

Supplementary Material

Table S1. Training and testing data sets. Number and total length of contigs with different labels. Only contigs with length greater than 100bp are shown.

Species	Samples #	Chromosome # Mbp	Plasmid # Mbp	Ambig. # Mbp	Unknown # Mbp
ESKAPEE training data					
ALL	140	25231	595.2	3845	22.7
(%)	140	73.7%	93.0%	11.2%	3.5%
<i>Enterococcus faecium</i>	20	3019	51.8	1191	4.5
<i>Staphylococcus aureus</i>	20	1578	55.8	52	0.4
<i>Klebsiella pneumoniae</i>	20	5991	101.0	1209	6.8
<i>Acinetobacter baumannii</i>	20	5238	74.8	747	3.8
<i>Pseudomonas aeruginosa</i>	20	5126	134.2	80	1.1
<i>Enterobacter spp.</i>	20	1991	81.8	328	3.2
<i>Escherichia coli</i>	20	2288	95.8	238	3.0
ESKAPEE testing data					
ALL	224	32761	936.8	3970	32.6
(%)	224	76.0%	90.4%	9.2%	3.1%
<i>Enterococcus faecium</i>	4	361	10.0	105	0.9
<i>Staphylococcus aureus</i>	62	3924	168.6	194	1.4
<i>Klebsiella pneumoniae</i>	92	21720	453.8	2587	21.8
<i>Acinetobacter baumannii</i>	10	1254	46.9	156	1.2
<i>Pseudomonas aeruginosa</i>	10	1165	65.4	0	0.0
<i>Enterobacter spp.</i>	30	1861	123.5	358	4.3
<i>Escherichia coli</i>	16	2476	68.6	570	3.1
Non-ESKAPEE testing data, close					
ALL	320	28989	1518.2	7901	65.3
(%)	320	69.0%	91.6%	18.8%	3.9%
<i>Citrobacter freundii</i>	100	8314	478.8	2227	16.8
<i>Escherichia fergusonii</i>	100	7234	415.6	1077	14.9
<i>Klebsiella oxytoca</i>	62	6923	348.7	3463	23.8
<i>Salmonella enterica</i>	58	6518	275.0	1134	9.8
Non-ESKAPEE testing data, distant					
ALL	188	46260	775.9	4260	30.7
(%)	188	87.3%	94.9%	8.0%	3.8%
<i>Mycobacteriaceae</i>	60	6296	338.3	222	6.2
<i>Campylobacter jejuni</i>	60	3521	100.8	238	3.7
<i>Bacillus spp.</i>	68	36443	336.8	3800	20.9

Table S2. Reference genomes for the homology-based gold-standard labeling

Accession	Description
NZ_CP038996.1	<i>Enterococcus faecium</i> strain SRR24
NC_007795.1	<i>Staphylococcus aureus</i> subsp. <i>aureus</i> NCTC 8325
NC_016845.1	<i>Klebsiella pneumoniae</i> subsp. <i>pneumoniae</i> HS11286
NZ_CP043953.1	<i>Acinetobacter baumannii</i> strain K09-14
NC_002516.2	<i>Pseudomonas aeruginosa</i> PAO1
NZ_CP009756.1	<i>Enterobacter cloacae</i> strain GGT036
NC_000913.3	<i>Escherichia coli</i> str. K-12 substr. MG1655
NC_002695.2	<i>Escherichia coli</i> O157:H7 str. Sakai DNA
NZ_AP019632.1	<i>Enterobacter asburiae</i> strain 1808-013
NZ_AP022508.1	<i>Enterobacter bugandensis</i> strain STN0717-56
NZ_CP081105.1	<i>Enterobacter cancerogenus</i> strain JY65
NZ_CP077392.1	<i>Enterobacter hormaechei</i> strain FDAARGOS 1435
NZ_CP017184.1	<i>Enterobacter roggenkampii</i> strain DSM 16690
NZ_CP044060.1	<i>Aeromonas veronii</i> strain FDAARGOS_632
NC_007530.2	<i>Bacillus anthracis</i> str. 'Ames Ancestor'
NC_002163.1	<i>Campylobacter jejuni</i> subsp. <i>jejuni</i> NCTC 11168 = ATCC 700819
NZ_CP033744.1	<i>Citrobacter freundii</i> strain FDAARGOS_549
NC_009495.1	<i>Clostridium botulinum</i> A str. ATCC 3502
NZ_CP057657.1	<i>Escherichia fergusonii</i> strain RHB19-C05
NZ_CP041925.1	<i>Klebsiella aerogenes</i> strain Ka37751
NZ_AP022547.1	<i>Klebsiella michiganensis</i> strain THO-011
NZ_CP033844.1	<i>Klebsiella oxytoca</i> strain FDAARGOS_500
NZ_CP065838.1	<i>Klebsiella quasipneumoniae</i> strain KqPF26
NC_003210.1	<i>Listeria monocytogenes</i> EGD-e
NZ_CP069288.1	<i>Priestia megaterium</i> strain ATCC 14581
NC_003197.2	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Typhimurium str. LT2
NC_004603.1	<i>Vibrio parahaemolyticus</i> RIMD 2210633 chromosome 1
NC_004605.1	<i>Vibrio parahaemolyticus</i> RIMD 2210633 chromosome 2
NC_017168.1	<i>Yersinia pestis</i> A1122
NZ_AP022570.1	<i>Mycolicibacterium poriferae</i> strain JCM 12603
NZ_AP022579.1	<i>Mycolicibacterium boenickei</i> strain JCM 15653
NZ_CP023147.1	<i>Mycobacterium marseillense</i> strain FLAC0026
NZ_CP025546.1	<i>Mycobacterium paragordoniae</i> strain 49061
NZ_CP034181.1	<i>Mycobacteroides abscessus</i> strain GZ002
NZ_LT906469.1	<i>Mycolicibacter terrae</i> strain NCTC10856 chromosome 1
NC_002163.1	<i>Campylobacter jejuni</i> subsp. <i>jejuni</i> NCTC 11168 = ATCC 700819
NZ_CP046317.1	<i>Campylobacter coli</i> strain FDAARGOS_735
NZ_CP017060.1	<i>Bacillus cereus</i> strain FORC_047
NC_007530.2	<i>Bacillus anthracis</i> str. 'Ames Ancestor'
NC_000964.3	<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168
NZ_CM000753.1	<i>Bacillus thuringiensis</i> serovar berliner ATCC 10792

