Supplementary Data

Supplementary Data S1 (Extending the number of complete genome sequences for *E. faecium*)

Illumina sequencing

Bacterial isolates were grown overnight (O/N) at 37°C on blood agar plates. Single colonies were picked up and grown O/N at 37°C with Brain Heart Infusion (BHI). Bacterial cell pellets were pretreated and incubated 1-4 hours with 180 µL of enzymatic lysis buffer. Subsequently, 0.75 mg proteinase K were added and incubated at 56°C until lysis completion. 20 µL of RNAse A (10mg/mL) were added and incubated for 5' at room-temperature (RT). Total DNA purification was performed using and following the protocol from NucleoSpin 96 Tissue Core Kit (Machery-Nagel), vacuum processing. DNA concentration was measured using Quant-it Picogreen (Thermo Fisher Scientific). Library preparation was carried out following Nextera DNA Library Prep Reference Guide. Finally, Nextera libraries were sequenced using Illumina NextSeq at USEQ, Utrecht, The Netherlands (http://www.useq.nl).

WGS short-read assemblies

Illumina reads were trimmed using nesoni clip, part of the nesoni toolkit (version 0.132), with the following settings: '--adaptor-clip yes --match 10 --max-errors 1 --clip-ambiguous yes --quality 10 --length 90 --trim-start 0 --trim-end 0 --gzip no --out-separate yes pairs:'. Trimmed reads were then assembled into scaffolds using SPAdes (version 3.5.0) with default settings. Scaffolds with an average coverage lower than 10 and/or a length smaller than 500bp were removed from the assemblies.

Table S1. SPAdes assembly statistics using Illumina MiSeq/NextSeq.

Technology	Number of isolates	Mean Coverage	Mean N50	Mean contig length	Median contig length	Average Number of contigs
Illumina MiSeq	63	98 X	54616 bp	21531 bp	6898 bp	169.1
Illumina NextSeq	1581	113 X	52256 bp	17989 bp	5356 bp	176.3

Isolate selection for ONT sequencing

A fraction (n=60) of the total number of isolates (n=1,644) was selected for long-read sequencing with ONT. The plasmid content of the isolates *in silico* were estimated using PlasmidSPAdes (version 3.8.2) *(1)*. Prokka (version 1.12) was used to annotate the putative plasmid contigs using the *Enterococcus* database included in Prokka (2). Orthologous clustered genes were estimated using Roary (version 3.8), splitting paralogues and defining a threshold of 95% amino-acid level similarity to cluster protein sequences (3). This multi-dimensionality matrix was then reduced and visualized to two dimensions using the t-Distributed Stochastic Neighbor Embedding (t-SNE) (theta = 0.5, iterations = 1000, dims = 2) using the implementation provided in the R (version 3.3.3) package Rtsne (version 0.13) (4, 5). k-means (iter.max = 1000) provided in the R package stats was used to allocate 50 centroids into the dimensionality reduced distribution given by tSNE. Euclidean distance of each isolate was calculated to extract the 50 isolates closest to each centroid.

To cover all plasmid replication genes not present in the first selection, 10 additional isolates were selected for ONT sequencing. This second selection was based on a reciprocal blast of the predicted plasmid orthologous genes against 76 previously described plasmid replication amino-acid sequences from the genus *Enterococcus* (6). Reciprocal blast allowed to identify miss-annotated genes corresponding to plasmid replication sequences. Isolates bearing plasmid replication genes not present in the first selection were sorted and selected based on highest number of orthologous genes.

ONT sequencing

E. faecium selected isolates were grown O/N at 37°C on blood agar plates, then single colonies were picked up and grown with BHI at 37°C. Genomic DNA was extracted using the Wizard Genomic DNA purification kit (Promega) following manufacturer's instructions. Isolated DNA was sheared (4000 rpm, 2x120 seconds) using G-tubes (Covaris). Library preparation was performed using Ligation Sequencing Kit 1D (SQK-LSK108) with the Native Barcoding Kit 1D (EXP-NBD103). Genomic libraries were loaded onto R9.4 (FLO-MIN106) flowcells using the MinION device (Mk2). Libraries were basecalled using Metrichor workflows (Run 1 ,2, 3), Albacore 1.01 (Run 4, 5) and Albacore 1.1.0 (Run 6). ONT Sequencing and basecalling were conducted at USEQ, Utrecht, The Netherlands (http://www.useq.nl)

ONT reads and hybrid assembly

Fastq files were obtained from base-called data using Poretools (version 0.6.0) except for Run6 in which fastq files were retrieved using Albacore (version 1.1.0). Distribution of read length and total number of reads were calculated using Bioawk (version 20110810, https://github.com/lh3/bioawk). We used Porechop (version 0.2.1, https://github.com/rrwick/Porechop) to trim reads and filter out chimeras from different bins specifying the flag "--discard_middle". Illumina reads were trimmed

using seqtk (version 1.2-r94, https://github.com/lh3/seqtk) with the command "--trimfq" prior to assembly.

Hybrid assembly was performed using Unicycler (version 0.4.1), specifying "bold" mode (7). Briefly, Unicycler uses SPAdes (version 3.6.2) to create different assembly graphs based on different k-mer size only considering Illumina reads (8). The best assembly graph was selected by Unicycler based on number of dead-ends and contiguity. Next, all ONT reads were used to scaffold and solve the assembly graph. Additionally, we specified the same file as described above (section 2.3) containing 76 known plasmid replication sequences to rotate and change the 0-coordinate of replicons resulting from hybrid assembly (6). Finally, Unicycler conducted several rounds of Pilon (version 1.22) to polish genome sequences using Illumina reads (9).

Categorization of Unicycler contigs

Unicycler contigs were labeled either as chromosome or plasmids based on size and circularity. Contigs were categorized as chromosome if they were larger than 350 kbp, regardless of circularity. However, only contigs were categorized as plasmids if they were circular and smaller than 350 kbp. Putative plasmids smaller than 350 kbp and lacking circularization signatures were not categorized. Draft annotation (Prokka - version 1.12) of plasmid sequences allowed to identify and discard four putative complete phage sequences present as circular contigs.

