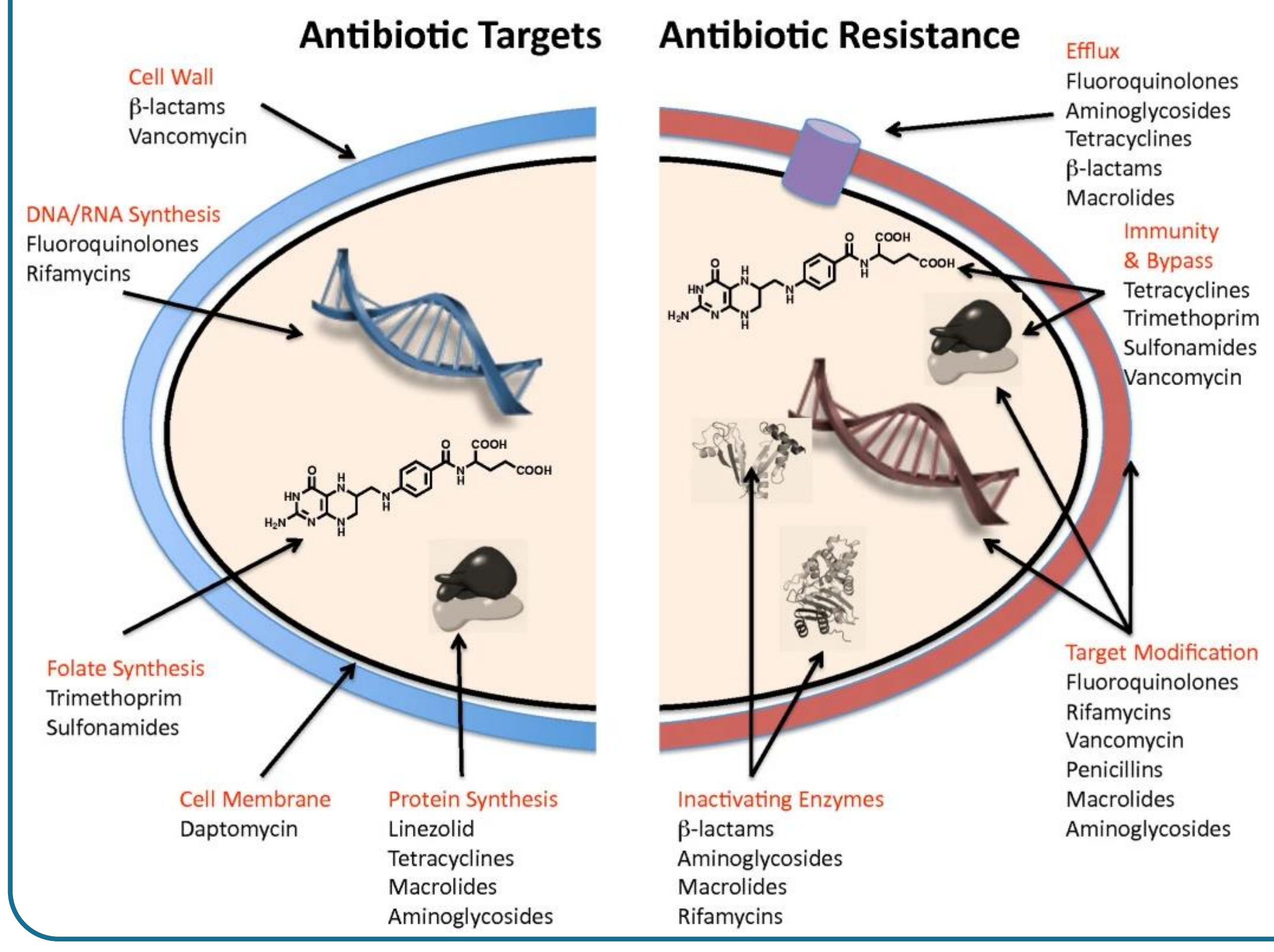
ARSENAL: Antimicrobial ReSistance prEdictioN by a mAchine Learning method

Simankov Nikolay[1,2,4,5,*], Ulysse Guyet[1,2], Léa Bientz[3], Véronique Dubois[3], Alexis Groppi[1,2], and Macha Nikolski[1,2]

[1]Univ. Bordeaux, CNRS, IBGC, UMR 5095, Bordeaux, 33077, France, [2]Univ. Bordeaux, Centre de Bioinformatique de Bordeaux (CBiB), Bordeaux, 33076, France, [3]MFP, CNRS 5234, Université de Bordeaux, Bordeaux,
[4]Laboratory of Plant Pathology – TERRA - Gembloux Agro-BioTech – University of Liège (ULiège) - 5030 Gembloux, Belgium, [5]Statistics, Computer Science and Modeling applied to bioengineering (SIMa) – TERRA – University of Liège (ULiège) 5030 Gembloux, Belgium.

