Mahé and Tournoud *BMC Bioinformatics* (2018) 19:383
https://doi.org/10.1186/s12859-018-2403-z

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Predicting bacterial resistance from whole-genome sequences using *k*-mers and stability selection

Pierre Mahé* and Maud Tournoud

Abstract

Background: Several studies demonstrated the feasibility of predicting bacterial antibiotic resistance phenotypes from whole-genome sequences, the prediction process usually amounting to detecting the presence of genes involved in antibiotic resistance mechanisms, or of specific mutations, previously identified from a training panel of strains, within these genes. We address the problem from the supervised statistical learning perspective, not relying on prior information about such resistance factors. We rely on a *k*-mer based genotyping scheme and a logistic regression model, thereby combining several *k*-mers into a probabilistic model. To identify a small yet predictive set of *k*-mers, we rely on the stability selection approach (Meinshausen et al., J R Stat Soc Ser B 72:417–73, 2010), that consists in penalizing logistic regression models with a Lasso penalty, coupled with extensive resampling procedures.

Results: Using public datasets, we applied the resulting classifiers to two bacterial species and achieved predictive performance equivalent to state of the art. The models are extremely sparse, involving 1 to 8 *k*-mers per antibiotic, hence are remarkably easy and fast to evaluate on new genomes (from raw reads to assemblies).

Conclusion: Our proof of concept therefore demonstrates that stability selection is a powerful approach to investigate bacterial genotype-phenotype relationships.

Keywords: Genotype-phenotype, Feature selection, *k*-mers, Lasso

Background

Recent advances in Next-Generation Sequencing (NGS) technologies provided new tools to sequence large amounts of DNA at a reasonable cost and in a limited period [1]. This technological breakthrough is expected to significantly modify the landscape and practices in the field of clinical microbiology. Microorganisms can now be characterized with unprecedented resolution, which can have a significant impact for both research and diagnostics purposes (see e.g., [2, 3]). In terms of diagnostics, NGS indeed holds the promise of addressing, in a single experiment, the main questions of clinical interest: identifying an isolate and determining its antibiotic resistance and virulence profile [4]. The genetic bases of antibiotic resistance and virulence remain partly unknown for most bacterial species, and it is still on open question whether the resistance or virulence of a microorganism can be inferred from its genome only. A recent study showed for example that an isogenic bacterial population exhibited heterogeneity in drug susceptibility due to random partitioning of efflux-pumps during cellular division [5].

Nevertheless, several works have demonstrated the feasibility of genotypic approaches for detection of antibiotic resistance, where a good concordance has been observed between resistance phenotypes predicted from microorganisms genomes, and their reference phenotypes, determined experimentally by assessing their ability to grow in the presence of antibiotics. The genetic bases of antibiotic resistances are for instance well known for *Staphylococcus aureus* and *Mycobacterium tuberculosis*, and accurate predictions could be achieved for these species by simply detecting specific genetic resistance determinants [6–11]. This strategy is notably implemented in well-established tools like `Mykrobe` [7] and `TBProfiler` [9], that leverage for instance catalogs of more than 200 and 1300 mutations, respectively, to predict resistance of *M. tuberculosis* to 9 and 10 antibiotics or antibiotic families. While this

*Correspondence: pierre.mahe@biomerieux.com
bioMérieux, Chemin de l'Orme, 69280 Marcy l'Etoile, France

approach proved to be effective for these relatively clonal species, it may suffer from two limitations if transposed to other bacterial species and/or drugs. The first limitation is that it intrinsically relies on prior knowledge of resistance determinants, which may not be available for all species. Non-coding regions are not explored either, which could be beneficial even for *M. tuberculosis*, for which resistance is mostly due to the presence and accumulation of mutations and indels within a limited number of core genes [12–14]. Secondly, resistance prediction rules typically rely on the presence of at least one resistance determinant, whereas it may be beneficial to combine several ones in a common prediction model to address complex multi-factorial resistance mechanisms, or to model the accumulation of several mutations eventually leading to resistance [12, 15]. With these two limitations in mind, we address the problem from the supervised learning perspective and build multi-factorial prediction rules from a panel of strains, for which whole-genome sequences and resistance phenotypes are available, without relying on any prior knowledge about the resistance determinants. We adopt a systematic *k*-mer based strain genotyping scheme, where any possible *k*-mer is a candidate determinant, and rely on the logistic regression model to combine several *k*-mers into a probabilistic prediction rule.

Genotyping strains with *k*-mers, where every sequence of length *k* found in a genome is a putative resistance determinant, allows to circumvent knowing the genes involved in antibiotic resistance. This approach offers the additional benefits of being alignment-free and able to capture various types of genetic determinants, like the presence of genes, as well as Single-Nucleotide Polymorphisms (SNPs) and indels that can be located in coding or non-coding regions [16]. Such *k*-mer based representations are therefore increasingly popular in this context, for both genome-wide association studies [16, 17] and predictive modelling [18, 19]. The probabilistic framework offered by the logistic regression model is also appealing. First, it naturally combines several genomic determinants in a global predictive model with weights modulating their respective effects, hence quantifying their relative predictive power and reflecting the fact that they can be associated with different levels of resistance [12]. It also provides a probabilistic prediction, which allows to quantify the confidence the user can have in the results provided.

Our approach is closely related to [18, 19], who recently relied on machine learning approaches to predict categorical antibiotic resistance phenotypes from *k*-mers. The resulting prediction rules are based on the detection of possibly multiple *k*-mers, which are automatically selected by the algorithm and are respectively combined by means of logical operations (conjuctions or disjunctions) or a linear combination. Likewise, [20] recently relied on a standard linear regression model to predict the Minimum Inhibitory Concentration (MIC) from candidate mutations found in pre-defined genes, and [11] explored several machine learning strategies to predict *M. tuberculosis* resistance from a pre-defined list of SNPs, with promising results. By combining *k*-mers and logistic regression, we therefore aim to bridge the gap between these two approaches, hence to build prediction models allowing to combine several genomic determinants within a versatile probabilistic framework, without relying on the prior knowledge of the underlying resistance mechanisms.

For the sake of interpretability and computational efficiency of the prediction, we have the utmost interest in building concise models, involving as few genetic determinants as possible. From the statistical learning perspective, the challenge is to identify a small yet predictive set of *k*-mers from a very large number of redundant and correlated ones (several 100.000's). We rely for this purpose on the so-called stability selection approach [21], that consists in penalizing the logistic regression model with the sparsity-promoting Lasso penalty, within an extensive resampling procedure. We present a proof of concept of this approach for *M. tuberculosis* and *S. aureus* using existing datasets [7, 19]. Our main contribution is to demonstrate that stability selection is a very efficient strategy in this context, leading to robust and extremely sparse signatures of resistance. The empirical results obtained allow furthermore to differentiate between the association and prediction perspectives on antibiotic resistance in bacteria, and suggest several promising leads for further work.

