



Universidade do Minho  
Escola de Engenharia

Metagenomic wastewater and freshwater core  
resistome analysis

Diogo Macedo Cachetas

**Metagenomic wastewater and  
freshwater core resistome analysis**

Diogo Macedo Cachetas

UMinho | 2022

outubro de 2022





**Universidade do Minho**

Escola de Engenharia

Diogo Macedo Cachetas

**Metagenomic wastewater and freshwater  
core resistome analysis**

Dissertação de mestrado

Mestrado em Bioinformática

Trabalho efetuado sob a orientação do(a)

**Professor Doutor Vítor Manuel Sá Pereira**

**Doutora Ivone Cristina Vaz Moreira**

outubro de 2022

## DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

### *Licença concedida aos utilizadores deste trabalho*



**Atribuição-NãoComercial-SemDerivações**

**CC BY-NC-ND**

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

## AGRADECIMENTOS

Gostaria de expressar os meus sentidos agradecimentos a quem em mim acreditou e que de alguma forma tenha contribuído e/ou apoiado durante o meu percurso académico.

Agradeço então, em primeiro lugar, à Dr.<sup>a</sup> Ivone Moreira, pela ajuda em todas as horas, para além de uma enorme orientadora e profissional uma excelente pessoa. Um pilar nos momentos mais críticos, grato pela amizade demonstrada e pela oportunidade de trabalhar com alguém assim. Obrigado por tudo.

Da mesma forma, agradeço ao Prof. Vítor Pereira por ter acompanhado e auxiliado na minha evolução desde seu aluno de mestrado até à conclusão da minha dissertação de mestrado. Por me ter desafiado a explorar os meus limites e por toda ajuda, obrigado.

A todos os meus colegas de trabalho da Escola Superior de Biotecnologia o meu obrigado, e com destaque, à Prof.<sup>a</sup> Dr.<sup>a</sup> Célia Manaia pela oportunidade de trabalhar com esta equipa e por todo o conhecimento que me foi transmitido.

Aos meus amigos, André e José, porque estiveram presentes desde os meus primeiros passos na Universidade do Minho e, com toda a certeza que estarão até ao último. Foram tantas as horas que juntos superamos os nossos obstáculos, fomos uma verdadeira equipa. O meu mais sincero obrigado. Agradeço ainda ao Flávio, por todo o apoio e inspiração.

Por fim, agradeço à minha família, em especial à minha Mãe e ao meu Pai, nos dias em que me faltava a motivação era neles que encontrava a força para continuar. Obrigado por sempre acreditarem em mim.

*“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of light, it was the season of darkness, it was the spring of hope, it was the winter of despair.”*

**Charles Dickens**

## **DECLARAÇÃO DE INTEGRIDADE**

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

## RESUMO

A escassez de água é atualmente uma grande preocupação. A reutilização de águas residuais tratadas, seja por descarga em ambientes hídricos (por exemplo, rios) ou pela utilização em irrigação, é apontada como umas das principais soluções. No entanto, é importante monitorizar o possível impacto desta reutilização, sobretudo ao nível da disseminação de contaminantes emergentes como as bactérias resistentes a antibióticos (ARB) e os seus genes de resistência a antibióticos (ARGs). Este estudo teve como objetivo determinar e comparar os resistomas de diferentes amostras de água (afluente, lamas ativadas, efluente e água doce), com base em metagenomas de diferentes geografias, de bases de dados públicas. O objetivo final foi identificar padrões e características distintas entre amostras. Estes permitirão identificar ARGs como possíveis biomarcadores para monitorizar a contaminação de ambientes aquáticos com agentes biológicos de origem antropogénica.

No total, 139 metagenomas (30 afluente, 30 lamas ativadas, 21 efluente, 58 de água doce) de 24 países foram analisados, usando métodos baseados em *assembly* e em *reads*. Os resultados mostraram que diferentes tipos de água partilham um grande número de ARGs. Uma nova abordagem foi usada para combinar a anotação de duas das bases de dados de ARGs mais abrangentes (CARD e ResFinder), superando a dificuldade que é lidar com anotações distintas provenientes de bases de dados diferentes. Esta abordagem permitiu determinar o resistoma *core* dos diferentes tipos de água, com o objetivo de obter genes biomarcadores para rastrear a contaminação em termos de resistência a antibióticos provocado pela descarga de águas residuais em ambientes recetores, como água doce. No final foram obtidos 60 possíveis biomarcadores, para os quais foram desenhadas sequências consenso que poderão ser usadas, por exemplo, para o desenho de *primers*.

Além disso, 7 modelos de *deep learning* foram desenvolvidos para classificar a transferibilidade de ARGs (genes adquiridos *versus* intrínsecos), dada a falta de informação sobre transferibilidade. Esta distinção é muito importante quer na monitorização quer na predição do risco, visto que os ARGs adquiridos são mais propensos à disseminação entre bactérias. O modelo de Redes Neurais Convolucionais superou os restantes com destaque (MCC de 0.881 e ROC-AUC de 0.906), o que é considerado um desempenho consistente.

**Palavras-chave:** biomarcador; *deep learning*; resistência a antibióticos; resistoma.

## ABSTRACT

Water scarcity is a major concern nowadays. The reuse of treated wastewater, by discharging in surface water bodies (e.g. rivers) or by utilization for irrigation, is pointed as one of the main solutions. However, it is important to monitor the possible impact this reuse, namely in terms of dissemination of contaminants of emerging concern such as antibiotic resistant bacteria (ARB) and their antibiotic resistance genes (ARGs). This study aimed to determine and compare the resistomes (set of genes associated with antibiotic resistance) of different water samples (urban wastewater influent, sewage sludge, final effluent and freshwater), based on metagenomes collected worldwide and available in public databases. The final goal was to identify overlaps and distinctive features among those compartments. This approach will permit the identification of ARGs to be used as possible biomarkers to monitor the contamination of aquatic environments with these biological contaminants of anthropogenic origin.

A total of 139 metagenomes (30 influent, 30 sludge, 21 effluent, 58 freshwater) from 24 countries were analysed, using assembled-based and reads-based methods. The results shown that different water types share a large number of ARGs. A new approach was used to combine the annotation of two of the most comprehensive ARGs databases (CARD and ResFinder), surpassing the difficulty that is to deal with different annotations coming from different databases. This approach allowed to determine the core resistomes, aiming to obtain biomarker genes to trace antibiotic resistance contamination from wastewater in receiving environments, such as freshwater. At the end 60 putative biomarkers were obtained, for which were designed consensus sequences that can be used, for example, to design primers for the genes monitoring.

Additionally, 7 deep learning models were developed and compared for classifying ARGs transferability (acquired versus intrinsic genes), motivated by the lack of information regarding transferability. This distinction may be very important in the monitoring and prediction of risk, since acquired ARGs are more prone to spread among bacteria. After validation, a Convolutional Neural Networks model outperformed the remaining with a 0.881 MCC and a 0.906 ROC-AUC, which is considered very consistent performance.

**Keywords:** antibiotic resistance; biomarker; deep learning; resistome.



## CONTENTS

1.	Introduction.....	1
1.1.	Context/Motivation.....	1
1.2.	Objectives.....	2
1.3.	Thesis Structure.....	3
2.	State of the art .....	4
2.1.	Antibiotic Resistance .....	4
2.2.	Dissemination of Antibiotic Resistance.....	6
2.3.	Antibiotic resistance in the environment.....	7
2.4.	Wastewater Treatment Plants .....	9
2.5.	Metagenomic Surveillance of Antibiotic Resistance Genes.....	11
2.6.	Core Resistome Analysis .....	13
2.7.	Bioinformatic methods for the analysis of antibiotic resistome.....	15
2.8.	Machine Learning.....	20
2.8.1.	Unsupervised Machine Learning.....	21
2.8.2.	Supervised Machine Learning .....	22
2.8.3.	Feature Selection.....	24
2.8.4.	Error Estimation .....	24
2.9.	Deep Learning (DL) .....	24
2.9.1.	Artificial Neural Networks (ANN) .....	25
2.10.	Machine Learning Applied to Metagenomic Surveillance .....	28
3.	Materials and Methods .....	28
3.1.	Metagenomic Dataset.....	29
3.2.	Reads Processing and Assembly.....	30
3.3.	Taxonomic Annotation .....	30
3.4.	ARG Database Alignment .....	32
3.4.1.	Case Study 1 (assembled reads).....	32
3.4.2.	Case Study 2 (raw reads).....	33
3.5.	Common Database: CARD & ResFinder .....	33
3.5.1.	Case Study 1 .....	34

3.5.2.	Case Study 2 .....	34
3.6.	Core Resistome Pipeline.....	35
3.6.1.	Core Resistome.....	35
3.6.2.	Statistical Analysis .....	36
3.6.3.	Identification of possible biomarkers and Consensus Sequences .....	36
3.6.4.	Biomarkers validation .....	37
3.7.	Classifying ARGs transferability through Deep Learning .....	37
4.	Results and Discussion .....	40
4.1.	Samples Selection.....	40
4.2.	Taxonomic Annotation.....	43
4.3.	ARGs Profiling and core resistome definition .....	46
4.4.	Biomarker ARGs and Consensus Sequences .....	53
4.5.	ARGs Transferability (Deep Learning) .....	54
5.	Conclusion and Future Work.....	58
6.	Bibliography.....	60

## LIST OF FIGURES

Figure 1. Antibiotic targets and examples of resistance mechanisms for selected antibiotics (adapted from [13]).....	5
Figure 2. Potential routes of creation of antibiotic residues in the environment and transmission to and from the environment of antibiotic residues, antibiotic-resistant bacteria, and antibiotic resistance genes (adapted from [61]).....	8
Figure 3. Bioinformatic workflow to conduct ARG metagenomic surveillance: after accessing the reads quality the assembly step is optional as the alignment tool can be chosen considering if reads or assembled reads are being aligned. ....	16
Figure 4. Proposed workflow to select a metagenome assembler based on the research question, the computational resources available, and the bioinformatics expertise of the researcher (adapted from [115]). ....	18
Figure 5. Examples of Supervised Learning (Linear Regression) and Unsupervised Learning (Clustering) [155].....	21
Figure 6. A typical ANN and a typical artificial neuron (from [167]).....	25
Figure 7. Convolutional neural network (from [167]).....	26
Figure 8. Recurrent neural network (from [167]). ....	27
Figure 9. Example of an autoencoder (from [169]).....	27
Figure 10. SILVA database processing for accurately quantifying 16s rRNA in metagenomic samples (from the SortMeRNA set construction [175]). ....	31
Figure 11. EMBL-EBI Custal Omega parameters. ....	36
Figure 12. EMBL-EBI Emboss Cons parameters. ....	37
Figure 13. ProPythia config.json file setup.....	39
Figure 14. a) SILVA reads per total reads ratio; SILVA reads per total contigs ratio: b) with a 75% coverage threshold; c) with a 50% coverage threshold; d) with a 10% coverage threshold. ....	44
Figure 15. PCA of taxonomic analysis with the relative abundance of Kaiju reads: a) phylum; b) class; c) order; d) family; e) genus and f) species. ....	45
Figure 16. Relative abundance of the 10 most abundant AMR classes, using KMA “-ef” tag with the ResFinder database.....	48

Figure 17. ARG reads relative abundance: on the left side using ResFinder database and on the right side using the CARD database. ....	49
Figure 18. ARG reads per 16S rRNA reads: on the left side using ResFinder database and on the right side using the CARD database. ....	49
Figure 19. ResFinder core Resistome for 90% prevalence using KMA and raw reads.....	51
Figure 20. CARD core Resistome for 90% prevalence using KMA and raw reads.....	51
Figure 21. Cluster Reunion core Resistome for 90% prevalence using KMA and raw reads...	52
Figure 22. Heatmaps of the Water Types Biomarkers presence in the freshwater and wastewater dataset.....	54
Figure 23. On the left is represented the scatter plot for the original dataset distribution using descriptors, on the right side is represented the dataset with the oversampling using SMOTE. ....	55

## LIST OF TABLES

Table 1. Summary of antimicrobial resistance reference databases (adapted from [131])....	18
Table 2. Filter set up for the construction of a freshwater and wastewater metagenomic dataset.....	29
Table 3. Core resistome (ARGs identified in >90% of the samples), using the ResFinder and CARD/RGI tools, using the assembled contigs and the raw reads, “Rf_to_CARD” uses the merged database annotation applied to the “Reunion” column. ....	47
Table 4. Reports of Deep learning models obtained through ProPythia: multilayer perceptron (MLP); convolutional neural networks (CNN); Long short-term memory (LSTM); gated recurrent unit (GRU).....	56
Table 5. Reports of the previously obtained models using Stratified kfold validation (k=5). ..	56

## ACRONYMS

### A

<b>AE</b>	Autoencoder
<b>AI</b>	Artificial Intelligence
<b>AMR</b>	Antimicrobial Resistance
<b>ANN</b>	Artificial Neural Networks
<b>AR</b>	Antibiotic Resistance
<b>ARB</b>	Antibiotic Resistant Bacteria
<b>ARG</b>	Antibiotic Resistance Gene
<b>AUC</b>	Area Under the Curve

### B

<b>BLAST</b>	Basic Local Alignment Search Tool
<b>BN</b>	Bayesian Networks

### C

<b>CARD</b>	Comprehensive Antibiotic Resistance Database
<b>CNN</b>	Convolutional Neural Networks

### D

<b>DL</b>	Deep Learning
<b>DNA</b>	Deoxyribonucleic acid
<b>DNN</b>	Deep Neural Networks
<b>DT</b>	Decision Trees

### E

<b>EMBL-EBI</b>	European Bioinformatics Institute
<b>EUCAST</b>	European Committee on Antimicrobial Susceptibility Testing

### G

<b>GRU</b>	Gated recurrent unit
------------	----------------------

### H

<b>HGT</b>	Horizontal Gene Transfer
------------	--------------------------

<b>HMM</b>	Hidden Markov models
<b>K</b>	
<b>KMA</b>	k-mer alignment
<b>L</b>	
<b>LSTM</b>	Long short-term memory
<b>M</b>	
<b>MCC</b>	Matthews Correlation Coefficient
<b>MGE</b>	Mobile Genetic Elements
<b>ML</b>	Machine Learning
<b>MLP</b>	Multilayer Perceptron
<b>N</b>	
<b>NB</b>	Naïve Bayes
<b>NCBI</b>	National Center for Biotechnology Information
<b>NGS</b>	Next-Generation Sequencing
<b>P</b>	
<b>PCA</b>	Principal Component Analysis
<b>PCR</b>	Polymerase chain reaction
<b>R</b>	
<b>RGI</b>	Resistance Gene Identifier
<b>RNA</b>	Ribonucleic acid
<b>RNN</b>	Recurrent Neural Networks
<b>ROC</b>	Receiver Operating Characteristic
<b>S</b>	
<b>SA</b>	Stacked Autoencoders
<b>SGS</b>	Second-Generation Sequencing
<b>SILVA</b>	Ribosomal RNA database
<b>SMOTE</b>	Synthetic Minority Oversampling Technique
<b>SRA</b>	Sequence Read Archive

**SVM** Support Vector Machines

**T**

**TGS** Third-Generation Sequencing

**V**

**VGT** Vertical Gene Transfer

**W**

**WGS** Whole Genome Sequencing

**WMS** Whole Metagenome Shotgun Sequencing

**WSL** Windows Subsystem for Linux

**WWTP** WasteWater Treatment Plant



# 1. Introduction

## 1.1. Context/Motivation

Antibiotic resistance (AR) showcases threatening follow-ups to human and wildlife health, resulting in an urge to minimize and control the risks of exposure to antibiotic resistant bacteria and genes (ARB & ARGs) [1]. Upon the introduction of antibiotics intake, their abusive consumption for human, veterinary, and agricultural purposes over the past years led to its ongoing release into the surrounding ecosystems, facilitating the appearance of antibiotic resistance. The increasing load and diversity of ARB & ARGs in aquatic environments, with an emphasis on wastewater [2], points to wastewater treatment plants (WWTP) as focal hotspots for their dissemination [3].

Over the past years, clinical settings have been prioritized over environmental resistance dissemination [4]. Due to the lack of information on resistance spread in the environment, transmission patterns need to be determined, resorting to a statistical characterization of the core resistome (here defined as a set of ARGs that are characteristic of a particular type of environment). The metagenomics of the urban wastewater and freshwater, as well as the human gut microbiota, need to be characterized and analyzed following detailed methods [5] while emphasizing the One-Health approach.

Although water treatment processes are thought to remove ARGs effectively, WWTPs are also considered major reservoirs of antibiotic resistance [6]. In this work, the resistomes of the wastewater influent, sewage sludge, and final effluent will be determined and compared with the freshwater resistome. This comparison will show if WWTPs act as crucial firewalls for the One-Health compartments or as focal points of resistance dissemination to the environment.

Furthermore, a dataset consisting of multiple sources of metagenomic AR environments was constructed where deep learning algorithms was applied to classify ARGs transferability.

## 1.2. Objectives

This project aimed to develop an efficient framework to evaluate the presence of ARGs in wastewater and freshwater through core resistome analysis from metagenomic datasets, and further comparison with the human gut microbiota resistome. More precisely, the specific objectives of this work were the following:

- Identification of the freshwater background resistome;
- Identification of the wastewater core resistome, aiming to identify possible signatures of wastewater contamination;
- Obtain biomarker genes for tracking and monitoring AR contamination;
- Use Deep learning methods to classify the potential for transferability of ARGs – intrinsic or acquired.

To reach those objectives the following tasks were developed:

- Construction of a metagenomic wastewater and freshwater dataset with an extensive geographic reach, including influent, sludge, and effluent samples;
- Identification of the ARGs present in the metagenomes, through the search in ARGs databases, namely, ResFinder and CARD;
- Identification of the core resistomes (for wastewater influent, sludge, effluent, and freshwater) through statistical differential analysis and network evaluation;
- Construction of an ARG annotation database and a querying system for core resistome characterization;
- Obtain an ARGs nucleotide dataset to build a deep learning-based model and validate the classification model.

### **1.3. Thesis Structure**

This thesis is grounded on a summary introduction to the topic and the definition of the objectives that motivated the development of this work, with the respective list of tasks developed. Chapter 2 presents an analysis of the state-of-the-art with particular attention to antibiotic resistance, wastewater treatment plants as important antibiotic resistance hotspots, bioinformatic tools available for resistome analysis and a brief description of machine learning and deep learning algorithms aiming towards the development of a deep learning model for classification of ARGs transferability. The same chapter further spans to identify the main persisting questions, current solutions, and critical problems. The remaining chapters target the development of the followed methodology, the obtained results and discussion, closing with the final remarks and future works.

## 2. State of the art

### 2.1. Antibiotic Resistance

Antibiotics are chemicals that affect metabolic pathways, inhibiting growth or eliminating microorganisms [7]. Considering the chemical structure of antibiotics, these can be divided into seven major groups, namely: macrolides,  $\beta$ -lactams, tetracyclines, quinolones and fluoroquinolones, sulfonamides, phenicols, and aminoglycosides [8]. With the discovery of penicillin in 1928 and until the 1950s, antibiotic discovery thrived. However, the recent gradual decline in antibiotic discovery and the evolution of antibiotic resistance as led to the present antimicrobial resistance crisis [9]. A resistance event happens when a bacterial strain shows tolerance to a higher minimal inhibitory concentration than the subsequent parental wild-type strain [10]. Therefore, when a resistance gene or resistance factor is present, it allows bacteria to tolerate higher antibiotic concentrations or, under the same ground, its absence increases susceptibility to an antibiotic [11]. There are four main mechanisms of microbial resistance (Figure 1): 1) antibiotic removal from the bacteria through an efflux pump; 2) creation of an alternative metabolic pathway acting similarly to the suppressed path or restraining the need for the metabolites produced in the inhibited pathway; 3) modification of the antibiotic target, and 4) enzymatic inactivation of the antibiotic [12]. AR is presented as one of the most significant hazards for human health, challenging health care treatment of threatening infections [14]. Recent impact assessments (2017) indicate that ARB are accountable for at least 23 000 deaths per year in the U.S. and nearly 25 000 deaths per year in Europe. The position of underdeveloped regions is pointed to be even worse [15]. It is expected that antimicrobial-resistant infections will kill globally as nearly as 700 000 people per year within 30 years and that the number of resistant infections will reach an estimate of 10 000 000 deaths each year, surpassing cancer death tolls [16]. Past studies, considering 71 countries, also report that there has been an increase in the usage of antibiotic drugs, from 54.1 billion standard units in 2000 to 73.6 billion standard units in 2010, translating into a total of 35% global increase [17]. Although antibiotics provide effective therapies against several types of infections, their abusive usage results in the spread of antibiotics and AR in the

environment [18]. There is still a significant lack of knowledge about the circumstances that motivate resistance development in the environment, which challenges control efforts to mitigate mobile resistance [19]. In response to emerging ARB, which could mean a step back into a pre-antibiotic era [20], surveillance programs have been developed for mitigating the spread of ARGs [21]. Consequently, based on the aforementioned and other reasons, the World Health Organization has considered AR a global public health crisis in the 21<sup>st</sup> century, among the biggest threats to human health [22].

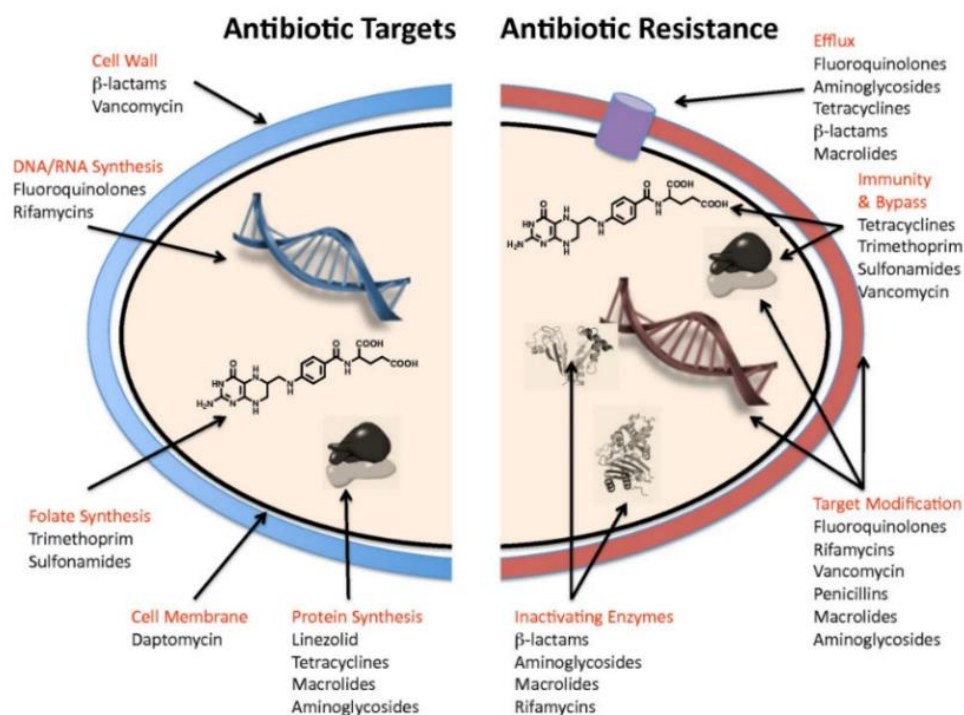


Figure 1. Antibiotic targets and examples of resistance mechanisms for selected antibiotics (adapted from [13]).

One-Health is defined as a concept that motivates worldwide collaboration, gathering efforts to achieve sustainable well-being between human, environment, and animal health [23]. To face AR, and according to the One-Health concept, surveillance and control measures need to be implemented in human, natural, and animal environments, based on the assumption that AR is spreading throughout these main compartments [24]. This underlines the need to prioritize research efforts to cope with AR, as many human lives, general well-being, and economic breakdown prevention depend on this matter of concern [25].