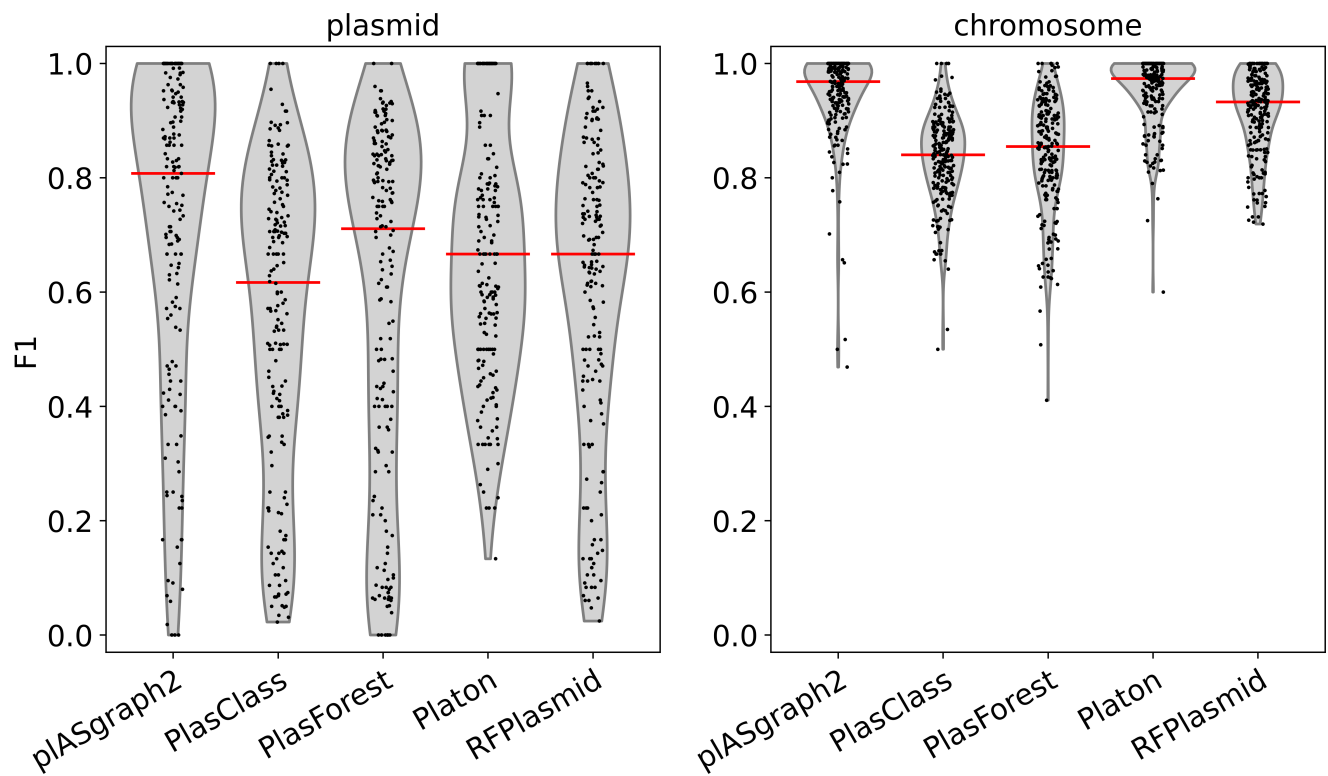


Figure S1. Comparison of F1-scores using samples from ESKAPEE species and considering all contigs longer than 100 bp. Each datapoint represents the F1-score of a single isolate. Median is shown as a horizontal line.

