UNIVERSITY OF TARTU

Faculty of Science and Technology

Institute of Bioengineering

Aleksander Roosimaa

GenePointer - A Software for Automated Identification of Resistance Markers in Bacterial Genomes

Bachelor's Thesis (12 ECTS)

Curriculum Science & Technology

Supervisors:

Prof. Maido Remm

MSc Erki Aun

Tartu 2025

GenePointer – A Software for Automated Identification of Resistance Markers in Bacterial Genomes

Abstract

Antimicrobial resistance has been a growing threat on the horizon for more than 70 years now, and recent improvements in sequencing technology and bioinformatics tools are making it easier to analyze bacterial genomes. To add upon the already existing functionality of PhenotypeSeeker, a program for phenotype prediction from genotypes, the author created the resistance marker identification pipeline, GenePointer. The program aims to identify the elements in bacterial genomes most associated with their resistance phenotypes, using mapping and alignment approaches and is made to provide insight into novel associations to make the study of resistance mechanisms faster and more efficient. The program managed to identify some of the resistance associated genes in two of the three species and antibiotic combinations analyzed. However, the current version of GenePointer did not identify any novel associations and lacks the statistical power to identify all the known resistance genes. The program in its current form is useful for analyzing and summarizing the resistance associated elements in large numbers of genomes at a time and has potential in future developments to improve on the detection of associations.

Keywords

Antimicrobial resistance, resistance marker identification, k-mer, bacteria, bioinformatics

CERCS: B110

Institute name: Institute of Molecular and Cell Biology

Research group: Chair of Bioinformatics

GenePointer - Bakteritel esinevate antibiootikumi resistentsuse näitajate tuvastamise automatiseeriv tarkvara

Lühikokkuvõte

Antimikrobiaalsete ravimite vastane resistents bakterite hulgas on olnud tõusuteel nüüd juba üle 70 aasta, kuid arenenud on ka paremad ja kiiremad süsteemid ja tarkvara lahendused, mida kasutatakse bakterite genoomide analüüsimiseks. Selleks, et bakterite genoomides resistentsuse elementide otsimist automatiseerida lõi autor programmi GenePointer laiendusena eksisteerivale programmile PhenotypeSeekerile. PhenotypeSeekeri põhifunktsioon on ennustada genoomide fenotüüpe, kuid programm ei annoteeri nende põhjustajaid. GenePointeri eesmärk on katsetada, kas k-meere mis PhenotypeSeeker tuvastab, on võimalik paigutada nende õigetesse asukohtadesse referentsgenoomil ja sealt saada vastavate resistentsusgeenide kirjeldused, kasutades otse referentsigenoomi paigutamist ja joondamist. Lisaks on GenePointer loodud, et katsetada kas programmi kahe k-meeri paigutamis meetoditega on võimalik ka uute resistentsusega seotud elementide tuvastamine ja selleks, et saada automaatset ülevaadet sisendgenoomides sisalduvatest resistentsuselementidest. Oma esimeses versioonis suutis GenePointer tuvastada ainult mõned teadaolevad resistentsusgeenid kahes paaris kolmest valitud bakteri ja antibiootikumi paarist ja uusi seoseid ei leitud. Sellegipoolest on programmil palju potentsiaali kiiremini elemente leida ja arengupotentsiaali, et selle efektiivsust parandada.

Võtmesõnad:
Antimikroobne resistents, resistentsi näitaja tuvastamine, kmeerid, bakterid, bioinformaatika

CERCS: B110

TABLE OF CONTENTS

TERMS AND ABBREVIATIONS

- AMR - Antimicrobial Resistance
- ARG - Antimicrobial Resistance Gene
- BLAST – Basic Local Alignment Search Tool
- FASTA - File Format for Nucleotide Sequences
- GFF - General Feature Format
- GWAS - genome-wide Association Study
- K-mer - Oligonucleotide of length k
- LD - Linkage Disequilibrium
- MDR-TB - Multi-Drug-Resistant Tuberculosis
- mGWAS - microbial genome-wide Association Study
- ML - Machine Learning
- PS - PhenotypeSeeker
- RF - Random Forest machine learning algorithm/classifier
- SAM - Sequence Alignment Map
- SNP - Single Nucleotide Polymorphism
- TDR - Totally Drug Resistant
- Unitig - Combined proximity k-mers
- WGS - Whole Genome Sequencing
- XDR - Cross/Extensively-Drug Resistant

INTRODUCTION

The rising prevalence of antimicrobial resistance (AMR) in bacteria around the world has put a significant burden on the shoulders of clinicians diagnosing patients. Administering the wrong antibiotic can have serious consequences on the outcome of a bacterial infection but choosing the right antibiotic requires extensive testing and evaluation of the properties of the infection at hand. The process of antimicrobial susceptibility (AMS) testing can be a time- and resource-intensive procedure where bacterial isolates must be cultured, sequenced and assayed separately for each antibiotic to determine resistance profiles. A more efficient world of AMS testing is now developing as DNA sequencing gathers speed, and more data is becoming available for machine learning and statistical approaches which enable accurate genome-wide association studies (GWAS). The number of genomes required for accurate statistical analysis and the inherent variation present in bacterial populations requires programs capable of processing large amounts of data while correcting for confounding effects and maintaining a user-friendly nature to make analysis easier for the people using them. A pipeline for finding the locations of resistance associated elements in genomes with minimal user input in tandem to the prediction of resistance in genomes of interest would provide a useful tool to clinicians diagnosing patients and in the research of resistance mechanisms.

The discovery and description of resistance mechanisms starts from identifying the associated genes in resistant isolates. The process of phenotyping genomes statistically instead of in a laboratory can be much faster, cheaper and sometime more accurate. The following is an overview of k-mer based GWAS programs and testing of the author's designed software tool, called GenePointer (GP), consisting of download, analysis and summarising programs, for identifying the elements associated with resistance mechanisms from large quantities of input bacterial genomes. The resulting data from GP is an overview of the input genomes and can be used to make laboratory analysis faster and more efficient.

1. LITERATURE REVIEW

1.1 ANTIBIOTIC RESISTANCE IN BACTERIA

Resistance is defined as the ineffectiveness of an antibiotic towards a bacterium. It can be measured as a binary feature (resistant/susceptible) but can also be presented quantitatively. The lowest concentration of an antibiotic that inhibits the visible growth of bacteria after overnight incubation is called the minimum inhibitory concentration (MIC) (Andrews, 2001) and is used to quantify the resistance level of bacteria and create standards for antibiotic administration. A bacterium is considered resistant to an antibiotic if it can grow in concentrations higher than the MIC breakpoint provided by EUCAST. Resistance genes can be present in bacterial genomes either intrinsically or by acquisition depending on if it developed the resistance or has it by nature (Reygaert et al., 2018, Manusco et al., 2021).

1.1.1 Evolution of Resistant Isolates

Resistance in bacteria appears due to various mutations in random places of their genome, followed by natural selection that eventually leaves only the most resilient bacteria. Bacteria have fast generation times, new variants appear often, and resistance can be the result of many different changes in bacterial cells. Proteins used by a bacterium can change, and so a protein-targeting antibiotic might stop working on the mutated strain. They can limit access to the cell or DNA for the antibiotic by changing properties of their membrane or cell wall, or by obstructing the area of DNA. Some bacteria have efflux pumps that remove the drug from their cells, and some use enzymes to manipulate the drugs before they can reach their target sites (Manusco et al. 2021). Additionally, resistant bacteria share their genes via horizontal gene transfer, and resistance can spread between bacteria of a population without proliferation. The ways in which bacteria can develop resistance are plentiful, and that creates a need for a large variety of antibiotics. Fortunately, all the information on the resistance mechanisms can theoretically be obtained from their genome sequences.

1.1.2 Antibiotic Classes

As the variation between bacteria and their resistance mechanisms is very large, many different types of antibiotics have been developed, each with a specific function and target in fighting bacterial infections. More widely, there are 2 categories: bactericidal, which kill bacteria, and bacteriostatic, which stop them from growing. Narrowing down, there is the group: macrolides, aminoglycosides, tetracyclines, oxazolidinones, and streptogramins, which all inhibit protein synthesis; glycopeptides and beta-lactams that target the synthesis or function of the cell membrane and wall; quinolones which interfere with DNA replication and transcription, and ansamycins that inhibit RNA synthesis. The analysis in the practical part of this paper focuses on the glycopeptide vancomycin, the fluoroquinolone ciprofloxacin and the Mycobacterium specific drug isoniazid. New antibiotics and antibiotic classes are being developed, but the rate at which bacteria become resistant to antibiotics is much faster, and so the drug developers are under a lot of pressure to find a solution to fight the ever-increasing AMR (Anderson et al., 2023).

