# MACHINE LEARNING METHODS FOR THE PREDICTION OF ANTIMICROBIAL RESISTANCE AND IDENTIFICATION OF NOVEL MARKERS OF RESISTANCE IN *ESCHERICHIA COLI*

**JANICE MOAT**
**Bachelor of Science, Simon Fraser University, 2019**

A thesis submitted
in partial fulfilment of the requirements for the degree of

**MASTER OF SCIENCE**

in

**BIOCHEMISTRY**

Department of Chemistry and Biochemistry
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

MACHINE LEARNING METHODS FOR THE PREDICTION OF ANTIMICROBIAL RESISTANCE AND IDENTIFICATION OF NOVEL MARKERS OF RESISTANCE IN *ESCHERICHIA COLI*

JANICE MOAT

Date of Defence: November 27, 2023

| | | |
|---|---|---|
| Dr. Athanasios Zovoilis<br>Dr. Chad Laing<br>Thesis Co-Supervisors | Associate Professor<br>Research Scientist | M.D., Ph.D.<br>Ph.D. |
| Dr. Tim McAllister<br>Thesis Examination Committee Member | Research Scientist | Ph.D. |
| Dr. Rahat Zaheer<br>Thesis Examination Committee Member | Research Scientist | Ph.D. |
| Dr. Theresa Burg<br>Internal External Examiner<br>Department of Biological Sciences<br>Faculty of Arts and Science | Professor | Ph.D. |
| Dr. Michael Gerken<br>Chair, Thesis Examination Committee | Professor | Ph.D. |

# Dedication

To family lost along the way: Dolores, Gary, Phoebe, & Sumo.

And to family found: Ilya & Tazik.

# Abstract

Antimicrobial resistant strains of pathogenic *Escherichia coli* are a burden on the health-care system, causing longer hospital stays and increased treatment costs compared to non-resistant strains. The proportion of *E. coli* infections in Canada caused by resistant strains producing extended-spectrum beta-lactamase rose from 3.4% in 2007 to 11.1% in 2017. Use of most antimicrobials in the treatment of Shiga-toxin producing *E. coli* infection is not recommended due to their propensity to increase toxin production. Rapid detection of resistant strains would improve both treatment and prevention of this pathogen. With whole genome sequencing (WGS) now ubiquitous in the analyses of outbreak and surveillance samples, in silico methods can be both faster and cheaper than traditional wet-lab methods. In this work, machine learning (ML) classification methods were used for the prediction of antimicrobial resistance and the identification of potentially novel genomic markers of resistance in *E. coli*.

There are four supplementary files to accompany Chapter 3. Supplementary Figure 1 is the Phylogenetic tree of the 4300 E. coli isolates with Salmonella as an outgroup; it is coloured by country of origin and serotype and paired with legends for each. Supplementary Table 1 contains the list of publicly available genomes collected and their corresponding metadata. Supplementary Table 2 contains the top features extracted from trained machine learning models and their annotations. Supplementary Table 3 contains the accuracy data for training machine learning models across a range of 100 to 5000 features. Supplementary Table 4 contains the complete performance data for the 11-mer and 25-mer machine learning models.

# Acknowledgments

I would like to thank my supervisors and committee members for their advice, supervision, and patience: Chad Laing, Athan Zovoilis, Tim McAllister, and Rahat Zaheer. I would also like to thank Matthew Whiteside, Jane Hobson, Vic Gannon, and everyone in Bacti. Res. for their support.

I would also like to thank the support of the Alberta government (as part of the Alberta Technology and Innovation Strategy (ATIS)), the Genomics Research and Development Initiative (GRDI), the Canadian Integrated Program for Antimicrobial Resistance Surveillance (CIPARS), and the Antimicrobial Resistance One Health Consortium (AMROH).

And a special thanks to all the cats that walked on my keyboard and shed all over my thesis: Munchkin, Jasper, Lewis, Houdini Hotdog, Minx, and Tazik.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| AAFC | Agriculture and Agri-Food Canada |
| AMBER | assessment of metagenome binners |
| AMC | co-amoxiclav |
| AMP | ampicillin |
| AMR | antimicrobial resistance |
| AMU | antimicrobial use |
| AMX | amoxicillin |
| ANN | artificial neural network |
| ANOVA | analysis of variance |
| APEC | avian pathogenic *E. coli* |
| ARDB | antibiotic resistance genes database |
| ARG | antibiotic resistance gene |
| ARG-ANNOT | antibiotic resistance gene-annotation |
| ARGs-OAP | antibiotic resistance genes online analysis pipeline |
| AST | antimicrobial susceptibility test |
| AZT | aztreonam |
| BLAST | basic local alignment search tool |
| CAMI | Critical Assessment of Metagenome Interpretation |
| CARD | Comprehensive Antibiotic Resistance Database |
| CET | cephalothin |
| CFZ | cefazolin |
| CGE | Centre for Genomic Epidemiology |
| cgMLST | core genome multilocus sequence typing |
| CHL | chloramphenicol |
| CIP | ciprofloxacin |
| CIPARS | Canadian Integrated Program for Antimicrobial Resistance Surveillance |
| CLSI | Clinical and Laboratory Standards Institute |
| CMY | cephamycinase (β-lactamase gene) |
| COCACOLA | composition, read coverage, co-alignment, and paired-end read linkage (binning tool) |
| CPM | cefepime |
| CRO | ceftriaxone |
| CST | colistin |

| | |
|---|---|
| CTX | cefotaxime |
| CTX-M | cefotaximase, first isolated in Munich (β-lactamase gene) |
| CTZ | ceftazidime |
| CXA | cefuroxime axetil |
| CXM | cefuroxime |
| DAEC | diffusely adherent *E. coli* |
| DAS Tool | dereplication, aggregation, and scoring strategy tool |
| DE | Deutschland / Germany |
| DeepARG | deep learning for predicting ARGs (pipeline) |
| DeepARG-LS | deep learning for predicting ARGs using long sequences (pipeline) |
| DeepARG-SS | deep learning for predicting ARGs using short sequences (pipeline) |
| DK | Denmark |
| DRIAMS | database of resistance against antimicrobials with MALDI-TOF mass spectrometry |
| EAEC | enteroaggregative *E. coli* |
| EFX | enrofloxacin |
| EHEC | enterohemorrhagic *E. coli* |
| EIEC | enteroinvasive *E. coli* |
| EPEC | enteropathogenic *E. coli* |
| ESBL | extended-spectrum beta-Lactamase/β-Lactamase |
| ETEC | enterotoxigenic *E. coli* |
| ETP | ertapenem |
| EUCAST | European Committee on Antimicrobial Susceptibility Testing |
| ExPEC | extraintestinal pathogenic *E. coli* |
| FAST | flow-cytometer method of antimicrobial susceptibility testing |
| FFC | florfenicol |
| FOX | cefoxitin |
| FTIR | Fourier-transform infrared (spectroscopy) |
| GEN | gentamicin |
| HUS | haemolytic uremic syndrome |
| IMP | imipenem |
| IPEC | intestinal pathogenic *E. coli* |
| KAN | kanamycin |
| LEE | locus of enterocyte effacement |
| LT | heat-labile enterotoxin |
| LVX | levofloxacin |

| | |
|---|---|
| MALDI-TOF/MS | matrix-assisted laser desorption/ionization time-of-flight mass spectrometry |
| MDR | multidrug resistance |
| MEGARes | Microbial Ecology Group antimicrobial resistance (tool) |
| MER | meropenem |
| MetaQUAST | quality assessment tool for metagenome assemblies |
| MetaSPAdes | St. Petersburg genome assembler for metagenomic data |
| MGE | mobile genetic element |
| MIC | minimum inhibitory concentration |
| ML | machine learning |
| MLST | multilocus sequence typing |
| MLVA | multiple-locus variable-number tandem repeat analysis |
| NAL | nalidixic acid |
| NCBI | National Center for Biotechnology Information |
| NIT | nitrofurantoin |
| NMEC | neonatal meningitis-causing *E. coli* |
| ONT | Oxford Nanopore Technologies |
| OXA | oxacillinase (β-lactamase gene) |
| PacBio | Pacific Biosciences |
| PCR | polymerase chain reaction |
| PFGE | pulsed-field gel electrophoresis |
| QUAST | quality assessment tool for genome assemblies |
| RAM | random access memory |
| RASE | resistance-associated sequence elements |
| SAM | ampicillin-sulbactam |
| SARG | structured antibiotic resistance gene (database) |
| SEPEC | sepsis-associated *E. coli* |
| SHV | sulfhydryl reagent variable (β-lactamase gene) |
| SIR | susceptible, intermediate, or resistant (classification) |
| SNP | single nucleotide polymorphism |
| SOX | sulfisoxazole |
| SPAdes | St. Petersburg genome assembler |
| ST | heat-stable enterotoxin |
| STEC | Shigatoxigenic *E. coli* / Shigatoxin-producing *E. coli* |
| STR | streptomycin |
| Stx | Shiga toxin |
| SVM | support vector machine |
| SXT | co-trimoxazole |
| TBM | tobramycin |

| | |
|---|---|
| TEM | Temoneira (β-lactamase gene) |
| TET | tetracycline |
| TGC | tigecycline |
| TIO | ceftiofur |
| TMP | trimethoprim |
| Toho | Toho University School of Medicine (β-lactamase gene) |
| TZP | piperacillin/tazobactam |
| UniProt | universal protein resource |
| UOE | University of Occupational and Environmental Health (β-lactamase gene) |
| UPEC | uropathogenic *E. coli* |
| VTEC | verotoxigenic *E. coli* / verotoxin-producing *E. coli* |
| wgMLST | whole genome multilocus sequence typing |
| WGS | whole genome sequencing / whole genome sequence(s) |
| WIMP | what's in my pot? (binning tool) |
| XGB | XGBoost (gradient boosted decision tree ensemble) |

# List of Abbreviations: Bacterial Species

| | |
|---|---|
| *C. difficile* | *Clostridium difficile* |
| *E. coli* | *Escherichia coli* |
| *K. pneumoniae* | *Klebsiella pneumoniae* |
| *M. tuberculosis* | *Mycobacterium tuberculosis* |
| *N. gonorrhoeae* | *Neisseria gonorrhoeae* |
| *P. aeruginosa* | *Pseudomonas aeruginosa* |
| *S. aureus* | *Staphylococcus aureus* |
| *S. enterica* | *Salmonella enterica* |
| *S. pneumoniae* | *Streptococcus pneumoniae* |

# Chapter 1

# Introduction

## 1.1 *Escherichia coli*

### 1.1.1 *E. coli* as a Pathogen

*Escherichia coli* are ubiquitous within the gut microbiome of mammals, colonizing human babies shortly after birth [1]. This bacterial species is important for the normal function of the digestive tract, and ordinarily causes no ill-effects to the host; however, not all strains are commensal [2].

Pathogenic *E. coli* are one of the leading causes of gastroenteritis in humans, and are also responsible for a variety of diseases outside of the gut, including meningitis and infections of the respiratory tract, urinary tract, and bloodstream [2]. The severity of infection ranges from typical food poisoning to kidney failure or death [2, 3]. Immunocompromised, young, and elderly people are at increased risk of severe infections and poorer outcomes [3, 4]. Disease prognosis is also impacted by the pathotype of the bacteria and the complement of specific virulence factors it harbours [4].

Outbreaks of pathogenic *E. coli* can commonly be sourced to meats and produce that have come in contact with contaminated manure, whether through transfer from dirty hides in processing plants, or its use as a fertilizer. Livestock, especially cattle, are a primary reservoir of *E. coli* linked to human infection.

Outside of humans, infection with *E. coli*, or colibacillosis, is especially problematic for young livestock. Colibacillosis causes diarrhea in sheep and goats, and is one of the leading causes of mortality for newborn lambs [5]. In pigs, post-weaning colibacillosis occurs when

*E. coli* infects the gastrointestinal tract of young pigs, resulting in symptoms such as diarrhea, dehydration, edema, slow growth rate, and possibly death [6]. Colibacillosis in calves less than 8 weeks old is a major cause of diarrhea, dehydration, and death [7]. Pathogenic *E. coli* can spread among calves within a herd, causing septicemia and rapid progression to death [7]. In poultry, *E. coli* can cause septicemia, pericarditis, conjunctivitis, cellulitis, and respiratory infections [8, 9, 10].

### 1.1.2 *E. coli* Pathogroups

Pathogenicity refers to the capacity to cause disease, while virulence refers to disease severity; however, definition disputes (primarily of "virulence") are ongoing, with debates dating back to at least the 1940s [11, 12, 13, 14, 15]. The term "virulence factor" is generally used to refer to a genomic feature which enables or enhances the ability of the organism to cause disease [13, 14, 15, 2].

Pathogenic *E. coli* strains can be grouped into pathotypes, which are defined by the presence of virulence factors, the site of infection of the host, and disease manifestation [2, 16]. Examples of virulence factors in *E. coli* are genes related to adhesion or toxin production [17]. These factors can often be mobilized between strains by horizontal transfer [18].

There are two broad groupings, Extraintestinal Pathogenic *E. coli* (ExPEC), whose members cause disease outside of the digestive tract, and Intestinal Pathogenic *E. coli* (IPEC), whose members cause disease of the digestive tract [19].

ExPEC include Uropathogenic *E. coli* (UPEC), Neonatal Meningitis-causing *E. coli* (NMEC), and Sepsis-Associated *E. coli* (SePEC). ExPEC strains are commonly opportunistic, causing infections in those with compromised immune systems [4]. An additional pathotype is Avian Pathogenic *E. coli* (APEC), an important group responsible for disease in domestic poultry [20].

The IPEC pathotype can be further split into more specific pathotypes: Enteropathogenic *E. coli* (EPEC), Enterotoxigenic *E. coli* (ETEC), Enteroaggregative *E. coli* (EAEC), En-

teroinvasive *E. coli* (EIEC), Diffusely Adherent *E. coli* (DAEC), and Shiga toxin-producing *E. coli* (STEC). EPEC strains cause diarrhoeal disease and harm the small intestine via an attaching and effacing modality that is coded in the locus of enterocyte effacement (LEE) pathogenicity island which aides in adherence to the intestinal epithelium where they destroy the lining and cause lesions [2]. ETEC also causes disease by targeting the small intestine. They produce heat-labile enterotoxins (LTs) and/or heat-stable enterotoxins (STs), which cause diarrhoeal disease [2]. EAEC primarily target the colon, and do not produce LT or ST, though they do produce other toxins [2]. EAEC can be identified in culture based on their ability to form "stacked brick" aggregates [2]. DAEC also show a distinctive scattered adherence pattern in culture. They can also be identified by the presence of adhesins belonging to the 'Dr family,' which are responsible for their adherence pattern. EIEC cause disease by entering intestinal epithelial cells, which leads to inflammation and lesions [2].

STEC are sometimes also referred to as Verotoxin-producing *E. coli* (VTEC). STEC strains are of great concern, as they are capable of causing severe disease and death in humans [21]. This pathogroup is defined by the production of Shiga toxin (Stx), which enters and kills intestinal cells by disrupting protein synthesis [2]. After being produced by STEC in the gastrointestinal tract, the toxins can enter the bloodstream. Haemolytic Uremic Syndrome (HUS) occurs when the toxins reach the kidneys and damage the renal cells, potentially resulting in renal failure. There are two groups of Stx: Stx1 and Stx2. They have the same mechanism for causing disease, however, Stx2 is associated with a greater risk of progression from intestinal infection to HUS in humans [22].

The ability to produce Stx does not immediately mean the strains are capable of causing disease: at least one additional virulence factor, such as the well known LEE, is required. [2]. LEE STEC strains are classified as Enterohemorrhagic *E. coli* (EHEC). As the 'hemorrhagic' in the name implies, this group causes attaching and effacing lesions and bloody diarrhoeal disease [2]. Non-LEE STEC strains may also cause disease, if other virulence factors are present [2]. Some of the currently known factors are STEC autoagglu-

tinating adhesion (*saa*), iron-regulated gene homolog adhesion (*iha*) and subtilase cytotoxin (*subAB*), and toxigenic invasion loci A (*tia*) [23, 24, 25].

Most antibiotics are not recommended for treatment of STEC infections, as they can increase the amount of toxins in the patient and potentially cause HUS [26]. Stx is encoded on lambdoid prophages in the chromosome, and its production is dependent on the lytic cycle [27]. Antibiotics can elicit a stress response by damaging DNA, which causes the lytic cycle to be triggered, increasing Stx production [26]. The bacterial cells then lyse, releasing toxins into the host. Within four hours the levels of toxin can increase 3-fold [26].

Historically, one of the most informative systems for classifying *E. coli* was serotype, which is antigenically defined. Primarily two antigens are considered: the O (somatic) antigens and H (flagellar) antigens. The K (capsular) antigens are usually not considered, as traditionally it was more difficult to type [28, 29]. As of 2016 there were 186 O-types and 53 H-types recognized for *E. coli* [28, 29]. O antigens are highly variable components of the outer cell membrane and serve as targets for the immune system of the host [28]. Traditional serotyping is performed in the wet-lab, with one agglutination test per possible O-type, but some O-types can not be determined with this method [28, 29].

Certain combinations of pathotype and serotype can be predictive of pathogenicity, and severity of infection. The seropathotype classification method was introduced by Karmali *et al.* [30], who divided STEC into five seropathotypes based upon association with certain serotypes, outbreaks, and HUS. They are designated A through E in order of their association with disease in humans: A is the most frequently associated with outbreaks and HUS, while E causes animal but not human disease [30]. Seropathotype A strains have a high incidence rate, are common causes of outbreaks, and are associated with severe disease [30]. This group consists of the *E. coli* O157:H7 and O157:NM serotypes which are commonly associated with foodborne outbreaks. Seropathotype B strains have a moderate incidence rate and are uncommon causes of outbreaks, but are still associated with severe disease [30]. This group consists of O26:H11, O103:H2, O111:NM, O121:H19,

and O145:NM [30]. Seropathotype C strains have a low incidence rate, rarely cause outbreaks, and are associated with severe disease to a lesser extent than A and B [30]. They include O91:H21, O104:H21, and O113:H21 [30]. Seropathotype D strains have a low incidence rate, rarely cause outbreaks, and are not associated with severe disease in humans. Seropathotype E have not been associated with infections in humans.

STEC O157 is a primary focus of many surveillance programs and treatment studies. These strains belong to seropathotype A and are responsible for the most frequent and severe human disease. In 2018, there were 1.15 cases of O157 infection per 100,000 people in Canada, a rate that has been consistent since 2010 [31]. Non-O157 strains are inconsistently surveyed; however, the incidence rate of non-O157 infections has recently increased, surpassing the number of O157 infections with a rate of 1.42 cases per 100,000 people in 2018 [31]. The estimated economic burden of O157 illnesses in Canada was $403.9 million in 2013, including the infection treatment and associated long term care [32]. As the predominant strains associated with human infection change over time, it is important to include all pathogenic strains in routine surveillance to be able to monitor and reduce the burden on the healthcare system.

One concerning trend is the occurrence of 'hybrid' strains, which belong to multiple pathotypes [33]. A 2011 outbreak in Germany of an O104:H4 *E. coli* strain belonging to both STEC and EAEC resulted in 900 cases of haemolytic uremic syndrome and 54 deaths [34, 21]. The average onset of haemolytic uremic syndrome caused by this strain was two days shorter than that caused by O157 [21]. The virulence of this strain was due to the combination of antibiotic resistance genes (ARGs), Shiga toxin production, and enhanced adhesion to intestinal tissues [21]. O104:H4 was not on the radar when the seropathotype classification was created in 2003 [30]. With different pathotypes having overlapping virulence factors, the defining set of factors varying by study, and the occurrence of hybrid strains, the pathotype classification is of debated usefulness. Additionally, serotypes, pathotypes, and seropathotypes are labelled as high risk based on past illnesses and out-

breaks and may not be as informative of future risks to human health. It has been suggested that other typing methods, including whole genome sequencing, be used to further support classification schemes [4].

As previously mentioned, antibiotics are commonly not used to treat STEC infections, because they can further harm the patient. A study by Wong *et al.* found that the risk to patients infected with an STEC O157 strain developing haemolytic uremic syndrome was increased by every antibiotic class examined [35]. However, another study, of STEC/EAEC O104, found that some antimicrobials did not cause an increase in toxin levels [36]. Meropenem, azithromycin, tigecycline, and rifaximin were identified as antibiotics which may be suitable for treating STEC O104 infections [36]. The effectiveness of select antibiotics on this strain could be due to its atypical nature, as it possesses characteristics of both STEC and EAEC.

With the limited safe antibiotics available for STEC infections, there may be no suitable antimicrobials for treatment if the level of resistance to them continues to increase. Accurately characterizing pathogenic *E. coli* would allow for improved antimicrobial stewardship practices, as well as the possibility of offering patients personalized antibiotic treatment, instead of only supportive care such as hydration for STEC infections.

### 1.1.3 Rise of Antimicrobial Resistance

#### Antibiotics and Resistance Mechanisms

Despite the relatively stable incidence rate of *E. coli* infections in Canada, treatment is becoming more difficult and costly as strains are becoming increasingly resistant to antibiotics [37]. Antimicrobial Resistance (AMR) can arise naturally, but selection pressure applied through the use of antibiotics speeds up its development and spread [38].

*E. coli* belong to the Enterobacteriaceae family of gram negative bacteria, which also includes *Klebsiella* spp. and *Salmonella* spp. Classes of antibiotics commonly used for the treatment of gram negative infections include cephalosporins, fluoroquinolones, tetra-

cyclines, aminoglycosides, carbapenems, and penicillins [39].

In Canada, antibiotics are grouped into Categories, numbered I through IV, according to their importance to human medicine. Category I antimicrobials are of the highest importance to human medicine, and their use in livestock and other animals is restricted [40, 41, 42, 43]. Carbapenems, cephalosporins (third and fourth generation), and fluoroquinolones belong to Category I [42, 43]. In contrast, Category IV are not used in human medicine, and are administered only to animals [42].

For treatment of *E. coli* infections, one of the largest groups of antibiotics is the beta-lactam group. This group includes penicillins (Category I-II), cephalosporins (Category I-II), carbapenems (Category I), and monobactams (Category I) [44, 42]. The mechanism of action for this group is through inhibition of bacterial cell wall synthesis [45]. The drugs bind to penicillin binding proteins in the cell membrane, which are required for cell wall synthesis [45]. The members of this class of antibiotics share a defining feature: a beta-lactam ring [44]. Resistant strains of *E. coli* produce beta-lactamases, enzymes which hydrolyze this ring, rendering the antibiotic ineffective [44]. As these drugs share this ring structure, this enzyme commonly confers resistance to multiple members of the beta-lactam class. Enzymes enabling multi-drug resistance are referred to as extended-spectrum beta-lactamases (ESBLs). Antibiotics may be prescribed with an inhibitor, such as clavulanic acid, in an effort to prevent hydrolysis of the beta-lactam ring; however, they are not always effective [46, 47]. Penicillins alone are Category II, but penicillin-inhibitor combinations are Category I [42]. In 2016, 11.1% of *E. coli* infections in Canadian hospitals were caused by resistant extended-spectrum beta-lactamase producing strains [48], costing the health-care systems three times more to treat than non-ESBL infections [37].

Another class of antimicrobials used to treat *E. coli* are the fluoroquinolones (Category I). Fluoroquinolones include enrofloxacin, ciprofloxacin, and levofloxacin. This class targets DNA topoisomerase II (DNA gyrase) or topoisomerase IV, enzymes which are required for bacterial DNA replication [49]. *gyrA*, *gyrB*, *parC*, and/or *parE* encode topoisomerase

subunits; resistance is commonly due to mutations in these genes which alter the target binding sites for fluoroquinolones [49, 50]. Resistance can also be due to efflux pumps which remove the antibiotic from the cell [49]. AcrAB-TolC is an efflux pump in *E. coli* that is able to remove fluoroquinolones and other antibiotics. Mutations in regulators, including *acrR* and *marA*, can increase the expression of this pump and thus resistance [49].

Aminoglycosides (Category II-III), including kanamycin, gentamicin, and streptomycin, have a mode of action where they bind to ribosomal RNA (rRNA), thereby inhibiting protein synthesis [51]. Resistant *E. coli* may produce enzymes which can modify the antibiotics, making them unable to bind to the rRNA target [51]. They may also have a mutation in the rRNA target site which prevents binding, or they may possess efflux pumps that remove the drug from the cell [51].

Azithromycin is a member of the macrolide class (Category II) of antibiotics. This antibiotic and others in its class are more commonly used for the treatment of gram positive infections, but also show activity against gram negative species such as *E. coli* [52]. *E. coli* is generally considered to be intrinsically resistant to macrolides because their outer membrane has low permeability [53]. However, azithromycin has increased activity against *E. coli*, compared to other macrolides; this has been attributed to the basicity of the antibiotic, which increases bacterial uptake [54, 55, 56]. Azithromycin binds to the bacterial ribosome and inhibits protein synthesis [57]. Known resistance determinants for this class are *erm*, *ere*, *mph*, *mef*, and *msr* genes [58, 52, 59]. The *erm* genes encode methylases which modify the antibiotic binding site through methylation [58, 52, 59]. The *ere* and *mph* genes encode esterases and phosphotransferases, respectively, which are able to inactivate the antibiotic [58, 52, 59]. Esterases can hydrolyze the lactone ring of the macrolide, while phosophotransferases phosphorylate one of the hydroxyl groups of the antibiotic [58, 52, 59]. The *mef* and *msr* genes encode efflux pumps, which remove the antibiotic from the cell [58, 52, 59].

**Spread of Resistance Genes**

Resistance develops due to selection pressure, where environmental stress on the bacteria causes those with mutations conferring higher tolerance to the stressor to thrive and multiply, while others die off [38]. Antibiotics are one such selection pressure, resulting in AMR strains [38].

Resistance and virulence genes are frequently transferred between strains of *E. coli*. One method is by transduction, wherein bacteriophage pick up DNA from one cell and carry it to another [60]. Another common method is by conjugation, where DNA is transferred between bacteria via a pilus [61, 60]. A third method of gene transfer is transformation, where bacteria pick up DNA from the surrounding environment [60].

A Mobile Genetic Element (MGE) is a portion of genetic material which includes sequences that promote their movement internally within a genome or between bacterial cells. Along with regions facilitating their movement, MGEs may carry many different genes, such as those conferring resistance to antibiotics. MGE types include plasmids, integrative conjugative elements (ICEs), transposons, integrons, and phages. Plasmids are segments of DNA which are typically circular and smaller than chromosomal DNA [60]. They are self-replicating and may have the ability to initiate conjugation by themselves [60]. ICEs use recombinases to integrate and excise themselves to and from chromosomes and plasmids [62]. They are also able to initiate conjugation to move to other cells, however, they are not self-replicative [62, 60]. Transposons are regions containing genes and a transposase which allows movement within a genome and also onto other elements, such as plasmids or ICEs [63]. If transposons are transferred between cells on a plasmid, the transposon may 'jump' from the plasmid to the chromosomal DNA of the receiving cell [63]. Integrons are similar to transposons and their transfer between bacteria requires other elements such as plasmids, but their integration is site-specific, whereas transposons move more randomly [63]. A bacteriophage is a virus capable of infecting bacteria. Upon infection of a cell, the phage may lyse (kill) the host, or it may integrate its genome into the host

chromosome or plasmid [64, 65]. This integrated genome, called a prophage, is a common location for toxin-encoding and virulence genes [60, 64, 65]. A prophage is latent, or inactive, and may be induced by signals, such as those due to stress [60, 64, 65]. Induction results in the production of new phages, and the expression of encoded virulence and toxin genes [60, 65].

Cross-resistance is when one gene is responsible for resistance to multiple antibiotics, and/or other substances, such as detergents [66]. Co-resistance is when genes conferring resistance are located within the same MGE [66]. As a result, selection pressure for one antibiotic or substance will also indirectly select for resistance to the other(s), resulting in a multidrug resistant (MDR) organism [66]. Heavy metals and cleaning agents have been identified as causing selection pressure for AMR in clinical and environmental settings [66].

Resistance can also be dependent on the expression of other genes. Beta-lactamases on plasmids tend to be continuously expressed, while those integrated in chromosomes have their expression regulated by nearby genes [67].

AMR can be spread not just among members of a species, but between different bacterial species. Beta-lactamase genes are transferrable between members of the family Enterobacteriaceae, such as *Salmonella* spp. and *Klebsiella* spp. via plasmids and transposons [44]. Beta-lactamase genes are classified into types, including SHV (sulfhydryl reagent variable), TEM (named for Temoneira, the patient who provided the first sample), CTX-M (cefotaximase; first isolated in Munich), and OXA (oxacillinase) [68, 69, 70, 71, 72, 73]. Commonly, the genes of each type follow the naming convention of type and numbered suffix; for example, CTX-M-1, CTX-M-2, and so forth. This is not a strict rule, for example, Toho-1 (named for Toho University School of Medicine) and UOE-1 (named for University of Occupational and Environmental Health) are CTX-M-type genes [74]. The genes within each of the SHV, TEM, and OXA types are highly similar, having arisen from the same gene through point mutations [67]. The "original" genes TEM-1, TEM-2, SHV-1, and OXA-10 are not responsible for the resistance; the point mutations in these genes con-

ferred the ability to hydrolyse beta-lactam antibiotics [75, 67]. The TEM and SHV genes conferring resistance were the most problematic ESBL *E. coli* genes throughout the 1980s and 1990s, but were overtaken by CTX-M in the mid-2000s [67]. CTX-M genes are believed to have originated in the chromosome of another bacterial genus, *Kluyvera*, and were subsequently passed to *Salmonella* and *E. coli* via MGEs [76, 77].

Fluoroquinolone resistance is typically due to spontaneous mutation in the topoisomerase subunit genes, or the efflux pump regulator genes [50]. Along with vertical transmission of mutations, there are fluoroquinolone resistance genes which spread horizontally. The plasmid-encoded Qnr protein is able to protect *E. coli* topoisomerases from fluoroquinolones by competitive inhibition, blocking the fluoroquinolone binding sites [50]. Plasmids may also carry a mutated aminoglycoside acetyltransferase gene [78, 79]. The variant aminoglycoside acetyltransferase acetylates the antibiotic, resulting in reduced activity [78, 79]. Dissemination of *qnr* genes was identified as an increasing worldwide issue by the early 2000s [78, 80].

Aminoglycoside resistance is also commonly spread by plasmids [51]. Aminoglycoside resistance genes potentially arose from actinobacteria, which produce aminoglycosides naturally [51]. For example, 16S rRNA methyltransferases (RMTs) are a group of enzymes conferring resistance to aminoglycosides by rRNA modification [51]. Plasmids carrying *rmtB* have been shown to be transferable between species, including from *Serratia marcescens* to *E. coli* [81, 82].

For the macrolide azithromycin, the genes carried on plasmids are of higher concern than those in chromosomes [52]. In a study of ETEC isolates in Shanghai, plasmids encoding *mphA* were shown to rapidly spread [83]. The plasmids can spread to other species; *Shigella sonnei* acquired *mphA* from *E. coli*, and this ARG has the potential to move to more Enterobacteriaceae [84, 58]. The *msr* and *mef* genes for efflux pumps may also be transmitted by plasmids [55, 52].

### 1.1.4 Antimicrobial Resistance and Agriculture

*E. coli* reservoirs include wild animals such as birds and rats [85], but livestock, especially cattle, are the primary reservoir for strains pathogenic to humans [86]. Humans become infected by direct contact with animals, or the consumption of contaminated food and water [2]. Person-to-person spread can also occur, although this is more common among children [87, 88].

*E. coli* strains pathogenic to humans may not be pathogenic to animals, but these non-pathogenic strains may still be problematic. As early as the 1970s, it has been known that resistance genes are able to spread from animals to humans [89]. If non-pathogenic strains develop resistance, they may pass the genes on to pathogenic strains [90, 91, 92]. *E. coli* and resistance genes can be carried through the meat production pipeline: from livestock farms, to slaughterhouses and packing plants, and then to consumers [93]. They may also pass to crops by use of contaminated manure, or into the environment through waste and runoff from farms and feedlots [94]. It is difficult to eliminate pathogenic bacteria from a farm setting. Even after thorough cleaning and disinfection, *E. coli* can be introduced or re-introduced via farm dust, machinery, rodents, insects, irrigation water, or farm workers [10, 95].

For livestock, antimicrobials are used both as therapeutic and non-therapeutic agents [93]. Non-therapeutic uses include disease prevention or growth promotion [93]. Even small doses have been shown to contribute to the development of resistance of bacteria in livestock and those who work with them [93]. In 1983, the streptothricin-class antibiotic nourseothricin began to be used as a growth promoter for pigs in the former German Democratic Republic [96]. This resulted in the appearance of plasmid-encoded nourseothricin resistance genes in the *E. coli* of these pigs. The plasmids facilitated rapid spread of the resistance genes among *E. coli*, and also conferred co-resistance to streptomycin and spectinomycin. By 1986, plasmids conferring resistance to nourseothricin were found in not only the pigs given the growth promoter, but the farmers who worked with

them and the families of the farmers. Alarmingly, they were also found in people who had no contact with the pigs or farmers. Although nourseothricin resistance is currently not a concern for human medicine, the streptomycin and spectinomycin co-resistance is, as they are commonly prescribed; the resistance genes could potentially spread to bacterial species which are treated with these antibiotics. This case demonstrates the potential of growth promoting antibiotics to cause rapid spread of resistance not just through farms, but through human communities.

Overcrowding of animals for food production also applies selective pressure for the development of AMR. Densely packed cattle in a feedlot allows for easy spread of bacteria and resistance genes between animals [97, 98, 90]. A study of poultry in the Netherlands examined the effect of administration of antibiotics on the resistance of isolates found in both poultry and the poultry farmers [99]. The intensively farmed flocks with regular antibiotic use harboured a significantly higher amount of resistant *E. coli* strains compared to less intensively farmed flocks with limited antibiotic use. The same pattern of resistance was seen in *E. coli* isolated from farmers for each type of flock.

Surveillance of all stages of the food production pipeline is important to prevent outbreaks, and also to monitor the spread of AMR. According to the 2019 Canadian Integrated Program for Antimicrobial Resistance Surveillance (CIPARS) annual report, 78% of antimicrobials were distributed to livestock, 22% to humans, $<$1% to companion animals, and $<$1% to crops [41]. Adjusting for biomass, the antimicrobial use (AMU) for livestock rose by 5% from 2017 to 2018 [41], declined by 12% from 2018 to 2019 [43], and rose again by 6% from 2019 to 2020 [100]. From 2014 to 2019, the primary purpose of antimicrobial administration in chickens, turkeys, and pigs was for disease prevention [41, 43].

In 2018, AMR surveillance was established by CIPARS for two feedlot and dairy cattle farms [41]: antimicrobial sales increased by 13% for beef cattle and 15% for dairy cattle between 2018 and 2019 [41, 43]. This surveillance effort also identified new and increasing resistance of *Salmonella* strains in cattle [43]. In 2019, Brault *et al.* examined AMU

in cattle in western Canada [101]. Between 2008 and 2012, a total of 2.6 million animals were surveyed, spanning 36 feedlots and representing nearly a quarter of the feedlot cattle in Canada over this time period. During the four years of the study, the main reason for antimicrobial administration was for the prevention of bovine respiratory disease, liver abscesses, and ruminitis. A study in Alberta of retail meat from 2007-2008 found that 32% of *E. coli* isolates from chicken products were multidrug resistant [102]. Beta-lactamase genes were found in 18% of poultry isolates, and 5% of beef isolates [102].

In 2014, Canada enacted stricter rules for AMU in the poultry industry in an attempt to reduce the use of Category I antimicrobials, which are important for human medicine [40, 41]. This change led to a decrease in resistance to third generation cephalosporins in chicken isolates throughout the production chain, from farms to consumers [40].

In 2018, Category I-III antimicrobials became prescription-only for veterinary use in Canada, and labels had growth promotion claims removed [103]. However, CIPARS reported that antibiotics in these categories were still used on some pig farms for growth promotion in 2019 and 2020 [41, 43]. The overall sales of Category I antimicrobials decreased between 2018 and 2019 for pigs and dairy cattle, but increased for beef cattle and companion animals [43]. The amounts remain relatively small, as Category I sales account for only 1% of all antimicrobials sold for animal use in Canada [43].

Along with reduced AMU, preventative measures taken during production can reduce the spread of bacteria and resistance. Slaughterhouse interventions in the beef production process has been shown to decrease the amount of bacteria and resistance genes [104, 90]. A review study by Greig *et al.* examined the efficacy of the different intervention methods within the abattoir, in terms of *E. coli* elimination [104]. The methods examined were carcass washing, steam and hot water pasteurization, and dry chilling. All interventions were shown to effectively decrease the amount of *E. coli* present, with pasteurization being the most important. Additional interventions of acid washing and pre-chilling carcasses provided an additional decrease, but not enough to be recommended for mandatory inclusion.

Another study, by Noyes *et al.*, monitored resistomes (present ARGs) during the production pipeline, collecting samples from both intensively farmed cattle and the surrounding environment [90]. They monitored the feedlot, sampling each cattle upon arrival and exit. Interventions during this stage included daily monitoring and removal of sick animals. Antimicrobials were primarily administered during the feeding stage and were found to impact resistome richness via selection pressure: resistance genes for administered antimicrobials persisted, while the levels of most other resistance genes decreased or disappeared. Pre-slaughter samples from transport trucks had up to 100 unique antimicrobial resistance determinants. Post-slaughter samples had no detected antimicrobial resistance determinants, though bacterial species were present. This suggests that abattoir interventions (including pasteurization, acid spray, carcass trimming, and carcass steam vacuuming) are effective at decreasing the levels of resistance determinants; however, the lack of resistance genes in the final product may have been due to amounts of bacteria being below the limits of detection [90].

## 1.2 Phenotypic Classification

### 1.2.1 Antimicrobial Susceptibility Testing

Antibiograms are laboratory diagnostic tests to identify the susceptibility and resistance characteristics of bacterial cultures. These tests are required in clinical settings to ensure that patients are treated with effective measures. The current gold standards for antibiograms are wet lab diagnostic methods [105]. These methods, along with species identification, include multi-day culturing steps, and require the appropriate reagents and qualified technicians [106].

