Quantifying fluoroquinolone resistance-inducing point mutations from metagenomics data

Simona Cernat
Molecular and Cellular Life Sciences

September 2023

Examiner and supervisor
Dr. Aldert Zomer
Associate Proffessor
Clinical Infectiology
Department

Second examiner
Dr. Alex Bossers
Assisstant Professor
Veterninary Medicine

Abstract

Objective: Bacterial resistance to fluoroquinolones primary arises from chromosomal point mutations in the g*yrA* and *parC* genes. Currently, bioinformatics tools designed for measuring resistance from metagenomics data are not able to detect resistance arising from point mutations. To address this gap, two methods for quantifying fluoroquinolone resistance-inducing SNPs from metagenomic reads were created.

Methods & Results: GyrAPointCounter is a software which can analyze the presence and abundance of point mutations present on *gyrA* that cause resistance and infer the relative proportion of reads carrying potential resistance causing sequences vs wildtype sequences. GyrAPointCounter is dependent on the alignment tool DIAMOND for specifically selecting reads mapping to *gyrA*. Additionally, we developed a classification tree machine learning model, which was trained using the physicochemical proprieties of amino acids. Both methods form their decision rules by using the available sequences and the mutational patterns of fluoroquinolone-resistant GyrAs present in the CARD database.

The two methods were validated using an *Escherichia coli* (*E. coli)* WGS dataset (n = 201) and an external dataset comprised of *gyrA* sequences (n = 40) belonging to species distinct from those present in the training data. The results of the analyses show that both methods display excellent concordance with the phenotypic data for *E. coli* sequences. The classification performance for novel species for GyrAPointCounter and the supervised learning model showed a True Positive Rate (TPR) of 0.87 and 1 and True Negative Rate (TNR) of 0.75 and 0.68 respectively. An in-house shotgun time series metagenomomics datatset containing *Illumina* short-reads from farm animals treated with enrofloxacin was submitted for the analysis with our two methods. The enrofloxacin-treated samples displayed higher average resistance levels compared to the control groups for both methods.

Conclusions: This research introduced the first version of GyrAPointCounter, which is a promising tool for monitoring the resistance levels from metagenomics data. However, stricter validation is needed before confidently evaluating the tool's performance. For the current version, we propose using the tool as a relative quantification method, rather than absolute.

Layman’s summary

Fluoroquinolones are a class of antibiotics widely used in medical practice. Bacterial resistance to antibiotics is a major public health concern which could become the cause treatment failure against infections. The primary cause of resistance to fluoroquinolones are mutations in two genes, namely *gyrA* and *parC*. Metagenomics is a field of biology where the DNA molecules of multiple microorganisms are analyzed together. For resistance monitoring, metagenomics methods are preferred over other methods. Currently, there are no tools which can screen the levels of fluoroquinolone resistance-inducing mutations from metagenomic data. In this project, we developed a software tool, GyrAPointCounter, which can report the fluoroquinolone resistance levels detected from metagenomic data when resistance is conferred by mutations in *gyrA*. Additionally, we trained a machine learning model for the same purpose. We evaluated the two methods on data where the resistant/susceptible status of the tested sequences was known. The results showed that GyrAPointCounter can detect the resistance in various species with promising accuracy, however, there was some degree of error. The machine learning model displayed satisfactory results, however, it showed a tendency to over-predict a significant number of sequences as resistant even when they were associated with susceptibility. When assessing a metagenomics dataset comprised of samples from enrofloxacin-treated and untreated farm chickens, both our methods detected higher resistance levels for the treatment groups in contrast to the control. This shows that both methods can identify resistant patterns in the data. However, both these bioinformatic methods are in their infant stages of development, and some refinement is needed.

Table of Contents

1. Introduction

In 2014, the World Health Organization acknowledged the emergence of antimicrobial resistance (AMR) as a significant global public health concern, with aggravating consequences expected to escalate in the future (Bengtsson-Palme et al., 2018). AMR is a complex phenomenon which primarily arises from the ability of bacteria to evolve and adapt through genetic mutations. Human activities such as clinical use of antibiotics promotes the development of resistant bacteria which pose consequences for human and animal health (Read and Woods, 2014). Solving the antibiotic resistance crisis requires multiple approaches, including reduced antibiotic use, novel treatment strategies, development of new antibiotics, and also improved surveillance.

Current screening strategies for antimicrobial-resistant organisms involve culture-based methods which have specific limitations. For instance, a significant number of bacteria are non-culturable under laboratory conditions, especially bacterial isolates belonging to environmental samples. These methods bias the results as resistance-associated genes found under this screening method are representative only for a fraction of the microbial population (Waskito et al., 2022). Moreover, clinical testing guidelines provide clear criteria for determining antimicrobial susceptibility, but such guidelines have not been developed for bacteria found in the environment, such as soil or water (Berendonk et al., 2015). The above-mentioned shortcomings of the culture-based methods can be mitigated using metagenomic approaches.

Metagenomics offer remarkable approaches to monitor resistance-linked bacterial genes, by offering a more accurate picture of the whole bacterial community’s resistome. Metagenomics-based screening has thus far been employed for various sources such as livestock, wastewater, the human microbiome, soils, permafrost, and other sources. In metagenomics applications, quantifying antimicrobial resistance typically involves matching the reads against established curated databases via pairwise alignment methods; however, recent strategies are exploring the use of supervised-learning-based methods (Arango-Argoty et al., 2018; Fahrenfeld et al., 2014; Hendriksen et al., 2019; Kim et al., 2022).

Fluoroquinolones are one of the most efficient classes of broad-spectrum antibiotics. However, their widespread use in humans has led to the development of fluoroquinolone resistance, which has become a major concern (Bush et al., 2020). Crucial for fluoroquinolone susceptibility, gyrase and topoisomerase IV are essential enzymes involved in DNA replication. They adjust the topological winding state of DNA by generating staggered cuts on the opposing strands. Fluoroquinolones exert their antimicrobial activity by forming a topoisomerase-drug complex which prevents the ligation process, thus rendering the DNA with an increased and lethal number of double-stranded cuts. Resistance to quinolones can be attributed to substitutions in gyrase and topoisomaraseIV genes such as *gyrA (DNA gyrase subunit A)*, *gyrB (DNA gyrase subunit B)*, *parC (subunit of the DNA topoisomerase IV)* and *parE (subunit of the DNA topoisomerase IV)*. Two of the most common mutations are at Ser83 and Glu87 (*E. coli* numbering) in *gyrA* (Redgrave et al., 2014)

By convention, the specific quinolone resistance-determining region (QRDR) is situated between amino acids 67-106 (*E. coli*). The presence of these mutations prevents the formation of the enzyme-

drug complex and mitigates the antimicrobial effects. While quniolone resistance can also be a result of gene acquisition, most of the times it is caused by single point mutations (SNPs) in the gyrase and topoisomerase genes (Levy et al., 2004 ; Bush et al., 2020).

The metagenomics tools presently available for antimicrobial resistance monitoring are unable to effectively quantify resistance to fluoroquinolone resulting from SNPs, thus new strategies are needed to tackle this issue. DeepARG (Arango-Argoty et al., 2018), for instance, can successfully categorize shotgun metagenomic sequences into resistant or non-resistant classes for various types of antibiotics. Nevertheless, this method does not account for resistance triggered by single point mutations. PointFinder (Zankari et al., 2017) is a web-based tool capable of detecting and quantifying point mutations, but it is species-specific, rendering it inappropriate for metagenomic data. Mumame (Magesh et al., 2019), on the other hand, tackles these issues, but is only specific to sequences already present in databases.