Supplementary Data S2 (Building a machine-learning model)

Decision trees, random forest and support-vector machines hyperparameters were optimized using random search in a predefined search space (Table 1). We performed 10-fold cross-validation to assess the quality of hyperparameters combination, using error rate as performance measure. For each object, posterior probabilities were generated and the class with a highest posterior probability was assigned.

Table S2. Hyperparameters optimized for decision trees, random forest, and support vector machine.

Classifier	Hyperparameter	Search space (min-max value)
Decision trees	minsplit	10/50
Decision trees	minbucket	5/50
Decision trees	cp	0.001/0.2
Random Forest	ntree	50/1000
Random Forest	mtry	3/10

Random Forest	nodesize	10/50
Support-vector machine	C	(-10)/10
Support-vector machine	sigma	(-10)/10

Supplementary Figures

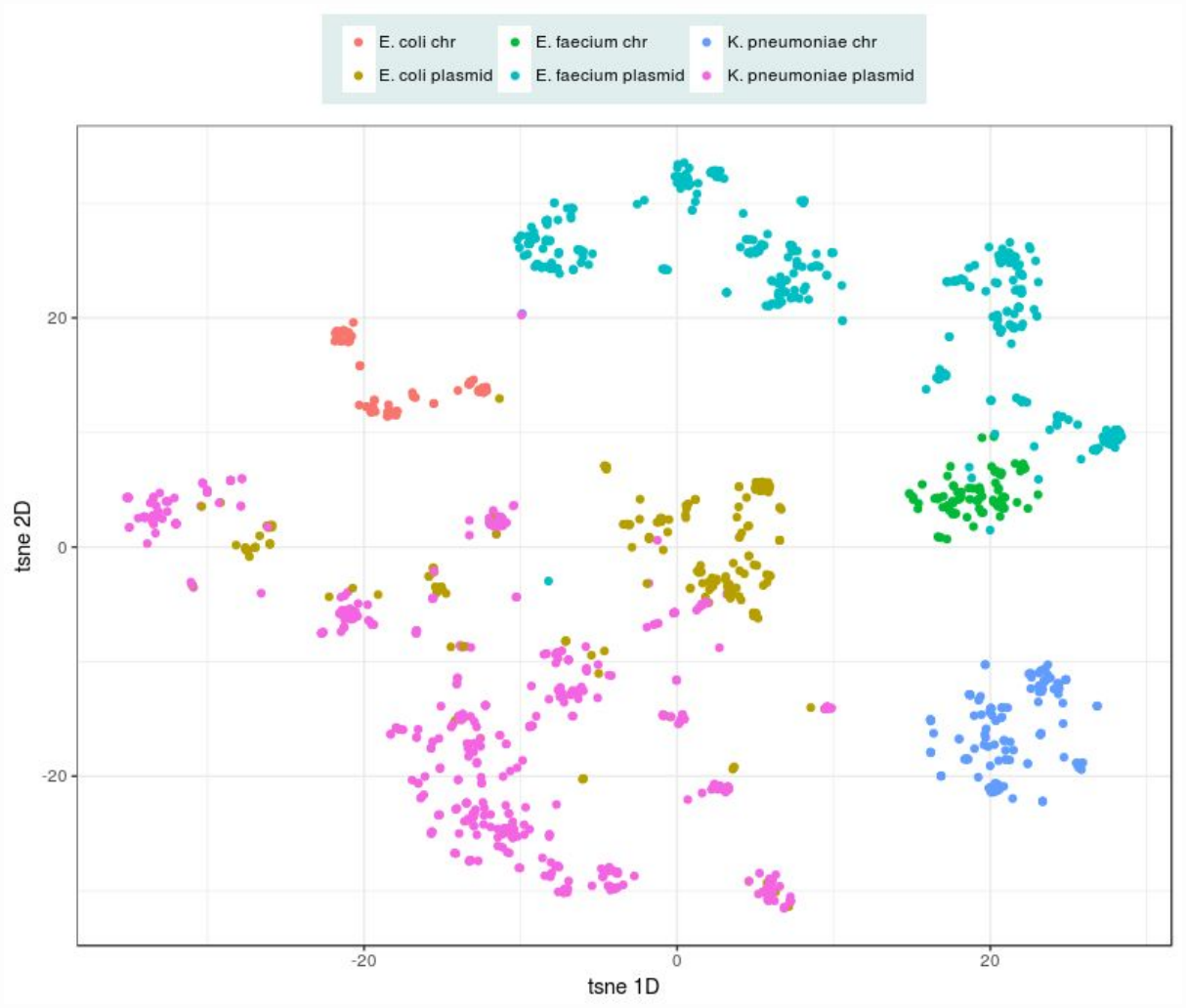

Figure S1. Computed mash distances between chromosome and plasmid sequences were visualized using t-sne. Each point in the graph corresponds to a different type replicon: *E. coli* chromosome (red), *E. coli* plasmid (yellow), *K. pneumoniae* chromosome (dark blue), *K. pneumoniae* plasmid (pink), *E. faecium* chromosome (green) and *E. faecium* plasmid (light blue).