Methods

Strain genotyping with *k*-mers

To genotype a training panel of n assembled bacterial genomes using *k*-mers, we first build a large matrix encoding the presence or absence of all *k*-mers of length $k = 31$ within each strain, using the Ray software [22]. Considering *k*-mers of length 31 is a safe default choice in this context, offering in general a good trade-off between sequence specificity and computational efficiency [18, 19]. Since *k*-mers occurring in too few or too many strains have a limited predictive power, we then discard *k*-mers found in one or two strains only, or in all the strains but one or two. Finally, since this *k*-mer based representation is highly redundant [18, 19], we define sets of equivalent *k*-mers from *k*-mers having the same presence/absence profile in all the strains. For the sake of model building, we can only keep one of them, which in practice dramatically reduces the number of *k*-mers to consider. We keep track of these sets of equivalent *k*-mers in order to better interpret the model obtained and carry out the predictions, as discussed in "Prediction from genome assembly or sequencing reads" and "Model interpretation" sections. Equivalent *k*-mers typically correspond to contiguous stretches of sequences conserved across strains, but can also correspond to non-contiguous stretches in total Linkage Disequilibrium (LD).

k-mer selection using L_1 penalty and stability selection

Logistic regression is a widely used generalized linear model addressing binary classification problems. In our case, it consists in building a linear function defined for a strain represented by a vector $x \in \{0,1\}^p$ as:

$$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j, \tag{1}$$

where p corresponds to the number of non-redundant k-mers obtained by the previous process, and x encodes the presence/absence of these k-mers. To estimate the model coefficients and simultaneously select a limited number of k-mers from a training panel of n strains, one can rely on the L_1 or Lasso penalty and consider the following optimization problem:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} L\left(y_i, f\left(x^{(i)}\right)\right) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2}$$

where $y_i = 1$ if the ith strain, represented by $x^{(i)}$, is resistant and 0 otherwise. The function L is the logistic loss function, which quantifies the discrepancy between the true phenotypes y_i of the strains and the predictions $f(x^{(i)})$ obtained by the model. The λ parameter achieves a trade-off between this empirical error and the Lasso regularization term, and is usually optimized by cross-validation.

Effectively tuning the Lasso regularization parameter is however a challenging problem. Cross-validation techniques usually succeed to build models of good predictive power, but the set of variables with non-zero coefficients is known to be unstable with respect to small variations in the training dataset or in the value of regularization parameter.

Several resampling based strategies have been proposed to cope with this issue [21, 23, 24]. We rely on the stability selection approach of [21], which is illustrated in Additional file 1: Figure S1. It consists in subsampling several times the entire dataset (step 1), solving for each subsample the Lasso-penalized logistic regression (step 2), and computing for each k-mer the proportion of models in which it was selected. This process is repeated for several values of the regularization parameter, which allows to define "stability paths" quantifying the probability of selecting each k-mer along the grid of regularization values (step 3). Instead of optimizing the regularization parameter λ itself, [21] propose to consider a threshold on the probability of being selected at some point on the regularization path: every k-mer whose stability path exceeds this threshold gets ultimately selected. Having identified stable k-mers, we finally fit a standard un-penalized logistic regression model that can be used to make predictions on new genome sequences (step 4). We note however that the issue of optimizing the regularization parameter of the Lasso is cast with this approach into that of optimizing the stability threshold, which can be done by cross-validation. To do so, we repeat the whole process described in Additional file 1: Figure S1 within each cross-validation fold, as detailed in "Models obtained" section. Using this optimized probability threshold, we repeat the process on the entire dataset to build the final model. In practice, we relied on the `R` software, and more precisely on the `glmnet` package [25], to compute regularization paths. Moreover, we considered subsamples involving half of the dataset and normalized each k-mer by its L_2 norm to ensure an homogeneous level of penalization, as suggested and discussed in [21].

Prediction from genome assembly or sequencing reads

Given sequencing data obtained from a new strain, the prediction process amounts to detecting the k-mers selected beforehand and making a prediction based on the score provided by the final logistic regression model, which can be turned into a probability of being resistant by the logistic function. We do not tolerate any mismatch to call a k-mer present and rely on the `nucmer` utility of the `Mummer` package [26] to do so. Predictions can be obtained from an assembled genome or directly from sequencing reads. In the latter case, however, a minimum threshold related to the sequencing depth must be considered to call a k-mer present from its number of occurrences, in order to be robust to sequencing errors.

Importantly, to call present each k-mer involved in a model, we rely on its entire set of equivalent k-mers. We detect each of them, and consider several strategies to ultimately call the k-mer present. The *stringent* strategy calls a k-mer present only if all its equivalent k-mers are detected. Conversely, the *conservative* strategy calls it present as soon as one of them is detected. Between these two possibilities, the *vote* strategy calls a k-mer present when more than half of its equivalent k-mers are detected, and the *smooth* strategy uses the proportion of equivalent k-mers detected instead of a binary presence/absence call. For instance, if we detected 8 k-mers out of a set of 10 equivalent ones, both the *conservative* and *vote* strategies would call the k-mer present, while the *stringent* would not. The *smooth* strategy would use a value of 0.8 – instead of 1 or 0 for presence or absence respectively – thereby effectively modulating the weight given to this k-mer by the logistic regression model, in order to account for the uncertainty in its detection. The optimal strategy to consider may in particular depend on the genomic plasticity of the bacterial species under study and the extent to which the training panel of strains properly accounts for it. Provided that the reads of the training data are available, it could easily be optimized by cross-validation techniques: the impact of the various strategies on the predictive performance could indeed be empirically measured and the best strategy retained. In this study, we solely assess the impact of the various strategies on the predictive performance, measured on independent validation sample, and leave the task of optimizing it automatically from the training data for future work. This prediction process takes a few seconds for an assembled genome and typically a couple of minutes for reads, depending on the sequencing depth (see "Predictive performance" section).

Model interpretation

We aim to annotate the k-mers included in the model, to identify whether they fall within known genes or regulatory regions, and provide their putative function when possible. We consider the set of equivalent k-mers associated to each k-mer included in the model and try to reconstruct the longer stretch(es) of sequence(s) that they originate from. A set of equivalent k-mers usually corresponds to a larger sequence perfectly conserved across several strains of the panel, and sometimes to several such sequences in total LD. Following the terminology used by genome assembly algorithms, we assemble equivalent k-mers into longer *unitigs*, defined as the longest sequence(s) that can be obtained when they overlap by exactly $k-1$ nucleotides, using the `bcalm2` software [27]. A unitig is interesting to represent a set of equivalent k-mers because it has the same presence/absence profile on the training genomes, but is in general longer that the individual equivalent k-mers, hence easier to annotate. Unitigs are finally aligned against one or several annotated reference genome(s) using the `blastn-short` program with a minimum of 80% identity and 85% coverage as filtering parameters, which, although not deeply optimized, gave satisfactory results in our experiments.

Results and discussion

We now present a proof of concept of the method described in the previous section on two bacterial species, *M. tuberculolis* and *S. aureus*.