Aim of the project

The aim of the project is to find a suitable method to quantify quinolone resistance-inducing point mutations from metagenomics data. To do this, we developed two methods for identifying and counting fluoroquinolone resistance levels arising from *gyrA*'s mutational hot-spot, the QRDR.

1. GyrAPointCounter is an alignment-based algorithmic tool which is able to quantify the relative proportion of *gyrA* sequences that carry quinolone resistance-inducing SNPs. The underlying algorithm utilizes a set of rules and assumptions to infer the resistant status of a sequence. These assumptions encompass the idea that closely related species are likely to share similar mutational patterns.
2. A tree-based supervised learning model was also trained for this purpose. The model utilizes the physicochemical properties of the amino acids present in the QRDR to predict the AMR phenotype. Both methods are suitable for metagenomics applications.

2. Materials and methods

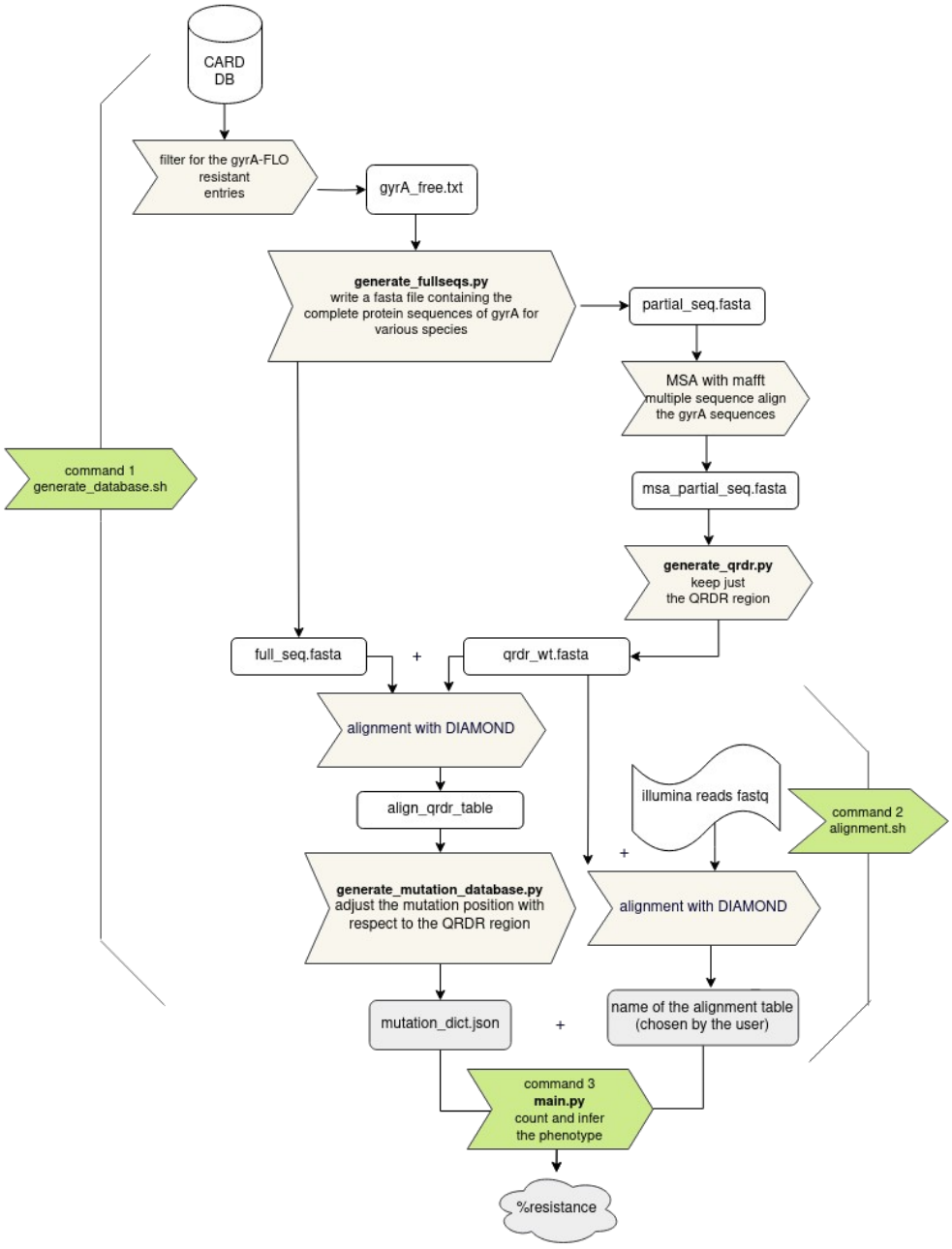


Figure 1. General workflow of the GyrAPointFinder. There are three main commands provided by the tool. The first command, *generate_database.sh* takes the CARD database files as input and outputs a file where species-specific substitutions are stored. This file is referred to as the mutational database (*mutation_dict.json*). The command also outputs a file containing the QRDR regions of the fluroquinolone-susceptible gyrases (qrdr_wt.fasta). The second command uses this file to align the reads provided by the user to the QRDR regions. Finally, the third command utilizes the alignment table generated in the previous step and the mutational database to count the number of potentially fluoroquinolone-resistant QRDR regions.

I. GyrAPointCounter implementation

GyrAPointCounter predicts the resistance phenotype of *gyrA* sequences from *Illumina* short sequencing reads and outputs the ratio of sequences inducing fluoroquinolone resistance mutations over wildtype sequences. The counting is based on a few assumptions which can be adjusted by the user. The main assumptions are: close species likely carry a similar mutational pattern and an evolutionary close amino acid to one which confers resistance, might also confer resistance. GyrAPointCounter uses DIAMOND (Buchfink et al., 2014) to align the reads to the gyrA QRDR reference sequences. Overall there are three main phases provided by three commands underlying our tool. An overview of the GyrAPointCounter steps can be found in **Figure 1** and below.

1. Generate the mutational database and isolate the QRDR region

The first command to run is generate_database.sh. The input for this command is the local path of the newest version of the CARD database. The CARD database (Comprehensive Antibiotic Resistance Database) can be downloaded from the following official webpage. The main purpose of this phase is to create a local database which stores the *gyrA* QRDR reference sequences and the corresponding SNPs (Single Nucleotide Polymorphisms).

a) Extraction the fluoroquinolone-resistant gyrA references and the corresponding mutations

The necessary files for this steps are protein_fasta_protein_variant_model.fasta and snps.txt from the CARD database file. The *protein_fasta_protein_variant_model.fasta* file is then filtered for the GyrA sequences resistant to fluoroquinolones and the *"single resistance variant"* or *"multiple resistance variants"* of the SNPs are selected from the snsps.txt file. The two tables are parsed in the python pandas environment.

b) Creation of the QRDR references file

The complete protein sequence of GyrA and the QRDR region of GyrA are generated (FASTA format). To isolate only the QRDR portion, the sequences were cut using the following approach. First, a file which contains the first 400 amino acids of GyrA proteins is generated. All the sequences in this file are multiple sequence aligned using MAFFT (Katoh and Standley, 2013). The start and the stop position of the QRDR is identified using the reference *E. coli* sequence. The QRDR is chosen to be no longer than 54 amino acids, around the same length as an Illumina short read. The sequences are cut accordingly while the gaps corresponding to insertions or deletions were removed. This resulting file constitutes the QRDR reference database.