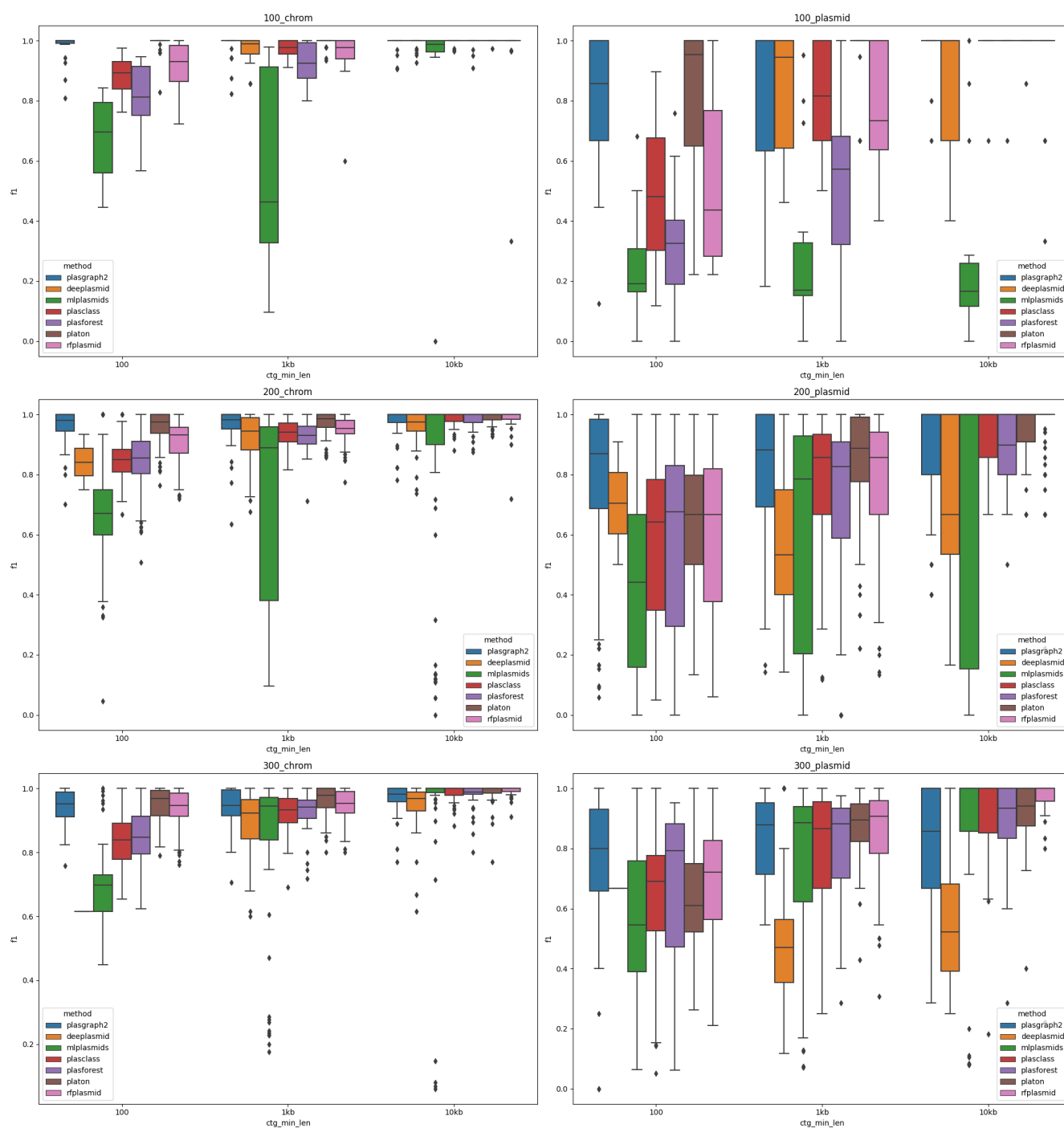


Figure S2. Comparison of F1-scores distribution over all samples for all methods considering only assemblies with a maximum number of 100, 200 or 300 contigs respectively. The plots on the left show the F1-scores for the chromosomal classification and the plots on the right show the results for the plasmid classification. From top to bottom the plots contain all samples that either considered maximally 100, 200 or 300 contigs. In total, only two ESKAPEE samples contained more than 300 contigs and are thus not considered in this figure. Within each plot the comparisons of all methods are shown considering only >100 bp, >1 kbp or >10 kbp contigs from left to right.

Table S3. Performance of contig classification algorithms on the ESKAPEE testing set using strict binary evaluation. The table shows the median values for each metric from among results on samples included in the testing set. The output of plASgraph2 was modified so that “ambiguous” predictions were replaced with the prediction of “plasmid” or “chromosome” depending on which option had a higher score. The contigs with “ambiguous” ground-truth label were removed from this evaluation. Fig. S3 shows the full distribution of F1 scores.

Method	SS	DB	AUROC	Precision	Recall	F1	Accuracy
A: Plasmid classification, contigs >100 bp, $n = 36,731$							
plASgraph2	–	–	0.995	0.961	0.900	0.827	0.955
mlplasmids	X	–	0.938	0.235	1.000	0.447	0.627
PlasClass	–	–	0.898	0.333	0.964	0.576	0.796
PlasForest	–	X	n/a	0.451	0.943	0.645	0.851
Platon	–	X	n/a	1.000	0.667	0.772	0.952
RFPlasmid	X	X	0.982	0.833	0.833	0.730	0.910
B: Chromosome classification, contigs >100 bp, $n = 36,731$							
plASgraph2	–	–	0.995	0.988	1.000	0.973	0.955
mlplasmids	X	–	0.938	1.000	0.568	0.718	0.627
PlasClass	–	–	0.898	1.000	0.771	0.861	0.796
PlasForest	–	X	n/a	0.992	0.855	0.894	0.851
Platon	–	X	n/a	0.954	1.000	0.973	0.952
RFPlasmid	X	X	0.982	0.982	0.974	0.942	0.907