## 2.2. Dissemination of Antibiotic Resistance

A bacterial host can acquire AR via three distinct routes: vertical gene transfer (VGT), *de novo* mutation, and horizontal gene transfer (HGT) [26]. Acquired AR translates into the ability of a bacteria to inhibit antibiotic activity, whilst similar or phylogenetically related bacteria are usually susceptible [27]. On the other hand, intrinsic AR is observed in mostly all bacteria of the same genera, being disseminated via VGT. It can be due to morphological traits or the presence of a distinct set of genes in a particular taxon [28]. ARGs usually have minimal effect on the bacteria overall fitness, as the presence of an antibiotic commonly induces their expression and therefore lowers the fitness cost of keeping the ARG [29].

Resistance attributes are extensively distributed throughout the microbiota community and can be disseminated by more than one well-known gene transfer mechanism in a wide range of pathogens and commensals [30]. New AR factors could potentially appear anywhere, on any occasion, due to an immense phylogenetic variability, which prompts opportunities for novel mutations, rearrangements, and HGT [31]. HGT is one of the most common mechanisms for novel and known ARGs spread, with the ability to promote the dissemination of AR between different bacterial communities [32], frequently beyond their boundaries [33]; additionally, this process is usually stimulated by stressors, often antibiotics [34]. Although HGT is more prone to occur among phylogenetically related bacteria [35], HGT from different environmental bacteria to pathogenic bacteria can occur if these share the same habitat [36]. There are four main identified pathways through which bacteria transfer ARGs horizontally, including conjugation, transformation, transduction, and vesiduction [37] [38]. Conjugation is pointed to as the most common pathway considering the remaining ones for HGT [39]. The ARGs are transferred through cell-to-cell contact via mobile genetic elements from a donor bacterium to a recipient bacterium. This process is also reported to happen between phylogenetically distant bacteria [40]. Differently from other pathways of HGT, transformation does not require viable donor cells [41], as it occurs through the uptake and integration of extracellular plasmid or chromosomal DNA by a recipient cell [42]. Thus, transformation enables ARG dissemination between different genera of bacteria [43]. Transduction is related to ARGs loaded bacteriophages, resulting in the spread of AR to an infected receiving bacterium [44]. Vesiduction occurs from membrane vesicles secreted from

a donor cell that carry ARG-containing DNA and fuse with the membrane of a receiver cell [45].

ARGs can be exchanged from environmental reservoirs to human pathogens in a complex process. The initial mobilisation is usually followed by more than one dissemination and adaptation stage, ARGs neighbouring sequences have been found to encode mobilisation elements (transposases and integrases), which are involved in the gene transfer process between bacterial genomes [46]. The initial mobilisation is often facilitated by mobile genetic elements (MGEs) [47], these can capture ARGs from chromosomes and horizontally transfer them through a plasmid or bacteriophage to another bacteria [48]. The ongoing acquisition of ARGs by human pathogens urges the development of new methodologies to predict the dissemination of ARGs, and such attempts have struggled to succeed due to the great number of ARGs held in environmental and artificial reservoirs [49].

### **2.3. Antibiotic resistance in the environment**

As described in the previous section, ARB can occur by mutations in the pre-existing bacterial genome or by the uptake of environmental DNA [50]. Although external environments are less contributive to mutation-based AR, the uptake of novel resistance is supported by precursor environments which provide an unmatched ecological niche with a substantial gene pool [51]. Along with antibiotic release facilitating HGT of ARGs, the presence of contaminants, such as heavy metals, plays a crucial role as it also induces selective pressure, affecting microbial communities promoting HGT [52]. Abiotic factors (e.g. pH, temperature, and nutrient abundance) also condition the spread of ARGs in the environment as they cause selective pressure and variability in bacterial communities [53]. Concerns over human health motivated by contaminated antibiotic environments are driven by antibiotic residues in the environment, which could result in its ingestion, altering the human microbiome, endorsing human gut resistant bacteria [54], and the occurrence of naturally developed resistance hotspots as a result of induced selective pressure [55]. The impact assessment of residual antibiotics in the environment shows its influence on reproduction, metabolism, changes in

the population structure, and ecological function of the ecosystem, including biomass and biodiversity [56].

The most significant proportion of released antibiotics is addressed to intake of antibiotics. Therefore, its excretion and release levels into the environment are limited by the seasonality of antibiotic usage, the doses used, and the processing metabolism (human or animal) [50]. Resistance dissemination between different environments has been stimulated by abusive antibiotic usage and its consequent release into the environment through anthropogenic activities [57]. A large number of intake antibiotics are excreted in their biologically active form [58], these are found in urine and faeces from both humans and animals, being released into complex environments such as soil and water [59] – hospital wastewater, wastewater treatment plants, sediments, animal manure, and agricultural soils are considered main precursors in AR spread as they act as natural containers for antibiotics, microbial communities, ARGs, and MGEs (Figure 2) [60].

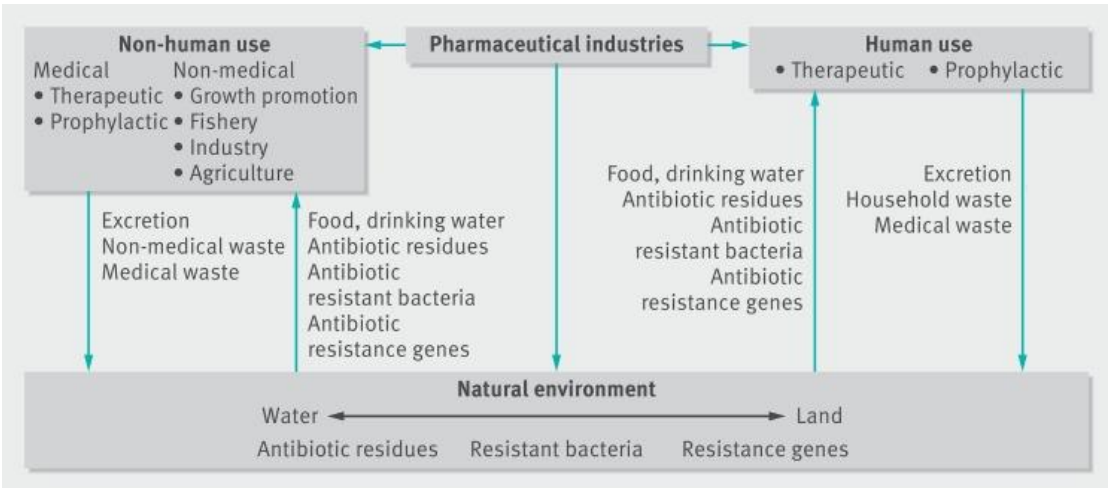


Figure 2. Potential routes of creation of antibiotic residues in the environment and transmission to and from the environment of antibiotic residues, antibiotic-resistant bacteria, and antibiotic resistance genes (adapted from [61]).

As described, ARB and their ARGs are being found in multiple and new contaminated environments [62]. There is a significant predominance of ARGs in soil. The primary sources are agricultural antibiotics found in manure from livestock and general wastewater (urban, hospital, aquaculture, and agricultural) [63]. Soil ARGs motivate the transmission of ARGs across the food chain, as microbial activities in soil, rhizosphere, and phyllosphere link the soil, microorganism, and plant to form an adjacent ecosystem crucial to the human food chain [64]- The accumulation of ARGs in plants is often associated with the application of manure as



fertilizer and irrigation with water containing antibiotics [65], the ecological and environmental effects of antibiotics show that some may affect plant growth and development [66]. At the same time, it has been recognized that livestock and its derivatives promote ARGs spread to humans, as most antibiotic applications are for agricultural use, encouraging ARGs appearance in the animal gut and faeces [63]. AR contaminants in soil can be transferred to water receiving environments and vice versa [67]. ARGs in water bodies are found in urban water systems as they hold human-associated ARGs, animal sources point mainly to livestock storage ponds and swine treatment lagoons [63]. Nowadays, besides freshwater and wastewater, contamination through AR factors also reaches main saltwater environments [68].

The routes and patterns for the transmission of AR to humans are still very unclear. The current state of the art indicates that a complex combination of variables is responsible for the transmission risks associated with humans. These include environmental compartments, ubiquitous bacteria, and human bacteria interaction [69]. Furthermore, few data is available regarding sources and mechanisms of ARGs transfer in clinical settings, although numerous cases of ARG detection on hospital surfaces and in patients were described [16]. The transmission of environmental ARGs to humans is believed to be carried through by pathogenic ARB or human commensal ARB, which can colonize and proliferate in the human gut. Additionally, pathogenic ARB can cause infectious disease, and commensal ARB are thought to be capable of transferring ARGs to other commensals in human microbiota [70]. Antibiotic residues pose a real potential threat to human health, and, in addition to promoting bacterial resistance, these include direct side effects causing distress in the human microbiome and triggering health problems, such as psoriasis, colitis, cardiovascular problems, asthma, diabetes, obesity, and colorectal carcinoma [54].

## **2.4. Wastewater Treatment Plants**

Antibiotic residues have been detected in various environments, such as wastewater treatment plants (WWTP), livestock farms, rivers, wastewater irrigated soils, groundwater, surface water, seawater, drinking water, water purification plants, landfills, and sediments [8],

[71]. Water, accordingly to the One-Health context, is an environmental compartment that is pointed to as the main link between human, animal, and natural environments, as it is an unrestrained route that conditions the transport and flux of abiotic factors offering bacteria a core habitat with high AR propagation potential [25]. The water environment represents a major microbial niche and is associated with the origin of resistance genes. It is also mentioned as an amplifier and reservoir of AR factors, acting as a container and enhancing the interaction of ARGs between different bacteria [72].

In the urban water cycle, wastewater is described as liquid discharge from human households. WWTPs are considered terminals of complex sewer systems where sewage is treated and the main pollutants are removed before returning to the environment. The result is an effective firewall for the One-Health compartments [73].

WWTPs are the main focal points for disseminating AR since most antibiotics used by human and veterinary medicine are excreted in partially metabolized forms, entering sewer networks, and reaching into WWTPs [74]. Antibiotics from pharmaceutical industries could also be conveyed in sewage systems to WWTPs. They tend to resist biodegradation and the wastewater treatment processes are not optimized for their removal. The presence of antibiotics in the sludge and final effluent of WWTPs causes their release to the surrounding recipient environments, namely, surface water, soil, and groundwater [8].

Diverse bacteria enter WWTPs systems, along with a high content of resistance factors, facilitating gene transfer within present bacteria. In addition, the existence of subclinical levels of antibiotics and heavy metals present in low concentrations in wastewater further increases the selective pressure of resistant strains in the urban water cycle [75]. Resistant bacteria in WWTPs have been recognized for several years, and techniques to mitigate their presence in effluent water have been highlighted as the main priority. Such techniques include water filtration that removes a high portion of bacteria, benefiting from recent membranes that further decreased their concentration [76]. Chlorination and UV treatment are also indicated as beneficial disinfection treatments regarding water quality and security [77]. Although numerous water treatment processes are thought to effectively inactivate ARB and remove ARGs, WWTPs are starting to be considered important reservoirs of resistance due to their functionality and specificity. Available studies underline the increasing abundance of AR

factors at WWTPs [6]. Even though the deletion of influent resistant bacteria is substantial, a high dose of resistant bacteria is still detected in the sludge and treated effluent [78].

The monitorization of AR in WWTPs is commonly associated with the populations of *E. coli* and *Enterococcus* spp. These bacteria are often used as faecal indicators given that they exhibit high resistance to conventional antibiotics such as aminopenicillins, sulphonamides, tetracyclines or tetracycline, and erythromycin respectively [79]. Regardless of the significance of *E. coli* and *Enterococcus* spp. as indicators of human faecal contamination, these bacteria are not considered the most common bacterial groups in wastewater. Curiously, *E. coli* and enterococci are indicated as minor representatives. The most abundant community members may also perform crucial roles regarding resistance dissemination [72]. Studies on monitorization of WWTPs and urban wastewaters are also reporting changes in the urban sewage resistance dissemination, usually related to seasonal variations, as most commonly winter surpasses the summer resistance loads [20] and the geographical distribution of WWTPs, being highlighted as an important factor in the variation of WWTPs bacterial communities, as network analysis indicate that bacteria from high-capacity WWTPs are further correlated than those from low-capacity WWTPs [80].

Recent studies show that WWTPs significantly reduce the abundance of ARGs and ARB in the treated effluent [81]. However, the treatment may also be responsible for an enrichment of ARGs and ARB in the final effluent, since it may promote a higher removal of susceptible bacteria. A recent study measured the impact of the discharges of 4 urban WWTPs in the receptor rivers [82]. A major conclusion was that the impact of the urban WWTPs on the river was not only determined by treatment efficiency and final effluent quality, but also by the background contamination of the river and/or dilution rate. This and other studies usually involve the quantification of ARB and ARGs through qPCR, emphasizing the need for representative biomarker sequences that allow good monitoring [83].

## **2.5. Metagenomic Surveillance of Antibiotic Resistance Genes**

The lack of supervision regarding AR is increasing the rate of AR dissemination at a regional and global scale [84]. Insufficient research funding and lack of surveillance programs

contribute to the dissemination of resistance, mainly in developing countries [85]. The solution may be the incorporation of metagenomics into frameworks for ARGs monitorization in the One-Health environments, which provides a novel approach for early assessment [86]. Through an epidemiological insight, the surveillance and analysis of AR in WWTPs has been recurring due to its significant advantages by delivering a broader perspective on the dissemination of ARGs, overcoming conventional surveillance restrictions due to the sampling size or the scarcity of data only provided by clinical environments [4].

Advances in culture-independent molecular biology techniques have facilitated the study of ARGs both qualitatively and quantitatively, such as correlation analysis, metagenomics, fluorescence-activated cell sorting (FACS), single-cell fusion PCR, genomic crosslinking, quantitative PCR (qPCR), high-throughput qPCR (HT-qPCR), and digital PCR (dPCR) [87], [88]. Currently, there are three main molecular methods to quantify ARGs in the environment, their distribution and propagation, namely, polymerase chain reaction (PCR) based amplification of ARGs, hybridization of DNA to ARG fragments, and metagenomic data analysis for ARGs [65]. Several tools based on PCR have been upgraded to be incorporated in microbial genetics in response to recent problems. ARG quantification amplification dependent methods, namely, qPCR, HT-qPCR, and dPCR, are extremely important due to their simple execution, robustness, specificity, and sensitivity [89]. On the other hand, metagenomics sequencing is a culture-independent method for characterizing microbial communities through shotgun or whole-genome sequencing [90]. This method is based on the assembly of contigs from initial raw reads or by the reconstruction of metagenomes from sequencing reads linking ARGs to a given taxonomy. ARG-hosts can be identified with host phylogenetic biomarkers or by annotating genes clustering with the target ARGs [87]. Metagenomics relies on next-generation sequencing (NGS) platforms, namely, second-generation sequencing (SGS) platforms such as Illumina, offering high throughput and accuracy. Although Illumina has the downside of generating short DNA segments, this problem has been overcome in third-generation sequencing (TGS) platforms (PacBio and Oxford Nanopore Technologies) which can sequence ultralong reads [91]. On the downside, TGS shows significantly higher sequencing errors compared to SGS. However, recent upgrades in errors and data processing indicate that TGS accuracy will improve with technological development [91]. Comprehensive resistome analyses are now using whole metagenome

shotgun sequencing (WMS) to obtain broader information concerning a high number of bacterial species and resistance genes [92]. Metagenomics is a powerful next-generation tool that is the main framework for compiling the resistome of several environments. Different metagenomic approaches can be used for the mining of ARGs and AR factors present in metagenomic samples [93]. Metagenomic mining revealed that bacteria have acquired resistance genes before the antibiotic era, providing valuable insight into the evolution of resistance mechanisms [94]. It is now established that ARGs are an ancient and natural part of many bacterial genomes [19].

Results obtained from short sequence reads assemblies must be carefully interpreted as they are prone to assembly errors [95]. Assemblies driven from long DNA reads can uncover further information about the genetic context of ARGs, such as the potential for mobility, evaluate if it is plasmid or chromosomal related and if it is clustered with MGEs [89]. Metagenomic methods have several advantages over conventional approaches for determining ARG-host relations, mainly because they are nontargeted and not restricted by selecting preset ARGs and MGEs. Additionally, metagenomics can yield multiple details regarding the variety of ARGs and MGEs in distinct environments [87]. However, any ARG-host relation produced from metagenomic data sets should be carefully interpreted, as metagenomic assemblies do not portray strain variation. Moreover, multiple host plasmids that have developed in several hosts may not be assigned with taxonomic marker genes, and therefore can't be attributed to a given host. This is a major downside, as plasmid carrying ARGs display a greater risk to human health than chromosome associated ARGs, as these can be transferred across different species reaching pathogenic bacteria [87].

Thanks to the evolution of genomics and metagenomics, the fight against pathogenic bacteria has considerably changed. With access to the complete genome of different bacteria populations, the more precise selection rather than an experiential selection of DNA fragments enabled the creation of a wide range of detection methods and dedicated bioinformatics tools to identify AR [96].

## **2.6. Core Resistome Analysis**

Both benign and pathogenic bacteria were explored at a community level, and metagenomics was integrated to discover resistance factors from several environments. Research efforts contributing to the discovery of new antibiotic therapies and surveillance programs of rising resistance hazards must be prioritized, focusing on ARGs that are the most probable contenders for HGT to pathogens [97].

The focal picture of ARGs sources for tracing environmental pollution in the past few years has been based on a few specific ARGs, in most cases, these studies resulted in a poor vision of AR spread. A comprehensive insight of AR in a certain environment is reflected in the overall repertoire of resistance genes, the so-called resistome, gathering both intrinsic and acquired ARGs, encoding proteins with distinctive resistance and acquired resistance through mutation or HGT, as well as precursor ARGs, encoding proteins with putative antibiotic resistance activity and proteins of under-expressed AR [98]. Along with phenotypic resistance genes, phenotypic sensitive genes (silent and proto genes) can evolve to intrinsic and acquired ARGs integrating into the resistome: although silent genes are functional these are not expressed, these can become clinically significant by mutation or mobilization if their expression occurs; proto-resistance genes have little or no activity against antibiotics, but can gain activity via mutations, though these are part of the environmental pool, they have little clinical importance due to the need for activation and mobilization [99]. The resistome translates into the total amount of the previous resistance genes associated with an ecosystem. Both soil and water environments hold complex resistomes, which act as a reservoir of resistance genes for many human pathogens [97]. Continuous exposure of these microbial communities to antibiotics residues, and other contaminants, induces a selective pressure on the core resistome [100].

The bacteria ARGs associated phenotype frequently acts as a defence mechanism against antibiotics or toxins produced by participants in the same environmental population [101]. The continuous increase in the prevalence of ARB has uncovered knowledge limitations regarding the undergoing evolutionary and environmental processes in different microbial ecosystems [102]. Thus, constant updating of ARG databases is necessary, facilitating comprehensive profiling of the antibiotic resistome but also contributing to primers design for clearer ARGs detection [103].

Regarding the resistome analysis, it is clear that when analysing samples of environmental and clinical origins, less comprehensive databases are less effective in finding complete ARGs profiles from the complex understudied environmental samples in comparison to the well documented clinical samples. Finding an adequate database and algorithm with the required retrieval capability is essential for analysing environmental AR. Choosing a suitable database based on sample type and study objective is necessary. Comprehensive databases containing most ARG variants are keen for recognising ARGs from complex environmental communities, while specialized databases better characterize specific and novel ARGs [98].

Metagenomics enables resistome analysis between multiple environments, reflecting that several ARGs are ubiquitous in host-associated and natural environments. Gut microbiota can exchange ARGs and interact with bacteria transiting through the colon, triggering these bacteria to acquire and transfer ARGs [104]. There is a massive reservoir of ARGs in human gut and their correlated environments, which can be mobilized from these compartments to human pathogens [105]. Thus, following the growing alarms on ARG surveillance, the assessment of WWTPs performance in eliminating ARGs underpins the need for easy access repositories of resistome data for both government stewardship and academic reference as a response to the progressively increasing volume of metagenomic datasets [106].

## **2.7. Bioinformatic methods for the analysis of antibiotic resistome**

In surveillance studies and following sample collection, metagenomic sequencing and processing occur, involving the extraction of DNA, metagenome sequencing, and analysis [107]. The steps highlighted in Figure 3 are carried through with bioinformatic tools, from trimming to statistical analysis.

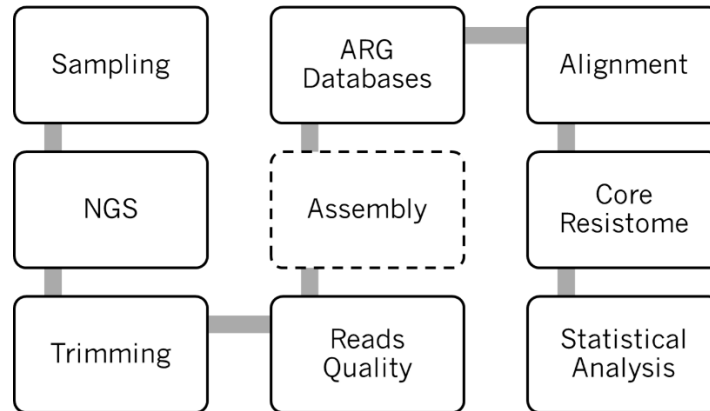


Figure 3. Bioinformatic workflow to conduct ARG metagenomic surveillance: after accessing the reads quality the assembly step is optional as the alignment tool can be chosen considering if reads or assembled reads are being aligned.

After sequencing, it is important to guarantee that just the high quality reads will be processed. For that, bioinformatic processing should do the trimming, removal of sequencing adaptors and low quality sequences, reducing bias information [108]. Although NGS technologies allow to sequence a metagenome with a high coverage, these technologies still have some limitations regarding the sequence length, rate of sequencing errors, and data with adapter sequences. Besides reducing the amount of usable data, these problems also impact the accuracy of further bioinformatics analysis [109]. So, it is really important the correct quality control of the raw reads. Many available tools can help to curate the raw reads, such as QcReads [109], AdapterRemoval [110], Cutadapt [111], Btrim [112] and Trimmomatic [113]. From the armour of NGS read preprocessing tools, when leveraging flexibility, paired-end data handling accuracy, and high performance, Trimmomatic is indicated as the more flexible and efficient preprocessing tool, especially when handling paired-end data [113]. FASTQC tool helps to evaluate the quality of the reads before and after processing, by assessing several parameters related to the reads quality [114].

The reads assembly usually follows the preprocessing. While shotgun metagenomic sequencing delivers inclusive access to microbial communities genomes, many of the encoded functional genes are considerably longer than the length of reads obtained via NGS [115]. Unassembled metagenomic data are more fragmented and prone to error with varying sequencing depths [116], although read-based methods can be used to detect ARGs without metagenomic assembly by applying pairwise alignment tools or splitting the reads into k-mers, it can prevent information loss but no positional information is retrieved, which is necessary



to analyse upstream and downstream factors of ARGs [117]. To achieve an accurate and comprehensive analysis, preprocessed reads are recommended to be assembled into larger DNA segments named contigs [118]. Several metagenome assembly tools (assemblers) were developed to handle this problem. Many assemble sequences via *de novo*, where metagenomic sequences are split into k-mers, overlapped into a network, and paths are crossed iteratively to obtain extended contigs [119]. Such a procedure allows for more confident gene predictions than those obtained from unassembled data [120].

