Quantification of the association of bacterial clones with antibiotic resistance

Supplementary Document 1 for 'Rapid heuristic inference of antibiotic resistance and susceptibility by genomic neighbor typing'

Karel Břinda, Alanna Callendrello, Kevin C Ma, Derek R MacFadden, Themoula Charalampous, Robyn S Lee, Lauren Cowley, Crista B Wadsworth, Yonatan H Grad, Gregory Kucherov, Justin O'Grady, Michael Baym, and William P Hanage

August 22, 2019

1 Introduction

In this document, we show how to quantify the association of bacterial clones with antibiotic resistance. For a given bacterial species and all antibiotics of interest, we construct optimal predictors of resistance from lineages and calculate the associated Receiver Operating Characteristics (ROC) curve and its Areas under the Curve (AUC). Comparing the resulting curves and areas among antibiotics helps to understand the different levels of associations.

We use this framework to show that for the pathogens *Streptococcus pneumoniae* and *Neisseria gonorrhoeae* antibiotic resistance is highly associated with the population structure. This provides evidence of the suitability of genomic neighbor typing as a diagnostic method for these pathogens.

2 Optimal lineage-to-resistance classifiers

2.1 Model

Let us have a bacterial species and assume that it has g lineages. For the purpose of this document, lineages are arbitrary classes of equivalance; they can be defined based on sequence typing (e.g., MLST (Maiden 2006)) or clustering (e.g., using BAPS (Cheng et al. 2013) or PopPUNK (Lees et al. 2019)). Assume that lineages are equally probable; i.e., a randomly drawn isolate x comes from every lineage i with the probability $\frac{1}{g}$. Assume that for every isolate x, we can always correctly determine its lineage $\ell(x)$. In a clinical setting, this could mean that we can always determine isolate's MLST sequence type.

Let us consider an antibiotic and assume that every isolate is either resistant or susceptible to this antibiotic. Assume that resistance within a lineage i is iid with the Bernoulli distribution and let r_i denote the probability of resistance. The constants r_i can be determined based on epidemiological data or as proportions of resistance isolates in individual lineages from population-level studies.

2.2 Lineage-to-resistance classifiers

For a given species, an antibiotic and fixed probabilities of resistance within lineages $r_1, \ldots, r_g$, we construct memoryless probabilistic resistance classifiers C of the form

$$C(\ell(x)) \to \{\text{'S'}, \text{'R'}\}.$$

In other words, for every isolate we identify its lineage and the classifier predicts resistance based on the knowledge of the lineage.

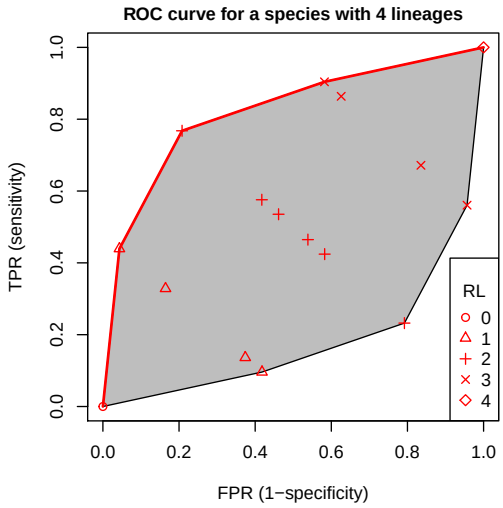

Figure 1: An illustration of the ROC diagram for a species with 4 lineages with resistance probabilities $(r_i)_{i=1}^4 = (0.1, 0.4, 0.8, 0.9)$. The grey area corresponds to all the possible lineage-to-resistance classifiers $C_{(R_1,R_2,R_3,R_4)}$ while the red points are the "vertex" classifiers with integer parameters; i.e., $R_i \in \{0,1\}$ for every $i \in \{1,\ldots,g\}$. RL denotes the number of lineages for which resistance is reported by an integer classifier, i.e., $\sum_i R_i$. The optimal classifiers that maximize the AUC lie on the piece-wise linear red curve, which we term the optimal ROC curve.