c) Creation of the mutational database adjusted to the QRDR region

In the CARD database, the positions of the resistance-conferring mutations are numbered based on the full length of GyrA. To adjust the SNPs positions relative to the QRDR, DIAMOND is used to align the QRDR sequences to the complete gyrase proteins generated in the previous step. The starting

alignment position, the one where the QRDR region starts to align to the reference, was subtracted from the initial, full-length-based mutational position. For example if the initial mutation was on position 86 (according to CARD) and the start position of the QRDR alignment to the full-length reference is the 45th amino acid, then the mutation position in the QRDR sequence is 42 (*86 – 45 + 1)*. The mutations outside the QRDR are excluded. The CARD database also contains combinations of mutations under "multiple resistance variants" meaning that the resistant phenotype was observed when the GyrA possessed all these substitutions. However, it is not confirmed that their synergistic effect is necessary to grant AMR, merely that their co-occurrence has been observed. For this reason, we treated those mutations as being able to grant resistance interdependently. By adopting this approach, some assumptions were introduced with the aim of enhancing the generalization power of our tool.

2. Aligning the reads to the references QRDR regions

To select only the reads belonging to *gyrA*, the *Illumina* sequencing reads provided by the user are aligned against the data base using DIAMOND. The command for the aligned table is the following:

```
diamond blastx -q <QUERY_FILE> --db <REFERENCE_SEQUENCES.dmd> --threads 16 -e
1E-1 --id 70 --query-cover 96 -o OUTPUT_FILE.out  -k 3 -f 6 sseqid qseqid
qstart qstrand qend sstart send qseq_translated sseq length bitscore mismatch
pident
```

A few reads can be smaller as a consequence of sequencing errors, in which case the confidence in those reads is reduced. For this reason, the tool next checks if the length of the queries is at least 45. The combination of minimum identity score of 70% and the minimum query cover of 96% were chosen after testing various different parameter scores.

3. Counting the resistance-conferring mutations

The command which counts and outputs the percentage of the resistance-conferring sequences is main.py. The command also contains an user-adjustable BLOSUM62 matrix threshold flag.

a) Assigning the most probable wild type references

The alignment table is parsed into python pandas environment. The table contains information such as the IDs of the query and the reference, their aligned sequences, the start and the stop positions of the alignments, the length of the alignment, the identity score and the bit score. To assign which mutations should be interrogated in a read, the top alignments of that particular read with the three references are checked. The top three alignments are selected based on the identity score. Only the mutations which are associated with the top three reference are searched for in the translated read (query).

b) Checking which mutations are present in the reads

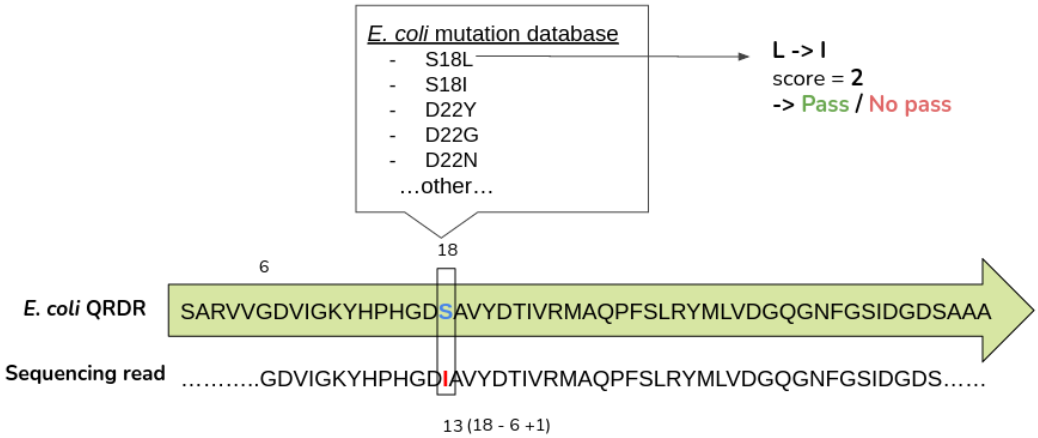


Figure 2. A demonstration of the GyrAPointCounter's computing steps for identifying mutations. A read aligns to the QRDR of a reference sequence (*E. coli* in this example). SNP positions belonging to the mutational database are interrogated for the given reference sequence. The first substitution checked in this example is S18L. To find the equivalent of the 18th position on the translated sequencing read, the position representing the start of the alignment on the reference (6), is subtracted from the mutation position relative to the QRDR (18). Once the position of the SNP is identified, the amino acid which is present at that position in the read is checked. If this amino acid shares a high similarity score on the BLOSUM2 matrix with the substituted one (including itself), then the sequencing read is classified as resistant. This process is repeated for all the mutations for that given reference if the identity score is at least 87%.

The alignment start position might not coincide with the first amino acid of the reference's QRDR. Therefore, the mutational positions within the database formed in the previous step are shifted in relation to the read sequence (Figure 2). To locate the mutation at the correct position, a comparable method involving the subtraction of the alignment start position from the mutational position within the database is employed.

Once the correct SNP location within the read sequence is identified, the position is checked for the presence of the mutation according the species' wildtype reference on which it aligned to. If the identity of the alignment is at least 87%, then the program scans for all the possible substitutions with respect to the reference sequence. If the identity is less, then only the substitutions in the 83, 84, 87 (*E. coli* numbering) positions are considered.

We hypothesized that not all possible resistance-conferring substitutions have been confirmed by research studies. Consequently, certain potential substitutions that could be valid are absent from the CARD database. For example when a mutation involving valine leads to resistance, it is probable that a a substitution to leucine would yield a similar effect due to their chemical similarity. To enact this idea, we integrated the ability to classify potential substitutions as ones that confer resistance. This classification utilizes numerical values derived from the BLOSUM62 substitution matrix. The user has

the flexibility to set a threshold for recognizing these probable substitutions (Figure 2). Finally, the number of resistant sequences is counted against the non-resistant ones. The numbers are reported to the user in the output, along with the percentages in a table.

II. Physicochemical classification tree model

A machine learning model was trained to predict the phenotype based on the sequence. First, the two CARD tables, protein_fasta_protein_variant_model.fasta and snps.txt, were parsed into the python environment as mentioned previously. A file containing the entire protein sequence of gyrA for all the available species in the CARD database was generated (FASTA format). The latest version of the CARD database during this internship contained information about curated resistance-conferring mutations from the following species: *Bartonella bacilliformis, Escherichia coli, Mycobacterium tuberculosis, Staphylococcus aureus, Mycobacterium leprae, Pseudomonas aeruginosa, Campylobacter jejuni, Acinetobacter baumannii, Haemophilus parainfluenzae, Salmonella enterica, Neisseria gonorrhoeae, Capnocytophaga gingivalis, Shigella flexneri, Cutibacterium acnes, Clostridioides difficile, Mycoplasma genitalium, Burkholderia dolosa* and *Helicobacter pylori*.

A suitable training dataset was constructed as follows. For every reference sequence and its corresponding mutation, a mutated sequence was generated. However, these sequences only contained one mutation each; multiple combinations of mutations were not created. The process of extracting the QRDR followed a similar approach as outlined earlier for GyrAPointFinder. In that approach we described how multiple sequence alignment (MSA) was employed. However, here the gaps representing the insertions and deletions were retained. Sequences carrying mutations outside the QRDR were disregarded. The mutated and wildtype QRDRs were parsed to a CSV table format where each position in the alignment was assigned a column. This file was then imported into the R environment.