Assembly quality has greatly extended the range of questions answered using NGS. These can be genome-centric questions (extraction of full genomes from metagenomes and genomics-informed microorganism isolation), requiring long contigs, or gene-centric questions (determination of microbial community composition as well as functional capacity and comparisons of microbial communities from various environments), requiring high-quality contigs and an assembly with extensive coverage of the metagenomic dataset [115]. It is important to understand the assembler performance considering the range of available assemblers. An assembler needs to produce long contigs to allow an accurate analysis of full genes within a genomic perspective and to enable the reconstruction of single genomes. Likewise, a good assembler should resort to the least computational resources possible along with an intuitive interface to allow minimal effort and rapid processing during the assembly process [115]. Currently, there are several open-source metagenome assemblers, such as: Velvet [121]; MetaVelvet [122]; SPAdes [123]; metaSPAdes [124]; Ray Meta [125]; IDBA-UD [126]; MEGAHIT [127]; and Omega [128]. Besides the differences between these assemblers, none proved to provide consistently superior assemblies. Therefore, it is proposed to select an appropriate assembler to consider the available computational resources, scientific research question, and bioinformatics expertise of the user (Figure 4) [115].

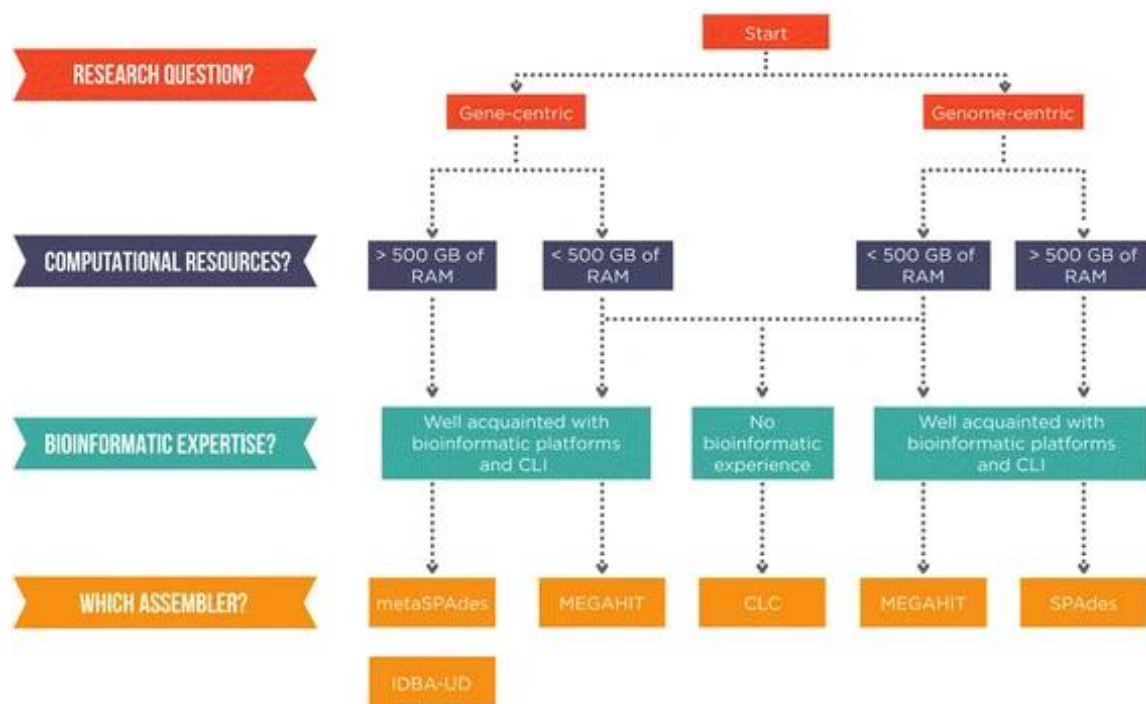


Figure 4. Proposed workflow to select a metagenome assembler based on the research question, the computational resources available, and the bioinformatics expertise of the researcher (adapted from [115]).

After assembling, the obtained contigs must be mapped against a resistance reference database. Besides assemblies, aligning reads directly to the reference databases can also be applied to detect ARGs from low-abundance bacteria in complex communities [129]. Additionally, there are bioinformatics tools for metagenomic data analysis that can be used for analyzing gene content and gene expression. There is, nevertheless, no unified standard analysis method, highlighting the importance of choosing an adequate approach to identify ARGs [130].

Table 1. Summary of antimicrobial resistance reference databases (adapted from [131]).

Database	Description
CARD [132]	Ontology-based database that provides comprehensive information on ARGs and their resistance mechanisms Currently contains >2,200 protein homologs and includes a curated set of resistance-conferring chromosomal mutations.
ResFinder [133]	Collection of acquired ARGs, frequently involved in HGT events
Resfams [134]	A profile HMM-based curated database confirmed for AR function
ARDB [135]	First centralized resource of ARGs information Manually curated; contains >4,500 ARGs sequences

MEGARes [136]	Collation of multiple databases (CARD, ARG-ANNOT, and ResFinder) to avoid redundancy between entries For high-throughput screening and statistical analysis
ARG-ANNOT [137]	Repository of >1,800 AR sequences collated from scientific literature and online resources It also includes point mutation data for select AR-associated chromosomal genes.
Mustard [138]	Resource containing 6,095 ARGs determinants from 20 families, including curated sets of AR genes, identified in functional metagenomics studies
FARME database [139]	A curated set of microbial sequences functionally screened to confer resistance in various functional metagenomics studies of different habitats.
SARG [140]	Hierarchically structured database derived from ARDB, CARD and NCBI-NR database It contains >12,000 AR genes; also includes profile HMMs for 189 ARGs subtypes.
Lahey list of $\beta$ -lactamases [141]	The first initiative to compile known $\beta$ -lactamases and assign nomenclature to novel ones
BLDB [142]	Manually curated database for AR enzymes classified by class, family, and subfamily
LacED [143]	A curated database of TEM and SHV $\beta$ -lactamases, including a curated set of known TEM and SHV variants
CBMAR [144]	Database that identifies and characterizes novel $\beta$ -lactamases based on Ambler classification

Numerous databases have been established to detect ARGs from metagenomic data (Table 1). As previously pointed out, selecting an adequate database is of great importance. Once chosen considering the research purpose, the core resistome design takes place. Very few bioinformatics tools are available for resistome determination. Nevertheless, some bioinformatics tools for statistical and exploratory analysis of resistome data have been

recently developed. For example, ResistoXplorer conjugates the current developments in statistics and visualization along with general functional annotations and phenotype library, allowing high-throughput analysis of joint outputs generated from metagenomic resistome data [145]. Similarly, sraX is a complete computerized analytical pipeline for precise resistome analysis, capable of inspecting numerous bacterial genomes to detect putative resistance factors [146]. Additionally, MetaCompare is a tool for ranking ARG risk through metagenomic data, estimating ARG applicants and retrieving a risk score to the respective resistome [147]. Moreover, further statistical analysis enables the possibility to study significant seasonal variation and the geographical impact of multiple bacterial communities and antibiotic resistomes [148]. Based on the previous tools for the resistome analysis, it is noticeable the lack of pipelines that are user-friendly and flexible when it comes to the resistome design and characteristics, such as ARGs annotation based on multiple databases, similarity, coverage and abundance, highlighting the need to develop new methods allowing to adjust and combine different parameters of the resistome.

## **2.8. Machine Learning**

Machine learning (ML) is a branch of the broader field of artificial intelligence that makes use of statistical models to develop predictions, using computational algorithms to shape empirical data into functional models [149]. Artificial Intelligence (AI) research, and breakthroughs in ML and deep learning (DL), have led to innovative advances in many research fields, such as radiology, pathology, genomics, etc. [150]. Based on the research purpose, the tradeoff between bias, variance and model complexity, the following types of ML models are debated as the central guide ideas of learning: neural network (feed-forward and recurrent), support vector machine, random forest, self-organizing map, and Bayesian network [151]. For example, ML/DL algorithms may be used to detect and quantify ARGs [152], and even track ARGs pollution from diverse sources [153].

The field of ML can be divided into supervised and unsupervised learning (Figure 5). Supervised algorithms refer to methods of implying a function from labelled training data;

unsupervised algorithms include methods that return a new set of features or patterns, strictly, from unlabelled data [154]. ML also includes a set of reinforcement learning algorithms and models where intelligent agents ought to take actions in an environment in order to maximize a cumulative reward.

The next sections present an overview of the most relevant ML concepts and nomenclature. It is not intended to be thorough (there are extensive resources available on the subject) but to provide the required basic notations to understand the work.

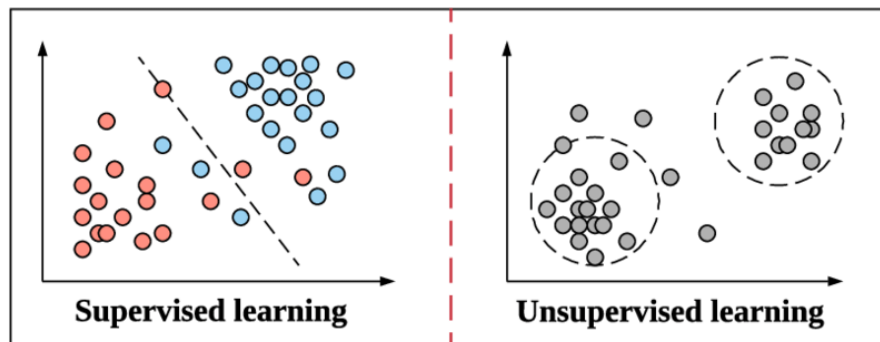


Figure 5. Examples of Supervised Learning (Linear Regression) and Unsupervised Learning (Clustering) [155].

### 2.8.1. Unsupervised Machine Learning

Unsupervised Learning algorithms were initially aimed to discover hidden patterns in unlabeled/unclassified sets of data. Since, and with more recent advances in DL, unsupervised methods have been employed to obtain compact data representations that capture underlying probabilistic distributions, these representations can then be used to distinguish patterns. Contrarily to supervised learning, unsupervised methods cannot be directly assigned to a regression or a classification problem [156]. The most commonly used unsupervised machine learning methods mainly include dimensionality reduction and clustering.

More recently, and with the advances of neural networks and deep learning, new unsupervised learning methods have been proposed such as autoencoders. Autoencoders compress data into a lower dimensional space, a latent space, and then recreate a new representation of the original data's input. They involve training two interlinked components, an encoder that compresses data, and a decoder that decompresses data (Figure 9).

### *Principal Component Analysis*

Dimensionality reduction of data can be achieved using Principal Component Analysis (PCA). PCA is a method that distinguishes data patterns while emphasizing similarities and differences. More precisely, PCA reduces the number of variables to the most significant factors providing an accurate summary of the original data through multivariate statistical data mining, while as many of the changes in the dataset are preserved for more efficient processing. With PCA, the sum of the squares of correlations is maximized, meaning that the first principal component vector has the highest sum of squares correlated with the variables, which is linearly related to the main variables [157].

### *Clustering*

Clustering is the most used unsupervised method and is applied to group objects by similarity-forming clusters. Clustering methods allow for finding hidden categories or patterns in the selected data. Some of the main algorithms for clustering include K-means, agglomerative clustering or hierarchical clustering [156].

#### **2.8.2. Supervised Machine Learning**

Supervised Learning deals with labelled data, that is, data where each sample has a corresponding signal or label. It aims to map, or model, the structure of an observation, or independent variable, into a label, a dependent variable. Once a model is learned, it can be used to predict the label of unseen observations. The nature of the label further defines the predictive task. Quantitative labels are more prone to regression tasks while qualitative labels are to classification. Examples of supervised learning methods include linear and logistic regressions, random forests, neural networks and support vector machines [158]. Some of the most frequently used supervised ML methods will be briefly described below.

### *Decision Trees (DT)*

DT consist of discrete classifiers, which can be used for both classification and regression tasks, these enable decision-making and risk analysis, and are usually represented

in the form of a graph, where the nodes represent the input variables and the branches the respective possible values, or even in the form of a list of rules. The DT architecture is simple to interpret and fast to learn through top-down algorithms [159].

### *Hidden Markov models (HMMs)*

HMMs are sequence models, where, given a succession of inputs an HMM will process a sequence of outputs with the same length. Visually, an HMM model is a graph: nodes are probability distributions over labels and edges return the probability of transitioning from the nodes [160].

### *Naïve Bayes (NB)*

Naïve Bayes classifier is a ML model that applies the Bayes theorem (Eq. 1), for probabilistic classification. By studying the input data of a given set of parameters (B), the NB classifier can calculate the likelihood of the input data belonging to a given class (A).

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)} \quad (\text{Eq. 1})$$

The classification of the input takes place when the probabilities of it belonging to each of the existing classes are determined and the highest probability will be the class to which the input data will fit. Thus, the class  $a$  with the highest probability must be found as expressed in Eq. 2 ( $b_i$  is one of the  $n$  features observed).

$$a = \text{argmax}_a P(a|b_1, \dots, b_n) \quad (\text{Eq. 2})$$

A NB classifier assumes that all variables are independent, requiring only a small set of training to estimate parameters necessary for classification [161].

### *Support Vector Machines (SVM)*

SVM is one of the most recent ML methods, with a major impact on neuroimaging analysis. In SVM, the input vector is mapped into a feature of higher dimensionality and the hyperplane that divides the data points into two classes is determined. The minimal distance between the decision hyperplane and the occurrences that are closest to the boundary is

maximized. The obtained classifier gains generalization capabilities and can be applied for consistent classification of other samples [162].

### *Bayesian Networks (BN)*

BN classifiers can produce probability estimations instead of predictions. BN is applied to represent knowledge paired with probabilistic dependencies including the variables of interest through a directed acyclic graph. These classifiers are being widely applied to several classification tasks and for knowledge representation as well as reasoning determinations [163].

#### **2.8.3. Feature Selection**

In ML algorithms, feature selection and feature extraction put together the so described dimensionality reduction. Feature selection is the term that describes the selection of a subset of features from a given set which enables the model construction. The three main methods to pick features consist of wrapper methods, filter methods and embedded methods. The wrapper methods optimize the classifier performance, although these are pointed as computationally more expensive due to repeated learning steps. Filter methods select properties of the features by statistically measuring or raking these, independently of the specific classifiers. Embedded methods are comparable to wrapper methods with the difference of making use of intrinsic model building metrics during the learning process [164].

#### **2.8.4. Error Estimation**

The generalized method to measure the performance of a learning algorithm correlates to its prediction ability on unbiased test data. Thus, meaning that the evaluation of the model performance is very important as it assists the choice of the learning method or model while providing a measure of the model quality [165].

## **2.9. Deep Learning (DL)**



Contrasting with shallow methods, DL methods have several layers of nonlinear segments of representation. These segments, individually, transform the representation at a given level into a further theoretical higher level. High-dimensional data with several degrees of freedom are a common outcome of these transformations, that allow to discover these sophisticated structures. Thus, it is quite promising that deep learning is outperforming the ML “shallow methods” in some areas. Though, DL only has significant advantages given the correct research purpose, as it is susceptible to overfitting [166]. In DL there are several architectures for a research purpose, the most common architectures used in DL will be briefly described below.

**2.9.1. Artificial Neural Networks (ANN)**

ANN is a piece of AI that is dedicated to imitating the learning approach that humans use to achieve specific types of understanding. Looking through a biological perspective, neurons, which are present in the brain, are used as artificial neurons in ANN, which are used to classify and hold data. ANN is made of input and output levels, as well as one or more hidden layers that transform the input into something that can be used in the output layer (Figure 6) [167].

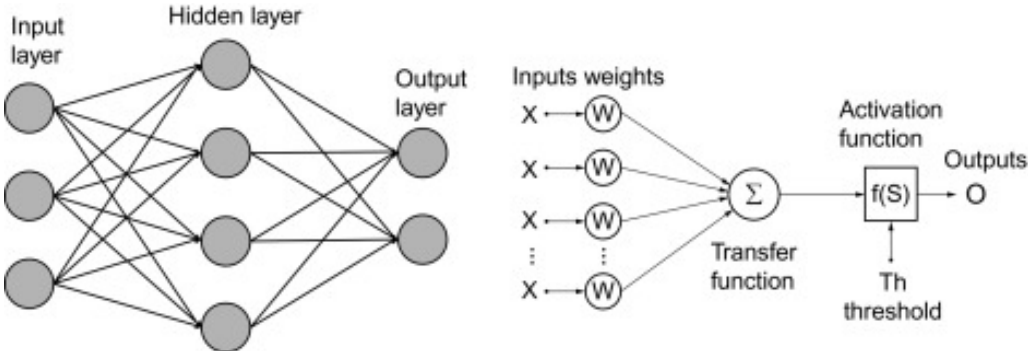


Figure 6. A typical ANN and a typical artificial neuron (from [167]).

In ANN the multiple artificial neuron nodes only have one type of link that connects the neurons. The input is kept by the neurons and simple operations are performed on the data, the information can be passed to other neurons, and this decision is determined by the activation function. Modulating feature extraction and classification. The dendrites/links vary in agreement with the significance of the inputs. ANN can be of many types, such as deep

neural networks, convolutional neural networks, stacked neural networks and recurrent neural networks [167].

### *Deep Neural Networks (DNN)*

DNN reached unmatched success in computer vision, their higher performance comes with the substantial cost of computational complexity and the possibility of overfitting caused by the usage of additional hidden layers to an ANN shallow approach. Consequently, methods that can boost the efficiency blockage while maintaining the high accuracy of DNN are in great need to enable several AI applications [168].

### *Convolutional Neural Networks (CNN)*

In CNN, a neuron is the consequence of multiple convolution tasks before getting triggered for feature extraction. A CNN has several steps of operation, which can be iterated numerous times resulting in a deep CNN. In the first stage, the convolution operation is completed, performing element-wise multiplication between the input (represented in a two-dimensional matrix) and the filter components (the filter is a two-dimensional matrix with a smaller size), where there may be a bias or weight filter. The sections of the weight filter are multiplied with the input and the portions of the bias filter are combined. In order to lower the dimension of the output, a pooling procedure is performed, generally max-pooling. Then a nonlinearity procedure is applied as an activation function. In the end, a fully connected layer can be applied for classification (Figure 7) [167].

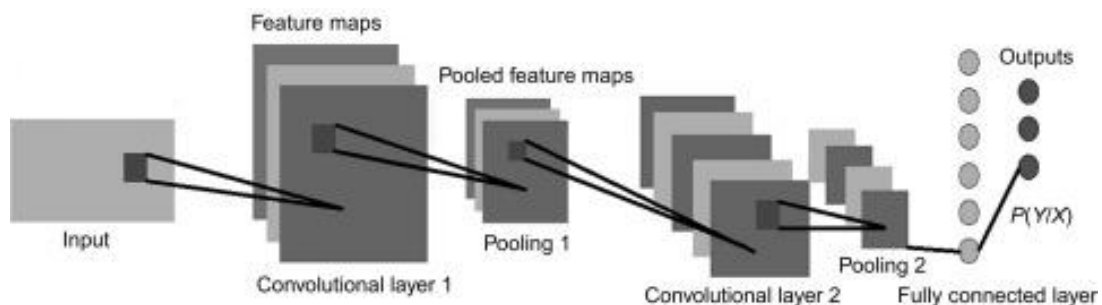


Figure 7. Convolutional neural network (from [167]).

*Recurrent Neural Networks (RNN)*

RNN make advantage of sequential information, theoretically, RNN can take advantage of data in randomly prolonged sequences, but in fact, they are limited to tracking back only a small number of moves (Figure 8) [167].

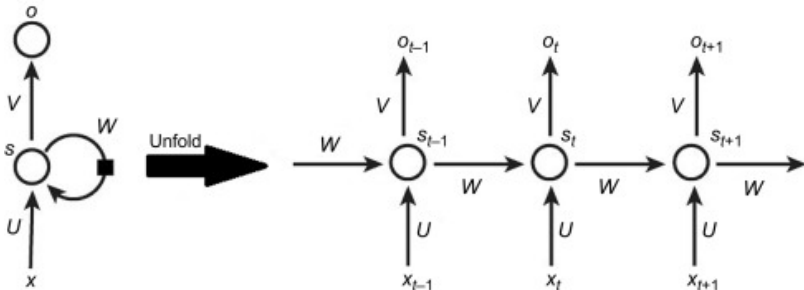


Figure 8. Recurrent neural network (from [167]).

*Stacked Autoencoders (SA)*

An autoencoder (AE) is a type of ANN used to learn useful information encrypting in an unsupervised conduct. An AE consists of two distinct parts: the encoder and the decoder (Figure 9). While the encoder is utilized to produce a reduced feature description from a starting input by a hidden layer, the decoder is applied to rebuild the original input from the encoder's output through the reduction of the loss function. Thus, as the AE transforms high-dimensional information to lower dimension, the AE is notably effective in noise removal, feature extraction and compression [169].

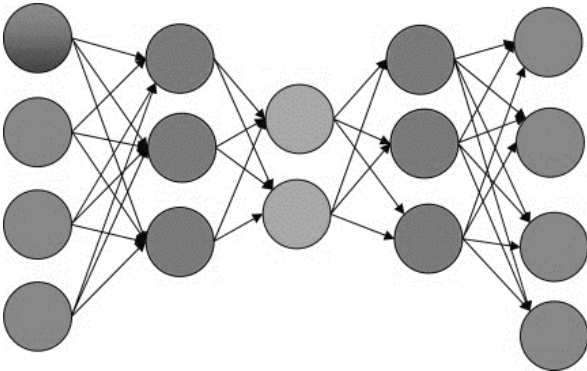


Figure 9. Example of an autoencoder (from [169]).

## 2.10. Machine Learning Applied to Metagenomic Surveillance

In the previous sections, the main ML and DL algorithms were introduced with the aim of choosing the most suited algorithms for building a multi label deep learning model, as it follows in the next chapters.

With the growing amount of ARGs being deposited in public databases, it is now possible to construct ML and DL models. Concerning metagenomic surveillance, particularly antibiotic resistance, ML/DL algorithms are commonly used to track ARGs pollution from different sources through classification tasks as well as predicting ARGs abundance [170]. On the other hand, DL algorithms can be also implemented as hierarchical multi-task for annotating antibiotic resistance genes in aminoacid format[171]. The combination between the lack of classification for gene potential for transferability, the absence of an intrinsic ARGs database and the ResFinder database being the only acquired ARGs database, makes room for a DL model for the classification of genes transferability.

## 3. Materials and Methods

This chapter firstly introduces the workflow from dataset construction to the definition of the core resistome and its statistical analysis using python. The endeavour comprised the exploration of different approaches to specific tasks.

Second, and related to DL, the chapter describes the actions taken from the dataset construction, model construction, training and validation for core resistome analysis. This task employed ProPythia, a platform for the classification of DNA using machine and deep learning (<https://github.com/BioSystemsUM/propythia>).

In order to run the following tools and command lines a complete Ubuntu terminal environment with Windows Subsystem for Linux (WSL) was used, Ubuntu 22.04 LTS from Canonical Group Limited, in a 16GB RAM computer with 8 logical processors and NVIDIA RTX 3050 GPU.

### 3.1. Metagenomic Dataset

A dataset was constructed from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (accessed on 25/07/2021), a public database for sequence data acquired from NGS platforms. Using the search terms “freshwater” and “wastewater”, 67391 samples were displayed. Regarding the research purpose, shotgun Illumina paired-end sequencing data with informative metadata that allowed us to classify the samples as raw wastewater (influent), treated wastewater (effluent), sludge and freshwater, from known geography were pre-selected (Table 2). After filtering, based on the previous criteria, if the resulting samples were from the same Bioproject, water type and location, only the sample with the highest number of reads were selected. Using those criteria, a dataset of 139 samples, 58 from freshwater and 81 from wastewater samples, ranging 24 different countries was obtained (Table S 1).