M. tuberculosis study

Dataset constitution

We gathered two datasets from previous studies. The training dataset was taken from [19], who recently made available a set of 1306 assembled *M. tuberculosis* genomes, together with binary resistance phenotypes relating to 7 drugs, namely ethambutol, ethionamide, isoniazid, kanamycin, ofloxacin, rifampicin, and streptomycin. The resistance phenotype of each strain was not always available for each drug, but for each drug the number of resistant and susceptible strains was reasonably high (Table 1). The validation dataset involved 1586 strains[1] that were previously used to validate the performance of the Mykrobe software [7]. Binary resistance phenotypes were available for 5 out of the 7 antibiotics included in the training panel, but not for ethionamide and ofloxacin. We used moxifloxacin as a proxy for ofloxacin resistance, since these two drugs belong to the same family and exhibit a very high level of cross-resistance [28]. We note that a variable number of each phenotype was available for each antibiotic, with in particular only nine strains resistant to kanamycin and ofloxacin (Table 2). This made estimating the sensitivity of the predictive models for these drugs uncertain. We also note that we directly worked from raw sequencing data to obtain predictions, that is, without relying on a prior step of genome assembly.

A total of 19,876,230 distinct k-mers of length 31 were obtained from the 1306 training (assembled) genomes. 5,113,633 k-mers remained after filtering those occurring in less than three strains or in all the strains but one or two, which approximately resulted in a fourfold reduction. This set of k-mers corresponded to 151,403 sets of equivalent k-mers, among which a single one was randomly picked to define the set of non-redundant candidate variables to learn the models. This drastic reduction of the number of k-mers was due to the high clonality of *M. tuberculosis* genomes.

Models obtained

As described in "k-mer selection using L_1 penalty and stability selection" section and illustrated in Fig. 1, the sole parameter to optimize in our approach is the probability threshold of the stability selection procedure, which defines the set of stable k-mers. In this study, it was optimized by cross-validation over the grid $\{0.6; 0.65; 0.7; 0.75; 0.8\}$, for each antibiotic. Three repetitions of a 10-fold cross-validation process were carried

Table 1 *M. tuberculosis* study: cross-validation results

	R / S	k-mers	AUC	*p*
ethambutol	333 / 712	8 (79)	91.0 (92.0)	0.75
ethioniamide	172 / 250	3 (9)	82.2 (85.3)	0.7
isoniazid	815 / 472	4 (5)	96.2 (96.0)	0.8
kanamycin	187 / 484	2 (49)	90.1 (93.2)	0.8
ofloxacin	238 / 458	1 (1)	91.2 (91.1)	0.8
rifampicin	668 / 533	7 (6)	96.6 (96.6)	0.8
streptomycin	492 / 678	7 (22)	90.8 (92.3)	0.8

R / S: number of resistant and susceptible strains. k-mers: number of k-mers in final models obtained with the stability selection approach and with the classical lasso penalty (between brackets). AUC: cross-validated AUC obtained with the stability selection approach and with the classical lasso penalty (within brackets). p: probability threshold selected for the stability approach

Table 2 *M. tuberculosis* study: validation results

		Stability		Mykrobe	
	R / S	sensi.	speci.	sensi.	speci.
ethambutol	194 / 1391	60.3 (6.9)	97.5 (0.8)	71.6 (6.3)	95.8 (1.1)
isoniazid	370 / 1216	89.7 (3.1)	97.5 (0.9)	84.3 (3.7)	98.6 (0.7)
kanamycin	9 / 460	33.3 (30.8)	98.9 (0.9)	33.3 (30.8)	99.6 (0.6)
ofloxacin	9 / 478	55.6 (32.5)	99.6 (0.6)	55.6 (32.5)	100 (0)
rifampicin	303 / 1262	94.1 (2.7)	99 (0.5)	93.7 (2.7)	99 (0.5)
streptomycin	353 / 1227	77.9 (4.3)	99.1 (0.5)	78.8 (4.3)	99.3 (0.5)

R / S: number of resistant and susceptible strains. Stability : sensitivity (sensi.) and specificity (speci.) values obtained with the stability-based final models. Mykrobe: sensitivity and specificity values obtained with the Mykrobe predictor. Figures into brackets correspond to half of the width of the 95% confidence intervals (CI) that shoud be added and substrated to get the upper and lower bounds of the 95% CI

out and the value of the parameter was chosen according to the average Area Under the (Receiver Operating Characteristic – ROC) Curve (AUC) obtained: the highest threshold allowing to reach the highest AUC value up to one point was retained, which allowed to favour sparser models for a comparable accuracy. The same cross-validation procedure was also applied to evaluate the standard L_1-penalized logistic-regression approach. In both cases we considered a grid of 200 candidate values of the regularization parameter. It was defined by the `glmnet` software and ranged in a log scale from a maximum value defined as the smallest value ensuring that at least one variable is selected (i.e., has a non-null coefficient), to a minimum value defined as this maximum value divided by 10^4. For the stability selection approach, we resampled 100 times the training dataset.

Table 1 summarizes the predictive performance and number of k-mers selected. As expected from [21], the stability procedure led to sparser signatures than the classical L_1 penalization, especially for the ethambutol, kanamycin, and streptomycin antibiotics, for a slight decrease in terms of AUC. The predictive performance remained comparable, with the largest drop of 3 points observed for ethionamide. Figure 1 illustrates the stability selection process for ethambutol. We noted in this example that some k-mers with a relatively high probability of being selected (e.g., $\geq$ 0.7, in orange), hence likely to be important to obtain accurate predictions, entered relatively late in the global regularization path of the Lasso, and vice-versa, which explained why the Lasso model involved many more k-mers. Additional file 1: Figures S2 and S3 show the same curves for the other antibiotics.

Predictive performance

We then evaluated the models obtained by stability-selection on the sequencing reads of the validation panel. We considered thresholds to call a k-mer present from its number of occurrences in reads ranging from 1 to 50, and the four strategies mentioned in "Prediction from assembly or sequencing reads" section to call a k-mer involved in a model present, based on the detection of its equivalent k-mers. While the threshold on the number of occurrences could be optimized for each sample, we systematically set it to 10. No major difference was observed as soon as it was not too low (e.g., 1 or 2 leading to false positive k-mer detection) or not too high (e.g., 25 or 50 missing some k-mers), as illustrated in Additional file 1: Figure S4. The differences observed between the various summarization strategies were also minor, and we systematically relied on the stringent approach.

