Identification of novel antibiotic resistance genes through the exploration of mobile genetic elements

Mohammad Razavi

Department of Infectious Diseases
Institute of Biomedicine
Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG
Gothenburg 2020

Identification of novel antibiotic resistance genes through the exploration of mobile genetic elements

Mohammad.Razavi@gu.se

ISBN 978-91-7833-902-0 (PRINT)
ISBN 978-91-7833-903-7 (PDF)

Printed in Gothenburg, Sweden 2020
Printed by Stema Specialtryck AB, Borås

To my family

Identification of novel antibiotic resistance genes through the exploration of mobile genetic elements

Mohammad Razavi

ABSTRACT

Backgrounds and aims: The evolution of multi-resistant pathogens is seriously threatening our ability to provide modern healthcare. Many of the mobile resistance factors in clinics appear to originate from environmental bacteria. In this thesis, strategies are developed and applied to explore and identify novel antibiotic resistance genes (ARGs) captured and carried by different mobile genetic elements on a large-scale. The primary aim is to identify novel, mobile ARGs that could become, or that already are, a threat to the public health.

Method: We used targeted amplicon sequencing (Paper I) and functional metagenomics of amplified gene cassettes (Papers II and III) that were recovered from two polluted environments. In Paper IV, we studied the associations between insertion sequences (ISs) and ARGs by analyzing all sequenced bacterial genomes. Moreover, several thousand metagenomic runs were analyzed to estimate the abundance of ARGs (including novel ARGs) and ISs.

Results and discussions: In Paper I, we found a novel mobile sulfonamide resistance gene providing a high level of resistance when expressed in *E. coli*. By using functional metagenomics (Papers II and III), we identified a completely new integron-borne aminoglycoside resistance gene that was already present, but previously not identified, in multi-resistant clinical isolates collected from patients in Italy, as well as in two food-borne *Salmonella enterica* isolates from the USA. Moreover, we described and characterized the first *ampC*s encoded as integron gene cassettes with increased transmission opportunities to move between different bacterial species. Metagenomic analysis showed that all three genes are spread in different geographical locations and were abundant in wastewater environments. In Paper IV, ISs and tentative composite transposons with strong associations with ARGs were identified, and we proposed that these could be explored further to discover novel ARGs, for example with an amplicon sequencing approach. Finally, metagenomic analyses shed light on the environments that potentially contain such ISs.

Conclusions: Targeted amplicon sequencing and its integration with functional metagenomics were successful in finding novel resistance gene cassettes that have already accumulated in pathogens or have the potential to do so. With a well-designed strategy, the content of ISs could be explored to identify unknown mobile ARGs in addition to those associated with integron gene cassettes. Finally, the information produced in this thesis is the initial seed for an accessible web application useful in studying the association between ISs and ARGs.

Keywords: Antibiotic resistance, resistome, integron, insertion sequences, metagenomics, environment

SAMMANFATTNING PÅ SVENSKA

Antibiotika är ett av vårdens allra viktigaste verktyg. Tyvärr utvecklar allt fler sjukdomsframkallande bakterier förmågan att motstå antibiotikabehandling. Bakterierna kan utveckla resistens mot antibiotika genom förändringar av bakteriernas eget DNA. Här är det allra största problemet är att de kan ta upp helt nytt DNA från andra bakterier. Man tror att många av de resistensgener som utgör stora problem i dag kommer från ofarliga bakterier i vår tarmflora eller vår omgivning. Att kunna förutspå eller tidigt upptäcka nya resistensgener i sjukdomsframkallande bakterier kan ha ett stort värde. Dels möjliggör det övervakning och därmed tidiga åtgärder för att begränsa genernas spridning, dels möjliggör det molekylär diagnostik, och dels kan det ge läkemedelsindustrin ett försprång när de försöker utveckla nya antibiotiska molekyler.

I denna avhandling har vi letat efter nya resistensgener med olika metoder. Gemensamt för de första tre delarbetena är att vi letat specifikt i DNA-strukturer som kallas integroner. Dessa har förmågan att klippa ut och klippa in gener i form av så kallade genkasetter. Ofta finns integroner både på kromosomer och på plasmider, och de senare kan ofta flytta sig mellan celler. På så sätt kan gener som förekommer i form av genkassetter öka sin rörlighet, och därmed ökar risken att de hamnar i sjukdomsframkallande bakterier.