Each amino acid letter was replaced by five physicochemical numerical values as determined by Atchley et. al, 2005 (Atchley et al., 2005), essentially expanding the number of columns by a factor of five. The columns where the variance was minimal were excluded. A pre-made table containing the Atchley factors was obtained from the following github page (https://github.com/vadimnazarov/kidera-atchley). The wildtype class imbalance was addressed by using oversampling method. A classification tree model was build on this dataset using *the rpart* (Recursive Partitioning And Regression Trees) package.

III. Testing datasets

1. WGS *E. coli* dataset

An in-house *E. coli* dataset was provided by Msc. Aram Swinkels. The dataset contains WGS (Whole Genome Sequencing) *Illumina* pair-end short reads of *E. coli* isolates for which the fluoroquinolone resistance phenotype was determined by selection on McConkey agar plates coated with enrofloxacin. The reads were analyzed using our two methods. After the pairwise alignment step, six reads mapped to the ParC sequence. To avoid this contamination, a filter was applied on read (query) sequences to match a few patterns unique to the *E. coli* ParC sequence. The reads that contained these patterns were discarded from further analysis. For more details about the dataset please contact Msc. Aram Swinkels.

2. External validation dataset

The external validation dataset was compiled from the BV-BRC database which can be found at the following link. This database was queried to identify genomes with recorded phenotypes related to fluoroquinolone treatment. A considerable number of genomes have been downloaded. Those genomes included species outside those belonging to the CARD database (the training set). GyrA sequences were further extracted and the rest of the genomes were discarded. Subsequently, these genes were subjected to pairwise sequence alignment utilizing DIAMOND against QRDR of the reference sequences of species featured within the CARD database.

The alignment table was analyzed using GyrAPointCounter. The pre-output table was analyzed in the python pandas environment. Normally, this table is hidden to the user. This table is identical to the alignment table but it additionally contains a classification prediction column consisting of True for resistant and False for susceptible. The table was filtered for identical isolate genomes, keeping just one of the duplicates. The BV-BRC database does not provide insight into the type of resistance mechanism underlying the resistant phenotype. Consequently a resistant phenotype might not necessarily correlate with mutations in the GyrA QRDR region. To minimize this error, the experimentally validated resistant genomes which are identical to the susceptible reference (100% identity score) were filtered out.

Further, for the validation of the machine learning model, the aligned and translated genome regions were extracted. They were multiple sequence aligned with the training sequences and converted into a CSV table as described previously for the training set, which was imported into R. Similarly to the training set, the the validation set was transformed from amino acid letters into numerical values.

3. Shotgun metagenomics dataset

To evaluate the feasibility of our methods on a real metagenomics usage, a shotgun metagenomics datasets was provided by Msc. Aram Swinkels. In his experiments, he tested the effect of enrofloxacin usage on the chicken gut microbiome. The chicken were divided into three treatment groups and three control groups. The treatment groups were treated enrofloxacin at day 0. A baseline sample was collected at the start of the treatment. *Illumina* shotgun metagenomics was carried to investigate the resistome. The sequencing was carried out by using *Illumina* Novaseq S2 sequencer, 60 million reads per sample, pair end short reads (150 base pairs). Faecal samples were collected at five timepoints

after the treatment for each group while a seacal sample was collected at the sixth timepoint (day 37). The dataset was analyzed with GyrAPointCounter and the tree-based model. For the tree-based model, the query sequences (the reads) were extracted from the alignment table and multiple sequenced aligned with the training set QRDR regions using MAFFT. The results were parsed into a csv file and imported into R for analysis as described above. For more details about the dataset please contact Msc. Aram Swinkels.

4. Notes on the versions

The project was carried out under the following software versions: DIAMOND v2.1.8, MAFFT v7, R v4.2.2, python v3.9.12 with pandas v1.4.2,. The versions of the python libraries and R packages used in this project are available on the github page (currently under construction).

3. Results

I. Internal validation of the CARD dataset

As previously described, GyrAPointCounter relies on a set of rules derived from a few unconfirmed but probable assumptions. While these assumptions are employed with the aim of enhancing the generalization power, this could also bias the algorithm towards the positive class (resistant). To check this hypothesis, GyrAPointCounter was tested on the same CARD dataset from which its rules are derived from. More precisely, we wished to see if some wildtype sequences can be misclassified as resistant due to the potential bias introduced by considering substitutions that surpass the threshold within the BLOSUM62 substitution matrix. To generate a working dataset for this evaluation, we generated all the possible mutated sequences, where just one mutation was incorporated per sequence.

The results showed that all the sequences are predicted correctly when the threshold option for the score of the BLOSUM62 substitution matrix is at least 3. When the threshold is at least a score of 2 then wildtype *Clostridioides difficile* is predicted as resistant. The predicted mutation was R26K (QRDR numbering). These results suggest that running GyrAPointCounter with this cutoff setting might lead to an over-prediction issue.

II. Internal validation on *E. coli* isolates

To assess the performance of our methods on species already available in the CARD database, we used in-house WGS *E. coli* sequencing data. The reads belong to fluoroquinolone-resistant (n = 93) and susceptible isolates (n = 111) for which the phenotype had been confirmed experimentally. ParC and GyrA are subunits of topoisomerases enzymes which both share high sequence similarity in their respective QRDR. The high cutoffs employed during the DIAMOND mapping step, such as the minimal percentage of query cover, were selected especially to minimize the off-target reads, such as those belonging to *parC*. Nonetheless, some *parC* reads still mapped to the *gyrA*'s QRDR region. Normally, GyrAPointCounter does not offer the option to remove the *parC* contamination. However, because this issue can be mitigated in future update of the tool by incorporating *parC* along with *gyrA* in the counting , we analyzed the dataset with and without the inclusion of the invasive *parC* reads. For

this experiment, the results were identical for the cutoffs of 2 and 3 of the substitution matrix. The results are presented in the Table 1.

Sample name	% resistance GPC without *parC* reads	% resistance GPC with *parC* reads	% resistance classification tree model
E. coli susceptible reads	0.09	0.09	0.09
E. coli resistant reads	91	86	91

Table 1. Table illustrating the performance of the supervised learning mode and GyrAPointCounter (GPC). In some cases, the reads belonging to *parC* were filtered out.

Irrespective of the quantification method, out of all the reads which belonged to the experimental susceptible phenotype, only one read was classified as resistant (1 out of 111). Upon visual inspection, the mentioned read contained the S83L mutation which should be indicative of a resistant phenotype. This implies a sequencing error rather than a miss-classification error (such as *bar-code bleeding*). The vast majority of reads associated with the resistant phenotype were correctly classified by our algorithm after excluding *parC* reads. More specifically, 85 out of 93 reads (91%) were labeled as belonging to the resistant isolates. Upon visual inspection, the the misclassified reads were identical to the wildtype reference, meaning that the gyrA gene was not responsible for the resistant phenotype. The machine learning model displayed identical results. Taking all of these findings together, both our methods are performing acculturate on *E. coli* isolates.

III. External validation using GyrAPointCounter on novel species

It is important to evaluate the capacity of our two methods to generalize beyond the training data (CARD data). Therefore, GyrAPointCounter and the machine-learning model were tested on genomes belonging to species different from the CARD dataset, which were downloaded from the BV-BRC database. The *gyrA* genes were extracted from the genomes and submitted to the analysis with our two methods.

