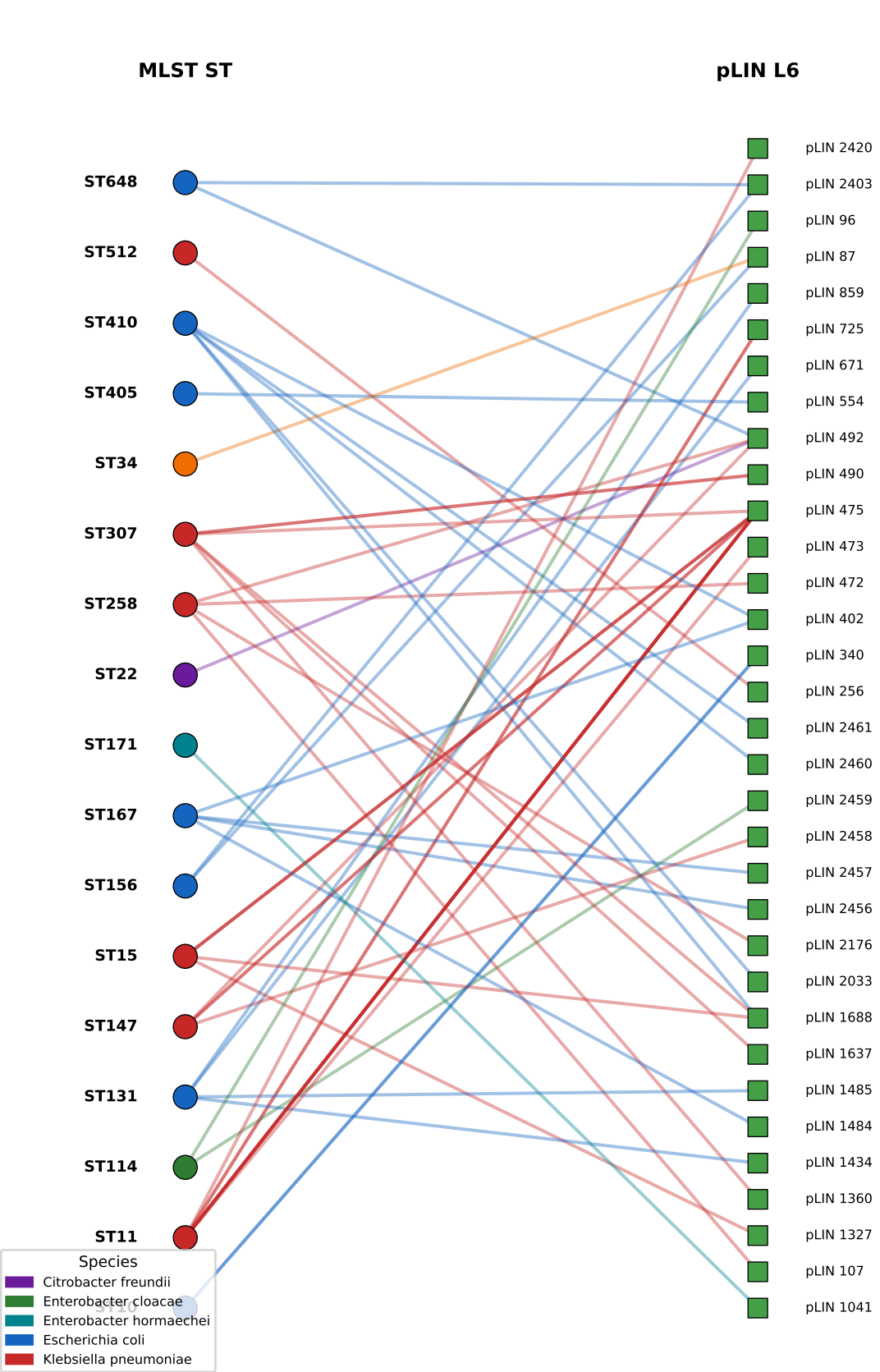
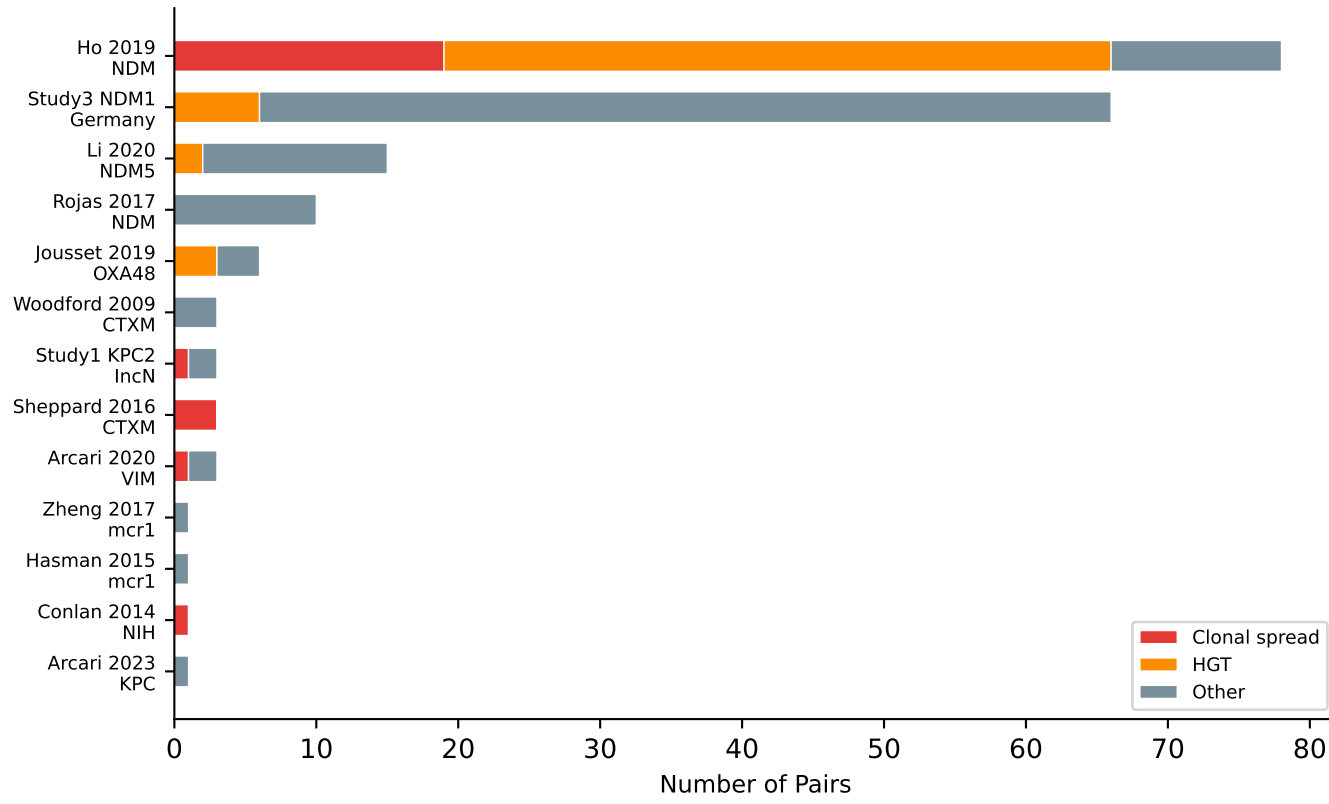