I den första studien fann vi en ny gen, *sul4*, som ger resistens mot sulfonamid-antibiotika. Tidigare kände man bara till tre gener som ger resistens mot denna mycket brett använda antibiotikaklass. I det andra delarbetet fann vi, med en annan metod, en ny resistensgen, *gar*, som ger resistens mot aminoglykosidantibiotika. Det visade sig att denna gen hade undgått upptäckt i flera kliniskt relevanta bakterier, däribland *Pseudomonas aeruginosa* som bland annat kan orsaka lunginflammation och *Salmonella entericia* som kan ge allvarliga mag-tarm infektioner. Såvitt vi vet är detta första gången man hittat en helt ny antibiotikaresistensgen som redan finns i kliniken genom att studera den yttre miljön. I det tredje arbetet fann vi en speciell typ av gen mot penicillin-liknande antibiotika i form av genkasetter. Denna grupp av gener (*ampC*) har tidigare inte hittas i form av genkasetter. Det öppnar upp fler möjligheter för spridning mellan arter. Slutligen utforskades hur starkt associerade resistensgener är med en annan typ av genetiska strukturer, så kallade ”Insertion Sequences” (IS). En möjlighet att upptäcka nya resistensgener skulle kunna vara att utforska den genetiska omgivningen hos de IS-sekvenser som i dag är associerade med en lång rad olika, kända resistensgener.

LIST OF PAPERS

This thesis is based on the following articles and manuscripts.

I. Mohammad Razavi, Nachiket P. Marathe, Michael R. Gillings, Carl-Fredrik Flach, Erik Kristiansson, and D. G. Joakim Larsson. **Discovery of the fourth mobile sulfonamide resistance gene**. Microbiome 2017; 5:160.

II. Maria-Elisabeth Böhm, Mohammad Razavi, Nachiket P. Marathe, Carl-Fredrik Flach, and D. G. Joakim Larsson. **Discovery of a novel integron-borne aminoglycoside resistance gene present in clinical pathogens by screening environmental bacterial communities**. Microbiome 2020; 8:41.

III. Maria-Elisabeth Böhm, Mohammad Razavi, Carl-Fredrik Flach, and D. G. Joakim Larsson. **A Novel, Integron-Regulated, Class C β-Lactamase**. Antibiotics 2020, 9(3), 123.

IV. Mohammad Razavi, Erik Kristiansson, Carl-Fredrik Flach, and D. G. Joakim Larsson. **Can the association with insertion sequences guide the discovery of novel antibiotic resistance genes?** Submitted.

CONTENT

ABBREVIATIONS

3'-CS	3'-conserved segments of class 1 integron
ARG	Antibiotic resistance gene
ARIA	Associations repository of ISs and ARGs
BacMet	Biocide and metal resistance database
*bla*IDC	Integron-derived cephalosporinase
BLAST	Basic local alignment search tools
CARD	Comprehensive antibiotic resistance database
CCD	Charge-coupled device
CDD	Conserved domain database
ddNTPs	Dideoxynucleotide triphosphates
dNTPs	Deoxyribonucleotide triphosphates
EBI	The European bioinformatics institute
EMBL	The European molecular biology laboratory
gar	Garosamine-specific aminoglycoside resistance
GTA	Gene transfer agent
HGT	Horizontal gene transfer
HMM	Hidden Markov model
ICE	Integrative and conjugative element
IR	Inverted repeat
NCBI	The national center for biotechnology information
NGS	Next-generation sequencing
MIC	Minimum inhibitory concentration
OLC	Overlap layout consensus
ORF	Open reading frame
PCR	Polymerase chain reaction
PETL	Patancheru Enviro Tech Ltd.
SMRT	Single-molecule real-time sequencing
SRA	Sequence read archive
TE	Transposable elements
VFDB	Virulence factor database
ZMW	Zero mode waveguide
WHO	World health organization

1. INTRODUCTION

1.1 BACTERIA

Bacteria are ubiquitous prokaryotic microorganisms that emerged on earth around three billion years ago. They have adapted to essentially all types of habitats, from highly acidic, hot volcanic lakes to the frozen sediment from glaciers in the Arctic and Antarctica. Life on earth is dependent on many bacterial-driven cycles, including nitrogen fixation, carbon assimilation, and recycling of organic materials, to name a few (Maier et al., 2009). Some bacteria that can live in/on humans and animals could invade their host tissues and induce infectious diseases. Several million people die each year due to bacterial infections and many more are severely affected (e.g., 1.5M deaths from tuberculosis, 1.9M deaths from lower respiratory infections caused by *Streptococcus pneumonia*, 0.25M deaths from diarrhea caused by bacteria, etc.) (Troeger et al., 2018; Troeger et al., 2017; WHO, 2019). Bacteria are often separated into two main categories based on their cell envelopes: Gram-positive and Gram-negative. The architecture of the cell wall, along with the cytoplasmic differences with eukaryotic cells, dictate the strategies for preventing or treating bacterial infections, most commonly using chemical agents called antibiotics.