All such classifiers can be parametrized as $C_{(R_1,\ldots,R_g)}$, where $R_i \in [0,1]$ is the probability of reporting resistance given the sample has been indentified as to belong to the lineage i. For instance, the classifier $C_{(0,\ldots,0)}$ always assigns 'S', $C_{(\frac{1}{2},\ldots,\frac{1}{2})}$ assigns 'R' and 'S' like a fair coin, $C_{(1,\ldots,1)}$ always assigns 'R', and $C_{(1,0,\ldots,0)}$ always assigns 'R' for the first lineage and 'S' for the other ones.

With the knowledge of the probabilities of resistance $(r_1,\ldots,r_g)$ within individual lineages $1,\ldots,g$, we can express false positive rate (FPR) and true positive rate (TPR) as a function of the classifier parameters $(R_1,\ldots,R_g)$.

Let us use the standard notation:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \qquad \text{and} \qquad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

where

$$\begin{aligned}
\mathrm{TP} &= \#\text{true positives; i.e., resistant isolates predicted as resistant}\\
\mathrm{FP} &= \#\text{false positives; i.e., susceptible isolates predicted as resistant}\\
\mathrm{FN} &= \#\text{false negatives; i.e., resistant isolates predicted as susceptible}\\
\mathrm{TN} &= \#\text{true negatives; i.e., susceptible isolates predicted as susceptible}
\end{aligned}$$

Given our assumptions, we can estimate the performance of a classifier $C_{R_1,\ldots,R_g}$ as

$$\begin{aligned}
\mathrm{TP} &\approx \tfrac{N}{g}\textstyle\sum_{i=1}^g r_i R_i & \mathrm{FP} &\approx \tfrac{N}{g}\textstyle\sum_{i=1}^g (1-r_i) R_i\\
\mathrm{FN} &\approx \tfrac{N}{g}\textstyle\sum_{i=1}^g r_i (1-R_i) & \mathrm{TN} &\approx \tfrac{N}{g}\textstyle\sum_{i=1}^g (1-r_i)(1-R_i)
\end{aligned}$$

where N is the number of samples tested. We then obtain the following estimates:

$$\mathrm{FPR} \approx \frac{\sum_{i=1}^g (1-r_i)R_i}{\sum_{i=1}^g (1-r_i)} \qquad \text{and} \qquad \mathrm{TPR} \approx \frac{\sum_{i=1}^g r_i R_i}{\sum_{i=1}^g r_i}$$