Figure 13. Combined Chromosomal-Plasmid Typing: Transmission Mode Discrimination

A. MLST ST — pLIN L6 Bipartite Network



B. Transmission Mode by Study



C. Validation Summary

Metric	Value
Studies evaluated	13
Host species	7
Unique MLST STs	20
Overall concordance	12/13 (92.3%)
Multi-ST detection	13/13 (100.0%)
Clonal detection	12/13 (92.3%)
Studies with HGT pairs	4
Studies with clonal pairs	5

Figure 14: Analytical Module Validation Summary — Expanded Classifier, Quality Assessment, Novelty Detection, and Mobile Genetic Element Analysis

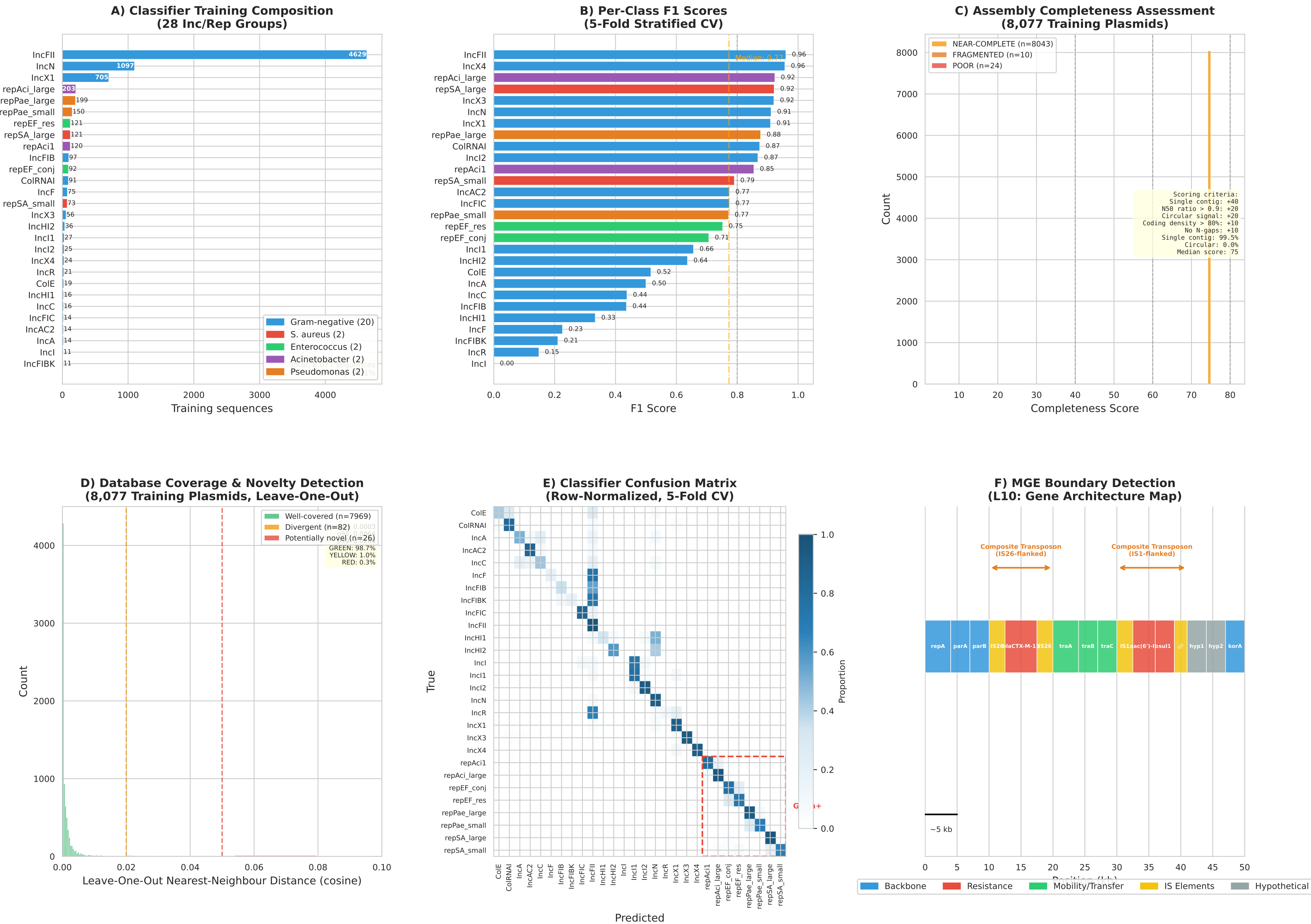
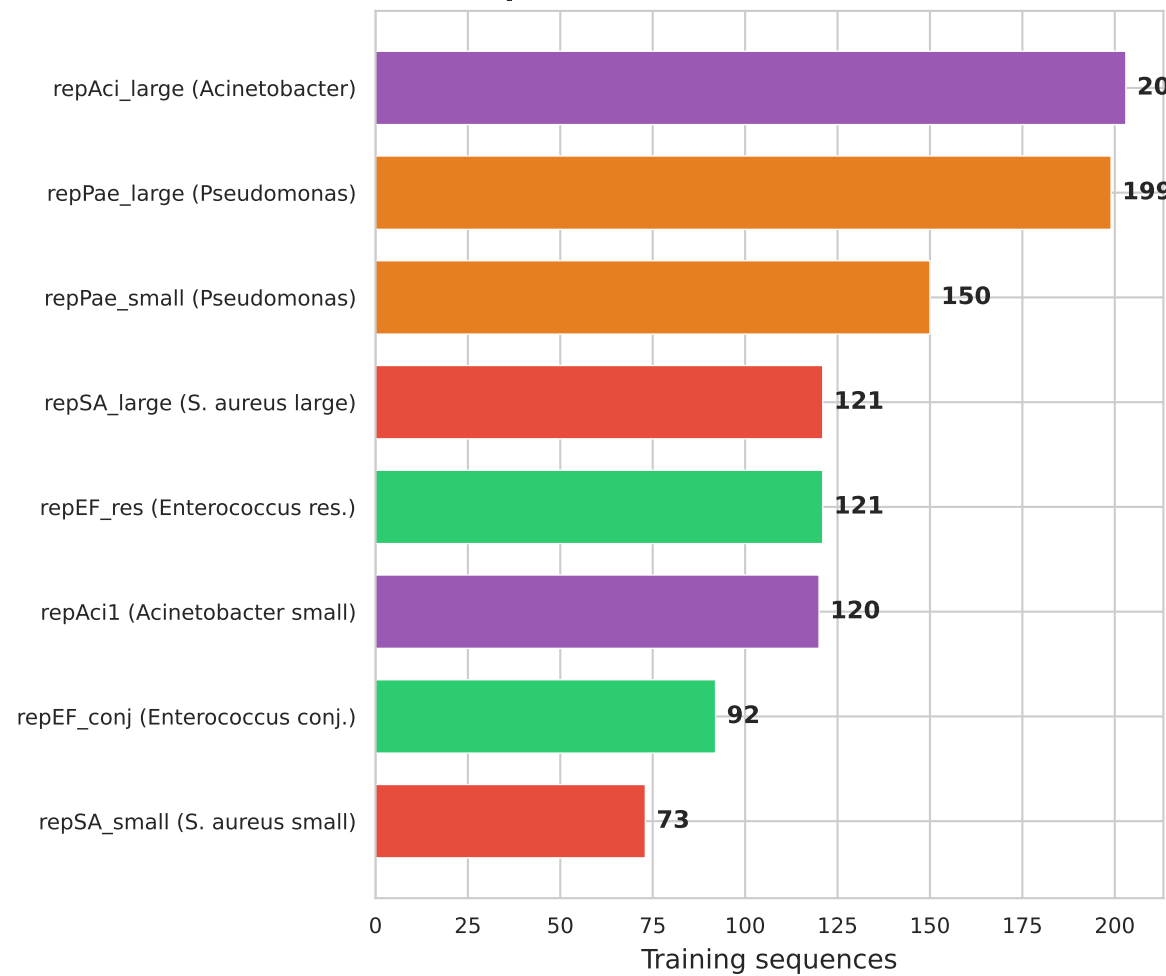
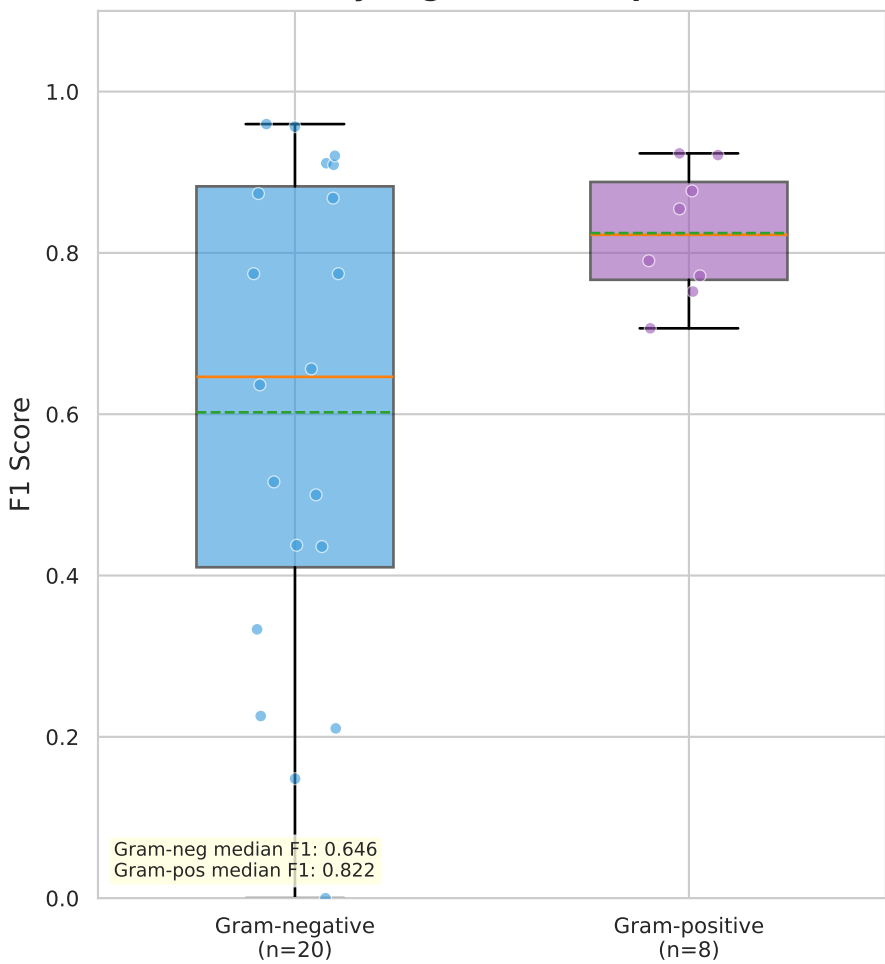


Figure 15: Novel Inc/Rep Group Discovery — Training Data, Performance, and Taxonomic Separation

A) Expanded Rep Type Groups
(Gram-positive, Acinetobacter, Pseudomonas)



B) Classification Performance
by Organism Group



C) Centroid Distances
(Gram-pos vs Selected Gram-neg)