*Table 2. Filter set up for the construction of a freshwater and wastewater metagenomic dataset.*

Label	Filter
Assay_Type	WGS or OTHER
LibraryLayout	PAIRED
LibrarySelection	RANDOM, Other or unspecified
LibrarySource	Genomic, Metagenomic; Other
Organism	metagenome or wastewater metagenome; activated sludge metagenome
Platform	ILLUMINA

### **3.2. Reads Processing and Assembly**

Two case studies were developed: in the first (described below), the raw reads were processed and assembled with KBase apps and further aligned with the proper tools; in the second, the raw reads were processed under the selected k-mers aligner. This approach enables comparison between two distinct methods for conducting metagenomic surveillance.

The raw metagenomic reads were imported to a KBase narrative [172], with the “Import SRA File as Reads From Web” (v1.0.7) app, using the direct download link and Illumina parameters. The resulting merged reads were trimmed using the “Trim Reads with Trimmomatic” (v0.36) app [113], with sliding window size 4 and sliding window minimum quality 15. The processed reads were assembled into contigs with the “Assemble Reads with metaSPAdes” (v3.15.3) app [124], the contig length was set to  $300\text{bp} \leq 2000\text{bp}$ . Due to processing problems, one of the freshwater samples (SRR14120374) was not assembled. The raw merged reads were exported in FASTQ format and the assemblies in FASTA format.

### **3.3. Taxonomic Annotation**

For the taxonomic annotation of the samples, three different approaches were used and compared: annotation of the raw reads and annotation of the assembled reads (contigs) against the SILVA database, and a final method which enabled taxonomic classification throughout all taxa levels using the raw reads with the Kaiju app [173] at KBase.

To determine the taxa abundance within the metagenomic samples, the SILVA rRNA database [174] was used, after processing it with the SortMeRNA tool [175] (Figure 10). As a result, the obtained representative 16S rRNA database was used as the template database.

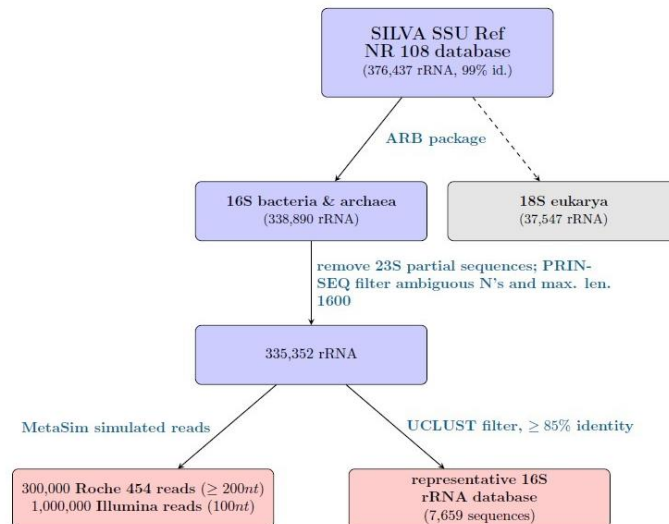


Figure 10. SILVA database processing for accurately quantifying 16s rRNA in metagenomic samples (from the SortMeRNA set construction [175]).

For the raw reads a k-mers aligner was used, namely, KMA [176], allowing to trim the reads, and match the k-mers between the query and the database, while simultaneously identifying regions with mismatches through the Needleman-Wunsch algorithm and using a scoring algorithm system named ConClave, to choose the best aligning template per query. The database was indexed through the following command line: “kma index -i set5-database.fasta -o silva\_db”, to align the reads against the database: “kma -i \*.fastq -o gs/silva/-t\_db silva\_db -t 8 -1t1 -mem\_mode -ef”, the flags: “-1t1” forces each query sequence to match to only one template; “-mem\_mode” KMA uses less memory as the ConClave algorithm is based on the mapping scores rather than alignment score; “-ef” stands for extended features, which creates an additional output file (\*.mapstat), enabling further analysis, such as the relative abundance of read hits with the database.

Similarly, for the contigs, the BLASTn [177] aligner was used as there was no need for mapping against the database, the representative 16S rRNA database was created using the command: “makeblastdb -in set5-database.fasta -dbtype nucl”, as for the alignment: “blastn -query \*.fa -db set5-database.fasta -outfmt 6 -out out/”, the flag “-outfmt 6” allows choosing the output format in this case the output 6 allows to create a tabular file, for a user-friendly approach when analysing the data, a threshold of 75%, 50% and 10% coverage was used, to verify the assembling process of the reads.

The metagenomic samples were also annotated in KBase with the “Classify Taxonomy of Metagenomic Reads with Kaiju” (v1.7.3), the classification was set to all taxonomic levels with the NCBI BLAST nr (no Euks) database the low abundance filter was set to 0.5 and subsample percent to 10.

While the two first approaches allow to quantify the bacterial load with the same database, the third approach allows to categorize the reads, accordingly, for the different taxonomic levels.

### **3.4. ARG Database Alignment**

The samples were aligned against two main ARG databases, the Comprehensive Antibiotic Resistance Database (CARD [133]) and the ResFinder database. Besides containing acquired genes like the ResFinder database, the CARD data gathers comprehensive information on ARGs and their resistance mechanisms consisting of protein homologs, including a curated set of resistance-conferring chromosomal mutations in protein-coding genes.

Two different approaches were explored, in “Case Study 1” the assembled reads were aligned with the embedded aligners in the ResFinder tool [133] and the Resistance Gene Identifier (RGI) from CARD [132], as for “Case Study 2” the raw reads were aligned with a k-mers aligner, KMA [176] and databases were indexed to the aligner.

#### **3.4.1. Case Study 1 (assembled reads)**

The assembled reads were firstly aligned with the ResFinder tool [133], from the Genomicepidemiology Bitbucket repository (<https://bitbucket.org/genomicepidemiology/resfinder/src/master/>), using the ResFinder database 7562716 version, under the following command line: “python3 run\_resfinder.py -o gs\_rf/ -s "Other" -l 0.6 -t 0.8 --acquired -ifa \*.fasta”, the default aligner for assemblies was used, namely, BLASTn [177]. The flags “-s” stand for the species, in this case, “Other” was used as it represents metagenomic samples or samples with unknown species, “-l” represents the desired minimum coverage and “-t” is the threshold for identity.



For the assembled reads, CARD Resistance Gene Identifier (RGI) tool was also used (<https://github.com/arpcard/rgi>). The command line used was: “rgi main -i \*.metaSPAdes.fa -o gs\_card/ -t contig -a DIAMOND --clean -d wgs --low\_quality --local”, the “--clean” tag removes temporary files, “-d” allows the user to choose the data type (wgs, plasmid, chromosome or NA), “--low\_quality” is used for short contigs to predict partial genes and the “--local” uses the local database in the executable directory. The input assemblies were in nucleotides and the template database of CARD is in aminoacids, for this case the DIAMOND [178] aligner was used (flag “-a”), as it is a sequence aligner for protein and translated DNA, allowing pairwise alignment of proteins and translated DNA at a much higher speed than BLAST and frameshift alignments for long read analysis. All the above features were built into the CARD RGI tool.

#### **3.4.2. Case Study 2 (raw reads)**

For comparison purposes, the CARD database, namely the file “nucleotide\_fasta\_protein\_homolog\_model.fasta”, and the compiled ResFinder database ([https://git@bitbucket.org/genomicepidemiology/resfinder\\_db.git](https://git@bitbucket.org/genomicepidemiology/resfinder_db.git) db\_resfinder) obtained with the cat command: “cat \*.fasta > rf\_db.fasta”, were individually indexed to KMA, using the following command line: “kma index -i \*.fasta -o database”.

The raw reads were then aligned to the reference databases using the following command line: “kma -i \*.fastq -o \*\_kma -t\_db -t 8 -1t1 -mem\_mode -ef”, the “-t\_db” allows the user to choose the template database, “-t” sets the number of threads, “-1t1” force each query sequence to match to only one template, “-mem\_mode” KMA uses less memory as the ConClave algorithm is carried out based on the mapping scores instead of the alignment scores, “-ef” creates an additional file (\*.mapstat) with extended features, holding additional information (KMA version, used database, number of fragments in the input, the date of the analysis and the command line), this file also allows to determine the relative abundance of reads that match the database.

### **3.5. Common Database: CARD & ResFinder**

As different databases may hold the same ARG sequence under distinct annotations (identification labels) it was necessary to uniformize the databases information in a common language. For this task, two approaches were developed, in the first case study, the concatenated ResFinder database was inputted into the CARD/RGI tool to annotate one database against the other, in the second case study both databases were concatenated into a single file and inputted to CD-HIT-est to create clusters of sequences based on similarity.

### **3.5.1. Case Study 1**

The concatenated ResFinder database was inputted to the CARD/RGI tool through the command line used before: “rgi main -i rf\_db.fasta -o common\_db -t contig -a DIAMOND --clean -d wgs --low\_quality --local”. The resulting file was filtered by a threshold of 90% minimum coverage and sequence identity. The file was then converted into a hashmap, where the key (CARD nomenclature) only matches a single and unique value (ResFinder nomenclature). A great number of genes between both databases did not match using this method.

### **3.5.2. Case Study 2**

The CARD database and the ResFinder database were concatenated using the “cat \*.fasta > rf\_db.fasta” command. The resulting database was inputted into the CD-HIT tool [179] available at <https://github.com/weizhongli/cdhit>. The program mode was set to CD-HIT-est, with a sequence identity threshold of 90%, and the following parameters: “-r No -G Yes -g Yes -b 20 -l 10 -s 0.0 -aL 0.0 -aS 0.0”, (-r: comparing both strands; -G: use global sequence identity; -g: sequence is clustered to the best cluster that meets the threshold; -b: bandwidth of alignment; -l: length of the sequence to skip; -s: minimal length similarity (fraction); -aL: minimal alignment coverage (fraction) for the longer sequence; -aS: minimal alignment coverage (fraction) for the shorter sequence).

The output returned 1278 clusters of sequences, with >90% similarity, in the combined database. Similarly, to the previous case study, the data was formatted into a hashmap, a dictionary containing a key with the cluster name and reference sequence, where the corresponding value is a list of genes belonging to the cluster key. This approach allowed to

uniform ARGs labels between different databases and further determination of consensus sequences, while also combining multiple ARGs variants in a common nomenclature.

### **3.6. Core Resistome Pipeline**

A Python pipeline was developed for the analysis of raw reads and contigs, focusing on the core resistome characterization and statistical analysis, with the aim of obtaining biomarker sequences for monitoring ARGs prevalence in the environment, the scripts can be consulted at <https://github.com/pg42866/Metagenomic-Analysis>.

The util.py file contains helper functions to assist in the taxonomic analysis, the construction of the merged database, to build and analyse the core resistomes, processing the biomarker sequences alignment and some other functions.

#### **3.6.1. Core Resistome**

The core\_resistome.py file allows to read the KMA output files from the ResFinder, CARD and SILVA alignments. The “make\_core” function allows to create core resistomes with adjustable prevalence percentage, in this case 90% of the samples. There is also a function (“soft\_full”) to analyze the differences between a full core (90% prevalence) and a soft core (75% prevalence). It also allows to create the Reunion and the Intersection annotations based on the CD-HIT-est database. If a given annotation was found in just one database (CARD or ResFinder) it was classified as Reunion annotation, if instead it was classified by both databases it was classified as Intersection annotation. The core resistomes can also be retrieved from these annotations.

Using the described methodology, were determined the core resistomes of the influent, sludge, effluent and freshwater samples, using the ResFinder annotation, the CARD annotation and as mentioned the Reunion annotation, for a 90% prevalence.

### 3.6.2. Statistical Analysis

The statistical.py and taxon.py files allow to conduct statistical analysis of the resistomes and taxonomic annotations. It is based on python libraries such as seaborn, sklearn, matplotlib and numpy. For the ARGs profiling using the KMA “-ef” tag, the function “get\_amr\_class” allows to create a dataframe of the most abundant AMR classes using the ResFinder annotation, from the ARGs to the phenotype nomenclature. The outputs are processed dataframes, PCA, k-means clustering, heatmaps, boxplots and charts for Elbow Method to find ideal k-means clusters.

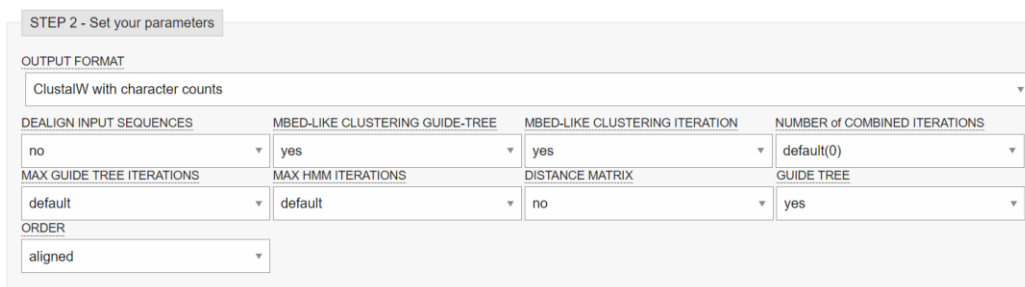
### 3.6.3. Identification of possible biomarkers and Consensus Sequences

The concatenated CARD+ResFinder database resulted in 1278 sequences clusters with a similarity of >90%. Of these 1278 clusters, 579 were clusters of a single sequence and 699 of 2 - 509 sequences.

Considering the core resistomes determined for wastewater (influent, sludge and effluent samples) and freshwater, were determined which genes are characteristic of wastewater. Those genes were classified as possible biomarkers to monitor freshwater contamination with wastewater. Due to the annotation being made with clusters, the cluster sequences of the core resistome were combined into consensus sequences.

For the ARGs observed to be part of the wastewater core resistome and simultaneously not part of the freshwater core resistome, and for which were observed clusters of more than one sequence (n=40) were determined the consensus sequences.

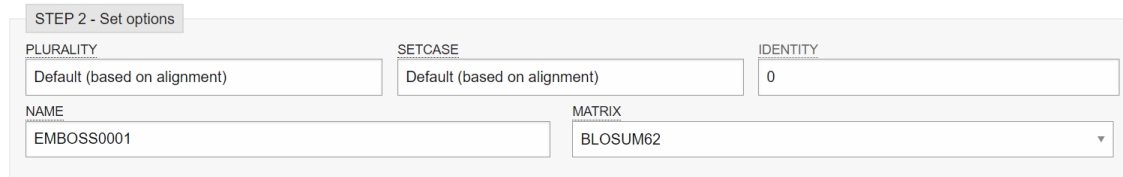
The sequences of each cluster were aligned with the EMBL-EBI Clustal Omega tool (v.1.2.4) [180], using the following parameters (Figure 11):



DEALIGN INPUT SEQUENCES	MBED-LIKE CLUSTERING GUIDE-TREE	MBED-LIKE CLUSTERING ITERATION	NUMBER of COMBINED ITERATIONS
no	yes	yes	default(0)
MAX GUIDE TREE ITERATIONS	MAX HMM ITERATIONS	DISTANCE MATRIX	GUIDE TREE
default	default	no	yes
ORDER			
aligned			

Figure 11. EMBL-EBI Clustal Omega parameters.

The resulting multiple alignment files were inputted to EMBL-EBI Emboss Cons [180] using the parameters below (Figure 12), to design the consensus sequence. To determine potential WWTP contamination in receiving environments, a set of 65 potential biomarker sequences, 25 single sequences and 40 consensus sequences, was obtained, and additionally 3 non-biomarker freshwater single sequences.



STEP 2 - Set options		
PLURALITY	SETCASE	IDENTITY
Default (based on alignment)	Default (based on alignment)	0
NAME	MATRIX	
EMBOSS0001	BLOSUM62	

Figure 12. EMBL-EBI Emboss Cons parameters.

#### 3.6.4. Biomarkers validation

In order to validate the biomarkers sequences, the 65 potential biomarkers were inputted as a reference database in KMA and aligned against the metagenomic samples for validation purposes. Similarly to the previously described methods using KMA. This approach allows to verify if there were variations in the detection sensitivity due to some of the biomarkers being representative consensus sequences of each core resistome cluster.

### 3.7. Classifying ARGs transferability through Deep Learning

A dataset of 2654 ARGs was constructed with two types of ARGs according to their transferability: 215 intrinsic genes (0) and 2439 acquired genes (1). The intrinsic genes were obtained from the HMD-ARG database [171], which is an aminoacid database with transferability classification, the aminoacid id from the intrinsic CARD ARGs was converted into the nucleotide version id returning 203 samples, 5 ARGs were obtained by using the search term “intrinsic” in the CARD database, the remaining 7 ARGs were obtained from the NCBI Genes database with the help of European Society of Clinical Microbiology and Infectious

Diseases (EUCAST) Expected Resistance Phenotypes (Version 1.1 March 2022). The 2439 acquired genes were obtained from the ResFinder database, as it is the only ARGs database fully made by acquired ARGs in DNA.

Before running ProPythia the following steps were made, the imbalanced dataset was submitted to a Synthetic Minority Oversampling Technique (SMOTE) [181] The resulting dataset was then used for the model training (60%), while maintaining the imbalanced dataset for validation (20%) and testing (20%).

The ProPythia pipeline allows to obtain models using descriptors or encoders. There are three types of encoders: One-hot encoding, single encoding for conversion of DNA alphabet into a binary vector; Chemical encoding, based on the chemical properties of the DNA combinations (ring structure, hydrogen bond and functional group); K-mer One-hot encoding, retrieving positional information of DNA sets, suited for larger sequences. Due to the length differences between the ARGs, a max length of 2000 bp was defined to prevent the encoders malfunction. Finally, to run ProPythia, the following config file was used (Figure 13). In the end, four models were obtained in combination with three modes: multilayer perceptron (MLP) with descriptors; CNN with one-hot encoders and chemical encoders; long short-term memory (LSTM) with one-hot encoders and chemical encoders; gated recurrent unit (GRU) with one-hot encoders and chemical encoders.

```

{
  "combination":{
    "model_label": ,
    "mode": ,
    "data_dir": "ARGs"
  },
  "do_tuning": false,
  "fixed_vals":{
    "epochs": 1000,
    "optimizer_label": "adam",
    "loss_function": "cross_entropy",
    "patience": 4,
    "output_size": 2,
    "cpus_per_trial": 2,
    "gpus_per_trial": 2,
    "num_samples": 10,
    "kmer_one_hot": 3,
    "dataset_file_format": "csv",
    "cutting_length": 2000,
    "save_to_pickle": true,
    "read_from_pickle": true
  },
  "hyperparameters": {
    "hidden_size": 32,
    "lr": 1e-3,
    "batch_size": 32,
    "dropout": 0.3,
    "num_layers": 1
  },
  "hyperparameter_search_space": {
    "hidden_size": [32, 64, 128],
    "lr": [1e-4, 1e-3, 1e-2],
    "batch_size": [16, 32, 64],
    "dropout": [0.2, 0.3, 0.4, 0.5],
    "num_layers": [1, 2, 3]
  }
}

```

Figure 13. ProPythia config.json file setup.

To deal with imbalanced data classes, special attention is given to the receiver operating characteristic curve (ROC). This analysis is particularly helpful due to its applicability when dealing with imbalanced class distributions as well as disproportionate classification error costs. The resulting metric from this analysis is obtained through the estimation of the area under the ROC curve (AUC), the AUC represents the probability of a randomly chosen positive occurrence will be rated higher than a negative occurrence [182]. To evaluate the DL models some other metrics were analyzed, namely, the classification accuracy, Matthews Correlation Coefficient (MCC), F1 score, precision and recall as well as the confusion matrix. The metrics were obtained from the sklearn Python library.

## 4. Results and Discussion

The following chapter describes and analyses the obtained results, since the dataset construction to the ARGs biomarkers validation and additionally the metrics of the DL algorithms used for the classification of ARGs transferability.

### 4.1. Samples Selection

As stated in the previous chapter, Table S 1 represents the selected samples based on the Table 2 filters. The geographic location of the samples encompasses 24 countries from distinct continents, socio-economic conditions and different sample types, with the purpose of reaching a global-scale metagenomic surveillance study.

*Table S 1. NCBI SRA sample selection for the construction of a metagenomic paired-end Illumina wastewater and freshwater dataset.*

<b>Accession n.</b>	<b>Sample type</b>	<b>Geographic Location</b>	<b>Reads n.</b>
SRR7614694	Influent	Antarctica	47870654
SRR10868593	Influent	Canada	13628676
SRR10868563	Influent	Canada	14669938
SRR10868583	Influent	Canada	16109110
SRR10868569	Influent	Canada	24671540
SRR14769839	Influent	China	55614748
SRR10688478	Influent	Czech Republic	2558208
SRR10688476	Influent	Czech Republic	2905240
SRR1616982	Influent	Germany	28295346
SRR8648012	Influent	Germany	28956220
SRR8648017	Influent	Germany	27404044
SRR11088457	Influent	Germany	44320206
SRR11088400	Influent	Germany	57139054
SRR11088463	Influent	Germany	58676406
SRR11088390	Influent	Germany	76746744
SRR8749025	Influent	Hong Kong	32907056
SRR8583493	Influent	Hong Kong	82968824
SRR14455375	Influent	Hong Kong	75416542
SRR8208344	Influent	Hong Kong	111696122
SRR8208343	Influent	Hong Kong	119850802
SRR10059215	Influent	India	29401180
SRR8749020	Influent	Philippines	29839314



SRR8944125	Influent	Portugal	8172260
SRR11567527	Influent	Puerto Rico	92853576
SRR10059214	Influent	Sweden	29569994
SRR8749021	Influent	Switzerland	31155086
SRR8573804	Influent	Uruguay	34630474
SRR8749023	Influent	USA	27889454
SRR13208900	Influent	USA	93321484
SRR13208907	Influent	USA	108997142
SRR15972293	Sludge	China	62539492
SRR11593520	Sludge	China	741557880
SRR8648015	Sludge	Germany	22764738
SRR8648014	Sludge	Germany	27371736
SRR8648013	Sludge	Germany	36551844
SRR11088394	Sludge	Germany	48022550
SRR11088425	Sludge	Germany	55797692
SRR11088423	Sludge	Germany	63918988
SRR11235434	Sludge	Germany	76346464
SRR11088424	Sludge	Germany	122539802
SRR11088415	Sludge	Germany	206564650
SRR8223441	Sludge	Hong Kong	82669882
SRR8205411	Sludge	Hong Kong	106889030
SRR8208348	Sludge	Hong Kong	101675980
SRR1544596	Sludge	Luxembourg	56920250
SRR9006530	Sludge	Luxembourg	59705684
SRR14610242	Sludge	Singapore	101381416
SRR9637883	Sludge	South Korea	39396822
SRR9637882	Sludge	South Korea	48162578
SRR9637884	Sludge	South Korea	93692160
SRR9827771	Sludge	USA	7064390
SRR16002673	Sludge	USA	49117576
SRR9827769	Sludge	USA	1467906
SRR13208905	Sludge	USA	91405164
SRR9827768	Sludge	USA	79498494
SRR9827758	Sludge	USA	96491226
SRR9827762	Sludge	USA	100101370
SRR9827761	Sludge	USA	106525534
SRR9827759	Sludge	USA	110151404
SRR8239393	Sludge	USA	538624998
SRR7638776	Effluent	China	11918058
SRR13287460	Effluent	China	86494806
SRR10688477	Effluent	Czech Republic	2854656
SRR10346178	Effluent	Germany	17726998
SRR1237782	Effluent	Germany	31723714
SRR8648016	Effluent	Germany	26212190
SRR8648011	Effluent	Germany	24443552
SRR11088403	Effluent	Germany	49003252

SRR11088420	Effluent	Germany	45787540
SRR11088367	Effluent	Germany	52685176
SRR11088386	Effluent	Germany	60064978
SRR8208349	Effluent	Hong Kong	91172100
SRR8584358	Effluent	India	31505234
SRR6158302	Effluent	Italy	20110778
SRR6158309	Effluent	Italy	19825978
SRR6158313	Effluent	Italy	23467288
DRR198516	Effluent	Japan	24927576
SRR11567528	Effluent	Puerto Rico	69412182
SRR8204324	Effluent	Spain	12766324
SRR13208893	Effluent	USA	88501942
SRR13208889	Effluent	USA	89023800
SRR10131203	Freshwater	Brazil	27024752
SRR12874346	Freshwater	Canada	40863560
SRR10868588	Freshwater	Canada	27593920
SRR8517159	Freshwater	Canada	62005698
SRR8517161	Freshwater	Canada	78978478
SRR12676972	Freshwater	Canada	128198488
SRR14576912	Freshwater	China	80424692
SRR10492803	Freshwater	China	69277724
SRR14368440	Freshwater	China	69789782
SRR14307622	Freshwater	China	94189640
SRR14576923	Freshwater	China	94668294
SRR14307628	Freshwater	China	97058800
SRR14307624	Freshwater	China	96303928
SRR10599111	Freshwater	China	101659050
SRR9924797	Freshwater	China	129695962
SRR9302965	Freshwater	China	392760680
SRR9302958	Freshwater	China	345656552
SRR9302961	Freshwater	China	426528358
SRR9302963	Freshwater	China	427445344
SRR9302960	Freshwater	China	403286044
SRR8894379	Freshwater	China	556564756
SRR14307626	Freshwater	China	95570234
SRR10688474	Freshwater	Czech Republic	2394602
SRR12274745	Freshwater	India	35988250
SRR12274752	Freshwater	India	35973620
SRR12274748	Freshwater	India	36807528
SRR12274739	Freshwater	India	37640628
SRR11700415	Freshwater	Nepal	90459524
SRR11567533	Freshwater	Puerto Rico	89732688
SRR12053438	Freshwater	Russia	195746714
SRR8561391	Freshwater	Russia	568594310
SRR13013687	Freshwater	Tanzania	98735066
SRR8436560	Freshwater	USA	11364226

SRR9289418	Freshwater	USA	18543892
SRR12197278	Freshwater	USA	18567990
SRR12197256	Freshwater	USA	20315966
SRR12197248	Freshwater	USA	22662822
SRR12197222	Freshwater	USA	23479332
SRR8075987	Freshwater	USA	37008680
SRR12197250	Freshwater	USA	30828778
SRR14371733	Freshwater	USA	76097452
SRR13490378	Freshwater	USA	62328892
SRR8075952	Freshwater	USA	46742322
SRR8075963	Freshwater	USA	52218732
SRR8075993	Freshwater	USA	39903310
SRR14240539	Freshwater	USA	40454146
SRR8075945	Freshwater	USA	83790046
SRR14240540	Freshwater	USA	51348758
SRR14240538	Freshwater	USA	51359896
SRR14240541	Freshwater	USA	51830844
SRR14240542	Freshwater	USA	57324576
SRR13434397	Freshwater	USA	213846214
SRR12261223	Freshwater	USA	189183200
SRR11472087	Freshwater	USA	188886590
SRR11555629	Freshwater	USA	201081816
SRR10520228	Freshwater	USA	195110656
SRR11557471	Freshwater	USA	298875516
SRR14120374	Freshwater	USA	76898234

## 4.2. Taxonomic Annotation

For the analysis of the raw reads, these were aligned using the KMA aligner against the SILVA database obtained from SortMeRNA. The matching 16S rRNA reads ratio with the total reads was displayed in the below sns.boxplot (Figure 14 a). It is noticeable that there is a lower bacterial diversity in the freshwater samples when comparing with the wastewater samples, which also have an increasing bacterial ratio from the WWTP start (influent) until the treated effluent.