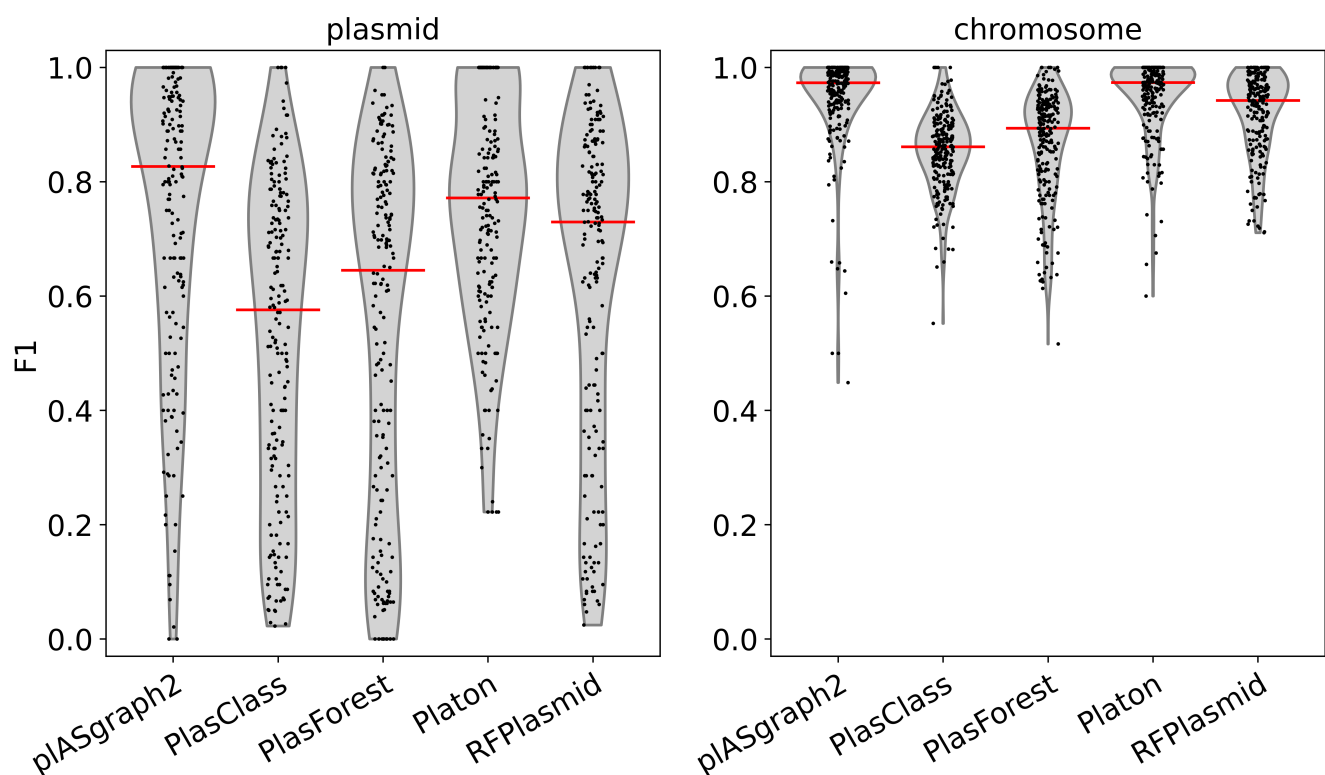


Figure S3. Comparison of F1-scores using samples from ESKAPEE species and considering all contigs longer than 100 bp using strict binary evaluation. Each datapoint represents the F1-score of a single isolate. Median is shown as a horizontal line.

Table S4. Performance of contig classification algorithms on the ESKAPEE testing set using ternary evaluation. The table shows the median values for each metric from among results on samples included in the testing set. Here, ambiguous contigs were considered as a class separate from chromosomes and plasmids. The accuracy of predictions of ambiguous class was evaluated only for plASgraph2, since no other tool predicts ambiguous contigs. Note that while F1-score of plASgraph2 in predicting ambiguous contigs is low, the task is very difficult since it is highly unbalanced (only 3.6% of contigs evaluated in this experiment are ambiguous). Fig. S4 shows the full distribution of F1 scores.

Method	SS	DB	Precision	Recall	F1	Accuracy
A: Plasmid classification, contigs >100 bp, $n = 38,110$						
plASgraph2	–	–	1.000	0.775	0.777	0.943
mlplasmids	X	–	0.223	1.000	0.420	0.609
PlasClass	–	–	0.302	0.964	0.522	0.766
PlasForest	–	X	0.333	0.943	0.549	0.795
Platon	–	X	1.000	0.667	0.764	0.952
RFPlasmid	X	X	0.685	0.833	0.667	0.894
B: Chromosome classification, contigs >100 bp, $n = 38,110$						
plASgraph2	–	–	0.987	0.984	0.959	0.935
mlplasmids	X	–	0.981	0.568	0.713	0.641
PlasClass	–	–	0.990	0.771	0.852	0.794
PlasForest	–	X	0.986	0.855	0.892	0.852
Platon	–	X	0.923	1.000	0.955	0.924
RFPlasmid	X	X	0.968	0.974	0.929	0.885
C: Ambiguous classification, contigs >100 bp, $n = 38,110$						
plASgraph2	–	–	0.167	0.500	0.400	0.941