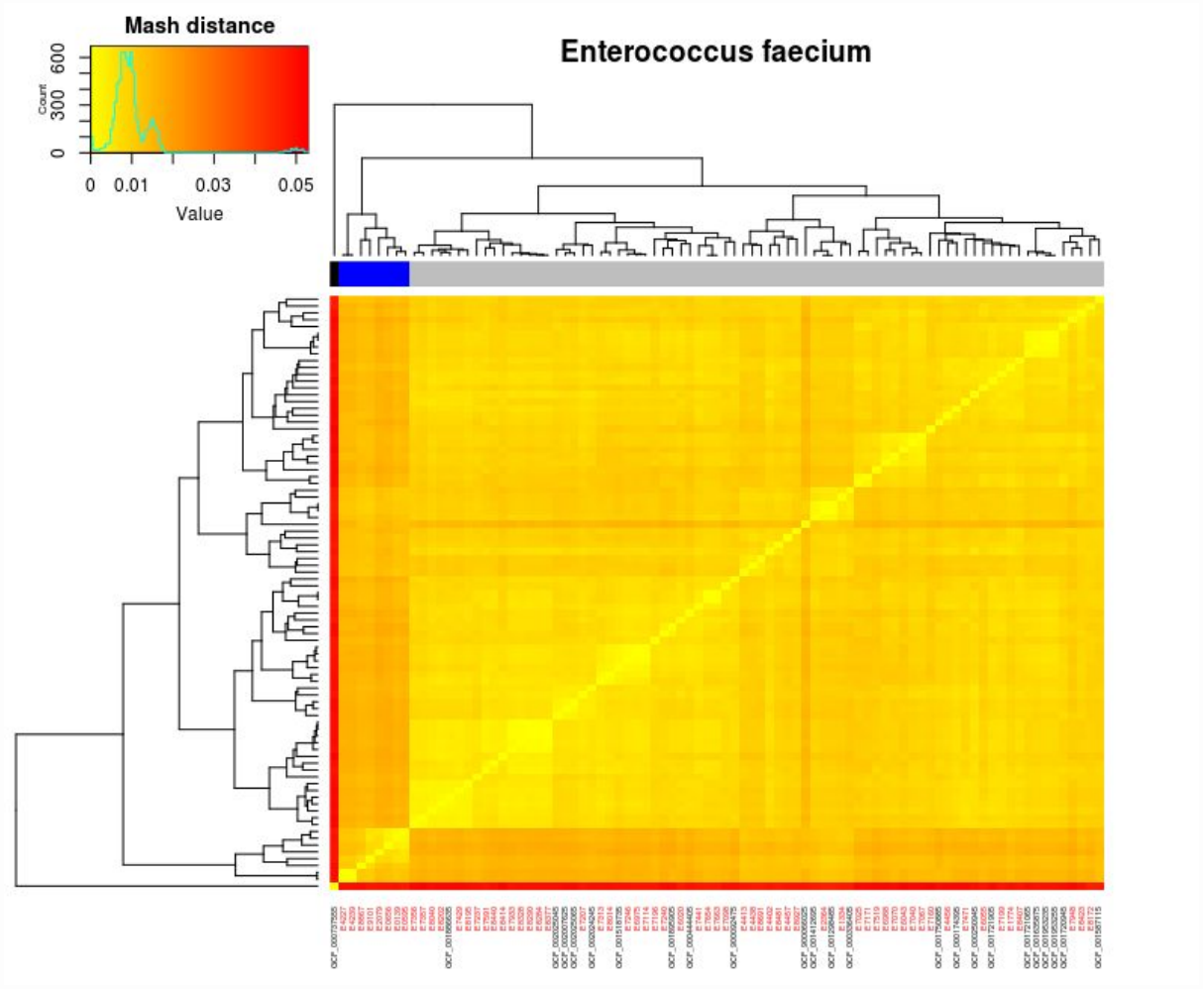

Figure S2. Ward hierarchical clustering of computed pairwise mash distances (k = 21 ; s = 1,000) from *E. faecium* isolates. Based on dendrogram branch lengths, we defined three clusters (black, blue and grey) and visualized mash distances using heatmap based on their genome content similarity. At the bottom y-axis, we coloured in red *E. faecium* isolates (n = 60) that were selected and completed using ONT sequencing and Illumina sequencing. Rest of the isolates corresponded to publicly available NCBI complete genomes from *E. faecium* (n = 24).

