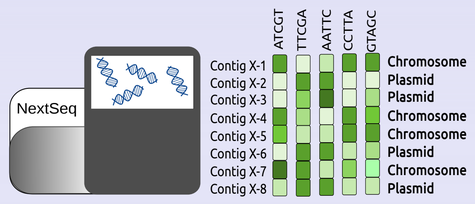
Predicting the plasmidome content of *Enterococcus faecium*

Sergio Arredondo-Alonso[1], Malbert C. Rogers[1], Iris Braat[1], Janetta Top[1], Jukka Corander[2,3,4]
Rob J. Willems[1], Anita C. Schurch[1]

1. Department of Medical Microbiology, University Medical Center Utrecht, Utrecht, The Netherlands.
2. Faculty of Medicine, Department of Biostatistics, University of Oslo, Norway.
3. Department of Mathematics and Statistics, University of Helsinki, Finland.
4. Infection Genomics, Wellcome Trust Sanger Institute, UK.



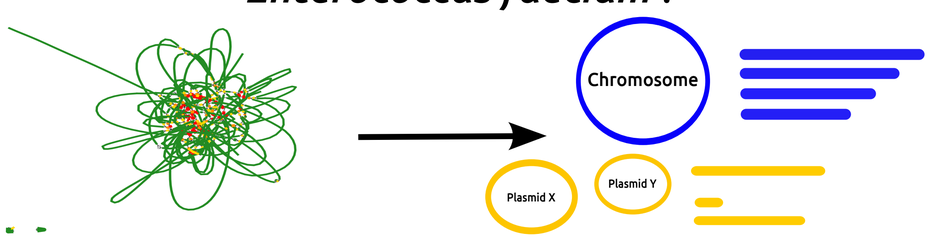
mlplasmids

UMC Utrecht

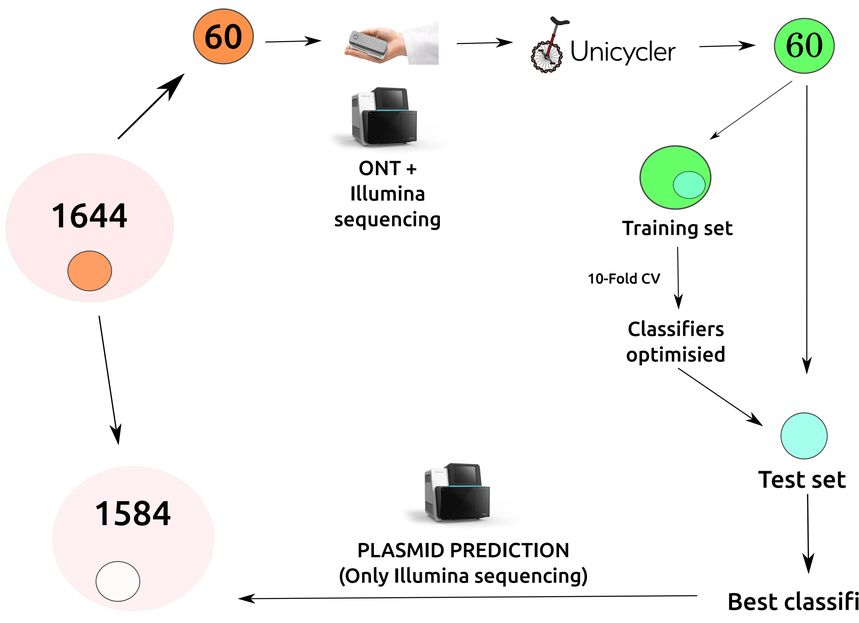
Introduction

Plasmid assembly from **short** whole-genome sequencing data **(WGS) results in an accurate but fragmented graph** consisting of hundreds of contigs. Determining whether a **contig is plasmid- or chromosome- derived is challenging** and error-prone with existing tools[1]. **Long-read sequencing** has emerged as a new solution to obtain complete plasmid sequences[2]. Information derived from **long-read sequencing can be used to label a dataset of short-read contigs** as chromosome- or plasmid- derived.

Can we accurately predict the plasmidome content of ***Enterococcus faecium*** ?