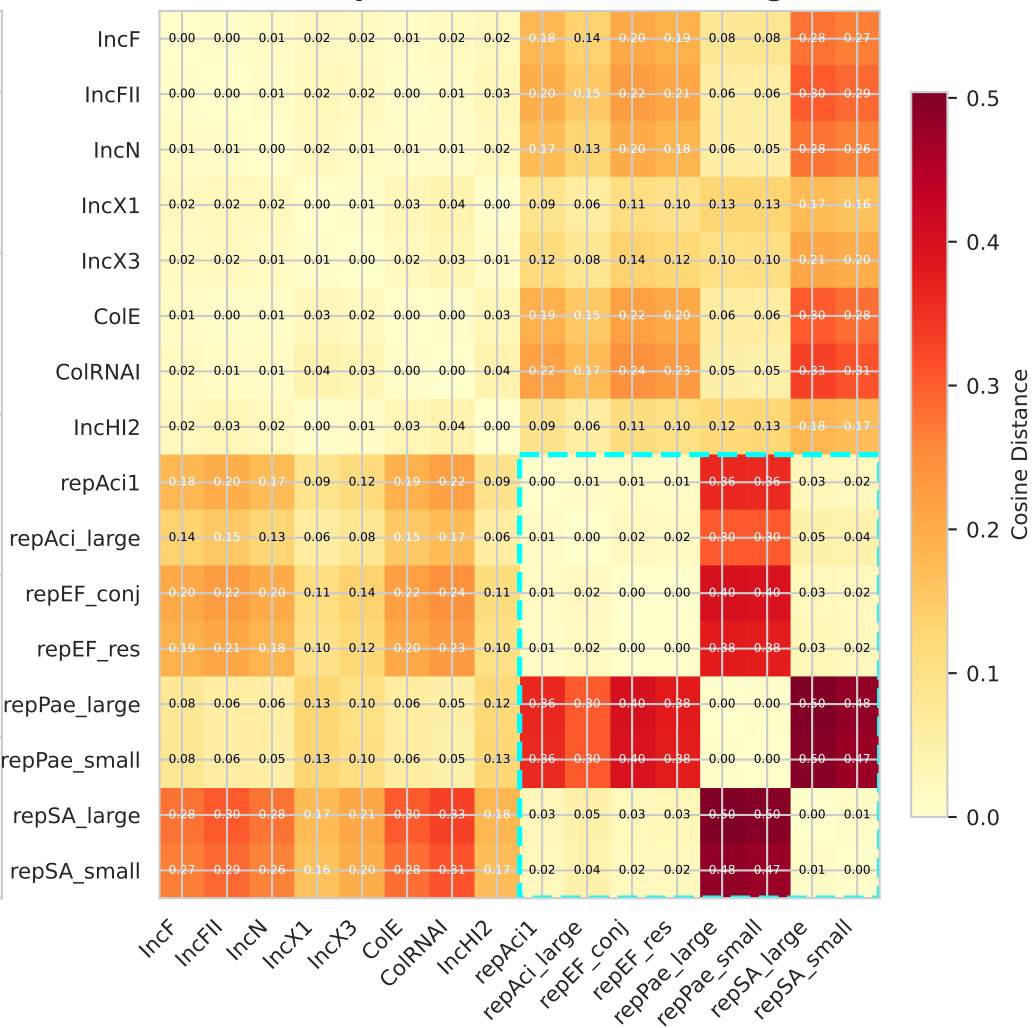


Figure 16: Adaptive Thresholds and Cluster Stability