To evaluate the assembly processing, the samples contigs were aligned using the previously 16S rRNA database and the BLASTn aligner with 80% similarity and 75%, 50% and 10% coverage, matching any database sequence with higher coverage than the selected threshold (Figure 14 b, c, d). By using 10% coverage threshold (corresponding to

approximately 150bp of the 1500bp of the 16S rRNA gene) some of the non-matching samples with the template database when using 75% or 50% coverage threshold have shown hits, as expected. This, may also be an indicator of data loss, meaning that some of the reads are not being properly assembled. In fact, that is a problem already described in the literature [183]. Being the 16S rRNA gene a gene with several repetitive regions the assembly may be challenging and result in several errors.

Comparing both alignments, the unassembled freshwater samples (Figure 4 a) that had the lower ratio average went to the second highest average in the assembled analysis (Figure 4 d). This result can be explained due to the lower complexity of the freshwater biome in comparison with the wastewater biome, indicating that with the higher environmental complexity the more prone the samples are to information loss in the assembly process.

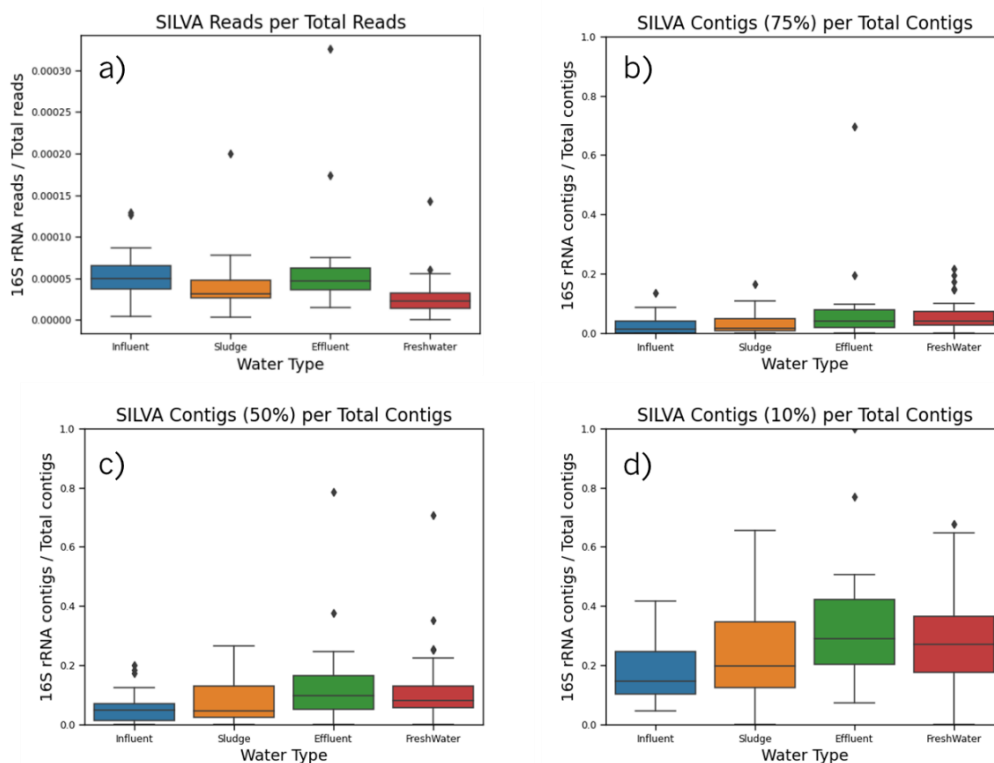


Figure 14. a) SILVA reads per total reads ratio; SILVA reads per total contigs ratio: b) with a 75% coverage threshold; c) with a 50% coverage threshold; d) with a 10% coverage threshold.

The PCA of taxonomic analysis with the relative abundance of Kaiju raw reads was conducted using the sklearn decomposition PCA library for Python. The PCA charts (Figure 15), of the previous dataset (Table S 1), do not show any evident clusters based on the water type across all taxonomic levels. As expected, with higher taxonomic specificity the more differentiated the samples get. However, also less reads will be classified, resulting for

example in what was observed for the order and family levels, where the taxonomy is more shared between the samples than in the higher taxonomic levels (phylum and class). These results show that there is no particular bacterial homology that can distinguish the samples by water type. It is noticeable that there is a wide variety of both unique combinations of bacteria and in some cases close similarity across all samples.

Given the outline of the taxonomic analysis, it is not expected to encounter a biomarker 16S rRNA gene. In fact, this result was more or less expected considering the close ecological proximity among the environments. For that reason, we tried to find other biomarkers, that will be explored in the next subchapter, with the characterization of the samples ARGs profile.

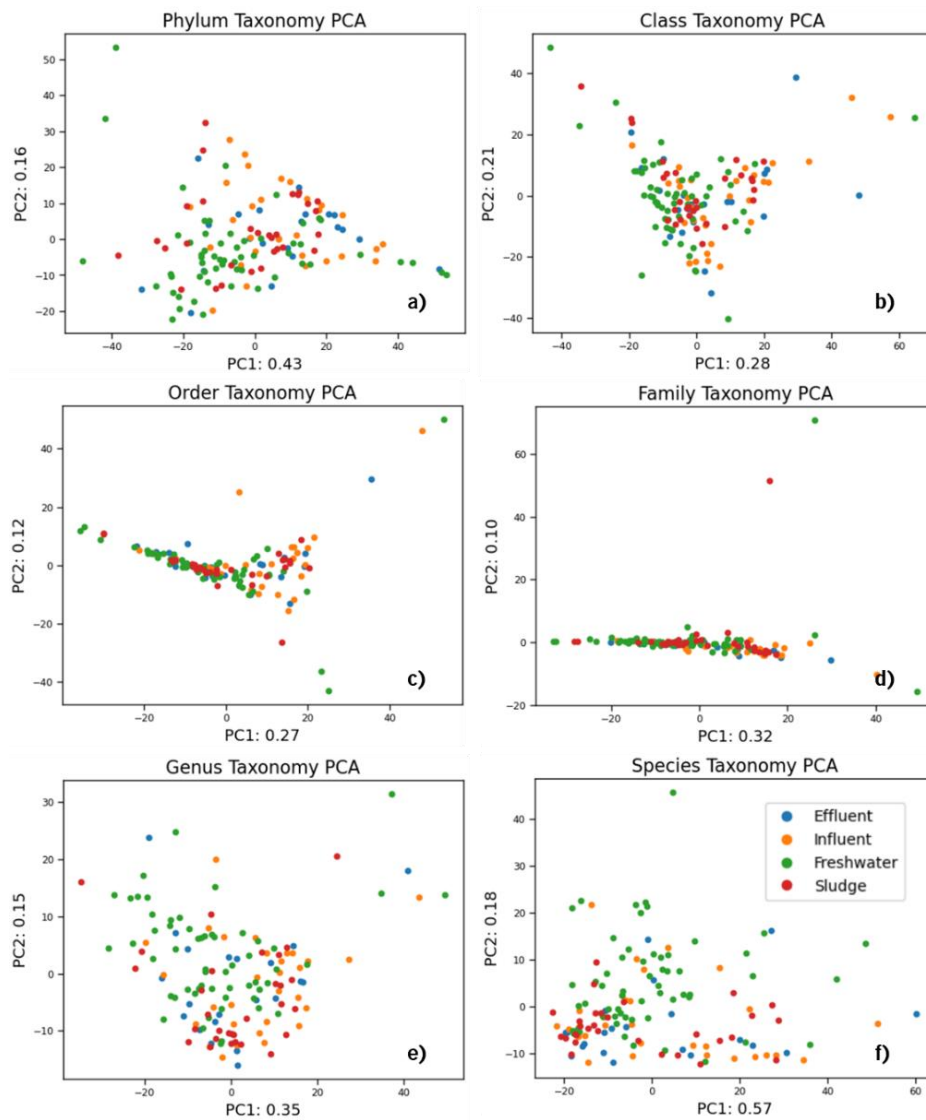


Figure 15. PCA of taxonomic analysis with the relative abundance of Kaiju reads: a) phylum; b) class; c) order; d) family; e) genus and f) species.

### **4.3. ARGs Profiling and core resistome definition**

The ResFinder database was inputted to the CARD/RGI tool with a threshold of 90% identity and 90% coverage. From the 3153 ARGs in ResFinder, 2917 were matched with ARGs from the CARD database, meaning that 236 ResFinder ARGs were unmatched along with 1717 CARD ARGs. In this first approach, the merged database resulted in a loss of 1953 ARGs. The second approach was based on creating ARGs clusters by similarity with a threshold of 90% through CD-HIT-est, guaranteeing that all the ARGs would be present in the merged database, individually or represented by a reference sequence. The result was a set of 1278 clusters, from which 579 were single sequence clusters.

The raw reads were inputted to the ResFinder tool, as it uses embedded KMA with the ResFinder default settings. This task was not performed in the CARD/RGI tool as it does not allow raw reads as inputs. The “Rf\_to\_CARD” column uses the first approach used for the merged databases and the Reunion column merges the annotations from the CARD and Rf\_to\_CARD columns. From the annotation using the CARD/RGI and ResFinder tool, from the original dataset, only 92 assembled samples had non-empty annotations from both tools, remaining: 28 influent, 28 sludge, 17 effluent and 18 freshwater. The highest number of ARGs for a given sample in the Reunion column was 329 ARGs and the lowest 5 ARGs. With an average of 63 ARGs per sample. The first annotations using the ResFinder and CARD/RGI tools, together with variant removal, resulted in the definition of the core resistome described in Table 3, considering a prevalence of 90% of the samples. The annotation was performed for the assembled contigs and for the raw reads (“ResFinder\_NP” (non-processed)).

At this stage, it was evident that information loss due to assembling was a problem not only during the taxonomic analysis but also showed major gaps in the ARGs annotation. As found elsewhere, the data loss in some cases mean that only 24.2-36.4% of reads are assembled in several metagenomic analyses [184].

Table 3. Core resistome (ARGs identified in >90% of the samples), using the ResFinder and CARD/RGI tools, using the assembled contigs and the raw reads, "Rf\_to\_CARD" uses the merged database annotation applied to the "Reunion" column.

Type of sample	ResFinder	CARD	Rf_to_CARD	Reunion	ResFinder_NP
	Assembled contigs				Raw reads
Influent	none	<i>adeF</i>	none	<i>adeF</i>	<i>msr(E), sul2, sul1, qacE, aph(6)-Id, erm(B), mph(E)</i>
Sludge	none	<i>adeF</i>	none	<i>adeF</i>	<i>sul2, sul1, qacE</i>
Effluent	none	<i>adeF</i>	none	<i>adeF</i>	<i>msr(E), sul2, sul1, qacE, mph(E), aph(6)-Id</i>
Freshwater	none	<i>adeF</i>	none	<i>adeF</i>	none

The raw reads were then aligned with KMA against the ResFinder Database using the extended features tag, to obtain the read hits relative abundance, allowing to profile the samples for acquired ARGs. From the raw reads annotation none of the samples had zero ARGs annotation, in contrast with the previous method where 44% of the assembled samples had no ARGs annotations.

The top 10 most abundant resistance phenotypes, according to ResFinder annotation are represented in Figure 16.

There was no considerable differences using the relative abundance of the 10 predominant AMR classes. Thus, the following ARGs analysis focused only on ARGs, using the KMA tool and raw reads (which are processed during the KMA run time) and the CD-HIT-est merged database. No variants were removed after the annotation process.

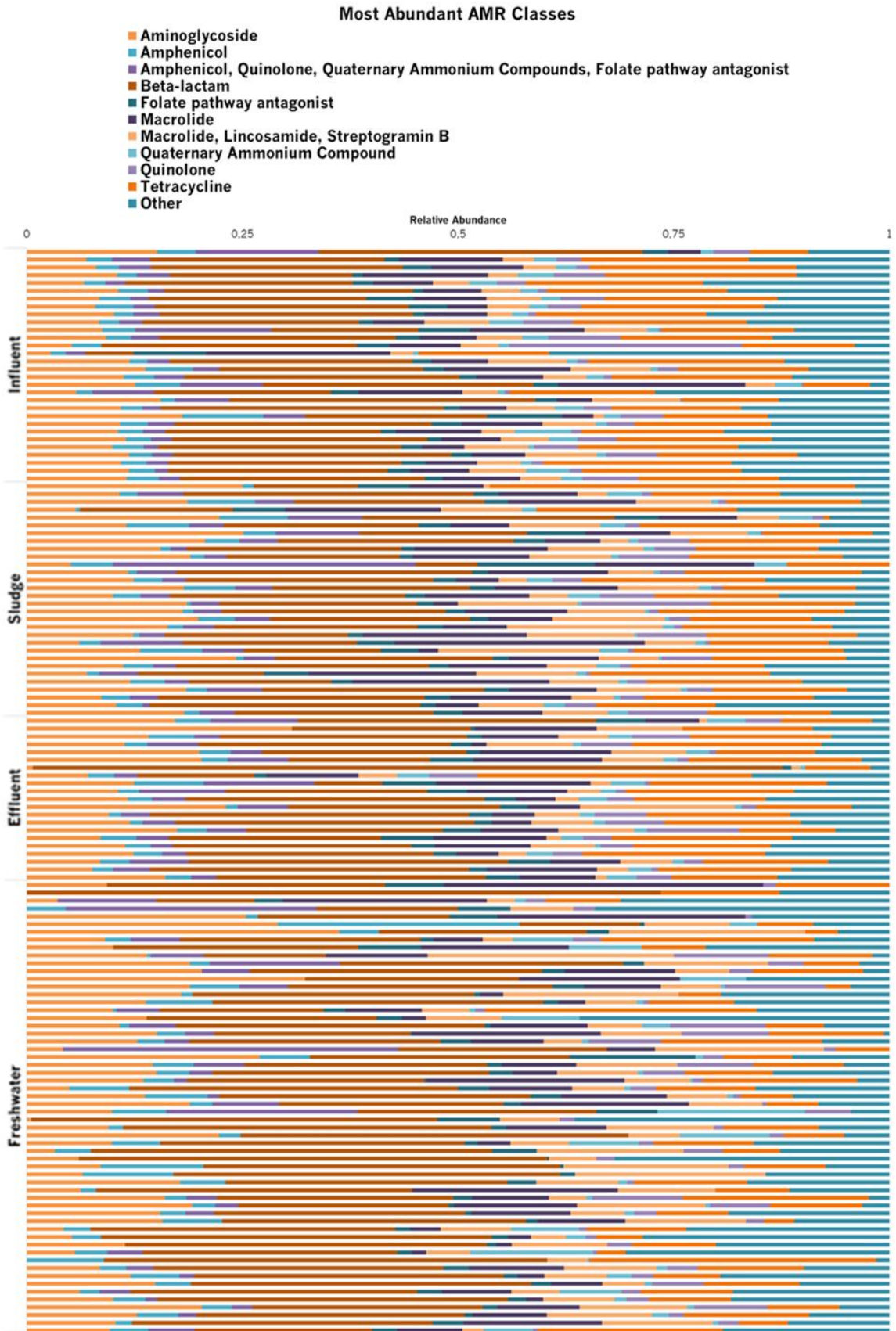


Figure 16. Relative abundance of the 10 most abundant AMR classes, using KMA “-ef” tag with the ResFinder database.



The boxplots below (Figure 17 and Figure 18), show that the ResFinder database detects fewer ARGs in comparison with the CARD database. The ARGs relative abundance decreases with the WWTP treatment along with the ARGs ratio per bacteria. The freshwater biome, which has the lowest 16S rRNA gene relative abundance (Figure 14 a), also has the lowest ARGs relative abundances considering the total number of reads of the sample or the bacteria abundance.

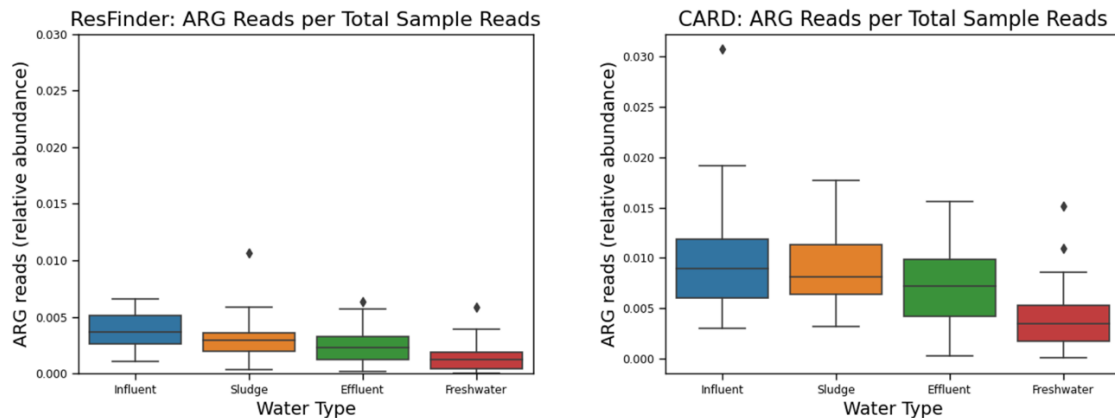


Figure 17. ARG reads relative abundance: on the left side using ResFinder database and on the right side using the CARD database.

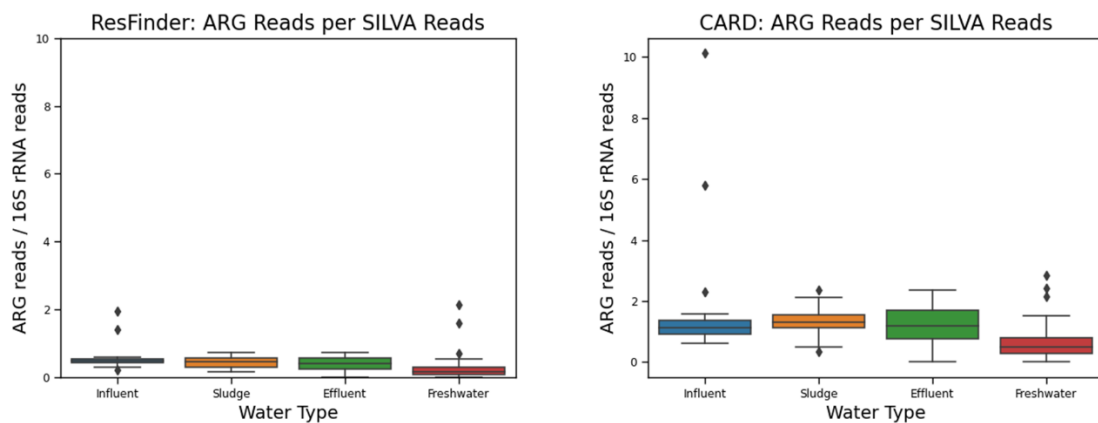


Figure 18. ARG reads per 16S rRNA reads: on the left side using ResFinder database and on the right side using the CARD database.

22 ARGs were identified in the resulting 90% prevalence core resistomes for each type of sample with the ResFinder database (Figure 19), from these: 14 in influent, 10 in sludge, 10 in effluent and 0 in freshwater. From these, 7 influent, 4 sludge and 3 effluent unshared ARGs.

In the 90% prevalence core resistome with the CARD database (Figure 20), 53 ARGs were found: 44 in influent, 29 in sludge, 31 in effluent and 3 in freshwater. As expected the untreated wastewater has the biggest core in comparison with the other water types. Only

the wastewater samples presented unshared ARGs, influent (17), sludge (3) and effluent (1), while all of the freshwater core ARGs were shared with all wastewater types.

Not surprisingly, the core resistome obtained with CARD or ResFinder annotations are proportional to the size of the respective databases. In both core resistomes, were identified shared ARGs across the wastewater samples (influent, sludge, effluent), 4 in the ResFinder core resistome and 20 in the CARD core resistome (Figure 19 and Figure 20).

Both annotations were converted using the CD-HIT-est database: the ResFinder resistome went from 22 core ARGs to 36 core ARGs; the CARD resistome went from 53 ARGs to 63 core ARGs. These results are due to the ARGs annotation conversion into the clusters reference ARGs.