*This Communication is supported by the Walloon Region as part of a FRIA grant.

Background



- ➢ Antimicrobial resistance (AMR) is a growing health threat responsible for an estimated 700 000 deaths per year and this number is projected to reach 10 million by 2050.
- ➢ Appropriate antibiotic therapy improves patient healing outcomes and is a key factor in preventing the emergence of antibiotic resistance.
- ➢ Antibiotic susceptibility testing (AST) from bacterial culture is the current clinical practice for assessing drug resistance by the determination of the minimum inhibitory concentration (MIC) corresponding to the lowest concentration of a specific antibiotic that inhibits bacterial growth. This method, fastidious and long (between 24h and 72h), is only applicable to cultivable bacteria, which excludes analysis of the emergence and spread of antimicrobial resistance in diverse and complex microbial communities with large fractions of currently uncultured bacteria.

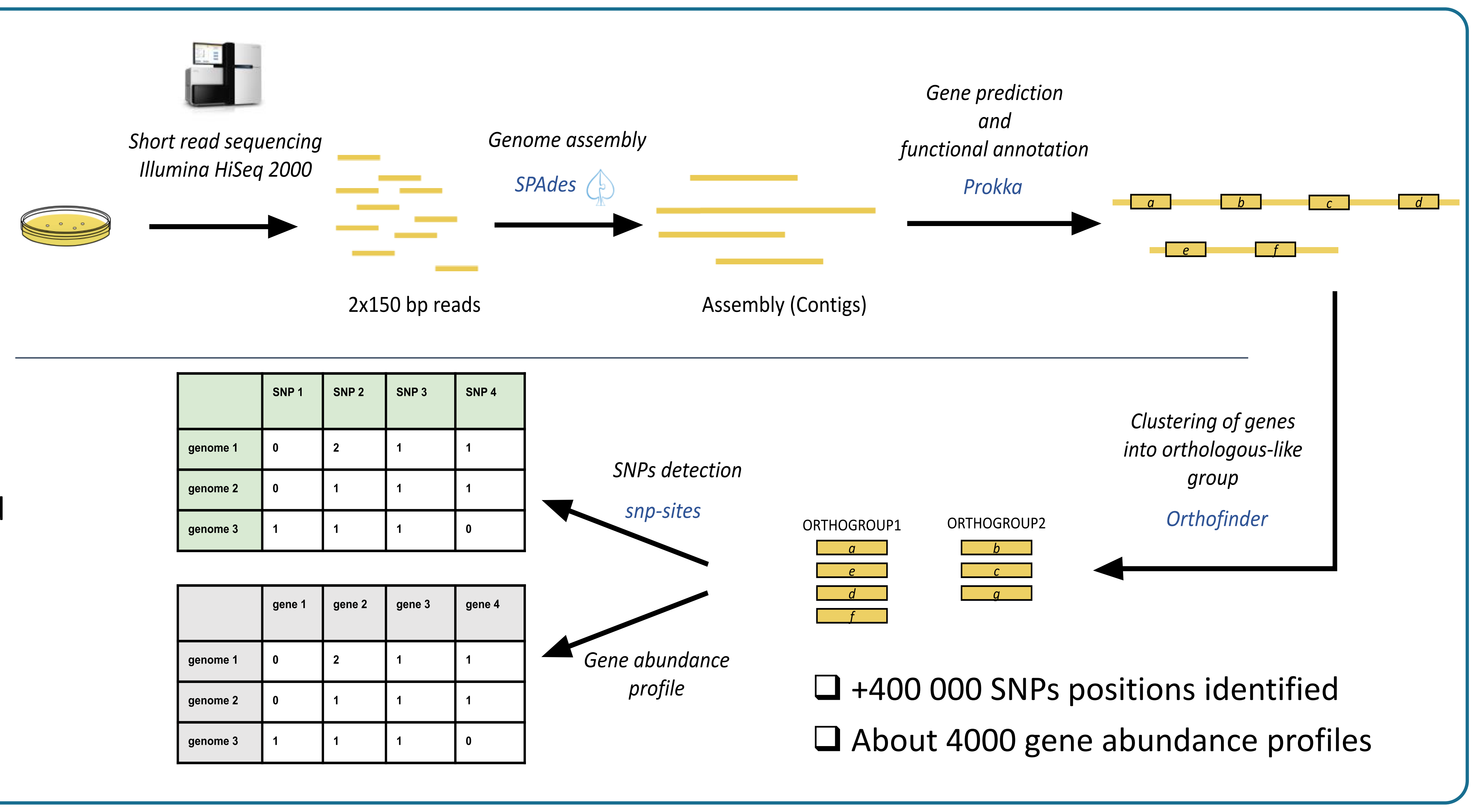
- ➢ Our method allows:
  - ✓ Fast screening of antibiotics to determine the most effective antibiotic and the right dose against a specific bacterial infection
  - ✓ Highlight new antibiotic-resistance genes and biomarkers