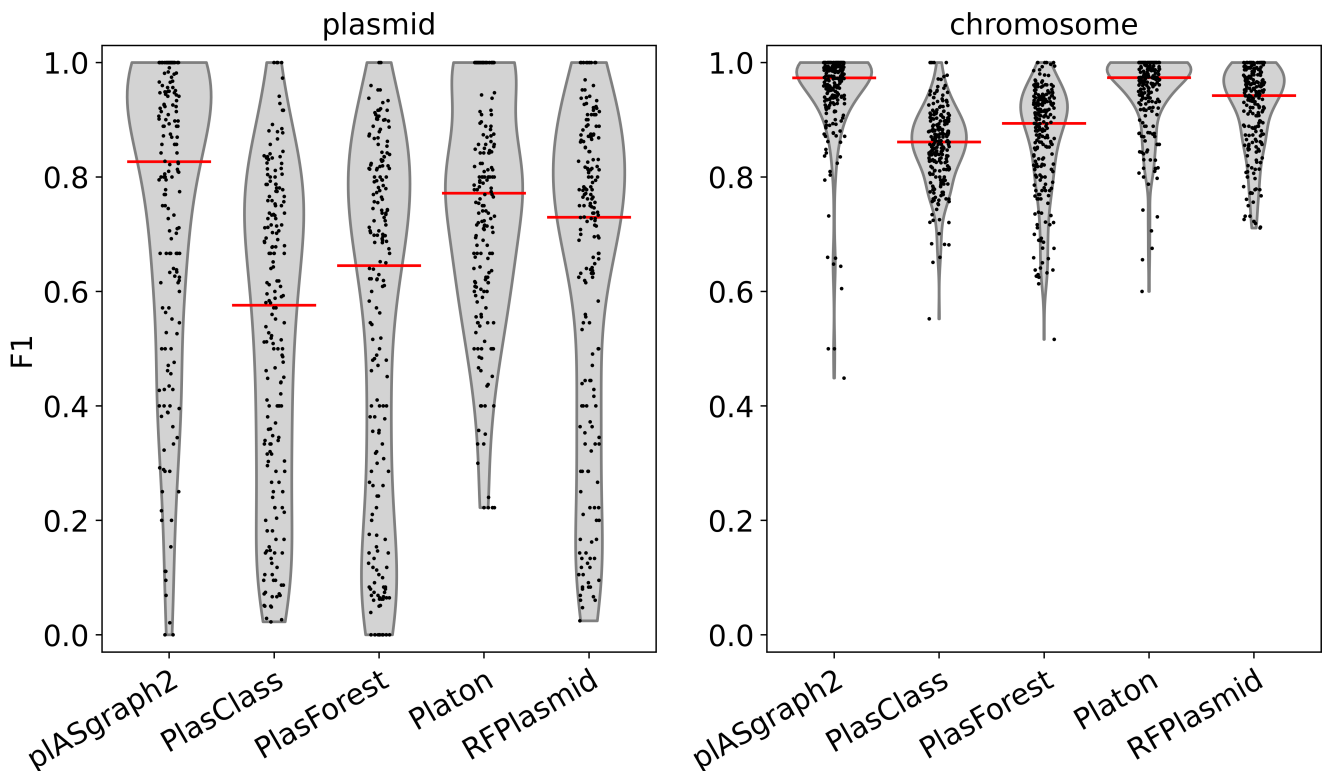


Figure S4. Comparison of F1-scores using samples from ESKAPEE species and considering all contigs longer than 100 bp using ternary evaluation. Each datapoint represents the F1-score of a single isolate. Median is shown as a horizontal line.