Figure 21 shows the combined core resistome for both databases after annotation conversion accordingly to the CD-HIT-est database, the “Cluster Reunion Core” for 90% prevalence was constituted by 68 ARGs. A noticeable increase was observed in the core ARGs while preserving the unshared ARGs and the 3 shared ARGs across all water types (Cluster 701|-|ARO:3004480|Bado\_rpoB\_RIF, Cluster 702|+|ARO:3000501|rpoB2 and Cluster 718|-|ARO:3004074|MuxB).

The ARGs that are part of the core resistomes of wastewater samples (influent, sludge and/or effluent) but not of freshwater samples were considered as the main candidates for retrieving biomarker genes, as presented in the next subchapter.

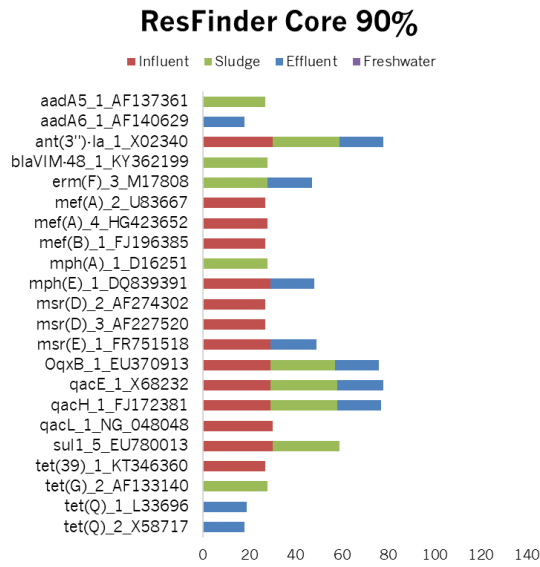


Figure 19. ResFinder core Resistome for 90% prevalence using KMA and raw reads.

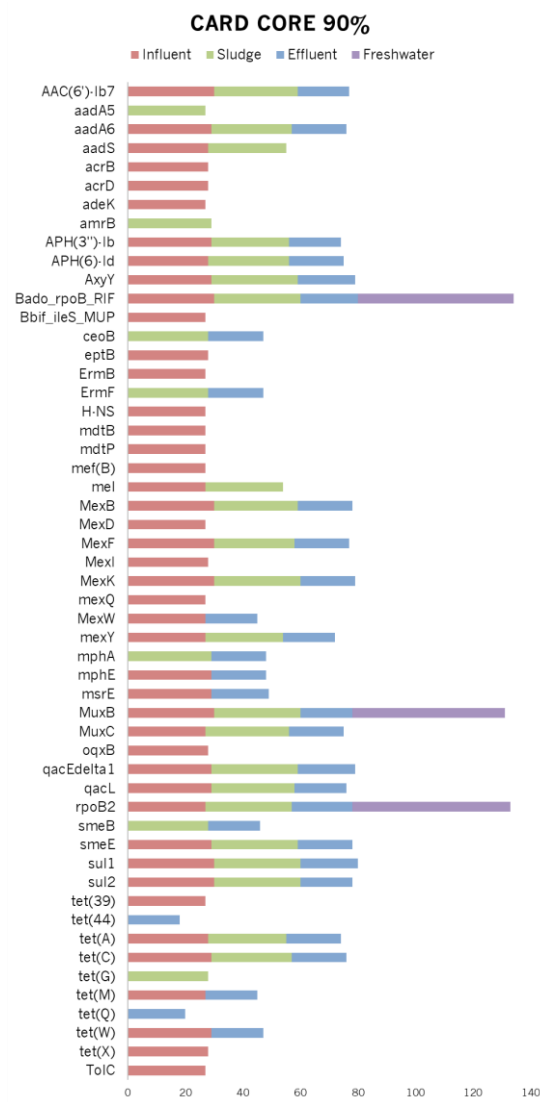


Figure 20. CARD core Resistome for 90% prevalence using KMA and raw reads.

## Cluster Reunion Core 90%

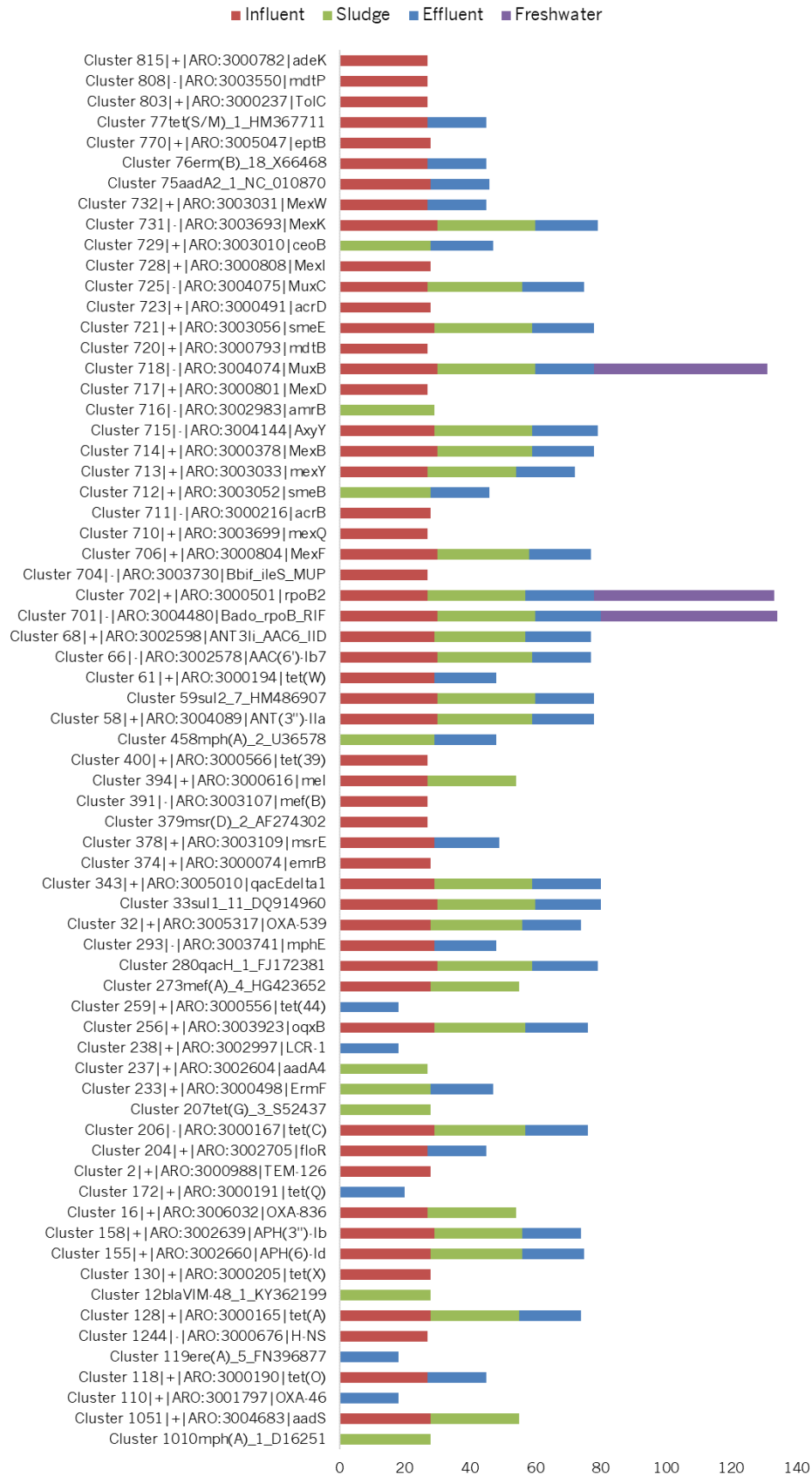


Figure 21. Cluster Reunion core Resistome for 90% prevalence using KMA and raw reads.

#### 4.4. Biomarker ARGs and Consensus Sequences

As described in the previous section 65 ARGs were identified as putative biomarkers, due to their presence in the core resistome of wastewater samples but not in freshwater samples. To increase the accuracy of this definition, a soft core (ARGs present in >75% of the samples) was also determined using the soft\_full function. The 4 ARGs initially classified as putative biomarkers but not present in the soft-core of the freshwater samples (Cluster 706|+|2810008-2813197|ARO:3000804|MexF, Cluster 715|-|23986-27124|ARO:3004144|AxyY, Cluster 725|-|2847775-2850886|ARO:3004075|MuxC, Cluster 731|-|4116187-4119265|ARO:3003693|MexK) were excluded from the list of possible biomarkers. Curiously, the excluded clusters were present in the core of all wastewater types.

From the 61 Cluster Reunion Core for 90% prevalence (Figure 21), 40 clusters were inputted to EMBL-EBI Emboss Cons obtaining cluster sequences, in combination with the remaining 21 single sequences, and a database with representative biomarker sequences was constructed.

As a validation, the biomarker database was used as a template database using KMA and the wastewater and freshwater samples were aligned against the biomarker database. Following the rationale that the genes should flow in the same direction of the water (influent > sludge > effluent > freshwater) the putative biomarkers for one type of water were searched in the other samples, expecting that for example a gene characteristic of freshwater should be also present in the upstream samples (influent, sludge and effluent). That analysis is represented in Figure 22 which shows the biomarkers found in wastewater aligned against all the water types in a heatmap. After this validation, the putative biomarker cluster12 is not indicated as a biomarker due to the loss of sensitivity as a result of the transition of the cluster to a consensus sequence. The remaining 60 potential biomarkers can be used to monitor ARGs contamination from WWTP to other receiving environments.

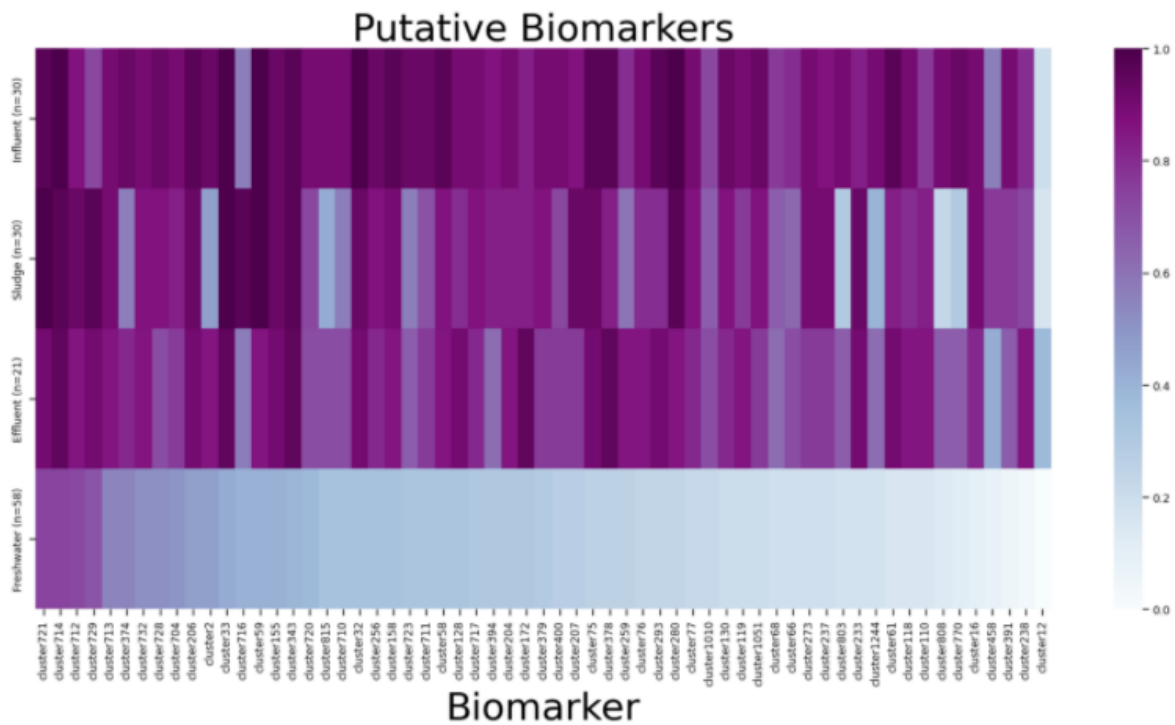


Figure 22. Heatmaps of the Water Types Biomarkers presence in the freshwater and wastewater dataset.

#### 4.5. ARGs Transferability (Deep Learning)

To develop a classifier able to distinguish ARGs transferability, between acquired and intrinsic, we trained and evaluated distinct DL models considering different encodings and descriptors. The intrinsic ARGs, as stated by EUCAST “Expected Resistant Phenotypes”, are the expected ARGs for a given species with an occurrence superior to 90%, which means that these ARGs, being expected, occur naturally. On the other hand, acquired ARGs must be closely monitored as new bacteria can develop AR by either genetic mutations or acquired ARGs. Thus, the correct distinction between the two ARG types is critical to monitorization and surveillance efforts.

As mentioned in chapter 3.7 the constructed dataset was inputted to Propyphia. The configuration settings were set as in Figure 13 The dataset consists of 2654 ARGs, according to their transferability the data divides in two labels: 215 intrinsic genes (0) and 2439 acquired genes (1). Given that the dataset is extremely unbalanced, we resorted to SMOTE to generate new synthetic samples based on the original samples (Figure 23).

Seven DL models were obtained through different conjugations of model type and descriptor or encoding type (Table 4). The training was composed by 60% of original dataset and the remaining with SMOTE synthetic samples. The test and validation datasets only used original ARGs samples, each was composed by 20% of the original dataset.

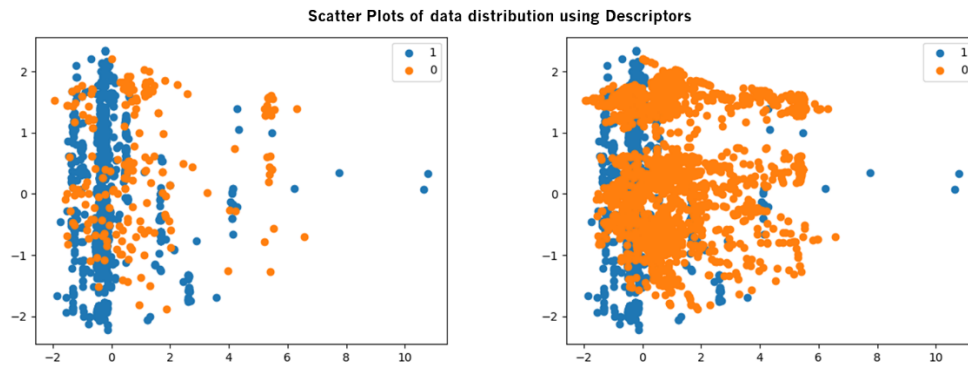


Figure 23. On the left is represented the scatter plot for the original dataset distribution using descriptors, on the right side is represented the dataset with the oversampling using SMOTE.

All the DL models tested shown promising results concerning ARGs transferability classification. The descriptor mode in combination with the multilayer perceptron (MLP) model, according to the confusion matrix, appears to have better results while using the synthetic samples for classifying the minority labels in the test dataset, although it is outperformed by the other models and modes combination when classifying the negative labels.

ProPythia, also had the option of k-mer one-hot encoding, due to the ARGs length it was considered that the one-hot encoding would be enough as it is the same as a k-mer one-hot encoder where  $k=1$ . The remaining models, CNN-LSTM and CNN-GRU, were not used as these result from combinations of the used models, and similar results would be expected to be obtained. No hyperparameter tuning was performed due to malfunction problems when using ProPythia, the default hyperparameters were used as shown in Figure 13.

Table 4. Reports of Deep learning models obtained through ProPythia: multilayer perceptron (MLP); convolutional neural networks (CNN); Long short-term memory (LSTM); gated recurrent unit (GRU).

	<b>DL Model</b>	<b>Acc.</b>	<b>MCC</b>	<b>ROC-AUC</b>	<b>F1</b>	<b>Precision</b>	<b>Recall</b>	<b>Conf. Matrix</b>
<b>MLP</b>	Descriptors	0.981	0.887	0.979	0.990	0.998	0.982	[42 1] [ 9 479]
<b>CNN</b>	One-hot encoding	0.983	0.881	0.906	0.991	0.984	0.998	[35 8] [ 1 487]
	Chemical encoding	0.987	0.909	0.929	0.993	0.988	0.998	[37 6] [ 1 487]
<b>LSTM</b>	One-hot encoding	0.976	0.837	0.923	0.987	0.988	0.986	[37 6] [ 7 481]
	Chemical encoding	0.987	0.909	0.929	0.993	0.988	0.998	[37 6] [ 1 487]
<b>GRU</b>	One-hot encoding	0.983	0.883	0.927	0.991	0.988	0.994	[37 6] [ 3 485]
	Chemical encoding	0.985	0.897	0.939	0.992	0.990	0.994	[38 5] [ 3 485]

To validate the models, stratified k-fold cross-validation was conducted (k=5), similarly to k-fold cross-validation, although it performs stratified sampling instead of random sampling. In Table 5 the mean values of the used metrics are displayed of the 5 folds per DL model. The same approach was used for the model construction as before, the training data was obtained through a conjugation of synthetic (SMOTE) and original data while the validation data and test data were exclusively generated with original data. As it is still considered an imbalanced dataset, the ROC-AUC metric is the most appropriate metric to evaluate the obtained models, in contrast with the conventional accuracy metric, which can lead to poor generalization results because the classifiers tend to predict the largest size class.

Table 5. Reports of the previously obtained models using Stratified kfold validation (k=5).

	<b>DL Model</b>	<b>Acc.</b>	<b>MCC</b>	<b>ROC-AUC</b>	<b>F1</b>	<b>Precision</b>	<b>Recall</b>
<b>MLP</b>	Descriptors	0.858	0.526	0.810	0.913	0.975	0.867
<b>CNN</b>	One-hot encoding	0.892	0.606	0.811	0.934	0.974	0.907
	Chemical encoding	0.888	0.586	0.796	0.933	0.970	0.906
<b>LSTM</b>	One-hot encoding	0.884	0.525	0.768	0.931	0.966	0.906
	Chemical encoding	0.885	0.514	0.755	0.932	0.964	0.910
<b>GRU</b>	One-hot encoding	0.886	0.544	0.778	0.933	0.968	0.907
	Chemical encoding	0.887	0.548	0.778	0.933	0.968	0.908

After assessing the Table 4 and Table 5 results, there are no significant differences between the usage of descriptors or encoders, though, in some cases the chemical encoding seems to be slightly outperformed by one-hot encoding, the small variation of results seems



to be related with the model type choice. Although, when looking at the stratified kfold validation table, the MLP and CNN (one-hot encoding) show the highest ROC-AUC, indicating that the classification consistency obtained in Table 4 is maintained when using smaller amounts of data and different data.

## 5. Conclusion and Future Work

As a consequence of the amount of data being deposited in public databases, metagenomic surveillance studies have thrived, as the need for a wet laboratory or culture dependent methods has been suppressed by the currently available tools, in combination with a wide range of bioinformatics tools also available. It is now possible to conduct metagenomic surveillance at a global scale with a suited dataset for a research purpose as well as an appropriate pipeline. Precisely, in this work, a metagenomic dataset was manually constructed, including wastewater and freshwater samples, from the NCBI SRA database and its analysis was conducted using available bioinformatics tools and a newly constructed Python pipeline.

On the other hand, there are still many challenges and efforts that must be surpassed, as shown in this study. First, to find good quality metagenomic samples requires a lot of manual curation and metadata validation. Second, the poor quality of the metadata available, impairs a good characterization of the samples. Third, the processing/assembly tools available do not seem to keep up with the complexity of some metagenomic environments leading to loss of information (although there are some tools that can handle raw data and deliver the expected results, such as KMA). Fourth, the growing amount of data deposited in public databases requires a major responsibility in the curation and organization of the data; different ARGs databases use different nomenclatures for the same ARGs under different accession numbers, which can differ if the sequence is in aminoacids and/or nucleotides.

In this work, some of the mentioned concerns were taken into consideration and the main focus was to solve and overcome them. The loss of ARGs information was prevented using raw reads and the KMA aligner, and, for the first time, it was possible to obtain a flexible and versatile core resistome, in a way that enabled multiple database analysis as well as their conjugation with adjustable ARGs prevalence in the core resistome, using the developed Python pipeline. The strategy of combining the CARD and ResFinder databases through ARGs similarity has led to major improvements regarding ARGs detection, the core resistomes of both annotations were extended with the clusters nomenclature and even combined resulting in a multidatabase core resistome. As far as we know, this approach was not yet described in

the literature and may bring an important contribution to the analysis of the resistome, not only from metagenomic samples but also for example for the characterization of the resistome of single strains.

As a result, a set of 60 biomarkers were obtained that, in the future, can be used in culture independent or *in silico* analysis for monitoring the presence and spread of ARGs from wastewater origin in multiple receiving environments. From the 68 core resistome sequences, special attention is given to the excluded freshwater and wastewater ARGs that are part of both core resistomes, these ARGs have shown to be present in the vast majority of samples throughout the complete urban water cycle and the in freshwater environments, therefore these are considered to be expected in both environments and thus not interpreted as biomarkers for AR in the receiving environments of WWTP. In the end, a set of 60 biomarkers was obtained that, in the future, can be used in culture independent or *in silico* analysis for monitoring the presence and spread of ARGs from wastewater origin in multiple receiving environments. This is particularly important nowadays, since treated wastewater is being pointed as one of the first solutions to overdue the water scarcity, being recommended their use for irrigation, namely of agricultural fields for the countries in a higher water stress. Figure 22

From the 7 DL models, the CNN (one-hot encoding) stands out as the most consistent model, after stratified kfold validation. Although, no hyperparameter tuning was performed the models showed promising results regarding ARGs transferability, even with an unbalanced dataset, it has been proven that with some other tools (SMOTE) it is possible to generate high quality models through their evaluation with the proper metrics, such as the ROC-AUC. Even though the metrics have considerably decreased, as a result of stratifying the data 5 times, the models still show robust results towards the classification of ARGs transferability.

For the future, there are some improvements and further steps that can be done, namely: i) the construction of an ARGs intrinsic database; ii) upgrades to the Python pipeline making it more user friendly and fully automated; iii) inclusion of samples from other origins to validate the biomarkers and try to reach a smaller number and more specific ARGs that improve the monitoring efficiency; iv) DL models optimization through hyperparameter tuning.