The structure of the validation dataset and the outcomes as predicted by GyrAPointCounter can be observed in Table 2. The results are identical when using the “2” and “3” cutoffs for the BLOSUM62 matrix. The Recall of GyrAPointCounter was 0.87 and the Specificity was 0.75. It is worth mentioning that BV-BRC data might contain errors. Upon inspection, the seemingly wrongly classified *Pseudomonas stutzeri* was identical to the wildtype references sequences to which it aligned to. It it highly likely that the resistance is not a result of a mutation in the QRDR, in which case GyrAPointCounter did not mislabel. One of the *Shigella sonnei* isolates contained the S83L mutation which is indicative of a resistant phenotype, however, while the mutation was called out by our algorithm, the BV-BRC label of the isolate was susceptible. Moreover, the two mislabeled *Enterobacter cloacae* isolates harbored the S83F and S83I mutations, while their QRDR regions shared

high similarity to those of *E. coli* and *Salmonella enterica.* For these two species, the S83L and S83F substitutions are resistance inducing, whereas the two *Enterobacter cloacae* reads were labeled as susceptible in the BV-BRC database.

Name species	Susceptible	Resistant	Correctly predicted
Acinetobacter pittii	1	1	Yes
Citrobacter freundii	1	1	Yes
Corynebacterium diphtheriae	1	4	One false negative
Corynebacterium striatum	1	1	Yes
Enterobacter asburiae	1	1	Yes
Enterobacter cloacae	2	4	Two false positives
Shigella sonnei	2	2	One false positive
Klebsiella aerogenes	2	2	Yes
Klebsiella oxytoca	1	0	Yes
Klebsiella pneumoniae	2	3	One false positive
Klebsiella pneumoniae subsp. pneumoniae	1	1	Yes
Proteus mirabilis	1	1	One false negative
Pseudomonas stutzeri	0	2	One false negative
Serratia marcescens	1	1	Yes

Table 2. The composition of the validation dataset and the results inferred by GyrAPointCounter. The number of isolates carrying one of the two phenotypes is indicated on the second and third columns. The last column shows the output classification, where the type of error is specified. The resistant phenotype is considered the positive class.

Finally, the validation dataset was tested on the classification-tree model. The reported Specificity and Sensitivity were 1 and 0.68 respectively. The miss-predicted species included *Enterobacter cloacae, Klebsiella pneumoniae, Shigella sonnei,* which were the same isolates that were wrongly labeled by GyrAPointCounter. While we maintain our doubts about the *Shigella sonnei* and *Enterobacter cloacae* isolates as mentioned earlier, these results are suggestive of a classification bias towards the resistant class.

IV. Shotgun metagenomics dataset validation

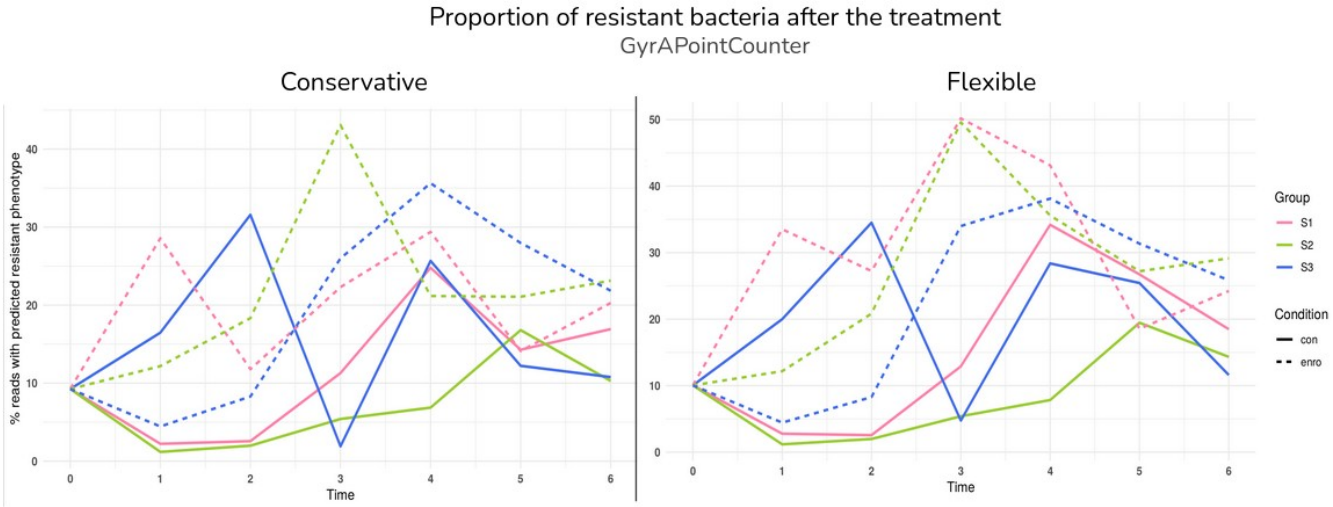


Figure 3. The proportions of reads bearing resistance signatures after the enrofloxacin treatment under the *conservative* and *flexible* options. For both options, the treated samples tend to exceed the resistance proportions of the control. Higher resistance scores can be observed for the *flexible* options. con - control samples where no antibiotic was used; enro - enrofloxacin-treated samples; S1-3 - subgroups

The aim of the project is to create a method to quantify resistance levels from shotgun metagenomics data. To test the feasibility of GyrAPointCounter on a metagenomics dataset, in-house data was analyzed, which has kindly been provided by Msc. Aram Swinkels. Farm chickens were divided into treatment, here named “enro”, and control groups and subsequently into three subgroups, namely S1, S2, S3. Fecal samples were collected at five timepoints after the treatment, while the sixth timepoint is represented by seacal samples. An additional sample was collected at the start of the treatment (timepoint 0).

GyrAPointCounter ran using two distinct configurations here referred to as "flexible" and "conservative." In the "flexible" setting, only substitutions involving amino acids that exhibited a minimum score of 2 on the BLOSUM62 matrix were considered as mutations which confer fluroquinolone resistance. Conversely, the "conservative" option utilized a substitution matrix threshold of 3 to identify AMR-conferring mutations.

Overall, for both options it can be noted that the control groups have lower resistance levels compared to the enrofloxacin-treated samples with few exceptions (Figure 3), which is in agreement with the experimental expectations. In general, the trends of the subgroups are similar for the two options. However, enro-S1 and enro-S3 subgroups were the exception, displaying significantly lower resistance levels at timepoints 3 and 5 respectively in the “conservative” model. For the exact number of reads that mapped to GyrA QRDR and the predictions under the “flexible” option please see the Supplementary table 1.

A

Nr.	Flexible option		Conservative option	
	Top 3 mutations	Occurrences	Top 3 mutations	Occurrences
1	R26K	204	D22H	60
2	D22H	58	T18A	20
3	T18A	20	N18A	16

B

VGDVLGKYHPHGDTAVYHAMV**R**MAQDFSTRGLLVDGHGNFGSVDGD
ARIVGDVLGKYHPHGDTAVYHAMV**R**MAQDFSTRGLLVDGHGNFGSV
IVGDVLGKYHPHGDTAVYHAMV**R**MAQDFSTRGLLVDGHGNFGSVDGDA

Figure 4. (A) Table comparing the top 3 mutations for the time point 3 S1-enro samples for the *flexible* and *conservative* options. **(B)** Three random reads classified as having R26K mutation by the flexible model. In bold it can be observed that the R26K is a false positive.