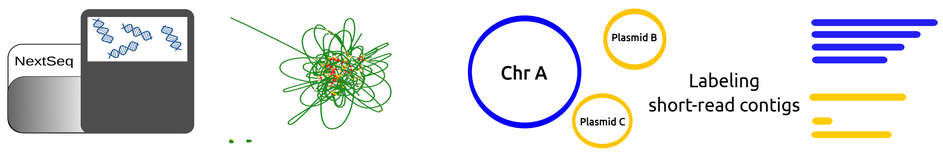
Predicting the plasmidome content of our collection



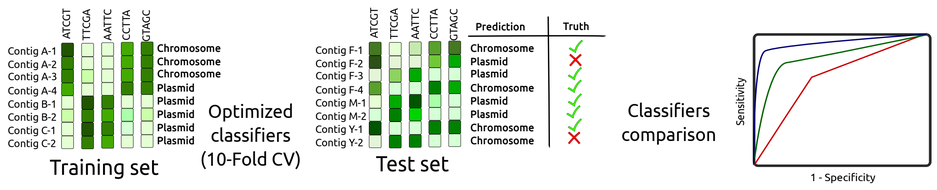
Developing a novel machine learning classifier

Hybrid assembly to obtain complete genome sequences (**Unicycler**[3])

Mapping short-read contigs against complete genomes (**bwa**[4] **+ samtools**[5])



Training and comparing different machine learning classifiers using **pentamer frequencies (R mlr package**[6]**)**



Benchmarking resulting classifier

Results

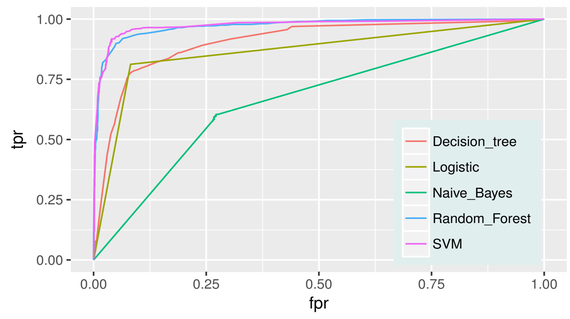


Figure 1. ROC Curves obtained for the tested machine learning algorithms

Support vector machine (SVM) was selected as **best classifier** (AUC = 0.97 ; F1-score = 0.95) using our test set (n = 2,010 contigs). Resulting model was **implemented in mlplasmids (Figure 1)**

Plasmidome prediction resulted in an **average of 58 plasmid contigs (~ 254,700 bp) and 119 chromosomal contigs (~ 2,632,470 bp)** with an associated average **posterior probability of 0.95 and 0.91** of belonging to the plasmid and chromosome class **(FIgure 2)**

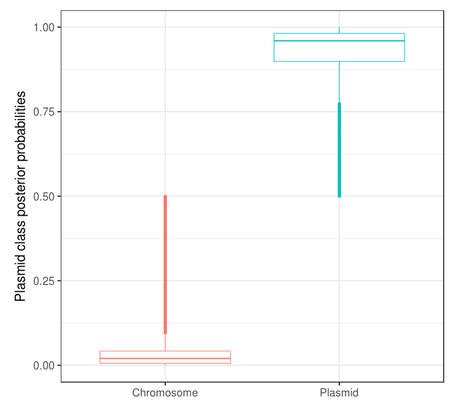


Figure 2. Boxplot showing the posterior probabilities distribution of belonging to the plasmid class

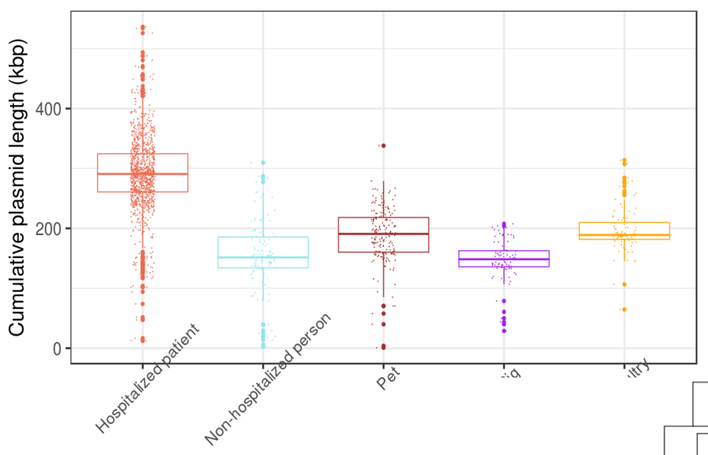


Figure 3. Box plot showing the distribution of the predicted cumulative plasmid length (kbp) per isolation source