## 6. Bibliography

- [1] T. U. Berendonk *et al.*, “Tackling antibiotic resistance: the environmental framework,” *Nat. Rev. Microbiol.* 2015 135, vol. 13, no. 5, pp. 310–317, Mar. 2015, doi: 10.1038/nrmicro3439.
- [2] M. Pazda, J. Kumirska, P. Stepnowski, and E. Mulkiewicz, “Antibiotic resistance genes identified in wastewater treatment plant systems – A review,” *Sci. Total Environ.*, vol. 697, p. 134023, Dec. 2019, doi: 10.1016/j.scitotenv.2019.134023.
- [3] R. Pallares-Vega *et al.*, “Determinants of presence and removal of antibiotic resistance genes during WWTP treatment: A cross-sectional study,” *Water Res.*, vol. 161, pp. 319–328, Sep. 2019, doi: 10.1016/J.WATRES.2019.05.100.
- [4] N. Sims and B. Kasprzyk-Hordern, “Future perspectives of wastewater-based epidemiology: Monitoring infectious disease spread and resistance to the community level,” *Environ. Int.*, vol. 139, p. 105689, Jun. 2020, doi: 10.1016/J.ENVINT.2020.105689.
- [5] C. Lal Gupta, R. Kumar Tiwari, and E. Cytryn, “Platforms for elucidating antibiotic resistance in single genomes and complex metagenomes,” *Environ. Int.*, vol. 138, p. 105667, May 2020, doi: 10.1016/J.ENVINT.2020.105667.
- [6] J. Bengtsson-Palme *et al.*, “Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics,” *Sci. Total Environ.*, vol. 572, pp. 697–712, Dec. 2016, doi: 10.1016/J.SCITOTENV.2016.06.228.
- [7] D. A. Vallero, “Environmental Biotechnology: An Overview,” *Environ. Biotechnol.*, p. 31, Jan. 2016, doi: 10.1016/B978-0-12-407776-8.00001-3.
- [8] S. Shao, Y. Hu, J. Cheng, and Y. Chen, “Research progress on distribution, migration, transformation of antibiotics and antibiotic resistance genes (ARGs) in aquatic environment,” *Crit. Rev. Biotechnol.*, vol. 38, no. 8, pp. 1195–1208, Nov. 2018, doi: 10.1080/07388551.2018.1471038.
- [9] M. Hutchings, A. Truman, and B. Wilkinson, “Antibiotics: past, present and future,” *Curr. Opin. Microbiol.*, vol. 51, p. 72, Oct. 2019, doi: 10.1016/J.MIB.2019.10.008.
- [10] J. L. Martínez, T. M. Coque, and F. Baquero, “What is a resistance gene? Ranking risk in resistomes,” *Nat. Rev. Microbiol.* 2014 132, vol. 13, no. 2, pp. 116–123, Dec. 2014, doi: 10.1038/nrmicro3399.
- [11] J. L. Martínez, F. Baquero, and D. I. Andersson, “Predicting antibiotic resistance,” *Nat. Rev. Microbiol.* 2007 512, vol. 5, no. 12, pp. 958–965, Dec. 2007, doi: 10.1038/nrmicro1796.
- [12] J. M. Munita and C. A. Arias, “Mechanisms of Antibiotic Resistance,” *Virulence Mechanisms of Bacterial Pathogens*, vol. 6, no. FEB. ASM Press, Washington, DC, USA, pp. 481–511, Apr. 09, 2016, doi: 10.1128/9781555819286.ch17.
- [13] G. D. Wright, “Q&A: Antibiotic resistance: Where does it come from and what can we do about it?,” *BMC Biology*, vol. 8, no. 1. BioMed Central, pp. 1–6, Sep. 20, 2010, doi:

10.1186/1741-7007-8-123.

- [14] M. Rahman and S. D. Sarker, "Antimicrobial natural products," *Annual Reports in Medicinal Chemistry*, vol. 55. Academic Press, pp. 77–113, Jan. 01, 2020, doi: 10.1016/bs.armc.2020.06.001.
- [15] B. P. Bougnom and L. J. V. Piddock, "Wastewater for Urban Agriculture: A Significant Factor in Dissemination of Antibiotic Resistance," *Environ. Sci. Technol.*, vol. 51, no. 11, pp. 5863–5864, Jun. 2017, doi: 10.1021/acs.est.7b01852.
- [16] N. A. Lerminiaux and A. D. S. Cameron, "Horizontal transfer of antibiotic resistance genes in clinical environments," doi: 10.1139/cjm-2018-0275.
- [17] T. P. Van Boeckel *et al.*, "Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data," *Lancet Infect. Dis.*, vol. 14, no. 8, pp. 742–750, Aug. 2014, doi: 10.1016/S1473-3099(14)70780-7.
- [18] L. Riaz *et al.*, "Treatment technologies and management options of antibiotics and AMR/ARGs," *Antibiotics and Antimicrobial Resistance Genes in the Environment: Volume 1 in the Advances in Environmental Pollution Research Series*, vol. 1. Elsevier, pp. 369–393, Jan. 01, 2019, doi: 10.1016/B978-0-12-818882-8.00023-1.
- [19] T. U. Berendonk *et al.*, "Tackling antibiotic resistance: The environmental framework," *Nat. Rev. Microbiol.*, vol. 13, no. 5, pp. 310–317, 2015, doi: 10.1038/nrmicro3439.
- [20] J. Q. Su *et al.*, "Metagenomics of urban sewage identifies an extensively shared antibiotic resistome in China," *Microbiome*, vol. 5, no. 1, Jul. 2017, doi: 10.1186/s40168-017-0298-y.
- [21] A. Pruden *et al.*, "Management Options for Reducing the Release of Antibiotics and Antibiotic Resistance Genes to the Environment," *Environ. Health Perspect.*, vol. 121, no. 8, pp. 1–8, Aug. 2013, doi: 10.1289/EHP.1206446.
- [22] World Health Organization, "Antimicrobial Resistance Global Report on Surveillance," Geneva, 2014. Accessed: Jan. 04, 2022. [Online]. Available: <https://apps.who.int/iris/bitstream/handle/10665/112642/?sequence=1>.
- [23] I. Samanta and S. Bandyopadhyay, "Antimicrobial resistance: one health approach," *Antimicrob. Resist. Agric.*, pp. 367–370, Jan. 2020, doi: 10.1016/B978-0-12-815770-1.00032-8.
- [24] S. A. McEwen and P. J. Collignon, "Antimicrobial Resistance: a One Health Perspective," *Microbiol. Spectr.*, vol. 6, no. 2, Apr. 2018, doi: 10.1128/MICROBIOLSPEC.ARBA-0009-2017.
- [25] A. Miłobedzka *et al.*, "Monitoring antibiotic resistance genes in wastewater environments: The challenges of filling a gap in the One-Health cycle," *J. Hazard. Mater.*, vol. 424, p. 127407, 2021, doi: 10.1016/j.jhazmat.2021.127407.
- [26] C. X. Hiller, U. Hübner, S. Fajnorova, T. Schwartz, and J. E. Drewes, "Antibiotic microbial resistance (AMR) removal efficiencies by conventional and advanced wastewater treatment processes: A review," *Sci. Total Environ.*, vol. 685, pp. 596–608, Oct. 2019, doi: 10.1016/J.SCITOTENV.2019.05.315.

- [27] H. Grundmann, M. Aires-de-Sousa, J. Boyce, and E. Tiemersma, "Emergence and resurgence of meticillin-resistant *Staphylococcus aureus* as a public-health threat," *Lancet*, vol. 368, no. 9538, pp. 874–885, Sep. 2006, doi: 10.1016/S0140-6736(06)68853-3.
- [28] A. Fajardo *et al.*, "The Neglected Intrinsic Resistome of Bacterial Pathogens," *PLoS One*, vol. 3, no. 2, p. e1619, Feb. 2008, doi: 10.1371/JOURNAL.PONE.0001619.
- [29] J. P. Coleman and C. J. Smith, "Microbial Resistance," *Ref. Modul. Biomed. Sci.*, Jan. 2014, doi: 10.1016/B978-0-12-801238-3.05148-5.
- [30] D. J and D. D, "Origins and evolution of antibiotic resistance," *Microbiol. Mol. Biol. Rev.*, vol. 74, no. 3, pp. 9–16, 2010, doi: 10.1128/MMBR.00016-10.
- [31] J. Bengtsson-Palme, E. Kristiansson, and D. G. J. Larsson, "Environmental factors influencing the development and spread of antibiotic resistance," *FEMS Microbiol. Rev.*, vol. 42, no. 1, pp. 68–80, Jan. 2018, doi: 10.1093/FEMSRE/FUX053.
- [32] A. Zarei-Baygi and A. L. Smith, "Intracellular versus extracellular antibiotic resistance genes in the environment: Prevalence, horizontal transfer, and mitigation strategies," *Bioresour. Technol.*, vol. 319, Jan. 2021, doi: 10.1016/J.BIORTECH.2020.124181.
- [33] J. L. Martínez, "Bottlenecks in the transferability of antibiotic resistance from natural ecosystems to human bacterial pathogens," *Front. Microbiol.*, vol. 2, no. JAN, p. 265, 2012, doi: 10.3389/FMICB.2011.00265/BIBTEX.
- [34] J. Jutkina, C. Rutgersson, C. F. Flach, and D. G. Joakim Larsson, "An assay for determining minimal concentrations of antibiotics that drive horizontal transfer of resistance," *Sci. Total Environ.*, vol. 548–549, pp. 131–138, Apr. 2016, doi: 10.1016/J.SCITOTENV.2016.01.044.
- [35] C. S. Smillie, M. B. Smith, J. Friedman, O. X. Cordero, L. A. David, and E. J. Alm, "Ecology drives a global network of gene exchange connecting the human microbiome," *Nat. 2011 4807376*, vol. 480, no. 7376, pp. 241–244, Oct. 2011, doi: 10.1038/nature10571.
- [36] J. Wiedenbeck and F. M. Cohan, "Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches," *FEMS Microbiol. Rev.*, vol. 35, no. 5, pp. 957–976, Sep. 2011, doi: 10.1111/J.1574-6976.2011.00292.X.
- [37] M. N. Alekshun and S. B. Levy, "Molecular Mechanisms of Antibacterial Multidrug Resistance," *Cell*, vol. 128, no. 6, pp. 1037–1050, Mar. 2007, doi: 10.1016/J.CELL.2007.03.004.
- [38] N. Soler and P. Forterre, "Vesiduction: the fourth way of HGT," *Environ. Microbiol.*, vol. 22, no. 7, pp. 2457–2460, Jul. 2020, doi: 10.1111/1462-2920.15056.
- [39] J. Rodríguez-Beltrán, J. DelaFuente, R. León-Sampedro, R. C. MacLean, and Á. San Millán, "Beyond horizontal gene transfer: the role of plasmids in bacterial evolution," *Nat. Rev. Microbiol. 2021 196*, vol. 19, no. 6, pp. 347–359, Jan. 2021, doi: 10.1038/s41579-020-00497-1.
- [40] Y. Wang *et al.*, "Antiepileptic drug carbamazepine promotes horizontal transfer of plasmid-borne multi-antibiotic resistance genes within and across bacterial genera,"

- ISME J.* 2018 132, vol. 13, no. 2, pp. 509–522, Oct. 2018, doi: 10.1038/s41396-018-0275-x.
- [41] M. C. Dodd, “Potential impacts of disinfection processes on elimination and deactivation of antibiotic resistance genes during water and wastewater treatment,” *J. Environ. Monit.*, vol. 14, no. 7, pp. 1754–1771, Jun. 2012, doi: 10.1039/C2EM00006G.
- [42] I. Chen and D. Dubnau, “DNA uptake during bacterial transformation,” *Nat. Rev. Microbiol.* 2004 23, vol. 2, no. 3, pp. 241–249, Mar. 2004, doi: 10.1038/nrmicro844.
- [43] S. Domingues, K. M. Nielsen, and G. J. da Silva, “Various pathways leading to the acquisition of antibiotic resistance by natural transformation,” *Mob. Genet. Elements*, vol. 2, no. 6, p. 260, Nov. 2012, doi: 10.4161/MGE.23089.
- [44] J. L. Balcázar, “How do bacteriophages promote antibiotic resistance in the environment?,” *Clin. Microbiol. Infect.*, vol. 24, no. 5, pp. 447–449, May 2018, doi: 10.1016/J.CMI.2017.10.010.
- [45] S. Domingues and K. M. Nielsen, “Membrane vesicles and horizontal gene transfer in prokaryotes,” *Curr. Opin. Microbiol.*, vol. 38, pp. 16–21, Aug. 2017, doi: 10.1016/J.MIB.2017.03.012.
- [46] M. M. H. Ellabaan, C. Munck, A. Porse, L. Imamovic, and M. O. A. Sommer, “Forecasting the dissemination of antibiotic resistance genes across bacterial genomes,” *Nat. Commun.*, vol. 12, no. 1, pp. 1–10, 2021, doi: 10.1038/s41467-021-22757-1.
- [47] X. Jiang *et al.*, “Dissemination of antibiotic resistance genes from antibiotic producers to pathogens,” *Nat. Commun.*, vol. 8, pp. 1–7, 2017, doi: 10.1038/ncomms15784.
- [48] P. M. Bennett, “Plasmid encoded antibiotic resistance: Acquisition and transfer of antibiotic resistance genes in bacteria,” *Br. J. Pharmacol.*, vol. 153, no. SUPPL. 1, pp. 347–357, 2008, doi: 10.1038/sj.bjp.0707607.
- [49] S. Hernando-Amado, T. M. Coque, F. Baquero, and J. L. Martínez, “Defining and combating antibiotic resistance from One Health and Global Health perspectives,” *Nat. Microbiol.* 2019 49, vol. 4, no. 9, pp. 1432–1442, Aug. 2019, doi: 10.1038/s41564-019-0503-9.
- [50] D. G. Joakim Larsson and C.-F. Flach, “Antibiotic resistance in the environment,” doi: 10.1038/s41579-021-00649-x.
- [51] F. Schulz *et al.*, “Towards a balanced view of the bacterial tree of life,” *Microbiome*, vol. 5, no. 1, p. 140, Oct. 2017, doi: 10.1186/s40168-017-0360-9.
- [52] J. Guo, J. Li, H. Chen, P. L. Bond, and Z. Yuan, “Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements,” *Water Res.*, vol. 123, pp. 468–478, Oct. 2017, doi: 10.1016/J.WATRES.2017.07.002.
- [53] A. R. Varela, S. André, O. C. Nunes, and C. M. Manaia, “Insights into the relationship between antimicrobial residues and bacterial populations in a hospital-urban wastewater treatment plant system,” *Water Res.*, vol. 54, pp. 327–336, May 2014, doi: 10.1016/J.WATRES.2014.02.003.

- [54] I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nat. Rev. Genet.* 2012 134, vol. 13, no. 4, pp. 260–270, Mar. 2012, doi: 10.1038/nrg3182.
- [55] Y. Qiao *et al.*, "A Miniature On-Chip Methane Sensor Based on an Ultra-Low Loss Waveguide and a Micro-Ring Resonator Filter," *Micromachines* 2017, Vol. 8, Page 160, vol. 8, no. 5, p. 160, May 2017, doi: 10.3390/MI8050160.
- [56] J. L. Martinez, "Environmental pollution by antibiotics and by antibiotic resistance determinants," *Environ. Pollut.*, vol. 157, no. 11, pp. 2893–2902, Nov. 2009, doi: 10.1016/J.ENVPOL.2009.05.051.
- [57] A. H. Holmes *et al.*, "Understanding the mechanisms and drivers of antimicrobial resistance," *Lancet*, vol. 387, no. 10014, pp. 176–187, Jan. 2016, doi: 10.1016/S0140-6736(15)00473-0.
- [58] M. E. Levison and J. H. Levison, "Pharmacokinetics and Pharmacodynamics of Antibacterial Agents," *Infect. Dis. Clin. North Am.*, vol. 23, no. 4, pp. 791–815, Dec. 2009, doi: 10.1016/J.IDC.2009.06.008.
- [59] A. C. Singer, H. Shaw, V. Rhodes, and A. Hart, "Review of Antimicrobial Resistance in the Environment and Its Relevance to Environmental Regulators," *Front. Microbiol.*, vol. 7, no. NOV, Nov. 2016, doi: 10.3389/FMICB.2016.01728.
- [60] M. Qiao, G. G. Ying, A. C. Singer, and Y. G. Zhu, "Review of antibiotic resistance in China and its environment," *Environ. Int.*, vol. 110, pp. 160–172, Jan. 2018, doi: 10.1016/J.ENVINT.2017.10.016.
- [61] C. S. Lundborg and A. J. Tamhankar, "Antibiotic residues in the environment of South East Asia," *BMJ*, vol. 358, pp. 42–45, Sep. 2017, doi: 10.1136/BMJ.J2440.
- [62] H. K. Allen, J. Donato, H. H. Wang, K. A. Cloud-Hansen, J. Davies, and J. Handelsman, "Call of the wild: antibiotic resistance genes in natural environments," *Nat. Rev. Microbiol.* 2010 84, vol. 8, no. 4, pp. 251–259, Mar. 2010, doi: 10.1038/nrmicro2312.
- [63] N. Skandalis *et al.*, "Environmental Spread of Antibiotic Resistance," *Antibiot.* 2021, Vol. 10, Page 640, vol. 10, no. 6, p. 640, May 2021, doi: 10.3390/ANTIBIOTICS10060640.
- [64] F. Wang *et al.*, "Antibiotic resistance in the soil ecosystem: A One Health perspective," *Curr. Opin. Environ. Sci. Heal.*, vol. 20, p. 100230, Apr. 2021, doi: 10.1016/J.COESH.2021.100230.
- [65] J. M. TIEDJE *et al.*, "Antibiotic Resistance Genes in the Human-Impacted Environment: A One Health Perspective," *Pedosphere*, vol. 29, no. 3, pp. 273–282, Jun. 2019, doi: 10.1016/S1002-0160(18)60062-1.
- [66] L. Aristilde, A. Melis, and G. Sposito, "Inhibition of photosynthesis by a fluoroquinolone antibiotic," *Environ. Sci. Technol.*, vol. 44, no. 4, pp. 1444–1450, 2010, doi: 10.1021/ES902665N.
- [67] F. Z. Gao *et al.*, "Swine farming elevated the proliferation of Acinetobacter with the prevalence of antibiotic resistance genes in the groundwater," *Environ. Int.*, vol. 136, Mar. 2020, doi: 10.1016/J.ENVINT.2020.105484.



- [68] S. M. Hatosy and A. C. Martiny, "The Ocean as a Global Reservoir of Antibiotic Resistance Genes," *Appl. Environ. Microbiol.*, vol. 81, no. 21, p. 7593, 2015, doi: 10.1128/AEM.00736-15.
- [69] C. M. Manaia, "Assessing the Risk of Antibiotic Resistance Transmission from the Environment to Humans: Non-Direct Proportionality between Abundance and Risk," *Trends Microbiol.*, vol. 25, no. 3, pp. 173–181, Mar. 2017, doi: 10.1016/J.TIM.2016.11.014.
- [70] Y. Ben, C. Fu, M. Hu, L. Liu, M. H. Wong, and C. Zheng, "Human health risk assessment of antibiotic resistance associated with antibiotic residues in the environment: A review," *Environ. Res.*, vol. 169, pp. 483–493, Feb. 2019, doi: 10.1016/J.ENVRES.2018.11.040.
- [71] M. C. Moreno-Bondi, M. D. Marazuela, S. Herranz, and E. Rodriguez, "An overview of sample preparation procedures for LC-MS multiclass antibiotic determination in environmental and food samples," *Anal. Bioanal. Chem.*, vol. 395, no. 4, pp. 921–946, Oct. 2009, doi: 10.1007/S00216-009-2920-8.
- [72] I. Vaz-Moreira, O. C. Nunes, and C. M. Manaia, "Bacterial diversity and antibiotic resistance in water habitats: Searching the links with the human microbiome," *FEMS Microbiol. Rev.*, vol. 38, no. 4, pp. 761–778, 2014, doi: 10.1111/1574-6976.12062.
- [73] H. Bürgmann *et al.*, "Water and sanitation: an essential battlefront in the war on antimicrobial resistance," *FEMS Microbiol. Ecol.*, vol. 94, no. 9, Sep. 2018, doi: 10.1093/FEMSEC/FIY101.
- [74] J. Wang, L. Chu, L. Wojnárovits, and E. Takács, "Occurrence and fate of antibiotics, antibiotic resistant genes (ARGs) and antibiotic resistant bacteria (ARB) in municipal wastewater treatment plant: An overview," *Sci. Total Environ.*, vol. 744, p. 140997, 2020, doi: 10.1016/j.scitotenv.2020.140997.
- [75] R. Lood, G. Ertürk, and B. Mattiasson, "Revisiting antibiotic resistance spreading in wastewater treatment plants - Bacteriophages as a much neglected potential transmission vehicle," *Front. Microbiol.*, vol. 8, no. NOV, pp. 1–7, 2017, doi: 10.3389/fmicb.2017.02298.
- [76] S. Purnell, J. Ebdon, A. Buck, M. Tupper, and H. Taylor, "Bacteriophage removal in a full-scale membrane bioreactor (MBR) - Implications for wastewater reuse," *Water Res.*, vol. 73, pp. 109–117, Apr. 2015, doi: 10.1016/J.WATRES.2015.01.019.
- [77] Q. Jiang, M. Feng, C. Ye, and X. Yu, "Effects and relevant mechanisms of non-antibiotic factors on the horizontal transfer of antibiotic resistance genes in water environments: A review," *Sci. Total Environ.*, vol. 806, p. 150568, 2022, doi: 10.1016/j.scitotenv.2021.150568.
- [78] W. Calero-Cáceres *et al.*, "Sludge as a potential important source of antibiotic resistance genes in both the bacterial and bacteriophage fractions," *Environ. Sci. Technol.*, vol. 48, no. 13, pp. 7602–7611, Jul. 2014, doi: 10.1021/ES501851S.
- [79] C. M. Manaia, I. Vaz-Moreira, and O. C. Nunes, "Antibiotic Resistance in Waste Water and Surface Water and Human Health Implications," *Handb. Environ. Chem.*, vol. 20,

- pp. 173–212, 2011, doi: 10.1007/698\_2011\_118.
- [80] Y. K. Kim *et al.*, “The capacity of wastewater treatment plants drives bacterial community structure and its assembly,” *Sci. Reports* 2019 91, vol. 9, no. 1, pp. 1–9, Oct. 2019, doi: 10.1038/s41598-019-50952-0.
- [81] L. Yang, Q. Wen, Z. Chen, R. Duan, and P. Yang, “Impacts of advanced treatment processes on elimination of antibiotic resistance genes in a municipal wastewater treatment plant,” *Front. Environ. Sci. Eng.* 2019 133, vol. 13, no. 3, pp. 1–10, May 2019, doi: 10.1007/S11783-019-1116-5.
- [82] C. Ferreira, J. Abreu-Silva, and C. M. Manaia, “The balance between treatment efficiency and receptor quality determines wastewater impacts on the dissemination of antibiotic resistance,” *J. Hazard. Mater.*, vol. 434, p. 128933, Jul. 2022, doi: 10.1016/J.JHAZMAT.2022.128933.
- [83] I. Buriánková *et al.*, “Antibiotic resistance in wastewater and its impact on a receiving river: A case study of wwtp brno-modřice, czech republic,” *Water (Switzerland)*, vol. 13, no. 16, 2021, doi: 10.3390/w13162309.
- [84] E. Singer *et al.*, “High-resolution phylogenetic microbial community profiling,” *ISME J.*, vol. 10, no. 8, pp. 2020–2032, Aug. 2016, doi: 10.1038/ISMEJ.2015.249.
- [85] R. M. Zellweger *et al.*, “A current perspective on antimicrobial resistance in Southeast Asia,” *J. Antimicrob. Chemother.*, vol. 72, no. 11, pp. 2963–2972, Nov. 2017, doi: 10.1093/JAC/DKX260.
- [86] J. A. Port, A. C. Cullen, J. C. Wallace, M. N. Smith, and E. M. Faustman, “Metagenomic frameworks for monitoring antibiotic resistance in aquatic environments,” *Environ. Health Perspect.*, vol. 122, no. 3, pp. 222–228, 2014, doi: 10.1289/EHP.1307009.
- [87] E. W. Rice, P. Wang, A. L. Smith, and L. B. Stadler, “Determining Hosts of Antibiotic Resistance Genes: A Review of Methodological Advances,” *Environ. Sci. Technol. Lett.*, vol. 7, no. 5, pp. 282–291, May 2020, doi: 10.1021/acs.estlett.0c00202.
- [88] S. Ishii, “Quantification of antibiotic resistance genes for environmental monitoring: Current methods and future directions,” *Curr. Opin. Environ. Sci. Heal.*, vol. 16, pp. 47–53, Aug. 2020, doi: 10.1016/J.COESH.2020.02.004.
- [89] A. Q. Nguyen *et al.*, “Monitoring antibiotic resistance genes in wastewater treatment: Current strategies and future challenges,” *Sci. Total Environ.*, vol. 783, p. 146964, Aug. 2021, doi: 10.1016/J.SCITOTENV.2021.146964.
- [90] T. J. Sharpton, “An introduction to the analysis of shotgun metagenomic data,” *Front. Plant Sci.*, vol. 5, no. JUN, p. 209, Jun. 2014, doi: 10.3389/FPLS.2014.00209/BIBTEX.
- [91] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil, “Opportunities and challenges in long-read sequencing data analysis,” *Genome Biol.* 2020 211, vol. 21, no. 1, pp. 1–16, Feb. 2020, doi: 10.1186/S13059-020-1935-5.
- [92] V. Pennone, J. F. Cobo-Díaz, M. Prieto, and A. Alvarez-Ordóñez, “Application of genomics and metagenomics to improve food safety based on an enhanced characterisation of antimicrobial resistance,” *Curr. Opin. Food Sci.*, vol. 43, pp. 183–188,