We next investigated the cause of the appearance of S1-enro peak at timepoint 3 compared to the conservative mode. When inspecting the top mutations for each sample, the flexible model presented additional R26K mutations (Figure 4A). The reads predicted as containing R26K mutations also mapped to the *Clostridioides difficile*, meaning that it employed its decision rules. As observed earlier for the internal validation dataset, the "flexible" model mislabels the *Clostridioides difficile* wildtypes by detecting the R26K, which is a false positive. This exact error is present at the S1-enro timepoint 3 sample. However, when inspecting the reads' sequences, no R26K mutation could be identified (Figure 4B).

We next tested the supervised learning model on the same dataset to see if the results are reproducible despite the two different approaches

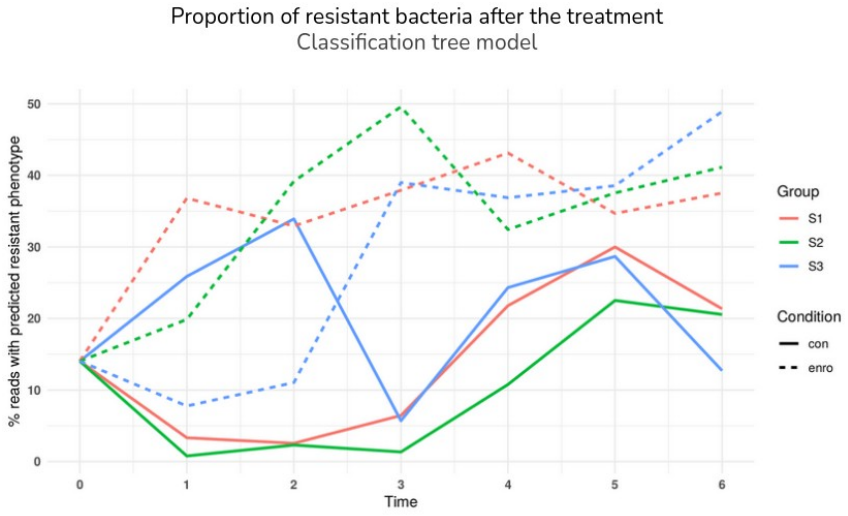


Figure 5. The proportions of reads bearing resistance signatures after the enrofloxacin treatment of farm chickens based on the predictions of the supervised learning classification tree model. Overall, the predicted resistance levels are higher when compared to GyrAPointCounter. con - control samples where no antibiotic was used; enro - enrofloxacin-treated samples

The predictions of the machine learning model show both agreements and disparities between the two methods (Figure 5). There are some consistent peaks predicted by all three models such as the enro-S1 group at timepoint 1, enro-S2 at timepoint 3, enro-S3 at timepoint 2. There are also troughs such as the enro-S2 at timepoint 3 group which are consistent with the previous models. The trends for the control-S1 and control-S3 groups are more similar to those of the “flexible” model. A major difference takes place at timepoint 5 where none of the control groups exceed the resistance levels of the treated samples, unlike the classifications predicted by GyrAPointCounter. Taken together, these results suggest that while the exact levels of fluoroquinolone resistance differ, there are common patterns in the data reported by both our methods.

4. Discussion

Fluoroquinolone resistance arises through either mutations in the QRDR regions of topoisomerase genes (mainly *gyrA and parC*), the acquisition of plasmid genes or overexpression of porins (Hooper & Jacoby, 2015; Cattoir et al., 2007). Currently, there are no existing tools that can quantify fluoroquinolone resistance-inducing point mutations from shotgun metagenomics data. Here, we present a novel tool, GyrAPointFinder and a machine learning model designed for this purpose.

PointFinder, a similar tool capable of detecting point mutations from WGS reads of specific species, employs a database mapping strategy (Zankari et al., 2017). GyrAPointCounter follows a similar approach but extends the database to include all available proteins for gyrA present in the CARD database. This strategy makes GyrAPointCounter suitable for metagenomics data. In contrast to PointFinder's best-hit approach, our classification decision of resistant/susceptible considers the precise position of the mutation compared to the reference gene. Moreover, our algorithm expands its generalization capacity by including a set of assumptions and rules which expand the repertoire of situations in which a SNP can be classified as resistant. For example the decision algorithm accommodates likely substitutions (for instance I → L) by integrating additional rules.

Our methods demonstrate agreement with *E. coli* experimental data. We tested GyrAPointCounter’s performance on an in-house dataset comprised of *E. coli* samples for which the phenotype was determined by antibiotic susceptibility testing. The tool exhibited a perfect classification performance when the *parC* reads were removed. However, the presence of *parC* reads indicates that those have not been filtered out completely during the pairwise-alignment step. The most prevalent mutation identified was S83L, which was previously reported to be the most common fluoroquinolone-resistant substitution in *E. coli (Johnning et al., 2015)*. The classification tree model showed identical results to our GyrAPointCounter. These results validate our methods’ ability to apply appropriate decision rules in the case of *E. coli*.

Our methods demonstrated a strong capacity to generalize effectively. *GyrA* genes belonging to species both related and distant to the ones selected in the CARD dataset were analyzed with our two methods.

Both methods showed an enhanced capacity to detect the resistant class. However, the specificity suffered (0.75 for GyrAPointCounter and 0.68 for the machine learning model). It is worth mentioning that the data was downloaded from a database containing non-curated phenotypic data and for this reason we believe that the Specificity values to be deflated. As mentioned in the results section, there are reasons to believe that phenotype is not directly a reflection of the *gyrA*'s QRDR. This means that some seemingly mislabeled sequences were in fact correctly classified. While we remain reserved in diagnosing the performance of our methods with respect to an absolute Specificity value, the results suggest that the supervised learning model might have a over-prediction bias for the positive class (resistant). This was expected, as the dataset suffered from imbalanced classes, specifically there were more resistant sequences than wildtypes. To solve this issue, an oversampling method was used, which poses the risk of over-training the model.

Lastly, we measured quinolone resistance in a metagenomics dataset where farm chickens were exposed to enrofloxacin and samples were collected after the treatment. Experimental data for *E. coli* (data not shown here; part of the PhD project of Msc. Aram Swinkels) indicated that the second day after treatment the resistant levels significantly rise and steadily remain up for the rest of the experiment. Our tool offers a view at the multi-species resistome and as such, the resistance levels significantly fluctuated between timepoints and conditions. However, except a few instances, the resistance levels in the enrofloxacin-treated chickens exceeded the control samples, demonstrating its capability to comparatively differentiate between different conditions. Some common patterns can be distinguished between the "flexible", "conservative" and the machine learning models, which raises the confidence in those results. However, while the trends are similar the resistance levels values are distinct.

Although the untreated control versus the treated sample at timepoint 3 displayed the most significant fluctuation in resistance, it is possible for the true levels to be different. Microbiome analysis (this data is part of the PhD project of Msc. Aram Swinkels) revealed *Lactobacillus johnsoni* as the most prevalent species in the second and third timeponts samples. The point mutations responsible for quinolone resistance in this species are unknown. Moreover, the QRDR region of *Lactobacillus johnsoni* is evolutionary distant compared to the available CARD reference sequences (Supplementary figure 1). For this species, no mutation could be detected by GyrAPointCounter (data not shown). It is possible that there are point mutations in this species which remain undetected by our approach due to data limitations. Additionally, while the *gyrA* QRDR is typically the primary hot-spot for such mutations, certain species primarily exhibit mutations in other genes, such as *parC* (Xiao et al., 2012), which our current tool version does not account for.