Table 2 showed that the performances obtained on the validation dataset by our approach and the Mykrobe predictor software were comparable. The most important differences were observed for the antibiotics ethambutol and isoniazid, where Mykrobe showed a higher and lower sensitivity, with an opposite effect on specificity. A ROC curve analysis indicated however that the sensitivity/specificity trade-offs achieved by Mykrobe could be

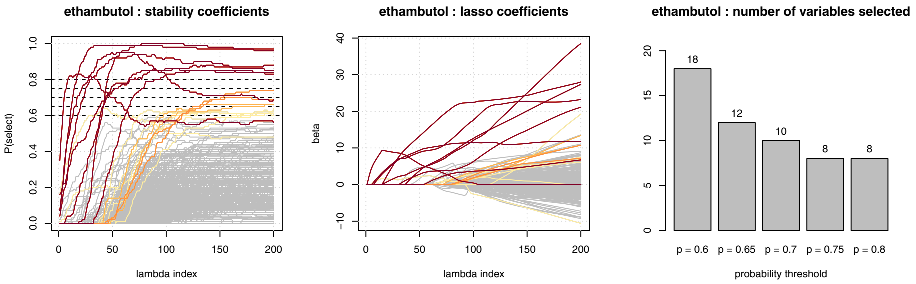


Fig. 1 Illustration of the stability selection process for ethambutol. Left: stability paths. Each curve corresponds to a k-mer and represents its selection frequency over all the resampled datasets, across the values of the regularization paramater. Darker red curves correspond to larger selection probabilities (from 0.6 to 0.8), while grey curves correspond to k-mers with probability of selection below 0.6. Middle : the regularization path obtained by fitting a L_1 penalized logistic regression model across the entire dataset, with k-mers colored according to the color code defined from the left panel. Right : number of k-mers selected by the stability selection approach for thresholds ranging from 0.6 to 0.8

met by modifying the decision threshold of the regression logistic model, which was set by default to 0.5. This is illustrated in Fig. 2 for ethambutol, and in Additional file 1: Figure S5 for the other antibiotics. The flexibility offered by the logistic regression model to control the trade-off between sensitivity and specificity can be useful in a diagnostics context to meet the target performance set by regulatory agencies.

We noted finally that the sequencing depth greatly varied in the validation set, with a minimum value around 20 and a maximum greater than 700. The prediction time increased linearly with the sequencing depth, showing that 1.25 s were necessary to process each coverage unit, which allowed to process a sample with a 100x coverage in about two minutes (Additional file 1: Figure S6).

Model interpretation

To interpret the predictive models obtained, sets of equivalent *k*-mers were assembled into unitigs, as described in "Model interpretation" section, and blasted against the H37Rv reference genome. The unitigs were highly conserved, with a coverage equal to 100% for all the unitigs and a minimum percent identify equal to 96.7% (rifampicin model, unitig #3). Each set of equivalent *k*-mers corresponded to a single unitig whose length ranged from 31 (the size of the individual *k*-mers) to 61 nucleotides, with a median value of 44.5 nucleotides (Additional file 1: Figure S7). These unitigs were easier to annotate than the individual *k*-mers because they were more specific of particular genomic regions, hence led to less ambiguous blast hits.

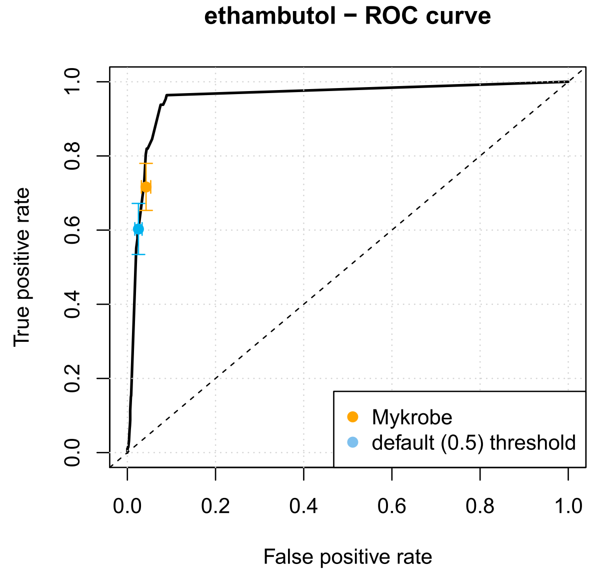


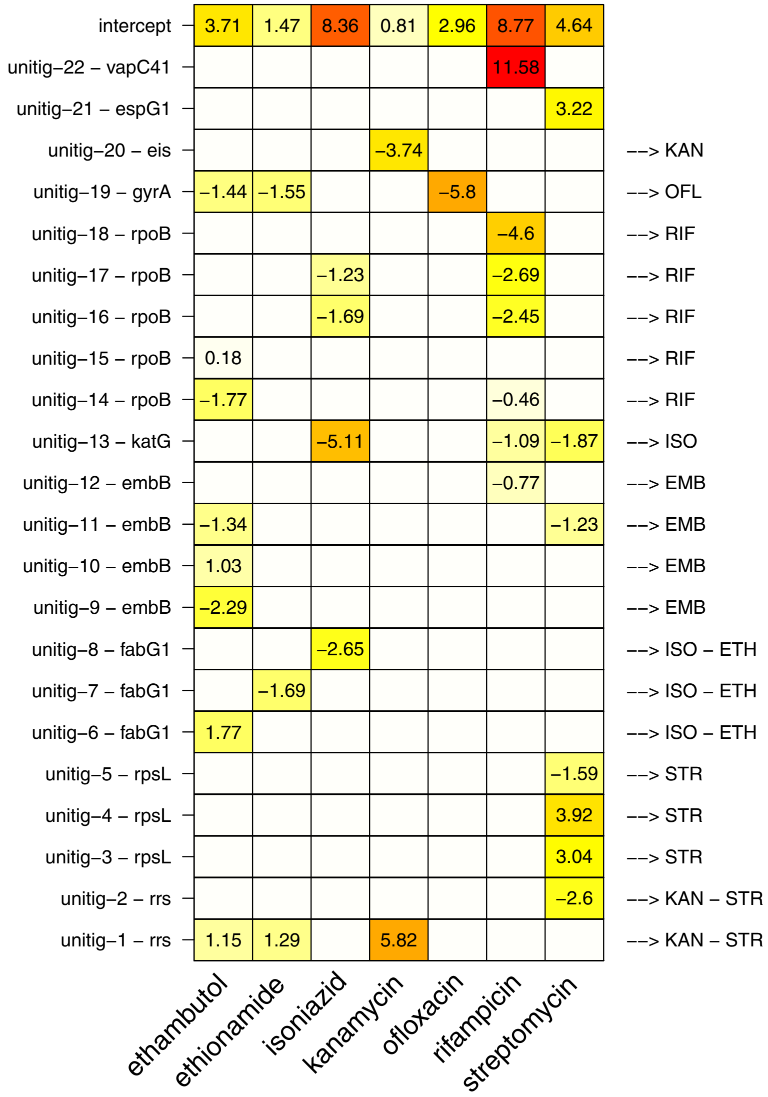
Fig. 2 Ethambutol ROC curve obtained using the *k*-mers based signature evaluated in the validation dataset. The orange dot represents the performance obtained by the Mykrobe predictor, and the blue one to our *k*-mer based approach when using the default threshold of 0.5 to predict a strain resistant based on the probability provided by the logistic regression model. Arrows represent 95% confidence intervals

Figure 3 represents the models and annotations obtained. Interestingly, without any any prior information of known resistance determinants, a total of 22 unitigs was retained (1 to 8 per antibiotic), relating to 10 genes or RNA often already known to be associated with *M. tuberculosis* antibiotic resistance, with unitigs originating from the *embB* gene for ethambutol, *fabG1* for ethionamide, *katG* and *fabG1* for isoniazid, *rss* and *eis* for kanamycin, *gyrA* for ofloxacin, *rpoB* for rifampicin, and *rss* and *rpsL* for streptomycin [12, 29]. Note that the *fabG1* gene is located just before the *inhA* gene, which is one of the two main markers of resistance to isoniazid and ethionamide. Some mutations associated to resistance said to originate from the promoter region of *inhA*, as for instance in [9], could actually be considered to fall within the *fabG1* gene.