Antimicrobial susceptibility test (AST) results are reported as either a minimum inhibitory concentration (MIC), which is the smallest amount of antibiotic required to inhibit visible bacterial growth, or as a classification of “susceptible,” “intermediate,” or “resistant” (SIR) to the antibiotic. The MIC results are compared with standardized breakpoint

tables to determine the SIR classification and appropriate treatment. When the MIC value is below the breakpoint, the antibiotic is effective, and the bacteria is declared susceptible; sometimes this case is instead referred to as "sensitive". If the MIC is above the breakpoint, the antibiotic is ineffective, and the bacteria is able to grow in spite of its presence, making it a poor choice of treatment for infection. If the MIC falls directly upon a breakpoint, or within a breakpoint range, then the antibiotic is neither effective nor ineffective, and the bacteria is deemed intermediately resistant/susceptible; such an antibiotic is generally not a favourable choice of treatment.

As resistance trends change, so do the breakpoints. Regional committees update and publish their standardized tables yearly. In North America, the breakpoints are set by the Clinical and Laboratory Standards Institute (CLSI) [107], and in Europe they are set by the European Committee on Antimicrobial Susceptibility Testing (EUCAST) [108]. As the committees are independent, a sample declared susceptible by CLSI standards may be considered resistant by EUCAST standards, or vice versa [109, 110, 111]. For example, if a bacterial strain had an experimental MIC value of $4\,\mu g/\mu L$ for Gentamicin, it would fall into the susceptible category for CLSI ($\leq 4\,\mu g/\mu L$), but the resistant category for EUCAST ($>2\,\mu g/\mu L$). The standards may also lack entries for certain antibiotics, especially if treatment with the drug is not recommended. As a current example, CLSI has azithromycin MICs for *E. coli*, while EUCAST does not. As of January 2020, EUCAST has redefined their "intermediate" classification to address confusion surrounding its interpretation [112]. The definition for the intermediate class is now "susceptible with high exposure," which means that the bacteria could potentially be treated with the antibiotic, provided that the dose is increased [112].

A typical diagnosis involves a multi-day process of culturing to identify and isolate the bacteria, then additional culturing to determine antibiotic susceptibility [113, 105]. Such delay in treatment administration can have significant impacts on the recovery and potential mortality of patients. In the case of septic shock from a bloodstream infection, the risk of

death increases with every hour of delay in drug administration, as after two days of delay the probability of death is nearly 100% [114].

A common practice for risky infections is to empirically prescribe broad-spectrum antibiotics, then adjust the treatment after an AST panel is completed [114, 115]. Whether this method improves survivability is disputed, and the overuse of broad-spectrum antibiotics also comes with drawbacks, including healthcare cost, preventable patient side effects, and increased AMR [115, 116, 117]. Antibiotics may cause adverse effects directly, for example ciprofloxacin can cause damage to tendons and nerves [118]. In 2018, the US FDA (Food and Drug Administration) announced that improved safety information would be required for fluoroquinolones, including ciprofloxacin, to warn that they can cause serious blood sugar and mental health problems [119]. Antibiotics can also cause indirect effects, such as disrupting the normal microbiome and promoting secondary infections. For example, antibiotic use is associated with subsequent vulvovaginal candidiasis, commonly known as a vaginal yeast infection, in women [120]. Side effects may also only be a risk for patients under specific circumstances. UTIs commonly occur during pregnancy, but as the drugs can cross the placenta, the effects on the fetus must also be considered when selecting treatment therapies [121, 122]. Tetracycline, streptomycin, and kanamycin are safe antibiotics for adults; however, they can cause birth defects such as hearing loss if taken during pregnancy [121, 122]. A faster method of AST allowing for more specific and efficient treatment would potentially reduce costs and improve patient prognosis, while minimizing the risk of resistance development.

### 1.2.2 Historical Culture Methods

Prior to conducting an AST, the bacteria must be cultured and isolated from a sample. The historical process of obtaining and identifying a pure culture can take up to two days [113, 123]. After isolation, the sample is subjected to an AST panel, where the protocol is repeated for each relevant antibiotic.

### Agar Methods

For the agar dilution method, multiple agar plates are prepared, each with a different concentration of antibiotic [124]. The bacterial culture is aliquoted onto each plate and spread evenly. Plates are incubated for between 18 and 48 hours, depending upon the type of media used in the protocol [105]. The plate with the lowest concentration of antibiotic that shows no visible bacterial growth indicates the MIC.

Another plating technique is disk diffusion, also known as the Kirby-Bauer method. Bacterial culture is evenly spread onto an agar plate, then small disks containing a known concentration of antibiotic are placed on top. After 24 hours of incubation, there will be a lawn of bacteria with areas of no growth, called plaques, surrounding the antibiotic disks. The measurement of the diameter of the plaque is called the zone of inhibition. This value is compared to a standardized table, such as those provided by CLSI or EUCAST, to determine the MIC [105].

The Etest strip method is very similar to disk diffusion. The plate is prepared in the same manner, but in place of disks, a plastic strip containing a gradient of antibiotic concentrations is used. The resulting plaque will vary in width in response to the gradient. The concentration where the plaque begins is the MIC [105].

### Broth Methods

Broth macrodilution is similar to agar dilution, but uses liquid media. Serial dilution is used to create an antibiotic concentration gradient across tubes of media. The bacterial culture is aliquoted into each tube, and they are incubated for 24 hours. The optical density of each tube is used as a measure of the bacterial growth [105].

Microdilution follows the same principle as macrodilution, but is performed using a 96-well plate [124]. The antibiotic concentration gradient is created by serial dilution across a row of wells instead of across tubes. After incubation, a plate reader is used to record the optical density of each well.

For macro and micro dilution, the MIC can be determined either visually or by measuring optical density, depending on the protocol used [105, 124]

### 1.2.3 Optical Methods

#### Flow Cytometry

In flow cytometry, sample characteristics are measured as bacteria move past a sensor. For AST with flow cytometry, a 96-well plate is prepared with standard concentration gradients of antibiotic, and then inoculated with the bacterial culture [113, 105]. After a two to four hour incubation, a fluorescent stain is added, and the plate is incubated again for a short time [113]. The flow cytometer is then used to measure the fluorescence of the stained cells. The intensity of fluorescence varies in response to the membrane potential changes in the bacterial cells [113, 105]. Increased emission is caused by disruption of the membrane potential due to antimicrobial activity [113, 105].

Flow cytometry is faster than traditional culture methods, and shows promising high accuracy [113, 105]. One of the biggest challenges with this method is that the fluorescent response to antibiotics will vary by bacterial strain and antibiotic used [105]. Standard protocols and guides to interpretation need to be developed if this method is to be used clinically [105]. An additional problem with flow cytometry is the difficulty of distinguishing a single cell from an aggregate of cells [105, 125].

#### MALDI-TOF

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF/MS) is a technique that can be applied for both species identification and resistance profiling. In brief, the expressed proteins in a bacterial sample are ionized and released into the gas phase to move along a tube towards a detector [126]. The movement speed is determined by the mass and charge of the particle type. The mass-to-charge ratios are recorded to create a spectrum for the sample [123]. This spectrum is like a bacterial fingerprint: it can be compared to a database of standard spectra to determine species, or even

strain [127, 123]. For *E. coli*, MALDI-TOF for AST has primarily been investigated for the beta-lactam antibiotics [123, 128, 129, 130].

MALDI-TOF spectra are representative of the whole sample, but only detect proteins within a set mass range. Some proteins associated with resistance may be too large for detection by MALDI-TOF. Such proteins include *E. coli* beta-lactamases [131], so their detection must be with indirect methods. Similar to the flow cytometry method, samples are examined at time points after incubation with antibiotics. If there is beta-lactamase present, it will hydrolyze the antibiotic, and the peaks for the resulting products will be seen on the spectra [123, 132, 129].

MALDI-TOF AST methods can be completed in just four hours; however, most still require an overnight culture before beginning the analysis [123, 133]. Though the cost per test is low, the equipment and its maintenance are expensive [105]. Drawbacks to AST by MALDI-TOF include the inability to distinguish between pathogenic strains and non-pathogenic strains that carry resistance markers [126], lack of resolution between *E. coli* and *Shigella* [126], and resistance-associated proteins potentially being outside of the detectable mass range [131]. As with other database methods, there is also an inability to account for novel resistance mechanisms [123].

**Fourier-Transform Infrared Spectroscopy**

The principle of Fourier-transform infrared (FTIR) spectroscopy is that the amount of infrared light absorbed by a molecule depends on its composition. Much like MALDI-TOF, the result of IR spectroscopy on a sample will be a spectra indicating the composition. This unique spectra of a bacterial culture can be used to determine characteristics such as species, subspecies, or serotype [134]. FTIR has been used for accurate capsular K-typing of *K. pneumoniae* isolates, and identifying K-types with clinical importance [135].

Many studies have attempted to create databases of standard spectra for various kinds of bacterial typing [136, 134, 137], but there is a lack of standardization, curation, centraliza-

tion, and public availability [138, 139]. FTIR shares similar advantages and disadvantages with MALDI-TOF: time is required to produce a pure culture prior to analysis, and the equipment and maintenance is expensive, though the cost-per-run is low [133].

## 1.3 Genotypic Classification

### 1.3.1 Conventional Molecular Subtyping

Conventional sequence-based subtyping methods examine only specific regions of the genome, while ignoring the rest. Methods of molecular subtyping include pulsed-field gel electrophoresis (PFGE), multilocus sequence typing (MLST), and multiple-locus variable-number tandem repeat analysis (MLVA). For MLST and MLVA, the polymerase chain reaction (PCR) is typically used. These methods are commonly used to establish relatedness between strains.

PFGE is a wet-lab method where whole-genome DNA is fragmented using rare-cutting restriction enzymes and run through an agarose gel alongside a standard sample. Like the other wet lab methods, PFGE requires a pure bacterial culture, which is time consuming to produce. The runtime of the PFGE itself can also be considerable; in a study involving ESBL *E. coli* typing, four hours were required for DNA digestion, followed by 20 hours for the PFGE itself [140]. Once complete, bands on the gel are visualized; fragments will move at different speeds along the gel depending on their size. The resulting profile of banding patterns can be compared to databases of known profiles. PFGE is usually used for determining relatedness, especially for tracing outbreaks. However, the result is not always informative. A study examining *E. coli* isolates from Canadian outbreaks spanning from 2008 to 2012 found that PFGE was not well suited for the task of discriminating between strains on its own [141]. The USA surveillance network for foodborne outbreaks, PulseNet, has been using PFGE as their gold standard since 1996. In 2015, 76,000 isolates were identified using their database of PFGE profiles [142]. PulseNet is currently working to replace PFGE with whole genome sequencing, as it has an improved ability to differentiate

between strains [143].

PCR-based methods are much faster than PFGE. They aim to amplify specific regions of DNA, such as resistance genes [105]. Primers specific to the regions of interest are added to the sample, along with nucleotides and a polymerase. Thermal cycling is used to drive the replication process; the number of copies increases exponentially with each cycle. Many commercial kits and machines are available, making this method a convenient option [105]. Not including the time to obtain a pure culture, the method can be completed within 2.5 hours [144, 105]. PCR can be used to detect the presence of bacterial species in a culture. A common target for this is the 16S rRNA gene [145]. The *Salmonella* spp., *Shigella* spp., and *Yersinia* spp. can be identified by PCR within three hours, instead of two to four days, which is required for selective culturing [146]. PCR can also be used for discerning characteristics within strains. For *E. coli*, PCR protocols have been created to differentiate between pathotypes [147, 148]. There are also protocols for serotyping, with a focus on the identification of STEC O157:H7 [149, 150, 151]. A drawback of PCR is the requirement of specific primers for each use case, either from a commercially available kit or designed by trained technicians. An additional challenge is multiplexing, which improves detection, but adds difficulty to experimental design [152]. Multiplexing is when multiple primers are designed for different locations within the same genome, and included in the same amplification; this allows multiple targets to be examined at once, including both the presence and absence of genes [153, 152]. In order to fully characterise a sample by PCR, many separate PCRs would need to be performed, each with their own set of primers.

MLVA is usually used alongside PFGE for verification and improved resolution [141]. In it, PCR is used to amplify variable-number tandem repeats at multiple locations in the genome; these are repetitive regions of DNA occurring beside each other [141]. Capillary electrophoresis is then used with the amplified DNA and a standard, and results can be determined similarly to PFGE [141]. A Canadian study of 2008-2012 outbreak isolates

found that MLVA in addition to PFGE improved resolution by 60% [141]. In Canada, there is only an MLVA protocol for a single *E. coli* serotype, O157:H7 [154, 155].

MLST also uses PCR. To establish relatedness, the targets for PCR are housekeeping genes which should be present in every isolate of a species. There are three commonly used MLST schemes for *E. coli*: the scheme defined by Achtman [156] uses seven housekeeping genes, the scheme by Pasteur [157] uses eight, and the scheme by Whittam [158] uses 15. Though usually used for strain-level resolution, it can also be used for examining resistance. For resistance, there are many possible gene targets, and the choice will depend on what antibiotics are of interest. According to a study involving the typing of ESBL *E. coli*, MLST is better at resolving strains than PFGE when both housekeeping and resistance genes are included [140].

For *E. coli*, the primary targets for AST panels are select ESBL producing genes [144, 140]. The panels may miss other important ESBL or non-ESBL genes [144]. Additionally, the resistance of an isolate may be complex, and unable to be predicted from only the genes in the panel [105, 144]. Resistance can also be due to pumps which remove the antibiotic from the cell, alterations that prevent antibiotic uptake, or novel mechanisms [144, 159].

The results of PCR AST are more informative for gram positive microbes than gram negative ones like *E. coli*, as their genotype may be less likely to match phenotype [144, 159]. If a resistance gene is present, the isolate may still be susceptible if that gene is underexpressed or not expressed at all [159]. Although the PCR results may be less certain for *E. coli*, they can aid in treatment selection while awaiting the results of traditional phenotypic methods [144].

The main drawback to all of these conventional typing methods is that there is a large number of genes to choose from, but only a portion can be examined at a time. Additional drawbacks include the requirement of highly trained personnel to run experiments and analyse data, and consumption of materials and time.

### 1.3.2 Whole Genome Sequencing Technology

As whole genome sequencing (WGS) is becoming increasingly available, cost-effective, and rapid, it is emerging as the new standard for pathogen diagnosis and surveillance [160]. FoodNet Canada and CIPARS are currently using whole genome sequencing in their surveillance of enteric pathogens [161, 162]. The American PulseNet is currently in the process of updating their standard to WGS from PFGE methods for foodborne outbreak surveillance [143]. The complete genome of an isolate allows for more thorough characterization than only examining select regions with PCR [160, 163, 164].

Many technologies are available, varying in sequencing time, error rate, and read length [160]. One of the most popular high-throughput technologies is Illumina sequencing, which has relatively low cost per base sequenced and low error rate [160]. A trade-off with Illumina is a longer runtime compared to other options, taking between four and 55 hours [160, 165]. As with previously discussed techniques, most sequencing technologies also have the prerequisite of a pure culture [164, 166].

Two WGS technologies capable of producing sequences in real-time were developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Both platforms produce longer reads than other next-generation sequencing platforms, which can improve the resolution of plasmids and resistance genes [167]. While Illumina sequencing produces reads of up to 300 bp in length [168], PacBio reads have an average length of 10 kbp to 25 kbp [169, 170], and ONT reads range between 500 bp and 2.3 Mbp [171]. Initially, a primary drawback of these technologies was low throughput and higher error rates, but higher throughput options, such as the ONT PromethION, are overcoming throughput limitations [172]. The accuracy of these technologies has also been steadily improving since the technologies were introduced [172].

One of the primary sequencers available from ONT is the MinION, which provides results in real-time [173]. Analysis can begin while sequencing is in progress, and complete AMR gene detection can be performed within six hours for plasmids [174]. The user can

adjust the run time [173]; longer run times can produce more data, however, a full 72 hours may not be required for diagnostics. In addition to speed, the device is portable, allowing for sequencing with a laptop instead of a laboratory [175]. As of 2021, the basic MinION costs only $1000, though there are extra costs including reagent kits and flow cells [175].

As the ONT technologies are easy to use, and relatively cost-effective, they have been investigated as replacements for traditional AST in clinical settings. Depending on the sample type, the time-consuming step of obtaining a pure culture may be skipped prior to sequencing. A study of *E. coli* spiked urine samples demonstrated that the MinION was able to identify the bacteria without culture [176]. Additionally, resistance profiles were determined by comparing the sequences to AMR gene databases. The results were compared to profiles created from Illumina sequencing analysis, which used a pure culture. The MinION detected 51 of the 55 markers found by the Illumina method. A common issue with the analysis of ONT data is difficulty identifying variants, though this may be improved upon through bioinformatic analysis rather than the sequencing itself [176, 177].

### 1.3.3 Whole Genome Sequencing Analysis

Following WGS, sequence reads are fed into bioinformatic pipelines for analysis. The computational analysis can vary greatly in runtime, from seconds to days, depending on the pipeline used. A primary benefit of *in silico* methods is their efficiency: multiple forms of analysis can be run concurrently by a single technician.

There is an overwhelming selection of bioinformatics software to choose from for typical analysis needs, such as AST, establishing phylogenetic relationships, or subtyping. With so many tools available, standardization is an issue; tools sharing a common purpose may have contradictory results. Another issue is that tools are commonly short-lived, either becoming inaccessible or unusable due to the lack of support and maintenance [178]. Although promising, WGS analysis methods are relatively new, and require more development before being able to replace traditional phenotyping [164].

**Resistance Databases**

There are numerous methods to perform AST by WGS. The most popular technique is comparison of the sequence with one of the many databases of resistance genes [164, 163]. Three of the most well-known databases of AMR genes are the Comprehensive Antibiotic Resistance Database (CARD) [179], the Bacterial Antimicrobial Resistance Reference Gene Database from the National Center for Biotechnology Information (NCBI) [180], and the ResFinder Database from Centre for Genomic Epidemiology (CGE) [181]. Tools such as ABRicate allow convenient access to multiple resistance databases using a single interface [182]. It is also common to create custom databases and use a tool such as BLASTn [183, 184] for comparisons [185].

Although databases can share data with one another [186], they are not equal in content. The variety of databases available creates standardization issues, where each database can use different nomenclature of genes and antibiotics. A comparison of AMRFinder (NCBI) with ResFinder (CGE) for the prediction of resistance in *Salmonella enterica*, *Campylobacter* spp., and *E. coli* found 1,229 discrepancies in the identified AMR genes [187]. A contributor to this was that at the time of the study, ResFinder did not include some of the drug classes in their scope [187].

Beta-lactamase gene names are a good example of discrepancies between databases, as their classification has changed over time. The blaBIL-1 gene discovered in the 1990s is now considered to be the same as blaCMY-2 (CMY: cephamycinase) [68], however, this is not always reflected in every database. There are attempts to create standardization; as of 2015, the NCBI has taken over curation of the many beta-lactamase families previously curated by the Lahey Clinic [186]. Some families are also curated by the Pasteur Institute [188]. Inconsistencies have also been found within databases, such as with the TEM beta-lactamases in the NCBI database, as their rapid expansion causes curation challenges [189, 190].

To ensure that all validated AMR genes are available for analysis, and named consis-

tently, a single standardized database is required [164]. Unfortunately, this ideal solution requires a great undertaking, with international collaboration beyond what is currently feasible [164]. Databases have been created in an attempt to have a central resource for naming and sequences, however, they are incomplete and inactive [189, 191]. A promising database, the Antibiotic Resistance Gene-ANNOTation (ARG-ANNOT) [192], was established in 2015, however, due to lack of maintenance it fell out of favour [193]. Although it may have discrepancies, the NCBI database is widely used, as it is curated as carefully and often as possible.

### Other Methods

Other subtyping methods can be run concurrently with AST, replacing both kinds of conventional testing methods with convenient *in silico* versions, requiring only one technician. Whole genome multilocus sequence typing (wgMLST) and core genome multilocus sequence typing (cgMLST) are typically used to establish relatedness of isolates. They are similar to the previously described MLST, but are not limited to a small set of genes. Reference genomes are used to determine the set of genes to use. Their presence or absence amongst the reference genomes determines if they are "core" or "accessory" genes. wgMLST includes both core and accessory genes, while cgMLST includes core genes only. wgMLST schemes can include tens of thousands of locations across the genome [155, 194], much higher than the traditional subtyping methods, which use 15 or fewer [156, 157, 158]. cgMLST schemes can also include thousands of loci [155]. When an isolate is assessed using one of these schemes, allelic variations can be determined along with presence or absence of a gene. A Canadian study found that wgMLST had improved resolution compared to PFGE with MLVA for strain detection of *E. coli* O157:H7 [155].

De Bruijn graphs have been used in *Staphylococcus aureus* and *Mycobacterium tuberculosis* to examine resistance and relatedness [195]. Another method for *M. tuberculosis* scans sequencing reads directly to look for single nucleotide polymorphisms (SNPs) to

identify resistance mutations [196]. More recently, machine learning methods have become increasingly popular, as they have great potential for the detection of new markers [164].

#### Future of AST with WGS

AST by WGS analysis is still a relatively new procedure and requires more development before being able to replace traditional phenotyping [164]. A primary benefit of using WGS is that no *a priori* information about resistance mechanisms is required [164]. Along with being able to detect known resistance markers, mutations and unknown markers may be identified [163, 164]. *In silico* serotyping or pathotyping, and phylogenetic analyses may supplement the results, and can be run in parallel to the AST.

WGS is increasingly replacing more limited testing methods in pathogen surveillance programs [160, 197]. WGS and *in silico* analysis can replace the many laboratory tests traditionally used in surveillance [197]. Reducing testing into fewer, centralized, tests, allows for more rapid response times [197].

## 1.4 Machine Learning

### 1.4.1 Overview of Machine Learning

Machine Learning (ML) is the use of algorithms to mimic a learning process, with the typical goal of grouping or classifying items. It has a broad range of applications, from identifying birds by their song [198] to diagnosing breast cancer with mammography images [199]. In bacteriology, applications include predicting the ability of phages to infect bacteria [200], host response to an infection [201], new antimicrobial compounds [202], host adaptations by bacteria [203], host or isolation source [204, 205], the association of human pathogens and plants [206], and the zoonotic potential of an isolate [207].

In general, training a model involves providing a dataset of items, such as a set of microbial isolates, animal sounds, or microscopic images. The supplied items have features, which are characteristics such as the presence or absence of a phenotype. The model works

to identify patterns of features which characterise different groups of the items.

ML can be either 'supervised' or 'unsupervised.' Supervised learning uses a labelled dataset, where the classifications of the input data are known. This method is used to create models to predict labels for previously unseen data [208]. Unsupervised learning uses an unlabelled dataset; the general goal of this method is to group the data based on inherent characteristics [208].

Unsupervised methods can be easier to implement than supervised methods, as they tend to be less customizable [209]. A primary benefit of unsupervised learning is also one of its biggest drawbacks: less metadata are required for these methods, as labels are not needed for training [208, 209]. However, without labels there is no direct way to determine if the training is successful [208, 209].

Unsupervised learning is typically used to cluster data into groups sharing similar characteristics. Common clustering methods include k-means, hierarchical clustering, and principal component analysis [208] These methods require a distance metric, which define how group similarity is determined [208, 210]. Metrics can have a large impact on the success of clustering, and there is no one method that is best for all situations [210].

Supervised methods are generally more complex than unsupervised methods and require more metadata, but they benefit from being able to be validated [209, 208]. Typically, the dataset with known labels is divided into parts, including a training set and a validation set. The validation set is withheld from training and used to determine if the trained model is successful. The trained model can then be used on unseen data to classify it. Three common ML models are support vector machines (SVMs), artificial neural networks (ANNs), and gradient boosted decision trees.

The goal of an SVM is to find a hyperplane which best separates the classes of data [208]. The data points closest to the division are the support vectors. The SVM aims to maximize the distance of the support vectors from the plane, to make the largest possible division between classes. An example of an SVM is shown in Figure 1.1.

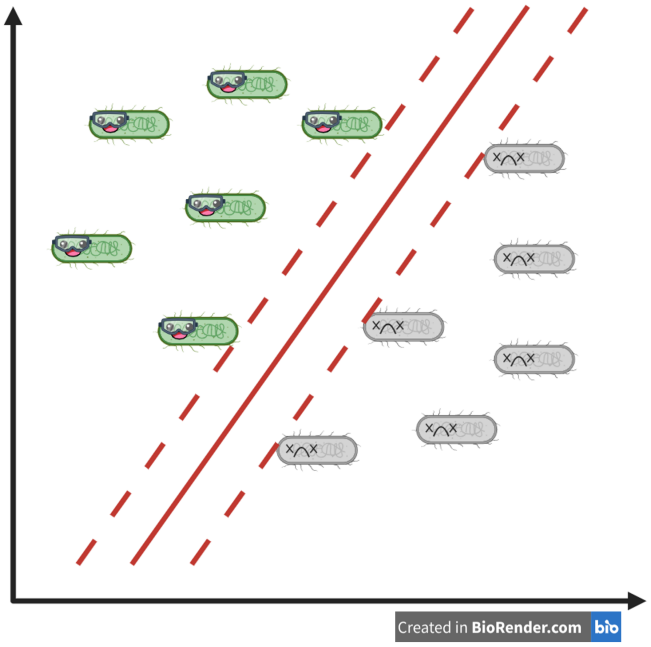



Figure 1.1: An example of a 2-dimensional support vector machine (SVM), depicting the separation of resistant (left, green) and susceptible (right, grey) bacteria. Data are plotted according to features, and the SVM determines the best line (solid line) by maximizing the size of the margin, which is the distance between points in each class (distance between dashed lines). The data points on the margin are called supporting vectors. This image was created with BioRender.com.

An example of ANN structure is shown in Figure 1.2. An ANN is made up of layers of neurons, with each neuron being interconnected to every neuron in the preceding and following layer [208]. The input layer consists of one neuron per input feature, and the output layer consists of one neuron per class [208]. In between the input and output layers can be any number of hidden layers. The connections between neurons have individual 'weights.' The input and weight are passed to an activation function and the output is compared to the neuron's threshold. If the resulting value makes it past the neuron's set activation threshold, then the neuron is activated, and the output is passed on to the next neuron to be used as input, and so forth. This is called "forward propagation". To train this type of model, known data is repeatedly passed through, and the output is compared with the true values; the neural net then adjusts the weights in such a way to improve the model's

accuracy. This is called “backward propagation.”

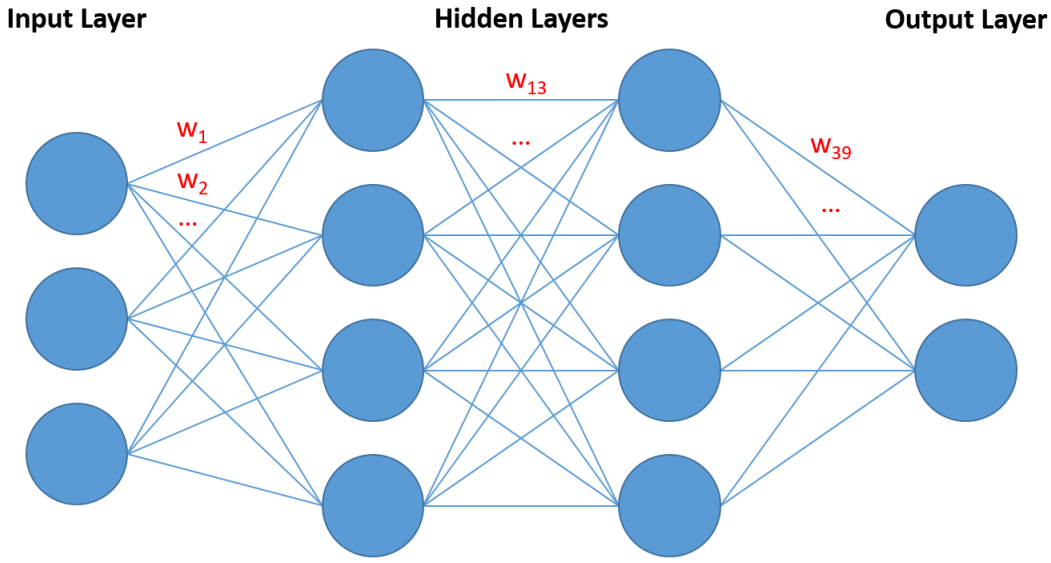



Figure 1.2: An example of an artificial neural network (ANN) structure, consisting of an input layer, two hidden layers, and an output layer. Each neuron is interconnected with every neuron in neighbouring layers. The connections have individual weights (w).

A decision tree is composed of nodes, each associated with a choice [208]. The tree starts with one node, and the possible choices lead to different nodes, and so forth, as in a flow chart. The output nodes are those which lead to no further choices; they represent the predicted class. During training, the choices included in the tree and their order are altered to find the best tree for predicting classes [208]. Trees perform better when used in groups, called ensembles. An example of this is a “forest,” which consists of multiple trees with different series of questions; the final classification is determined by majority or averaging the predictions of the individual trees. An example of a forest is shown in Figure 1.3. Another ensemble method is gradient boosted decision trees. Gradient boosting involves sequentially combining the weak trees to create a better tree [208].

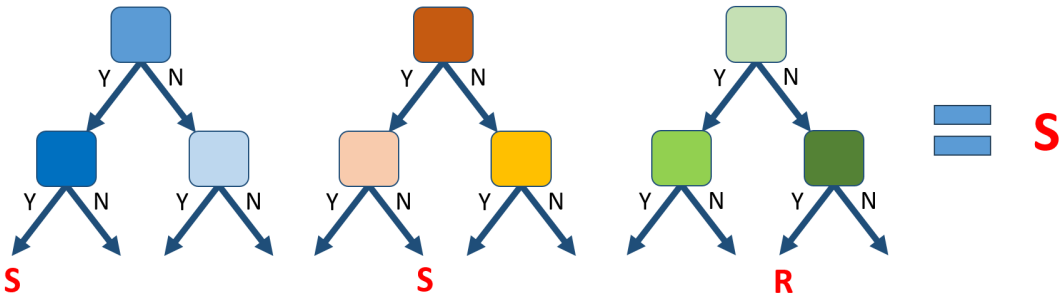



Figure 1.3: An example of an ensemble of decision trees ("forest"). Each decision tree consists of a different series of branching yes (Y) or no (N) questions, which lead to a prediction such as susceptible (S) or resistant (R). In an ensemble, a final classification is determined by majority or by averaging.

ML methods are becoming more accessible, with many able to run on standard desktop computers. Some models can even be run with phones, such as Merlin, which uses a ML model to identify birds by their songs or images [211, 212]. Supervised methods are often used with WGS data, as the required labels, such as isolation source, are usually readily available. Additionally, a reusable model to classify new WGS data is desirable for use in diagnostics or surveillance.

### 1.4.2 Machine Learning and WGS for AMR Prediction

There are many ways to train ML models to predict resistance profiles from sequencing data. The training data may be specific, such as using the amino acid sequences from genes encoding one kind of protein, or broad, such as using the pan-genome, or all present subsequences (*k*-mers).

#### Amino Acids

A study on *Streptococcus pneumoniae* extracted only the genes encoding penicillin-binding proteins from whole genome sequences, and their variability in amino acid sequence, along with MICs, were used to predict resistances to antibiotics [213]. Another study, on *M. tuberculosis*, used SNPs within 23 genes and their upstream regions as input

for machine learning models to predict SIR classification [214]. Instead of hand-picking genes, a reference database (discussed above) can be used to find the known AMR determinants present in the dataset, and also for deducing the variants present [215, 216, 163].

**Pan-genome**

While examining a set of known genes can provide insight into variants' association with resistance, they may overlook other genes or regions contributing to AMR. Broader approaches make use of the pan-genome, or a set of all genes present across the dataset. As the pan-genome may be very large, it is common to break it down into core genes and accessory genomes, or by clustering, or both. In general, the presence/absence or copy number of each gene, gene cluster, or the variants, is used as input for training ML models. This approach has been successful in multiple bacterial species, including *E. coli*, *S. aureus*, *Pseudomonas aeruginosa*, *Elizabethkingia* spp., *Neisseria gonorrhoeae*, and *M. tuberculosis* [217, 218, 219, 220, 221, 222].

An advantage of this method is that it excels in the ability to detect novel markers or mechanisms conferring susceptibility or resistance, on top of accurate predictions [219, 222]. Kavvas *et al.* found that 24 of the 57 identified targets were previously undescribed for resistance in *M. tuberculosis*; however, as independent models for each drug were not made, associations of these targets with specific antibiotics could not be made [222]. In contrast, Hyun *et al.* trained one model per antibiotic for each of *S. aureus*, *P. aeruginosa*, and *E. coli* [219]. They extracted and annotated the genes most important for classification from trained models and selected potentially novel genes associated with AMR. Across all species and antibiotics tested, they selected 25 novel candidates. The candidates included both core and accessory genes; for *E. coli*, all nine candidates identified were accessory genes. Some of the candidates were novel for the antibiotic and species combination, but the mechanism of resistance had been previously described in other species. For example, one candidate for *S. aureus* was a gene encoding for a *HflX* protein that was previously

found to confer resistance in *Listeria monocytogenes*, with Hyun *et al.* proposing that a similar mechanism may be responsible for resistance in *S. aureus* [219, 223].

A drawback is that there is no consensus on how to filter the pan-genome for the most meaningful set of genes for prediction, so any filtering may discard important genomic features. Two recent studies with *E. coli* have demonstrated that accessory genes may be the most important for accurate prediction [217, 218], however, core genes can also be responsible for resistance and should not be ignored [219].

### *k*-mers

An alternative to pan-genome determination and clustering is the use of *k*-mers, or subsequences of a specified length k. The frequency of occurrence of each *k*-mer within a genome may be determined, or just the simple presence or absence of each *k*-mer. They may also be passed in groups instead of individually when training models [224]. A primary advantage of using *k*-mers is that it allows for the examination of all areas of the genome, with no prior knowledge of resistance mechanisms required. As with the previously mentioned pan-genomic approach, this means that the *k*-mer methods are well suited to the detection of novel markers of resistance. An advantage that *k*-mer methods have over gene-based approaches is the ability to also investigate intergenic regions, which can also contain important information, such as markers of host-specificity [225].

There is no single best *k*-mer length for all analyses. The optimal size will vary depending on what species is being examined, how much computational power is available, and how specific the sequence fragments need to be. Shorter *k*-mers may produce informative models, but they are harder to locate within a genome, as they are more likely to appear by chance. Longer *k*-mers are more specific and easier to annotate; however, the total number of unique subsequences of length *k* grows exponentially with each increase of *k*, leading to extreme computational costs. The total number of possible unique *k*-mers is calculated with $4^k$; there are 4,194,304 unique 11-mers, 67,108,864 unique 13-mers, and $4.6 \times 10^{18}$

unique 31-mers. The number of unique *k*-mers and their frequency of occurrence will differ depending on the species being examined, the diversity of the dataset, and the size of the dataset.

Aun *et al.* developed PhenotypeSeeker and predicted phenotypes for *Pseudomonas aeruginosa*, *Klebsiella pneumoniae*, and *Clostridium difficile* [226]. Their dataset had approximately 10 million unique 13-mers and 28 million 31-mers. Prior to training, they used filtering to reduce the number of input features to 1000. They determined that 13-mers were ideal for the performance of PhenotypeSeeker, as the RAM usage did not exceed 2 GB for their dataset. This is within the capabilities of a standard desktop computer. They noted that increasing the number of unique *k*-mers would increase memory requirements. The runtime of model building for PhenotypeSeeker scales linearly, and their dataset of 200 assembled *P. aeruginosa* genomes took just over four hours to process per phenotype, using 13-mers [226]. If it was run on a dataset of 4000, this would take over three days to complete.

Drouin *et al.* developed Kover, and predicted phenotypes for *C. difficile*, *Mycobacterium tuberculosis*, *P. aeruginosa*, and *Streptococcus pneumoniae* [227]. It was limited to binary classification, and cannot additionally predict intermediate classifications [227, 228, 229]. 31-mers were determined to provide optimal performance and specificity for their datasets. The full set of input 31-mers was unable to be fit into the memory of their supercomputer. Three of their four tested model types required that the input undergo feature reduction prior to use. The fourth type was an out-of-core method, which used hard disk space instead of loading the input into memory. As with PhenotypeSeeker, the runtime of model building for Kover increases linearly with the number of input genomes and *k*-mers.

Feature reduction with statistical tests is useful for reducing memory requirements, as shown with PhenotypeSeeker [226]. The number of *k*-mers required for accurate prediction may be small. Mahe and Tournoud trained models for predicting the resistance of *Staphylococcus aureus* and *M. tuberculosis*, and found that only 1 to 8 31-mers were required for

successful classification [224].

### 1.4.3 Machine Learning and Optical Methods for AMR Prediction

#### Flow Cytometry

The results of flow cytometry for AST represent the changes in bacteria in response to the presence of antibiotics. These observations in conjunction with labels can be passed to machine learning models for training. A study developed a flow-cytometer method of antimicrobial susceptibility testing (FAST) that was able to predict MIC within 3 hours, but was limited by the requirement of a specialized technician for analysis [230]. Recently, they improved the throughput of their FAST method by the inclusion of machine learning methods [231]. The hands-on time was reduced to just 10 to 25 minutes, and expert training was no longer required. For their validation test of their decision tree, 83% had an SIR classification in agreement with a traditional broth microdilution method. The sample size was small, but supports that the method should be investigated further in the future [231].

#### MALDI-TOF

MALDI-TOF spectra and resistance labels can be passed to machine learning models for AMR prediction. A 2015 study achieved 89-98% accuracy using SVMs paired with spectra for the SIR classification of *S. aureus* isolates to vancomycin [232]. Another *S. aureus* study trained a SVM and in a blind test, correctly identified 93.3% of methicillin-resistant isolates and 86.7% of methicillin-susceptible isolates [233]. Both of these studies had a relatively small sample size, but illustrated proof of concept.

Recently, a much larger study of AST with multiple model types was performed for *E. coli*, *K. pneumoniae*, *P. aeruginosa*, and *S. aureus* [131]. Training data consisted of thousands of samples taken from their database of MALDI-TOF spectra: the Database of Resistance against Antimicrobials with MALDI-TOF mass Spectrometry (DRIAMS). The database should be publicly available soon, but at this time it is still undergoing review. Models trained on samples from one location could not be used to predict an SIR class for

samples from another location. As their database only contains samples from institutions in Switzerland, classifiers produced with it are currently not globally applicable.

Their machine learning models can produce a classification within one day, with the time constraint being the 24 hours required for culture isolation [131]. One model was created per bacterial species and antimicrobial combination, and they compared logistic regression, random forests, support vector machines, and gradient-boosted decision trees for SIR classification.