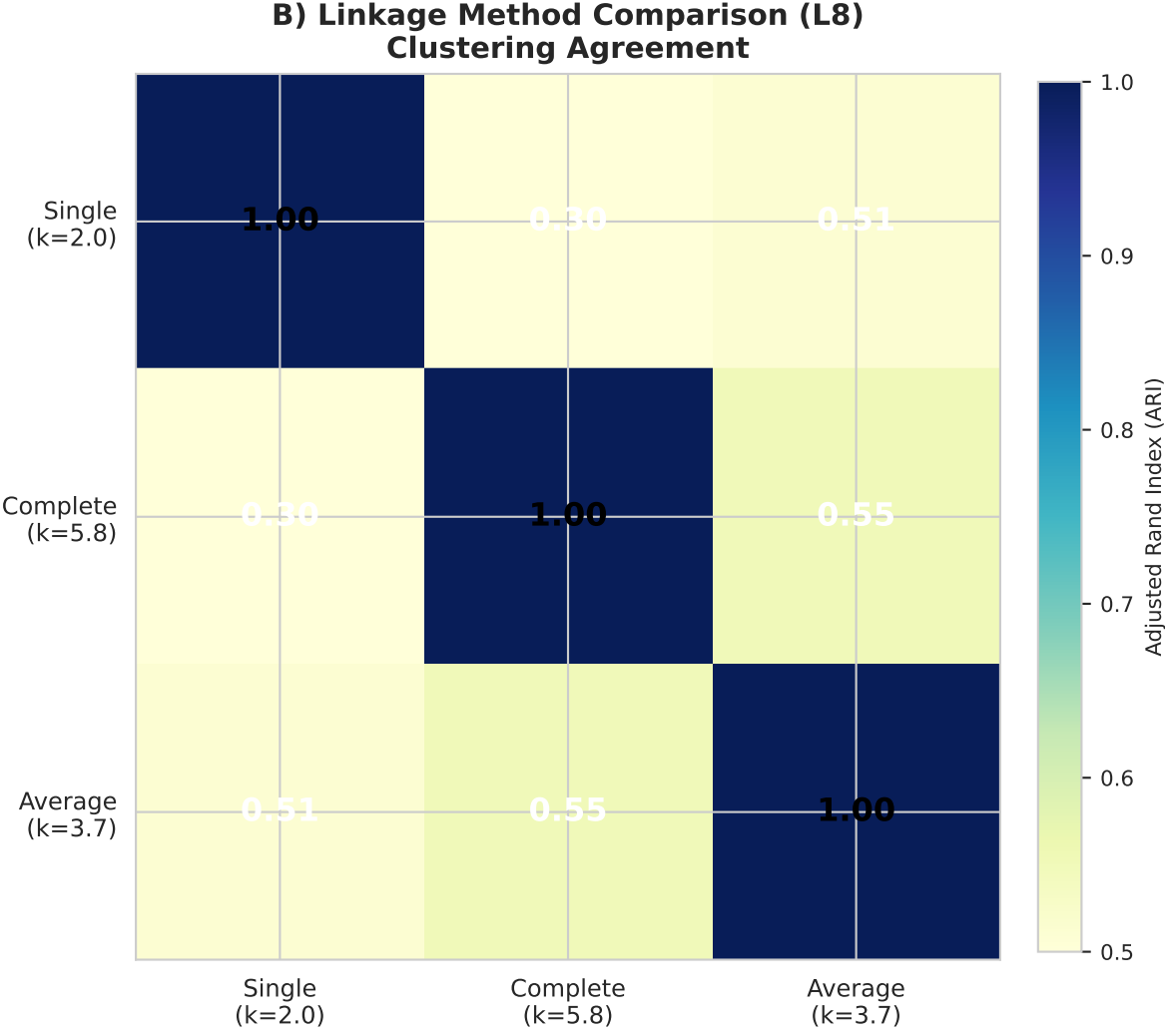
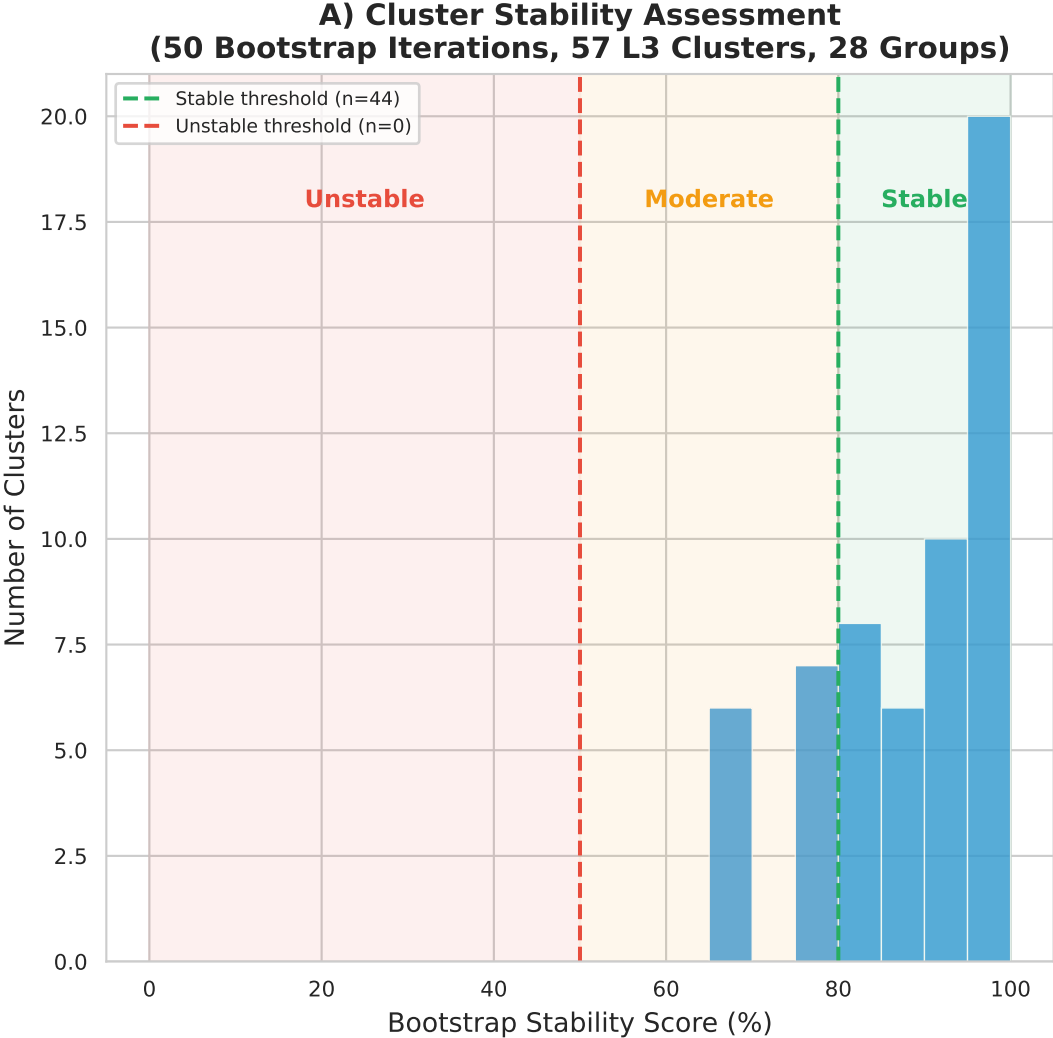