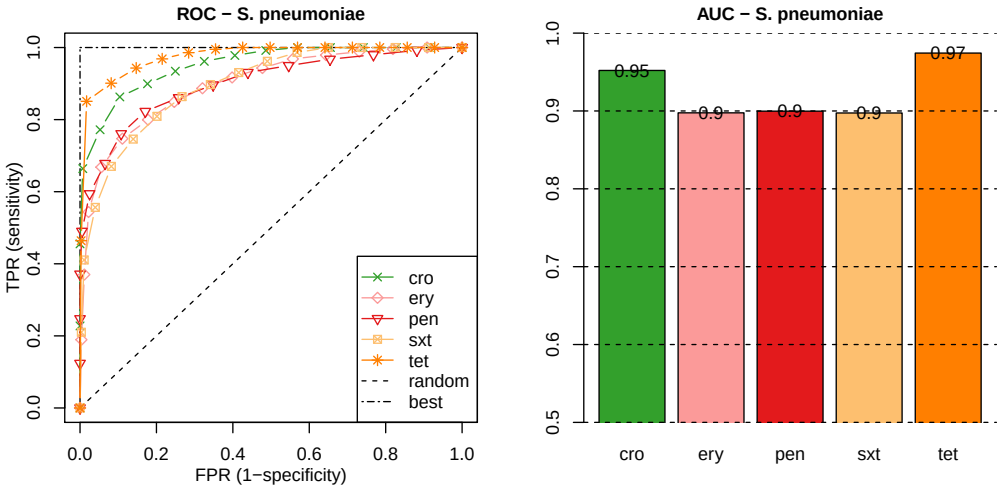

Figure 2: *S. pneumoniae* ROC curves and the corresponding AUCs for ceftriaxone (cro), erythromycin (ery), penicillin (pen), trimethoprim (sxt), and tetracycline (tet).

Using these formulae, we can map every classifier $C_{(R_1,\ldots,R_g)}$ to the ROC space by

$$f(R_1,\ldots,R_g) \to \left(\frac{\sum_{i=1}^{g}(1-r_i)R_i}{\sum_{i=1}^{g}(1-r_i)}, \frac{\sum_{i=1}^{g} r_i R_i}{\sum_{i=1}^{g} r_i} \right)$$

It is easy to see that the map f is linear. Since the set of all possible classifiers is the cube $[0,1]^g$ in the parameter space, it is convex and its image in the ROC space must be convex too; an example is provided in Figure 1. Moreover, the image of the cube is equal to the convex hull of the images of individual cube vertices (red points in Figure 1).

2.3 Optimal ROC curves

Our aim now is to find the optimal ROC curve maximizing the AUC (the red curve in Figure 1). Even though the curve corresponds to infinitely many classifiers, it is a piece-wise linear function which is fully defined by the $g+1$ "vertex" classifiers $C^{(0)},\ldots,C^{(g)}$ lying on the line intersections. Enumerating this classifier sequence corresponds to putting lineages to the order in which they are switched from susceptible to resistant (i.e., $R_i = 0 \to R_i = 1$) along the red line. The first classifier $C^{(0)}$, $(0,0)$ in the ROC diagram, corresponds to all lineages being marked as susceptible, while the last classifier, $C^{(g)}$, $(1,1)$ in the diagram to all lineages being predicted marked as resistant.

The optimal ROC curve can be computed in multiple different ways. For instance, we can enumerate all cube vertices in the parameter space, map them to the ROC space, compute the convex hull, and extract its upper part. More efficiently, we can construct the classifier sequence directly in the ROC diagram by the following iterative process. We start at $(0,0)$; i.e., with all lineages marked as susceptible, and at every step we switch one susceptible lineage to resistant so that the corresponding step in the ROC graph has maximal possible slope, and we continue until all lineages are marked as resistant. If lineages are equally probable, it is easy to see that this order corresponds to sorting lineages by r_i.

3 Results

We applied the method to 616 pneumococcal genomes from a carriage study in Massachusetts children (Croucher et al. 2013, 2015) and 1102 clinical gonococcal isolates collected from 2000 to 2013 by the Centers