Over 90% sensitivity and specificity were achieved for *E. coli* classification; however, in order to prevent false negatives, stringent prediction rejection criteria were used, resulting in rejection percentages of over 40%. This means that the models only produced an approved classification for 60% of tested samples, with the remaining 40% requiring traditional or alternative AST methods.

**Infrared Spectroscopy**

As with MALDI-TOF, the spectra from FTIR can be used with resistance labels to train models for AMR prediction. These models achieve similar accuracies to those made with MALDI-TOF data, as is the time required for prediction after obtaining a pure culture. A study with a small *E. coli* dataset used ANNs to classify isolates as SIR to cephalothin with 83.4% accuracy. When classification was switched to binary susceptible/resistant, the accuracy increased to 92.6% [234]. Subsequent larger *E. coli* studies have used SVMs, obtaining 66% to 91% accuracy, depending on the study and antibiotic [235, 236, 237, 238].

Regions that contributed the most to the accuracy of model prediction could be identified, but their informational content was limited. The wavelength where absorption bands appear in a spectrum are informative of certain component characteristics, such as fatty acids, or carbohydrates [234], but are not always specific. The carbohydrate region 1200-900 $cm^{-1}$ region was found to be important for predicting the presence or absence of ESBL in *K. pneumoniae*, but the significance of this region can only be speculated on [239]. The

relevant ranges may also be small: bands at 1452 $cm^{-1}$ have been associated with nalidixic acid and ofloxacin resistance, and the region at 990–1170 $cm^{-1}$ has been associated with resistance to gentamicin, ceftazidime, nitrofurantoin, and ofloxacin [236]. In general, for *E. coli*, the 1800-900 $cm^{-1}$ region has been shown to be a good target for SVM classification [235, 236].

## 1.5 Metagenomics

The previously described techniques for AMR determination primarily involve the time-consuming step of obtaining a pure culture. A way to bypass this is by sequencing the metagenome, or all of the mixed DNA within a sample. The two primary methods of this are 16S rRNA gene sequencing and shotgun sequencing. The former method is older, examining only the ribosomal RNA gene to identify species [240]. 16S rRNA gene sequencing is used mainly to investigate the composition of microbial communities as a whole [240], as opposed to more specific species characteristics. It can also suffer from poor resolution at the species level of classification [240].

In contrast, the more modern shotgun sequencing aims to sequence everything in the sample, allowing for higher taxonomic resolution [241], and a wider variety of analyses. Illumina technologies are most popular for metagenome sequencing, because of the cost per sequence size, throughput, depth of sequence, and ability to sequence low DNA quantities [242, 243]. Although accurate, the resulting sequence is highly fragmented, making assembly more challenging [244, 245]

As long-read technologies, PacBio and ONT, develop, they are becoming more promising for metagenome sequencing. Compared to Illumina, there are more sequencing errors, greater costs, and a larger quantity of required input DNA [242, 246, 244], but the assembled longer sequences are more complete due to the long reads [244].

Metagenomes sequenced from human blood with the MinION can be analyzed within an hour to detect the presence of Ebola, chikungunya, and hepatitis C viruses [246].

### 1.5.1 Sequence Processing

Though time is saved by skipping culturing steps, it can potentially be lost again in complicated computational processing steps. The volume of bioinformatic tools available for processing and analysis is overwhelming, and one tool may not be the best choice for all sample types [242, 247]. The most common processing required before genome analysis can begin are assembly and taxonomic binning.

The Critical Assessment of Metagenome Interpretation (CAMI) challenge was created to provide a method of metagenomic pipeline benchmarking without bias [247]. The Challenge is run over a set time period, and participating developers may choose whether or not to have results included in the CAMI publication; however, the standard datasets and procedures are also publicly available for anyone to use outside of the official challenge [247, 248]. The first CAMI challenge consisted of only short-read Illumina sequences and began in 2015, with results published in 2017 [247]. The second challenge, CAMI2, includes both short- and long-read data; it began in 2019 and was only recently concluded [249, 250, 248].

#### Assembly

Six short-read assemblers had their results published as part of the first CAMI Challenge [247]. MEGAHIT, Minia, Meraga, A*STAR, Ray Meta and Velour were compared by resulting assembly size, number of unaligned bases, misassemblies, and the ability to resolve closely related genomes. By their metrics, MEGAHIT, Minia, and Meraga were the best performing assemblers, however, the resolution of closely related species from a metagenomic sample was problematic for all of the assemblers. They also found that methods using multiple lengths of $k$-mers performed better than those using a single $k$-mer size [247]. In the first CAMI Challenge, the evaluation of assemblers included the use of MetaQUAST, a version of QUAST (quality assessment tool for genome assemblies) optimized for metagenomics. It assesses assemblies by comparison to reference

genomes. The MetaQUAST developers previously tested their tool on four assemblers, IDBA-UD, SPAdes, Ray Meta, and SOAPdenovo2, with one of their datasets originating from the CAMI Challenge. They concluded that all were equally capable of assembling metagenomes [251].

These studies focused on the assembly quality, but this is not the only deciding factor when choosing a tool; runtime and computational requirements are also important to consider. Another study investigated these additional factors on top of quality for the assemblers MetaSPAdes, IDBA-UD, MEGAHIT, MetaVelvet, Ray Meta, SOAPdenovo2, and Omega [242]. Instead of using CAMI Challenge data, they used two metagenomic samples of differing complexity. They provided four cores to each assembler, with the exception of MetaVelvet which was given eight. Runtimes varied from 0.8 to over 100 hours, and memory usage varied from 5 to 267 GB. Time, RAM, and assembly quality were not only affected by the sample composition and assembler chosen, but also by user specified parameters such as *k*-mer length. They also found that there was a trade-off between assembly size and sensitivity in detecting diversity in the sample. For example, MetaVelvet excelled at resolving diversity, but had relatively low assembly sizes. For this comparison study, MetaSPAdes was considered to be the best, as the assemblies were large while also detecting much of the diversity in the samples.

Fewer benchmarking studies have been conducted for long-read sequencing, as they have only recently been gaining popularity. A recent comparison for ONT sequence assembly was performed using two low-complexity samples containing the same species but at different abundances [245]. They compared three short-read assemblers: metaSPAdes, MEGAHIT, and Minia; and ten long-read assemblers: Canu, HINGE, Miniasm, Unicycler, Pomoxis, Raven, Redbean, Shasta, and two versions of MetaFlye. Evaluating completeness and alignment to reference genomes was performed with QUAST [252], MetaQUAST, and minimap2 [253]. Testing was run on a desktop computer with 8 cores and 64 GB of RAM, much less than the previously mentioned benchmarking studies. MetaSPAdes, one of the

best assemblers for Illumina reads, was unable to run with such limited memory. With ONT platforms, the ability to run analysis on simple desktop computers is desirable, as a big advantage of the MinION is its portability.

Shasta was the fastest, able to produce an assembly in just 5 minutes, but at the cost of having the worst quality. The long-read assemblers MetaFlye, Raven, and Canu were able to produce high quality assemblies from ONT reads. Canu produced some of the highest quality results, but took nearly a day to complete. Raven and both versions of MetaFlye were also able to produce high quality assemblies, but required 3 hours to do so. Although many assemblers were compared, only two mock samples were used. Further studies are needed to assess the performance of ONT data and assemblers for more complex samples.

**Binning**

Along with assembly, taxonomic classification of metagenomic samples to determine composition and/or abundance is commonly performed. There are two methods of binning. In "genome binning", sequences are grouped independently of taxonomic labels, instead using similarity to one another for sorting. Meanwhile, "Taxonomic binning" groups sequences into taxonomic labels by comparing them to reference sequences.

The first CAMI Challenge compared five genome binners, MyCC, MaxBin 2.0, MetaBAT, Metawatt 3.5, CONCOCT, and four taxonomic binners, PhyloPythiaS+, taxator-tk, MEGAN6, and Kraken [247]. They applied the programs to three datasets of differing complexity, and examined the purity and completeness of the genomes in the resulting bins. For the genome binners, the average purity of the resulting genomes varied from 70% to 90%, while the average completeness varied from 34% to 80%. Sample complexity affected performance, with most programs failing as complexity increased. MaxBin 2.0 was able to achieve 96% purity on the most complex dataset, but its completeness was 48%, similar to the other four tools. Metawatt 3.5 was close behind with 90% purity and 49% completeness. Both the genome binners and the taxonomic binners struggled with resolving closely

related strains into different bins. Taxonomic binners performed well at higher taxonomic levels, but less well at the genus and species levels due to the small number of reference genomes available for the more specific ranks. For the low-complexity dataset, the top two performing tools were PhyloPythiaS+ with over 75% completeness and accuracy up to the family level, and Kraken with over 50% completeness and accuracy up to the family level. The performance of Kraken, MEGAN, and taxator-tk improved for the medium-complexity dataset.

A package called Assessment of Metagenome Binners (AMBER) was developed after the CAMI Challenge, to incorporate their evaluation methods and provide a streamlined method to compare binners [254]. They evaluated 11 binning tools using a low- and a high-complexity CAMI dataset. Five tools examined were also used in the first CAMI Challenge: CONCOCT, MyCC, MaxBin 2.0.2, MetaBAT, and Metawatt 3.5. They also tested three other tools, BinSanity, BinSanity-wf, COCACOLA, and DAS Tool 1.1, and newer versions of two tools, MaxBin 2.2.4 and MetaBAT 2.11.2. DAS Tool is unique in that it takes the results of other binners as input; for this study all of the results except those from COCACOLA were used as input.

The newer version of MetaBAT performed much better than the version tested 3 years earlier. On the high-complexity dataset, it had bins with the highest purity and completeness compared to the other standalone binners. This highlights that when choosing a tool, it is important to consider how it is being maintained, and the dates of benchmarking publications. DAS Tool was able to achieve higher performance than MetaBAT on the high-complexity dataset, but was on par with other binners for the low complexity dataset [254]. Although DAS Tool may provide higher purity and completeness than using a single binner, if each program could take up to an hour to run [255] and could not be run in parallel, then the extra time required for analysis may not be worth the small increase in performance. Additionally, such an approach would introduce more variables into the choice of binners; decisions would include which and how many binners to include, as well as the numerous

settings available for each binner.

A 2019 comparison [255] found that the 20 tools they compared all performed similarly, but their evaluation metrics were not the same as the prior studies, so the results are not directly comparable despite additional tests with CAMI datasets. They found that the group of classifiers using DNA criteria performed better than those using proteins, as the latter lacked non-coding regions in their databases. As with the prior studies, the classifiers struggled to classify sequences at the species level. They also measured the runtime and memory requirements of the programs. A tool may be fast but require an amount of memory beyond a standard computer; for example, KrakenUniq completed in just 1 minute, but required 200 Gb of memory. The programs varied in memory usage from under 1 Gb to 200 Gb, and in runtime from under 1 minute to 40 minutes. Computer resources may affect the choice of binner, but there are many options available for even basic desktops. They concluded that presently there is no single best tool and that improvements to increase performance, especially at lower taxonomic ranks, are needed for the future.

As with assembly, there is a lack of large benchmarking studies on the binning of long-read data. A 2019 study [256] compared the ONT and Illumina platforms in the analysis of cattle rumen samples, using MEGAN for taxonomic binning. The ONT data allowed for better genus and species resolution compared to Illumina. When detecting diversity at lower ranks is important, such as for pathogen detection, ONT may be a better platform choice; alternatively, the two platforms may be used in conjunction through hybrid assemblies, although cost may be prohibitive [256]. The tool they used, MEGAN, also has a derivative tool called MEGAN-LR which was optimized for long read sequencing, including the PacBio platform as well as ONT [257].

A binning workflow, “What’s In My Pot?” (WIMP), was developed specifically for the ONT platform [258]. This workflow is included in their online analysis program EPI2ME [259]. WIMP is an attractive option for users with a limited bioinformatics background, as it has a user-friendly graphical interface, and does not require use of the com-

mand line like the majority of other programs available. It uses the Centrifuge program, which may also be used for short-read data [260, 255], and NCBI databases to perform classifications. At the time of release, it took approximately 3.5 hours to complete a run [258], which is over three times longer than the longest runtime for Illumina data in the study by Ye *et al.* [255]. However, WIMP can begin data analysis while sequencing is still underway [258]. The MinION Detection Software (MINDS) is a similar workflow to WIMP, and also uses Centrifuge [261]. It was developed to remove the requirement of an internet connection, making it well suited for the portability of the MinION.

### 1.5.2 Detecting Antimicrobial Resistant Determinants in Metagenomes

#### Database Methods

The previously mentioned database methods for detecting AMR determinants in WGS data can also be applied to metagenomes from any platform. Some studies also merge databases, such as ResFinder, CARD, and ARG-ANNOT, into one master database, though differences in nomenclature make this challenging [90]. One of the programs available through the ONT online EPI2ME portal is the AMRA workflow, which uses CARD for real-time detection of AMR genes [262].

A database called MEGARes, and the pipeline designed to interface with it called AMRPlusPlus, were created to provide a more optimized database and analysis pipeline for processing large metagenomic samples [263]. The developers argue that existing AMR databases, originally created for use primarily with single genomes, are not optimized for automation and metagenomic applications. They compiled the information of the databases ResFinder, ARG-ANNOT, CARD, and the National Center for Biotechnology Information (NCBI) Lahey Clinic beta-lactamase archive. Along with automated merging of identical entries across databases, manual curation was used to ensure the quality of the database. MEGARes uses strictly hierarchical annotation, in contrast to CARD, whose entries may belong to multiple groups and thus cause counting errors [263]. These errors may not matter

for individual genome analysis, but may cause skewed results for complex samples [263]. MEGARes is well suited to analyze the overall resistome, but for more detailed SNP analysis, supplementary tools are recommended.

The Structured Antibiotic Resistance Gene (SARG) Database and associated pipeline ARGs-OAP (antibiotic resistance genes online analysis pipeline) were published the same year as MEGARes [264]. Similarly to MEGARes, they aimed to make a database optimized for metagenomic analysis. They used the CARD and ARDB to form their database, performed similar size reduction methods, and created a hierarchical annotation.

A different database approach is to use a database of genomes of the organism(s) of interest to establish phenotype by relatedness. This method has been used to predict a susceptible or resistant classification for *S. pneumoniae* and *N. gonorrhoeae* [265]. In the program called Resistance-Associated Sequence Elements (RASE), *k*-mers are used to rapidly match sequences to a database of strains [265]. They demonstrated that RASE can function in real-time, producing predictions within 10 minutes of starting to sequence clinical metagenomic samples with the ONT platform [265]. For this kind of analysis, a database for every organism of interest is required. As such, this method is better suited for targeted clinical diagnostics, where only a few specific organisms may need to be profiled, as opposed to broader surveillance.

The database methods readily identify true positives, but can struggle with false negatives, especially when identity cutoffs are high [264, 266]. Identity cutoff is a debated user setting; one study suggests 80% for metagenomic samples to avoid false positives [266], while the ARGs-OAP pipeline recommends 60% to decrease the number of missed ARGs [264]. A study which applied ResFinder for examining bacteria in permafrost used a cutoff of just 50% [267].

### Machine Learning

A pipeline called DeepARG was created to decrease the number of false negatives, an identified problem with database methods [266]. A consolidated non-redundant database was created from the CARD, ARDB (antibiotic resistance genes database), and UniProt (universal protein resource) databases. Their machine learning approach was based on similarity, where sequences are assigned distances based upon how similar they are to each known ARG. A dissimilarity distance matrix was created by comparing the UniProt genes to those in CARD and ARDB. A total of 30% of the reads were withheld for validation while the remaining were used for training. Separate neural network models were created to accept long (DeepARG-LS) and short sequences (DeepARG-SS). In the full DeepARG pipeline, when a sequence is submitted by a user, it is first filtered to remove non-ARG content before being passed to the trained models which give a prediction.

For DeepARG-SS, the average precision was 97%, and for DeepARG-LS it was 99%. Models performed poorly for categories, such as triclosan, where only a few known ARGs were in the database. They additionally tested the models on the ARGs contained in the MEGARes database. The only modification to the MEGARes database was the removal of SNPs. The models performed well, with an average precision of 94%. This validation method is problematic, as both MEGARes and the DeepARG training sets have CARD content, so the validation set includes data that the models have already seen.

A set of 76 candidate genes conferring resistance were run through the models, and 85% were correctly classified. The models were able to classify these genes, while other databases could not, due to the lower sequence similarity to beta-lactamase genes. However, this validation is limited, as they were all from a single group: metallo beta-lactamase genes. This group had a large training dataset; a smaller group may not have the same ability to recognize these novel genes because of a lack of diversity.

A problem with the validation in DeepARG is that they only used a held out set for validation; further validation of sequences external to ARG databases should also be used.

Their models are limited by the quantity and quality of the ARGS in the source databases, with some categories suffering more than others. Additionally, as the models were trained on ARG databases, they lack the capability to identify novel ARGs.

### 1.5.3 Metagenomics for AMR Surveillance

One of the most important applications of metagenomics is in antimicrobial resistance surveillance. A resistome consists of all of the antimicrobial resistant genes in a metagenomic sample, regardless of whether the carrying organism is pathogenic or not. Detecting resistance markers in non-pathogenic strains is crucial, as they could be transmitted to strains which pose a higher risk to human or animal health [92, 91]. Monitoring changes in this resistome can be used to detect the emergence of resistant pathogens and allow for the efficient implementation or alteration of methods to prevent spread.

Metagenomic shotgun sequencing has been used to examine resistome changes in beef production systems [90, 268]. It was found that antimicrobial use, surrounding environment, and prevention methods can impact the microbial community composition and resistome diversity [90, 268].

## 1.6 Thesis Overview

As antimicrobial resistant bacteria put increasing strain on the healthcare system, it is desirable to find ways to detect threats before outbreaks occur. Traditional methods of resistance profiling are time-consuming as they involve culturing in the laboratory. A more rapid method of resistance profiling would enable faster response times to prevent outbreaks, and more tailored treatment options. WGS is becoming cheaper and more widespread; it is being integrated into surveillance programs. WGS data can be used alongside databases to determine resistances; however, databases rely on constant curation of known markers, and are primarily focused on genes (as opposed to intergenic regions). As resistance is constantly changing and developing, there needs to be a way to discover new markers.

ML analysis methods don't reply on *a priori* information, so they can be used to identify potentially novel markers of resistance as well as determine resistance profiles.

In this work, ML classification methods were used for the prediction of antimicrobial resistance in *E. coli*. Additionally, the trained ML models were used to identify genomic markers of resistance, including potentially novel features. The whole genome sequences were used by way of *k*-mers, so that all regions (both genes and intergenic regions) were considered for classification.

# Chapter 2

# Application of artificial intelligence to the *in silico* assessment of antimicrobial resistance and risks to human and animal health presented by priority enteric bacterial pathogens

## 2.1 Preface

Chapter 2 has been published in a peer-reviewed journal, the Canada Communicable Disease Report: Steinkey R, Moat J, Gannon V, Zovoilis A, Laing C. Application of artificial intelligence to the *in silico* assessment of antimicrobial resistance and risks to human and animal health presented by priority enteric bacterial pathogens. Can Commun Dis Rep. 2020 Jun 4;46(6):180-185. doi: 10.14745/ccdr.v46i06a05.

For this project, I was responsible for the *Escherichia coli* components, which include data curation, training of machine learning models, data visualization, and manuscript writing. I additionally contributed to the *Salmonella* components alongside Rylan Steinkey; I contributed to data curation, training of machine learning models, data visualization, and manuscript writing. Dr. Chad Laing supervised the project and contributed to its design and development, and to the writing of the manuscript. Dr. Victor Gannon and Dr. Athanasios Zovoilis contributed to the project design, development, review, and writing of the final manuscript.

## 2.2 Abstract

Each year, approximately one in eight Canadians are affected by foodborne illness, either through outbreaks or sporadic illness, with animals being the major reservoir for the pathogens. Whole genome sequence analyses are now routinely implemented by public and animal health laboratories to define epidemiological disease clusters and to identify potential sources of infection. Similarly, a number of bioinformatics tools can be used to identify virulence and antimicrobial resistance (AMR) determinants in the genomes of pathogenic strains.

Many important clinical and phenotypic characteristics of these pathogens can now be predicted using machine learning algorithms applied to whole genome sequence data. In this overview, we compare the ability of support vector machines, gradient-boosted decision trees and artificial neural networks to predict the levels of AMR within *Salmonella enterica* and extended-spectrum β-lactamase (ESBL) producing *Escherichia coli*. We show that minimum inhibitory concentrations (MIC) for each of 13 antimicrobials for *S. enterica* strains can be accurately determined, and that ESBL-producing *E. coli* strains can be accurately classified as susceptible, intermediate or resistant for each of seven antimicrobials.

In addition to AMR and bacterial populations of greatest risk to human health, artificial intelligence algorithms hold promise as tools to predict other clinically and epidemiologically important phenotypes of enteric pathogens.

## 2.3 Introduction

Every year, about one in eight Canadians will be affected by a foodborne illness, resulting in an average of 11,600 hospitalizations and 238 deaths nationwide [269]. Animals are often the reservoir for major bacterial pathogens such as *Salmonella enterica* and *Escherichia coli*. These pathogens are associated with both sporadic cases and outbreaks of foodborne disease. Antimicrobial resistance (AMR) among these organisms is a growing concern, with treatment being more difficult and expensive. For example, extended-

spectrum β-lactamase (ESBL) producing *E. coli* are multidrug resistant, with treatment costs up to three times that of non-ESBL-producing *E. coli* [270].

National and provincial public health agencies are very effective at identifying sources and halting exposure to pathogens. Historically, AMR determination has been performed in a wet lab setting [105, 271]. Two of the most commonly used diagnostic methods are diffusion and dilution tests. Diffusion methods, such as the Kirby–Bauer method, require growing a bacterial lawn in either a disk of known concentration of antimicrobials or a strip with a gradient of concentrations of antimicrobials; the zone of growth inhibition around the antimicrobial is compared with a standard to determine the resistance of the bacteria [105]. Dilution methods involve liquid cultures in serial dilution of each antimicrobial, where growth of the organism is used to determine the minimum inhibitory concentration (MIC) [105, 271].

These methods are time consuming because they rely on the growth of bacteria, and expensive because they require trained personnel and specialized equipment to carry out.

Whole genome sequence (WGS) analyses have become integral to public health work flows. *In silico* tests have largely replaced many costly and time-consuming wet lab tests in outbreak response and routine surveillance [272, 273, 274]. Artificial intelligence is being increasingly used to analyse these datasets.

Artificial intelligence involves training machines to make predictions based on large amounts of data. It has been used in fields as disparate as handwriting recognition [275] and autonomous weapons systems [276].

Supervised machine learning (ML) better describes the application of artificial intelligence to the prediction of bacterial phenotypes based on WGS data. ML algorithms are trained on known data ("features") and subsequently predict or classify unknown data using the trained models. In general, data used for ML training are application specific and can include images or information about weather or outbreaks of infectious disease. Biological data, and in particular WGS data from populations of organisms, provide an extremely

large number of features for training ML models and predicting phenotypes of interest. Use of these algorithms in infectious disease research has not yet been fully exploited but holds significant promise.

ML algorithms have been used to predict important phenotypes such as AMR [277, 217] and to determine if different groups of pathogens from the same species pose different risks to human health [201, 207, 205]. The ability to predict important bacterial phenotypes based solely on WGS data would be of enormous benefit to both Canadian public health and the animal agriculture industry.

In this study, we trained three ML models on WGS data to predict the levels of resistance to 13 antimicrobials in *S. enterica* isolates and to classify ESBL-producing *E. coli* strains as susceptible, intermediate or resistant (SIR) to seven antimicrobials.

## 2.4 Methods

*S. enterica* WGS was collected from the National Center for Biotechnology Information GenBank. These 5,853 sequences were primarily isolated within North America between 2002 and 2017; the data included 63 serotypes with at least five members, along with phenotypic MICs for 13 antimicrobials [278]. WGSs were decomposed into sequence substrings, called *k*-mers, of length 11, and their occurrences were counted using Jellyfish [279]. To limit the selection of features to those most associated with the phenotype being examined, we used an ANOVA F-value, keeping the top 1,000 *k*-mers most associated with each antimicrobial agent prior to model training. This feature selection allows the model to focus on statistically important *k*-mers, which can improve accuracy and saves substantial amounts of time and computing resources.

We implemented gradient-boosted decision trees using XGBoost [280] and support vector machines using SciKit-learn [281]. Data analyses were conducted using five-fold cross-validation where 80% of the data was used to train a model and the remaining 20% was withheld to evaluate model performance. This was repeated five times, with each 20%

being used once for evaluating performance. An average of the accuracy for the five evaluations was calculated for each experimental replicate. Ten separate experimental replicates with random assignment of genomes to each fold were performed, with the total model accuracy and standard deviation calculated from these.

Artificial neural networks were implemented using Keras [282] with a TensorFlow [283] backend and hyperparameter optimizations conducted with Hyperas [284]. The five-fold cross-validation for the neural network consisted of a 60-20-20 split for training, hyperparameter optimization and testing, respectively, for each fold. Early stopping mechanisms were used to prevent over-fitting by monitoring diminishing or negative returns with successive training epochs. In addition, a random selection of nodes in the network and their connections were removed via dropout to prevent over-fitting or co-adaptation [285].

As shown in Figure 2.1, MICs were predicted within one dilution with an accuracy of 97.88% ($\pm$ 1.13) using XGBoost, 97.48% ($\pm$ 1.20) using support vector machines and 97.16% ($\pm$ 1.48) using artificial neural networks. XGBoost classifiers averaged a major error and major error rate of 0.19% ($\pm$ 0.19) and 0.71% ($\pm$ 0.60), respectively. To prevent inflating model accuracies, co-trimoxazole, ciprofloxacin and ceftriaxone, which had low MIC class diversity, were removed from these averages. XGBoost classifiers trained to predict MICs for a single antimicrobial used eight cores (Intel Xeon Gold 6154 CPU), had a mean training time of 15 minutes and 12 seconds, and peaked at 84.74 GB of random access memory (RAM).

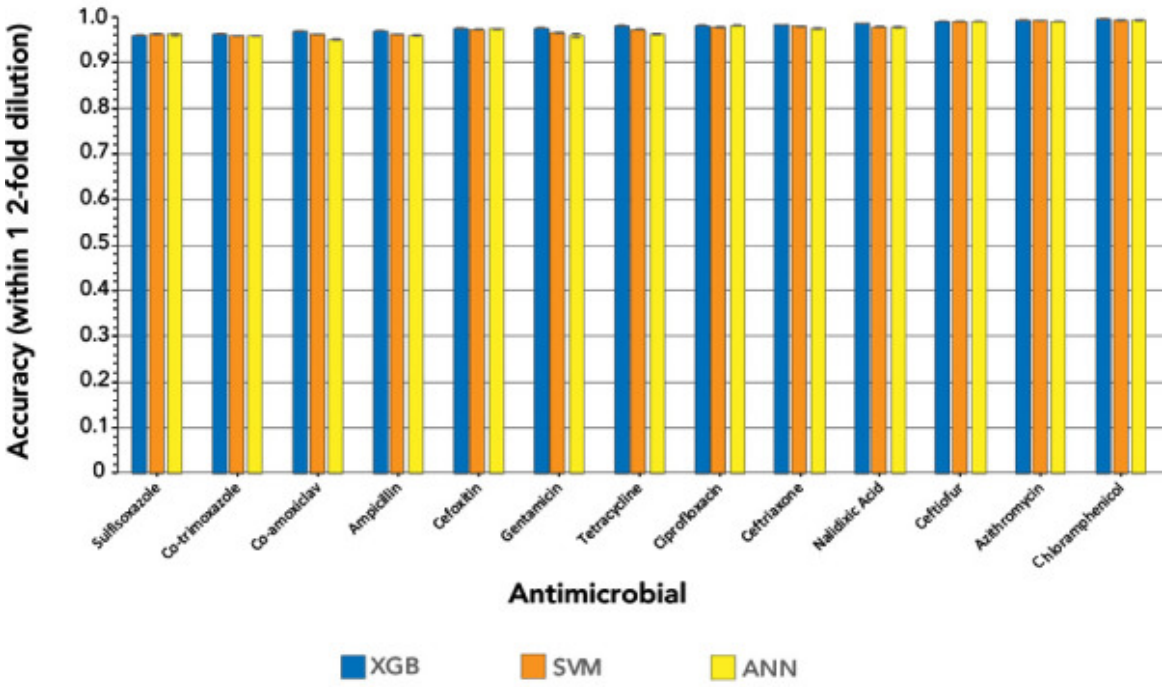



Figure 2.1: Accuracies within one two-fold dilution for three machine learning models trained on the top 1,000 11-mers and used to predict minimum inhibitory concentrations for 13 *Salmonella enterica* antimicrobials. Abbreviations: ANN, artificial neural network; SVM, support vector machine; XGB, XGBoost

We also examined a set of 2,413 *E. coli* sequences containing ESBL producers, but no MIC data were available for these strains. Instead, they were classified as SIR for seven antimicrobials. The set included bovine, clinical and environmental samples isolated between 1970 and 2017 in Canada, Thailand and the United Kingdom [217, 286, 287]. We analyzed the sequences with the *k*-mer approach described above and used them to train models to classify isolates as SIR for each antimicrobial. The average accuracies of the models across the seven antimicrobials were 89.18% ($\pm$ 5.44) for XGBoost, 89.25% ($\pm$ 4.43) for support vector machines and 89.18% ($\pm$ 5.20) for artificial neural networks (Figure 2.2).

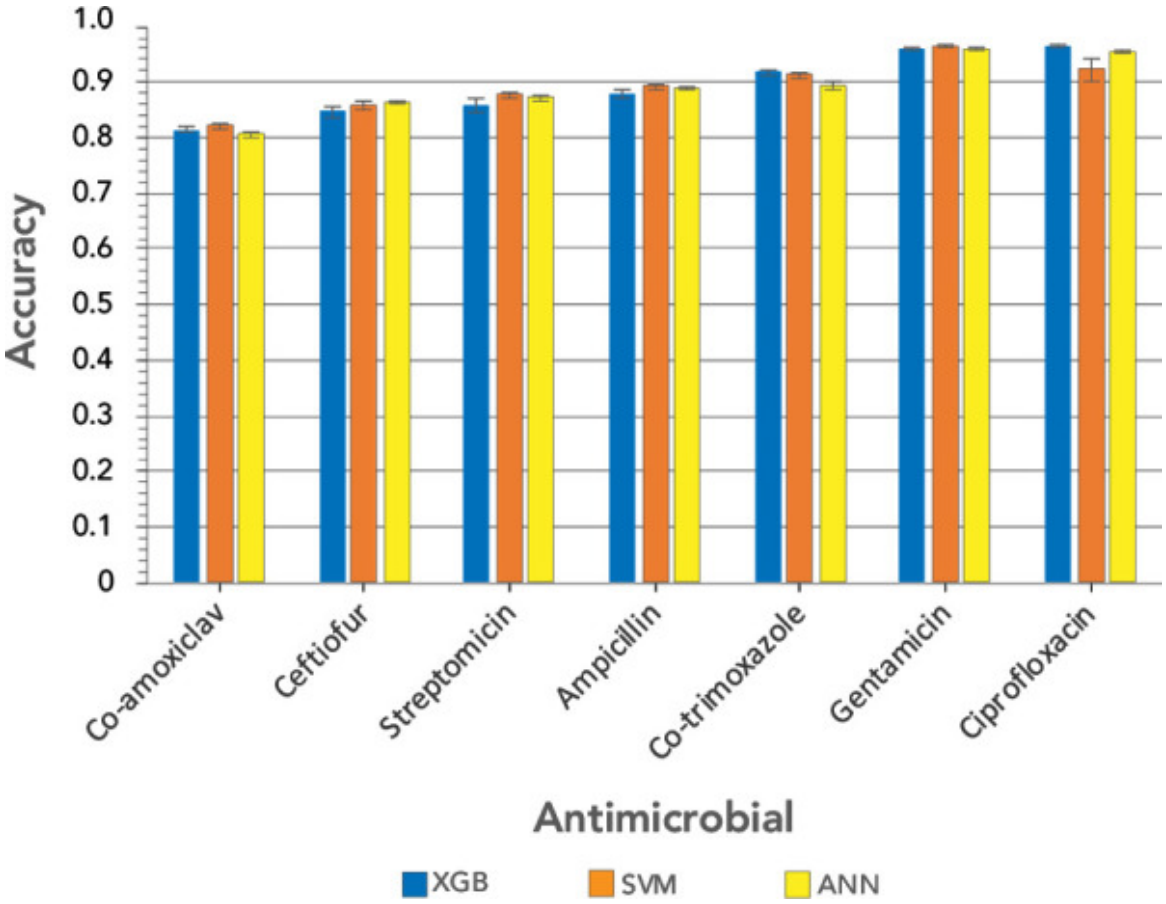



Figure 2.2: Accuracies of three machine learning models trained on the top 1,000 11-mers, and used to predict susceptible, intermediate and resistant classifications for seven *Escherichia coli* antimicrobials. Abbreviations: ANN, artificial neural network; SVM, support vector machine; XGB, XGBoost

## 2.5 Discussion

As we have shown, the ML methods we employed did not rely on specific reference genomes, or *a priori* knowledge of the mechanisms of resistance, but on the classification of organisms into broad phenotypic groups. It is the ML models that identify the underlying genomic differences that are most associated with the phenotype. This has the double benefit of not requiring mechanistic knowledge and has the potential for identifying novel genomic determinants of the phenotype under study. These novel features extracted from the models have enormous potential benefit: as in the case of AMR, they can be used to grow established public databases of resistance mechanisms, and they can be used as

potential targets for rapid diagnostics in subsequent *in silico* or wet lab assays.

ML models can rapidly and accurately predict AMR using WGS data, from SIR classification to quantitative MIC values. For AMR predictions, XGBoost models were shown to train faster, use less memory and be more accurate than deep-learning methods. In addition, XGBoost and support vector machine models can be used to determine the specific regions of the genome that are most predictive of a phenotype. This is very difficult with the "black box" implementation of a neural network; however, artificial neural networks still excel in complicated network modelling and therefore should not be excluded from future studies in genomics.

AMR data typically suffer from substantial class imbalance, which can result in high accuracy models that are of no value, such as the case of co-trimoxazole in our *Salmonella* data, where more than 95% of the samples were within one dilution of each other, resulting in a model capable of 95% accuracy without learning anything from the underlying data.

Nguyen et al. [277] trained XGBoost regressors on a dataset containing 4,500 non-typhoidal *S. enterica* whole genome sequences (from a larger dataset of 5,278 samples, of which 4,595 were also in our dataset). These models had a cross-validation accuracy of 95% for the same 10 antimicrobials included in our current study. Nguyen et al. [277] used a single regressor trained on all 15 antimicrobials at once, which took 51 hours to train and peaked at 1,184 GB of memory on 170 cores (Intel Xeon E5-4669v4 CPU) [277]. The XGBoost classifiers trained in our current study improved upon these training times as well as memory usage and accuracy. The XGBoost classifiers did this by creating per-antimicrobial models and initially selecting only the 1,000 most statistically important features. To better compare the accuracies of these models, an independent dataset should be used instead of relying on the reported cross-validation accuracies.

The *E. coli* dataset included 1,935 isolates from a previous study by Moradigaravand et al. [217]. Their methods required the isolation year for each sequence and data preprocessing in the form of pan-genome determination and population structure calculation [217]. In

contrast, our methods required only the genome sequence paired with laboratory-determined resistance phenotype, which allows classification as well as identification of novel regions not currently known to be associated with AMR. The regions could be used for subsequent *in silico* or wet lab diagnostic tests.

While broader classifications, such as SIR, are common for laboratory diagnostics, and useful for establishing treatment guidelines for a bacterial infection, the breakpoint criteria for these categories are established by committees, with some disparity between regions. The prediction of quantified values in the form of MICs will be of most use in future, even if they are subsequently used for classifying bacteria into broader categories such as SIR.

Though the results of these studies are encouraging, over-interpretation of results is a problem with genomic data due to the high number of features used to make predictions relative to the smaller sample size of the number of genomes. This can lead to over-fitting of data and poor performance of models, both of which we have tried to address in the methods of this study [208].

Use of ML has proved successful for AMR prediction in other pathogens, including *Mycobacterium tuberculosis*, where new resistant genetic signatures were identified [222]. ML has also proved useful in the identification of novel antimicrobial compounds, which has historically been fraught with high failure rates in pharmaceutical companies [202].

ML research on *S. typhimurium* found that more than 80% of host source could be attributed using protein variants. This result was obtained using support vector machine (SVM), artificial neural networks and Random Forest models [209]. What is particularly interesting from this study is the overlap between the animal reservoir and human cases. This indicates that not all isolates of a particular pathogen represent the same disease risk and suggests that more specific points of control could limit human infection. In addition, as more than 60% of human pathogens are of zoonotic origin, ML holds promise for identifying emerging pathogens by analyses of host adaptation of current animal pathogens [288].

Despite the proven usefulness of ML, bacteria are constantly evolving, and so our mod-

els, as they are only as good as the data they are trained on. The power of these techniques must be tempered by their judicious use. In addition, class and species-specific models are still required to generate meaningful results, for example, one model per drug per species for predicting AMR [289].

It should be noted that ML does not always accurately capture complex interactions and that improved modelling alone cannot compensate for sampling bias or an incomplete or error-prone dataset.

## 2.6 Conclusion

As demonstrated in this overview, artificial intelligence has already improved infectious disease identification and characterization, the benefits of which will affect public health and animal health laboratories around the world. For example, genomic regions identified as predictive for specific AMR classes could be used for rapid downstream identification and classification, including *in silico* pipelines and wet lab applications such as polymerase chain reaction.

The near-future promises exciting developments, such as using ML to identify bacteriophages that lyse specific groups of pathogenic bacteria, enabling phage therapy in place of traditional antimicrobials [200]. Lastly, "whole phenotype" characterization, with the ability to predict integral membrane protein expression, is becoming more likely [290]; and biofilm formation [291].

Despite this, the size of the datasets required to effectively train ML models mean that desktop computers are often incapable of analyzing the data. Those without access to the necessary resources must instead use analytical approaches that reduce the computational burden [292]. Fittingly, the use of ML itself has led to an increase in speed of mechanistic models, in some cases over four orders of magnitude [293].

We are just at the beginning of the coupling of vast amounts of genomic data and artificial intelligence, with the promise of new discoveries that will improve most aspects of

animal and human health from the burden of enteric bacterial pathogens.