1.1.3 AMR Genes Targeted in the Practical Part

Resistance associated genes have various ways to contribute to its mechanism. Sometimes there is just one gene with one mutation that makes a particular strain resistant, sometimes it is a series of mutations in the same gene, and often it is a combination of genes and mutations that result in resistance towards an antimicrobial. The genes may be native or acquired from horizontal gene transfer and so the resistance mechanism is often hard to determine from just the core genes.

1.1.3.1 *Mycobacterium Tuberculosis* Resistant to Isoniazid

Isoniazid is a bactericidal antibiotic effective against susceptible members of the Mycobacterium genus. It inhibits the synthesis of mycolic acid, a component of the cell wall. The main resistance-associated gene is *katG*, which, upon mutation, induces resistance to the drug. KatG is a catalase-peroxidase that catalyses the oxidative activation of isoniazid. The gene *inhA*, encoding the isoniazid target enoyl-acyl carrier, has also been reported as resistance-conferring when mutated (Heym et al., 1993; Rostamian et al., 2023). Mutations vary, and different combinations of mutations in both genes provide varying resistance levels quantified

with MIC values in (Ghodousi et al. 2019). *katG* and *inhA* are used as positive controls to determine whether the analysis was successful in the practical section of this paper.

1.1.3.2 *Enterococcus Faecium* Resistant to Vancomycin

Vancomycin and glycopeptide antibiotics, in general, work by binding an intermediate in the synthesis of peptidoglycan, which puts stress on the membrane of a bacterium, eventually killing it. Multiple gene clusters associated with vancomycin resistance have been identified for one, VanA, which contains the reading frames of *vanR, -S, -H, -A, -X, -Y*. These are the genes used as positive controls in the practical section of this paper (O'Toole et al., 2023).

1.1.3.3 *Escherichia Coli* Resistant to Ciprofloxacin

Ciprofloxacin is a fluoroquinolone that inhibits DNA topoisomerase and DNA gyrase and thus DNA replication. Resistance usually forms due to mutations in *gyrA, gyrB, parC,* and *parE* genes, which affect the binding affinity of fluoroquinolones to DNA gyrase and DNA topoisomerase, respectively. *gyrA, gyrB, parC,* and *parE* will be used as indicators of true detection in the practical section of this paper (Neyestani et al., 2023; Thai et al., 2023).

1.1.4 Available Databases of Antimicrobial Resistance Genes

AMR genes have now been extensively annotated, and large databases have been curated. These databases can be used to identify resistance genes present in bacterial genomes, find resistance-conferring mutations, and some provide services for analysis of isolates to determine resistance genes. Some of the larger databases are:

- CARD. The Comprehensive Antibiotic Resistance Database is a source for data about AMR genes, mutations conferring resistance, and algorithms for resistance gene identification.
- ResFinder. A database and program for the identification of acquired antimicrobial resistance genes. Works on whole and partial genomes of bacterial isolates.

- PATRIC. The Pathosystems Resource Integration Centre has bacterial genomes with resistance phenotype annotations and services for analysis of genomes.

1.2 THE SPREAD OF ANTIMICROBIAL RESISTANCE (AMR)

AMR has been on the rise ever since the first antibiotics were created and used publicly. The first antibiotic was created over 100 years ago in 1910 and is called salvarsan, which started the revolution that extended human lifespans drastically (Fleming, 1929). Penicillin, a once very popular antibiotic, first produced in 1928, started the search for better and more diverse antibiotics that peaked in the mid-1950s (Hutchings et al., 2019), but itself has become rather ineffective. Many antibiotics were created and discovered in the 1950s, but the regulation of their use was lacking, which was the start of the problem we are facing today. Antibiotics were used without limitations, disposed of incorrectly, and used for the wrong infections (J. Davies & D. Davies, 2010).

Antimicrobial resistance is rising because the misuse and ever-growing industries' use of antimicrobials are putting a selective force on bacteria, making them rapidly become resistant or in some cases even untreatable (Prescott et al., 2014). Large industries such as agriculture, which use antibiotics to keep animals healthy and their products safe for us to consume, self-medication in countries where antibiotics are sold over the counter and just-in-case diagnosis of antibiotics to patients are among the many causes of increasing AMR rates (Salam et al. 2023). Now, new drug-resistant, multidrug-resistant (MDR), and even totally drug-resistant (TDR) strains are emerging worldwide, putting a lot of pressure on healthcare and many other industries. In 2019, there were an estimated 4.95 million deaths associated with AMR, of which 1.27 million were directly attributable (Murray et al., 2022). Low and middle-income countries are most affected because more specialised drugs are expensive, often inaccessible, and the amount of susceptibility testing is reduced. The region with the highest all-age death rates was sub-Saharan Africa, with 27.3 deaths per 100,000 people. 6.2% of the 490,000 MDR tuberculosis cases were extensively drug-resistant tuberculosis (XDR-TB), which is when, in addition to MDR, tuberculosis has resistance to at least one fluoroquinolone and to one second-line drug (Murray et al., 2022; Polsfuss et al., 2019). A case study from 2016 describes a patient with pulmonary tuberculosis with detectable resistance towards delamanid and bedaquiline,

two next-generation antibiotics for treating XDR (extensively drug-resistant) tuberculosis, showing that the bacteria can develop resistance quickly, and even the most expensive antibiotics won't keep us safe for a long time (Polsfuss et al., 2019). Some of the more widely spread resistant bacteria are bacteria that many people have in their organism, but which don't necessarily cause illness in healthy individuals. One example is MRSA (Methicillin-resistant *Staphylococcus aureus*), which spreads mostly in hospitals via personal contact. MRSA becomes a threat when a person is injured or has a weakened immune system, as it can develop into a serious infection, and the treatment is long and complicated (Coello et al., 1997). Another bacterial type with low treatment success rates is MDR-TB (multidrug-resistant tuberculosis). Resistant to rifampicin and isoniazid, it is hard to treat with success around 63% as of 2020. With around 410,000 people developing it in 2022, it is an infection in need of better treatment (World Health Organisation, 2024). Determining the right treatments for patients is not easy, and many factors need to be considered. Susceptibility testing is important in determining which antibiotic a bacterial infection can be treated with, but the process currently requires microbial culturing, which takes time. Because of this, clinicians often rely on empirical judgment like previous treatment success when diagnosing patients (Weinstein, 2001; Nguyen et al., 2020). If the progress of AMR can't be slowed soon, a study from 2014 estimates it can cause deaths upwards of 10,000,000 per year by 2050, essentially putting us back to the pre-antibiotic era (O'Neill et al., 2014; Mosquera-Rendon et al., 2023).

1.3 GENOME-WIDE ASSOCIATION STUDIES

The ongoing research into identifying antimicrobial resistance mechanisms is wide, and one of the most popular methods for phenotyping bacterial strains is a genome-wide association study (GWAS) originally meant for identifying the single nucleotide polymorphism (SNPs) associated with phenotypes in the human genome. This means the analytical process was optimised for the properties and size of the human genome, but recently it has been optimised for the use on bacterial genomes with great success, and this version is often called microbial GWAS (mGWAS).

1.3.1 Data Acquisition for GWAS

Like all data analysis, GWAS begins with collecting a sufficiently large dataset of genomes and their phenotype values. Larger genomes are harder to sequence hence microarrays are often used to find SNPs and phenotypes can be determined experimentally or observed directly. For a successful study, the collection of genomes should be diverse in relation to the phenotype of interest, quality-controlled for errors, and as large as possible. Usually, imbalances in the distribution of input data phenotypes can lead to skewed associations and so diversity directly affects the results (Loya et al., 2025). When dealing with AMR and considering only binary phenotypes, the results obtained from isolate analysis in the laboratory, which are usually MIC (minimum inhibitory concentration) values, must be translated to binary resistant/susceptible form following the standard EUCAST or CLSI breakpoints. As raw data of phenotyped bacterial genomes is becoming more and more abundant in public databases, this step is becoming less time-consuming, and often isolate genome sequencing and phenotyping are not required. Whole genome studies require a lot of computational power and time, the larger the dataset and genome lengths, increasing the need for better algorithms for association testing. Some studies have successfully shown associations from unassembled sequencing reads and partial genomes, but often this has shortcomings in the genome coverage the results provide (Voichek et al., 2020). The number of input genomes has always been proportional to the accuracy of the results but in some cases the large number of input genomes can be detrimental to the outcome, specifically when using machine learning. The computational cost also depends on the elements, often either SNPs or k-mers (the markers that genetic variants are measured by).

1.3.2 GWAS Methodology

Variations in genomes can be measured as single-nucleotide polymorphisms and short insertions and deletions (SNPs and InDels), presence/absence patterns of genes or the copy number differences of genes. None of the types of variations of the genome mentioned can provide detection power towards all the mutations that may be in a genome. When mapping these variations to a phenotype, it is important to choose the right subdivision of the genome that can encompass as many of these types of variations as possible otherwise some variants may go unnoticed when associating the genotype with the phenotype.