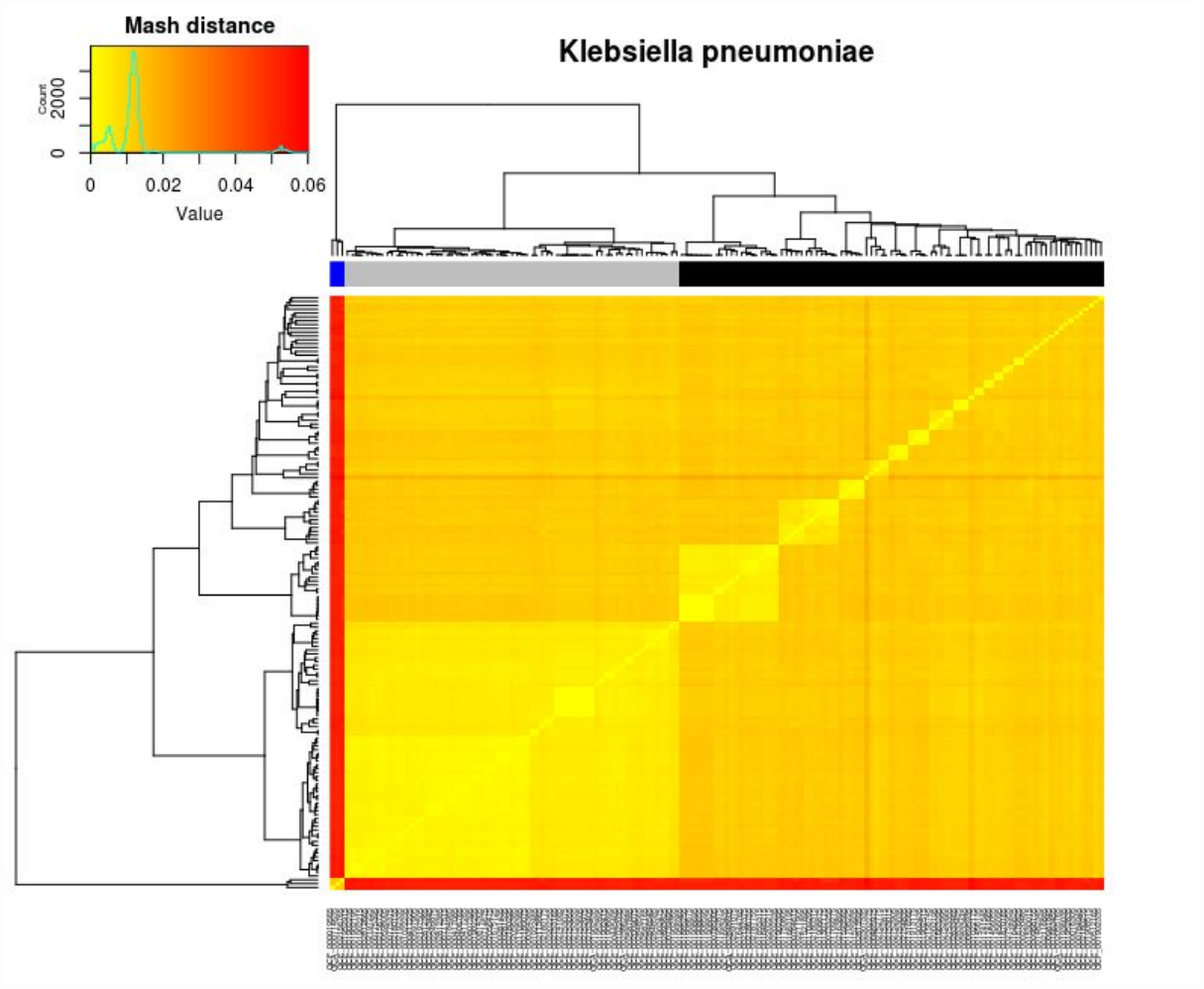

Figure S3. Ward hierarchical clustering of computed pairwise mash distances (k = 21 ; s = 1,000) from *K. pneumoniae* isolates retrieved from Assembly Entrez NCBI database (n = 156). Based on dendrogram branch lengths, we defined three clusters of isolates (blue, grey and black) and visualized mash distances using heatmap to group isolates based on their genome content similarity.

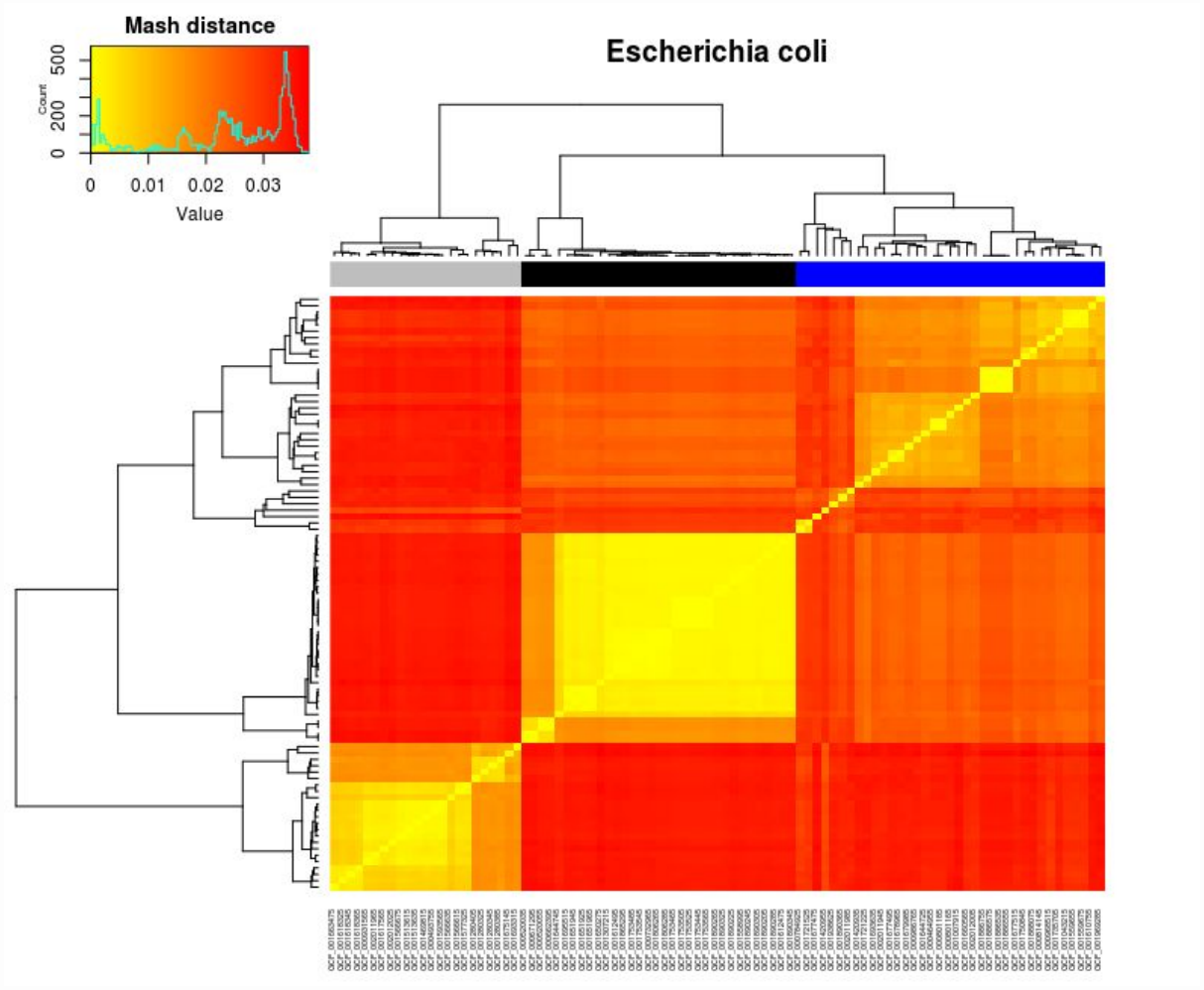

Figure S4. Ward hierarchical clustering of computed pairwise mash distances (k = 21 ; s = 1,000) from *E. coli* isolates retrieved from Assembly Entrez NCBI database (n = 168). Based on dendrogram branch lengths, we defined three clusters of isolates (grey, black and blue) and visualized mash distances using heatmap to group isolates based on their genome content similarity.