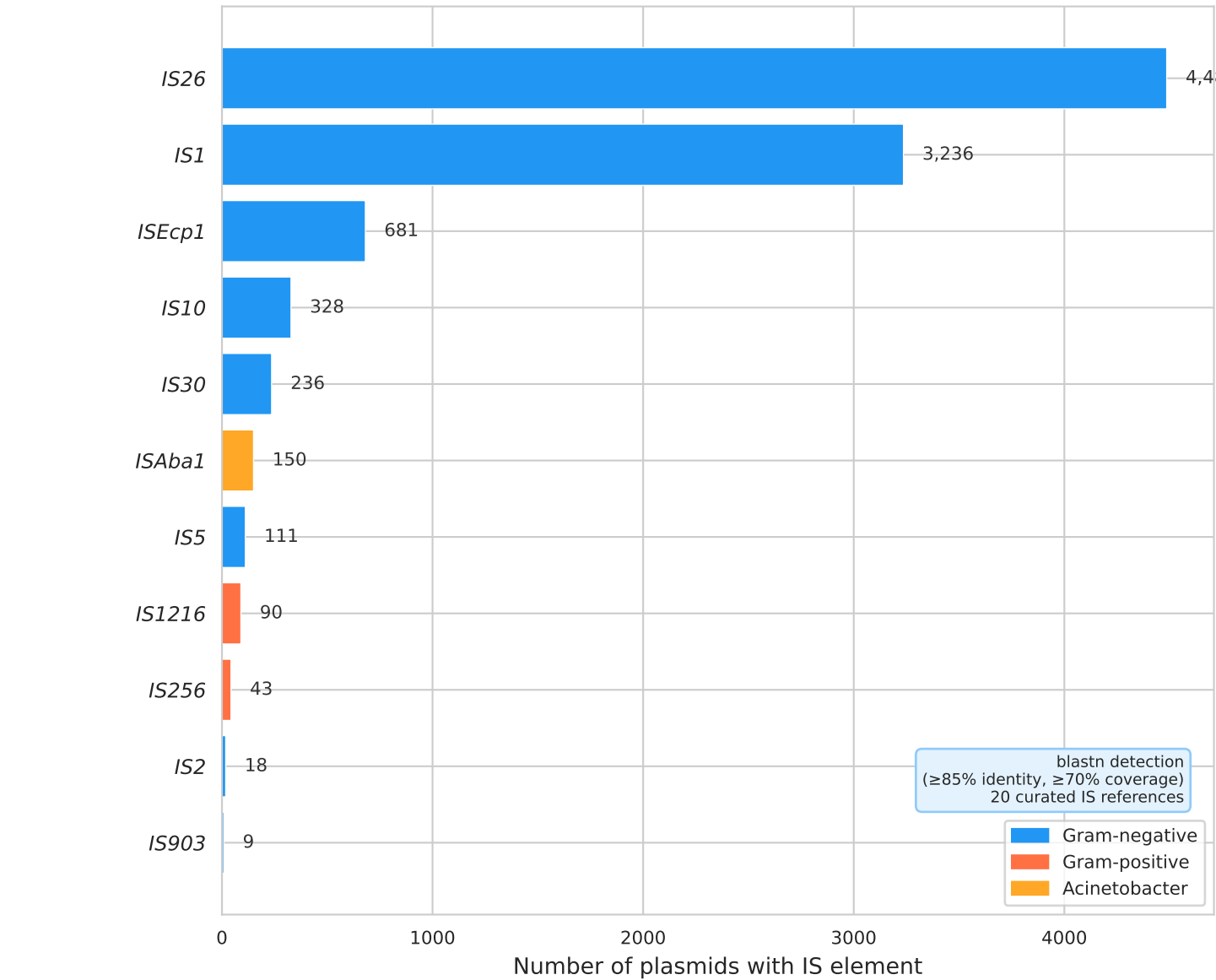
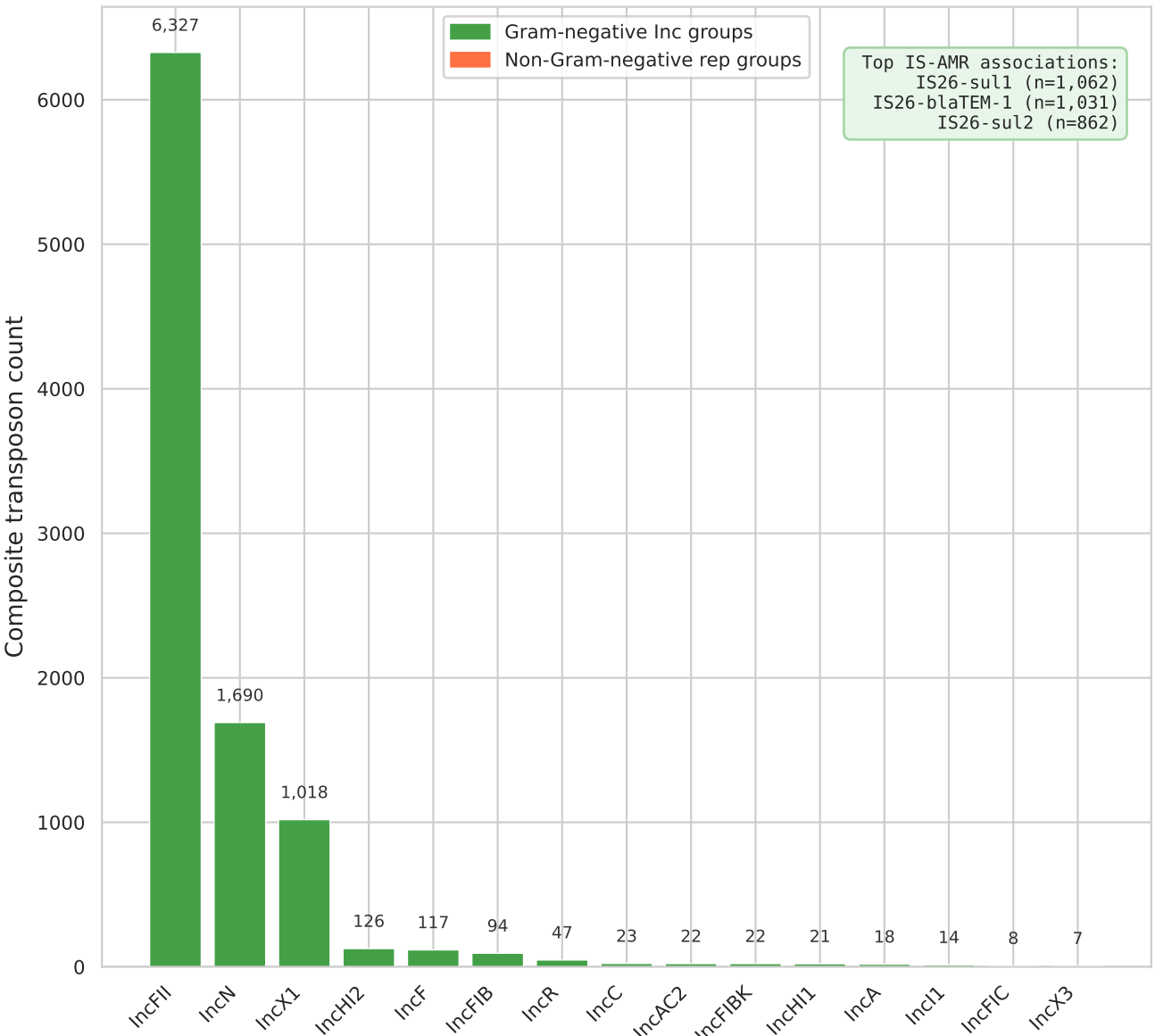