We noted that some *k*-mers were part of in several signatures. This was probably due in large part to the level of correlation between resistance phenotypes observed within the training panel (Additional file 1: Figure S8). While in some cases such correlation is expected, for instance between rifampicin and isoniazid or between ethionamide and isoniazid [12], it may also be a consequence of a peculiar constitution of the training dataset due to a sampling bias, as discussed in "Conclusion" section.

Interestingly, two *k*-mers involved in four signatures originated from the 16S RNA (*rrs*). While this could be expected for kanamycin and streptomycin [12], we noted also that these *k*-mers captured general population or "clade" effects, as illustrated in Additional file 1: Figure S9. This suggested that resistance may in some cases be intrinsically related to the evolutionary history of the strains, which makes identifying causal resistance determinants difficult [17].

A striking observation was that the predictive models involved much fewer genetic determinants than alternative approaches relying on catalogues of mutations. For instance, [9] compiled a library of 1345 mutations used to predict the resistance to 15 antibiotics. Narrowing the list to the 7 antibiotics considered in this study led to a list of 745 mutations, much greater than the number of *k*-mers obtained by our approach. There are at least two reasons for that. First, the library of mutations was obtained by compiling results of several studies, involving altogether probably a much greater number of strains than used here. Reproducing our approach on a more exhaustive dataset would most likely lead to selecting more *k*-mers. However, we noted that *k*-mers allowed to represent in a concise way complex patterns of genetic variations, as illustrated in Fig. 4 for 3 genes included in the ofloxacin, isoniazid,



	ethambutol	ethionamide	isoniazid	kanamycin	ofloxacin	rifampicin	streptomycin	
intercept	3.71	1.47	8.36	0.81	2.96	8.77	4.64	
unitig−22 – vapC41						11.58		
unitig−21 – espG1							3.22	
unitig−20 – eis				−3.74				−−> KAN
unitig−19 – gyrA	−1.44	−1.55			−5.8			−−> OFL
unitig−18 – rpoB						−4.6		−−> RIF
unitig−17 – rpoB			−1.23			−2.69		−−> RIF
unitig−16 – rpoB			−1.69			−2.45		−−> RIF
unitig−15 – rpoB	0.18							−−> RIF
unitig−14 – rpoB	−1.77					−0.46		−−> RIF
unitig−13 – katG			−5.11			−1.09	−1.87	−−> ISO
unitig−12 – embB						−0.77		−−> EMB
unitig−11 – embB	−1.34						−1.23	−−> EMB
unitig−10 – embB	1.03							−−> EMB
unitig−9 – embB	−2.29							−−> EMB
unitig−8 – fabG1			−2.65					−−> ISO – ETH
unitig−7 – fabG1		−1.69						−−> ISO – ETH
unitig−6 – fabG1	1.77							−−> ISO – ETH
unitig−5 – rpsL							−1.59	−−> STR
unitig−4 – rpsL							3.92	−−> STR
unitig−3 – rpsL							3.04	−−> STR
unitig−2 – rrs							−2.6	−−> KAN – STR
unitig−1 – rrs	1.15	1.29		5.82				−−> KAN – STR

Fig. 3 Signatures annotation at the unitig level. In total, 22 unitigs falling in 10 genes were retained. Known target / antibiotics association are shown on the right hand side of the figure. Figures correspond to β coefficients in the unpenalized final logistic model and colors to their magnitude (in absolute value). A negative coefficient leads to a decreased risk of resistance, *i.e.*, the presence of the corresponding unitig in the strain genome is associated with a decreased risk of antibiotic resistance, and conversely for a positive coefficient. A strain is predicted as resistant is the resulting score, taking into account the intercept of the model, is positive

and rifampicin signatures. These graphs were obtained by mapping the unitigs corresponding to the k-mers of the signatures against the training genomes, extracting and aligning the unique hits obtained and eventually representing the unitigs on the aligned haplotypes. In the upper panel, the unitig falling in the *gyrA* gene (obtained from 3 equivalent k-mers) captured a single SNP: a nucleotide other than "G" at the SNP position increased the risk of resistance to ofloxacin. In this simple example, 1 k-mer (or equivalently 1 unitig) indeed corresponded to a single mutation. It appears at position 7581 of the H37Rv chromosome and has already been widely described [9]. In the middle and lower panels, however, unitigs captured more complex resistance determinant patterns than a single SNP. The middle panel showed that a single unitig captured the increased risk of resistance to isoniazid brought by two distinct well known SNPs [9], occurring at positions 2155167 and 2155168 in the *katG* gene, and defining 5 distinct haplotypes. The genomic variability observed in the *rpoB* gene in the lower panel was even more complex. We noted indeed that the unitigs corresponding to three k-mers of the signature fell in the same region of the gene. Eight SNPs were observed in the training panel in this region between positions 761109 and 761161. They fall in