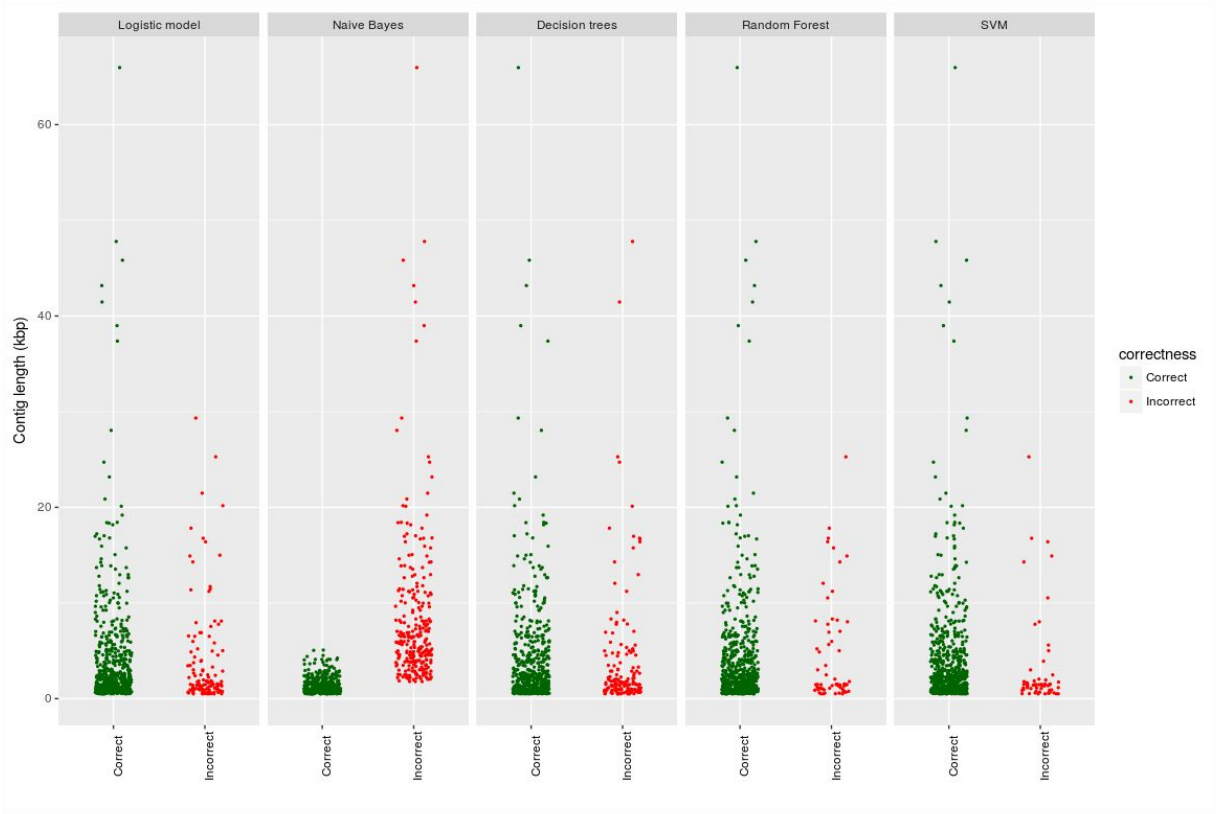

Figure S5. Distribution of correct- and miss- classified short-reads contigs for: Logistic Model, Bayesian Classifier (Naive Bayes), Decision trees, Random Forest, and Support-Vector Machine (SVM). Except for the Bayesian classifier, misclassification most notably occured in contigs with