GWASs on larger genomes, mostly human, are usually done by comparing the differences between fixed, predetermined sets of SNPs. The SNPs are single-nucleotide differences with respect to the reference genome and finding them is either done by microarrays that can detect a specific set of SNPs without having to sequence the genome or by sequencing the genome and using alignment against the reference genome to identify nucleotide differences. As of now alignment is one of the major limitations of performing GWAS successfully on larger genomes, being computationally slow and prone to false results. Other kinds of subdivisions include k-mers, unitigs, and sequence variants (Lemay et al., 2023).

As mentioned before, GWAS depends on choosing the right subdivision to measure variation in genomes. K-mers (oligonucleotides of length k) have gained popularity for smaller genome analyses with their benefit of covering larger mutations than an SNP can (Voichek et al., 2020; CRyPTIC Consortium, 2022). For larger genomes, SNPs are still used because a set of fixed SNPs per analysis can minimise the number of potential variants to just the ones of interest and reduce computational costs. Each SNP gives uniqueness to the genome and the set of SNPs is compared with other genomes' sets of SNPs to determine the sets that are associated with the phenotype under investigation. Whether the feature set is made of SNPs or k-mers, the individual features are statistically tested for any association with the phenotype of the genome and the most significant of them are considered the ones to look for when phenotyping an unphenotyped genome. The most intuitive representation of a GWA study result comes from the Manhattan plot, which presents the p-values of each subdivision and their locations in the genome, showing the potential source of the phenotype and spread of resistance conferring elements over the genome. The existing GWAS approaches can be divided into 4 groups: non-phylogenetic where phylogeny of populations is not considered, alignment free where a reference genome is not necessary, phylogenetic which utilises analysis methods based on convergent evolution and the mixed approach which tries to provide deeper insight into the mutations and secondary mutations that can develop the existing phenotype. Further development of faster and more efficient solutions to GWAS is ongoing, many of them aimed at removing the need to make a compromise between computational speed and input data size (Loya et al., 2025).

1.3.3 SNP-based GWAS

The SNP-focused genome-wide association study is the first developed GWA method, with its focus on determining phenotype-conferring genotypes in large genomes when sequencing wasn't available or too expensive. SNPs can be profiled in genomes using microarrays that can detect thousands to millions of predefined SNPs fast and inexpensively. This made SNP-based GWAS applicable before next-generation sequencing, however this method only gives a partial overview of the genotype, and the detection of new variants is less likely. With the innovations in the DNA sequencing world came whole genome sequencing, and a lot more SNPs can be profiled for GWAS, making it possible to detect novel and rarer variants. This approach is popular because it doesn't require expensive laboratory work, and computers are becoming faster at analysing the data. Aligning large genomes is a complicated process and computationally the most expensive, but when complete, the SNP analysis steps are rather straightforward when compared to, for example, k-mers. SNPs found by alignment are filtered for ones present in too many or too few genomes to exclude sequencing errors, alignment errors, or ones that aren't tied to the phenotype. Depending on the species under study, the SNPs can be further filtered for linkage disequilibrium to remove unassociated SNPs. The amount of data produced during analysis isn't larger than the input data because SNPs are single-nucleotide variations, which don't take up a lot of memory. The genotypes are then statistically tested for association with the phenotype to determine the kinds of variations that need to be present for a certain phenotype to arise. These results can further be used to build ML models for phenotype prediction or deeper analysis of the variations causing the phenotype.

1.3.4 Alignment-free GWAS

K-mers are the rising new approach to GWA studies, showing that they can detect more variants with higher confidence (Lemay et al., 2023; He et al., 2021). The usage of k-mers is enabled by the increase in available whole genome data and increasing ease of sequencing, and their effectiveness seems to depend on the quality and length of the sequenced genome. K-mer based GWAS starts with finding all the k-mers present in the genomes and then either forming an absence/presence matrix or a k-mer count matrix. This step is different from SNPs because the input data gets inflated drastically, and the process requires a lot of memory. This inflation of data is often associated with overfitting in machine learning because the input number of

genomes gets cut into a much larger amount of data which can reduce the generalisability of ML models. The choice of matrix usually depends on the study because there are subtle differences between associating the presence/absence or count of a k-mer with the phenotype. For example, when looking for copy number variations, a simple presence/absence pattern of k-mers will not detect them. On the other hand, shorter k-mers can be redundant in larger genomes, and so k-mer counts are more effective to analyse larger genomes. Similarly to SNPs, to remove excess data, k-mers with a low prevalence in genomes may be excluded as they can usually be attributed to sequencing errors, and k-mers present in most of the genomes likely aren't related to the phenotype.

1.3.5 Results and Interpretation of GWAS

GWASs provide insight into the connection between the genotype and phenotype of genomes. The set of features obtained from a GWAS of a species can be used to make much faster predictions of whether the genome of interest is of a specific phenotype or not. These studies also reveal the potential sources of phenotypes for research in a laboratory setting. The results of GWAS are often in the form of tables of mutations or variations and associated loci are usually presented on Manhattan plots. While the plots and tables are good summaries of the data, something that GWAS still needs to improve on is the interpretability of the results in terms of the molecular mechanisms that these phenotypes arise from and which variants in what combination cause them. An example attempt to improve on this was made by (Jaillard et al., 2018), where DeBruijn graphs were used to visualize the series of k-mer differences between genomes that led to a specific phenotype in the form of decision trees.

1.4 PHENOTYPESEEKER

PhenotypeSeeker (Aun et al, 2018) is a program for the prediction of phenotypes based on genotypes. It produces a machine learning model that can take genomes as input and provides a probability of observing a specific phenotype. So far, it has been utilised for the prediction of AMR in bacteria and the prediction of colonization efficiency of root nodules by

Sinorhizobium meliloti (Bellabarba et al., 2021), but the software can be applied to predict any kind of phenotype, binary or continuous.

The program begins by processing input genomes into k-mers, filtered by minimum count and appearance in genomes. A presence/absence matrix in the genomes is created, and the k-mers are statistically associated with the phenotypes, using a chi-squared test when analysing binary phenotypes. Up to a 1000 significant k-mers are provided for building the ML model with a classifier of choice; a smaller amount is given if the number of significant k-mers above a set threshold is lacking. For binary phenotypes, classifiers such as Random Forest, Logistic Regression, Decision Trees, Naive Bayes, and XGBoost can be used. In this study, Random Forest (RF) and Logistic regression are used for the practical part. However, the program doesn't go any further to determine and annotate the genomic locations where these significant k-mers originate from. Therefore, an extension software was developed by the author as the practical goal of this thesis. The resulting software program, GenePointer, utilises the output of significant k-mers provided by PhenotypeSeeker to pinpoint the exact locations on a reference genome that these k-mers originate from. GenePointer shows the location of significant k-mers in the genome and, by doing that, potentially reveals novel associations to genes, intergenic regions, or even mobile elements while also providing the specific sequences, per genome, from these elements.

2. AIMS

In order to take the functionality of PhenotypeSeeker one step beyond just predicting phenotypes and provide deeper analysis of the resistance associated k-mers, potentially any phenotype associated k-mers, an extension was designed by the author. GenePointer is a pipeline built on top of PhenotypeSeeker to provide annotations, where possible, to all the significant k-mers that are currently provided as just an intermediate output from PhenotypeSeeker. It simplifies the transition into detailed analysis of known and possibly unknown ARGs with potential in the analysis of phenotypes in general and provides a future means of removing unassociated elements for the building of more accurate phenotype predictors.

- The first aim of developing GenePointer is to determine whether direct mapping and alignment for placing statistically associated k-mers to the reference genome are sufficient to identify the resistance conferring genes.

- The second aim of GenePointer was to go even further and to determine if the program can provide insights into the genomic changes that are present in resistant isolates and identify novel associations with other genes and genetic elements.

- Finally, GenePointer aims to summarise the extended k-mer alignments obtained to provide a broad sense understanding of the population of isolates whose genomes were used as input and to help in further research of AMR mechanisms.

3. PRACTICAL PART

3.1 GENEPOINTER

The author created GenePointer as an extension program to PhenotypeSeeker to determine whether the statistical analysis of genomes and their k-mers as done by PhenotypeSeeker and other programs alike is suitable for identifying causal elements in resistant genomes. The program processes the significant k-mers found by PS and provides insight into the kind of elements that these significant k-mers originate from and provides appropriate annotations when available on the reference genome of a species.