Resistance mechanisms are mainly acquired by genome modifications, e.g., mutations and horizontal gene transfers.

Data Generation

- ➢ 1312 ***Streptococcus pneumoniae*** newly sequenced genomes of strains isolated from patients
  - ❑ Different sequence types (ST) and 64 distinct serotypes
  - ❑ Many strains are multidrug resistant
  - ❑ Diversity of geographical origin (isolated from 28 hospitals in 18 provinces of China between 2007 and 2020)

- ➢ MIC values (Minimum Inhibitory Concentration) of each strain measured by antibiotic susceptibility testing mainly for 6 Beta-lactam antibiotics :
  - ❑ Penicillin
  - ❑ Amoxicillin
  - ❑ Cefuroxime
  - ❑ Ceftriaxone
  - ❑ Cefepime
  - ❑ Imipenem

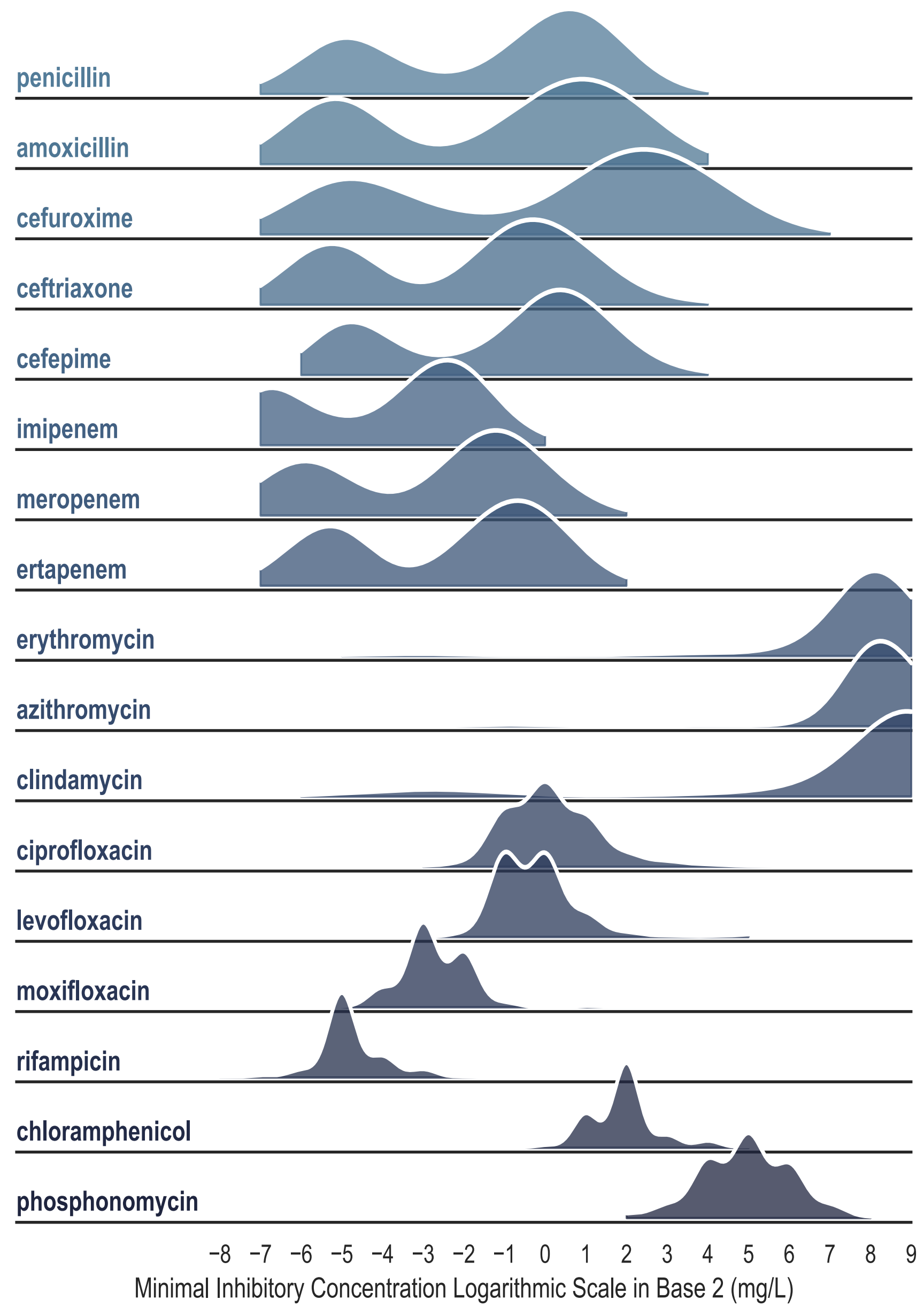


	SNP 1	SNP 2	SNP 3	SNP 4
genome 1	0	2	1	1
genome 2	0	1	1	1
genome 3	1	1	1	0

	gene 1	gene 2	gene 3	gene 4
genome 1	0	2	1	1
genome 2	0	1	1	1
genome 3	1	1	1	0

- ❑ +400 000 SNPs positions identified
- ❑ About 4000 gene abundance profiles

Feature Selection



The Challenge :

- ❑ The identification of resistance-related genes supports the relevance of the model.
- ❑ GWAS-based approach resulted in a highly correlated matrix with insufficient resolution for MIC prediction.
- ❑ Collinear features are a thread for Machine Learning efficiency.
- ❑ The initial number of SNPs does not allow us to compute the correlation matrix.
- ❑ Different population types required different approaches.

The Results :

- ✓ We reduced the input data from 1,000,000 to 1,000 features
- ✓ Clusters contain all highly correlated features that allow us to study them afterward

Low Variance Filtering

- SNPs with a variance lower than 0.05% are removed
- ≈ 0.95 > prevalence > 0.05

Per Gene SNP Clustering

- SNPs from the same orthogroup are clustered together according to their Spearman correlation (>95%)