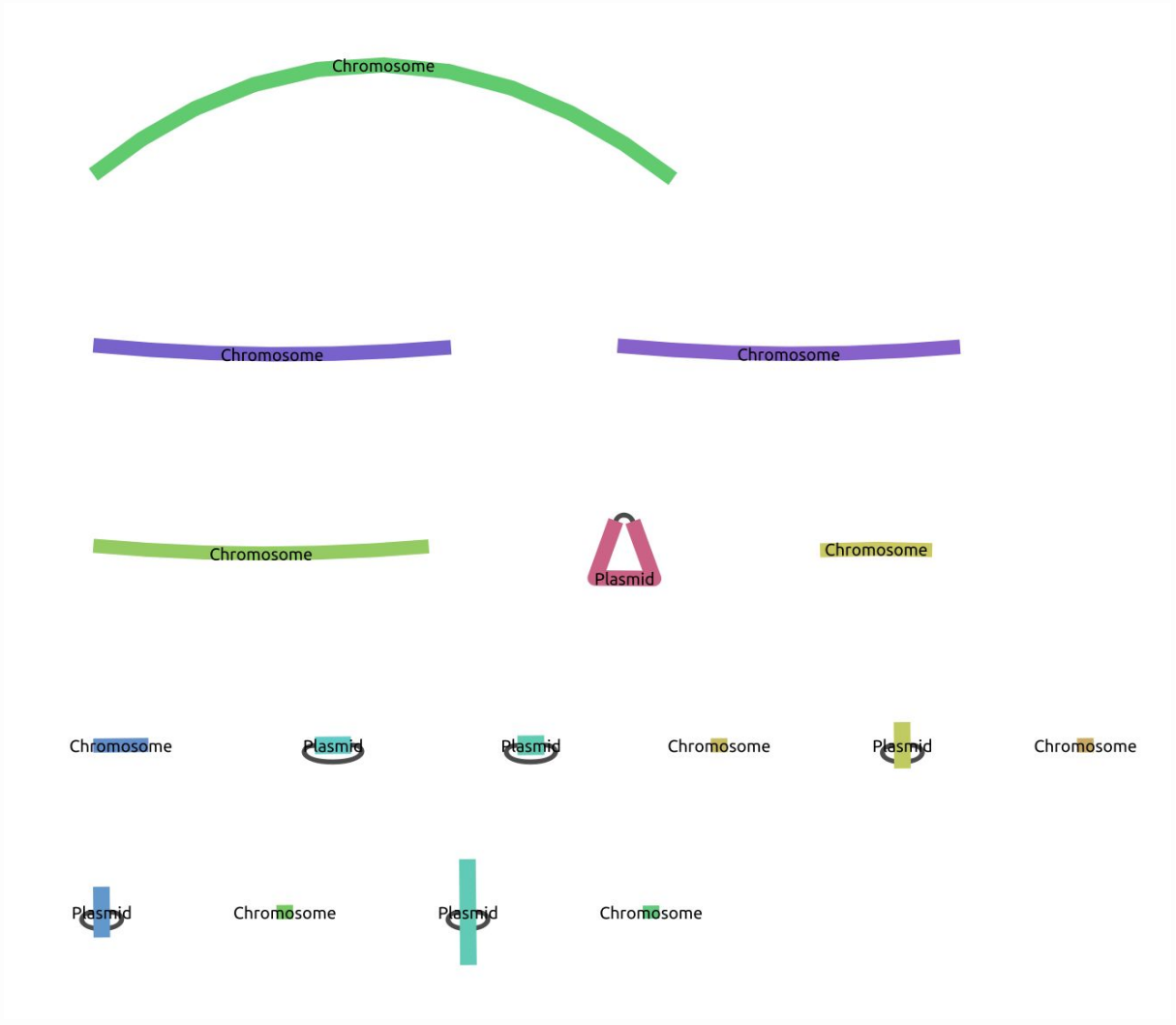

Figure S6. Bandage visualization of the hybrid assembly obtained for the *E. faecium* isolate E7070. For this isolate, hybrid assembly using Unicycler did not result in a complete assembly (chromosome and plasmids in single and circular components). Resulting contigs were labeled based on mlplasmids prediction.

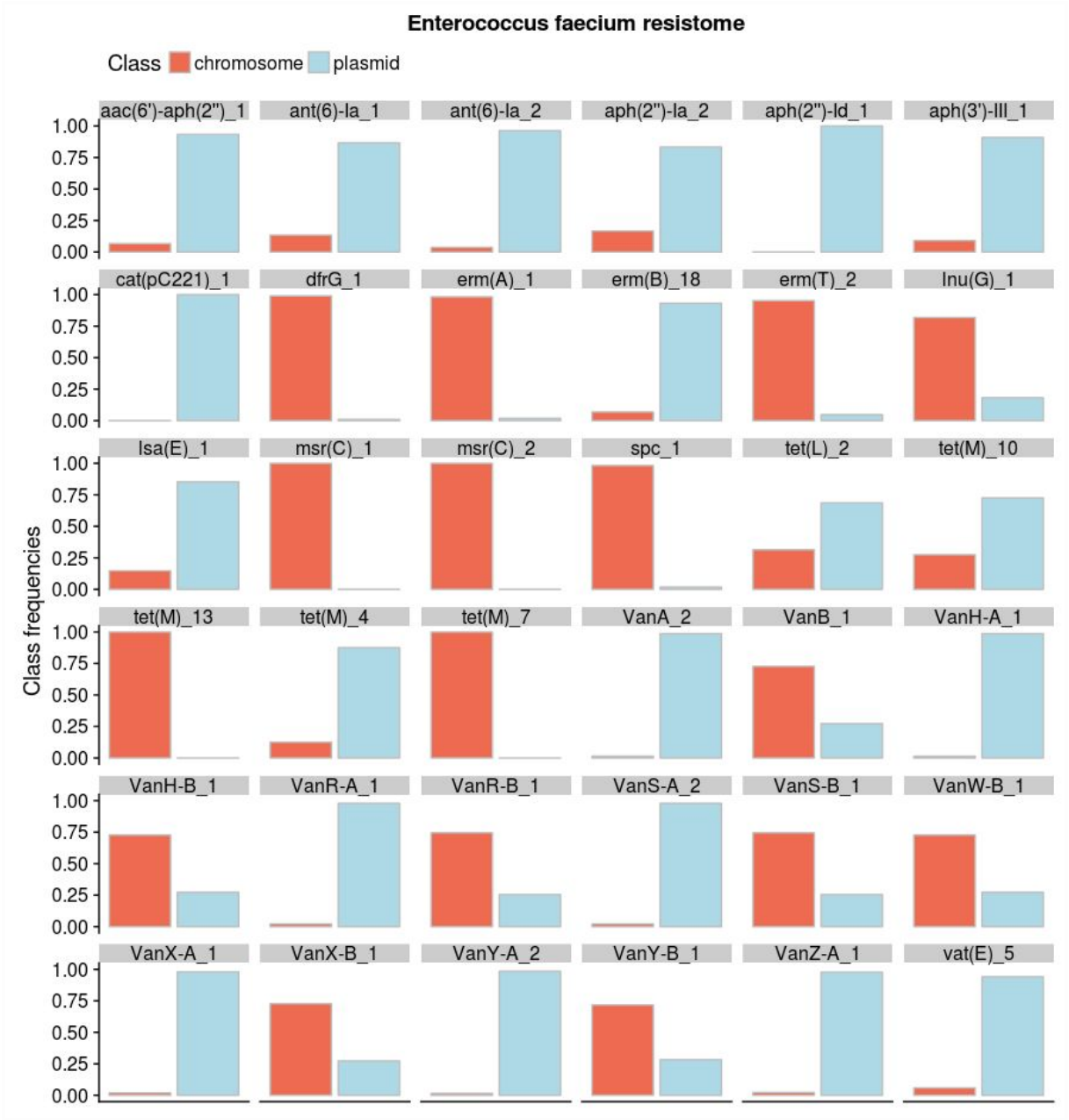

Figure S7. ***Enterococcus faecium*** **resistome.** Draft genomes available in NCBI Genomes FTP (n = 369) were downloaded and screened using Abricate and ResFinder for the presence of antibiotic resistance genes. Each contig containing a resistance gene was predicted with mlplasmids to predict plasmid- or chromosome-origin. For visualization purposes, only antibiotic resistance genes present more than five times are shown.

Figure S8. ***Klebsiella pneumoniae*** **resistome.** Draft genomes available in NCBI Genomes FTP (n = 1,346) were downloaded and screened using Abricate and ResFinder for the presence of antibiotic resistance genes. Each contig containing a resistance gene was predicted with mlplasmids to predict plasmid- or chromosome-origin. For visualization purposes, only antibiotic resistance genes present more than five times are shown.