The program provides 2 options for placing k-mers on the reference genome, the first method obtains the locations of the k-mers from the reference genome index and looks for available annotations found in the general feature format file (.GFF). The reference genome index is made with GenomeTester4 and it gives the locations of every k-mer and the general feature format file contains all annotations currently available on the reference genome. While this approach is simpler and maps some of the significant k-mers back to the reference genome it doesn't find a location for all of them. In addition, there is a chance that the k-mers don't originate from the identified locations if they contain mutations. To try to improve on the efficiency of k-mer mapping and confidence in their locations the second approach was implemented.

The second approach attempts to align the k-mers back to the reference genome, but for this the 13 bp k-mers as used in this analysis aren't sufficient to provide confident alignments, hence the k-mers are extended prior to alignment. To extend them, each k-mer is mapped back to their original genomes, in a similar manner as described before, and flanking pieces of DNA are added to the k-mers to contain more of the surrounding region information and these extended sequences are aligned back to the reference genome to obtain another set of locations. These locations are searched for in the GFF file and any available annotations are extracted. A flank length of 50 bp was chosen to make 113 bp sequences from each k-mer because bowtie2 specialises on aligning shorter sequences.

The results are presented as a CSV file containing information on the k-mer, aligned sequence, location in the reference genome and the annotation found, considering any plasmids

or contigs and k-mers that map to multiple locations. The GFF annotation file can provide information about genes, pseudogenes, repeat-regions, mobile genetic elements and more. Upon success the genes of interest are identified and highlighted in Tables 1-4.

3.2 METHODS AND DATA

3.2.1 Validation

To test whether GenePointer can identify these genetic elements of interest the genomes of 3 species, *Mycobacterium tuberculosis*, *Enterococcus faecium* and *Escherichia coli* with resistances to isoniazid, vancomycin and ciprofloxacin respectively were input into GenePointer with 1000 input genomes. These organisms are well studied and the main resistance associated genes are well known and make it easier to verify if GenePointer was successful. The resulting data, in the form of potential genes with significance values, are validated or rejected using existing literature.

3.2.2 Data

The genomes obtained for validation of the program's functionality are downloaded from PATRIC in FASTA format because the database provides a large collection of phenotyped genomes for *Mycobacterium tuberculosis, Enterococcus faecium* and *Escherichia coli* and the quality of genome sequencing is not considered in this analysis. The reference genome and general feature format (.gff) file for each species were downloaded from the NCBI database [https://www.ncbi.nlm.nih.gov/datasets/genome/].

Because some genome files downloaded from PATRIC can be empty, a shell script was used to remove the empty files from the download folder and their corresponding lines from the data.pheno file.

Analysis on all three species was done using 1000 and 200 genomes evenly split into 500 and 100 of both phenotype isolates, respectively. It has been shown that an uneven distribution of phenotypes, including an overly high proportion of resistant isolates in the training dataset can lead to worse specificity and sensitivity of prediction models (The

CRyPTIC Consortium, 2018; Nguyen et al., 2020) and based on this the analysis put an emphasis on using an even number of both phenotypes. The usage of input data sizes of 1000 and 200 was motivated by the results obtained during testing of the program where smaller amounts of genomes in some cases identified more of the correct genes.

3.2.3 Parameters

Data was downloaded without intermediate phenotype isolates with the intermediate_phenos parameter set to False. The download step requires a metadata file provided by PATRIC with all the information needed to find the right genomes from their database. A separate script is provided to download it from PATRIC.

Data analysis steps were done with k-mers of length 13 as it was confirmed the best performing when considering computational speed and model accuracy by Aun et al. 2018. The machine learning models used are logistic regression and Random Forest. When testing the analysis process, the two models produced different results and to determine which model works better for identifying elements conferring resistance both were used and compared.

3.2.4 Implementation

GenePointer is implemented in Python on a Linux machine. Because PhenotypeSeeker is only Python 3.8 compatible, GenePointer including the PhenotypeSeeker processing step and all the modules used, work with Python 3.8. The most important modules include pandas for data frames, NumPy for matrices, BioPython for light sequence analysis and patric-tools(modified) for downloading genomes from PATRIC (https://www.patricbrc.org/). PhenotypeSeeker and GenePointer are developed for Linux. GenePointer uses the binaries of: Bowtie2 and GenomeTester4's programs glismaker and glistquery. Bowtie2 is an alignment program used to align the extended k-mers to the reference genome. GenomeTester4 is the program used to index genomes and make lists of k-mers from input genomes.

Analysis was done on a Linux server with more than 300GB of free space to fit the data created when processing 1000 genomes.

3.3 RESULTS

3.3.1 Pipeline and Performance

GenePointer contains 4 separate Python scripts, one for downloading data, one for finding genes without alignment, one for finding genes with alignment and finally a script to summarise the data obtained. This provides some separation of tasks and each script can be changed to match one's needs. The scripts themselves use functions that further separate the analysis process into pieces that can be useful when repeating segments of the analysis and when updating code.

The download script is used to download genomes from PATRIC by inputting the species, antibiotic, and number of genomes to be downloaded, *Figure* 1(A). After downloading the genomes, the program generates the required input file for PS (data_*.pheno) which contains the names of the genomes, their address and phenotype label (1 or 0 for yes or no). The program downloads an even number of both phenotype isolates(assuming an even number is available for download) and if possible, creates a selection of data.pheno files with varying amounts of genomes (10,30,100,200,350,500,1000, and larger, ex. data_100.pheno) so that a subset of genomes can be used when working on smaller computers and for testing the effect of input genome numbers on gene detection. The download script also provides the option for downloading “intermediate” phenotyped genomes, of which PS is capable of processing, but for the initial validation of GenePointer strictly binary values were used.

The first data analysis script, finding genes without alignment, takes the data.pheno file and the parameters: k-mer length, ML classifier for the binary phenotype, species and antibiotic and runs PS, *Figure* 1(B). This step can take time, depending on the amount of input genomes, numbers above 1000 will most likely take 3 hours or more because statistics on the millions of k-mers produced takes time. This step also requires enough free memory for storing the k-mer lists generated by GenomeTester4, for 1000 genomes this was more than 100GB. After PS finishes it generates a file containing the k-mers most associated with the phenotype, their coefficients in the ML model and the genomes that they exist in. This is the file that GenePointer uses as its input. The next steps include indexing the reference genome if not already done and reducing the size of the .GFF file for easier parsing in later steps. The k-mers are then searched in the reference genome index only considering 0 mismatched k-mers because the goal is to find exact matches in the reference genome and a mismatch is the same as a completely different

k-mer at this size of k. The k-mers that were found are separated into a file along with their locations and the locations are finally used to determine if any annotations exist for them. The results are in a csv file with the k-mers, their location, identified element type and any available description of the element usually containing the id, name and some notes from literature. If the k-mers truly are part of the locations that confer resistance then the genes that have already been determined causal should be present in the results file. Exceptions are horizontally acquired genes because the program in its current form only looks for genes in the same species' reference genome and blasting would be required to determine any foreign genes. Additionally, this approach makes the strong assumption that these associated k-mers aren't mutated with respect to the reference genome and to take all k-mers into account a deeper k-mer mapping step was implemented.

The alignment-based k-mer mapping step takes the same significant k-mers file output by PS but first extends them from their original genomes. For each genome that the k-mer is present in, the k-mer locations are found and the desired length of DNA, in this study 50 nucleotides from either side, is taken from around the locations so that the k-mer is in the centre of that sequence. These extended pieces of DNA are recorded in a k-mers * genomes (columns * rows) size matrix (.csv file) that enables comparative analysis of DNA. All the extended pieces of DNA are combined into a single custom FASTA file that gets input into Bowtie2 for fast alignment of all the sequences. Bowtie2 is run with the --very-sensitive-local flag to make sure that the extended k-mers align to the reference genome as accurately as possible while allowing for mutations. Bowtie2 outputs a SAM file that is processed for only the desired columns and an exhaustive results file is created for sequences that aligned and ones that didn't. The file contains the original k-mer, the aligned sequence, its location on the reference and the genes corresponding to this location.

The results files from the previous 2 steps can be quite large and so a separate script is used to summarise the results. The final step makes a list of all the unique genes identified and shows how many unique k-mers from PS were in them with the significance values of the k-mers summed to give that gene a significance value. If the associated k-mers with the highest coefficients in the ML model are part of a gene then that gene also has the highest probability of influencing the phenotype.