Predicted cumulative plasmid length is higher in hospitalized patients (average 290.51 kbp vs 174.46 kbp) **(Figure 3)**

Plasmid DNA coding capacity differs between **hospitalized patients** and other isolation sources **(Figure 4)**

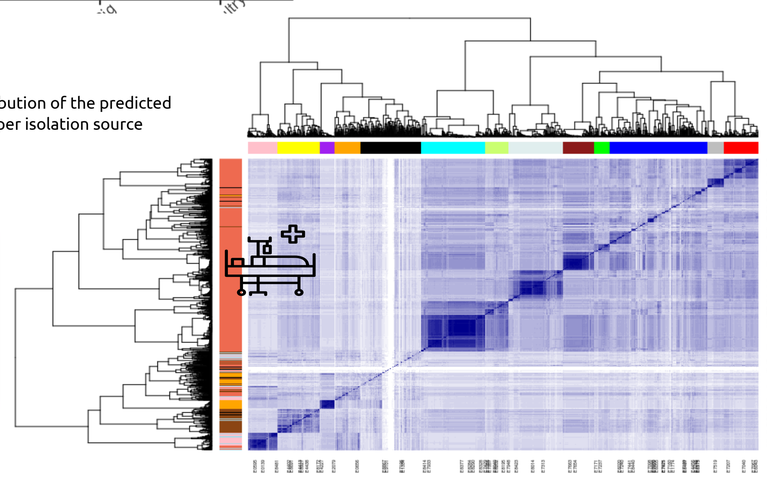


Figure 4. Heatmap showing the distance similarity between strains based on their predicted plasmid gene content (mlplasmids + Prokka[6]+ Roary[7]). Isolation sources are indicated on the left-vertical bar using the same colours defined in Figure 3. We defined 13 plasmid clusters (top-horizontal bar) to represent all the plasmid subpopulations present in our collection.

Conclusions

Genomic structure information resolved by ONT sequencing can be used to build a model capable to accurately classify plasmid sequences in *E. faecium*

E. faecium hospital isolates have a different pool of plasmid genes than isolates from other sources

mlplasmids is publicly available and allows the assignment of a particular contig/gene of interest as plasmid- or chromosome- encoded

References

1. Orlek, Alex, Nicole Stoesser, Muna F. Anjum, Michel Doumith, Matthew J. Ellington, Tim Peto, Derrick Crook, et al. 2017. “Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology.” Frontiers in Microbiology 8 (February). Frontiers: 182.
2. George, Sophie, Louise Pankhurst, Alasdair Hubbard, Antonia Votintseva, Nicole Stoesser, Anna E. Sheppard, Amy Mathers, et al. 2017. “Resolving Plasmid Structures in Enterobacteriaceae Using the MinION Nanopore Sequencer: Assessment of MinION and MinION/Illumina Hybrid Data Assembly Approaches.” Microbial Genomics 3 (8). Microbiology Society. https://doi.org/10.1099/mgen.0.000118.
3. Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. 2017. “Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads.” PLoS Computational Biology 13 (6). Public Library of Science: e1005595.
4. Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM,” March. http://arxiv.org/abs/1303.3997.
5. Li, H., B. Handsaker, A. Wysoker, T. Fennell, and J. Ruan. n.d. “692 (2009). The Sequence Alignment/Map Format and SAMtools.” Bioinformatics .
6. Bischl, B., M. Lang, L. Kotthoff, and J. Schiffner. 2016. “Mlr: Machine Learning in R.” Of Machine Learning …. jmlr.org. http://www.jmlr.org/papers/volume17/15-066/15-066.pdf.
6. Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” Bioinformatics 30 (14): 2068–69.
7. Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. “Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis.” Bioinformatics 31 (22): 3691–93.

GitLab

https://gitlab.com/sirarredondo/mlplasmids

Shiny by RStudio

Future releases

Escherichia coli and *Klebsiella pneumoniae* models