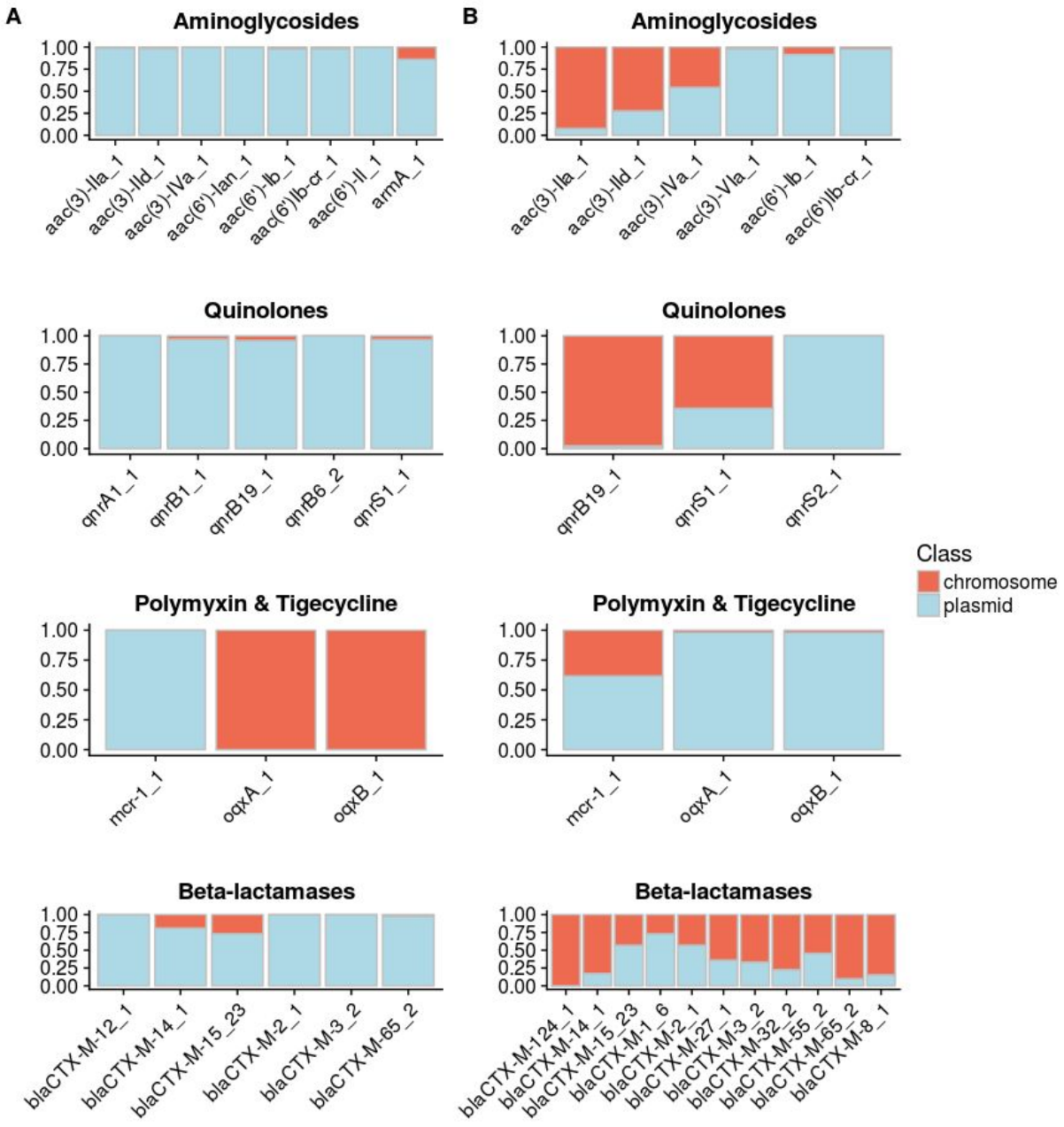

Figure S9. Highlighted genes for *Klebsiella* pneumoniae (panel A) and *Escherichia coli* (panel B).