# Chapter 3

# Machine Learning Methods for the Prediction of AMR in *E. coli*

## 3.1 Preface

For Chapter 3, I was responsible for all components, which include data collection and curation, training of machine learning models, data visualization, and writing. Dr. Chad Laing and Dr. Athanasios Zovoilis supervised the project. Dr. Chad Laing, Dr. Athanasios Zovoilis, Dr. Rahat Zaheer, and Dr. Tim McAllister contributed to the design, development, and review of the project.

## 3.2 Abstract

Antimicrobial resistant strains of pathogenic *Escherichia coli* are a burden on the health-care system, causing longer hospital stays and increased treatment costs compared to non-resistant strains [48, 294, 295, 296]. The proportion of *E. coli* infections in Canada caused by the resistant strains producing extended-spectrum beta-lactamase rose from 3.4% in 2007 to 11.1% in 2017 [48]. Rapid detection of resistant strains would be beneficial for both surveillance and clinical settings. Faster response times would make outbreak containment and prevention easier. The speed would allow for simultaneous testing of more antimicrobials than traditional methods, allowing for the administration of the most effective treatment available. Whole genome sequencing (WGS) is decreasing in cost and becoming more available to labs: as of 2021, the basic MinION costs only $1000. When paired with *in silico* methods for diagnostics, the MinION is faster and cheaper than tradi-

tional wet-lab methods.

In this study, WGS was paired with machine learning methods and used for rapid prediction of resistance profiles. A total of 4300 *E. coli* isolates with whole genome sequence data and lab acquired resistance classifications were collected. Their sequences were decomposed into *k*-mers of length 11 and 25, which were used to train three types of machine learning models for the prediction of susceptible, intermediate, or resistant (SIR) classifications to 36 different antimicrobials. For 11-mers, gradient boosted decision trees (XGB), support vector machines (SVM), and artificial neural networks (ANN) achieved an average accuracy of 93.2%, 92.3%, and 92.3% respectively, and average F1-score of 86.0%, 83.8%, and 84.9% respectively. For 25-mers, the average accuracies for XGB, SVM, and ANN models were 87.8%, 87.9%, and 68.9% respectively, and F1-scores were 75.0%, 73.6%, and 59.5% respectively.

## 3.3 Introduction

Pathogenic strains of *Escherichia coli* are one of the leading causes of gastroenteritis in humans; they are also responsible for many other diseases, including meningitis, septicemia, and urinary tract infections [2]. Foodborne outbreaks are commonly due to contaminated manures used on produce [94].

Shiga toxin-producing *E. coli* (STEC) are primary targets of many surveillance programs, as they typically cause the most severe forms of *E. coli* infection. In 2018, there were 1.15 cases of O157 infection per 100,000 people in Canada, a rate that has been consistent since 2010 [31]. Non-O157 strains are inconsistently reported to surveillance programs, but the incidence of non-O157 infections has been increasing and surpassing the number of O157 infections, with a rate of 1.42 cases per 100,000 people in 2018 [31]. In 2011, a foodborne outbreak in Germany was caused by *E. coli* serotype O104:H4; the outbreak resulted in 900 cases of haemolytic uremic syndrome and 54 deaths [34, 21].

*E. coli* infections are becoming increasingly problematic to treat due to the rise of AMR.

As most resistance genes within *E. coli* are carried on mobile genetic elements, resistance can arise and spread rapidly within a population [60, 61]. AMR causes constant strain on the healthcare system, with ESBL infections costing up to three times more than non-ESBL infections [37]. In 2017, 11.1% of *E. coli* infections in Canada were by resistant ESBL producing strains [48].

As WGS is being integrated into diagnostics and surveillance programs, it is being investigated as replacement for traditional time-consuming wet-lab AST methods. WGS can be used to train ML models for the prediction of resistance profiles. They are additionally capable of discovering previously unknown associations of genomic regions with AMR. The features responsible for classification can be extracted from trained models, along with their level of contribution. This enables the discovery of novel markers and mechanisms and the identification of associations of known regions and AMR. An advantage over AMR databases is that the ML models can identify markers that are associated with susceptibility, in addition to those associated with resistance. Once trained, models can make predictions on unseen sequences in less than a minute on a standard desktop computer. The time gate for this method is the sequencing step; however, the Nanopore system has previously been used to perform sequencing and diagnostics within 24 hours [297].

In this study, the aim was to predict the SIR classification of *E. coli* whole genome sequences for 36 different antimicrobials. We trained support vector machines (SVMs), artificial neural networks (ANNs), and gradient boosted decision tree ensembles, using *k*-mers as input features. Additionally, external datasets were used to validate models for 16 of the antimicrobials. Models trained for SIR prediction could be applied to diagnostics or surveillance, to improve testing speed and decrease the requirement of highly trained personnel.

## 3.4 Methods

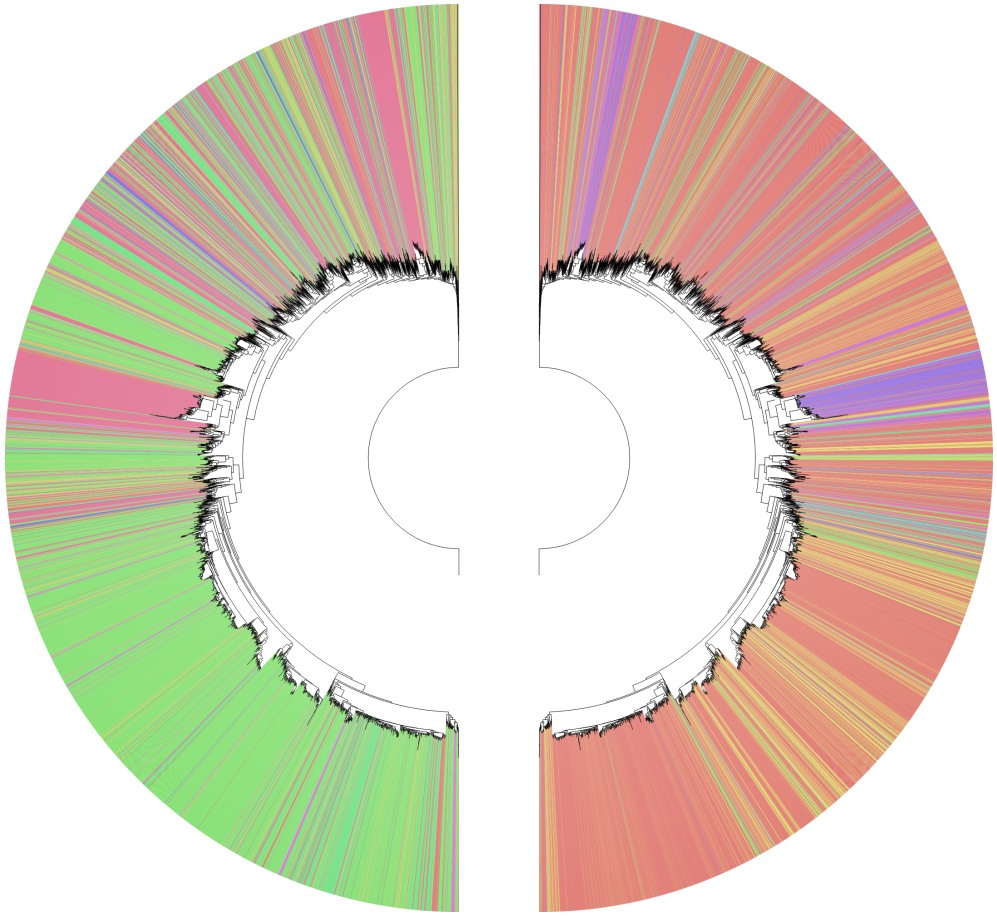
### 3.4.1 Data Collection

#### Whole Genome Sequences

The dataset of 4300 *E. coli* whole genome sequences consisted of public isolates and those provided for this study. The study-specific data consisted of 636 Canadian isolates provided by Agriculture and Agri-Food Canada (AAFC) and the Canadian Integrated Program for Antimicrobial Resistance Surveillance (CIPARS). The public data consisted of 3664 samples obtained from the European Nucleotide Archive (ENA), National Center for Biotechnology Information (NCBI), and Pathosystems Resource Integration Center (PATRIC). The phylogenetic relationships between the isolates can be seen in Figure 3.1. Complete genome information for all study data, including source, resistance, SIR, host, origin, country, and date of isolation is available in Supplementary Table 1.

#### Antimicrobial Resistance Data

Each genome was paired with metadata including their laboratory based classification of susceptible, intermediate or resistant (SIR) for up to 36 antimicrobials (Table 3.1). The 36 antimicrobials were: co-amoxiclav (AMC), ampicillin (AMP), amoxicillin (AMX), aztreonam (AZT), cephalothin (CET), cefazolin (CFZ), chloramphenicol (CHL), ciprofloxacin (CIP), cefepime (CPM), ceftriaxone (CRO), colistin (CST), cefotaxime (CTX), ceftazidime (CTZ), cefuroxime axetil (CXA), cefuroxime (CXM), enrofloxacin (EFX), ertapenem (ETP), florfenicol (FFC), cefoxitin (FOX), gentamicin (GEN), imipenem (IMP), kanamycin (KAN), levofloxacin (LVX), meropenem (MER), nalidixic acid (NAL), nitrofurantoin (NIT), ampicillin-sulbactam (SAM), sulfisoxazole (SOX), streptomycin (STR), co-trimoxazole (SXT), tobramycin (TBM), tetracycline (TET), tigecycline (TGC), ceftiofur (TIO), trimethoprim (TMP), and piperacillin/tazobactam (TZP).



**A** **B**

Figure 3.1: Phylogenetic tree of 4300 *E. coli* isolates with *Salmonella* as an outgroup; generated with IQ-TREE (v1.6.12). A) Colouration is by country of origin; data was from 16 different countries, with the most frequent being the United Kingdom and Ireland (N=2159), Canada (N=774), and USA (N=498). B) Colouration is by serotype (typed by ECTyper); the dataset had 162 unique O groups and 48 unique H types; the most frequent serotypes were O25:H4 (N=544), O6:H1 (N=252), and O1:H7 (N=152). Trees with full legends are available in Supplementary Figure 1

Table 3.1: The Susceptible (S), Intermediate (I), and Resistant (R) classification for the thirty-six antimicrobials among the 4300 isolates of this study. Each genome was paired with an SIR classification for between one and 36 antimicrobials.

| Code | Antimicrobial | Total Number of Isolates | S | I | R |
|---|---|---|---|---|---|
| AMP | ampicillin | 2937 | 1184 | | 1753 |
| AMX | amoxicillin | 1297 | 516 | 30 | 751 |
| AZT | aztreonam | 748 | 453 | | 295 |
| CET | cephalothin | 1011 | 372 | 188 | 451 |
| CFZ | cefazolin | 548 | 116 | 37 | 395 |
| CHL | chloramphenicol | 893 | 678 | | 215 |
| CIP | ciprofloxacin | 3258 | 2317 | 86 | 855 |
| CPM | cefepime | 1052 | 798 | 43 | 211 |
| CRO | ceftriaxone | 1262 | 484 | | 778 |
| CST | colistin | 226 | 33 | | 193 |
| CTX | cefotaxime | 2393 | 1637 | | 756 |
| CTZ | ceftazidime | 2792 | 1929 | 177 | 686 |
| CXA | cefuroximeaxetil | 414 | 352 | | 62 |
| CXM | cefuroxime | 1955 | 1429 | | 526 |
| EFX | enrofloxacin | 157 | 76 | | 81 |
| ETP | ertapenem | 817 | 752 | | 65 |
| FFC | florfenicol | 157 | 83 | | 74 |
| FOX | cefoxitin | 1621 | 957 | 59 | 605 |
| GEN | gentamicin | 3259 | 2507 | 42 | 710 |
| IMP | imipenem | 1583 | 1533 | | 50 |
| KAN | kanamycin | 568 | 495 | | 73 |
| LVX | levofloxacin | 401 | 108 | | 293 |
| MER | meropenem | 1133 | 993 | | 140 |
| NAL | nalidixic acid | 835 | 703 | | 132 |
| NIT | nitrofurantoin | 294 | 232 | 29 | 33 |
| SAM | ampicillin/sulbactam | 190 | 78 | | 112 |
| SOX | sulfisoxazole | 921 | 278 | | 643 |
| STR | streptomycin | 991 | 387 | | 604 |
| SXT | co-trimoxazole | 1335 | 786 | | 549 |
| TBM | tobramycin | 1200 | 807 | 59 | 334 |
| TET | tetracycline | 1087 | 474 | | 613 |
| TGC | tigecycline | 1747 | 1662 | | 85 |
| TIO | ceftiofur | 783 | 347 | 60 | 376 |
| TMP | trimethoprim | 884 | 496 | | 388 |
| TZP | piperacillin/tazobactam | 2200 | 1874 | 100 | 226 |

### Converting MIC to SIR

The categories of SIR were selected for classifying genomes, as the metadata were more abundant than minimum inhibitory concentration (MIC). Isolates were subjected to different testing panels, so they did not have a SIR classification for every antimicrobial. In cases where an isolate lacked a SIR classification for an antibiotic, but had a MIC or zone diameter (ZD) available, a SIR classification was made by comparison to a standard. The CLSI M100 standard was used where available. The following conversions were used when CLSI standards were unavailable: the MIC-to-SIR for TIO and STR used NARMS, and the MIC-to-SIR and ZD-to-SIR for TGC used EUCAST; there was no ZD-to-SIR standard for TIO nor CST.

### Converting I into R for low abundance classes

To ensure adequate data for training the machine learning models, each SIR class was required to have at least 25 members; if there were less than 25 isolates with intermediate resistance to an antibiotic, they were grouped into the resistant class. The total number of isolates and number of isolates in each SIR class are compiled in Table 3.1.

Classifying an intermediate isolate as susceptible to an antibiotic is risky: if the antibiotic is used for treatment, it may be ineffective in its typical dose. Classifying an intermediate isolate as resistant poses less risk, as it only restricts the available treatment options. Intermediate isolates have factors setting them apart from the susceptible class as they may share resistance markers with resistant isolates, but exhibit lower expression. Binning intermediate with susceptible isolates could mask these markers.

## 3.4.2 Data Processing

Raw reads of all genomes were used where available; pre-assembled genomes were used in cases where raw data were unavailable. Raw *E. coli* sequences were trimmed using Trim-galore (v0.6.5) [298], which is a wrapper for Cutadapt [299] and FastQC [300]. Fast length adjustment of short reads or FLASh (v2.2.00), was used to merge the short paired-

end reads [301]. The genomes were then assembled with SPAdes (v3.14.0) [302].

Jellyfish (v2.2.10) [279] was used to count the occurrences of *k*-mers, or subsequences of length k, in each genome. Only canonical *k*-mers were counted, meaning reverse compliments were skipped. The counts were stored in a frequency matrix, where rows correspond to genomes, columns correspond to *k*-mers, and cells contain the frequencies of *k*-mers within the genomes. Two frequency matrices were prepared, one for 11-mers, and another for 25-mers. All possible 11-mer sequences can be pre-computed, for a total of 4,194,304; however, the 1,125,899,906,842,620 possible 25-mers were impractical to pre-compute, so only those present in one or more genomes were included in the frequency matrix.

### 3.4.3 Model Training

Three types of machine learning models were trained to predict SIR classifications based upon *E. coli* whole genome sequences. Scikit-learn (v0.23.2) [281] was used as a wrapper to implement the three methods: support vector machines (SVM) [281], gradient boosted decision trees from XGBoost (XGB) (v1.0.2) [280], and artificial neural networks (ANN) from Keras [282]. Separate models, using each of the three methods, were trained for each antimicrobial. For each model, the input was the *k*-mer frequency matrix paired with the SIR classifications of the genomes.

Parameter optimization was performed for 11-mers. 20% of the genomes were withheld for validation, while the remaining 80% were used for nested five-fold cross-validation, including parameter optimization. The outer loop of nested cross-validation used the cross_validate method from scikit-learn, and the inner loop used GridSearchCV from scikit-learn [281]. F1 score was used as the evaluation parameter of each loop. A final model was trained on the 80%, using the selected parameters, and validated with the withheld 20%.

Nested five-fold cross-validation of the 25-mer matrix was not possible as the frequency matrix was too large to be manipulated with the available computational resources. Instead, default parameters were used.

Feature selection was performed to reduce the size of the matrix. This step was inside of the nested cross-validation, as part of parameter optimization. The most important *k*-mers (features) were selected via the SelectKBest method of scikit-learn, using the ANOVA F-value [281]. Feature selection must be performed only on the training data to prevent bias, so the size of the matrix could not be reduced in advance. For 11-mers, the following values were tested for selection: 100, 500, 1000, 2000, 3000, 4000, and 5000 features; the best number of features was determined per antimicrobial. For 25-mers, a value of 1000 was used.

### 3.4.4 Feature Extraction and Annotation

The contribution of each *k*-mer to model training was reported by the importance measure of XGB. The top ten k-mers were extracted from the XGB models for each antimicrobial. For annotation, blastn-short was used to identify perfect matches within the Comprehensive Antibiotic Resistance Database (CARD) [179], as well as a custom database consisting of the 4300 assembled genomes annotated with Prokka (v1.14.6) [303].

### 3.4.5 Comparison to Database Methods and Validation

The performance of the trained 11-mer ML models were compared to two popular database methods. AMRFinderPlus (v3.8.4) [180] and ResFinder (v4.1.11) [181] were run on the WGS dataset to predict resistance profiles.

Two public datasets used by Bortolaia *et al.* for the validation of ResFinder 4.0 were obtained to validate the trained models [181]. These datasets were independent from the dataset of 4300 genomes used to train the ML models. The German (DE) dataset consisted of 390 *E. coli* whole genome sequences from a mix of animal and human sources [181]; antibiotics with available data included AMP, CHL, CIP, CPM, CTX, CTZ, ETP, GEN, IMP, MER, NAL, TET, and TGC. The Denmark (DK) dataset consisted of 95 *E. coli* whole genome sequences from a mix of animal and food sources [304]; antibiotics with available data included AMP, CHL, CIP, CPM, CST, CTX, CTZ, ETP, FOX, GEN, IMP, MER, NAL,

TET, and TMP. ResFinder was run on the datasets to replicate and confirm the published results, for both raw and assembled genomes. The trained ML models were then run on the raw and assembled genomes, and the performance was compared to ResFinder. The laboratory-determined resistance classifications were made as susceptible or resistant only; they did not use an intermediate definition. Therefore, any model predictions of intermediate were binned into the resistant class for evaluation.

## 3.5 Results

### 3.5.1 11-mer Models

To determine the optimal number of features for training the 11-mer models to classify genomes into SIR categories, performance was compared across a range of 100 to 5000 features. The variation in model accuracy across this range is shown for each antimicrobial in Figure 3.2, with the complete feature range and accuracy data for each antimicrobial and model available in Supplementary Table 3.

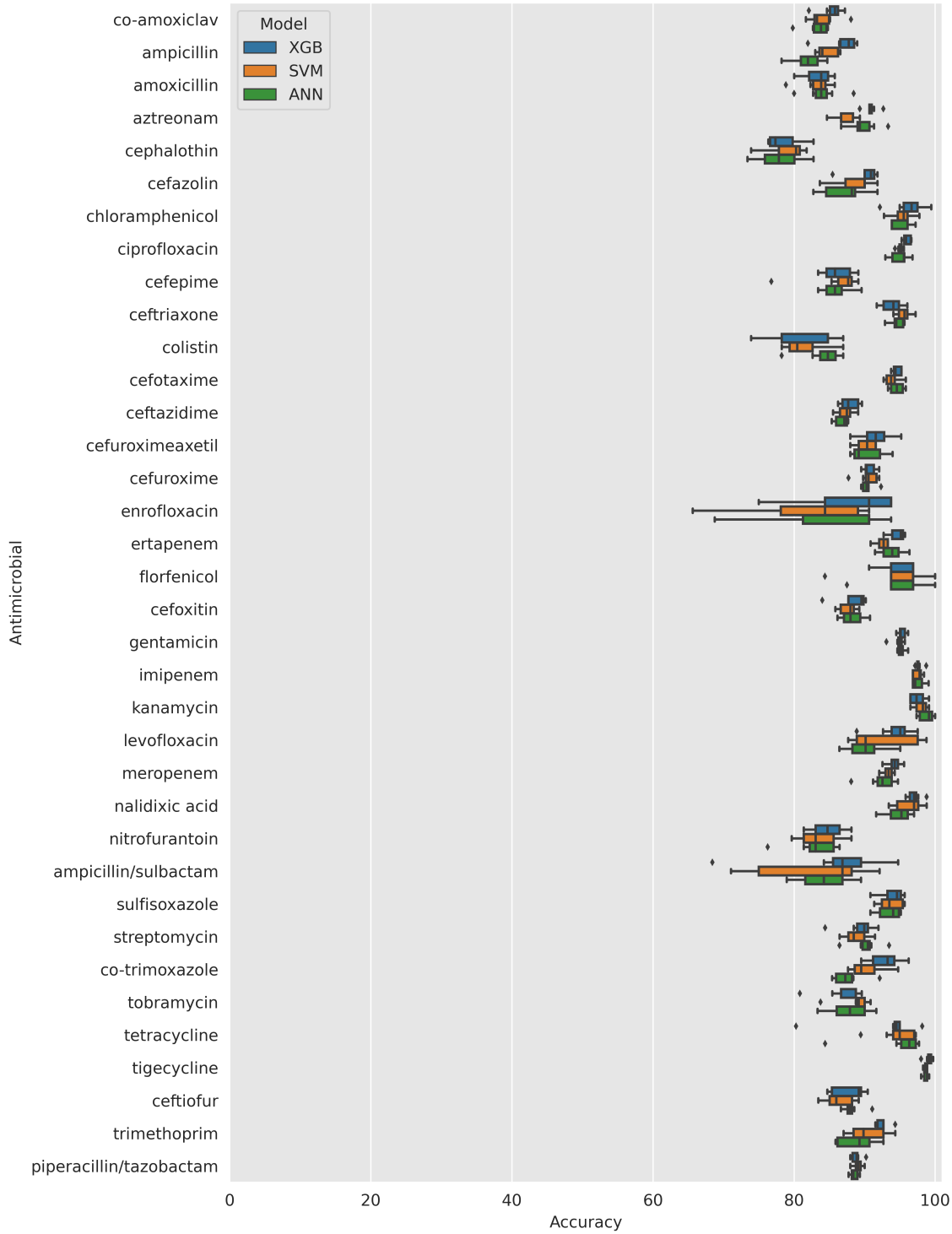



Figure 3.2: Variation in accuracy of predicting Susceptible, Intermediate, or Resistant classification to antimicrobials depending on number of features (range of 100 to 5000 features). Feature selection was performed with SelectKBest. XGBoost (XGB), support vector machines (SVM), and artificial neural networks (ANN) were trained using the 11-mer frequency matrix.

For most of the antimicrobials, changing the number of features affected the accuracy by only a few percent; however, the biggest exceptions to this were enrofloxacin and ampicillin/sulbactam, which varied over 20% depending on the model and number of features. Both had one of the smallest datasets, with enrofloxacin having only 157 tested isolates from a single study (AAFC), and ampicillin/sulbactam having only 190, but from a number of sources. Based on these data, the top-performing model for each antimicrobial was used for subsequent analyses.

The performance of 11-mer models is shown in Figure 3.3. In general, XGB models performed the best, achieving an average accuracy of 93.2% (range 82.8% to 99.7%) and average F1-score of 86.0%. SVM models had an average accuracy of 92.3% (range 81.8% to 100%) and average F1-score of 83.8%, while ANN models had an average accuracy of 92.3% (range 82.8% to 100%) and average F1-score of 84.9%.

The top-performing model overall based on average F1 score of 98.9% was florfenicol, with individual scores of 96.8% for XGB, 100.0% for SVM, and 100.0% for ANN. The worst-performing model was piperacillin/tazobactam with an average F1 score of 54.8%, with individual scores of 58.8% for XGB, 53.4% for SVM, and 52.1% for ANN.

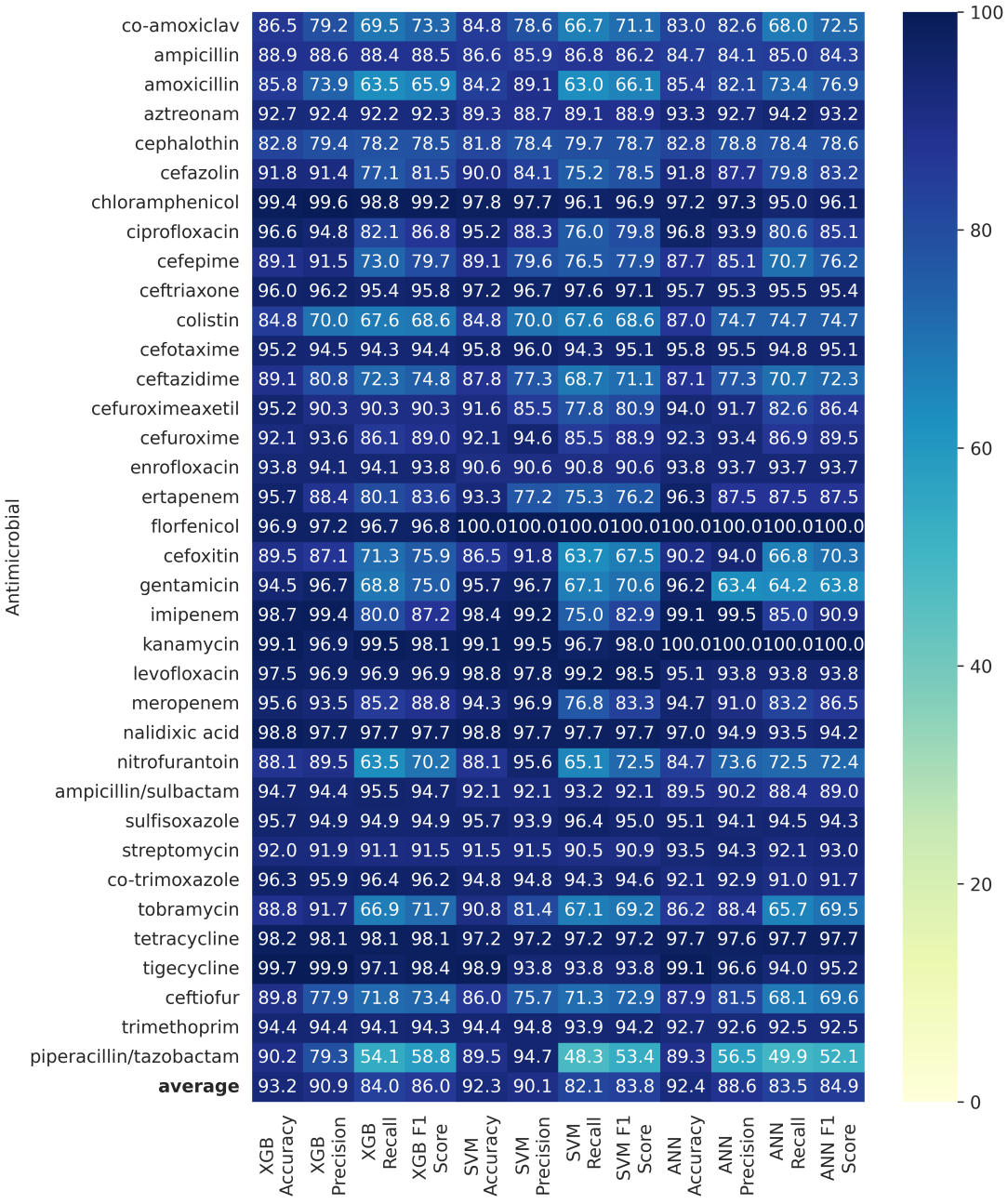



Figure 3.3: Accuracy, precision, recall, and F1-score for the prediction of Susceptible, Intermediate, or Resistant classification to 36 different antimicrobials. XGBoost (XGB), support vector machines (SVM), and artificial neural networks (ANN) were trained using the 11-mer frequency matrix.

### 3.5.2 25-mer Models

The use of 25-mer features were investigated due to their greater specificity within genomes; however, due to the constraints of the available computational resources the models were not able to be parameter tuned and exhibited lower overall performance when compared to the 11-mer models. The performance of 25-mer models is shown in Figure 3.4, and the performance comparison to 11-mer models is shown in Figure 3.5. The complete performance data for the 11-mer and 25-mer models are available in Supplementary Table 4.

Of the 25-mer models, XGB and SVM were close in performance, with average accuracies of 87.8% (range 73.7% to 99.1%) and 87.9% (73.9% to 98.9%) respectively, and average F1-scores of 75.0% and 73.6% respectively. The ANN models performed poorly in comparison, with an average accuracy of 68.9% (range 10.2% to 98.6%) and average F1-score of 59.5%.

Comparing the 11-mer and 25-mer models in terms of accuracy: 35 of 36 XGB, 34 of 36 SVM, and 35 of 36 ANN models performed better for 11-mers. In terms of F1-score, 35 of 36 XGB, 35 of 36 SVM, and 34 of 36 ANN models performed better for 11-mers.

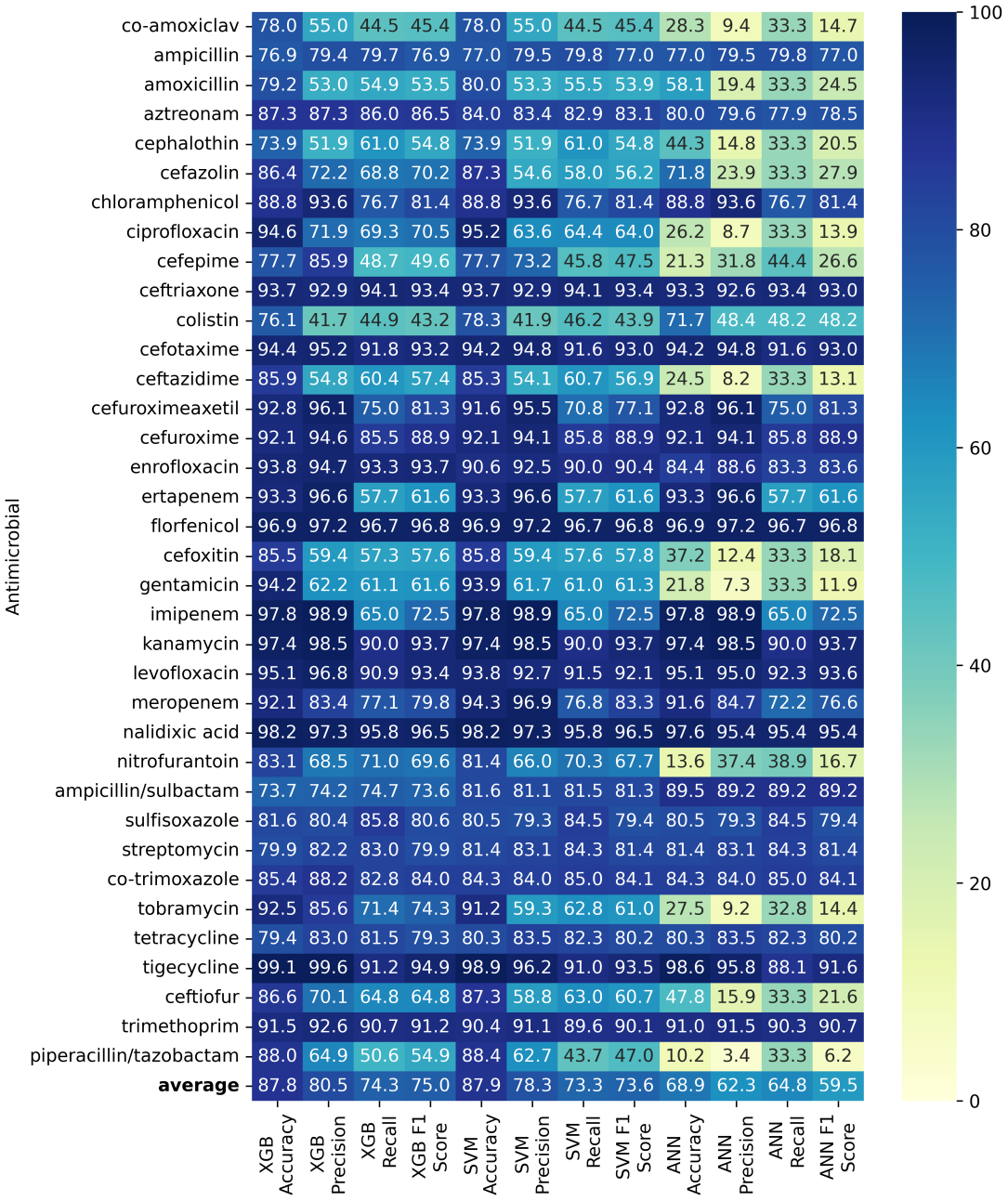



Figure 3.4: Accuracy, precision, recall, and F1-score for the prediction of Susceptible, Intermediate, or Resistant classification to 36 different antimicrobials. XGBoost (XGB), support vector machines (SVM), and artificial neural networks (ANN) were trained using the 25-mer frequency matrix.

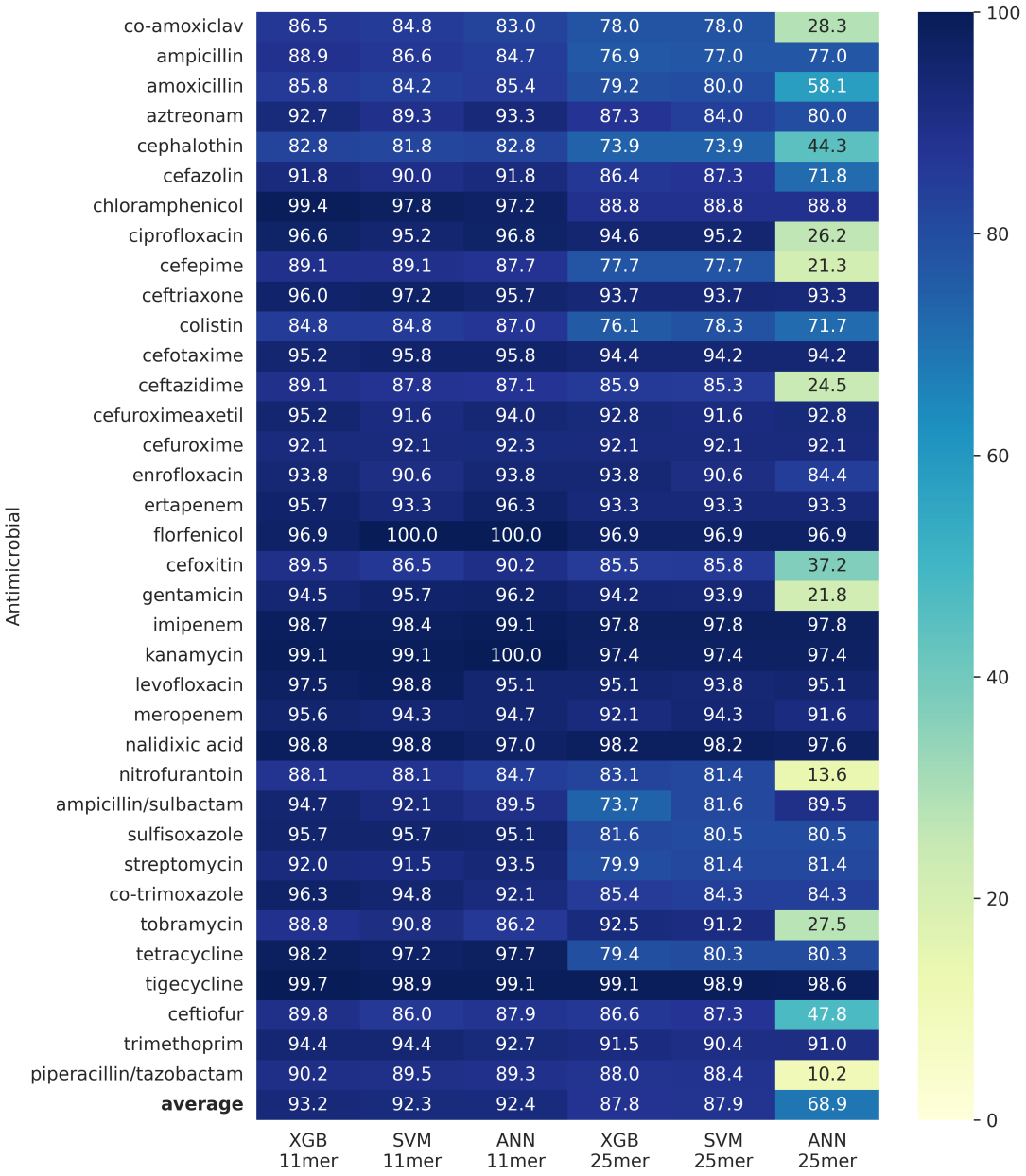



Figure 3.5: Accuracy of predicting Susceptible, Intermediate, or Resistant classification to 36 different antimicrobials. XGBoost (XGB), support vector machines (SVM), and artificial neural networks (ANN) were trained using frequency matrices with *k*-mers of length 11 or 25.

### 3.5.3 Comparison to Database Methods

Two leading tools for *in silico* classification of susceptibility and resistance are AMRFinderPlus and ResFinder; they use curated databases of known resistance markers to make classifications. The tools were applied to the 4300 genomes of this study. A comparison of their accuracies with those of the ML 11-mer models is shown in Figure 3.6.

Compared to AMRFinderPlus, 11-mer XGB models performed better for 34 of 36 antimicrobials, SVM performed better for 33 of 36, and ANN performed better for 32 of 36. Compared to ResFinder, 11-mer XGB models performed better for 34 of 36 antimicrobials, SVM performed better for 32 of 36, and ANN performed better for 32 of 36.

Compared to AMRFinderPlus, 25-mer XGB models performed better for 28 of 36 antimicrobials, SVM performed better for 28 of 36, and ANN performed better for 17 of 36. Compared to ResFinder, 25-mer XGB models performed better for 23 of 36 antimicrobials, SVM performed better for 23 of 36, and ANN performed better for 15 of 36.