- Feb. 2022, doi: 10.1016/J.COFS.2021.12.002.
- [93] S. Yadav and A. Kapley, "Antibiotic resistance: Global health crisis and metagenomics," *Biotechnol. Reports*, vol. 29, p. e00604, Mar. 2021, doi: 10.1016/J.BTRE.2021.E00604.
- [94] E. M. H. Wellington *et al.*, "The role of the natural environment in the emergence of antibiotic resistance in Gram-negative bacteria," *Lancet Infect. Dis.*, vol. 13, no. 2, pp. 155–165, Feb. 2013, doi: 10.1016/S1473-3099(12)70317-1.
- [95] G. A. Arango-Argoty, D. Dai, A. Pruden, P. Vikesland, L. S. Heath, and L. Zhang, "NanoARG: A web service for detecting and contextualizing antimicrobial resistance genes from nanopore-derived metagenomes," *Microbiome*, vol. 7, no. 1, pp. 1–18, Jun. 2019, doi: 10.1186/S40168-019-0703-9/TABLES/4.
- [96] R. Padmanabhan, A. K. Mishra, D. Raoult, and P. E. Fournier, "Genomics and metagenomics in medical microbiology," *J. Microbiol. Methods*, vol. 95, no. 3, pp. 415–424, Dec. 2013, doi: 10.1016/J.MIMET.2013.10.006.
- [97] T. S. Crofts, A. J. Gasparri, and G. Dantas, "Next-generation approaches to understand and combat the antibiotic resistome," *Nat. Rev. Microbiol.* 2017 157, vol. 15, no. 7, pp. 422–434, Apr. 2017, doi: 10.1038/nrmicro.2017.28.
- [98] L. G. Li, Q. Huang, X. Yin, and T. Zhang, "Source tracking of antibiotic resistance genes in the environment — Challenges, progress, and prospects," *Water Res.*, vol. 185, p. 116127, Oct. 2020, doi: 10.1016/J.WATRES.2020.116127.
- [99] J. A. Perry, E. L. Westman, and G. D. Wright, "The antibiotic resistome: what's new?," *Curr. Opin. Microbiol.*, vol. 21, pp. 45–50, Oct. 2014, doi: 10.1016/J.MIB.2014.09.002.
- [100] S. Ebmeyer, E. Kristiansson, and D. G. J. Larsson, "A framework for identifying the recent origins of mobile antibiotic resistance genes," *Commun. Biol.*, vol. 4, no. 1, pp. 1–10, 2021, doi: 10.1038/s42003-020-01545-5.
- [101] R. I. Aminov, "The role of antibiotics and antibiotic resistance in nature," *Environ. Microbiol.*, vol. 11, no. 12, pp. 2970–2988, Dec. 2009, doi: 10.1111/J.1462-2920.2009.01972.X.
- [102] H. Waseem, M. R. Williams, R. D. Stedtfeld, and S. A. Hashsham, "Antimicrobial Resistance in the Environment," *Water Environ. Res.*, vol. 89, no. 10, pp. 921–941, Oct. 2017, doi: 10.2175/106143017X15023776270179.
- [103] L. Ma, B. Li, and T. Zhang, "New insights into antibiotic resistome in drinking water and management perspectives: A metagenomic based study of small-sized microbes," *Water Res.*, vol. 152, pp. 191–201, Apr. 2019, doi: 10.1016/J.WATRES.2018.12.069.
- [104] A. A. Salyers, A. Gupta, and Y. Wang, "Human intestinal bacteria as reservoirs for antibiotic resistance genes," *Trends Microbiol.*, vol. 12, no. 9, pp. 412–416, Sep. 2004, doi: 10.1016/J.TIM.2004.07.004.
- [105] J. M. Rolain, "Food and human gut as reservoirs of transferable antibiotic resistance encoding genes," *Front. Microbiol.*, vol. 4, no. JUN, p. 173, 2013, doi: 10.3389/FMICB.2013.00173/BIBTEX.
- [106] X. Yin *et al.*, "An assessment of resistome and mobilome in wastewater treatment

- plants through temporal and spatial metagenomic analysis," *Water Res.*, vol. 209, p. 117885, Feb. 2022, doi: 10.1016/J.WATRES.2021.117885.
- [107] E. A. Rodríguez, D. Ramirez, J. L. Balcázar, and J. N. Jiménez, "Metagenomic analysis of urban wastewater resistome and mobilome: A support for antimicrobial resistance surveillance in an endemic country," *Environ. Pollut.*, vol. 276, p. 116736, May 2021, doi: 10.1016/J.ENVPOL.2021.116736.
- [108] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from sampling to analysis," *Nat. Biotechnol.* 2017 359, vol. 35, no. 9, pp. 833–844, Sep. 2017, doi: 10.1038/nbt.3935.
- [109] Y. Ma, H. Xie, X. Han, D. M. Irwin, and Y. P. Zhang, "QcReads: An Adapter and Quality Trimming Tool for Next-Generation Sequencing Reads," *J. Genet. Genomics*, vol. 40, no. 12, pp. 639–642, Dec. 2013, doi: 10.1016/J.JGG.2013.11.001.
- [110] S. Lindgreen, "AdapterRemoval: Easy cleaning of next-generation sequencing reads," *BMC Res. Notes*, vol. 5, 2012, doi: 10.1186/1756-0500-5-337.
- [111] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, pp. 10–12, May 2011, doi: 10.14806/EJ.17.1.200.
- [112] Y. Kong, "Btrim: A fast, lightweight adapter and quality trimming program for next-generation sequencing technologies," *Genomics*, vol. 98, no. 2, pp. 152–153, Aug. 2011, doi: 10.1016/J.YGENO.2011.05.009.
- [113] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/BIOINFORMATICS/BTU170.
- [114] M. I. J. Maran and D. J. Davis G., "Benefits of merging paired-end reads before pre-processing environmental metagenomics data," *Mar. Genomics*, vol. 61, p. 100914, Feb. 2022, doi: 10.1016/j.margen.2021.100914.
- [115] A. J. van der Walt, M. W. van Goethem, J. B. Ramond, T. P. Makhalanyane, O. Reva, and D. A. Cowan, "Assembling metagenomes, one community at a time," *BMC Genomics*, vol. 18, no. 1, pp. 1–13, Jul. 2017, doi: 10.1186/s12864-017-3918-9.
- [116] N. Nagarajan and M. Pop, "Sequence assembly demystified," *Nat. Rev. Genet.*, vol. 14, no. 3, pp. 157–167, Mar. 2013, doi: 10.1038/NRG3367.
- [117] M. Boolchandani, A. W. D, G. Dantas, and A. M. contributions, "Sequencing-based methods and resources to study antimicrobial resistance HHS Public Access," doi: 10.1038/s41576-019-0108-4.
- [118] K. Anantharaman *et al.*, "Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system," *Nat. Commun.* 2016 71, vol. 7, no. 1, pp. 1–11, Oct. 2016, doi: 10.1038/ncomms13219.
- [119] P. E. C. Campeau, P. A. Pevzner, and G. Tesler, "How to apply de Bruijn graphs to genome assembly," *Nat. Biotechnol.*, vol. 29, no. 11, pp. 987–991, Nov. 2011, doi: 10.1038/NBT.2023.
- [120] N. J. Loman *et al.*, "A culture-independent sequence-based metagenomics approach to

- the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4,” *JAMA*, vol. 309, no. 14, pp. 1502–1510, Apr. 2013, doi: 10.1001/JAMA.2013.3231.
- [121] D. R. Zerbino and E. Birney, “Velvet: algorithms for de novo short read assembly using de Bruijn graphs,” *Genome Res.*, vol. 18, no. 5, pp. 821–829, May 2008, doi: 10.1101/GR.074492.107.
- [122] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara, “MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads,” *Nucleic Acids Res.*, vol. 40, no. 20, p. e155, Nov. 2012, doi: 10.1093/NAR/GKS678.
- [123] A. Bankevich *et al.*, “SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing,” *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, May 2012, doi: 10.1089/CMB.2012.0021.
- [124] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, “MetaSPAdes: A new versatile metagenomic assembler,” *Genome Res.*, vol. 27, no. 5, pp. 824–834, May 2017, doi: 10.1101/GR.213959.116/-/DC1.
- [125] S. Boisvert, F. Raymond, É. Godzaridis, F. Laviolette, and J. Corbeil, “Ray Meta: Scalable de novo metagenome assembly and profiling,” *Genome Biol.*, vol. 13, no. 12, pp. 1–13, Dec. 2012, doi: 10.1186/gb-2012-13-12-r122.
- [126] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, “IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth,” *Bioinformatics*, vol. 28, no. 11, pp. 1420–1428, Jun. 2012, doi: 10.1093/BIOINFORMATICS/BTS174.
- [127] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph,” doi: 10.1093/bioinformatics/btv033.
- [128] B. Haider, T. H. Ahn, B. Bushnell, J. Chai, A. Copeland, and C. Pan, “Omega: an overlap-graph de novo assembler for metagenomics,” *Bioinformatics*, vol. 30, no. 19, pp. 2717–2722, Apr. 2014, doi: 10.1093/BIOINFORMATICS/BTU395.
- [129] Y. Wang, Y. Hu, and G. F. Gao, “Combining metagenomics and metatranscriptomics to study human, animal and environmental resistomes,” *Med. Microecol.*, vol. 3, p. 100014, Mar. 2020, doi: 10.1016/J.MEDMIC.2020.100014.
- [130] S. Y. Niu, J. Yang, A. McDermaid, J. Zhao, Y. Kang, and Q. Ma, “Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes,” *Brief. Bioinform.*, vol. 19, no. 6, pp. 1415–1429, Nov. 2018, doi: 10.1093/BIB/BBX051.
- [131] M. Boolchandani, A. W. D’Souza, and G. Dantas, “Sequencing-based methods and resources to study antimicrobial resistance,” *Nat. Rev. Genet.* 2019 206, vol. 20, no. 6, pp. 356–370, Mar. 2019, doi: 10.1038/s41576-019-0108-4.
- [132] B. P. Alcock *et al.*, “CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D517–D525, Jan. 2020, doi: 10.1093/NAR/GKZ935.

- [133] V. Bortolaia *et al.*, “ResFinder 4.0 for predictions of phenotypes from genotypes,” *J. Antimicrob. Chemother.*, vol. 75, no. 12, pp. 3491–3500, Dec. 2020, doi: 10.1093/JAC/DKAA345.
- [134] M. K. Gibson, K. J. Forsberg, and G. Dantas, “Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology,” *ISME J. Adv. online Publ.*, 2014, doi: 10.1038/ismej.2014.106.
- [135] B. Liu and M. Pop, “ARDB--Antibiotic Resistance Genes Database,” *Nucleic Acids Res.*, vol. 37, no. Database issue, 2009, doi: 10.1093/NAR/GKN656.
- [136] S. M. Lakin *et al.*, “MEGARes: an antimicrobial resistance database for high throughput sequencing,” *Nucleic Acids Res.*, vol. 45, no. Database issue, p. D574, Jan. 2017, doi: 10.1093/NAR/GKW1009.
- [137] S. K. Gupta *et al.*, “ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes,” *Antimicrob. Agents Chemother.*, vol. 58, no. 1, pp. 212–220, Jan. 2014, doi: 10.1128/AAC.01310-13.
- [138] E. Ruppé *et al.*, “Prediction of the intestinal resistome by a three-dimensional structure-based method,” *Nat. Microbiol.* 2018 41, vol. 4, no. 1, pp. 112–123, Nov. 2018, doi: 10.1038/s41564-018-0292-6.
- [139] J. C. Wallace, J. A. Port, M. N. Smith, and E. M. Faustman, “FARME DB: a functional antibiotic resistance element database,” *Database J. Biol. Databases Curation*, vol. 2017, no. 1, Jan. 2017, doi: 10.1093/DATABASE/BAW165.
- [140] X. Yin *et al.*, “ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes,” *Bioinformatics*, vol. 34, no. 13, pp. 2263–2270, Jul. 2018, doi: 10.1093/BIOINFORMATICS/BTY053.
- [141] K. Bush and G. A. Jacoby, “Updated functional classification of beta-lactamases,” *Antimicrob. Agents Chemother.*, vol. 54, no. 3, pp. 969–976, Mar. 2010, doi: 10.1128/AAC.01009-09.
- [142] T. Naas *et al.*, “Beta-lactamase database (BLDB) – structure and function,” *J. Enzyme Inhib. Med. Chem.*, vol. 32, no. 1, pp. 917–919, Jan. 2017, doi: 10.1080/14756366.2017.1344235.
- [143] Q. K. Thai, F. Bös, and J. Pleiss, “The lactamase engineering database: A critical survey of TEM sequences in public databases,” *BMC Genomics*, vol. 10, no. 1, p. 390, Aug. 2009, doi: 10.1186/1471-2164-10-390.
- [144] A. Srivastava, N. Singhal, M. Goel, J. S. Viridi, and M. Kumar, “CBMAR: a comprehensive  $\beta$ -lactamase molecular annotation resource,” *Database*, vol. 2014, Jan. 2014, doi: 10.1093/database/bau111.
- [145] A. Dhariwal, R. Junges, T. Chen, and F. C. Petersen, “ResistoXplorer: a web-based tool for visual, statistical and exploratory data analysis of resistome data,” *NAR Genomics Bioinforma.*, vol. 3, no. 1, 2021, doi: 10.1093/nargab/lqab018.
- [146] L. G. Panunzi, “sraX: A Novel Comprehensive Resistome Analysis Tool,” *Front.*

- Microbiol.*, vol. 11, Feb. 2020, doi: 10.3389/fmicb.2020.00052.
- [147] M. Oh, A. Pruden, C. Chen, L. S. Heath, K. Xia, and L. Zhang, "MetaCompare: a computational pipeline for prioritizing environmental resistome risk," *FEMS Microbiol. Ecol.*, vol. 94, no. 7, Jul. 2018, doi: 10.1093/femsec/fiy079.
- [148] J. Q. Su *et al.*, "Metagenomics of urban sewage identifies an extensively shared antibiotic resistome in China," *Microbiome*, vol. 5, no. 1, pp. 1–15, Jul. 2017, doi: 10.1186/s40168-017-0298-y.
- [149] T. W. Edgar and D. O. Manz, "Machine Learning," *Res. Methods Cyber Secur.*, pp. 153–173, Jan. 2017, doi: 10.1016/B978-0-12-805349-2.00006-6.
- [150] P. Schneider and F. Xhafa, "Machine learning: ML for eHealth systems," *Anom. Detect. Complex Event Process. over IoT Data Streams*, pp. 149–191, Jan. 2022, doi: 10.1016/B978-0-12-823818-9.00019-5.
- [151] P. Bangert, "Machine Learning," *Mach. Learn. Data Sci. Oil Gas Ind.*, pp. 37–67, Jan. 2021, doi: 10.1016/B978-0-12-820714-7.00003-0.
- [152] R. R. C. Cuadrat, M. Sorokina, B. G. Andrade, T. Goris, and A. M. R. Dávila, "Global ocean resistome revealed: exploring Antibiotic Resistance Genes (ARGs) abundance and distribution on TARA oceans samples through machine learning tools," *bioRxiv*, p. 765446, Sep. 2019, doi: 10.1101/765446.
- [153] L. G. Li, X. Yin, and T. Zhang, "Tracking antibiotic resistance gene pollution from different sources using machine-learning classification," *Microbiome*, vol. 6, no. 1, p. 93, May 2018, doi: 10.1186/S40168-018-0480-X/FIGURES/5.
- [154] E. F. Morales and H. J. Escalante, "A brief introduction to supervised, unsupervised, and reinforcement learning," *Biosignal Process. Classif. Using Comput. Learn. Intell. Princ. Algorithms, Appl.*, pp. 111–129, Jan. 2022, doi: 10.1016/B978-0-12-820125-1.00017-8.
- [155] B. Qian *et al.*, "Orchestrating the Development Lifecycle of Machine Learning-based IoT Applications: A Taxonomy and Survey," *ACM Comput. Surv.*, vol. 53, no. 4, 2020, doi: 10.1145/3398020.
- [156] K. El Bouchefry and R. S. de Souza, "Learning in Big Data: Introduction to Machine Learning," *Knowl. Discov. Big Data from Astron. Earth Obs. Astrogeoinformatics*, pp. 225–249, Jan. 2020, doi: 10.1016/B978-0-12-819154-5.00023-0.
- [157] M. Vafakhah and S. Janizadeh, "Application of artificial neural network and adaptive neuro-fuzzy inference system in streamflow forecasting," *Adv. Streamflow Forecast.*, pp. 171–191, Jan. 2021, doi: 10.1016/B978-0-12-820673-7.00002-0.
- [158] L. Wang, Z. Zhang, X. Zhang, X. Zhou, P. Wang, and Y. Zheng, "A Deep-forest based approach for detecting fraudulent online transaction," *Adv. Comput.*, vol. 120, pp. 1–38, Jan. 2021, doi: 10.1016/BS.ADCOM.2020.10.001.
- [159] J. Djuris, S. Ibric, and Z. Djuric, "Neural computing in pharmaceutical products and process development," *Comput. Appl. Pharm. Technol.*, pp. 91–175, Jan. 2013, doi: 10.1533/9781908818324.91.
- [160] J. Vasilakes, S. Zhou, and R. Zhang, "Natural language processing," *Mach. Learn.*

- Cardiovasc. Med.*, pp. 123–148, Jan. 2021, doi: 10.1016/B978-0-12-820273-9.00006-3.
- [161] G. F. M. de Souza, A. Caminada Netto, A. H. de Andrade Melani, M. A. de Carvalho Michalski, and R. F. da Silva, “Engineering systems’ fault diagnosis methods,” *Reliab. Anal. Asset Manag. Eng. Syst.*, pp. 165–187, Jan. 2022, doi: 10.1016/B978-0-12-823521-8.00006-2.
- [162] D. A. Pisner and D. M. Schnyer, “Support vector machine,” *Mach. Learn. Methods Appl. to Brain Disord.*, pp. 101–121, Jan. 2020, doi: 10.1016/B978-0-12-815739-8.00006-7.
- [163] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015, doi: 10.1016/J.CSBJ.2014.11.005.
- [164] A. Benkessirat and N. Benblidia, “Fundamentals of feature selection: An overview and comparison,” *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, vol. 2019–November, Nov. 2019, doi: 10.1109/AICCSA47632.2019.9035281.
- [165] L. Oneto, “Model Selection and Error Estimation in a Nutshell,” vol. 15, 2020, doi: 10.1007/978-3-030-24359-3.
- [166] P. Nousi, M. Tzelepi, N. Passalis, and A. Tefas, “Lightweight deep learning,” *Deep Learn. Robot Percept. Cogn.*, pp. 131–164, Jan. 2022, doi: 10.1016/B978-0-32-385787-1.00012-9.
- [167] H. S. Das and P. Roy, “A Deep Dive Into Deep Learning Techniques for Solving Spoken Language Identification Problems,” *Intell. Speech Signal Process.*, pp. 81–100, Jan. 2019, doi: 10.1016/B978-0-12-818130-0.00005-2.
- [168] H. Cai, J. Lin, and S. Han, “Efficient methods for deep learning,” *Adv. Methods Deep Learn. Comput. Vis.*, pp. 159–190, Jan. 2022, doi: 10.1016/B978-0-12-822109-9.00013-8.
- [169] S. S. Kunapuli and P. C. Bhallamudi, “A review of deep learning models for medical diagnosis,” *Mach. Learn. Big Data, IoT Med. Informatics*, pp. 389–404, Jan. 2021, doi: 10.1016/B978-0-12-821777-1.00007-0.
- [170] L. G. Li, X. Yin, and T. Zhang, “Tracking antibiotic resistance gene pollution from different sources using machine-learning classification,” *Microbiome*, vol. 6, no. 1, p. 93, 2018, doi: 10.1186/s40168-018-0480-x.
- [171] Y. Li *et al.*, “HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes,” *Microbiome*, vol. 9, no. 1, pp. 1–12, Dec. 2021, doi: 10.1186/S40168-021-01002-3/FIGURES/4.
- [172] A. P. Arkin *et al.*, “KBase: The United States department of energy systems biology knowledgebase,” *Nature Biotechnology*, vol. 36, no. 7. Nature Publishing Group, pp. 566–569, Jul. 06, 2018, doi: 10.1038/nbt.4163.
- [173] P. Menzel, K. L. Ng, and A. Krogh, “Fast and sensitive taxonomic classification for metagenomics with Kaiju,” *Nat. Commun.* 2016 71, vol. 7, no. 1, pp. 1–9, Apr. 2016, doi: 10.1038/ncomms11257.
- [174] C. Quast *et al.*, “The SILVA ribosomal RNA gene database project: improved data



- processing and web-based tools,” *Nucleic Acids Res.*, vol. 41, no. D1, pp. D590–D596, Jan. 2013, doi: 10.1093/NAR/GKS1219.
- [175] E. Kopylova, L. Noé, and H. Touzet, “SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data,” *Bioinformatics*, vol. 28, no. 24, pp. 3211–3217, Dec. 2012, doi: 10.1093/BIOINFORMATICS/BTS611.
- [176] P. T. L. C. Clausen, F. M. Aarestrup, and O. Lund, “Rapid and precise alignment of raw reads against redundant databases with KMA,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–8, Aug. 2018, doi: 10.1186/S12859-018-2336-6/TABLES/2.
- [177] C. Camacho *et al.*, “BLAST+: Architecture and applications,” *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–9, Dec. 2009, doi: 10.1186/1471-2105-10-421/FIGURES/4.
- [178] B. Buchfink, C. Xie, and D. H. Huson, “Fast and sensitive protein alignment using DIAMOND,” *Nat. Methods* 2014 121, vol. 12, no. 1, pp. 59–60, Nov. 2014, doi: 10.1038/nmeth.3176.
- [179] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, “CD-HIT Suite: a web server for clustering and comparing biological sequences,” *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010, doi: 10.1093/BIOINFORMATICS/BTQ003.
- [180] H. McWilliam *et al.*, “Analysis Tool Web Services from the EMBL-EBI,” *Nucleic Acids Res.*, vol. 41, no. W1, pp. W597–W600, Jul. 2013, doi: 10.1093/NAR/GKT376.
- [181] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–16, Mar. 2013, doi: 10.1186/1471-2105-14-106/FIGURES/7.
- [182] K. Gajowniczek and T. Ząbkowski, “ImbTreeAUC: An R package for building classification trees using the area under the ROC curve (AUC) on imbalanced datasets,” *SoftwareX*, vol. 15, p. 100755, 2021, doi: 10.1016/j.softx.2021.100755.
- [183] E. Gonzalez, F. E. Pitre, and N. J. B. Brereton, “ANCHOR: a 16S rRNA gene amplicon pipeline for microbial analysis of multiple environmental samples,” *Environ. Microbiol.*, vol. 21, no. 7, pp. 2440–2468, Jul. 2019, doi: 10.1111/1462-2920.14632.
- [184] F. Maguire, B. Jia, K. L. Gray, W. Y. V. Lau, R. G. Beiko, and F. S. L. Brinkman, “Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands,” *Microb. Genomics*, vol. 6, no. 10, pp. 1–12, 2020, doi: 10.1099/mgen.0.000436.