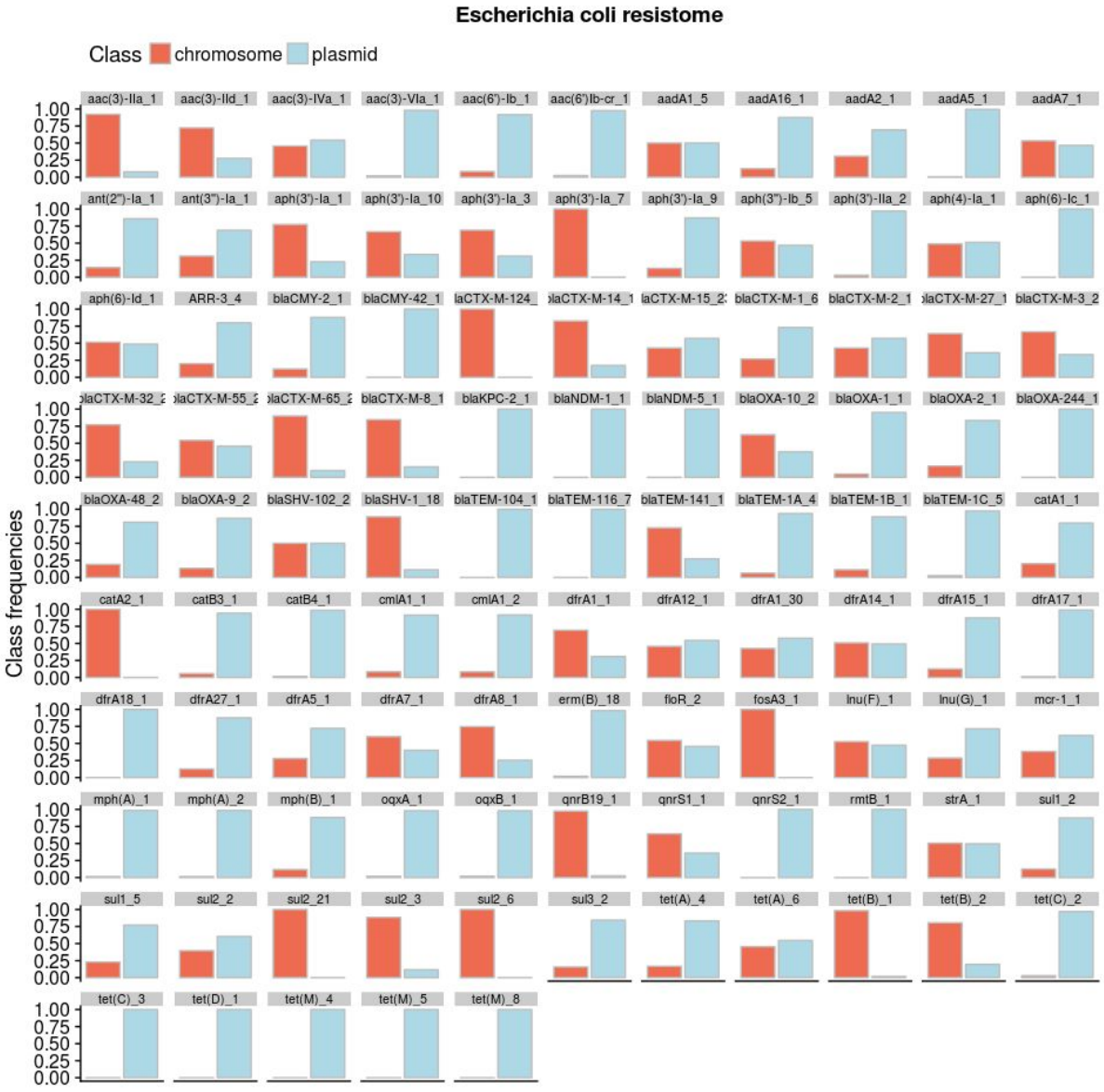

Figure S10. ***Escherichia coli*** **resistome.** Draft genomes available in NCBI Genomes FTP (n = 5,234) were downloaded and screened using Abricate and ResFinder for the presence of antibiotic resistance genes. Each contig containing a resistance gene was predicted with mlplasmids to predict plasmid- or chromosome-origin. For visualization purposes, only antibiotic resistance genes present more than five times are shown.

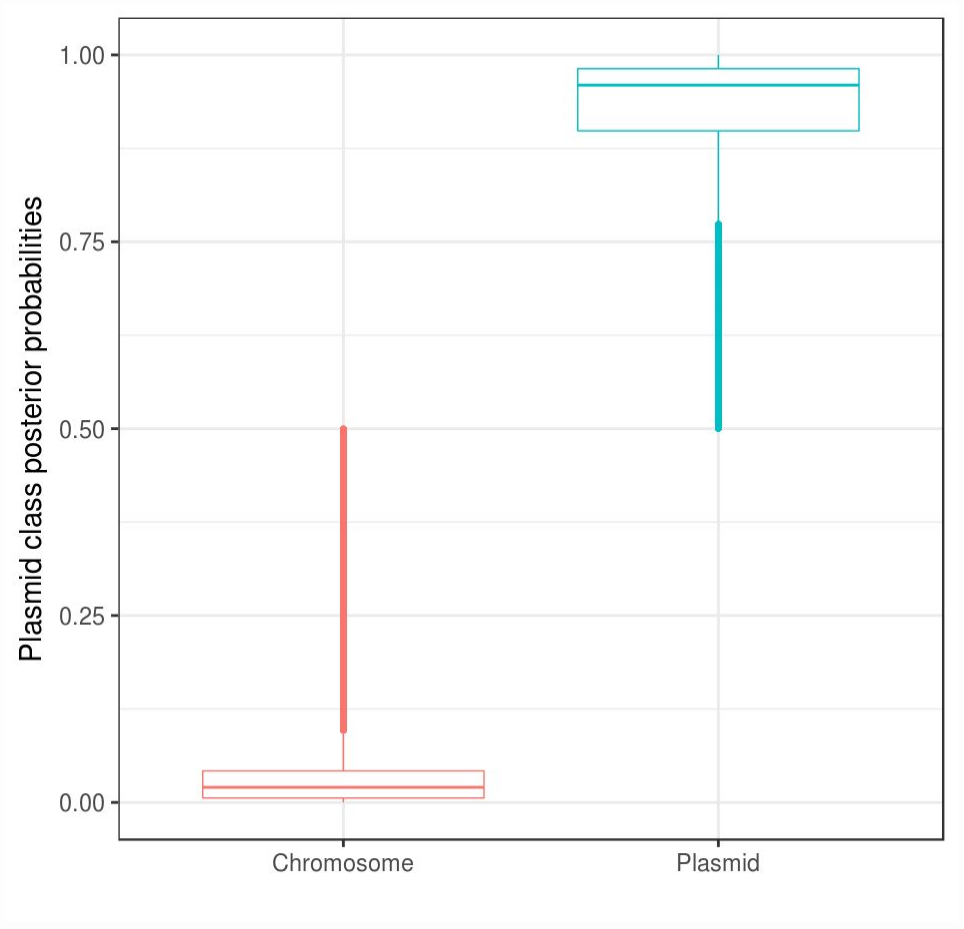

Figure S11. Posterior probabilities of short-read contigs (n= 289,369) of belonging to chromosome- or plasmid- class using our optimized mlplasmids *E. faecium* model for our collection of 1,644 Illumina sequenced *E. faecium* isolates.

References

1. Antipov,D., Hartwick,N., Shen,M., Raiko,M. and Pevzner,P.A. (2016) plasmidSPAdes : Assembling Plasmids from Whole Genome Sequencing Data. *Bioinformatics*, **32**, 3380–3387.

2. Seemann,T. (2014) Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

3. Page,A.J., Cummins,C.A., Hunt,M., Wong,V.K., Reuter,S., Holden,M.T.G., Fookes,M., Falush,D., Keane,J.A. and Parkhill,J. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691–3693.

4. Maaten,L. van der and Hinton,G. (2008) Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

5. Krijthe,J. (2015) Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation (R package version 0.10). *Computer Software*.

6. Clewell,D.B., Weaver,K.E., Dunny,G.M., Coque,T.M., Francia,M.V. and Hayes,F. (2014) Extrachromosomal and Mobile Elements in Enterococci: Transmission, Maintenance, and Epidemiology.

7. Wick,R.R., Judd,L.M., Gorrie,C.L. and Holt,K.E. (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.*, **13**, e1005595.

8. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A. a., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S., Prjibelski,A.D., *et al.* (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.*, **19**, 455–477.

9. Walker,B.J., Abeel,T., Shea,T., Priest,M., Abouelliel,A., Sakthikumar,S., Cuomo,C.A., Zeng,Q., Wortman,J., Young,S.K., *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.