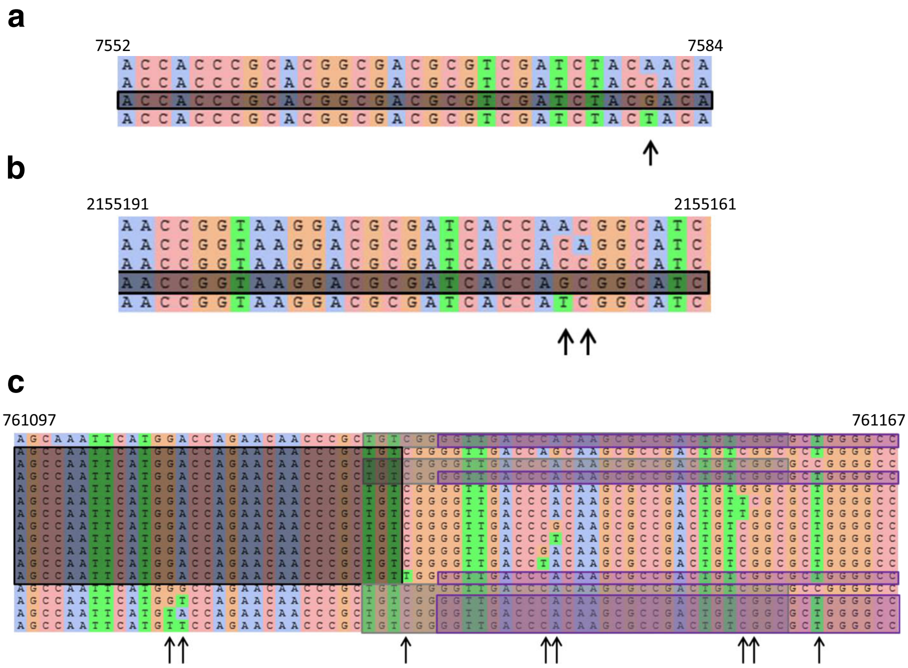

a
7552
7584
ACCACCCGCACGGCGACGCGTCGATCTACAACA
ACCACCCGCACGGCGACGCGTCGATCTACCACA
ACCACCCGCACGGCGACGCGTCGATCTACGACA
ACCACCCGCACGGCGACGCGTCGATCTACTACA
b
2155191
2155161
AACCGGTAAGGACGCGATCACCAACGGCATC
AACCGGTAAGGACGCGATCACCACAGGCATC
AACCGGTAAGGACGCGATCACCACCGGCATC
AACCGGTAAGGACGCGATCACCAGCGGCATC
AACCGGTAAGGACGCGATCACCATCGGCATC
c
761097
761167
AGCAAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCCACAAGCGCCGACTGTCGGCGCTGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCAGCAAGCGCCGACTGTCGGCGCTGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCCACAAGCGCCGACTGTCGGCGCCGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCCACAAGCGCCGACTGTCGGCGCTGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCCACAAGCGCCGACTGTGGGCGCTGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCCACAAGCGCCGACTGTTCGCGCTGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCCACAAGCGCCGACTGTTGGCGCTGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCCGCAAGCGCCGACTGTCGGCGCTGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCCTCAAGCGCCGACTGTCGGCGCTGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCGACAAGCGCCGACTGTCGGCGCTGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTCGGGGTTGACCTACAAGCGCCGACTGTCGGCGCTGGGGCC
AGCCAATTCATGGACCAGAACAACCCGCTGTTGGGGTTGACCCACAAGCGCCGACTGTCGGCGCTGGGGCC
AGCCAATTCATGGGCCAGAACAACCCGCTGTCGGGGTTGACCCACAAGCGCCGACTGTCGGCGCCGGGGCC
AGCCAATTCATGGTCCAGAACAACCCGCTGTCGGGGTTGACCCACAAGCGCCGACTGTCGGCGCTGGGGCC
AGCCAATTCATGTACCAGAACAACCCGCTGTCGGGGTTGACCCACAAGCGCCGACTGTCGGCGCTGGGGCC
AGCCAATTCATGTTCCAGAACAACCCGCTGTCGGGGTTGACCCACAAGCGCCGACTGTCGGCGCTGGGGCC

Fig. 4 Examples of genomic variation patterns captured by *k*-mers in the ofloxacin, isoniazid, and rifampicin signatures. For each signature, mutliple alignments of haplotypes found in the training dataset are shown. Unitigs are surrounded by colored boxes and coordinates refer to nucleic positions on the H37Rv chromosome (NC_000962.3). **a** Ofloxacin - DNA gyrase subunit A - gene gyrA: a single SNP in *gyrA* predicts oxfloxacin resistance. At the SNP position, the 4 bases can be observed in the training dataset, the haplotype with the "G" being wild-type sensitive phenotype. **b** Isoniazid - catalase peroxidase - gene katG: 2 SNPs in *katG* predict isoniazid resistance, and these 2 SNPs are captured by a single unitig. **c** Rifampicin - DNA directed RNA polymerase beta subunit - gene rpoB: 8 SNPs in *rpoB* predict rifampicin resistance, and these 8 SNPs are captured by 3 unitigs

the well-known rifampicin resistance-determining region [12], and all of them have been described in [9], except that observed at position 761156. Using only 3 *k*-mers, the model could therefore account for a much greater number of SNPs combinations, these eight SNPs defining 16 haplotypes in the training panel.

We note finally that the model coefficients allowed to measure the relative importance of the various *k*-mers defining the signatures. These coefficients could indeed be interpreted as odds ratio, a well known statistical indicator measuring the strength of association between a variable and an outcome. We emphasize also that they defined, through the logistic regression model, a probability of resistance, a positive coefficient indicating a higher risk of resistance. This property was interesting because it allowed us to associate a level of confidence to the prediction, and to control the trade-off between sensitivity and specificity that can be achieved by the model (see "Predictive performance" section).

S. aureus study

We applied the same procedure to *S. aureus*. The training dataset was taken from [6]. It involved 501 genomes and 6 antibiotics, namely ciprofloxacin, erythromycin, fusidic acid, methicillin, penicillin and tetracyclin. Three other antibiotics were not considered here because the number of resistant strains was too limited (7 and 3 out of 501 for rifampicin and gentamicin respectively, and 2 out of 176 for mupirocin). The validation dataset included 470 genomes and phenotypes, and was also used in [7] to demonstrate the performance of the Mykrobe predictor software. Additional file 1: Table S1 gives the number of resistant and susceptible strains for these 6 antibiotics.

Table 3 summarizes the results obtained. We first noted that the stability selection approach also led to very sparse models (1 to 5 *k*-mers per model), often sparser than the classical Lasso. Predictive performance was in general comparable or slightly lower than that obtained by Mykrobe, except for fusidic acid where our model significantly lacked sensitivity. This was also the case, to a lesser extent, for ciprofloxacin where a drop of almost 6 points was observed, which corresponded to a difference of 4 strains out of 65 resistant ones and was not significant, as can be seen from the associated confidence intervals. We noted an effect of the strategy used to call present a *k*-mer involved in a signature from its set of equivalent *k*-mers. We observed indeed that while the stringent strategy was appropriate in most

Table 3 *S. aureus* study: validation results

	Stability			Mykrobe	
	k-mers	sensi.	speci.	sensi.	speci.
ciprofloxacin	1 (18)	89.2 (7.5)	99.8 (0.5)	95.4 (5.1)	99.8 (0.5)
erythromycin	3 (8)	96.2 (4.2)	99.5 (0.7)	98.7 (2.5)	100 (0)
fusidic acid	3 (26)	78 (12.7)	100 (0)	100 (0)	99.1 (0)
methicillin	1 (1)	98.1 (3.6)	100 (0)	100 (0)	100 (0)
penicillin	1 (1)	99.7 (0.5)	88.3 (6.5)	99.7 (0.5)	88.3 (6.5)
tetracyclin	5 (7)	100 (0)	99.8 (0.4)	100 (0)	99.8 (0.4)

R / S: number of R/S strains. Stability: sensitivity (sensi.) and specificity (speci.) values obtained with the stability-based final models. Mykrobe: sensitivity and specificity values obtained with the Mykrobe predictor. Figures into brackets correspond to half of the width of the 95% confidence intervals (CI) that shoud be added and substrated to get the upper and lower bound of the 95% CI

cases, better results could be obtained for penicillin using the smooth strategy and for tetracyclin using either the vote or the conservative strategy. Additional file 1: Figure S10 shows the impact of the various strategies, as well as the threshold on the number of *k*-mer occurrences, on the predictive performance. Table 3 gives the best result obtained for each antibiotic, with a threshold set to ten.