Based on the currently validated data at hand, our tool demonstrated its ability to effectively differentiate resistance levels among variously treated samples for *gyrA*. Future updates might incorporate *parC, gyrB* and *parE* into the measurement. Additionally, hyper parameters such as the maximum number of hits to analyze and the BLOSUM62 substitution values thresholds should be tested further and optimized. If our algorithmic approach is validated as accurate using new meatgenomics datasets with clear experimental outputs, the same principle can be applied to other point mutations conferring resistance to different antibiotics such as tetracycline or streptomycin

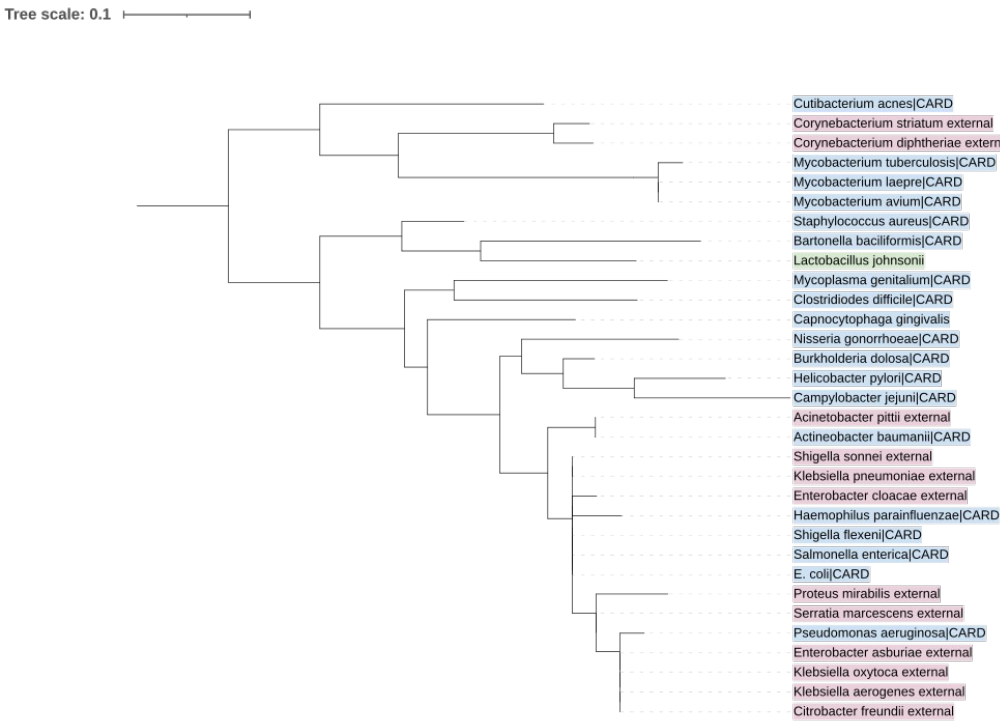
among others (Melançon et al., 1998, Yassin et al., 2005). However, in the case of point mutations in the RNA-encoding genes such as 16S or 23S, a different approach should be investigated, as our current method generalizes based on the BLOSUM62 matrix, which will not be feasible for non-protein coding genes. Finally, more validated sequences available in the online databases, specifically CARD, have a great potential for directly improving the efficiency and reliability of our algorithm.

It was already shown in this project that “flexible” option is prone to *Clostridioides difficile* false positives and might bias the results. This suggests that GyrAPointCounter performed more reliably with a more stringent threshold. The extent on which this conservative threshold affects the generalization power is unknown. Nevertheless, we advise users to interpret the results cautiously. Since the actual resistance levels are still uncertain, we recommend utilizing the tool as a means of relative quantification rather than absolute measurement.

5. Acknowledgments

Of course, this project would not have come to fruition without the help of my supervisor. I would like to thank him for maintaining a very positive attitude throughout the project, especially in the face of constant disappointments during the first months.

6. Supplementary



Supplementary figure 1. Phylogenetic tree of wildtype GyrA QRDR of multiple species. In blue are the species which can be found in CARD while in pink are the QRDR sequences of species validated in our external dataset. The tree was build using IQ-TREE (Nguyen et al., 2015) an visualized using iTOL. The QRDR sequences were aligned using MAFFT.

file_name	trues	falses	condition	%	time	S
./reads/T2-Con-S4S2_merged.out	6	296	con	1.98675496688742	2	2
./reads/T5-Con-S4S2_merged.out	47	215	con	17.9389312977099	5	2
./reads/T2-Con-S4S3_merged.out	55	116	con	32.1637426900585	2	3
./reads/T3-enro-S1S1_merged.out	127	142	enro	47.2118959107807	3	1
./reads/T3-enro-S1S2_merged.out	60	63	enro	48.7804878048781	3	2
./reads/T4-enro-S1S2_merged.out	68	154	enro	30.6306306306306	4	2
./reads/T2-enro-S1S3_merged.out	14	167	enro	7.73480662983425	2	3
./reads/T6-enro-S1S3_merged.out	82	243	enro	25.2307692307692	6	3
./reads/T5-Con-S4S1_merged.out	127	363	con	25.9183673469388	5	1
./reads/T5-Con-S4S3_merged.out	102	299	con	25.4364089775561	5	3
./reads/T5-enro-S1S1_merged.out	36	212	enro	14.5161290322581	5	1
./reads/T6-Con-S4S3_merged.out	40	322	con	11.0497237569061	6	3
./reads/T3-enro-S1S3_merged.out	32	68	enro	32	3	3
./reads/T1-Con-S4S2_merged.out	3	250	con	1.18577075098814	1	2
./reads/T1-enro-S1S1_merged.out	58	124	enro	31.8681318681319	1	1
./reads/T4-Con-S4S1_merged.out	56	178	con	23.9316239316239	4	1
./reads/T6-enro-S1S1_merged.out	103	326	enro	24.009324009324	6	1
./reads/T4-Con-S4S3_merged.out	46	176	con	20.7207207207207	4	3
./reads/T6-enro-S1S2_merged.out	90	260	enro	25.7142857142857	6	2
./reads/T1-enro-S1S2_merged.out	19	137	enro	12.1794871794872	1	2
./reads/T2-enro-S1S2_merged.out	24	96	enro	20	2	2
./reads/T1-enro-S1S3_merged.out	12	258	enro	4.44444444444444	1	3
./reads/T3-Con-S4S1_merged.out	6	56	con	9.67741935483871	3	1
./reads/T4-enro-S1S1_merged.out	76	135	enro	36.0189573459716	4	1
./reads/T2-enro-S1S1_merged.out	66	180	enro	26.8292682926829	2	1
./reads/T4-Con-S4S2_merged.out	8	94	con	7.84313725490196	4	2
./reads/T4-enro-S1S3_merged.out	56	104	enro	35	4	3
./reads/T5-enro-S1S2_merged.out	68	193	enro	26.0536398467433	5	2
./reads/T3-Con-S4S3_merged.out	4	101	con	3.80952380952381	3	3
./reads/T2-Con-S4S1_merged.out	1	116	con	0.854700854700855	2	1
./reads/T1-Con-S4S1_merged.out	3	177	con	1.66666666666667	1	1
./reads/T6-Con-S4S1_merged.out	60	324	con	15.625	6	1
./reads/T5-enro-S1S3_merged.out	73	163	enro	30.9322033898305	5	3
./reads/T3-Con-S4S2_merged.out	3	71	con	4.05405405405405	3	2
./reads/T6-Con-S4S2_merged.out	41	280	con	12.7725856697819	6	2
./reads/T1-Con-S4S3_merged.out	33	137	con	19.4117647058824	1	3