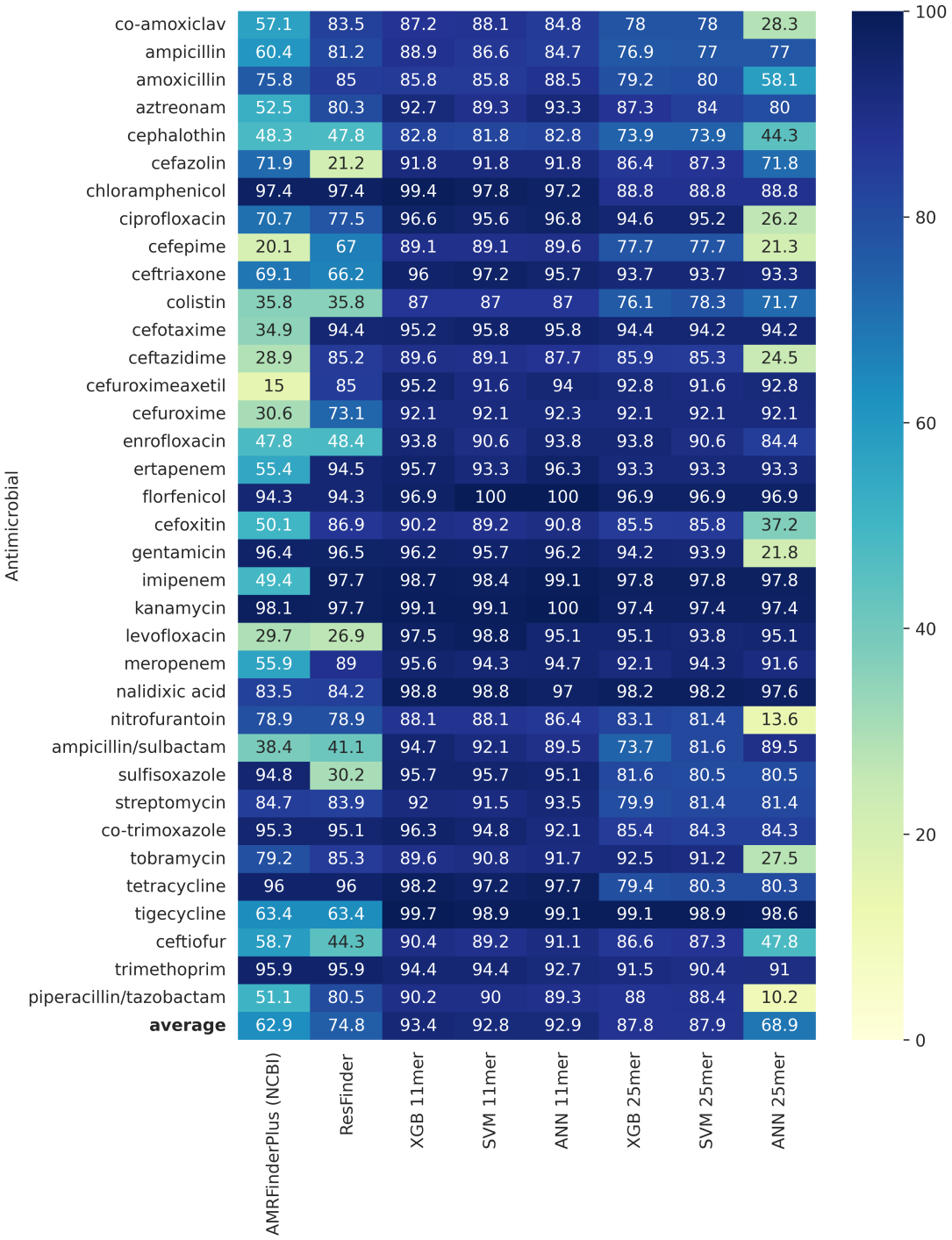



Figure 3.6: Accuracy of predicting Susceptible, Intermediate, or Resistant classification for three machine learning models and two database methods. The machine learning methods, XGBoost (XGB), support vector machines (SVM), and artificial neural networks (ANN) were trained using the 11-mer and 25-mer frequency matrices. The database methods are ResFinder and AMRFinderPlus.

### 3.5.4 Validation Datasets

The pre-trained 11-mer ML models were validated for 16 antimicrobials using two external public datasets, DE and DK, which were also used by ResFinder for validation. The results from the ResFinder publication for these two datasets were replicated, as shown in Table 3.2 and Table 3.3. There was a 0.1% difference between the paper and replication for CIP and FOX; this is likely due to the ResFinder database being updated since publication.

The 11-mer ML models were run on the DE and DK datasets, and their performance is shown in Table 3.4 and Table 3.5, respectively. For the DE dataset, 2 of 13 XGB models, 3 of 13 SVM models, and 1 of 13 ANN models outperformed ResFinder. The ML models had 0 correct predictions for MER; however, the sample size for this drug in the DE dataset was only two. For the DK dataset, 4 of 15 XGB models, 4 of 15 SVM models, and 3 of 15 models matched or exceeded the accuracy of ResFinder. ResFinder had 100% accuracy for CST, GEN, IMP, MER, TET, and TMP. XGB and SVM had 100% accuracy for GEN, IMP, and MER; ANN had 100% for GEN, MER, TET. The worst performing ML models for the DK dataset were those for CPM and CST, with below 22% accuracy.

Table 3.2: Replication of the results from the ResFinder publication for the DK (Denmark) dataset.

| Code | Antimicrobial | Total Samples | Num. R | Num.S | Paper | Replicate |
|---|---|---|---|---|---|---|
| AMP | Ampicillin | 95 | 95 | 0 | 100 | 100 |
| CHL | Chloramphenicol | 95 | 8 | 87 | 100 | 100 |
| CIP | Ciprofloxacin | 95 | 29 | 66 | 87.3 | 87.4 |
| CPM | Cefepime | 95 | 78 | 17 | 71.6 | 71.6 |
| CST | Colisitin | 95 | 0 | 95 | 100 | 100 |
| CTX | Cefotaxime | 95 | 95 | 0 | 100 | 100 |
| CTZ | Ceftazidime | 95 | 94 | 1 | 98.9 | 98.9 |
| ETP | Ertapenem | 95 | 1 | 94 | 98.9 | 98.9 |
| FOX | Cefoxitin | 95 | 46 | 49 | 97.8 | 97.9 |
| GEN | Gentamicin | 95 | 15 | 80 | 100 | 100 |
| IMP | Imptamicin | 95 | 0 | 95 | 100 | 100 |
| MER | Meropenem | 95 | 0 | 95 | 100 | 100 |
| NAL | Nalidixic Acid | 95 | 25 | 70 | 98.9 | 98.9 |
| TET | Tetracycline | 95 | 53 | 42 | 100 | 100 |
| TGC | Tigecycline | – | – | – | – | – |
| TMP | Trimethoprim | 95 | 29 | 66 | 100 | 100 |

Table 3.3: Replication of the results from the ResFinder publication for the DE (Germany) dataset.

| Code | Antimicrobial | Total Samples | Num. R | Num.S | Paper | Replicate |
|---|---|---|---|---|---|---|
| AMP | Ampicillin | 202 | 202 | 0 | 100 | 100 |
| CHL | Chloramphenicol | 130 | 63 | 67 | 73.1 | 73.1 |
| CIP | Ciprofloxacin | 275 | 275 | 0 | 99.2 | 99.2 |
| CPM | Cefepime | 137 | 137 | 0 | 100 | 100 |
| CST | Colisitin | – | – | – | – | – |
| CTX | Cefotaxime | 370 | 370 | 0 | 98.6 | 98.6 |
| CTZ | Ceftazidime | 282 | 282 | 0 | 99.2 | 99.2 |
| ETP | Ertapenem | 130 | 60 | 70 | 54.6 | 54.6 |
| FOX | Cefoxitin | – | – | – | – | – |
| GEN | Gentamicin | 387 | 129 | 258 | 97.6 | 96.9 |
| IMP | Imptamicin | 194 | 2 | 192 | 100 | 100 |
| MER | Meropenem | 2 | 2 | 0 | 100 | 100 |
| NAL | Nalidixic Acid | 127 | 99 | 28 | 99.2 | 99.2 |
| TET | Tetracycline | 171 | 138 | 33 | 98.8 | 98.8 |
| TGC | Tigecycline | 152 | 5 | 147 | 96.7 | 96.7 |
| TMP | Trimethoprim | – | - | – | – | – |

Table 3.4: Comparison of accuracy of ML models and ResFinder for predicting Susceptible, Intermediate, or Resistant classification using *E. coli* whole genome sequences from the DK validation dataset.

| Code | Antimicrobial | Total Samples | ResFinder | XGB | SVM | ANN |
|---|---|---|---|---|---|---|
| AMP | Ampicillin | 95 | 98.9 | 89.5 | 73.7 | 86.3 |
| CHL | Chloramphenicol | 95 | 98.9 | 98.9 | 100 | 94.7 |
| CIP | Ciprofloxacin | 95 | 87.4 | 70.5 | 77.9 | 73.7 |
| CPM | Cefepime | 95 | 70.5 | 21.1 | 20.0 | 18.9 |
| CST | Colisitin | 95 | 100 | 1.1 | 6.3 | 3.2 |
| CTX | Cefotaxime | 95 | 98.9 | 71.6 | 73.7 | 75.8 |
| CTZ | Ceftazidime | 95 | 97.9 | 72.6 | 66.3 | 82.1 |
| ETP | Ertapenem | 95 | 98.9 | 97.9 | 88.4 | 95.8 |
| FOX | Cefoxitin | 95 | 97.9 | 84.2 | 84.2 | 83.2 |
| GEN | Gentamicin | 95 | 100 | 100 | 100 | 100.0 |
| IMP | Imptamicin | 95 | 100 | 100 | 100 | 98.9 |
| MER | Meropenem | 95 | 100 | 100 | 100 | 100.0 |
| NAL | Nalidixic Acid | 95 | 98.9 | 96.8 | 88.4 | 85.3 |
| TET | Tetracycline | 95 | 100 | 97.9 | 97.9 | 100 |
| TGC | Tigecycline | – | – | – | – | – |
| TMP | Trimethoprim | 95 | 100 | 94.7 | 86.3 | 88.4 |

Table 3.5: Comparison of accuracy of ML models and ResFinder for predicting Susceptible, Intermediate, or Resistant classification using *E. coli* whole genome sequences from the DE validation dataset.

| Code | Antimicrobial | Total Samples | ResFinder | XGB | SVM | ANN |
|---|---|---|---|---|---|---|
| AMP | Ampicillin | 202 | 100 | 99. 5 | 95.0 | 99.0 |
| CHL | Chloramphenicol | 130 | 68.5 | 72.3 | 69.2 | 65.4 |
| CIP | Ciprofloxacin | 275 | 99.3 | 89.1 | 89.5 | 91.3 |
| CPM | Cefepime | 137 | 100 | 41.6 | 38.7 | 43.1 |
| CST | Colisitin | – | – | – | – | – |
| CTX | Cefotaxime | 370 | 97.0 | 96.5 | 95.1 | 96.2 |
| CTZ | Ceftazidime | 282 | 98.9 | 84.0 | 88.7 | 93.3 |
| ETP | Ertapenem | 130 | 54.6 | 58.5 | 55.4 | 56.2 |
| FOX | Cefoxitin | – | – | – | – | – |
| GEN | Gentamicin | 387 | 97.4 | 89.9 | 88.4 | 88.6 |
| IMP | Imptamicin | 194 | 100 | 99.0 | 94.8 | 95.9 |
| MER | Meropenem | 2 | 100 | 0 | 0 | 0 |
| NAL | Nalidixic Acid | 127 | 98.4 | 97.6 | 99.2 | 93.7 |
| TET | Tetracycline | 171 | 98.2 | 96.5 | 97.1 | 96.5 |
| TGC | Tigecycline | 152 | 96.1 | 82.2 | 89.5 | 90.1 |
| TMP | Trimethoprim | – | – | – | – | – |

### 3.5.5 Feature Annotation

Based on the 25-mer ML models, the top ten predictive features of the models for each of the 36 antimicrobials were determined. The annotated genome regions that these features correspond to were found using both CARD and Prokka. Below are examples of extracted features which include k-mers with known AMR associations, as well as a k-mer currently suspected, but unconfirmed, to contribute to AMR. The full set of top features and annotations are shown in Supplementary Table 2. The majority of the predictive features are more associated with resistant strains, and overall, most of the identified predictive features were from genes previously associated with AMR.

The top features extracted for AMC are shown in Table 3.6. With CARD, all of the features were annotated as CMY beta-lactamase. With Prokka, nine of the features were annotated as *ampC* (AmpC: beta-lactamase), and one as *blc* (Blc: outer membrane lipopro-

tein). A result of Web BLAST of the top features against *E. coli* is shown in Figure 3.7.

Table 3.6: Top 25-mers for co-amoxiclav (AMC), their associated Susceptible/Intermediate/Resistant (SIR) classification, and their annotations via CARD and Prokka.

| *k*-mer | SIR | CARD | Prokka |
|---|---|---|---|
| AGCCGCATTACTGCATTTTTATCAA | R | CMY:beta-lactamase | *ampC*:beta-lactamase |
| AGGGATTAGGCTGGGAGATGCTGAA | R | CMY:beta-lactamase | *ampC*:beta-lactamase |
| ATCAACGGGGTGATGGTGCGATTAA | R | CMY:beta-lactamase | *ampC*:beta-lactamase |
| ATCACCCCGTTGATGCAGGAGCAGG | R | CMY:beta-lactamase | *ampC*:beta-lactamase |
| ATCCCTGGTACATATCGCCAATACG | R | CMY:beta-lactamase | *ampC*:beta-lactamase |
| CACGTTTCTCCGGGACAACTTGACG | R | CMY:beta-lactamase | *ampC*:beta-lactamase |
| CCTCGACACGGACAGGGTTAGGATA | R | CMY:beta-lactamase | *ampC*:beta-lactamase |
| CTGAACCCAGCGGGCCATATCAATA | R | CMY:beta-lactamase | *ampC*:beta-lactamase |
| CTGGTTGTCGCGTGTAGCTCCCCGA | R | CMY:beta-lactamase | *blc*: outer membrane lipoprotein |
| GCATAACGATTTTTTCATCATGAAA | R | CMY:beta-lactamase | *ampC*:beta-lactamase |

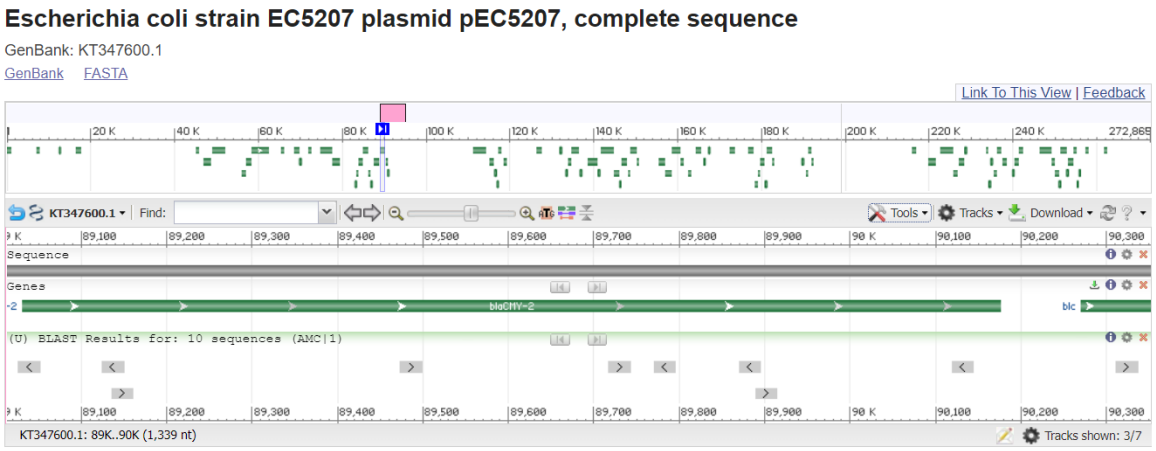



Figure 3.7: An example Web BLAST result of the top 10 features for co-amoxiclav (AMC) against *E. coli*. Nine of the features aligned to CMY, and one to *blc*.

The top features extracted for AMP are shown in Table 3.7. CARD only annotated two of the ten features: one feature as beta-lactamase TEM-55 and one as *ramR*. Prokka annotated four features as "IS*1380* family transposase IS*Ecp1*," and the remainder as "hypothetical proteins." When compared with *E. coli* available in Web BLAST, the features align to *tnpA* and the region between *tnpA* and a CTX-M beta-lactamase. Example Web BLAST results are shown in Figures 3.8 and 3.9

Table 3.7: Top 25-mers for ampicillin (AMP), their associated Susceptible/Intermediate/Resistant (SIR) classification, and their annotations via CARD and Prokka.

| *k*-mer | SIR | CARD | Prokka |
|---|---|---|---|
| AAAATTGAGTGTTGCTCTGTGGATA | R | – | hypothetical protein |
| AAATTATCGAAACGATAGAATCTCT | R | – | none: IS*1380* family transposase IS*Ecp1* |
| AAGCAGTCTAAATTCTTCGTGAAAT | R | – | hypothetical protein |
| AAGGTGGTTGTAAATAATGTTACAA | R | TEM-55: beta-lactamase | hypothetical protein |
| AGACCATGCTCTGCGGTCACTTCAT | R | – | none: IS*1380* family transposase IS*Ecp1* |
| AGCAGTCTAAATTCTTCGTGAAATA | R | – | hypothetical protein |
| ATTAGCTTCAAAAATCACTATTTCA | R | *ramR*: repressor regulating RamA expression | hypothetical protein |
| CAAAAATGATTGAAAGGTGGTTGTA | R | – | hypothetical protein |
| CATATAACCTATTTTTGTTGTTCAA | R | – | none: IS*1380* family transposase IS*Ecp1* |
| CGCCCAAAATGACTTTAGCAAGAGA | R | – | none: IS*1380* family transposase IS*Ecp1* |

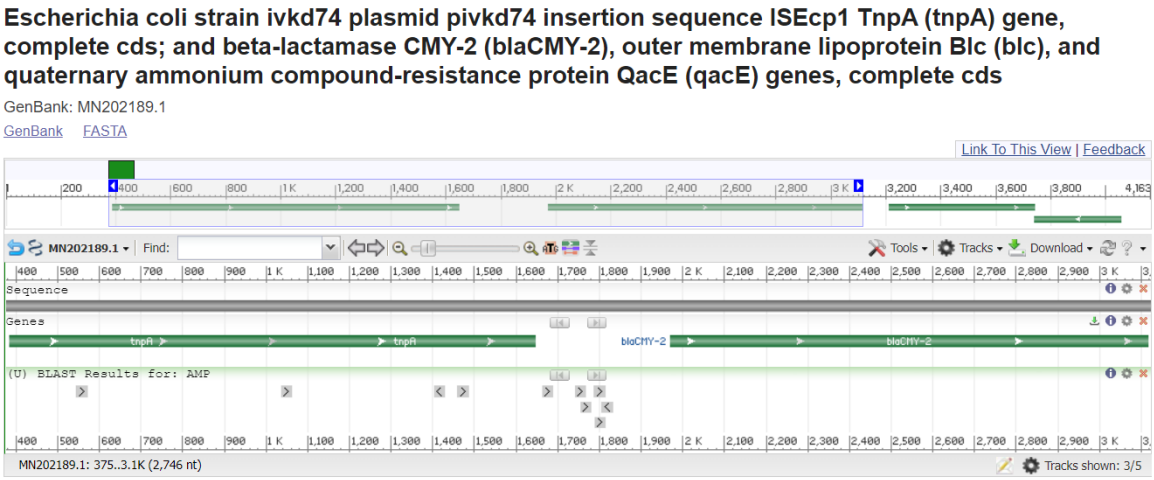



Figure 3.8: An example Web BLAST result of the top 10 features for ampicillin (AMP) against *E. coli*. Nine of the features aligned with the reference genome: three within *tnpA* and the remainder between *tnpA* and CTX-M.

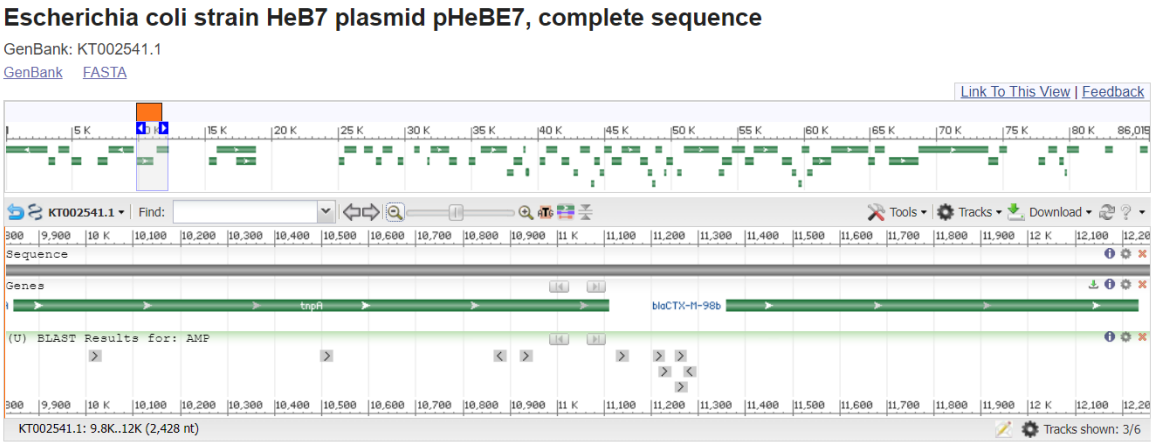



Figure 3.9: A Web BLAST result of the top 10 features for ampicillin (AMP) against *E. coli*. Four *k*-mers aligned with *tnpA* and six with the region between *tnpA* and CTX-M.

The top features extracted for CIP are shown in Table 3.8. CARD and Prokka annotated seven of the top ten *k*-mers as point mutations in *gyrA* (DNA gyrase subunit A), and two as point mutations in *parC* (ParC subunit of topoisomerase IV). The two *parC* *k*-mers differed by a single base; one *k*-mer was associated with resistance to ciprofloxacin in the dataset, while the other was associated with susceptibility. These bases are bolded in Table 3.8. Five of the *gyrA* *k*-mers also showed a single base difference between susceptible and resistant association (bolded in Table 3.8). All of the resistant-associated *gyrA* *k*-mers overlapped with one another, and/or were the reverse complement of one another.

Table 3.8: Top 25-mers for ciprofloxacin (CIP), their associated Susceptible/Intermediate/Resistant (SIR) classification, and their annotations via CARD and Prokka. Bolded bases differ in the k-mers associated with S vs those with R.

| *k*-mer | SIR | CARD | Prokka |
|---|---|---|---|
| AAATACCATCCCCATGGTGACT**T**GG | R | point mutation of *gyrA* | *gyrA*:DNA gyrase subunit A |
| CCCATGGTGACT**T**GGCGGTTTATAA | R | point mutation of *gyrA* | *gyrA*:DNA gyrase subunit A |
| CCGCC**A**AGTCACCATGGGGATGGTA | R | point mutation of *gyrA* | *gyrA*:DNA gyrase subunit A |
| AATACCATCCCCATGGTGACT**C**GGC | S | point mutation of *gyrA* | *gyrA*:DNA gyrase subunit A |
| ATACGGACGATCGTGTCATAAACCG | S | point mutation of *gyrA* | *gyrA*:DNA gyrase subunit A |
| CATAAACCGCCG**A**GTCACCATGGGG | S | point mutation of *gyrA* | *gyrA*:DNA gyrase subunit A |
| CTGCGCCATACGGACGATCGTGTCA | S | point mutation of *gyrA* | *gyrA*:DNA gyrase subunit A |
| **A**TATCGCCGTGCGGATGGTATTTAC | R | point mutation of *parC* | *parC*:DNA topoisomerase 4 subunit A |
| **C**TATCGCCGTGCGGATGGTATTTAC | S | point mutation of *parC* | *parC*:DNA topoisomerase 4 subunit A |
| ACGGGAAGAACGCACTACTGCAGTG | R | – | *ansB*:L-asparaginase 2 |

## 3.6 Discussion

### 3.6.1 Model Performance

11-mer ML models performed well for SIR prediction from *E. coli* WGS data; only one antimicrobial, TZP, had an F1-score below 60%. 25-mer models had lower accuracies and F1-scores for the majority of antimicrobials. This decline in performance may be due to the lack of parameter tuning. ANN models in particular performed poorly; as they have the highest number of variables, they may have suffered the most from the lack of tuning.

The 11-mer frequency matrix was 9.5 GB, while the 25-mer was 650 GB and too large for five-fold nested cross-validation. To lower computational resource requirements, the 25-mer matrix was modified and saved for each antimicrobial: as not all genomes have data for every antimicrobial, the irrelevant rows were trimmed for each antibiotic. After removing irrelevant genomes, any all-0 frequency columns were also trimmed. For AMP, this reduced the 650 GB matrix to 411 GB. This lowers resource usage during training at the tradeoff of requiring increased storage space, as one matrix is created for each antimicrobial. The high resource consumption by long *k*-mers is a universal problem for ML model training.

PhenotypeSeeker [226] uses *k*-mers to train logistic regression models to predict bacterial phenotypes. It has predicted virulence phenotypes of *Klebsiella pneumoniae* (F1 0.88), ciprofloxacin resistance of *Pseudomonas aeruginosa* (F1 0.88), and azithromycin resistance of *Clostridium difficile* (F1 0.97). The default *k*-mer length for PhenotypeSeeker is 13, as training models with longer *k*-mers were found to take too much RAM and time for their datasets, which ranged in size from 167 to 459 genomes. Additionally, they found that building models with assembled genomes was faster than raw genomes, with no decrease in model performance. The models built by PhenotypeSeeker use presence binary presence/absence of *k*-mers in the samples, as opposed to a frequency matrix. By default, the samples are also weighted to account for multiple highly similar genomes. To build models, 25% of the data is withheld for testing and the remainder is used for training. Neither cross-validation nor parameter optimization is performed.

Kover is another *k*-mer based prediction tool which uses Set Covering Machines (SCM) [227]. Drouin et al validated Kover by predicting AMR of *Clostridium difficile*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa*, and *Streptococcus pneumoniae* for 17 antibiotics. Like PhenotypeSeeker, Kover uses *k*-mer presence/absence as opposed to *k*-mer frequency. They tested six *k*-mer lengths, ranging from 15 to 91, and found no impact on accuracy. An advantage of Kover is its low memory usage; however, it does require storage space. Instead of loading all the data into RAM at once, it accesses stored data in pieces. *k*-mers responsible for the models' rules were extracted and compared to annotated genomes with BLAST. The models may identify one resistance marker, as was the case for isoniazid resistance of *M. tuberculosis*, or multiple markers. Multiple markers may be for the same region or gene, as with rifampicin resistance of *M. tuberculosis*. They may also be for different regions or genes, as with erythromycin resistance of *S. pneumoniae*.

The present study examined three different ML model types other than those used by PhenotypeSeeker (logistic regression) and Kover (SCM). The ML models were trained with *k*-mer frequencies and are able to classify genomes into more than two classes, while the other tools used only *k*-mer presence/absence, and perform binary classifications.

### 3.6.2 Top Features

The highest importance 25-mers were extracted from XGB models and located within the Prokka-annotated whole genomes. A maximum of 10 top features were extracted from each model for further examination; however, there may be more or less than 10 features contributing to predictions for a model. Having a single top *k*-mer is not abnormal. Kover commonly identifies only one or two top targets as well [227].

#### Penicillins

Penicillins are beta-lactam antibiotics which inhibit cell wall synthesis, resulting in bacterial death [45]. The most common mechanism of resistance to beta-lactams is hydrolysis of the antibiotics by beta-lactamases [44].

Co-amoxiclav (AMC) is composed of amoxicillin and the inhibitor clavulanic acid, which binds to beta-lactamase enzymes to inactivate them [46, 47]. The top 10 25-mers for AMC were annotated as CMY-type beta-lactamase via the CARD, while Prokka annotated nine as *ampC* (AmpC beta-lactamase) and one as *blc* (Blc: outer membrane lipoprotein) (Figure 3.6). Web BLAST against *E. coli* located nine of the *k*-mers within a CMY beta-lactamase, and one within *blc* (Figure 3.7). CMY-types belong to the group of AmpC beta-lactamases, which are capable of hydrolyzing numerous beta-lactam antibiotics, including amoxicillin [305]. The presence of a beta-lactam antibiotic, such as amoxicillin, induces the expression of AmpC [306]. Clavulanic acid is a poor inhibitor of AmpC beta-lactamases, and can instead induce expression of the enzymes [307, 308, 309, 306]. The *blc* gene is commonly conserved downstream of a CMY gene [310, 311, 312]. Blc is an outer membrane lipoprotein which is expressed under stress conditions [313]. Its role in resistance is believed to be assisting in membrane maintenance through storing or transporting lipids [314, 313]

Ampicillin is another beta-lactam antibiotic which inhibits cell-wall synthesis. Of the top 10 25-mers for AMP, two were able to be annotated with CARD: one as TEM-55, and one as *ramR* from *Klebsiella pneumoniae*. Prokka annotated four as belonging to IS*Ecp1*, an insertion sequence of the IS*1380* family; the remaining six were labelled "hypothetical proteins." A web BLAST search places the four annotated as IS*Ecp1* by Prokka within *tnpA*, and the remaining six between *tnpA* and a beta-lactamase gene (Figures 3.8 and 3.9). IS*Ecp1* is commonly followed by a resistance gene such as for CTX-M or CMY, and often also supplies a promotor for this gene [315, 316].

TEM-type beta-lactamases are able to cleave beta-lactam antibiotics through hydrolysis [305]. RamR represses the gene for RamA which controls an efflux pump; when *ramR* is repressed, expression of *ramA* is increased, which in turn causes an increase in AMR via efflux pumps [317]. As CARD flagged the RamR-annotated *k*-mer as being from *Klebsiella*, and the BLAST results support that the *k*-mer is in the same region as the others, it

is likely an incorrect annotation.

An IS*Ecp1* together with a resistance gene is referred to as a transposition unit; these units are often found within Tn2, or with Tn2 components (such as *tnpA*) [318, 319]. IS*Ecp1* is able to mobilize the transposition unit in and out of both plasmids and chromosomes, facilitating spread of resistance genes [315, 319]. IS*Ecp1* and its relatives can also be found in other Enterobacteriaceae, and are a major driver behind the worldwide spread of CTX-M [320, 315, 318, 321].

**Cephalosporins**

Along with penicillins, cephalosporins are beta-lactam antibiotics which inhibit cell wall synthesis. Cephalosporins included in this study were cefoxitin, ceftiofur, and ceftriaxone.

Eight of the top 10 *k*-mers for cefoxitin were annotated as CMY-type beta-lactamase genes by CARD and Prokka, and also as *blc* by Prokka. The results are similar to those for co-amoxiclav, described above; the models can identify important features for individual antimicrobials as well as broader groups (such as the beta-lactam group).Two *k*-mers were identified by neither CARD nor Prokka. A BLAST search of the kmers places them in regions surrounding CMY, and potentially within a hypothetical protein.

The top *k*-mers for ceftriaxone also had similarities to those for the penicillins. Three were identified by Prokka as belonging to the insertion sequence IS*Ecp1*, and one was identified by CARD as TEM-55 beta-lactamase. Two *k*-mers were annotated as belonging to either *ramR* or *marR*. As with ampicillin, the *ramR* annotations were noted as being from *Klebsiella*. *marR* was noted as being from *E. coli*, and therefore likely the correct annotation. A third top *k*-mer was annotated as only *ramR*, but it overlaps with these two *k*-mers, and thus is likely also *marR*. The chromosomal *marR* gene encodes MarR, a repressor of the operon *marRAB*. The *marRAB* operon includes MarA, which positively regulates the expression of AcrAB, a multidrug efflux pump [322]. *marRAB* is typically repressed,

but expression may be induced by stressors, such as by some antibiotics and chemical compounds [323]. Additionally, mutations in *marR* result in less repression, and increased marRAB expression, and thus increased resistance [323]. Three ceftriaxone *k*-mers had no hits from Prokka nor CARD; a BLAST search suggests that they occur within or near the insertion sequence IS*Ecp1*.

Three of the top 10 *k*-mers for ceftiofur were for beta-lactamases: two for TEM-55 and one for CMY. Two *k*-mers were annotated by Prokka as belonging to the insertion sequence IS*Ecp1*. One of these two was annotated by CARD as *tetM*; however, it is involved in tetracycline resistance, and unlikely to be involved in ceftiofur resistance [324]. The remaining five *k*-mers had no hits from Prokka nor CARD. A BLAST search places them around IS*Ecp1*: and one occurs just upstream, two overlap with its end, and two occur just downstream (and just upstream of a CTX-M beta-lactamase). The three non-overlapping *k*-mers may indicate important assistance locations for IS*Ecp1* or CTX-M.

**Aminoglycosides**

The top *k*-mers for the four aminoglycoside antibiotics (kanamycin, streptomycin, gentamicin, and tobramycin) primarily were annotated as aminoglycoside acetyltransferase (AAC) and aminoglycoside phosphotransferase (APH) genes. Aminoglycoside acetyltransferases catalyze the acetylation of amine groups of aminoglycoside antibiotics, and aminoglycoside phosphotransferases catalyze the addition of phosphate to aminoglycoside antibiotics [325]. Both of these types of modification block the antibiotic from binding to its target in the bacterial cell.

For kanamycin, seven of the top nine *k*-mers were annotated as *aph(3')-Ia* or *aph(3')-Ic* genes; the phosphotransferases encoded by these genes catalyzes the addition of a phosphate to the 3' hydroxyl group of an aminoglycoside [325]. For streptomycin, the top nine *k*-mers were annotated as *aph(6)-Id* and *aph(3")-Ib*; the enzymes encoded by these genes catalyzes the addition of a phosphate to the 6' location or 3' location, respectively [325].

For gentamicin, seven of the top *k*-mers were annotated as *aac(3)*, *aph(3")*, and *qacEdelta1*. AAC(3) catalyzes the addition of an amine group in the 3 position of the antibiotic [325]. *qacEdelta1* is known to confer resistance to antiseptics, not antimicrobials; however, it is also a marker of class 1 integrons [326]. These integrons are mobile elements known to carry other resistance genes - most notably *sul1* for sulphonamide resistance [326]. The *k*-mers annotated as *qacEdelta1* support that gentamicin resistance genes may be located in class 1 integrons.

All of the top *k*-mers for tobramycin were annotated; some had multiple annotations. Similar to the other aminoglycosides, hits included *aac(3)*, *aph(3')-Ia*, *qacEdelta1*, and *aph(6)-Id*. Additionally, there were hits for *aac(6')-IB* which encodes an acetyltransferase that catalyzes the addition of an amine group to the 6 position of the antibiotic [325]. Two *k*-mers had hits for *aac(6')-Ib-cr*, encoding a variant acetyltransferase which confers resistance to both aminoglycosides and fluoroquinolones in *Enterobacteriaceae* [327, 328]. *aac(6')-Ib-cr* is commonly found within class 1 integrons [329, 328], which also include *qacEdelta1* [326].

**Quinolones**

The quinolones inhibit bacterial DNA replication through binding to DNA gyrase (subunits encoded by *gyrA* and *gyrB*) and/or topoisomerase IV (subunits encoded by *parC* and *parE*) [330]. Mutations in the quinolone resistance-determining regions (QRDRs) of *gyrA*, *gyrB*, *parC*, and *parE* can cause amino acids substitutions which change the conformation of their enzymes [330]. The conformation change impacts the ability of quinolones to bind and inhibit them [330].

Two of the top *k*-mers for ciprofloxacin, two for enrofloxacin, two for levofloxacin, and one for NAL were annotated as *parC*. Each of them mapped to the same region in *parC*, with the susceptible and resistant *k*-mers differing by a single base. Resistance arises from the point mutation from G to T in the coding strand; it causes a substitution from serine to

isoleucine at ParC position 80 [331, 332].

Across the quinolones, all five of the resistance-associated *gyrA* *k*-mers, and seven of the susceptible-associated *k*-mers mapped to the same location. The resistance- and susceptible-associated *k*-mers differed by a point mutation, from C to T in the coding strand. This mutation causes an amino acid substitution from serine to leucine at GyrA position 83 [332]. The two other susceptible-associated *gyrA* *k*-mers, one from ciprofloxacin and one from levofloxacin, contained another point mutation location. They have the GAC codon; a mutation from G to C would confer resistance by causing an amino acid substitution from aspartic acid (aspartate) to asparagine [332].

Ciprofloxacin had one top *k*-mer annotated by Prokka as *ansB*, encoding L-asparaginase II. This enzyme catalyses the hydrolysis of the amino acid asparagine to aspartic acid (aspartate) and ammonia [333]. L-asparaginase II is suspected to contribute to virulence or resistance of *E. coli* [334, 335]; it is a known factor for other bacterial species including *Salmonella enterica* [333], *Heliobacter pylori* [336], and *Campylobacter jejuni* [337]. Levels of the enzyme have been shown to increase when *E. coli* is exposed to ciprofloxacin [335]. Recently, L-asparaginase II isolated from an STEC strain was shown to have 93% similarity in amino acid sequence to L-asparaginase II from *Salmonella enterica* Typhimurium [334]. It was also shown that this STEC L-asparaginase II has a similar role to the enzyme from *Salmonella*: it inhibits T cell proliferation, thereby lowering the immune response of the host to infection [333, 334].

Enrofloxacin had one top resistance-associated *k*-mer annotated as *abaF*; AbaF is an efflux pump in *Acinetobacter baumannii* known to confer resistance to fosfomycin [338]. However, a web BLAST of the *k*-mer instead places it within *shiA*, encoding a shikimate transporter. ShiA is a known virulence factor in *E. coli*, as it suppresses the immune response of the host to infection [339, 340, 341].

**Monobactams**

Aztreonam is a monobactam belonging to the larger group of beta-lactam antibiotics. As with other beta-lactams, it interferes with the synthesis of bacterial cell walls [342, 343]. Aztreonam has functional groups which provide stability for its beta-lactam ring, helping to resist hydrolysis by some bacterial beta-lactamases [342, 343].

CARD annotated one top *k*-mer as *marR*, which is involved in resistance via efflux pump, as described in the cephalosporins section above. Prokka annotated three other *k*-mers as IS*Ecp1*, which is also described above, in the penicillins section. Six *k*-mers had no annotation. BLAST revealed that the two of these *k*-mers are neighbours of IS*Ecp1*, though with no direct gene hit of their own. The remaining four *k*-mers may be uninformative: they were present in every isolate, and BLAST maps them to tRNA.

### 3.6.3 Comparison to Database Methods

When run on the dataset of 4300 isolates, the database methods ResFinder and AMRFinderPlus struggled with some antibiotics. ResFinder had above 80% accuracy for 22 of 36 antimicrobials and AMRFinderPlus had above 80% accuracy for only 10 of 36 antimicrobials. For NIT, AMRFinderPlus did not detect any resistance markers, and classified all isolates as susceptible. ResFinder did the same for CFZ, CXA, CXM, EFX, LVX, NAL, NIT, SAM, SOX, and TIO. Databases need to be expanded to include more resistance markers for their existing antimicrobials, and to include more antimicrobials overall. ML methods could be used as an alternative to database methods, or to identify resistance markers and expand the databases.