Finally, Additional file 1: Table S2 presents the annotations of the unitigs obtained. Starting without a priori from more than 18 million *k*-mers, reduced to 335.238 filtered and non-redundant ones, known resistance determinants were identified for all drugs. We noted however that sets of equivalent *k*-mers were sometimes assembled into several unitigs, which therefore corresponded to non contiguous stretches of sequences in total LD within the training dataset. This was in particular the case for penicillin and tetracyclin, which may explain why the stringent *k*-mer summarization strategy was not appropriate for these antibiotics.

Conclusion

We applied a machine learning approach to predict bacterial resistance phenotypes, starting from their whole genome sequences and without any prior information about the underlying resistance determinants. Using a penalized logistic regression model, coupled with a stability selection approach, we obtained predictive models involving a very limited number of *k*-mers, yet allowing to reach a performance comparable to alternative state of the art bioinformatics strategies for two bacterial species. The *k*-mers obtained uncovered previously known resistance determinants, thereby confirming that such a data driven strategy is promising to unravel bacterial genotype-phenotype relationships [18, 19]. The approach is generic and could readily be transposed to new bacterial species, and/or phenotypes, provided that adequate training data is available.

By selecting them in a discriminative fashion, our data-driven strategy allows to build complex prediction rules from a limited number of *k*-mers, as shown here for *M. tuberculosis*. A potential drawback of this approach, however, is that it intrinsically relies on the level of information provided by the training dataset. In particular, if the level of genomic variability around a causal determinant (e.g., a SNP) is under-represented in the training data because of a sampling bias, our approach will lead us to build too large sets of equivalent *k*-mers, that will not be detected as such in a new strain showing a different genomic context around the determinant. The various strategies proposed to summarize equivalent *k*-mers may therefore be useful to cope with this issue. Moreover, the study led on *M. tuberculosis* revealed that this approach is sensitive to the level of correlation between phenotypes. Some *k*-mers capturing resistance determinants within the target of a given antibiotic were involved in the model predicting resistance to another antibiotic, simply because strains of the training panels tended to be resistant to both antibiotics. *M. tuberculosis* strains may be simultaneously resistant to several antibiotics, which is in part due to the fact that therapies involve antibiotic "cocktails" [13]. In a predictive context where correlation between antibiotics is a biological reality, we consider that such correlation patterns may actually be helpful and leveraged by the model. The overlap we observed between signatures may actually capture some synergistic effects driving simultaneously the resistance to several antibiotics. Coll et al. [9] showed that specific mutations tended to co-occur among multi-drug resistant strains, and evidence has been reported on other bacterial species that a mutation conferring resistance to a given antibiotic could also increase the level of resistance to other antibiotics [30]. This observation also suggests, however, that explicitly learning jointly these predictive models within a multi-task learning framework may be a promising way to exploit such correlation patterns. Several extensions of the Lasso have been proposed to enforce tasks to share a common support, depending on their level of correlation [31]. They could provide an interesting way to study and leverage such cross-resistance mechanisms. If however this correlation is specific to the training dataset, hence results from a sampling bias, it can clearly compromise the generalization of the model. Drouin et al. [18] and Davis et al. [19] proposed different strategies to compensate for this correlation while learning predictive models, relying respectively on a sub-sampling of the dataset, or a post-processing of the list of *k*-mers selected. Within the penalized regression framework considered here, an alternative strategy could be to rely on multi-task appproaches enforcing tasks to have disjoint supports [32]. More generally, correlation between phenotypes is an issue to establish causal relationships between genetic determinants and antibiotic resistance. It represents a confounding factor for bacterial genome-wide association studies that should be taken into account, as it is commonplace for population structure [17].

The versatile framework of the penalized logistic regression offers many perspectives to further investigate bacterial genotypes/phenotypes relationships. Besides the extensions to the multi-task setting mentioned above to model cross-resistance mechanisms, it can easily be extended to consider semi-quantitative measurement of antibiotic resistance, using the MIC as outcome for an ordinal regression model [33]. It is indeed known that mutations can sometimes induce a variable level of resistance [12], and working directly from MICs may lead to better models. Likewise, correlation between strains due to their underlying population structure should be taken into account, by reducing the loss incurred by close strains or relying on mixed-model strategies [34].

In terms of diagnostics, the ability to carry out the prediction from reads coupled with the emergence of nanopore technologies paves the way to real-time sequencing-based applications [35]. A recent proof of concept, led on *M. tuberculosis* and starting from direct respiratory samples, demonstrated the feasibility of this approach [36]. How such *k*-mer based strategies could be transposed to metagenomics settings, in order to predict resistance directly from a sample, with a higher level of sequencing noise, remains an open question and will be the purpose of future work.

Endnote

[1] The original panel involved 1609 strains but 23 of them have missing phenotypes for all the drugs considered in this study.

Additional files

Additional file 1: Supplementary information and figures. (PDF 2415 kb)

Additional file 2: Summary of the training dataset involved in the *M. tuberculosis* study. (CSV 189 kb)

Additional file 3: Summary of the test dataset involved in the *M. tuberculosis* study. (CSV 352 kb)

Additional file 4: Summary of the dataset (training and test) involved in the *S. aureus* study. (CSV 213 kb)

Abbreviations

AUC: Area under the (ROC) curve; LD: Linkage disequilibrium; MIC: Minimum inhibitory concentration; NGS: Next-generation sequencing; ROC: Receiver operating characteristic; SNP: Single-nucleotide polymorphism

Availability of data and materials

All the datasets involved in this study are available in public repositories, as described in Additional file 1: Section S4. CSV files providing links to sequence data and reference phenotypes are available as supplementary materials (Additional file 2, Additional file 3, Additional file 4). A R function implementing the stability selection procedure is available upon request to the authors.