Figure 17: Mobile Genetic Element Boundary Detection

A) IS Element Detection Summary
(5,370/8,077 plasmids, 66.5%)



C) Composite Transposon Frequency by Inc/Rep Group
(9,559 total across 8,077 plasmids)



B) Representative Composite Transposon Structures

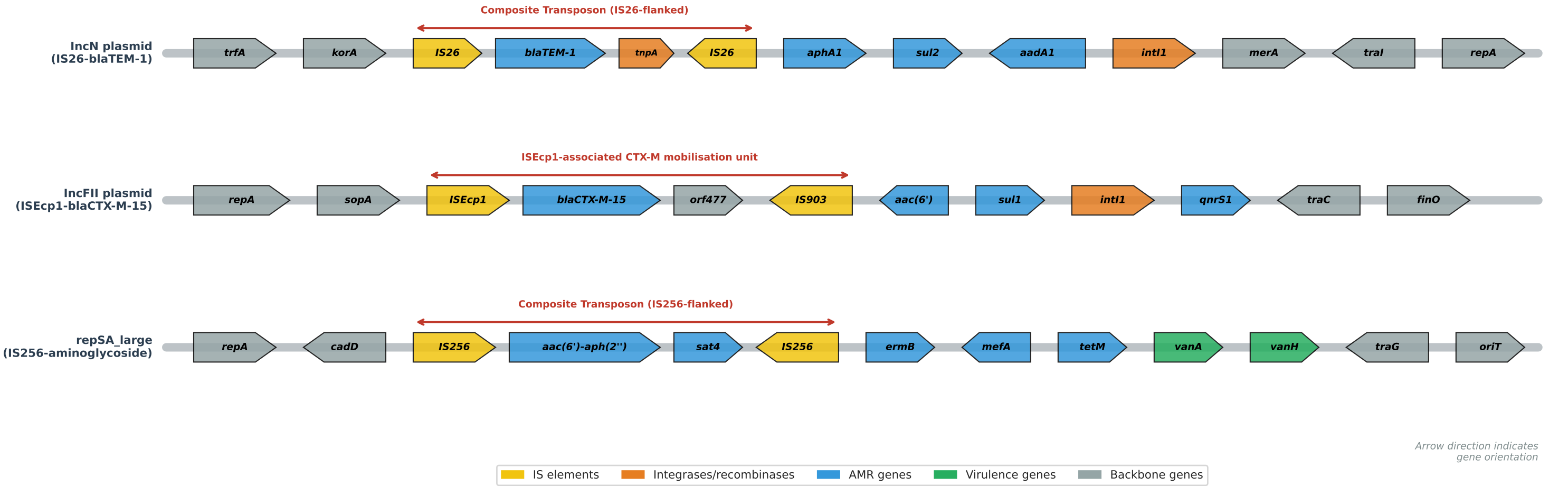
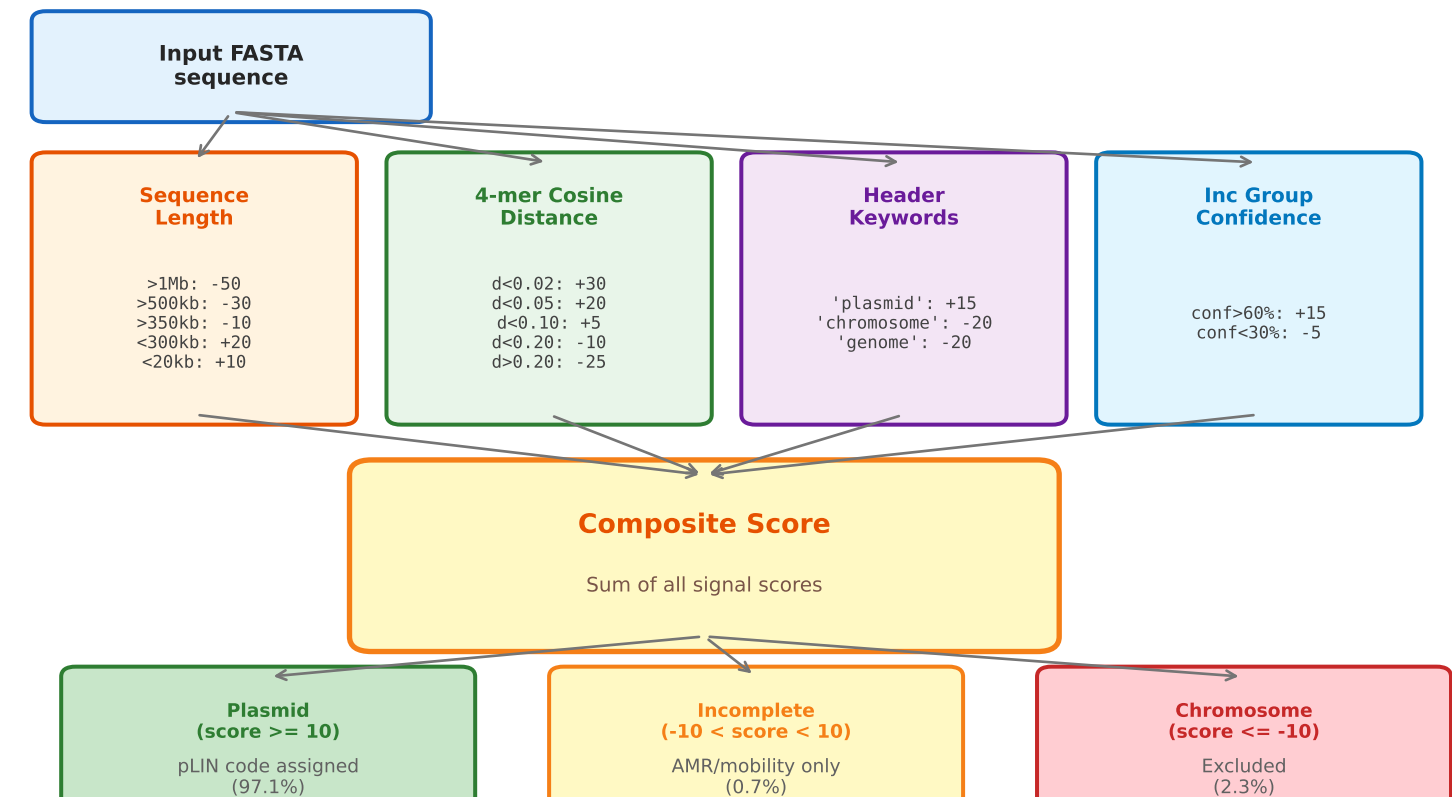
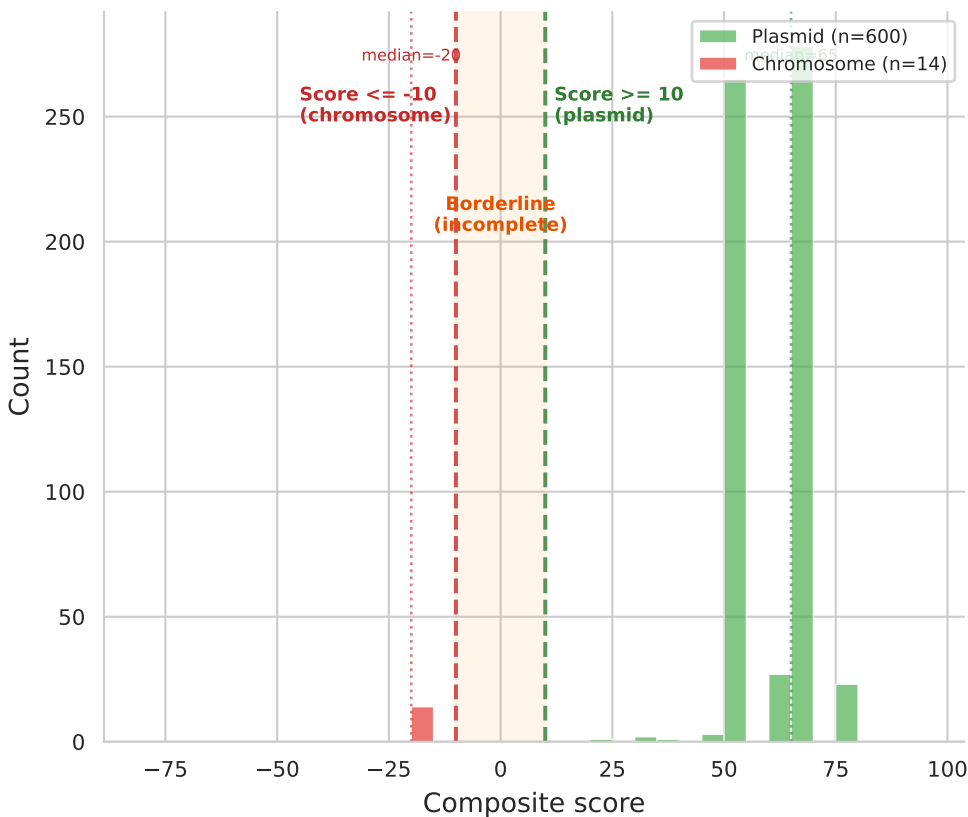