Another popular database tool is the Resistance Gene Identifier (RGI), which uses the CARD database. It was not used for comparison because the command line version presently resolves only to class level, and not specific antibiotic. The accuracy of using the CARD database may be higher if accessed through different means.

When predicting SIR for the ResFinder DE and DK validation datasets, some ML mod-

els displayed poor performance. For the DK dataset and the antimicrobial CPM, XGB, SVM, and ANN models had 21.1%, 20.0%, and 18.9% accuracy, respectively; however, this does not immediately indicate that the models should be discarded. For the ResFinder publication, SIR was determined by comparison of MICs to the ECOFF. The ECOFF for CPM is 0.125, the CLSI breakpoints are S$\leq$2 and R$>$16, and the EUCAST breakpoints are S$\leq$1 and R$>$4. When the isolates are reclassified as SIR following EUCAST or CLSI breakpoints, the predictions of the model are more accurate, jumping to 61.1% and 63.2%, respectively; the accuracy of ResFinder drops to 73.7% and 68.4%, respectively. As SIR breakpoints are committee-defined, they vary between regions (committees). Databases and models could alternatively predict MIC values instead, as they are stable experimentally determined measurements; however, SIR classifications are more widely available than MICs for public data currently.

## 3.7 Conclusion

The ML models used in this study were on average more accurate than the database methods, and do not rely on curated databases for their prediction. This ability allows us to move closer to "agnostic diagnostics", where specific outcomes do not need to be known for them to be tested for. However, the ML models are only as good as the data they are trained on, and also require periodic updating based on new high-quality sequencing and laboratory phenotypic data. This approach has the potential to allow cheaper, faster, and constant surveillance activities across the one health continuum. ML approaches also have the ability to increase our understanding of novel mechanisms of microbial resistance, and to quantitatively predict minimum inhibitory concentrations of antimicrobials. Lastly, when paired with on-site sequencing methods, ML methods are poised to make rapid AMR diagnoses in clinical and agricultural settings a reality.

# Chapter 4

# Conclusion and Future Directions

As WGS is becoming more accessible, it is replacing the traditional time-consuming and expensive wet-lab methods for pathogen diagnostics and surveillance. For resistance profiling, the *in silico* tools for the analysis of WGS data need to be rapid and accurate. A common method is comparison against AMR databases; however, these databases have limitations. As discussed in Chapter 1, databases suffer from standardization issues, and they can quickly become irrelevant without constant curation. They also focus on markers of resistance, when markers of susceptibility are also valuable when selecting an effective antibiotic. ML models are an attractive alternative to databases, as no *a priori* information about resistance mechanisms is required. They are also beneficial to use alongside databases as they can identify novel markers of susceptibility or resistance to assist in keeping databases current.

In this thesis, ML models were successfully trained for the rapid and accurate prediction of SIR classifications for *E. coli* data. Their performance met or exceeded leading database methods. The models identified markers of both resistance and susceptibility. For example, they were able to identify SNPs in *parC* and *gyrA*, which confer resistance (as discussed in Chapter 3). In contrast to similar studies, our models were trained with the entire set of k-mers, without pre-selection for core genes, which allows for the identification of important intergenic markers as well as those within genes. Our models are also capable of multi-class classification (including the Intermediate class), while others commonly only perform binary classifications. Additionally, instead of only $k$-mer presence/absence, our input con-

sisted of *k*-mer frequencies, which other tools were unable to handle due to computational constraints. Although we were able to use 25-mers, there were tradeoffs. Unlike the 11-mer models, the 25-mer models did not include parameter optimization; though SVM and XGB models exhibited good performance, the ANN models suffered from a lack of parameter tuning, with an average accuracy of 68.9%. Parameter optimization may be more feasible with large *k*-mers in the future, as tools for storing and handling the frequency matrices are developed and improved.

A limitation of the ML models is the availability of public data. ML models are more accurate when trained on a larger amount of data, so they would benefit from periodic updates as new data are released. Presently, genomes more commonly have SIR companion metadata, as opposed to MIC metadata. If more MIC metadata becomes available, it would be advantageous to additionally train models for MIC classification. MICs are experimentally-determined values which do not change between years, while SIR classifications are based on regional committee-defined breakpoints which can have some variation between countries and years. Though unchanging, MICs still have challenges: the values are subjective, as they are interpretations by lab personnel; they can also be difficult to unify between countries, as antibiotic concentrations used for panels are also set by regional committees. The former challenge can be compensated for by considering the MIC to be correct within one dilution.

The ML models can be applied to more than just AMR if additional metadata are made publicly available. Training models for the classification of WGS data by host/source and location of origin would be advantageous for outbreak tracing and predicting zoonotic potential. Lupolova *et al.* [205] previously investigated the use of SVMs with protein presence/absence for the host/source classification of *Salmonella* and *E. coli*. For *Salmonella*, the accuracy of predicting host/source was 89% for avian, 67% for bovine, 90% for human, and 75% for swine isolates. For *E. coli*, the accuracy was 72% for bovine and 89% for human; they were unable to classify isolates for other groups due to the small number

of isolates available. A recent study by Bayliss *et al.* [344] looked at the use of a hierarchical ML model for location-based classification of clinical *S. enterica* serovar Enteritidis isolates. Prediction at the regional level (Europe, Africa, Americas, Asia) had the largest sample size and was the most accurate. Prediction at the sub-regional level and country level was less accurate, as less data were available for these levels, and it was more variable. With increased data availability, our ML models can be applied to host/source and location classification using WGS *E. coli*. Our models would come with the additional benefit of feature extraction and annotation, to identify genomic markers contributing to host and location. An additional find by Balyiss *et al.* [344] was that prediction accuracies improve when the training dataset contains isolates collected over multiple years. For their dataset of *S. enterica* serovar Enteritidis, a two year window for isolate collection was found to be optimal. In future work, it would be beneficial to investigate the effect on accuracy of our ML models when specific years are selected, as our dataset of 4300 genomes spanned from 1970 to 2018. Additionally, the models may be used to investigate the correlation of AMR with the other categories, to support the surveillance of AMR across the one health continuum. This would allow for rapid detection and tracing spanning from a local to a global scale, and covering all sources, including human, animal, and environmental.

The aforementioned methods are typically applied to WGS data obtained by sequencing the DNA of an isolated culture. As discussed in chapter 1, much time can be saved by eliminating the wet-lab isolation steps and replacing it with shotgun sequencing to obtain the metagenome from a sample. The use of ONT sequencing can further reduce the time for sequencing from over three days (as with Illumina) to as little as one day. Although a full ONT sequencing run can take multiple days, analysis can be performed in real-time. Combining metagenomic sequencing with a rapid *in silico* analysis pipeline would provide a powerful tool for the identification of resistant pathogens, so interventions could be established to prevent outbreaks. In addition, ONT sequencing is extremely portable, with field sequencing kits and sequencers able to run with a laptop. Future studies would run

an analysis pipeline at various time points after commencing an ONT run, to determine the earliest time at which species and resistance profiles could be identified.

Metagenomics comes with many challenges, including the samples themselves. The sample type greatly impacts the complexity of analysis. It is typically faster to identify and analyze bacteria of interest within a sample with relatively low bacterial background, such as a blood sample, and harder to do the same for a sample with a high bacterial background, such as fecal samples. This is not a set rule: if the bacterial load of a clinical sample is low, longer sequencing times, addition of culture amplification, or use of background (for example host DNA) depletion may be needed [345, 346, 347]. Background depletion is also advantageous for samples with high bacterial background, as it can allow for higher coverage of rarer bacterial species [348]. ONT offers an adaptive sampling during sequencing for this purpose; however, it requires more computational resources to run. In general, sample pre-processing, and sequencing run settings need to be tailored to sample type. Future studies would need to investigate which options are optimal for different sample types.

Differences in sample type and sequencing technology additionally leads to difficulty in the development of analysis pipelines universally applicable to metagenomes. Future work for the development of such a pipeline would need to include options to handle different sample types and sequencing technologies. The options would need to affect every stage. Commonly, the first post-sequencing step is trimming and filtering. For Illumina runs, this may include the removal of adapters, primers, poly-A tails, sequences longer than a set value, sequences shorter than a set value, and/or low-quality ends of reads; tools for this task include Trim-galore [298] (wrapper for Cutadapt [299] and FastQC [300]) and Trimmomatic [349]. For PacBio and ONT runs, this may include removal of sequences below a set quality threshold, sequences longer than a set value, sequences shorter than a set value, and/or a number of nucleotides at the start or end of a read; tools for this task include Chopper and nanoq, which are part of NanoPack [350]. The other commonly required tasks are assembly and taxonomic binning; as with post-processing, tool choice is impacted by

sequencing technology, and there are many parameters available for customization. Tools for these tasks are discussed in detail in Chapter 1. Pipeline development would need to consider processing order as well, such as whether to assemble a metagenome first, or to bin the data and assemble individual species - it may also be optimal to instead work with raw data and avoid assembly altogether. Measurements of success would include amount of recovery of a species from a metagenome, and how well *in silico* tools, such as ML models, are able to perform on the recovered data.

In this thesis I demonstrated the predictive power of ML models trained for SIR classifications of WGS *E. coli*. In the future these rapid and accurate approaches can be extended to classify organisms extracted from metagenomes, and to classify WGS data based on other criteria, such as country and host. They also have the potential for integration into clinical diagnostics and surveillance programs, to aid in the replacement of traditional wet-lab methods with sequencing and *in silico* analysis.

# Bibliography


[1] K. Bettelheim and S. Lennox-King. The acquisition of *Escherichia coli* by new-born babies. *Infection*, 4(3):174–179, September 1976.

[2] James B. Kaper, James P. Nataro, and Harry L. T. Mobley. Pathogenic *Escherichia coli*. *Nature Reviews Microbiology*, 2(2):123, February 2004.

[3] L. Hannah Gould, Linda Demma, Timothy F. Jones, Sharon Hurd, Duc J. Vugia, Kirk Smith, Beletshachew Shiferaw, Suzanne Segler, Amanda Palmer, Shelley Zansky, Patricia M. Griffin, and the Emerging Infections Program FoodNet Working Group. Hemolytic uremic syndrome and death in persons with *Escherichia coli* O157:H7 infection, foodborne diseases active surveillance network sites, 2000–2006. *Clinical Infectious Diseases*, 49(10):1480–1485, November 2009.

[4] Roy M. Robins-Browne, Kathryn E. Holt, Danielle J. Ingle, Dianna M. Hocking, Ji Yang, and Marija Tauschek. Are *Escherichia coli* pathotypes still relevant in the era of whole-genome sequencing? *Frontiers in Cellular and Infection Microbiology*, 6:141, November 2016.

[5] Labib Sharif, J. Obeidat, and Falah Al-Ani. Risk factors for lamb and kid mortality in sheep and goat farms in Jordan. *Bulgarian Journal of Veterinary Medicine*, 8(2):99–108, January 2005.

[6] Francois Madec, Nathalie Bridoux, Stéphane Bounaix, Roland Cariolet, Yvonne Duval-Iflah, David J. Hampson, and André Jestin. Experimental models of porcine post-weaning colibacillosis and their relationship to post-weaning diarrhoea and digestive disorders as encountered in the field. *Veterinary Microbiology*, 72(3):295–310, March 2000.

[7] G. Bashahun. Colibacillosis in calves: A review of literature. *Journal of Animal Science and Veterinary Medicine*, 2:62–71, March 2017.

[8] W. J. Sojka and R. B. A. Carnaghan. *Escherichia coli* infection in poultry. *Research in Veterinary Science*, 2(4):340–352, October 1961.

[9] Zachary R. Stromberg, James R. Johnson, John M. Fairbrother, Jacquelyn Kilbourne, Angelica Van Goor, Roy Curtiss III, and Melha Mellata. Evaluation of *Escherichia coli* isolates from healthy chickens to determine their potential risk to poultry and human health. *PLOS ONE*, 12(7), July 2017.

[10] Randall S. Singer, Joan S. Jeffrey, Tim E. Carpenter, Cara L. Cooke, E. Rob Atwill, Wesley O. Johnson, and Dwight C. Hirsh. Persistence of cellulitis-associated *Escherichia coli* DNA fingerprints in successive broiler chicken flocks. *Veterinary Microbiology*, 75(1):59–71, July 2000.

[11] D. W. Watson and C. A. Brandly. Virulence and pathogenicity. *Annual Review of Microbiology*, 3(1):195–220, 1949.

[12] H. D. Isenberg. Pathogenicity and virulence: Another view. *Clinical Microbiology Reviews*, 1(1):40–53, January 1988.

[13] Trudy M. Wassenaar and Wim Gaastra. Bacterial virulence: Can we draw the line? *FEMS Microbiology Letters*, 201(1):1–7, July 2001.

[14] Arturo Casadevall and Liise-anne Pirofski. Host-pathogen interactions: The attributes of virulence. *The Journal of Infectious Diseases*, 184(3):337–344, August 2001.

[15] Stephen R. Thomas and Joseph S. Elkinton. Pathogenicity and virulence. *Journal of Invertebrate Pathology*, 85(3):146–151, March 2004.

[16] Raghavan U. M. Palaniappan, Yu Zhang, David Chiu, Alfonso Torres, Chobi DebRoy, Thomas S. Whittam, and Yung-Fu Chang. Differentiation of *Escherichia coli* pathotypes by oligonucleotide spotted array. *Journal of Clinical Microbiology*, 44(4):1495–1501, April 2006.

[17] B. B. Finlay and S. Falkow. Common themes in microbial pathogenicity revisited. *Microbiology and Molecular Biology Reviews*, 61(2):136–169, June 1997.

[18] Andreas Leimbach, Jörg Hacker, and Ulrich Dobrindt. *E. coli* as an all-rounder: The thin line between commensalism and pathogenicity. In Ulrich Dobrindt, Jörg H. Hacker, and Catharina Svanborg, editors, *Between Pathogenicity and Commensalism*, Current Topics in Microbiology and Immunology, pages 3–32. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[19] Bjørn-Arne Lindstedt, Misti D. Finton, Davide Porcellato, and Lin T. Brandal. High frequency of hybrid *Escherichia coli* strains with combined intestinal pathogenic *Escherichia coli* (IPEC) and extraintestinal pathogenic *Escherichia coli* (ExPEC) virulence factors isolated from human faecal samples. *BMC Infectious Diseases*, 18(1):544, November 2018.

[20] Charlotte Collingwood, Kirsty Kemmett, Nicola Williams, and Paul Wigley. Is the concept of avian pathogenic *Escherichia coli* as a single pathotype fundamentally flawed? *Frontiers in Veterinary Science*, 1:5, October 2014.

[21] Helge Karch, Erick Denamur, Ulrich Dobrindt, B. Brett Finlay, Regine Hengge, Ludgers Johannes, Eliora Z. Ron, Tone Tønjum, Philippe J. Sansonetti, and Miguel Vicente. The enemy within us: Lessons from the 2011 European *Escherichia coli* O104:H4 outbreak. *EMBO Molecular Medicine*, 4(9):841–848, September 2012.

[22] Angela R. Melton-Celsa. Shiga toxin (Stx) classification, structure, and function. *Microbiology Spectrum*, 2(2), July 2014.

[23] Hayley J. Newton, Joan Sloan, Dieter M. Bulach, Torsten Seemann, Cody C. Allison, Marija Tauschek, Roy M. Robins-Browne, James C. Paton, Thomas S. Whittam, Adrienne W. Paton, and Elizabeth L. Hartland. Shiga toxin–producing *Escherichia coli* strains negative for locus of enterocyte effacement. *Emerging Infectious Diseases*, 15(3):372–380, March 2009.

[24] Eelco Franz, Angela H. A. M. van Hoek, Mark Wuite, Fimme J. van der Wal, Albert G. de Boer, E. I. Bouw, and Henk J. M. Aarts. Molecular hazard identification of non-O157 Shiga toxin-producing *Escherichia coli* (STEC). *PLOS ONE*, 10(3), March 2015.

[25] Roslen Bondì, Paola Chiani, Valeria Michelacci, Fabio Minelli, Alfredo Caprioli, and Stefano Morabito. The gene *tia*, harbored by the subtilase-encoding pathogenicity island, is involved in the ability of locus of enterocyte effacement-negative Shiga toxin-producing *Escherichia coli* strains to invade monolayers of epithelial cells. *Infection and Immunity*, 85(12), November 2017.

[26] P. T. Kimmitt, C. R. Harwood, and M. R. Barer. Toxin gene expression by Shiga toxin-producing *E. coli* the role of antibiotics and the bacterial SOS response. *Emerging Infectious Diseases*, 6(5):458–465, September 2000.

[27] Melody N. Neely and David I. Friedman. Functional and genetic analysis of regulatory regions of coliphage H-19B: Location of Shiga-like toxin and lysis genes suggest a role for phage functions in toxin release. *Molecular Microbiology*, 28(6):1255–1267, November 1998.

[28] Chitrita DebRoy, Pina M. Fratamico, Xianghe Yan, GianMarco Baranzoni, Yanhong Liu, David S. Needleman, Robert Tebbs, Catherine D. O'Connell, Adam Allred, Michelle Swimley, Michael Mwangi, Vivek Kapur, Juan A. Raygoza Garay, Elisabeth L. Roberts, and Robab Katani. Comparison of O-antigen gene clusters of all O-serogroups of *Escherichia coli* and proposal for adopting a new nomenclature for O-typing. *PLOS ONE*, 11(1), January 2016.

[29] Pina M. Fratamico, Chitrita DebRoy, Yanhong Liu, David S. Needleman, Gian Marco Baranzoni, and Peter Feng. Advances in molecular serotyping and subtyping of *Escherichia coli*. *Frontiers in Microbiology*, 7:664, May 2016.

[30] Mohamed A. Karmali, Mariola Mascarenhas, Songhai Shen, Kim Ziebell, Shelley Johnson, Richard Reid-Smith, Judith Isaac-Renton, Clifford Clark, Kris Rahn, and James B. Kaper. Association of genomic O island 122 of *Escherichia coli* EDL 933 with verocytotoxin-producing *Escherichia coli* seropathotypes that are linked to epidemic and/or serious disease. *Journal of Clinical Microbiology*, 41(11):4930–4940, November 2003.

[31] Government of Canada. *National Enteric Surveillance Program Annual Summary 2018*. Public Health Agency of Canada, Guelph, 2020.

[32] P. Sockett, S. E. Goebel, N. P. Varela, A. Guthrie, J. Wilson, L. A. Guilbault, and W. F. Clark. Verotoxigenic *Escherichia coli*: Costs of illness in Canada, including long-term health outcomes. *Journal of Food Protection*, 77(2):216–226, February 2014.

[33] Ana Carolina de Mello Santos, Fernanda Fernandes Santos, Rosa Maria Silva, and Tânia Aparecida Tardelli Gomes. Diversity of hybrid- and hetero-pathogenic *Escherichia coli* and their potential implication in more severe diseases. *Frontiers in Cellular and Infection Microbiology*, 10:339, July 2020.

[34] Martina Bielaszewska, Alexander Mellmann, Wenlan Zhang, Robin Köck, Angelika Fruth, Andreas Bauwens, Georg Peters, and Helge Karch. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: A microbiological study. *The Lancet Infectious Diseases*, 11(9):671–676, September 2011.

[35] Craig S. Wong, Jody C. Mooney, John R. Brandt, Amy O. Staples, Srdjan Jelacic, Daniel R. Boster, Sandra L. Watkins, and Phillip I. Tarr. Risk factors for the hemolytic uremic syndrome in children infected with *Escherichia coli* O157:H7: A multivariable analysis. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 55(1):33–41, July 2012.

[36] Martina Bielaszewska, Evgeny A. Idelevich, Wenlan Zhang, Andreas Bauwens, Frieder Schaumburg, Alexander Mellmann, Georg Peters, and Helge Karch. Effects of antibiotics on Shiga toxin 2 production and bacteriophage induction by epidemic *Escherichia coli* O104:H4 strain. *Antimicrobial Agents and Chemotherapy*, 56(6):3277–3282, June 2012.

[37] Teresa M. Wozniak, Louise Barnsbee, Xing J. Lee, and Rosana E. Pacella. Using the best available data to estimate the cost of antimicrobial resistance: A systematic review. *Antimicrobial Resistance and Infection Control*, 8:26, February 2019.

[38] Francesca Prestinaci, Patrizio Pezzotti, and Annalisa Pantosti. Antimicrobial resistance: A global multifaceted phenomenon. *Pathogens and Global Health*, 109(7):309–318, October 2015.

[39] Radha Rangarajan and Rasika Venkataraman. Chapter 3 - antibiotics targeting gram-negative bacteria. In Prashant Kesharwani, Sidharth Chopra, and Arunava Dasgupta, editors, *Drug Discovery Targeting Drug-Resistant Bacteria*, pages 39–70. Academic Press, January 2020.

[40] Government of Canada (Public Health Agency of Canada). Reductions in antimicrobial use and resistance: Preliminary evidence of the effect of the Canadian chicken industry's elimination of use of antimicrobials of

very high importance to human medicine. https://www.canada.ca/en/public-health/services/publications/drugs-health-products/canadian-integrated-program-antimicrobial-resistances-surveillance-bulletin.html, August 2016.

[41] Government of Canada (Public Health Agency of Canada). Canadian Integrated Program for Antimicrobial Resistance Surveillance (CIPARS) 2018: Integrated findings. https://www.canada.ca/en/public-health/services/surveillance/canadian-integrated-program-antimicrobial-resistance-surveillance-cipars/cipars-reports/2018-annual-report-integrated-findings.html, December 2020.

[42] Government of Canada (Health Canada). Categorization of antimicrobial drugs based on importance in human medicine. https://www.canada.ca/en/health-canada/services/drugs-health-products/veterinary-drugs/antimicrobial-resistance/categorization-antimicrobial-drugs-based-importance-human-medicine.html, September 2009.

[43] Government of Canada (Public Health Agency of Canada). CIPARS 2019: Integrated findings. https://www.canada.ca/en/public-health/services/surveillance/canadian-integrated-program-antimicrobial-resistance-surveillance-cipars/cipars-2019-integrated-findings.html, July 2022.

[44] Deepti Rawat and Deepthi Nair. Extended-spectrum β-lactamases in gram negative bacteria. *Journal of Global Infectious Diseases*, 2(3):263–274, 2010.

[45] N. H. Georgopapadakou. Penicillin-binding proteins and bacterial resistance to beta-lactams. *Antimicrobial Agents and Chemotherapy*, 37(10):2045–2053, October 1993.

[46] J. T. Freeman, D. A. Williamson, H. Heffernan, M. Smith, J. E. Bower, and S. A. Roberts. Comparative epidemiology of CTX-M-14 and CTX-M-15 producing *Escherichia coli*: Association with distinct demographic groups in the community in New Zealand. *European Journal of Clinical Microbiology & Infectious Diseases*, 31(8):2057–2060, August 2012.

[47] Priyanka Bajaj, Nambram S. Singh, and Jugsharan S. Virdi. *Escherichia coli* β-lactamases: What really matters. *Frontiers in Microbiology*, 7:417, March 2016.

[48] Andrew J. Denisuik, James A. Karlowsky, Heather J. Adam, Melanie R. Baxter, Philippe R. S. Lagacé-Wiens, Michael R. Mulvey, Daryl J. Hoban, George G. Zhanel, and Canadian Antimicrobial Resistance Alliance (CARA) and CANWARD. Dramatic rise in the proportion of ESBL-producing *Escherichia coli* and *Klebsiella pneumoniae* among clinical isolates identified in Canadian hospital laboratories from 2007 to 2016. *Journal of Antimicrobial Chemotherapy*, 74(Supplement_4):iv64–iv71, August 2019.

[49] Maria Karczmarczyk, Marta Martins, Teresa Quinn, Nola Leonard, and Séamus Fanning. Mechanisms of fluoroquinolone resistance in *Escherichia coli* isolates from

food-producing animals. *Applied and Environmental Microbiology*, 77(20):7113–7120, October 2011.

[50] Katie L. Hopkins, Robert H. Davies, and E. John Threlfall. Mechanisms of quinolone resistance in *Escherichia coli* and *Salmonella*: Recent developments. *International Journal of Antimicrobial Agents*, 25(5):358–373, May 2005.

[51] Kevin M. Krause, Alisa W. Serio, Timothy R. Kane, and Lynn E. Connolly. Aminoglycosides: An overview. *Cold Spring Harbor Perspectives in Medicine*, 6(6), June 2016.

[52] Cláudia Gomes, Lidia Ruiz-Roldán, Judit Mateu, Theresa J. Ochoa, and Joaquim Ruiz. Azithromycin resistance levels and mechanisms in *Escherichia coli*. *Scientific Reports*, 9(1):6089, April 2019.

[53] M Vaara. Outer membrane permeability barrier to azithromycin, clarithromycin, and roxithromycin in gram-negative enteric bacteria. *Antimicrobial Agents and Chemotherapy*, 37(2):354–356, February 1993.

[54] Susan Farmer, Zusheng Li, and Robert E. W. Hancock. Influence of outer membrane mutations on susceptibility of *Escherichia coli* to the dibasic macrolide azithromycin. *Journal of Antimicrobial Chemotherapy*, 29(1):27–33, January 1992.

[55] Cláudia Gomes, Sandra Martínez-Puchol, Noemí Palma, Gertrudis Horna, Lidia Ruiz-Roldán, Maria J Pons, and Joaquim Ruiz. Macrolide resistance mechanisms in Enterobacteriaceae: Focus on azithromycin. *Critical Reviews in Microbiology*, 43(1):1–30, January 2017.

[56] J. Retsema, A. Girard, W. Schelkly, M. Manousos, M. Anderson, G. Bright, R. Borovoy, L. Brennan, and R. Mason. Spectrum and mode of action of azithromycin (CP-62,993), a new 15-membered-ring macrolide with improved potency against gram-negative organisms. *Antimicrobial Agents and Chemotherapy*, 31(12):1939–1947, December 1987.

[57] Alexandros D. Petropoulos, Ekaterini C. Kouvela, Agata L. Starosta, Daniel N. Wilson, George P. Dinos, and Dimitrios L. Kalpaxis. Time-resolved binding of azithromycin to *Escherichia coli* ribosomes. *Journal of Molecular Biology*, 385(4):1179–1192, January 2009.

[58] Minh Chau Phuc Nguyen, Paul-Louis Woerther, Mathilde Bouvet, Antoine Andremont, Roland Leclercq, and Annie Canu. *Escherichia coli* as reservoir for macrolide resistance genes. *Emerging Infectious Diseases*, 15(10):1648–1650, October 2009.

[59] Corey Fyfe, Trudy H. Grossman, Kathy Kerstein, and Joyce Sutcliffe. Resistance to macrolide antibiotics in public health pathogens. *Cold Spring Harbor Perspectives in Medicine*, 6(10), October 2016.

[60] C. Gyles and P. Boerlin. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Veterinary Pathology*, 51(2):328–340, March 2014.

[61] Suhartono Suhartono and Mary Savin. Conjugative transmission of antibiotic-resistance from stream water *Escherichia coli* as related to number of sulfamethoxazole but not class 1 and 2 integrase genes. *Mobile Genetic Elements*, 6(6):e1256851, November 2016.

[62] Vincent Burrus, Guillaume Pavlovic, Bernard Decaris, and Gérard Guédon. Conjugative transposons: The tip of the iceberg. *Molecular Microbiology*, 46(3):601–610, October 2002.

[63] Angela van Hoek, Dik Mevius, Beatriz Guerra, Peter Mullany, Adam Roberts, and Henk Aarts. Acquired antibiotic resistance genes: An overview. *Frontiers in Microbiology*, 2:203, September 2011.

[64] Heather E. Allison. Stx-phages: Drivers and mediators of the evolution of STEC and STEC-like pathogens. *Future Microbiology*, 2(2):165–174, March 2007.

[65] Darren L. Smith, David J. Rooks, Paul CM Fogg, Alistair C. Darby, Nick R. Thomson, Alan J. McCarthy, and Heather E. Allison. Comparative genomics of Shiga toxin encoding bacteriophages. *BMC Genomics*, 13(1):311, July 2012.

[66] Craig Baker-Austin, Meredith S. Wright, Ramunas Stepanauskas, and J. V. McArthur. Co-selection of antibiotic and metal resistance. *Trends in Microbiology*, 14(4):176–182, April 2006.

[67] Rafael Cantón, José María González-Alba, and Juan Carlos Galán. CTX-M enzymes: Origin and diffusion. *Frontiers in Microbiology*, 3:110, April 2012.

[68] George A. Jacoby. β-lactamase nomenclature. *Antimicrobial Agents and Chemotherapy*, 50(4):1123–1129, April 2006.

[69] M. Matthew, R. W. Hedges, and J. T. Smith. Types of beta-lactamase determined by plasmids in gram-negative bacteria. *Journal of Bacteriology*, 138(3):657–662, June 1979.

[70] Naomi Datta and Polyxeni Kontomichalou. Penicillinase synthesis controlled by infectious R factors in Enterobacteriaceae. *Nature*, 208(5007):239–241, October 1965.

[71] D. Sirot, J. Sirot, R. Labia, A. Morand, P. Courvalin, A. Darfeuille-Michaud, R. Perroux, and R. Cluzel. Transferable resistance to third-generation cephalosporins in clinical isolates of *Klebsiella pneumoniae*: Identification of CTX-1, a novel β-lactamase. *Journal of Antimicrobial Chemotherapy*, 20(3):323–334, September 1987.

[72] R. W. Hedges, Naomi Datta, Polyxeni Kontomichalou, and J. T. Smith. Molecular specificities of R factor-determined beta-lactamases: Correlation with plasmid compatibility. *Journal of Bacteriology*, 117(1):56–62, January 1974.

[73] M. Matthew and R. W. Hedges. Analytical isoelectric focusing of R factor-determined beta-lactamases: Correlation with plasmid compatibility. *Journal of Bacteriology*, 125(2):713–718, February 1976.

[74] Patricia S. Langan, Venu Gopal Vandavasi, Kevin L. Weiss, Jonathan B. Cooper, Stephan L. Ginell, and Leighton Coates. The structure of Toho1 β-lactamase in complex with penicillin reveals the role of Tyr105 in substrate recognition. *FEBS Open Bio*, 6(12):1170–1177, November 2016.

[75] Catherine Branger, Oana Zamfir, Sabine Geoffroy, Geneviève Laurans, Guillaume Arlet, Hoang Vu Thien, Stéphanie Gouriou, Bertrand Picard, and Erick Denamur. Genetic background of *Escherichia coli* and extended-spectrum β-lactamase type. *Emerging Infectious Diseases*, 11(1):54–61, January 2005.

[76] Laurent Poirel, Peter Kämpfer, and Patrice Nordmann. Chromosome-encoded Ambler class A β-lactamase of *Kluyvera georgiana*, a probable progenitor of a subgroup of CTX-M extended-spectrum β-lactamases. *Antimicrobial Agents and Chemotherapy*, 46(12):4038–4040, December 2002.

[77] Christel Humeniuk, Guillaume Arlet, Valerie Gautier, Patrick Grimont, Roger Labia, and Alain Philippon. β-lactamases of *Kluyvera ascorbata*, probable progenitors of some plasmid-encoded CTX-M types. *Antimicrobial Agents and Chemotherapy*, 46(9):3045–3049, September 2002.

[78] Ari Robicsek, George A. Jacoby, and David C. Hooper. The worldwide emergence of plasmid-mediated quinolone resistance. *The Lancet Infectious Diseases*, 6(10):629–640, October 2006.

[79] Ari Robicsek, Jacob Strahilevitz, George A. Jacoby, Mark Macielag, Darren Abbanat, Chi Hye Park, Karen Bush, and David C. Hooper. Fluoroquinolone-modifying enzyme: A new adaptation of a common aminoglycoside acetyltransferase. *Nature Medicine*, 12(1):83–88, January 2006.

[80] Patrice Nordmann and Laurent Poirel. Emergence of plasmid-mediated resistance to quinolones in Enterobacteriaceae. *Journal of Antimicrobial Chemotherapy*, 56(3):463–469, September 2005.

[81] Yohei Doi, Keiko Yokoyama, Kunikazu Yamane, Jun-ichi Wachino, Naohiro Shibata, Tetsuya Yagi, Keigo Shibayama, Haru Kato, and Yoshichika Arakawa. Plasmid-mediated 16S rRNA methylase in *Serratia marcescens* conferring high-level resistance to aminoglycosides. *Antimicrobial Agents and Chemotherapy*, 48(2):491–496, February 2004.

[82] Kunikazu Yamane, Jun-ichi Wachino, Yohei Doi, Hiroshi Kurokawa, and Yoshichika Arakawa. Global spread of multiple aminoglycoside resistance genes. *Emerging Infectious Diseases*, 11(6):951–953, June 2005.

[83] Ying Xiang, Feng Wu, Yinghui Chai, Xuebin Xu, Lang Yang, Sai Tian, Haoran Zhang, Yinxia Li, Chaojie Yang, Hongbo Liu, Shaofu Qiu, Hongbin Song, and Yansong Sun. A new plasmid carrying mphA causes prevalence of azithromycin resistance in enterotoxigenic *Escherichia coli* serogroup O6. *BMC Microbiology*, 20(1):247, August 2020.

[84] Robert V. Tauxe, Timothy R. Cavanagh, and Mitchell L. Cohen. Interspecies gene transfer in vivo producing an outbreak of multiply resistant Shigellosis. *The Journal of Infectious Diseases*, 160(6):1067–1070, December 1989.

[85] Matthew A. Croxen, Robyn J. Law, Roland Scholz, Kristie M. Keeney, Marta Wlodarska, and B. Brett Finlay. Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clinical Microbiology Reviews*, 26(4):822–880, October 2013.

[86] Dale Hancock, Tom Besser, Jeff Lejeune, Margaret Davis, and Dan Rice. The control of VTEC in the animal reservoir. *International Journal of Food Microbiology*, 66(1):71–78, May 2001.

[87] B. Boudailliez, P. Berquin, P. Mariani-Kurkdjian, D. D. Ilef, B. Cuvelier, I. Capek, B. Tribout, E. Bingen, and C. Piussan. Possible person-to-person transmission of *Escherichia coli* O111 – associated hemolytic uremic syndrome. *Pediatric Nephrology*, 11(1):36–39, January 1997.

[88] Edward A. Belongia, Michael T. Osterholm, John T. Soler, David A. Ammend, Jane E. Braun, and Kristine L. MacDonald. Transmission of *Escherichia coli* 0157:H7 infection in Minnesota child day-care facilities. *JAMA*, 269(7):883–888, February 1993.

[89] Stuart B. Levy, George B. Fitzgerald, and Ann B. Macone. Spread of antibiotic-resistant plasmids from chicken to chicken and from chicken to man. *Nature*, 260(5546):40–42, March 1976.

[90] Noelle R. Noyes, Xiang Yang, Lyndsey M. Linke, Roberta J. Magnuson, Adam Dettenwanger, Shaun Cook, Ifigenia Geornaras, Dale E. Woerner, Sheryl P. Gow, Tim A. McAllister, Hua Yang, Jaime Ruiz, Kenneth L. Jones, Christina A. Boucher, Paul S. Morley, and Keith E. Belk. Resistome diversity in cattle and the environment decreases during beef production. *eLife*, 5.

[91] Jean-Marc Rolain. Food and human gut as reservoirs of transferable antibiotic resistance encoding genes. *Frontiers in Microbiology*, 4:173, June 2013.

[92] Kevin J. Forsberg, Alejandro Reyes, Bin Wang, Elizabeth M. Selleck, Morten O.A. Sommer, and Gautam Dantas. The shared antibiotic resistome of soil bacteria and human pathogens. *Science (New York, N.Y.)*, 337(6098):1107–1111, August 2012.

[93] Bonnie M. Marshall and Stuart B. Levy. Food animals and antimicrobials: Impacts on human health. *Clinical Microbiology Reviews*, 24(4):718–733, October 2011.

[94] Justin Falardeau, Roger P. Johnson, Franco Pagotto, and Siyun Wang. Occurrence, characterization, and potential predictors of verotoxigenic *Escherichia coli*, Listeria monocytogenes, and *Salmonella* in surface water used for produce irrigation in the Lower Mainland of British Columbia, Canada. *PLOS ONE*, 12(9):e0185437, September 2017.

[95] M. F. Miller, G. H. Loneragan, D. D. Harris, K. D. Adams, J. C. Brooks, and M. M. Brashears. Environmental dust exposure as a factor contributing to an increase in *Escherichia coli* O157 and *Salmonella* populations on cattle hides in feedyards. *Journal of Food Protection*, 71(10):2078–2081, October 2008.

[96] Ruth Hummel, H. Tschäpe, and W. Witte. Spread of plasmid-mediated nourseothricin resistance due to antibiotic use in animal husbandry. *Journal of Basic Microbiology*, 26(8):461–466, 1986.

[97] G. G. Khachatourians. Agricultural use of antibiotics and the evolution and transfer of antibiotic-resistant bacteria. *CMAJ: Canadian Medical Association Journal*, 159(9):1129–1136, November 1998.

[98] S. Harris Ali. A socio-ecological autopsy of the *E. coli* O157:H7 outbreak in Walkerton, Ontario, Canada. *Social Science & Medicine*, 58(12):2601–2612, June 2004.

[99] A. E. van den Bogaard, N. London, C. Driessen, and E. E. Stobberingh. Antibiotic resistance of faecal *Escherichia coli* in poultry, poultry farmers and poultry slaughterers. *Journal of Antimicrobial Chemotherapy*, 47(6):763–771, June 2001.

[100] Government of Canada (Health Canada). 2020 veterinary antimicrobial sales highlights report. https://www.canada.ca/en/health-canada/services/publications/drugs-health-products/2020-veterinary-antimicrobial-sales-highlights-report.html, February 2022.

[101] Stephanie A. Brault, Sherry J. Hannon, Sheryl P. Gow, Brian N. Warr, Jessica Withell, Jiming Song, Christina M. Williams, Simon J. G. Otto, Calvin W. Booker, and Paul S. Morley. Antimicrobial use on 36 beef feedlots in western Canada: 2008–2012. *Frontiers in Veterinary Science*, 6:329, October 2019.