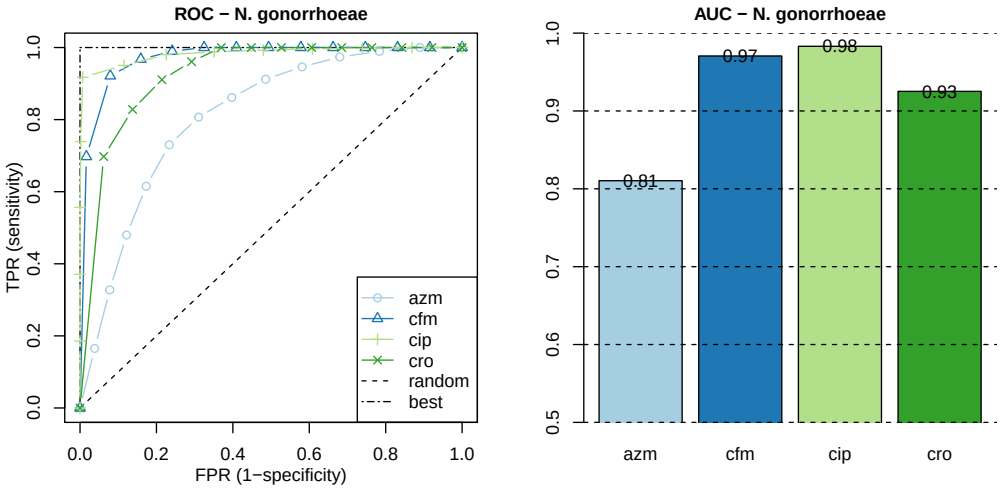

Figure 3: *N. gonorrhoeae* ROC curves and the corresponding AUCs for azithromycin (azm), cefixime (cfm), ciprofloxacine (cip), and ceftriaxone (cro).

for Disease Control and Prevention's Gonococcal Isolate Surveillance Project (GISP) (Grad et al. 2016). For all isolates, we inferred the resistance categories as described in the main manuscript, including the ancestral state responstruction step. As lineages we used the sequenced clusters computed using Bayesian Analysis of Population Structure (BAPS) (Cheng et al. 2013).

Lineages of *S. pneumoniae* are predictive for benzylpenicillin, ceftriaxone, trimethoprim-sulfamethoxazole, erythromycin, and tetracycline resistance with AUC ranging from 0.90 to 0.97 (Figure 2). In *N. gonorrhoeae* ciprofloxacin, ceftriaxone, and cefixime attained comparably large AUCs (from 0.93 to 0.98) whereas azithromycin demonstrated lower association (AUC 0.80) (Figure 3).

Bibliography

Cheng, L., T. R. Connor, J. Siren, D. M. Aanensen, and J. Corander. 2013. "Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software." *Molecular Biology and Evolution* 30 (5): 1224–8. https://doi.org/10.1093/molbev/mst028.

Croucher, Nicholas J, Jonathan A Finkelstein, Stephen I Pelton, Patrick K Mitchell, Grace M Lee, Julian Parkhill, Stephen D Bentley, William P Hanage, and Marc Lipsitch. 2013. "Population Genomics of Post-Vaccine Changes in Pneumococcal Epidemiology." *Nature Genetics* 45 (6): 656–63. https://doi.org/10.1038/ng.2625.

Croucher, Nicholas J., Jonathan A. Finkelstein, Stephen I. Pelton, Julian Parkhill, Stephen D. Bentley, Marc Lipsitch, and William P. Hanage. 2015. "Population Genomic Datasets Describing the Post-Vaccine Evolutionary Epidemiology of Streptococcus Pneumoniae." *Scientific Data* 2 (1). https://doi.org/10.1038/sdata.2015.58.

Grad, Yonatan H., Simon R. Harris, Robert D. Kirkcaldy, Anna G. Green, Debora S. Marks, Stephen D. Bentley, David Trees, and Marc Lipsitch. 2016. "Genomic Epidemiology of Gonococcal Resistance to Extended-Spectrum Cephalosporins, Macrolides, and Fluoroquinolones in the United States, 20002013." *Journal of Infectious Diseases* 214 (10): 1579–87. https://doi.org/10.1093/infdis/jiw420.

Lees, John A., Simon R. Harris, Gerry Tonkin-Hill, Rebecca A. Gladstone, Stephanie W. Lo, Jeffrey N. Weiser, Jukka Corander, Stephen D. Bentley, and Nicholas J. Croucher. 2019. "Fast and Flexible Bacterial Genomic Epidemiology with PopPUNK." *Genome Research* 29 (2): 304–16. https://doi.org/10.1101/gr.241455.118.

Maiden, Martin C. J. 2006. "Multilocus Sequence Typing of Bacteria." *Annual Review of Microbiology* 60 (1): 561–88. https://doi.org/10.1146/annurev.micro.59.030804.121325.