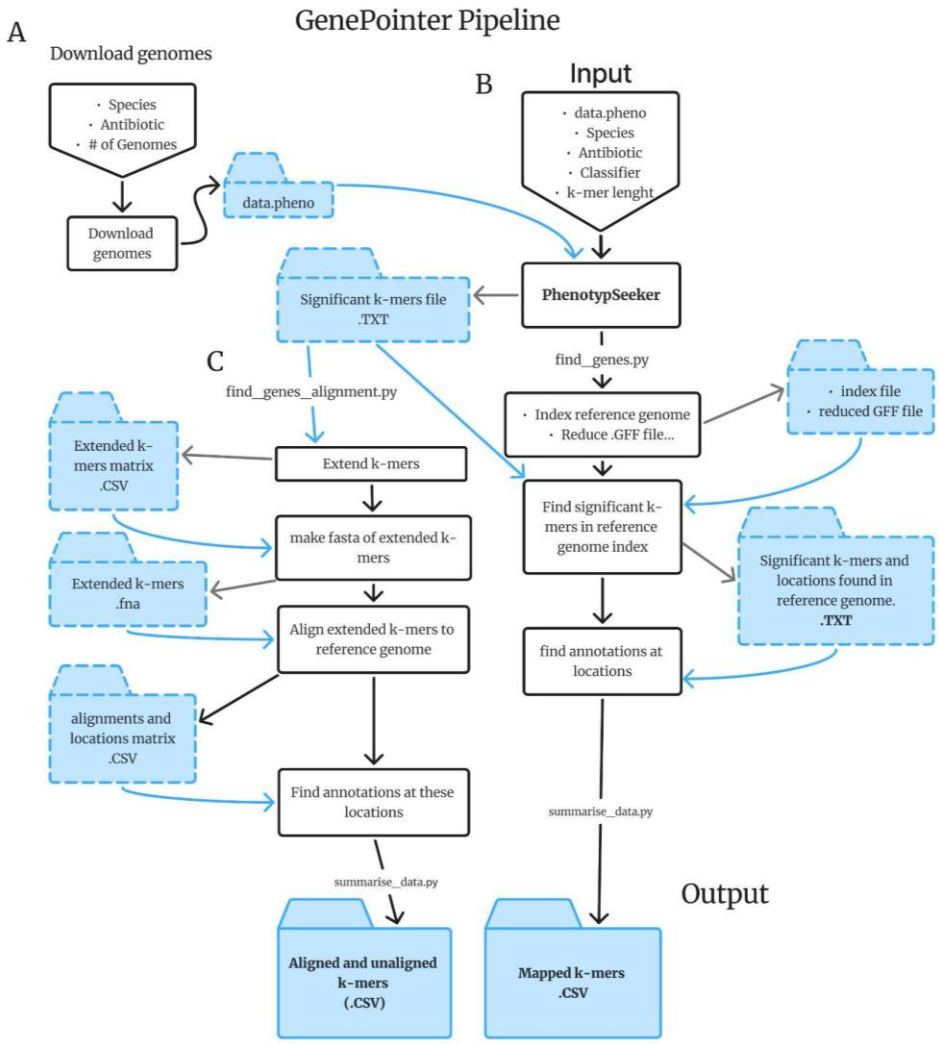

Figure 1. Diagram of GenePointer processes. Pipeline begins with the download script (A). Direct mapping and alignment of k-mers both start with PhenotypeSeeker (B). Alignment of k-mers splits off the mapping pipeline after significant k-mers are found (C). Input and output files are shown in blue, dashed lines indicate intermediate files and solid lines indicate final outputs of the pipeline. Black bordered boxes are parts of the pipeline, like functions or steps of the analysis.

GenePointer completed the 1000 genome analysis direct map method, including the time PhenotypeSeeker takes, on average in about 4 hours. The slowest step in GenePointer turned out to be looking for k-mers in the reference genome index file and 1000 k-mers took about 1 hour to map to the reference genome. Alignment based k-mer analysis was slower due

to k-mer expansion. Both methods depend on the number of k-mers output by PhenotypeSeeker and in alignment-based analysis, speed is dependent on the number of k-mers and the number of genomes. For this reason, the number *#k-mers * #genomes*, which is the number of expanded sequences produced, was limited to 300 000 by reducing the number of genomes used in the expansion step. 1000 k-mers and 300 genomes took 4 hours to expand and the time increases non-linearly with increasing amounts of genomes and k-mers. As the results using logistic regression and RF are almost identical in most cases, only logistic regression results are presented with only mentions of RF analysis results if significant.

3.3.2 *Mycobacterium Tuberculosis*' Resistance to Isoniazid

The gene of interest in isoniazid resistant *Mycobacterium* isolates is katG. When running GenePointer with 1000 genomes and both logistic regression and RF as classifiers, *katG* was not identified. *katG* was identified only in the 200-genome analysis. Tables shown are partial and show only a few top rows from tables containing the most significant k-mers. The most significant genes and intergenic regions identified are shown in *Table* 1 and 2. Alignment of k-mers also produces k-mers that didn't align to the reference genome and the most significant of these k-mers are provided on the bottom of each alignment analysis table. Differences between tables from alignment-based k-mer placement and direct mapping come from the analysis structure. GenePointer doesn't associate k-mers with their original genomes when performing direct mapping and so gene prevalence in genomes is not shown in direct mapping results. In some cases, GenePointer provided the RefSeq accession number of the protein sequence instead of the name of the gene and these cells in Tables 1-6, have been changed for the name of the gene using UniProt. Duplicate rows caused by redundancy in the GFF file are also removed.

Table 1. Analysis results of 1000 and 200 *Mycobacterium tuberculosis* isolates with ciprofloxacin resistance/susceptibility using direct mapping and logistic regression. Bottom rows contain most significant k-mers that didn't align.

Mycobacterium tuberculosis analysis with Logistic Regression. **Aligned** k-mers					
Gene (inter_Start…End)	**Minimum p-value of k-mers**	**# of Unique k-mers**	**Most significant k-mer frequency (Total/Res/Sus)**	**Significant k-mer location**	**Gene prevalence in genomes (Total/Res/Sus)**
1000 genomes analysis					
embB	1.00E-59	1	825/325/500	4247371	810/313/497
inter_542852...543174	1.00E-36	1	924/486/438	542876	2/0/2
PPE26	1.00E-36	1	924/486/438	2027251	909/476/433
PE18	1.00E-36	1	924/486/438	2026593	907/476/431
200 genomes analysis					
katG	8.86E-19	2	123/24/99	2155107	123/24/99
ugpC	1.28E-11	23	69/12/57	3138380	1/0/1
trpE	1.28E-11	23	69/12/57	1808569	1/0/1
Most significant k-mers that failed to align					
K-mer	**P-value**	**Prevalence in genomes**	**Extended sequences sample**		
GCCATGCCCAGGA	1.00E-59	825/325/500	CTTCCTGCTCTGGCATTTCATCGGCGCGAATTCGTCGGACG ACGGCTACATCCTGGGCATGGCCCGAGTCGCCGACCACGC CGGCTACATGTCCAACTATTTCCGCTGGTTCG		

Table 2. Analysis results of 1000 and 200 *Mycobacterium tuberculosis* isolates with isoniazid resistance/susceptibility using direct mapping and logistic regression. The "…" means more items exist in the cell. Intergenic regions, where found, are presented on the bottom of the table.

Mycobacterium tuberculosis analysis with Logistic Regression. **Mapped** k-mers					
Genes	**Minimum p-value of k-mer**	**# of Unique k-mers**	**Significant k-mer prevalence (Total/Res/Sus)**	**Unique k-mers**	**Locations on reference genome**
1000 genome analysis					
PPE26	1.00E-36	1	924/486/438	CGCAGATTGCCAA	2027300
PE18	1.00E-36	1	924/486/438	CGGCACACGCCCA	2026642
Rv1762c	1.00E-35	1	466/172/294	CGCCGTCTCGCTC	1995202
PE_PGRS6	1.00E-26	1	805/341/464	GAGCGGGGCGGCC	623397
cut1	1.00E-23	1	448/170/278	CGCTATATGCAGA	1989420
200 genome analysis					
katG	8.86E-19	1	123/24/99	ATGCCGCTGGTGA	2155161
Rv2655c	5.96E-10	6	179/97/82	CGACCGCACAGTA	2978132 ....
relA	1.80E-09	1	184/99/85	CGGCCGAGGTCAC	2907830
wag22	5.99E-09	1	72/15/57	ATCCCCGAATCCG	1991689
Intergenic regions					
Inter_Start_End		**Significant k-mer p-value**	**Significant k-mer prevalence (Total/Res/Sus)**	**K-mer**	**Locations**
1000 genome analysis					
-		-	-	-	-
200 genome analysis					
inter_854157...854267		5.03E-10	21/4/17	CGGCCCCCCCCA	854249

The most significant k-mer found was in the *embB* gene which is known to be the source of ethambutol resistance in *Mycobacterium* when mutated (Sreevatsan et al., 1997). In many cases isoniazid and ethambutol resistances have been identified together, meaning where there is ethambutol resistance there is often also isoniazid resistance (Gupta et al., 2006). The next few k-mers in *Table* 1, 1000 genome section, have the same p-value and they aren't considered significant enough for conclusions. The other genes like PPE26, PE18 have roles in immune system responses and inflammation respectively.