Authors' contributions

PM and MT conceived the study. PM conducted the experiments. PM and MT analyzed the results and wrote the manuscript. Both authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 13 February 2018 Accepted: 1 October 2018
Published online: 17 October 2018

References

1. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity,. Nat Rev Microbiol. 2012;10(9):599–606.
2. Chan JZM, Pallen MJ, Oppenheim B, Constantinidou C. Genome sequencing in clinical microbiology. Nat Biotechnol. 2012;30(11):1068–71. https://doi.org/10.1038/nbt.2410.
3. Bertelli C, Greub G. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. Clin Microbiol Infect. 2013;19(9): 803–13. https://doi.org/10.1111/1469-0691.12217.
4. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook D. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012;13(9):601–12. https://doi.org/10.1038/nrg3226.
5. Bergmiller T, Andersson AM, Tomasek K, Balleza E, Kiviet DJ, Hauschild R, Tkačik G, Guet CC. Biased partitioning of the multidrug efflux pump *AcrAB-TolC* underlies long-lived phenotypic heterogeneity. Science. 2017;356(6335):311–5.
6. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney F, Kearns AM, Pichon B, Young B, Wilson DJ, Llewelyn MJ, Paul J, Peto TEA, Crook D, Walker AS, Golubchika T. Prediction of *Staphylococcus aureus* Antimicrobial Resistance by Whole-Genome Sequencing. J Clin Microbiol. 2014;52(4):1182–91.
7. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, Earle S, Pankhurst L, Anson L, de Cesare M, Piazza P, Votintseva AA, Golubchik T, Wilson DJ, Wyllie DH, Diel R, Niemann S, Feuerriegel S, Kohl TA, Ismail N, Omar SV, Smith EG, Buck D, McVean G, Walker AS, Peto T, Crook D, Iqbal Z. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat Commun. 2015;6:10063.
8. Walker TM, Kohl TA, Omar SV, Hedge J, Elias CDO, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ, Clifton DA, Kapatai G, Ip CLC, Bowden R, Drobniewski FA, Allix-Béguec C, Gaudin C, Parkhill J, Diel R, Supply P, Crook D, Smith EG, Walker AS, Ismail N, Niemann S, Peto TEA. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis. 2015;15:1193–202.
9. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, Mallard K, Nair M, Miranda A, Alves A, Perdigão J, Viveiros M, Portugal I, Hasan Z, Hasan R, Glynn JR, Martin N, Pain A, Clark TG. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 2015;7(1):51.
10. Schleusener V, Köser CU, Beckert P, Niemann S, Feuerriegel S. *Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. Bioinformatics. 2018;4(10):1666–71. (see https://www.ncbi.nlm.nih.gov/pubmed/29240876).
11. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, Peto TEA, Crook D, Smith EG, Zhu T, Clifton DA. Machine learning for classifying

tuberculosis drug-resistance from DNA sequencing data. Bioinformatics. 2017;801.

12. Palomino JC, Martin A. Drug resistance mechanisms in *Mycobacterium tuberculosis*. Antibiotics. 2014;3:317–40.
13. Zhang Y, Yew WW. Mechanisms of drug resistance in *Mycobacterium tuberculosis*. Int J Tuberc Lung Dis. 2009;13:1320–30.
14. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, Liu Z, Li C, Galwey N, Deng J, Zhou Y, Zhu Y, Gao Y, Wang T, Wang S, Huang Y, Wang M, Zhong Q, Zhou L, Chen T, Zhou J, Yang R, Zhu G, Hang H, Zhang J, Li F, Wan K, Wang J, Zhang X-E, Bi L. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet. 2013;45: 1255–60.
15. Palmer AC, Kishony R. Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. Nat Rev Genet. 2013;14: 243–8.
16. Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Honkela A, Parkhill J, Bentley SD, Corander J. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun. 2016;7:12797.
17. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CCA, Iqbal Z, Clifton DA, Hopkins KL, Woodford N, Smith EG, Ismail N, Llewelyn MJ, Peto TE, Crook D, McVean G, Walker AS, Wilson DJ. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016;1: 16041.
18. Drouin A, Giguère S, Déraspe M, Marchand M, Tyers M, Loo VG, Bourgault A-M, Laviolette F, Corbeil J. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics. 2016;17(1):754.
19. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, Overbeek R, Santerre J, Shukla M, Wattam AR, Will R, Xia F, Stevens R. Antimicrobial resistance prediction in PATRIC and RAST. Sci Rep. 2016;6:27930.
20. Eyre DW, De Silva D, Cole K, Peters J, Cole MJ, Grad YH, Demczuk W, Martin I, Mulvey MR, Crook D, et al. WGS to predict antibiotic MICs for *Neisseria gonorrhoeae*. J Antimicrob Chemother. 2017;72(7):1937–47.
21. Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc Ser B. 2010;72:417–73.
22. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biology. 2012;13(12):122. https://doi.org/10.1186/gb-2012-13-12-r122.
23. Bach FR. Bolasso: model consistent lasso estimation ,through the bootstrap. In: Cohen WW, Mccallum A, Roweis ST, editors. International Conference on Machine Learning; 2008. p. 33–40. http://doi.acm.org/10.1145/1390156.1390161.
24. Lim C, Yu B. Estimation stability with cross-validation (ESCV). J Comput Graph Stat. 2016;25(2):464–92.
25. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
26. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):12.
27. Chikhi R, Limasset A, Medvedev P. Compacting De Bruijn graphs from sequencing data quickly and in low memory. Bioinformatics. 2016;32(12): 201–8.
28. Li J, Gao X, Luo T, Wu J, Sun G, Liu Q, Jiang Y, Zhang Y, Mei J, Gao Q. Association of *gyrA/B* mutations and resistance levels to fluoroquinolones in clinical isolates of *Mycobacterium tuberculosis*. Emerg Microbes Infect. 2014;3(3):19.
29. Jnawali HN, Ryoo S. First- and second-line drugs and drug resistance. In: Mahboub BH, Vats MG, editors. Tuberculosis- Current Issues in Diagnosis and Managment. London: IntechOpen; 2013. Chap. 10.
30. Lázár V, Nagy I, Spohn R, Csörgö B, Györkei A, Nyerges A, Horváth B, Vörös A, Busa-Fekete R, Hrtyan M, Bogos B, Méhi O, Fekete G, Szappanos B, Kégl B, Papp B, Pál C. Genome-wide analysis captures the determinants of the antibiotic cross-resistance interaction network. Nat Commun. 2014;5. https://doi.org/10.1038/ncomms5352.
31. Kim S, Xing EP. Tree-guided group lasso for multi-task regression with structured sparsity. In: International Conference on Machine Learning. USA: Omnipress; 2010. p. 543–550.
32. Vervier K, Mahé P, D'Aspremont A, Veyrieras J-B, Vert J-P. On learning matrices with orthogonal columns or disjoint supports. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer Berlin Heidelberg; 2014. p. 274–289.
33. McCullagh P. Regression models for ordinal data. J R Stat Soc Ser B. 1980;42:109–42.
34. Dundar M, Krishnapuram B, Bi J, Rao RB. Learning classifiers when the training data is not IID. In: International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc.; 2007. p. 756–761.
35. van der Helm E, Imamovic L, Hashim Ellabaan MM, van Schaik W, Koza A, Sommer MOA. Rapid resistome mapping using Nanopore sequencing. Nucleic Acids Res. 2017;45(8):61. https://doi.org/10.1093/nar/gkw1328.
36. Votintseva AA, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, Chatterjee A, Smith EG, Sanderson N, Walker TM, Morgan MR, Wyllie DH, Walker AS, Peto TEA, Crook D, Iqbal Z. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. J Clin Microbiol. 2017;55(5):1285–98. https://doi.org/10.1128/jcm.02483-16.