[102] Ali Ahmad Sheikh, Sylvia Checkley, Brent Avery, Gabhan Chalmers, Valerie Bohaychuk, Patrick Boerlin, Richard Reid-Smith, and Mueen Aslam. Antimicrobial resistance and resistance genes in *Escherichia coli* isolated from retail meat purchased in Alberta, Canada. *Foodborne Pathogens and Disease*, 9(7):625–631, July 2012.

[103] Government of Canada (Public Health Agency of Canada). Responsible use of medically important antimicrobials in animals. https://www.canada.ca/en/public-health/services/antibiotic-antimicrobial-resistance/animals/actions/responsible-use-antimicrobials.html, October 2017.

[104] J. D. Greig, L. Waddell, B. Wilhelm, W. Wilkins, O. Bucher, S. Parker, and A. Rajić. The efficacy of interventions applied during primary processing on contamination of beef carcasses with *Escherichia coli*: A systematic review-meta-analysis of the published research. *Food Control*, 27(2):385–397, October 2012.

[105] A. Schumacher, T. Vranken, A. Malhotra, J. J. C. Arts, and P. Habibovic. In vitro antimicrobial susceptibility testing methods: Agar dilution to 3D tissue-engineered models. *European Journal of Clinical Microbiology & Infectious Diseases*, 37(2):187–208, February 2018.

[106] Susan M. Novak and Elizabeth M. Marlowe. Automation in the clinical microbiology laboratory. *Clinics in Laboratory Medicine*, 33(3):567–588, September 2013.

[107] CLSI. *Performance Standards for Antimicrobial Susceptibility Testing, CLSI Guideline M100.* Clinical and Laboratory Standards Institute, 27 edition, 2017.

[108] The European Committee on Antimicrobial Susceptibility Testing. Clinical breakpoints and guidance. http://www.eucast.org/clinical_breakpoints/.

[109] Arne Rodloff, Torsten Bauer, Santiago Ewig, Peter Kujath, and Eckhard Müller. Susceptible, intermediate, and resistant – the intensity of antibiotic action. *Deutsches Ärzteblatt International*, 105(39):657–662, September 2008.

[110] Michael Hombach, Brice Mouttet, and Guido V. Bloemberg. Consequences of revised CLSI and EUCAST guidelines for antibiotic susceptibility patterns of ESBL- and AmpC β-lactamase-producing clinical Enterobacteriaceae isolates. *Journal of Antimicrobial Chemotherapy*, 68(9):2092–2098, September 2013.

[111] C. O'Halloran, N. Walsh, M. C. O'Grady, L. Barry, C. Hooton, G. D. Corcoran, and B. Lucey. Assessment of the comparability of CLSI, EUCAST and Stokes antimicrobial susceptibility profiles for *Escherichia coli* uropathogenic isolates. *British Journal of Biomedical Science*, 75(1):24–29, January 2018.

[112] Gunnar Kahlmeter, Christian G. Giske, Thomas J. Kirn, and Susan E. Sharp. Point-counterpoint: Differences between the European Committee on Antimicrobial Susceptibility Testing and Clinical and Laboratory Standards Institute recommendations for reporting antimicrobial susceptibility results. *Journal of Clinical Microbiology*, 57(9), September 2019.

[113] Alexandra Mihaela Velican, Luminiţa Măruţescu, Crina Kamerzan, Violeta Corina Cristea, Otilia Banu, Elvira Borcan, and Mariana-Carmen Chifiriuc. Rapid detection and antibiotic susceptibility of uropathogenic *Escherichia coli* by flow cytometry. *Microorganisms*, 8(8):1233, August 2020.

[114] Anand Kumar, Daniel Roberts, Kenneth E. Wood, Bruce Light, Joseph E. Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Sergio Zanotti, Leo Taiberg, David Gurka, Aseem Kumar, and Mary Cheang. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine*, 34(6):1589–1596, June 2006.

[115] Leibovici, Shraga, Drucker, Konigsberger, Samra, Pitlik, and Pitlik. The benefit of appropriate empirical antibiotic treatment in patients with bloodstream infection. *Journal of Internal Medicine*, 244(5):379–386, 1998.

[116] Kerri A. Thom, Marin L. Schweizer, Regina B. Osih, Jessina C. McGregor, Jon P. Furuno, Eli N. Perencevich, and Anthony D. Harris. Impact of empiric antimicrobial therapy on outcomes in patients with *Escherichia coli* and *Klebsiella pneumoniae* bacteremia: A cohort study. *BMC Infectious Diseases*, 8(1):116, September 2008.

[117] Gokhan Metan, Pinar Zarakolu, Banu Cakir, Gulsen Hascelik, and Omrum Uzun. Clinical outcomes and therapeutic options of bloodstream infections caused by extended-spectrum beta-lactamase-producing *Escherichia coli*. *International Journal of Antimicrobial Agents*, 26(3):254–257, September 2005.

[118] Government of Canada (Health Canada). Summary safety review - fluoroquinolones - assessing the potential risk of persistent and disabling side effects. https://www.canada.ca/en/health-canada/services/drugs-health-products/medeffect-canada/safety-reviews/summary-safety-review-fluoroquinolones-assessing-potential-risk-persistent-disabling-effects.html, January 2017.

[119] Office of the Commissioner. FDA updates warnings for fluoroquinolone antibiotics on risks of mental health and low blood sugar adverse reactions. https://www.fda.gov/news-events/press-announcements/fda-updates-warnings-fluoroquinolone-antibiotics-risks-mental-health-and-low-blood-sugar-adverse, July 2018.

[120] Arsenio Spinillo, Ezio Capuzzo, Salvatore Acciano, Antonella De Santolo, and Francesca Zara. Effect of antibiotic use on the prevalence of symptomatic vulvovaginal candidiasis. *American Journal of Obstetrics and Gynecology*, 180(1):14–17, January 1999.

[121] Errol R. Norwitz and James A. Greenberg. Antibiotics in pregnancy: Are they safe? *Reviews in Obstetrics and Gynecology*, 2(3):135–136, 2009.

[122] Joanna Matuszkiewicz-Rowińska, Jolanta Małyszko, and Monika Wieliczko. Urinary tract infections in pregnancy: Old and new unresolved diagnostic and therapeutic problems. *Archives of Medical Science*, 11(1):67–77, March 2015.

[123] Georgia Vrioni, Constantinos Tsiamis, George Oikonomidis, Kalliopi Theodoridou, Violeta Kapsimali, and Athanasios Tsakris. MALDI-TOF mass spectrometry technology for detecting biomarkers of antimicrobial resistance: Current achievements and future perspectives. *Annals of Translational Medicine*, 6(12):240, June 2018.

[124] Irith Wiegand, Kai Hilpert, and Robert E. W. Hancock. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nature Protocols*, 3(2):163–175, February 2008.

[125] Julie G. Burel, Mikhail Pomaznoy, Cecilia S. Lindestam Arlehamn, Gregory Seumois, Pandurangan Vijayanand, Alessandro Sette, and Bjoern Peters. The challenge of distinguishing cell-cell complexes from singlet cells in non-imaging flow cytometry and single-cell sorting. *bioRxiv*, 97(11):1127–1135, November 2020.

[126] Andrew E. Clark, Erin J. Kaleta, Amit Arora, and Donna M. Wolk. Matrix-assisted laser desorption ionization–time of flight mass spectrometry: A fundamental shift in the routine practice of clinical microbiology. *Clinical Microbiology Reviews*, 26(3):547–603, July 2013.

[127] John J. Bright, Martin A. Claydon, Majeed Soufian, and Derek B. Gordon. Rapid typing of bacteria using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry and pattern recognition software. *Journal of Microbiological Methods*, 48(2):127–138, February 2002.

[128] Katrin Sparbier, Sören Schubert, Ulrich Weller, Christiane Boogen, and Markus Kostrzewa. Matrix-assisted laser desorption ionization–time of flight mass spectrometry-based functional assay for rapid detection of resistance against β-lactam antibiotics. *Journal of Clinical Microbiology*, 50(3):927–937, March 2012.

[129] Irene Burckhardt and Stefan Zimmermann. Using matrix-assisted laser desorption ionization-time of flight mass spectrometry to detect carbapenem resistance within 1 to 2.5 hours. *Journal of Clinical Microbiology*, 49(9):3321–3324, September 2011.

[130] Markus Kostrzewa, Katrin Sparbier, Thomas Maier, and Sören Schubert. MALDI-TOF MS: An upcoming tool for rapid detection of antibiotic resistance in microorganisms. *Proteomics Clinical Applications*, 7(11-12):767–778, October 2013.

[131] Caroline Weis, Aline Cuénod, Bastian Rieck, Felipe Llinares-López, Olivier Dubuis, Susanne Graf, Claudia Lang, Michael Oberle, Kirstine K. Soegaard, Michael Osthoff, Karsten Borgwardt, and Adrian Egli. Direct antimicrobial resistance prediction from MALDI-TOF mass spectra profile in clinical isolates through machine learning. *Nature Medicine*, 28:164–174, January 2022.

[132] Jaroslav Hrabák, Radka Walková, Vendula Študentová, Eva Chudáčková, and Tamara Bergerová. Carbapenemase activity detection by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Journal of Clinical Microbiology*, 49(9):3222–3227, September 2011.

[133] Georgia D. Kaprou, Ieva Bergšpica, Elena A. Alexa, Avelino Alvarez-Ordóñez, and Miguel Prieto. Rapid methods for antimicrobial resistance diagnostics. *Antibiotics*, 10(2):209, February 2021.

[134] Dieter Naumann, Dieter Helm, and Harald Labischinski. Microbiological characterizations by FT-IR spectroscopy. *Nature*, 351(6321):81–82, May 1991.

[135] Carla Rodrigues, Clara Sousa, João A. Lopes, Ângela Novais, and Luísa Peixe. A front line on *Klebsiella pneumoniae* capsular polysaccharide knowledge: Fourier

transform infrared spectroscopy as an accurate and fast typing tool. *mSystems*, 5(2), April 2020.

[136] Dieter Helm, Harald Labischinski, and Dieter Naumann. Elaboration of a procedure for identification of bacteria using Fourier-transform IR spectral libraries: A stepwise correlation approach. *Journal of Microbiological Methods*, 14(2):127–142, November 1991.

[137] Nicole R. Büchl, Mareike Wenning, Herbert Seiler, Henriette Mietke-Hofmann, and Siegfried Scherer. Reliable identification of closely related *Issatchenkia* and *Pichia* species using artificial neural network analysis of Fourier-transform infrared spectra. *Yeast (Chichester, England)*, 25(11):787–798, November 2008.

[138] Mareike Wenning and Siegfried Scherer. Identification of microorganisms by FTIR spectroscopy: Perspectives and limitations of the method. *Applied Microbiology and Biotechnology*, 97(16):7111–7120, August 2013.

[139] Ângela Novais, Ana R. Freitas, Carla Rodrigues, and Luísa Peixe. Fourier transform infrared spectroscopy: Unlocking fundamentals and prospects for bacterial strain typing. *European Journal of Clinical Microbiology & Infectious Diseases*, 38(3):427–448, March 2019.

[140] Lucia L. Nemoy, Mamuka Kotetishvili, Justine Tigno, Ananda Keefer-Norris, Anthony D. Harris, Eli N. Perencevich, Judith A. Johnson, Dave Torpey, Alexander Sulakvelidze, J. Glenn Morris, and O. Colin Stine. Multilocus sequence typing versus pulsed-field gel electrophoresis for characterization of extended-spectrum beta-lactamase-producing *E. coli* isolates. *Journal of Clinical Microbiology*, 43(4):1776–1781, April 2005.

[141] Jillian Leigh Rumore, Lorelee Tschetter, and Celine Nadon. The impact of multilocus variable-number tandem-repeat analysis on PulseNet Canada *Escherichia coli* O157:H7 laboratory surveillance and outbreak support, 2008–2012. *Foodborne Pathogens and Disease*, 13(5):255–261, May 2016.

[142] Efrain M. Ribot and Kelley B. Hise. Future challenges for tracking foodborne diseases. *EMBO Reports*, 17(11):1499–1505, November 2016.

[143] Efrain M. Ribot, Molly Freeman, Kelley B. Hise, and Peter Gerner-Smidt. PulseNet: Entering the age of next-generation sequencing. *Foodborne Pathogens and Disease*, 16(7):451–456, July 2019.

[144] Jennifer Dien Bard and Francesca Lee. Why can't we just use PCR? the role of genotypic versus phenotypic testing for antimicrobial resistance testing. *Clinical Microbiology Newsletter*, 40(11):87–95, June 2018.

[145] Robert J. Clifford, Michael Milillo, Jackson Prestwood, Reyes Quintero, Daniel V. Zurawski, Yoon I. Kwak, Paige E. Waterman, Emil P. Lesho, and Patrick Mc Gann. Detection of bacterial 16S rRNA and identification of four clinically important bacteria by real-time PCR. *PLOS ONE*, 7(11):e48558, November 2012.

[146] Scott A. Cunningham, Lynne M. Sloan, Lisa M. Nyre, Emily A. Vetter, Jayawant Mandrekar, and Robin Patel. Three-hour molecular detection of *Campylobacter, Salmonella, Yersinia*, and *Shigella* species in feces with accuracy as high as that of culture. *Journal of Clinical Microbiology*, 48(8):2929–2933, August 2010.

[147] Claudia Toma, Yan Lu, Naomi Higa, Noboru Nakasone, Isabel Chinen, Ariela Baschkier, Marta Rivas, and Masaaki Iwanaga. Multiplex PCR assay for identification of human diarrheagenic *E. coli*. *Journal of Clinical Microbiology*, 41(6):2669–2671, June 2003.

[148] N. G. Tornieporth, J. John, K. Salgado, P. de Jesus, E. Latham, M. C. Melo, S. T. Gunzburg, and L. W. Riley. Differentiation of pathogenic *Escherichia coli* strains in Brazilian children by PCR. *Journal of Clinical Microbiology*, 33(5):1371–1374, May 1995.

[149] Thomas E. Grys, Lynne M. Sloan, Jon E. Rosenblatt, and Robin Patel. Rapid and sensitive detection of Shiga toxin-producing *Escherichia coli* from nonenriched stool specimens by real-time PCR in comparison to enzyme immunoassay and culture. *Journal of Clinical Microbiology*, 47(7):2008–2012, July 2009.

[150] William C. Rice. Design and evaluation of PCR primers which differentiate *Escherichia coli* O157:H7 and related serotypes. *Journal of Applied Microbiology*, 106(1):149–160, January 2009.

[151] P. Feng and S. R. Monday. Multiplex PCR for detection of trait and virulence factors in enterohemorrhagic *Escherichia coli* serotypes. *Molecular and Cellular Probes*, 14(6):333–337, November 2000.

[152] Samuel Yang and Richard E. Rothman. PCR-based diagnostics for infectious diseases: Uses, limitations, and future applications in acute-care settings. *The Lancet. Infectious Diseases*, 4(6):337–348, June 2004.

[153] Elfath M. Elnifro, Ahmed M. Ashshi, Robert J. Cooper, and Paul E. Klapper. Multiplex PCR: Optimization and application in diagnostic virology. *Clinical Microbiology Reviews*, 13(4):559–570, October 2000.

[154] D. Boxrud, K. Pederson-Gulrud, J. Wotton, C. Medus, E. Lyszkowicz, J. Besser, and J. M. Bartkus. Comparison of multiple-locus variable-number tandem repeat analysis, pulsed-field gel electrophoresis, and phage typing for subtype analysis of *Salmonella enterica* serotype Enteritidis. *Journal of Clinical Microbiology*, 45(2):536–543, February 2007.

[155] Jillian Rumore, Lorelee Tschetter, Ashley Kearney, Rima Kandar, Rachel McCormick, Matthew Walker, Christy-Lynn Peterson, Aleisha Reimer, and Celine Nadon. Evaluation of whole-genome sequencing for outbreak detection of verotoxigenic *Escherichia coli* O157:H7 from the Canadian perspective. *BMC Genomics*, 19(1):870, December 2018.

[156] Thierry Wirth, Daniel Falush, Ruiting Lan, Frances Colles, Patience Mensa, Lothar H. Wieler, Helge Karch, Peter R. Reeves, Martin C.J. Maiden, Howard Ochman, and Mark Achtman. Sex and virulence in *Escherichia coli*: An evolutionary perspective. *Molecular Microbiology*, 60(5):1136–1151, June 2006.

[157] Françoise Jaureguy, Luce Landraud, Virginie Passet, Laure Diancourt, Eric Frapy, Ghislaine Guigon, Etienne Carbonnelle, Olivier Lortholary, Olivier Clermont, Erick Denamur, Bertrand Picard, Xavier Nassif, and Sylvain Brisse. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics*, 9(1):560, November 2008.

[158] Weihong Qi, David Lacher, Alyssa Bumbaugh, Katie Hyma, Lindsey Ouellette, Teresa Bergholz, Cheryl Tarr, and Thomas Whittam. EcMLST: An online database for multi locus sequence typing of pathogenic *Escherichia coli*. In *IEEE Computational Systems Bioinformatics Conference*, pages 520–521, Stanford, CA, USA, August 2004.

[159] Marina R. Pulido, Meritxell García-Quintanilla, Reyes Martín-Peña, José Miguel Cisneros, and Michael J. McConnell. Progress on the development of rapid methods for antimicrobial susceptibility testing. *Journal of Antimicrobial Chemotherapy*, 68(12):2710–2717, December 2013.

[160] J. C. Kwong, N. Mccallum, V. Sintchenko, and B. P. Howden. Whole genome sequencing in clinical and public health microbiology. *Pathology*, 47(3):199–210, April 2015.

[161] Government of Canada (Public Health Agency of Canada). FoodNet Canada annual report 2018. https://www.canada.ca/en/public-health/services/surveillance/foodnet-canada/publications/foodnet-canada-annual-report-2018.html, December 2019.

[162] Government of Canada (Public Health Agency of Canada). Canadian Integrated Program for Antimicrobial Resistance Surveillance (CIPARS). https://www.canada.ca/en/public-health/services/surveillance/canadian-integrated-program-antimicrobial-resistance-surveillance-cipars.html, September 2023.

[163] Gregory H. Tyson, Patrick F. McDermott, Cong Li, Yuansha Chen, Daniel A. Tadesse, Sampa Mukherjee, Sonya Bodeis-Jones, Claudine Kabera, Stuart A. Gaines, Guy H. Loneragan, Tom S. Edrington, Mary Torrence, Dayna M. Harhay, and Shaohua Zhao. WGS accurately predicts antimicrobial resistance in *Escherichia coli*. *Journal of Antimicrobial Chemotherapy*, 70(10):2763–2769, October 2015.

[164] M. J. Ellington, O. Ekelund, F. M. Aarestrup, R. Canton, M. Doumith, C. Giske, H. Grundman, H. Hasman, M. T. G. Holden, K. L. Hopkins, J. Iredell, G. Kahlmeter, C. U. Köser, A. MacGowan, D. Mevius, M. Mulvey, T. Naas, T. Peto, J. M. Rolain, Ø. Samuelsen, and N. Woodford. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: Report from the EUCAST subcommittee. *Clinical Microbiology and Infection*, 23(1):2–22, January 2017.

[165] Illumina. Sequencing platforms — Illumina NGS platforms. https://www.illumina.com/systems/sequencing-platforms.html.

[166] Xavier Didelot, Rory Bowden, Daniel J. Wilson, Tim E. A. Peto, and Derrick W. Crook. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics*, 13(9):601–612, September 2012.

[167] Narjol González-Escalona, Marc A. Allard, Eric W. Brown, Shashi Sharma, and Maria Hoffmann. Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*. *PLOS ONE*, 14(7):e0220494, July 2019.

[168] Illumina. Sequencing read length — how to calculate NGS read length. https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html.

[169] Aaron M. Wenger, Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D. Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M. Phillippy, Michael C. Schatz, Gene Myers, Mark A. DePristo, Jue Ruan, Tobias Marschall, Fritz J. Sedlazeck, Justin M. Zook, Heng Li, Sergey Koren, Andrew Carroll, David R. Rank, and Michael W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162, October 2019.

[170] Ting Hon, Kristin Mars, Greg Young, Yu-Chih Tsai, Joseph W. Karalius, Jane M. Landolin, Nicholas Maurer, David Kudrna, Michael A. Hardigan, Cynthia C. Steiner, Steven J. Knapp, Doreen Ware, Beth Shapiro, Paul Peluso, and David R. Rank. Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, 7(1):399, November 2020.

[171] Alexander Payne, Nadine Holmes, Vardhman Rakyan, and Matthew Loose. BulkVis: A graphical viewer for Oxford Nanopore bulk FAST5 files. *Bioinformatics*, 35(13):2193–2198, 20197.

[172] Richard M. Leggett and Matthew D. Clark. A world of opportunities with Nanopore sequencing. *Journal of Experimental Botany*, 68(20):5419–5429, November 2017.

[173] Oxford Nanopore Technologies. The Nanopore sequencing workflow. http://nanoporetech.com/how-it-works/nanopore-sequencing-workflow.

[174] Jamie K. Lemon, Pavel P. Khil, Karen M. Frank, and John P. Dekker. Rapid Nanopore sequencing of plasmids and resistance gene detection in clinical isolates. *Journal of Clinical Microbiology*, 55(12):3530–3543, December 2017.

[175] Oxford Nanopore Technologies. Product comparison. http://nanoporetech.com/products/comparison.

[176] K. Schmidt, S. Mwaigwisya, L. C. Crossman, M. Doumith, D. Munroe, C. Pires, A. M. Khan, N. Woodford, N. J. Saunders, J. Wain, J. O'Grady, and D. M. Livermore. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by Nanopore-based metagenomic sequencing. *Journal of Antimicrobial Chemotherapy*, 72(1):104–114, January 2017.

[177] Pranita D. Tamma, Yunfan Fan, Yehudit Bergman, Geo Pertea, Abida Q. Kazmi, Shawna Lewis, Karen C. Carroll, Michael C. Schatz, Winston Timp, and Patricia J. Simner. Applying rapid whole-genome sequencing to predict phenotypic antimicrobial susceptibility testing results among carbapenem-resistant *Klebsiella pneumoniae* clinical isolates. *Antimicrobial Agents and Chemotherapy*, 63(1), January 2019.

[178] Serghei Mangul, Thiago Mosqueiro, Richard J. Abdill, Dat Duong, Keith Mitchell, Varuni Sarwal, Brian Hill, Jaqueline Brito, Russell Jared Littman, Benjamin Statz, Angela Ka-Mei Lam, Gargi Dayama, Laura Grieneisen, Lana S. Martin, Jonathan Flint, Eleazar Eskin, and Ran Blekhman. Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLOS Biology*, 17(6):e3000333, June 2019.

[179] Baofeng Jia, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, Biren M. Dave, Sheldon Pereira, Arjun N. Sharma, Sachin Doshi, Mélanie Courtot, Raymond Lo, Laura E. Williams, Jonathan G. Frye, Tariq Elsayegh, Daim Sardar, Erin L. Westman, Andrew C. Pawlowski, Timothy A. Johnson, Fiona S. L. Brinkman, Gerard D. Wright, and Andrew G. McArthur. CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1):D566–D573, January 2017.

[180] Michael Feldgarden, Vyacheslav Brover, Daniel H. Haft, Arjun B. Prasad, Douglas J. Slotta, Igor Tolstoy, Gregory H. Tyson, Shaohua Zhao, Chih-Hao Hsu, Patrick F. McDermott, Daniel A. Tadesse, Cesar Morales, Mustafa Simmons, Glenn Tillman, Jamie Wasilenko, Jason P. Folster, and William Klimke. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrobial Agents and Chemotherapy*, 63(11), November 2019.

[181] Valeria Bortolaia, Rolf S. Kaas, Etienne Ruppe, Marilyn C. Roberts, Stefan Schwarz, Vincent Cattoir, Alain Philippon, Rosa L. Allesoe, Ana Rita Rebelo, Alfred Ferrer Florensa, Linda Fagelhauer, Trinad Chakraborty, Bernd Neumann, Guido Werner, Jennifer K. Bender, Kerstin Stingl, Minh Nguyen, Jasmine Coppens, Basil Britto Xavier, Surbhi Malhotra-Kumar, Henrik Westh, Mette Pinholt, Muna F. Anjum, Nicholas A. Duggett, Isabelle Kempf, Suvi Nykäsenoja, Satu Olkkola, Kinga Wieczorek, Ana Amaro, Lurdes Clemente, Joël Mossong, Serge Losch, Catherine Ragimbeau, Ole Lund, and Frank M. Aarestrup. ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy*, 75(12):3491–3500, December 2020.

[182] Torsten Seemann. Abricate. https://github.com/tseemann/abricate, November 2019.

[183] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[184] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: Architecture and applications. *BMC Bioinformatics*, 10:421, December 2009.

[185] N. Stoesser, E. M. Batty, D. W. Eyre, M. Morgan, D. H. Wyllie, C. Del Ojo Elias, J. R. Johnson, A. S. Walker, T. E. A. Peto, and D. W. Crook. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. *Journal of Antimicrobial Chemotherapy*, 68(10):2234–2244, October 2013.

[186] NCBI. Bacterial antimicrobial resistance reference gene database, bioproject prjna313047. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047, February 2016.

[187] Michael Feldgarden, Vyacheslav Brover, Daniel H. Haft, Arjun B. Prasad, Douglas J. Slotta, Igor Tolstoy, Gregory H. Tyson, Shaohua Zhao, Chih-Hao Hsu, Patrick F. McDermott, Daniel A. Tadesse, Cesar Morales, Mustafa Simmons, Glenn Tillman, Jamie Wasilenko, Jason P. Folster, and William Klimke. Using the NCBI AMRFinder tool to determine antimicrobial resistance genotype-phenotype correlations within a collection of NARMS isolates. *Antimicrobial Agents and Chemotherapy*, 63(11), February 2019.

[188] Institut Pasteur. Klebsiella sequence typing home page. https://bigsdb.pasteur.fr/klebsiella.

[189] Quan Ke Thai, Fabian Bös, and Jürgen Pleiss. The Lactamase Engineering Database: A critical survey of TEM sequences in public databases. *BMC Genomics*, 10(1):390, August 2009.

[190] Quan K. Thai and Juergen Pleiss. SHV Lactamase Engineering Database: A reconciliation tool for SHV β-lactamases in public databases. *BMC Genomics*, 11(1):563, October 2010.

[191] Abhishikha Srivastava, Neelja Singhal, Manisha Goel, Jugsharan Singh Virdi, and Manish Kumar. CBMAR: A comprehensive β-lactamase molecular annotation resource. *Database*, 2014, January 2014.

[192] Sushim Kumar Gupta, Babu Roshan Padmanabhan, Seydina M. Diene, Rafael Lopez-Rojas, Marie Kempf, Luce Landraud, and Jean-Marc Rolain. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial Agents and Chemotherapy*, 58(1):212–220, 2014.

[193] Enrique Doster, Steven M. Lakin, Christopher J. Dean, Cory Wolfe, Jared G. Young, Christina Boucher, Keith E. Belk, Noelle R. Noyes, and Paul S. Morley. MEGARes 2.0: A database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Research*, 48(D1):D561–D569, January 2020.

[194] Angela J. Taylor, Victoria Lappi, William J. Wolfgang, Pascal Lapierre, Michael J. Palumbo, Carlota Medus, and David Boxrud. Characterization of foodborne outbreaks of *Salmonella enterica* serovar Enteritidis with whole-genome sequencing single nucleotide polymorphism-based analysis for surveillance and outbreak detection. *Journal of Clinical Microbiology*, 53(10):3334–3340, October 2015.

[195] Phelim Bradley, N. Claire Gordon, Timothy M. Walker, Laura Dunn, Simon Heys, Bill Huang, Sarah Earle, Louise J. Pankhurst, Luke Anson, Mariateresa de Cesare, Paolo Piazza, Antonina A. Votintseva, Tanya Golubchik, Daniel J. Wilson, David H. Wyllie, Roland Diel, Stefan Niemann, Silke Feuerriegel, Thomas A. Kohl, Nazir Ismail, Shaheed V. Omar, E. Grace Smith, David Buck, Gil McVean, A. Sarah Walker, Tim E. A. Peto, Derrick W. Crook, and Zamin Iqbal. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nature Communications*, 6(1):10063, December 2015.

[196] Andreas Steiner, David Stucki, Mireia Coscolla, Sonia Borrell, and Sebastien Gagneux. KvarQ: Targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics*, 15(1):881, October 2014.

[197] Heather A. Carleton and Peter Gerner-Smidt. Whole-genome sequencing is taking over foodborne disease surveillance: Public health microbiology is undergoing its biggest change in a generation, replacing traditional methods with whole-genome sequencing. *Microbe Magazine*, 11(7):311–317, July 2016.

[198] Dan Stowell, Michael D. Wood, Hanna Pamuła, Yannis Stylianou, and Hervé Glotin. Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3):368–380, March 2019.

[199] Raúl Ramos-Pollán, Miguel Angel Guevara-López, Cesar Suárez-Ortega, Guillermo Díaz-Herrero, Jose Miguel Franco-Valiente, Manuel Rubio-del-Solar, Naimy González-de-Posada, Mario Augusto Pires Vaz, Joana Loureiro, and Isabel Ramos. Discovering mammography-based machine learning classifiers for breast cancer diagnosis. *Journal of Medical Systems*, 36(4):2259–2269, August 2012.

[200] Diogo Manuel Carvalho Leite, Xavier Brochet, Grégory Resch, Yok-Ai Que, Aitana Neves, and Carlos Peña-Reyes. Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics*, 19(14):420, November 2018.

[201] Daniel Fisch, Artur Yakimovich, Barbara Clough, Joseph Wright, Monique Bunyan, Michael Howell, Jason Mercer, and Eva Frickel. Defining host-pathogen interactions employing an artificial intelligence workflow. *eLife*, 8, February 2019.

[202] Yan A. Ivanenkov, Alex Zhavoronkov, Renat S. Yamidanov, Ilya A. Osterman, Petr V. Sergiev, Vladimir A. Aladinskiy, Anastasia V. Aladinskaya, Victor A. Terentiev, Mark S. Veselov, Andrey A. Ayginin, Victor G. Kartsev, Dmitry A. Skvortsov, Alexey V. Chemeris, Alexey Kh Baimiev, Alina A. Sofronova, Alexander S. Malyshev, Gleb I. Filkov, Dmitry S. Bezrukov, Bogdan A. Zagribelnyy, Evgeny O. Putin, Maria M. Puchinina, and Olga A. Dontsova. Identification of novel antibacterials using machine learning techniques. *Frontiers in Pharmacology*, 10, August 2019.

[203] Nicole E. Wheeler, Paul P. Gardner, and Lars Barquist. Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. *PLOS Genetics*, 14(5):e1007333, May 2018.

[204] Eric A. Dubinsky, Steven R. Butkus, and Gary L. Andersen. Microbial source tracking in impaired watersheds using PhyloChip and machine-learning classification. *Water Research*, 105:56–64, November 2016.

[205] Nadejda Lupolova, Tim J. Dallman, Nicola J. Holden, and David L. Gally. Patchy promiscuity: Machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. *Microbial Genomics*, 3(10), October 2017.

[206] Pedro Manuel Martínez-García, Emilia López-Solanilla, Cayo Ramos, and Pablo Rodríguez-Palenzuela. Prediction of bacterial associations with plants using a supervised machine-learning approach. *Environmental Microbiology*, 18(12):4847–4861, May 2016.

[207] Nadejda Lupolova, Timothy J. Dallman, Louise Matthews, James L. Bono, and David L. Gally. Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proceedings of the National Academy of Sciences*, 113(40):11312–11317, October 2016.

[208] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition, 2009.

[209] Nadejda Lupolova, Samantha J. Lycett, and David L. Gally. A guide to machine learning for bacterial host attribution using genome sequence data. *Microbial Genomics*, 5(12), November 2019.

[210] Zhen Yang, Feng Xu, Hongdou Li, and Yungang He. Beyond samples: A metric revealing more connections of gut microbiota between individuals. *Computational and Structural Biotechnology Journal*, 19:3930–3937, January 2021.

[211] Cornell Lab of Ornithology. Merlin - eBird. https://ebird.org/species/merlin.

[212] Brian L. Sullivan, Christopher L. Wood, Marshall J. Iliff, Rick E. Bonney, Daniel Fink, and Steve Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, October 2009.

[213] Chaodong Zhang, Yingjiao Ju, Na Tang, Yun Li, Gang Zhang, Yuqin Song, Hailing Fang, Liang Yang, and Jie Feng. Systematic analysis of supervised machine learning as an effective approach to predicate β-lactam resistance phenotype in *Streptococcus pneumoniae*. *Briefings in Bioinformatics*, 21(4):1347–1355, July 2020.

[214] Yang Yang, Katherine E. Niehaus, Timothy M. Walker, Zamin Iqbal, A. Sarah Walker, Daniel J. Wilson, Tim E. A. Peto, Derrick W. Crook, E. Grace Smith, Tingting Zhu, and David A. Clifton. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics*, 34(10):1666–1671, May 2018.

[215] Saskia Neuert, Satheesh Nair, Martin R. Day, Michel Doumith, Philip M. Ashton, Kate C. Mellor, Claire Jenkins, Katie L. Hopkins, Neil Woodford, Elizabeth de Pinna, Gauri Godbole, and Timothy J. Dallman. Prediction of phenotypic antimicrobial resistance profiles from whole genome sequences of non-typhoidal *Salmonella enterica*. *Frontiers in Microbiology*, 9, March 2018.

[216] Patrick F. McDermott, Gregory H. Tyson, Claudine Kabera, Yuansha Chen, Cong Li, Jason P. Folster, Sherry L. Ayers, Claudia Lam, Heather P. Tate, and Shaohua Zhao. Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal *Salmonella*. *Antimicrobial Agents and Chemotherapy*, 60(9):5515–5520, September 2016.

[217] Danesh Moradigaravand, Martin Palm, Anne Farewell, Ville Mustonen, Jonas Warringer, and Leopold Parts. Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLOS Computational Biology*, 14(12):e1006258, December 2018.

[218] Hsuan-Lin Her and Yu-Wei Wu. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains. *Bioinformatics*, 34(13):i89–i95, July 2018.

[219] Jason C. Hyun, Erol S. Kavvas, Jonathan M. Monk, and Bernhard O. Palsson. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. *PLoS Computational Biology*, 16(3), March 2020.

[220] Bryan Naidenov, Alexander Lim, Karyn Willyerd, Nathanial J. Torres, William L. Johnson, Hong Jin Hwang, Peter Hoyt, John E. Gustafson, and Charles Chen. Pan-genomic and polymorphic driven prediction of antibiotic resistance in *Elizabethkingia*. *Frontiers in Microbiology*, 10, July 2019.

[221] Xuefei Li, Jingxia Lin, Yongfei Hu, and Jiajian Zhou. PARMAP: A pan-genome-based computational framework for predicting antimicrobial resistance. *Frontiers in Microbiology*, 11, October 2020.

[222] Erol S. Kavvas, Edward Catoiu, Nathan Mih, James T. Yurkovich, Yara Seif, Nicholas Dillon, David Heckmann, Amitesh Anand, Laurence Yang, Victor Nizet,

Jonathan M. Monk, and Bernhard O. Palsson. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nature Communications*, 9(1):4306, October 2018.

[223] Mélodie Duval, Daniel Dar, Filipe Carvalho, Eduardo P. C. Rocha, Rotem Sorek, and Pascale Cossart. HflXr, a homolog of a ribosome-splitting factor, mediates antibiotic resistance. *Proceedings of the National Academy of Sciences of the United States of America*, 115(52):13359–13364, December 2018.

[224] Pierre Mahé and Maud Tournoud. Predicting bacterial resistance from whole-genome sequences using k-mers and stability selection. *BMC Bioinformatics*, 19(1):383, October 2018.

[225] Shuai Zhi, Qiaozhi Li, Yutaka Yasui, Graham Banting, Thomas A. Edge, Edward Topp, Tim A. McAllister, and Norman F. Neumann. An evaluation of logic regression-based biomarker discovery across multiple intergenic regions for predicting host specificity in *Escherichia coli*. *Molecular Phylogenetics and Evolution*, 103:133–142, October 2016.

[226] Erki Aun, Age Brauer, Veljo Kisand, Tanel Tenson, and Maido Remm. A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Computational Biology*, 14(10), October 2018.

[227] Alexandre Drouin, Sébastien Giguère, Maxime Déraspe, Mario Marchand, Michael Tyers, Vivian G. Loo, Anne-Marie Bourgault, François Laviolette, and Jacques Corbeil. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, 17(1):754, September 2016.

[228] James Emmanuel San, Shakuntala Baichoo, Aquillah Kanzi, Yumna Moosa, Richard Lessells, Vagner Fonseca, John Mogaka, Robert Power, and Tulio de Oliveira. Current affairs of microbial genome-wide association studies: Approaches, bottlenecks and analytical pitfalls. *Frontiers in Microbiology*, 10, 2020.

[229] J. Jeukens, I. Kukavica-Ibrulj, J. G. Emond-Rheault, L. Freschi, and R. C. Levesque. Comparative genomics of a drug-resistant *Pseudomonas aeruginosa* panel and the challenges of antimicrobial resistance prediction from genomes. *FEMS Microbiology Letters*, 364(18):fnx161, September 2017.

[230] K. T. Mulroney, J. M. Hall, X. Huang, E. Turnbull, N. M. Bzdyl, A. Chakera, U. Naseer, E. M. Corea, M. J. Ellington, K. L. Hopkins, A. L. Wester, O. Ekelund, N. Woodford, and T. J. J. Inglis. Rapid susceptibility profiling of carbapenem-resistant *Klebsiella pneumoniae*. *Scientific Reports*, 7(1):1903, May 2017.

[231] Timothy J. J. Inglis, Teagan F. Paton, Malgorzata K. Kopczyk, Kieran T. Mulroney, and Christine F. Carson. Same-day antimicrobial susceptibility test using acoustic-enhanced flow cytometry visualized with supervised machine learning. *Journal of Medical Microbiology*, 69(5):657–669, May 2020.

[232] Cheryl A. Mather, Brian J. Werth, Shobini Sivagnanam, Dhruba J. SenGupta, and Susan M. Butler-Wu. Rapid detection of vancomycin-intermediate *Staphylococcus aureus* by matrix-assisted laser desorption ionization–time of flight mass spectrometry. *Journal of Clinical Microbiology*, 54(4):883–890, April 2016.