Both alignment and direct mapping methods in the 200-genome analysis, *Table* 1 and *Table* 2, revealed *katG*, responsible for isoniazid resistance but because the input data size is much lower so is the resulting p-value of the k-mers. As seen from the frequency of the most significant k-mer it is the absence of this k-mer that can be associated with the resistance phenotype, likely due to a substitution mutation (Narmandakh et al., 2020).

K-mers that didn't align to the reference genome are presented on the bottom rows of *Table* 1. Only one row is presented because the other rows either had the same k-mer with minor variations of the same extended sequence, caused by variations between input genomes, or k-mers with lower p-values that are likely insignificant to the resistance mechanism. GCCATGCCCAGGA in *Table* 1 has the same p-value as the most significant k-mer from the *embB* gene suggesting that it is as strongly associated with the resistance mechanism. The fact that it didn't align to the reference genome may be because it is part of an unannotated plasmid or larger sequence mutation that leads to the resistance towards isoniazid.

3.3.3 Enterococcus Faecium's Resistance to Vancomycin

The genes of interest in Enterococcus's resistance mechanism towards vancomycin can be found in multiple gene clusters. The known gene clusters include VanA, VanB, VanD, VanM (Stogios & Savchenko, 2020). VanA specifically contains the genes: *vanR, -S, -H, -A, -X, -Y, -Z.* The analysis of Enterococcus' resistance towards vancomycin provided a lot more positive hits in all methods and dataset sizes.

Table 3. Analysis results of 1000 and 200 *Enterococcus faecium* isolates with ciprofloxacin resistance/susceptibility by alignment and logistic regression. Bottom rows contain most significant k-mers that didn't align.

Enterococcus faecium analysis with Logistic Regression. **Aligned** k-mers					
Gene (inter_Start…End)	**Minimum p-value of k-mers**	**# of Unique k-mers**	**Most significant k-mer frequency (Total/Res/Sus)**	**Significant k-mer location**	**Gene prevalence in genomes (Total/Res/Sus)**
1000 genomes analysis					
inter_11278...11493	1.00E-118	1	639/484/155	11291	90/90/0
VanY	1.00E-113	1	432/416/16	15404	85/84/1
vanZ-A	1.00E-100	1	438/407/31	16627	85/85/0
trxA	1.00E-14	1	958/499/459	2115952	198/100/98
200 genome analysis					
WP_010729418.1	1.01E-35	1	101/94/7	1872436	14/8/2006
vanA	1.01E-34	3	111/97/14	12768	96/94/2
arcD	1.01E-34	1	111/97/14	518422	36/22/14
vanR-A	1.01E-32	4	110/94/16	9687	97/94/3
Most significant k-mers that failed to align					
K-mer	**P-value**	**Prevalence in genomes**	**Extended sequences sample**		
CAATTTAATATTA	1.00E-118	639/484/155	ATTTTTTAGGAAAATCTCAAGGTTATCTTTACTTTTTCTT AGGAAATTAACAATTTAATATTAAGAAACGGCTCGTTC TTACACGGTAGACTTAATACCGTAAGAACGAGCCG		
TACTATCAAGCAA	1.00E-113	432/416/16	CAGATATCGTGAATTTATCTAAACATGACGAATTAATA AATGGATACGGGTTGCTTGATAGTAATATTTATATGTCA AAAGAAATAGCACAAAAATTTTCAGAGATGGTCAAT		

Table 4. Analysis results from 1000 and 200 *Enterococcus faecium* isolates with vancomycin resistance/susceptibility using direct mapping and logistic regression. The "…" means more items exist in the cell. Intergenic regions, where found, are presented on the bottom of the table.

Enterococcus faecium analysis with Logistic Regression. **Mapped** k-mers					
Genes	**Minimum p-value of k-mer**	**# of Unique k-mers**	**Significant k-mer prevalence (Total/Res/Sus)**	**Unique k-mers**	**Locations on reference genome**
1000 genome analysis					
B6S05_RS00050	1.00E-118	1	639/484/155	CAATTTAATATTA	11340
addA	1.00E-14	1	972/499/473	CGATTTTTGATCA	1237921
B6S05_RS14200	1.00E-14	1	572/343/229	ATGAATGCAAGTG	2756505
200 genome analysis					
rplI	1.01E-38	1	97/94/3	GCTAACATTAATA	12537
ssb	1.01E-37	1	100/94/6	ACTTATTGTGGAT	9464
B6S05_RS00050	1.01E-36	1	102/94/8	GATGTGAGCAGGA	11520
brxL	1.01E-19	2	103/80/23	AATTCAGATGCTG ...	583278 ...
Intergenic regions					
Inter_Start_End		**Significant k-mer p-value**	**Significant k-mer prevalence (Total/Res/Sus)**	**K-mer**	**Locations**
1000 genome analysis					
inter_16325...17112		1.00E-100	438/407/31	ACCCATTTAAGAA	16676
200 genome analysis					
inter_12612...12886		1.01E-35	101/94/7	AGCTTTGCATGGC	12740
inter_17975...18104		1.01E-21	91/75/16	AGATTATATATAT	18029

Both 1000 and 200 isolate analyses found the key genes associated with vancomycin resistance, *vanY, vanZ, vanH, vanR* that are all part of the VanA gene cluster where only *vanY* is not directly involved in the vancomycin resistance mechanism (Stogios & Savchenko, 2020; Arthur et al., 1992). The direct map method couldn't identify any known associated genes but the most significant B6S05_RS00050 expresses a Cyclic-di-AMP phosphodiesterase. Direct map results didn't identify any of the van genes, likely because the k-mers associated with them all contain mutations of some kind and are thus unmappable to the reference genome. The genes *rplL* is the 50S ribosomal protein L7/L12 (https://www.ncbi.nlm.nih.gov/gene/60894703) and

ssb is a single stranded DNA binding protein (https://www.ncbi.nlm.nih.gov/gene/60892571) and both don't have a clear role in the resistance mechanism.

Two of the most significant k-mers are intergenic with the most significant of them being just before the *vanH-A* gene. The 2 sequences that didn't align to the reference also have low p-values and may reveal more Van genes. BLAST-ing is not considered in this paper but it is a future improvement for GenePointer.

3.3.4 Escherichia Coli Resistance to Ciprofloxacin

The genes of focus in Escherichia coli's resistance mechanism are *gyrA, gyrB, parE* and *parC*. Ciprofloxacin resistance is often identified as combinations of mutations in these 4 genes and if the analysis is successful k-mers from these mutated regions should be identified. Almost all significant k-mers, however, have been identified from within intergenic regions with only one exception for genes found.

Table 5. Analysis results of 1000 and 200 *Escherichia coli* isolates with ciprofloxacin resistance/susceptibility using direct mapping and logistic regression. Bottom rows contain the most significant k-mers that didn't align.

Escherichia coli analysis with Logistic Regression. **Aligned** k-mers					
Gene (inter_Start…End)	**Minimum p-value of k-mers**	**# of Unique k-mers**	**Most significant k-mer frequency (Total/Res/Sus)**	**Significant k-mer location**	**Gene prevalence in genomes (Total/Res/Sus)**
1000 genomes analysis					
inter_1255952...1256721	1.00E-111	1	573/463/110	1256582	187/150/37
inter_3111325...3112054	1.00E-102	1	548/443/105	3111967	58/52/6
ytfT	1.00E-14	1	608/351/257	4453039	6/0/6
1000 genomes analysis with RF					
inter_1255952...1256721	1.00E-111	1	573/463/110	1256582	187/150/37
inter_3111325...3112054	1.00E-102	1	548/443/105	3111967	58/52/6
ydcS	1.00E-101	1	390/354/36	1511996	116/106/10
200 genome analysis					
yghT	1.00E-18	1	126/91/35	3134422	125/91/34
ydcU	1.00E-17	1	122/34/88	1513952	48/27/21
Most significant k-mers that failed to align					
K-mer	**P-value**	**Prevalence in genomes (Total/Res/Sus)**	**Extended sequences sample**		
GAAATGAAAAATC	1.00E-111	573/463/110	GGCAATAAGCACAAAAGTGTAGGATGTTACAAGAATGATTAGGACTCGGTGAAATGAAAAATCCACGCAATTGCGTGGATTATATATACTTTTGCGCTCTTCATGAGATTTAG		
TTCAACGGGGCAA	1.00E-102	548/443/105	GCATATGCTGCTGGTGAATCACAGCTACGGTTGCCAGTTGCTGGTCGATTTTCAACGGGGCAAAATGCAGTTCATCGATATCCTGGTTCAGCTCTTCTTCCAGCAAGGCGGGC		

Table 6. Analysis results of 1000 and 200 *Escherichia coli* isolates with ciprofloxacin resistance/susceptibility using direct mapping using logistic regression. Intergenic regions are presented on the bottom section of the table.