Supplementary table 1. Except from the excel table displaying the results of the metagenomics dataset validation experiment under the *flexible* option. Trues represents the number of resistant reads and falses the wildtype predicted reads. S- subgroup, time - timepoint, condition - weather the samples belong to the treatment or control groups, % - percentage of the resistant reads from the total

7. References

Arango-Argoty, G., Garner, E., Pruden, A., Heath, L. S., Vikesland, P., & Zhang, L. (2018). DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, *6*(1), 1–15. https://doi.org/10.1186/S40168-018-0401-Z/FIGURES/9

Atchley, W. R., Zhao, J., Fernandes, A. D., & Drüke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(18), 6395–6400.

Berendonk, T. U., Manaia, C. M., Merlin, C., Fatta-Kassinos, D., Cytryn, E., Walsh, F., Bürgmann, H., Sørum, H., Norström, M., Pons, M. N., Kreuzinger, N., Huovinen, P., Stefani, S., Schwartz, T., Kisand, V., Baquero, F., & Martinez, J. L. (2015). Tackling antibiotic resistance: the environmental framework. *Nature Reviews Microbiology 2015 13:5*, *13*(5), 310–317. https://doi.org/10.1038/NRMICRO3439

Bengtsson-Palme J., Kristiansson E., Larsson D.G.J. (2018). Environmental factors influencing the development and spread of antibiotic resistance. FEMS Microbiol. Rev. 2018;42:68–80. doi: 10.1093/femsre/fux053.

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods 2014 12:1*, *12*(1), 59–60.

Bush NG, Diez-Santos I, Abbott LR, Maxwell A. (2020) Quinolones: Mechanism, Lethality and Their Contributions to Antibiotic Resistance. Molecules. 2020 Dec 1;25(23):5662. doi: 10.3390/molecules25235662. PMID: 33271787; PMCID: PMC7730664.

Cattoir, V., Weill, F.-X., Poirel, L., Fabre, L., Soussy, C.-J., & Nordmann, P. (2007). Prevalence of qnr genes in Salmonella in France. *Journal of Antimicrobial Chemotherapy*, *59*(4), 751–754. https://doi.org/10.1093/jac/dkl547

Fahrenfeld, N., Knowlton, K., Krometis, L. A., Hession, W. C., Xia, K., Lipscomb, E., Libuit, K., Green, B. L., & Pruden, A. (2014). Effect of manure application on abundance of antibiotic resistance genes and their attenuation rates in soil: field-scale mass balance approach. *Environmental Science & Technology*, *48*(5), 2643–2650. https://doi.org/10.1021/ES404988K

Hendriksen, R. S., Munk, P., Njage, P., van Bunnik, B., McNally, L., Lukjancenko, O., Röder, T., Nieuwenhuijse, D., Pedersen, S. K., Kjeldgaard, J., Kaas, R. S., Clausen, P. T. L. C., Vogt, J. K., Leekitcharoenphon, P., van de Schans, M. G. M., Zuidema, T., de Roda Husman, A. M., Rasmussen, S., Petersen, B., … Aarestrup, F. M. (2019). Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nature Communications*, *10*(1). https://doi.org/10.1038/S41467-019-08853-3

Hooper, D. C., & Jacoby, G. A. (2015). Mechanisms of drug resistance: quinolone resistance. *Annals of the New York Academy of Sciences*, *1354*(1), 12. https://doi.org/10.1111/NYAS.12830

Johnning, A., Kristiansson, E., Fick, J., Weijdegård, B., & Larsson, D. G. J. (2015). Resistance Mutations in gyrA and parC are Common in Escherichia Communities of both Fluoroquinolone-Polluted and Uncontaminated Aquatic Environments. *Frontiers in Microbiology*, *6*.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772-80. doi: 10.1093/molbev/mst010. Epub 2013 Jan 16. PMID: 23329690; PMCID: PMC3603318.

Kim, H., Kim, M., Kim, S., Lee, Y. M., & Shin, S. C. (2022). Characterization of antimicrobial resistance genes and virulence factor genes in an Arctic permafrost region revealed by metagenomics. *Environmental Pollution*, *294*, 118634.

Levy, D. D., Sharma, B., & Cebula, T. A. (2004). Single-nucleotide polymorphism mutation spectra and resistance to quinolones in Salmonella enterica serovar enteritidis with a mutator phenotype. *Antimicrobial Agents and Chemotherapy*, *48*(7)

Magesh, S., Jonsson, V., & Bengtsson-Palme, J. (2019). Mumame: a software tool for quantifying gene-specific point-mutations in shotgun metagenomic data. *Metabarcoding and Metagenomics 3: E36236*, *3*, e36236-. https://doi.org/10.3897/MBMG.3.36236

Melançon P, Lemieux C & Brakier-Gingras L (1988). A mutation in the 530 loop of Escherichia coli 16S ribosomal RNA causes resistance to streptomycin. Nucleic Acids Res. 1988 Oct 25;16(20):9631-9. doi: 10.1093/nar/16.20.9631. PMID: 3054810; PMCID: PMC338768.

Nguyen L.T., Schmidt, H. A., von Haeseler, A., & Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268-74. doi: 10.1093/molbev/msu300. Epub 2014 Nov 3. PMID: 25371430; PMCID: PMC4271533.

Read, A.F. & Woods, R. J. (2014) Antibiotic resistance management. Evol Med Public Health. 2014;2014(1):147

Redgrave, L. S., Sutton, S. B., Webber, M. A., & Piddock, L. J. V. (2014). Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. *Trends in Microbiology*, *22*, 438–445.

Waskito, L. A., Rezkitha, Y. A. A., Vilaichone, R. K., Wibawa, I. D. N., Mustika, S., Sugihartono, T., & Miftahussurur, M. (2022). Antimicrobial Resistance Profile by Metagenomic and Metatranscriptomic Approach in Clinical Practice: Opportunity and Challenge. *Antibiotics*, *11*(5). https://doi.org/10.3390/ANTIBIOTICS11050654

Yassin A, Fredrick K, Mankin AS. Deleterious mutations in small subunit ribosomal RNA identify functional sites and potential targets for antibiotics. Proc Natl Acad Sci U S A. 2005 Nov 15;102(46):16620-5. doi: 10.1073/pnas.0508444102. Epub 2005 Nov 3. PMID: 16269538; PMCID: PMC1283848.

Xiao, L., Crabb, D. M., Duffy, L. B., Paralanov, V., Glass, J. I., & Waites, K. B. (2012). Chromosomal mutations responsible for fluoroquinolone resistance in Ureaplasma species in the United States. *Antimicrobial Agents and Chemotherapy*, *56*(5), 2780–2783. https://doi.org/10.1128/AAC.06342-11

Zankari, E., Allesøe, R., Joensen, K. G., Cavaco, L. M., Lund, O., & Aarestrup, F. M. (2017). PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *Journal of Antimicrobial Chemotherapy*, *72*(10), 2764–2768. https://doi.org/10.1093/JAC/DKX217