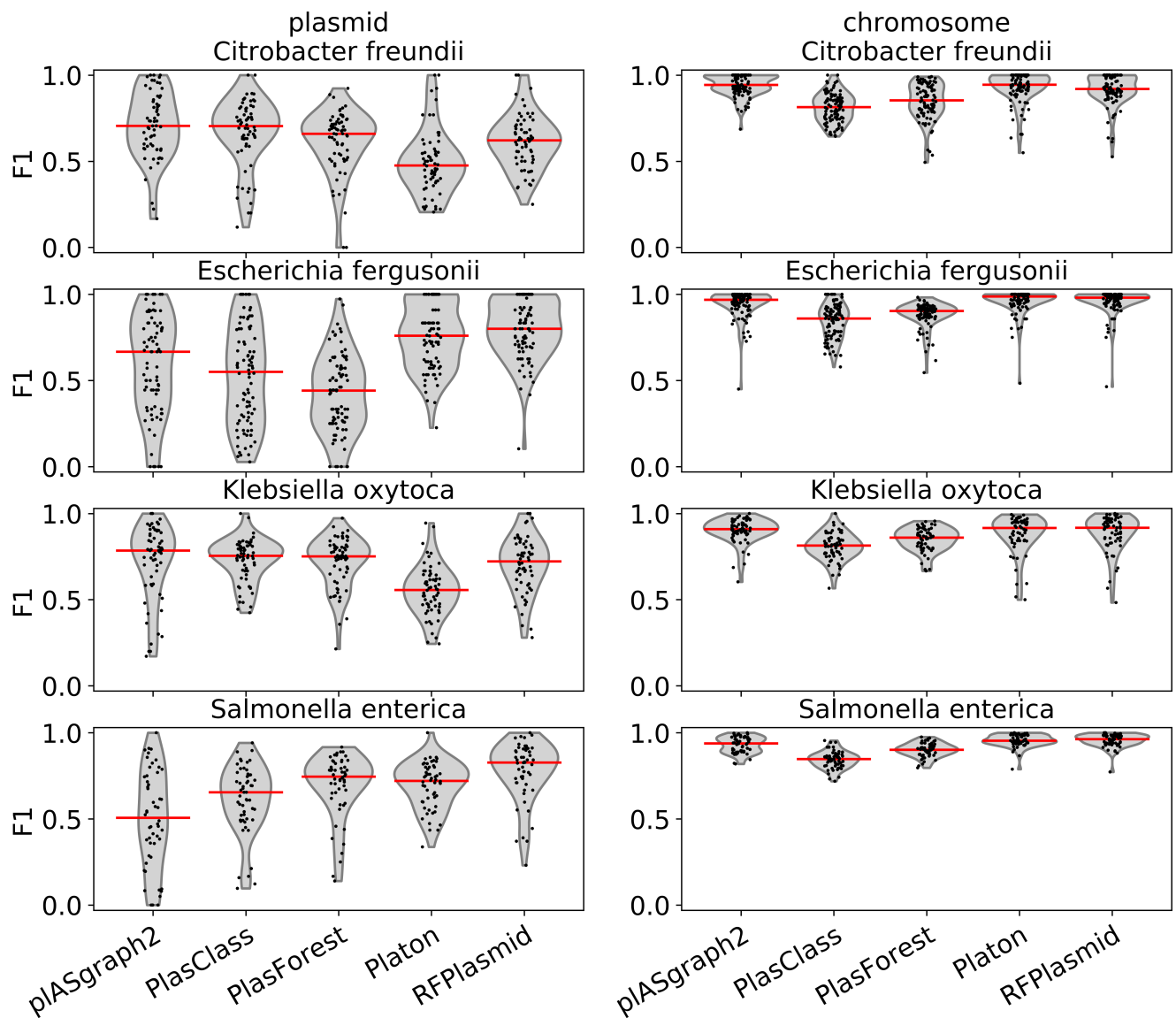


Figure S5. Comparison of F1-scores using samples of close non-ESKAPEE species and considering all contigs longer than 100 bp. Each datapoint represents the F1-score of a single isolate. The horizontal lines represent median.