[233] Kazuyuki Sogawa, Masaharu Watanabe, Takayuki Ishige, Syunsuke Segawa, Akiko Miyabe, Syota Murata, Tomoko Saito, Akihiro Sanda, Katsunori Furuhata, and Fumio Nomura. Rapid discrimination between methicillin-sensitive and methicillin-resistant *Staphylococcus aureus* using MALDI-TOF mass spectrometry. *Biocontrol Science*, 22(3):163–169, 2017.

[234] Lukasz Lechowicz, Mariusz Urbaniak, Wioletta Adamus-Białek, and Wiesław Kaca. The use of infrared spectroscopy and artificial neural networks for detection of uropathogenic *Escherichia coli* strains' susceptibility to cephalothin. *Acta Biochimica Polonica*, 60(4):713–718, 2013.

[235] Ahmad Salman, Uraib Sharaha, Eladio Rodriguez-Diaz, Elad Shufan, Klaris Riesenberg, Irving J. Bigio, and Mahmoud Huleihel. Detection of antibiotic resistant *Escherichia coli* bacteria using infrared microscopy and advanced multivariate analysis. *Analyst*, 142(12):2136–2144, June 2017.

[236] Uraib Sharaha, Eladio Rodriguez-Diaz, Klaris Riesenberg, Irving J. Bigio, Mahmoud Huleihel, and Ahmad Salman. Using infrared spectroscopy and multivariate analysis to detect antibiotics' resistant *Escherichia coli* bacteria. *Analytical Chemistry*, 89(17):8782–8790, September 2017.

[237] Uraib Sharaha, Eladio Rodriguez-Diaz, Orli Sagi, Klaris Riesenberg, Ahmad Salman, Irving J. Bigio, and Mahmoud Huleihel. Fast and reliable determination of *Escherichia coli* susceptibility to antibiotics: Infrared microscopy in tandem with machine learning algorithms. *Journal of Biophotonics*, 12(7):e201800478, March 2019.

[238] Uraib Sharaha, Eladio Rodriguez-Diaz, Orli Sagi, Klaris Riesenberg, Itshak Lapidot, Yoram Segal, Irving J. Bigio, Mahmoud Huleihel, and Ahmad Salman. Detection of extended-spectrum β-lactamase-producing *Escherichia coli* using infrared microscopy and machine-learning algorithms. *Analytical Chemistry*, 91(3):2525–2530, February 2019.

[239] Manal Suleiman, George Abu-Aqil, Uraib Sharaha, Klaris Riesenberg, Orli Sagi, Itshak Lapidot, Mahmoud Huleihel, and Ahmad Salman. Rapid detection of *Klebsiella pneumoniae* producing extended spectrum β lactamase enzymes by infrared microspectroscopy and machine learning algorithms. *Analyst*, 146(4):1421–1429, February 2021.

[240] J. Michael Janda and Sharon L. Abbott. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, 45(9):2761–2764, September 2007.

[241] Francesco Durazzi, Claudia Sala, Gastone Castellani, Gerardo Manfreda, Daniel Remondini, and Alessandra De Cesare. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Scientific Reports*, 11(1):3030, February 2021.

[242] John Vollmers, Sandra Wiegand, and Anne-Kristin Kaster. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLOS ONE*, 12(1):e0169662, January 2017.

[243] Martin Ayling, Matthew D. Clark, and Richard M. Leggett. New approaches for metagenome assembly with short reads. *Briefings in Bioinformatics*, 21(2):584–594, February 2019.

[244] Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, 13(6):e1005595, June 2017.

[245] Adriel Latorre-Pérez, Pascual Villalba-Bermell, Javier Pascual, and Cristina Vilanova. Assembly methods for Nanopore-based metagenomic sequencing: A comparative study. *Scientific Reports*, 10(1):13588, August 2020.

[246] Alexander L. Greninger, Samia N. Naccache, Scot Federman, Guixia Yu, Placide Mbala, Vanessa Bres, Doug Stryke, Jerome Bouquet, Sneha Somasekar, Jeffrey M. Linnen, Roger Dodd, Prime Mulembakani, Bradley S. Schneider, Jean-Jacques Muyembe-Tamfum, Susan L. Stramer, and Charles Y. Chiu. Rapid metagenomic identification of viral pathogens in clinical samples by real-time Nanopore sequencing analysis. *Genome Medicine*, 7(1):99, September 2015.

[247] Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, Andreas Bremges, Adrian Fritz, Ruben Garrido-Oter, Tue Sparholt Jørgensen, Nicole Shapiro, Philip D. Blood, Alexey Gurevich, Yang Bai, Dmitrij Turaev, Matthew Z. DeMaere, Rayan Chikhi, Niranjan Nagarajan, Christopher Quince, Fernando Meyer, Monika Balvočiūtė, Lars Hestbjerg Hansen, Søren J. Sørensen, Burton K. H. Chia, Bertrand Denis, Jeff L. Froula, Zhong Wang, Robert Egan, Dongwan Don Kang, Jeffrey J. Cook, Charles Deltel, Michael Beckstette, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Dominique Lavenier, Yu-Wei Wu, Steven W. Singer, Chirag Jain, Marc Strous, Heiner Klingenberg, Peter Meinicke, Michael D. Barton, Thomas Lingner, Hsin-Hung Lin, Yu-Chieh Liao, Genivaldo Gueiros Z. Silva, Daniel A. Cuevas, Robert A. Edwards, Surya Saha, Vitor C. Piro, Bernhard Y. Renard, Mihai Pop, Hans-Peter Klenk, Markus Göker, Nikos C. Kyrpides, Tanja Woyke, Julia A. Vorholt, Paul Schulze-Lefert, Edward M. Rubin, Aaron E. Darling, Thomas Rattei, and Alice C. McHardy. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature Methods*, 14(11):1063–1071, November 2017.

[248] Fernando Meyer, Till-Robin Lesker, David Koslicki, Adrian Fritz, Alexey Gurevich, Aaron E. Darling, Alexander Sczyrba, Andreas Bremges, and Alice C. McHardy.

Tutorial: Assessing metagenomics software with the CAMI benchmarking toolkit. *Nature Protocols*, 16:1785–1801, March 2021.

[249] Fernando Meyer, Adrian Fritz, Zhi-Luo Deng, David Koslicki, Till Robin Lesker, Alexey Gurevich, Gary Robertson, Mohammed Alser, Dmitry Antipov, Francesco Beghini, Denis Bertrand, Jaqueline J. Brito, C. Titus Brown, Jan Buchmann, Aydin Buluç, Bo Chen, Rayan Chikhi, Philip T. L. C. Clausen, Alexandru Cristian, Piotr Wojciech Dabrowski, Aaron E. Darling, Rob Egan, Eleazar Eskin, Evangelos Georganas, Eugene Goltsman, Melissa A. Gray, Lars Hestbjerg Hansen, Steven Hofmeyr, Pingqin Huang, Luiz Irber, Huijue Jia, Tue Sparholt Jørgensen, Silas D. Kieser, Terje Klemetsen, Axel Kola, Mikhail Kolmogorov, Anton Korobeynikov, Jason Kwan, Nathan LaPierre, Claire Lemaitre, Chenhao Li, Antoine Limasset, Fabio Malcher-Miranda, Serghei Mangul, Vanessa R. Marcelino, Camille Marchet, Pierre Marijon, Dmitry Meleshko, Daniel R. Mende, Alessio Milanese, Niranjan Nagarajan, Jakob Nissen, Sergey Nurk, Leonid Oliker, Lucas Paoli, Pierre Peterlongo, Vitor C. Piro, Jacob S. Porter, Simon Rasmussen, Evan R. Rees, Knut Reinert, Bernhard Renard, Espen Mikal Robertsen, Gail L. Rosen, Hans-Joachim Ruscheweyh, Varuni Sarwal, Nicola Segata, Enrico Seiler, Lizhen Shi, Fengzhu Sun, Shinichi Sunagawa, Søren Johannes Sørensen, Ashleigh Thomas, Chengxuan Tong, Mirko Trajkovski, Julien Tremblay, Gherman Uritskiy, Riccardo Vicedomini, Zhengyang Wang, Ziye Wang, Zhong Wang, Andrew Warren, Nils Peder Willassen, Katherine Yelick, Ronghui You, Georg Zeller, Zhengqiao Zhao, Shanfeng Zhu, Jie Zhu, Ruben Garrido-Oter, Petra Gastmeier, Stephane Hacquard, Susanne Häußler, Ariane Khaledi, Friederike Maechler, Fantin Mesny, Simona Radutoiu, Paul Schulze-Lefert, Nathiana Smit, Till Strowig, Andreas Bremges, Alexander Sczyrba, and Alice Carolyn McHardy. Critical assessment of metagenome interpretation: The second round of challenges. *Nature Methods*, 19(4):429–440, April 2022.

[250] Andreas Bremges and Alice C. McHardy. Critical assessment of metagenome interpretation enters the second round. *mSystems*, 3(4), August 2018.

[251] Alla Mikheenko, Vladislav Saveliev, and Alexey Gurevich. MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics*, 32(7):1088–1090, April 2016.

[252] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, April 2013.

[253] Heng Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018.

[254] Fernando Meyer, Peter Hofmann, Peter Belmann, Ruben Garrido-Oter, Adrian Fritz, Alexander Sczyrba, and Alice C McHardy. AMBER: Assessment of metagenome binners. *GigaScience*, 7(6):giy069, June 2018.

[255] Simon H. Ye, Katherine J. Siddle, Daniel J. Park, and Pardis C. Sabeti. Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178(4):779–794, August 2019.

[256] Beatriz Delgado, Magdalena Serrano, Carmen González, Alex Bach, and Oscar González-Recio. Long reads from Nanopore sequencing as a tool for animal microbiome studies. *bioRxiv*, page 2019.12.21.886028, December 2019.

[257] Daniel H. Huson, Benjamin Albrecht, Caner Bağcı, Irina Bessarab, Anna Górska, Dino Jolic, and Rohan B. H. Williams. MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct*, 13(1):6, April 2018.

[258] Sissel Juul, Fernando Izquierdo, Adam Hurst, Xiaoguang Dai, Amber Wright, Eugene Kulesha, Roger Pettett, and Daniel J. Turner. What's in my pot? real-time species identification on the MinION. *bioRxiv*, page 030742, November 2015.

[259] Oxford Nanopore Technologies. Nanopore sequencing data analysis. http://nanoporetech.com/nanopore-sequencing-data-analysis.

[260] Daehwan Kim, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12):1721–1729, December 2016.

[261] Samir V. Deshpande, Timothy M. Reed, Raymond F. Sullivan, Lee J. Kerkhof, Keith M. Beigel, and Mary M. Wade. Offline next generation metagenomics sequence analysis using MinION detection software (MINDS). *Genes*, 10(8):578, July 2019.

[262] Oxford Nanopore Technologies. Real-time detection of antibiotic-resistance genes using Oxford Nanopore Technologies' MinION. http://nanoporetech.com/resource-centre/real-time-detection-antibiotic-resistance-genes-using-oxford-nanopore-technologies, November 2016.

[263] Steven M. Lakin, Chris Dean, Noelle R. Noyes, Adam Dettenwanger, Anne Spencer Ross, Enrique Doster, Pablo Rovira, Zaid Abdo, Kenneth L. Jones, Jaime Ruiz, Keith E. Belk, Paul S. Morley, and Christina Boucher. MEGARes: An antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Research*, 45(D1):D574–D580, January 2017.

[264] Ying Yang, Xiaotao Jiang, Benli Chai, Liping Ma, Bing Li, Anni Zhang, James R. Cole, James M. Tiedje, and Tong Zhang. ARGs-OAP: Online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. *Bioinformatics*, 32(15):2346–2351, August 2016.

[265] Karel Břinda, Alanna Callendrello, Kevin C. Ma, Derek R. MacFadden, Themoula Charalampous, Robyn S. Lee, Lauren Cowley, Crista B. Wadsworth, Yonatan H. Grad, Gregory Kucherov, Justin O'Grady, Michael Baym, and William P. Hanage. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nature Microbiology*, 5(3):455–464, March 2020.

[266] Gustavo Arango-Argoty, Emily Garner, Amy Pruden, Lenwood S. Heath, Peter Vikesland, and Liqing Zhang. DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6(1):23, February 2018.

[267] Kortine Annina Kleinheinz, Katrine Grimstrup Joensen, and Mette Voldby Larsen. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage*, 4(1):e27943, January 2014.

[268] Rahat Zaheer, Steven M. Lakin, Rodrigo Ortega Polo, Shaun R. Cook, Francis J. Larney, Paul S. Morley, Calvin W. Booker, Sherry J. Hannon, Gary Van Domselaar, Ron R. Read, and Tim A. McAllister. Comparative diversity of microbiomes and resistomes in beef feedlots, downstream environments and urban sewage influent. *BMC Microbiology*, 19(1):197, August 2019.

[269] Government of Canada (Public Health Agency of Canada). Yearly food-borne illness estimates for canada. https://www.canada.ca/en/public-health/services/food-borne-illness-canada/yearly-food-borne-illness-estimates-canada.html, July 2016.

[270] J. A. Maslikowska, S. A. N. Walker, M. Elligsen, N. Mittmann, L. Palmay, N. Daneman, and A. Simor. Impact of infection with extended-spectrum β-lactamase-producing *Escherichia coli* or *Klebsiella* species on outcome and hospitalization costs. *Journal of Hospital Infection*, 92(1):33–41, January 2016.

[271] Jennifer M. Andrews. Determination of minimum inhibitory concentrations. *Journal of Antimicrobial Chemotherapy*, 48(suppl_1):5–16, July 2001.

[272] Lucie Collineau, Patrick Boerlin, Carolee A. Carson, Brennan Chapman, Aamir Fazil, Benjamin Hetman, Scott A. McEwen, E. Jane Parmley, Richard J. Reid-Smith, Eduardo N. Taboada, and Ben A. Smith. Integrating whole-genome sequencing data into quantitative risk assessment of foodborne antimicrobial resistance: A review of opportunities and challenges. *Frontiers in Microbiology*, 10, 2019.

[273] John M. Besser, Heather A. Carleton, Eija Trees, Steven G. Stroika, Kelley Hise, Matthew Wise, and Peter Gerner-Smidt. Interpretation of whole-genome sequencing for enteric disease surveillance and outbreak investigation. *Foodborne Pathogens and Disease*, 16(7):504–512, June 2019.

[274] Rene S. Hendriksen, Valeria Bortolaia, Heather Tate, Gregory H. Tyson, Frank M. Aarestrup, and Patrick F. McDermott. Using genomics to track global antimicrobial resistance. *Frontiers in Public Health*, 7, 2019.

[275] Guenter Muehlberger, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel, Basilis Gatos, Albert Greinoecker, Tobias Grüning, Guenter Hackl, Vili Haukkovaara, Gerhard Heyer, Lauri Hirvonen, Tobias Hodel, Matti Jokinen, Philip Kahle, Mario Kallio, Frederic Kaplan, Florian Kleber, Roger Labahn, Eva Maria

Lang, Sören Laube, Gundram Leifert, Georgios Louloudis, Rory McNicholl, Jean-Luc Meunier, Johannes Michael, Elena Mühlbauer, Nathanael Philipp, Ioannis Pratikakis, Joan Puigcerver Pérez, Hannelore Putz, George Retsinas, Verónica Romero, Robert Sablatnig, Joan Andreu Sánchez, Philip Schofield, Giorgos Sfikas, Christian Sieber, Nikolaos Stamatopoulos, Tobias Strauß, Tamara Terbul, Alejandro Héctor Toselli, Berthold Ulreich, Mauricio Villegas, Enrique Vidal, Johanna Walcher, Max Weidemann, Herbert Wurster, and Konstantinos Zagoris. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976, January 2019.

[276] Amanda Sharkey. Autonomous weapons systems, killer robots and human dignity. *Ethics and Information Technology*, 21(2):75–87, June 2019.

[277] Marcus Nguyen, S. Wesley Long, Patrick F. McDermott, Randall J. Olsen, Robert Olson, Rick L. Stevens, Gregory H. Tyson, Shaohua Zhao, and James J. Davis. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *Journal of Clinical Microbiology*, 57(2), February 2019.

[278] Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. Genbank. *Nucleic Acids Research*, 41(D1):D36–D42, November 2012.

[279] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, March 2011.

[280] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, August 2016.

[281] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, October 2011.

[282] François Chollet. Keras. https://github.com/keras-team/keras, 2015.

[283] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[284] Max Pumperla. Hyperas. https://github.com/maxpumperla/hyperas, May 2019.

[285] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, January 2014.

[286] Chakkaphan Runcharoen, Kathy E. Raven, Sandra Reuter, Teemu Kallonen, Suporn Paksanont, Jeeranan Thammachote, Suthatip Anun, Beth Blane, Julian Parkhill, Sharon J. Peacock, and Narisara Chantratita. Whole genome sequencing of ESBL-producing *Escherichia coli* isolated from patients, farm waste and canals in Thailand. *Genome Medicine*, 9, September 2017.

[287] Teemu Kallonen, Hayley J. Brodrick, Simon R. Harris, Jukka Corander, Nicholas M. Brown, Veronique Martin, Sharon J. Peacock, and Julian Parkhill. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Research*, 27(8):1437–1449, August 2017.

[288] Samuel K. Sheppard, David S. Guttman, and J. Ross Fitzgerald. Population genomics of bacterial host adaptation. *Nature Reviews Genetics*, 19(9):549–565, September 2018.

[289] Allison L. Hicks, Nicole Wheeler, Leonor Sánchez-Busó, Jennifer L. Rakeman, Simon R. Harris, and Yonatan H. Grad. Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. *PLOS Computational Biology*, 15(9):e1007349, September 2019.

[290] Shyam M. Saladi, Nauman Javed, Axel Müller, and William M. Clemons. A statistical model for improved membrane protein expression using sequence-derived features. *Journal of Biological Chemistry*, 293(13):4913–4927, March 2018.

[291] Jinyuan Yan, Maxime Deforet, Kerry E. Boyle, Rayees Rahman, Raymond Liang, Chinweike Okegbe, Lars E. P. Dietrich, Weigang Qiu, and Joao B. Xavier. Bow-tie signaling in c-di-GMP: Machine learning in a simple biochemical network. *PLOS Computational Biology*, 13(8):e1005677, August 2017.

[292] Alexandre Drouin, Gaël Letarte, Frédéric Raymond, Mario Marchand, Jacques Corbeil, and François Laviolette. Interpretable genotype-to-phenotype classifiers with performance guarantees. *Scientific Reports*, 9(1):4071, March 2019.

[293] Shangying Wang, Kai Fan, Nan Luo, Yangxiaolu Cao, Feilun Wu, Carolyn Zhang, Katherine A. Heller, and Lingchong You. Massive computational acceleration by using neural networks to emulate mechanism-based biological models. *Nature Communications*, 10(1):1–9, September 2019.

[294] Amy J. Mathers, Gisele Peirano, and Johann D. D. Pitout. Chapter four - *Escherichia coli* ST131: The quintessential example of an international multiresistant high-risk clone. In Sima Sariaslani and Geoffrey Michael Gadd, editors, *Advances in Applied Microbiology*, volume 90, pages 109–154. Academic Press, January 2015.

[295] Joshua T. Freeman, Joseph Rubin, Gary N. McAuliffe, Gisele Peirano, Sally A. Roberts, Dragana Drinković, and Johann D. D. Pitout. Differences in risk-factor profiles between patients with ESBL-producing *Escherichia coli* and *Klebsiella pneumoniae*: A multicentre case-case comparison study. *Antimicrobial Resistance and Infection Control*, 3:27, September 2014.

[296] Patrice Nordmann, Thierry Naas, and Laurent Poirel. Global spread of carbapenemase-producing Enterobacteriaceae. *Emerging Infectious Diseases*, 17(10):1791–1798, October 2011.

[297] Joshua Quick, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, Raymond Koundouno, Gytis Dudas, Amy Mikhail, Nobila Ouédraogo, Babak Afrough, Amadou Bah, Jonathan H. J. Baum, Beate Becker-Ziaja, Jan Peter Boettcher, Mar Cabeza-Cabrerizo, Álvaro Camino-Sánchez, Lisa L. Carter, Juliane Doerrbecker, Theresa Enkirch, Isabel García Dorival, Nicole Hetzelt, Julia Hinzmann, Tobias Holm, Liana Eleni Kafetzopoulou, Michel Koropogui, Abigael Kosgey, Eeva Kuisma, Christopher H. Logue, Antonio Mazzarelli, Sarah Meisel, Marc Mertens, Janine Michel, Didier Ngabo, Katja Nitzsche, Elisa Pallasch, Livia Victoria Patrono, Jasmine Portmann, Johanna Gabriella Repits, Natasha Y. Rickett, Andreas Sachse, Katrin Singethan, Inês Vitoriano, Rahel L. Yemanaberhan, Elsa G. Zekeng, Trina Racine, Alexander Bello, Amadou Alpha Sall, Ousmane Faye, Oumar Faye, N'Faly Magassouba, Cecelia V. Williams, Victoria Amburgey, Linda Winona, Emily Davis, Jon Gerlach, Frank Washington, Vanessa Monteil, Marine Jourdain, Marion Bererd, Alimou Camara, Hermann Somlare, Abdoulaye Camara, Marianne Gerard, Guillaume Bado, Bernard Baillet, Déborah Delaune, Koumpingnin Yacouba Nebie, Abdoulaye Diarra, Yacouba Savane, Raymond Bernard Pallawo, Giovanna Jaramillo Gutierrez, Natacha Milhano, Isabelle Roger, Christopher J. Williams, Facinet Yattara, Kuiama Lewandowski, James Taylor, Phillip Rachwal, Daniel J. Turner, Georgios Pollakis, Julian A. Hiscox, David A. Matthews, Matthew K. O' Shea, Andrew McD. Johnston, Duncan Wilson, Emma Hutley, Erasmus Smit, Antonino Di Caro, Roman Wölfel, Kilian Stoecker, Erna Fleischmann, Martin Gabriel, Simon A. Weller, Lamine Koivogui, Boubacar Diallo, Sakoba Keïta, Andrew Rambaut, Pierre Formenty, Stephan Günther, and Miles W. Carroll. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530(7589):228–232, February 2016.

[298] Felix Krueger. TrimGalore. https://github.com/FelixKrueger/TrimGalore, December 2019.

[299] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, May 2011.

[300] Simon Andrews. FastQC. https://github.com/s-andrews/FastQC, April 2020.

[301] Tanja Magoč and Steven L. Salzberg. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, November 2011.

[302] Sergey Nurk, Anton Bankevich, Dmitry Antipov, Alexey Gurevich, Anton Korobeynikov, Alla Lapidus, Andrey Prjibelsky, Alexey Pyshkin, Alexander Sirotkin, Yakov Sirotkin, Ramunas Stepanauskas, Jeffrey McLean, Roger Lasken, Scott R. Clingenpeel, Tanja Woyke, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. Assembling genomes and mini-metagenomes from highly chimeric reads. In Minghua Deng, Rui Jiang, Fengzhu Sun, and Xuegong Zhang, editors, *Research in Computational Molecular Biology*, volume 7821 of *Lecture Notes in Computer Science*, pages 158–170, Berlin, Heidelberg, 2013. Springer.

[303] Torsten Seemann. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, July 2014.

[304] Rene S. Hendriksen, Susanne Karlsmose Pedersen, Pimlapas Leekitcharoenphon, Burkhard Malorny, Maria Borowiak, Antonio Battisti, Alessia Franco, Patricia Alba, Virginia Carfora, Antonia Ricci, Eleonora Mastrorilli, Carmen Losasso, Alessandra Longo, Sara Petrin, Lisa Barco, Tomasz Wołkowicz, Rafał Gierczyński, Katarzyna Zacharczuk, Natalia Wolaniuk, Dariusz Wasyl, Magdalena Zajac, Kinga Wieczorek, Katarzyna Półtorak, Liljana Petrovska-Holmes, Rob Davies, Yue Tang, Kathie Grant, Anthony Underwood, Timothy Dallman, Anaïs Painset, Hassan Hartman, Ali Al-Shabib, and Lauren Cowley. Final report of ENGAGE - establishing next generation sequencing ability for genomic analysis in europe. *EFSA Supporting Publications*, 15(6), June 2018.

[305] Karen Bush and George A. Jacoby. Updated functional classification of β-lactamases. *Antimicrobial Agents and Chemotherapy*, 54(3):969–976, March 2010.

[306] George A. Jacoby. AmpC β-lactamases. *Clinical Microbiology Reviews*, 22(1):161–182, January 2009.

[307] S. Peter-Getzlaff, S. Polsfuss, M. Poledica, M. Hombach, J. Giger, E. C. Böttger, R. Zbinden, and G. V. Bloemberg. Detection of AmpC beta-lactamase in *Escherichia coli*: Comparison of three phenotypic confirmation assays and genetic analysis. *Journal of Clinical Microbiology*, 49(8):2924–2932, August 2011.

[308] M. Akova, Youjun Yang, and D. M. Livermore. Interactions of tazobactam and clavulanate with inducibly- and constitutively-expressed Class I β-lactamases. *Journal of Antimicrobial Chemotherapy*, 25(2):199–208, February 1990.

[309] G S Babini, F Danel, S D Munro, P A Micklesen, and D M Livermore. Unusual tazobactam-sensitive AmpC beta-lactamase from two *Escherichia coli* isolates. *Journal of Antimicrobial Chemotherapy*, 41(1):115–118, January 1998.

[310] Tania S. Darphorn, Keshia Bel, Belinda B. Koenders-van Sint Anneland, Stanley Brul, and Benno H. Ter Kuile. Antibiotic resistance plasmid composition and architecture in *Escherichia coli* isolates from meat. *Scientific Reports*, 11(1):2136, January 2021.

[311] C. Verdet, V. Gautier, E. Chachaty, E. Ronco, N. Hidri, D. Decré, and G. Arlet. Genetic context of plasmid-carried blaCMY-2-like genes in Enterobacteriaceae. *Antimicrobial Agents and Chemotherapy*, 53(9):4002–4006, September 2009.

[312] Tania S. Darphorn, Yuanqing Hu, Belinda B. Koenders-van Sintanneland, Stanley Brul, and Benno H. ter Kuile. Multiplication of *ampC* upon exposure to a beta-lactam antibiotic results in a transferable transposon in *Escherichia coli*. *International Journal of Molecular Sciences*, 22(17):9230, August 2021.

[313] Valérie Campanacci, Russell E. Bishop, Stéphanie Blangy, Mariella Tegoni, and Christian Cambillau. The membrane bound bacterial lipocalin Blc is a functional dimer with binding preference for lysophospholipids. *FEBS letters*, 580(20):4877–4883, September 2006.

[314] Valérie Campanacci, Didier Nurizzo, Silvia Spinelli, Christel Valencia, Mariella Tegoni, and Christian Cambillau. The crystal structure of the *Escherichia coli* lipocalin Blc suggests a possible role in phospholipid binding. *FEBS Letters*, 562(1-3):183–188, 2004.

[315] Sally R. Partridge, Stephen M. Kwong, Neville Firth, and Slade O. Jensen. Mobile genetic elements associated with antimicrobial resistance. *Clinical Microbiology Reviews*, 31(4):e00088–17, August 2018.

[316] Salome N. Seiffert, Markus Hilty, Andreas Kronenberg, Sara Droz, Vincent Perreten, and Andrea Endimiani. Extended-spectrum cephalosporin-resistant *Escherichia coli* in community, specialized outpatient clinic and hospital settings in Switzerland. *Journal of Antimicrobial Chemotherapy*, 68(10):2249–2254, October 2013.

[317] Suguru Yamasaki, Eiji Nikaido, Ryosuke Nakashima, Keisuke Sakurai, Daisuke Fujiwara, Ikuo Fujii, and Kunihiko Nishino. The crystal structure of multidrug-resistance regulator ramr with multiple drugs. *Nature Communications*, 4(1):2078, June 2013.

[318] Ivana Jamborova, Brian D. Johnston, Ivo Papousek, Katerina Kachlikova, Lenka Micenkova, Connie Clabots, Anna Skalova, Katerina Chudejova, Monika Dolejska, Ivan Literak, and James R. Johnson. Extensive genetic commonality among wildlife, wastewater, community, and nosocomial isolates of *Escherichia coli* sequence type 131 (H30R1 and H30Rx subclones) that carry blaCTX-M-27 or blaCTX-M-15. *Antimicrobial Agents and Chemotherapy*, 62(10):e00519–18, September 2018.

[319] Nicole Stoesser, Anna E. Sheppard, Louise Pankhurst, Nicola De Maio, Catrin E. Moore, Robert Sebra, Paul Turner, Luke W. Anson, Andrew Kasarskis, Elizabeth M. Batty, Veronica Kos, Daniel J. Wilson, Rattanaphone Phetsouvanh, David Wyllie, Evgeni Sokurenko, Amee R. Manges, Timothy J. Johnson, Lance B. Price, Timothy E. A. Peto, James R. Johnson, Xavier Didelot, A. Sarah Walker, and Derrick W. Crook. Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *mBio*, 7(2):e02162–15, March 2016.

[320] Laurent Poirel, Jean-Winoc Decousser, and Patrice Nordmann. Insertion sequence ISEcp1B is involved in expression and mobilization of a blaCTX-M β-lactamase gene. *Antimicrobial Agents and Chemotherapy*, 47(9):2938–2945, September 2003.

[321] Amy J. Mathers, Gisele Peirano, and Johann D. D. Pitout. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant Enterobacteriaceae. *Clinical Microbiology Reviews*, 28(3):565–591, July 2015.

[322] Lisa Praski Alzrigat, Douglas L Huseby, Gerrit Brandis, and Diarmaid Hughes. Fitness cost constrains the spectrum of marR mutations in ciprofloxacin-resistant *Escherichia coli*. *Journal of Antimicrobial Chemotherapy*, 72(11):3016–3024, November 2017.

[323] S P Cohen, H Hächler, and S B Levy. Genetic and functional analysis of the multiple antibiotic resistance (mar) locus in *Escherichia coli*. *Journal of Bacteriology*, 175(5):1484–1492, March 1993.

[324] Getahun E. Agga, John W. Schmidt, and Terrance M. Arthur. Antimicrobial-resistant fecal bacteria from ceftiofur-treated and nonantimicrobial-treated comingled beef cows at a cow–calf operation. *Microbial Drug Resistance*, 22(7):598–608, October 2016.

[325] Maria S. Ramirez and Marcelo E. Tolmasky. Aminoglycoside modifying enzymes. *Drug resistance updates: reviews and commentaries in antimicrobial and anticancer chemotherapy*, 13(6):151–171, December 2010.

[326] Michael R Gillings, William H Gaze, Amy Pruden, Kornelia Smalla, James M Tiedje, and Yong-Guan Zhu. Using the class 1 integron-integrase gene as a proxy for anthropogenic pollution. *The ISME Journal*, 9(6):1269–1279, June 2015.

[327] Ilaria Frasson, Antonietta Cavallaro, Cristina Bergo, Sara N Richter, and Giorgio Palù. Prevalence of aac(6')-Ib-cr plasmid-mediated and chromosome-encoded fluoroquinolone resistance in Enterobacteriaceae in italy. *Gut Pathogens*, 3(1):12, August 2011.

[328] María de Toro, Irene Rodríguez, Beatriz Rojo-Bezares, Reiner Helmuth, Carmen Torres, Beatriz Guerra, and Yolanda Sáenz. pMdT1, a small ColE1-like plasmid mobilizing a new variant of the aac(6′)-Ib-cr gene in *Salmonella enterica* serovar Typhimurium. *Journal of Antimicrobial Chemotherapy*, 68(6):1277–1280, June 2013.

[329] Sally R. Partridge, Guy Tsafnat, Enrico Coiera, and Jonathan R. Iredell. Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiology Reviews*, 33(4):757–784, July 2009.

[330] David C. Hooper. Bacterial topoisomerases, anti-topoisomerases, and anti-topoisomerase resistance. *Clinical Infectious Diseases*, 27(Supplement_1):S54–S63, August 1998.

[331] P Heisig. Genetic evidence for a role of parC mutations in development of high-level fluoroquinolone resistance in *Escherichia coli*. *Antimicrobial Agents and Chemotherapy*, 40(4):879–885, April 1996.

[332] G.P. Pazhani, S. Chakraborty, K. Fujihara, S. Yamasaki, A. Ghosh, G.B. Nair, and T. Ramamurthy. QRDR mutations, efflux system & antimicrobial resistance genes in enterotoxigenic *Escherichia coli* isolated from an outbreak of diarrhoea in Ahmedabad, India. *The Indian Journal of Medical Research*, 134(2):214–223, August 2011.

[333] Amy L. Kullas, Michael McClelland, Hee-Jeong Yang, Jason W. Tam, AnnMarie Torres, Steffen Porwollik, Patricio Mena, Joseph B. McPhee, Lydia Bogomolnaya, Helene Andrews-Polymenis, and Adrianus W.M. van der Velden. L-asparaginase II produced by *Salmonella* Typhimurium inhibits T cell responses and mediates virulence. *Cell Host & Microbe*, 12(6):791–798, December 2012.

[334] Alexia N. Torres, Nayaret Chamorro-Veloso, Priscila Costa, Leandro Cádiz, Felipe Del Canto, Sebastián A. Venegas, Mercedes López Nitsche, Roberto F. Coloma-Rivero, David A. Montero, and Roberto M. Vidal. Deciphering additional roles for the EF-Tu, l-asparaginase II and OmpT Proteins of Shiga Toxin-Producing *Escherichia coli*. *Microorganisms*, 8(8):1184, August 2020.

[335] Sónia Ramos, Ingrid Chafsey, Michel Hébraud, Margarida SOUSA, Patrícia Poeta, and Gilberto Igrejas. Ciprofloxacin stress proteome of the extended-spectrum beta-lactamase producing *Escherichia coli* from slaughtered pigs. *Current Proteomics*, 13:285–289, 2016.

[336] Claudia Scotti, Patrizia Sommi, Maria Valentina Pasquetto, Donata Cappelletti, Simona Stivala, Paola Mignosi, Monica Savio, Laurent Roberto Chiarelli, Giovanna Valentini, Victor M. Bolanos-Garcia, Douglas Scott Merrell, Silvia Franchini, Maria Luisa Verona, Cristina Bolis, Enrico Solcia, Rachele Manca, Diego Franciotta, Andrea Casasco, Paola Filipazzi, Elisabetta Zardini, and Vanio Vannini. Cell-cycle inhibition by *Helicobacter pylori* L-asparaginase. *PLOS ONE*, 5(11):e13892, November 2010.

[337] Dirk Hofreuter, Veronica Novik, and Jorge E. Galán. Metabolic diversity in *Campylobacter jejuni* enhances specific tissue colonization. *Cell Host & Microbe*, 4(5):425–433, November 2008.

[338] Atin Sharma, Rajnikant Sharma, Tapas Bhattacharyya, Timsy Bhando, and Ranjana Pathania. Fosfomycin resistance in *Acinetobacter baumannii* is mediated by efflux through a major facilitator superfamily (MFS) transporter—AbaF. *Journal of Antimicrobial Chemotherapy*, 72(1):68–74, January 2017.

[339] Gabriel Kambale Bunduki, Eva Heinz, Vincent Samuel Phiri, Patrick Noah, Nicholas Feasey, and Janelisa Musaya. Virulence factors and antimicrobial resistance of uropathogenic *Escherichia coli* (UPEC) isolated from urinary tract infections: A systematic review and meta-analysis. *BMC Infectious Diseases*, 21(1):753, August 2021.

[340] Bin-Hsu Mao, Yung-Fu Chang, Joy Scaria, Chih-Ching Chang, Li-Wei Chou, Ni Tien, Jiunn-Jong Wu, Chin-Chung Tseng, Ming-Cheng Wang, Chao-Chin Chang, Yuan-Man Hsu, and Ching-Hao Teng. Identification of *Escherichia coli* genes associated with urinary tract infections. *Journal of Clinical Microbiology*, 50(2):449–456, February 2012.

[341] Muhamad Ali K. Shakhatreh, Samer F. Swedan, Ma'en A. Al-Odat, and Omar F. Khabour. Uropathogenic *Escherichia coli* (UPEC) in Jordan: Prevalence of urovirulence genes and antibiotic resistance. *Journal of King Saud University - Science*, 31(4):648–652, October 2019.

[342] Harold C. Neu. Aztreonam activity, pharmacology, and clinical uses. *The American Journal of Medicine*, 88(3, Supplement 3):S2–S6, March 1990.

[343] D. P. Bonner and R. B. Sykes. Structure activity relationships among the monobactams. *The Journal of Antimicrobial Chemotherapy*, 14(4):313–327, October 1984.

[344] Sion C. Bayliss, Rebecca K. Locke, Claire Jenkins, Marie Anne Chattaway, Timothy J. Dallman, and Lauren A. Cowley. Rapid geographical source attribution of salmonella enterica serovar enteritidis genomes using hierarchical machine learning. *eLife*, 12:e84167, April 2023.

[345] Michelle J. Bauer, Anna Maria Peri, Lukas Lüftinger, Stephan Beisken, Haakon Bergh, Brian M. Forde, Cameron Buckley, Thom Cuddihy, Patrice Tan, David L. Paterson, David M. Whiley, and Patrick N. A. Harris. Optimized method for bacterial nucleic acid extraction from positive blood culture broth for whole-genome sequencing, resistance phenotype prediction, and downstream molecular applications. *Journal of Clinical Microbiology*, 60(11):e01012–22, October 2022.

[346] Wei Gu, Xianding Deng, Marco Lee, Yasemin D. Sucu, Shaun Arevalo, Doug Stryke, Scot Federman, Allan Gopez, Kevin Reyes, Kelsey Zorn, Hannah Sample, Guixia Yu, Gurpreet Ishpuniani, Benjamin Briggs, Eric D. Chow, Amy Berger, Michael R. Wilson, Candace Wang, Elaine Hsu, Steve Miller, Joseph L. DeRisi, and Charles Y. Chiu. Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nature Medicine*, 27(1):115–124, January 2021.

[347] Satomi Mitsuhashi, Kirill Kryukov, So Nakagawa, Junko S. Takeuchi, Yoshiki Shiraishi, Koichiro Asano, and Tadashi Imanishi. A portable system for rapid bacterial composition analysis using a Nanopore-based sequencer and laptop computer. *Scientific Reports*, 7(1):5657, July 2017.

[348] Samuel Martin, Darren Heavens, Yuxuan Lan, Samuel Horsfield, Matthew D. Clark, and Richard M. Leggett. Nanopore adaptive sampling: A tool for enrichment of low abundance species in metagenomic samples. *Genome Biology*, 23(1):11, January 2022.

[349] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

[350] Wouter De Coster and Rosa Rademakers. NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics*, 39(5):btad311, May 2023.