Figure 18. Plasmid vs. Chromosome Contig Classification

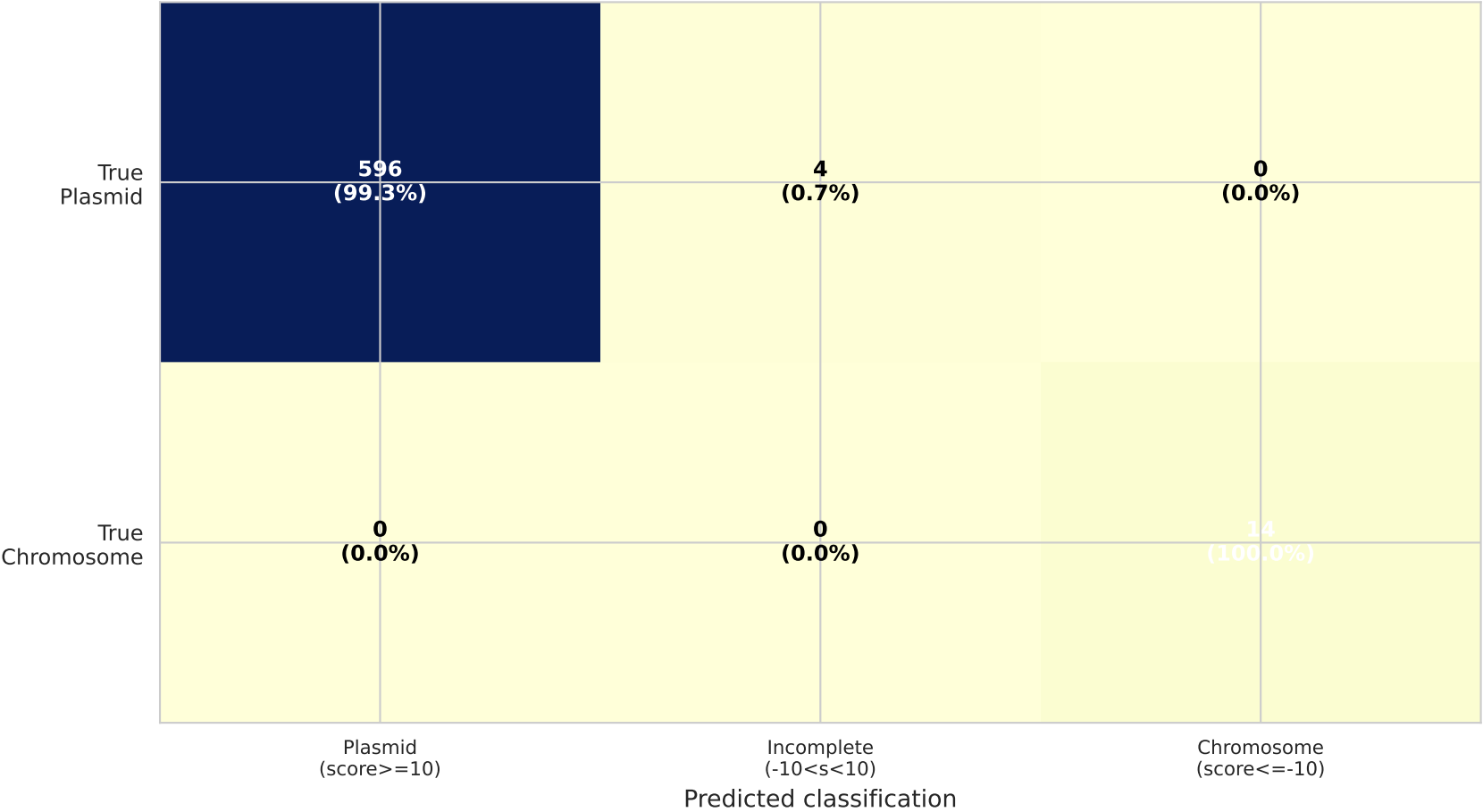
A Multi-signal scoring system



B Score distribution (n=614)

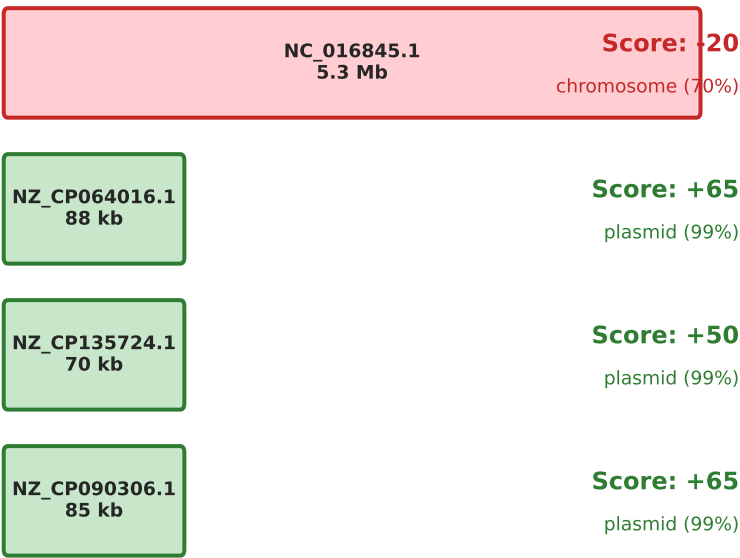


C Classification results



Sensitivity: 100.0% | Specificity: 100.0% | Accuracy: 100.0%

D Application example: K. pneumoniae assembly



Chromosome correctly excluded | All plasmid contigs retained for pLIN assignment

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [pLINManuscriptTables16.xlsx](#)
- [pLINSupplementaryDataTables.xlsx](#)
- [SupplementaryTableS3Perclassclassifierperformance.tsv](#)
- [SupplementaryTableS1AccessionNumbers.txt](#)