- All features inside each cluster are represented by one single SNP

Inter Gene Clustering

- Representative SNPs that share the same Spearman correlation to a random vector are clustered together (>95%)
- All features inside each cluster are represented by one single SNP

Statistical Univariate Cluster Ranking

- Union of the *n* best clusters based on the KendallTau and Spearman correlations, Mutual information test, and ANOVA F-test
- *n* will depend on the number of available samples

Optimized Multivariate Feature Selection

- Union of the best feature combinations based on Random Forest, Linear Regressions and Relevance Vector Machine
- Feature numbers are chosen to optimize predictive model performance

Conclusion and Perspectives

- ➢ ARSENAL is a new method that can predict a precise level of resistance in bacterial strains based on genomic data in a significantly shorter period of time but also pave the way for the discovery of new potential links between phenotype and genotype by identifying new genes and SNPs involved in resistance mechanisms

- ➢ We aim to identify some genes whose functions are related to antimicrobial resistance, but the function of many of the output genes of the model remains to be characterized (e.g., by site-directed mutagenesis) and may allow us to identify new resistance biomarkers.

- ➢ The method has been applied to strains of *Streptococcus pneumoniae* (gram-positive) and will be soon applied on several gram-negative Bacteria :
  - ❑ *Pseudomonas aeruginosa*
  - ❑ *Helicobacter pylori*
  - ❑ *Salmonella enterica*
  - ❑ *Klebsiella pneumoniae*
  - ❑ *Campylobacter jejuni*
  - ❑ *Escherichia coli*

jpiamr cbib BIOINFORMATIQUE wgc LIÈGE université université de BORDEAUX