Escherichia coli analysis with Logistic Regression. **Mapped** k-mers					
Genes	**Minimum p-value of k-mer**	**# of Unique k-mers**	**Significant k-mer prevalence (Total/Res/Sus)**	**Unique k-mers**	**Locations on reference genome**
1000 genome analysis					
ilvA	1.00E-14	1	324/107/217	GGATTTATTCGAA	3956154
ccmF	1.00E-14	1	290/79/211	CCCACGTACCCCA	2294356
yqiI	1.00E-13	1	137/109/28	CTAAAAATGCCAA	3191133
200 genome analysis					
ybeU	1.00E-12	1	114/29/85	AGCGGGAAGCGAT	680118
yfiF	1.00E-11	1	85/69/16	ACGGCTCTGGAAA	2718159
paaE	1.00E-11	1	108/27/81	CGGACGTTTCTCC	1456659
Intergenic regions					
Inter_Start_End		**Significant k-mer p-value**	**Significant k-mer prevalence (Total/Res/Sus)**	**K-mer**	**Locations**
1000 genome analysis					
inter_1255952...1256721		1.00E-111	573/463/110	GAAATGAAAAATC	1256631
200 genome analysis					
-		-	-	-	-

The performed analysis and annotation didn't reveal any of the known resistance associated genes and the main significant k-mers originate from intergenic regions. The RF model associated the resistance mechanism to the *ydcS* gene. The gene is involved in natural transformation and dsDNA transport (Sun D, 2016) but any connections to the ciprofloxacin resistance mechanism remain unclear and require deeper analysis and more statistics to determine. The table of un-aligned k-mers contains 2 significant elements suggesting that the resistance conferring genes may be too different to their original versions on the reference genome. The spread of the resistance strength onto mutations on multiple genes may also be the cause of the poor statistical performance.

3.4 DISCUSSION

GenePointer was designed by the author because there is potential in resistance marker identification from PhenotypeSeeker's output data and because PS can technically work on any kind of phenotype, GP has great potential in phenotype marker identification in general. The benefits of annotations to the k-mers output by PS help in focusing laboratory work, supplementing the predictions of PS with resistance element descriptions and their use in improving the machine learning process by adaptively removing unwanted elements from being used as features for predicting resistance mechanisms. GenePointer successfully provides annotations to k-mers that are alignable to the reference genome and creates separate tables for k-mers that need further analysis to determine their source. While the antimicrobial resistance element identification efficiency of GenePointer is lacking, it can be used as a starting point to develop a more accurate analysis program, utilising more of the available tools and databases, aimed at mapping all the k-mers to the right locations and finding all the known resistance associated genes.

As per the first aim of GenePointer, testing alignment and direct mapping to the reference genome showed that alignment of k-mers always found more significant genes, intergenic regions and for *M. tuberculosis* and *E. faecium*, identified some of the known genes associated with the resistance mechanisms analysed, making direct mapping almost redundant. This means GenePointer can be simplified to use only the alignment-based pipeline and the only added cost of using alignment is the computation time, which for smaller numbers of genomes is not significantly more.

Identifying novel associations in the isolate genomes proved to be hard as the statistical significance of unknown resistance associated elements is often too low and so they get lost in the noise of all the other insignificant elements identified. Detecting these elements needs more sophisticated k-mer filtering to remove noise and make sure that the k-mers provided by PhenotypSeeker have the highest possible probability of originating from resistance conferring elements. This is something that GenePointer still aims to improve upon with future versions.

The testing methodology of GenePointer should be expanded to encompass more kinds of bacterial species and different combinations of parameters manipulating different parts of the analysis. PhenotypeSeeker and its available parameters were not explored in depth for this analysis either. PS has options for population structure correction and can use other machine

learning models. One of the potential sources of error to the analysis results can be due to the former. Bacteria have complex population structure and phylogenetic trees which can cause false associations when testing k-mers for statistical associations to the phenotype. The experimental design should be supplemented with larger sets of input data, population structure correction and testing on other kinds of machine learning model to determine the right conditions for resistance marker identification using GenePointer.

Future plans to improve GenePointer include implementing intergenic region annotation by trying to predict whether an intergenic region k-mer corresponds to a promoter, terminator or another regulatory region. The program also needs a script to run BLAST on all the k-mers that didn't align to the reference genome to determine their potential loci.

The development process showed that the analysis and pinpointing of genetic elements is not straightforward and many approaches can be applied to achieve the same goal of gene identification. This raised questions about what could be improved and added to create the ideal resistance marker identification pipeline. GenePointer can be explored deeper in places like k-mer extension which for now has only been tested with a flanking DNA length of 50 nucleotides either side of the k-mer. It could be that longer or shorter sequences provide better results in alignment. The extended k-mers are also a good source for information on the mutations present in the input genomes for analysing the molecular mechanisms of the phenotype under study. GenePointer will be extended to provide a file where the extended k-mer sequences can be directly compared to the sequences in the matching locus on the reference genome. Alignment of these extended sequences is currently done by Bowtie2 and the parameters used for alignment are likely not fine-tuned for matching mutated pieces of DNA to a reference genome.

CONCLUSION

The prior was an overview of antimicrobial resistance and software used to analyse bacterial genomes followed by the analysis of the authors designed software tool GenePointer to determine if the programs approach to identifying resistance markers in bacterial genomes is effective and what kind of potential it has in finding novel associations.

The rising prevalence of AMR towards antibiotics is threatening to make healthcare ineffective, expensive and inaccessible. Fortunately, the approaches for determining AMR in bacterial isolates are becoming faster, cheaper and more accurate. Bioinformatics has huge potential in the analysis of bacterial genomes and there are a large variety of programs already available. PhenotypeSeeker is an effective program for prediction of phenotypes from genotypes. The author's implementation of GenePointer as an extension to PhenotypeSeeker is a potentially powerful tool for the identification of resistance markers in bacterial genomes and potentially any phenotype markers in genomes.

GenePointer successfully showed that it can highlight the genes associated with a resistance mechanism but currently it lacks statistically power in finding the more significant locations in the large amount of noise that genome analysis generates. The program as it is now didn't manage to clearly identify elements that could be considered novel associations and to improve this the noise needs to be reduced and deeper analysis of the genes with unclear association to resistance must be performed. The results given by GenePointer provided an interpretation to almost all the k-mers provided by PhenotypeSeeker, whether they were aligned or un-aligned to the reference or found in intergenic regions. The k-mers that didn't align still need a functionality in GP that would automatically run them through BLAST to find annotations to them.

Overall GenePointer is a simple to use program, capable of analysing multiple genomes at once and can identify the most significant resistance markers often. In its current form, GenePointer can be used as a starting point for much more sophisticated software solutions to aid in the research of antimicrobial resistance mechanisms but also shows potential in phenotype marker discovery in general which could aid in the research of any phenotype.

REFERENCES

Andrews, J. M. (2001). Determination of minimum inhibitory concentrations. *Journal of Antimicrobial Chemotherapy*, *48*(SUPPL. 1). https://doi.org/10.1093/jac/48.suppl_1.5

Nguyen, M., Olson, R., Shukla, M., VanOeffelen, M., & Davis, J. J. (2020). Predicting antimicrobial resistance using conserved genes. PLoS Computational Biology, 16(10). https://doi.org/10.1371/journal.pcbi.1008319

C Reygaert, W. (2018). An overview of the antimicrobial resistance mechanisms of bacteria. AIMS Microbiology, 4(3), 482–501. https://doi.org/10.3934/microbiol.2018.3.482

Mancuso, G., Midiri, A., Gerace, E., & Biondo, C. (2021). Bacterial antibiotic resistance: the most critical pathogens. In *Pathogens* (Vol. 10, Issue 10). https://doi.org/10.3390/pathogens10101310

Anderson, M., Panteli, D., van Kessel, R., Ljungqvist, G., Colombo, F., & Mossialos, E. (2023). Challenges and opportunities for incentivising antibiotic research and development in Europe. In *The Lancet Regional Health - Europe* (Vol. 33). https://doi.org/10.1016/j.lanepe.2023.100705

Fleming, A. (2001). On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. 1929. *Bulletin of the World Health Organization*, *79*(8). https://doi.org/10.1093/clinids/2.1.129

Hutchings, M., Truman, A., & Wilkinson, B. (2019). Antibiotics: past, present and future. In *Current Opinion in Microbiology* (Vol. 51). https://doi.org/10.1016/j.mib.2019.10.008

Davies, J. (1996). Origins and evolution of antibiotic resistance. In *Microbiología (Madrid, Spain)* (Vol. 12, Issue 1). https://doi.org/10.1128/mmbr.00016-10

Prescott, J. F. (2014). The resistance tsunami, antimicrobial stewardship, and the golden age of microbiology. *Veterinary Microbiology*, *171*(3–4). https://doi.org/10.1016/j.vetmic.2014.02.035