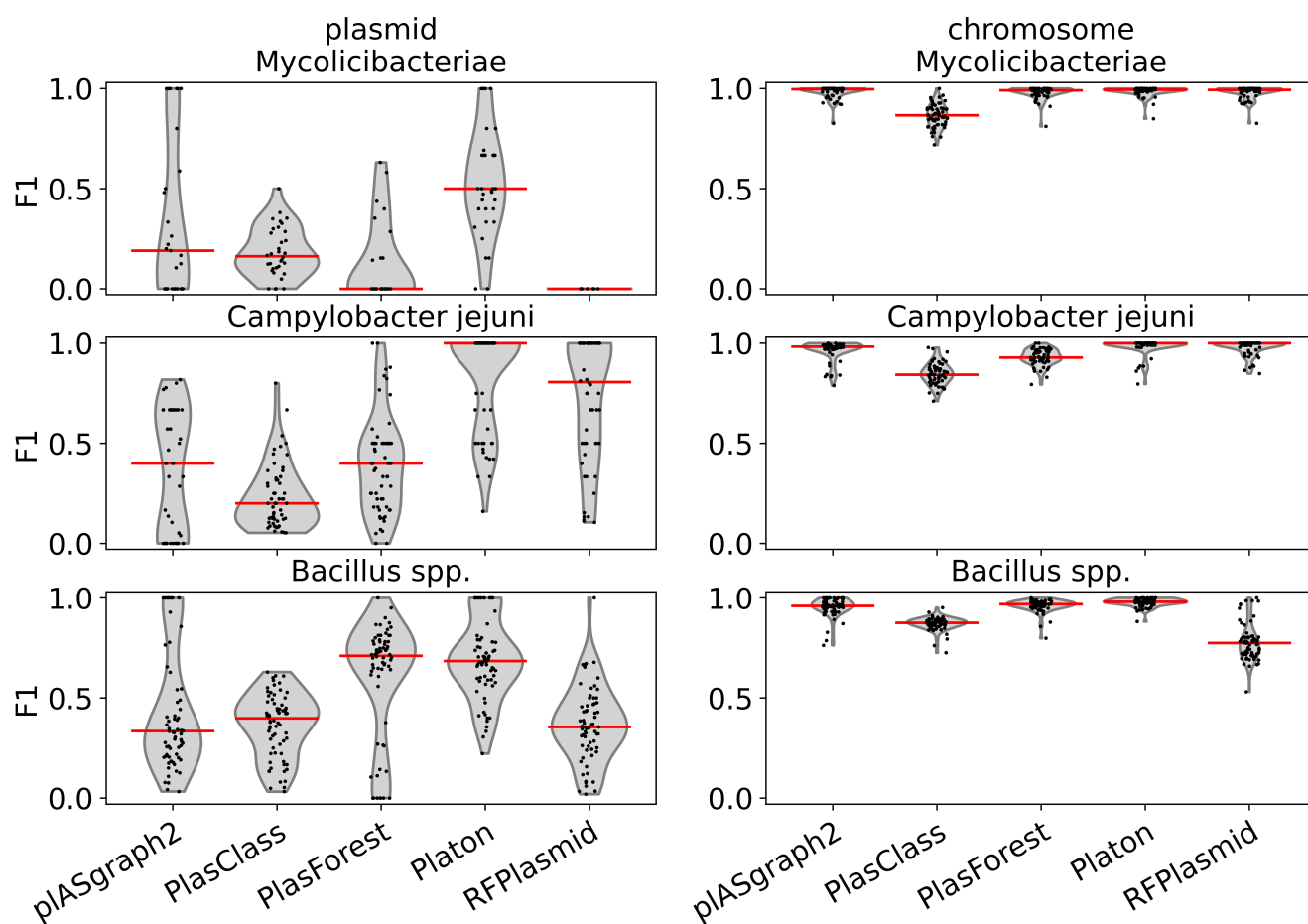


Figure S6. Comparison of F1-scores using samples of more distant non-ESKAPEE species and considering all contigs longer than 100 bp. Each datapoint represents the F1-score of a single isolate. The horizontal lines represent median.

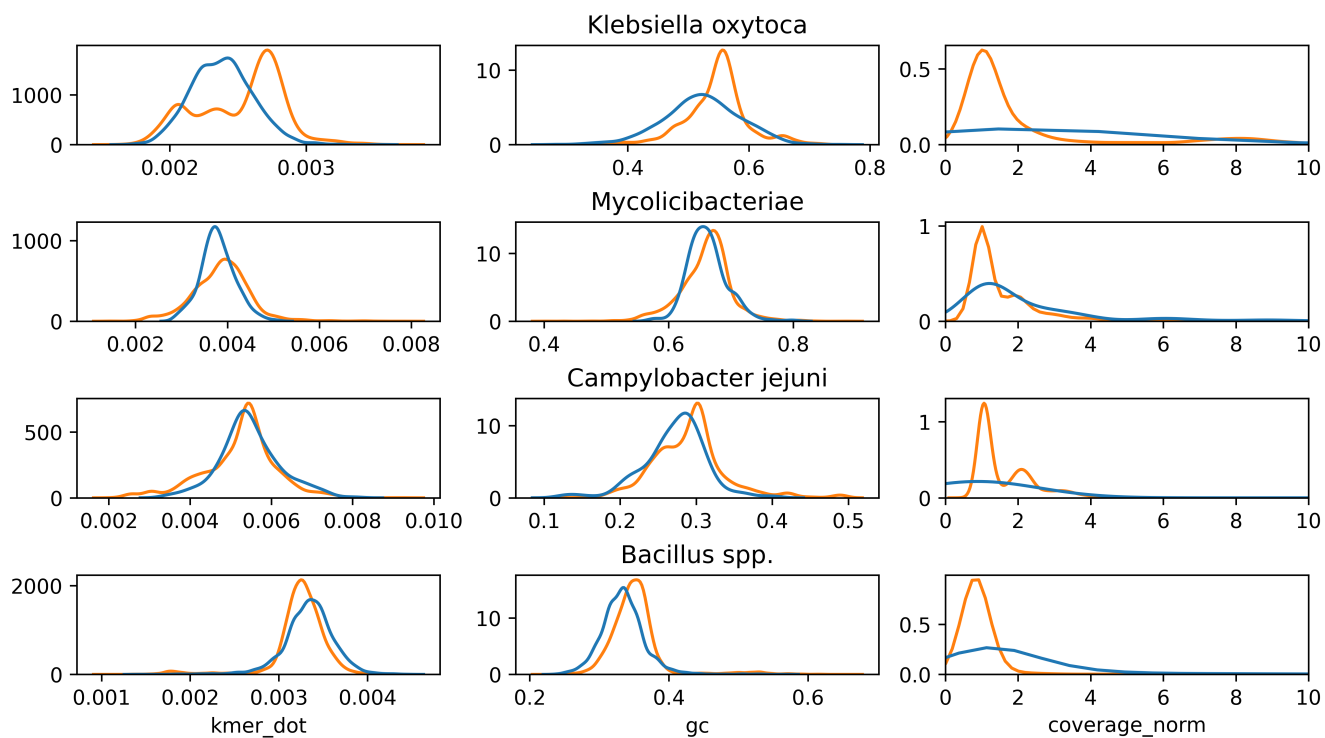


Figure S7. Distribution of sequence-related features of chromosome and plasmid contigs. Orange: chromosome contigs. Blue: plasmid contigs. Ambiguous and unlabeled contigs are not included. Left: relative pentamer content distribution. Middle: GC content distribution. Right: Relative coverage.