Salam, M. A., Al-Amin, M. Y., Salam, M. T., Pawar, J. S., Akhter, N., Rabaan, A. A., & Alqumber, M. A. A. (2023). Antimicrobial Resistance: A Growing Serious Threat for Global Public Health. In *Healthcare (Switzerland)* (Vol. 11, Issue 13). https://doi.org/10.3390/healthcare11131946

Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S. C., Browne, A. J., Chipeta, M. G., Fell, F., Hackett, S., Haines-Woodhouse, G., Kashef Hamadani, B. H., Kumaran, E. A. P., McManigal, B., … Naghavi, M. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, *399*(10325). https://doi.org/10.1016/S0140-6736(21)02724-0

Polsfuss, S., Hofmann-Thiel, S., Merker, M., Krieger, D., Niemann, S., Rüssmann, H., Schönfeld, N., Hoffmann, H., & Kranzer, K. (2019). Emergence of Low-level Delamanid and Bedaquiline Resistance during Extremely Drug-resistant Tuberculosis Treatment. *Clinical Infectious Diseases*, *69*(7). https://doi.org/10.1093/cid/ciz074

Coello, R., Glynn, J. R., Gaspar, C., Picazo, J. J., & Fereres, J. (1997). Risk factors for developing clinical infection with methicillin-resistant Staphylococcus aureus (MRSA) amongst hospital patients initially only colonized with MRSA. *Journal of Hospital Infection*, *37*(1). https://doi.org/10.1016/S0195-6701(97)90071-2

World Health Organisation. (2024, May). *Tuberculosis: Multidrug-resistant (MDR-TB) or rifampicin-resistant TB (RR-TB)*. https://www.who.int/news-room/questions-and-answers/item/tuberculosis-multidrug-resistant-tuberculosis-(mdr-tb)

Florensa, A. F., Kaas, R. S., Clausen, P. T. L. C., Aytan-Aktug, D., & Aarestrup, F. M. (2022). ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microbial Genomics*, *8*(1). https://doi.org/10.1099/mgen.0.000748

Weinstein, R. A. (2001). Controlling antimicrobial resistance in hospitals: Infection control and use of antibiotics. *Emerging Infectious Diseases*, *7*(2). https://doi.org/10.3201/eid0702.010206

O'Neill, J. (2016). Antimicrobial Resistance : Tackling a crisis for the health and wealth of nations. *Review on Antimicrobial Resistance*, *December*.

Mosquera-Rendón, J., Moreno-Herrera, C. X., Robledo, J., & Hurtado-Páez, U. (2023). Genome-Wide Association Studies (GWAS) Approaches for the Detection of Genetic Variants Associated with Antibiotic Resistance: A Systematic Review. In *Microorganisms* (Vol. 11, Issue 12). https://doi.org/10.3390/microorganisms11122866

Heym, B., Zhang, Y., Poulet, S., Young, D., & Cole, S. T. (1993). Characterization of the *katG* gene encoding a catalase-peroxidase required for the isoniazid susceptibility of *Mycobacterium tuberculosis*. In *Journal of Bacteriology* (Vol. 175, Issue 13). https://doi.org/10.1128/jb.175.13.4255-4259.1993

Ghodousi, A., Tagliani, E., Karunaratne, E., Niemann, S., Perera, J., Köser, C. U., & Cirillo, D. M. (2019). Isoniazid resistance in *Mycobacterium tuberculosis* is a heterogeneous phenotype composed of overlapping mic distributions with different underlying resistance mechanisms. *Antimicrobial Agents and Chemotherapy*, *63*(7). https://doi.org/10.1128/AAC.00092-19

O'Toole, R. F., Leong, K. W. C., Cumming, V., & van Hal, S. J. (2023). Vancomycin-resistant *Enterococcus faecium* and the emergence of new sequence types associated with hospital infection. In *Research in Microbiology* (Vol. 174, Issue 4). https://doi.org/10.1016/j.resmic.2023.104046

Neyestani, Z., Khademi, F., Teimourpour, R., Amani, M., & Arzanlou, M. (2023). Prevalence and mechanisms of ciprofloxacin resistance in *Escherichia coli* isolated from hospitalized patients, healthy carriers, and wastewaters in Iran. *BMC Microbiology*, *23*(1). https://doi.org/10.1186/s12866-023-02940-8

Thai T, Salisbury BH, Zito PM. Ciprofloxacin. [Updated 2023 Aug 28]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK535454/

Loya, H., Kalantzis, G., Cooper, F. *et al.* A scalable variational inference approach for increased mixed-model association power. *Nat Genet* 57, 461–468 (2025). https://doi.org/10.1038/s41588-024-02044-7

Voichek, Y., & Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nature Genetics*, *52*(5). https://doi.org/10.1038/s41588-020-0612-7

Lemay, M. A., de Ronne, M., Bélanger, R., & Belzile, F. (2023). k-mer-based GWAS enhances the discovery of causal variants and candidate genes in soybean. *Plant Genome*, *16*(4). https://doi.org/10.1002/tpg2.20374

CRyPTIC Consortium. (2022). Genome-wide association studies of global *Mycobacterium tuberculosis* resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. *PLoS Biology*, *20*(8). https://doi.org/10.1371/journal.pbio.3001755

He, C., Washburn, J. D., Schleif, N., Hao, Y., Kaeppler, H., Kaeppler, S. M., Zhang, Z., Yang, J., & Liu, S. (2024). Trait association and prediction through integrative k-mer analysis. *Plant Journal*. https://doi.org/10.1111/tpj.17012

Jaillard, M., Lima, L., Tournoud, M., Mahé, P., van Belkum, A., Lacroix, V., & Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genetics*, *14*(11). https://doi.org/10.1371/journal.pgen.1007758

Aun, E., Brauer, A., Kisand, V., Tenson, T., & Remm, M. (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLOS Computational Biology*, *14*(10), e1006434. https://doi.org/10.1371/JOURNAL.PCBI.1006434

Bellabarba, A., Bacci, G., Decorosi, F., Aun, E., Azzarello, E., Remm, M., Giovannetti, L., Viti, C., Mengoni, A., & Pini, F. (2021). Competitiveness for Nodule Colonization in Sinorhizobium meliloti: Combined In Vitro -Tagged Strain Competition and Genome-Wide Association Analysis . *MSystems*, *6*(4). https://doi.org/10.1128/msystems.00550-21

Sreevatsan, S., Stockbauer, K. E., Pan, X., Kreiswirth, B. N., Moghazeh, S. L., Jacobs, W. R., Telenti, A., & Musser, J. M. (1997). Ethambutol resistance in *Mycobacterium tuberculosis*: Critical role of *embB* mutations. *Antimicrobial Agents and Chemotherapy*, *41*(8). https://doi.org/10.1128/aac.41.8.1677

Gupta, P., Jadaun, G. P. S., Das, R., Gupta, U. D., Srivastava, K., Chauhan, A., Sharma, V. D., Chauhan, D. S., & Katoch, V. M. (2006). Simultaneous ethambutol & isoniazid resistance in clinical isolates of *Mycobacterium tuberculosis*. *Indian Journal of Medical Research*, *123*(2).

Narmandakh, E., Tumenbayar, O., Borolzoi, T., Erkhembayar, B., Boldoo, T., Dambaa, N., Burneebaatar, B., Nymadawa, N., Mitarai, S., Jav, S., & Chiang, C. Y. (2020). Genetic mutations associated with isoniazid resistance in *Mycobacterium tuberculosis* in Mongolia. *Antimicrobial Agents and Chemotherapy*, *64*(7). https://doi.org/10.1128/AAC.00537-20

Stogios, P. J., & Savchenko, A. (2020). Molecular mechanisms of vancomycin resistance. In *Protein Science* (Vol. 29, Issue 3). https://doi.org/10.1002/pro.3819

Arthur, M., Molinas, C., & Courvalin, P. (1992). Sequence of the *vanY* gene required for production of a vancomycin-inducible D,D-carboxypeptidase in *Enterococcus faecium* BM4147. *Gene*, *120*(1). https://doi.org/10.1016/0378-1119(92)90017-J

Sun, D. (2016). Two different routes for double-stranded DNA transfer in natural and artificial transformation of *Escherichia coli*. *Biochemical and Biophysical Research Communications*, *471*(1). https://doi.org/10.1016/j.bbrc.2016.01.137

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Aleksander Roosimaa ,
(author´s name)

1. grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis

   GenePointer - A Software for Automated Identification of Resistance Markers in Bacterial Genomes ,
   (title of thesis)

   supervised by Maido Remm and Erki Aun ;
   (supervisor´s name)

2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;

3. am aware of the fact that the author retains the rights specified in points 1 and 2;

4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Aleksander Roosimaa

19/05/2025