1.2 ANTIBIOTICS

Antibiotics are natural or synthetic chemical compounds that could be used against bacteria to kill or inhibit their growth, while similar concentrations have no or limited effects on eukaryotic cells. Natural antibiotics are primarily secondary metabolites produced by bacteria (e.g., *Streptomyces* spp.) or fungi (e.g., *Penicillium* spp.). Their original function was likely the regulation of growth and to act in the competition for resources with neighboring microorganisms (Hibbing et al., 2010). However, in the clinic, they are used to kill and stop the growth of infectious bacteria by targeting various bacterial structures or processes, including the cell wall, ribosomes, and specific biosynthetic pathways. Antibiotics are divided into different families based on their targets and chemical structures. For instance, beta-lactams are natural or semi-synthetic agents that inhibit peptidoglycan transpeptidases and, thereby, halt the cell wall assembly. Aminoglycosides are composed of natural or semi-synthetic antibiotics produced mostly by soil bacteria. They inhibit the growth of bacteria by binding to the 30S ribosomal unit and disrupting protein

synthesis. Sulfonamides are synthetic drugs that utilize the inability of bacteria to take up folic acid by interfering with their folate biosynthesis pathways. However, bacteria could withstand the toxic effect of antibiotics in different ways and confer resistance.

1.3 ANTIBIOTIC RESISTANCE

The ability of bacteria to resist the effect of antibiotics could largely be divided into the forms of intrinsic resistance and acquired resistance. Intrinsic resistance refers to the case when all members of a species are resistant, often due to the incongruity of the antibiotic mode of action (Cox & Wright, 2013). For instance, the inability of anaerobic bacteria to take up aminoglycosides makes them intrinsically resistant to this family of antibiotics (Mingeot-Leclercq et al., 1999). When bacterial evolution is considered over very long times, one may argue that even traits that are common to all members of a species today could have been horizontally acquired at some point in its evolution, though this possibility is of less clinical and practical relevance. In contrast, acquired resistance involves traits that could appear in bacteria at any point in time through mutations or via horizontal gene transfer mechanisms. Important acquired resistance mechanisms include target site modification or protection (e.g., mutation of DNA gyrase in the case of fluoroquinolone resistance or methylation of ribosomal RNA in the case of macrolide resistance), antibiotic modification (e.g., hydrolyzing beta-lactam using beta-lactamase enzymes or modifying aminoglycoside with different chemical groups), reduction of permeability (e.g., via the overexpression of pumps extruding antibiotics), and overproduction of targets (e.g., the overproduction of dihydrofolate reductase due to the alteration of the gene promotor in trimethoprim resistance). Antibiotic resistance is thought to be in balance with antibiotic production in pristine environments (Martínez, 2008). The presence of resistance genes in antibiotic-producing strains may be related to their protective roles against their own/neighbors' products, or even involvement in the biosynthetic pathways of antibiotics (Sengupta et al., 2013). However, the massive use of antibiotics in human medicine and agriculture (i.e., since the discovery of sulfonamides) has disrupted the balance and led to the accumulation of antibiotic resistance genes (ARGs) in non-producer strains like human commensals and pathogens.

1.4 ANTIBIOTIC RESISTOME

The group of all existing known and unknown ARGs in clinical and environmental bacteria is sometimes referred to as the antibiotic *resistome* (Perry et al., 2014). It encompasses acquired ARGs that provide acquired resistance (see section 1.3) and intrinsic ARGs that are mostly immobile chromosomal ARGs transferred vertically. In addition, resistome contains two other groups of genes: those with potential resistance functions that are not expressed in their current hosts (i.e., silent ARGs) and those that require further evolution to provide a resistance functions by, for example, mutations (i.e., proto-resistance genes). For instance, chromosomal *ampC* in some strains of *Citrobacter freundii* was a silent ARG that, through mutation of its regulator (i.e., *ampR* gene), was expressed and provided with a resistance phenotype. In contrast, while proteins encoded by proto-resistance genes might have structural similarities with those in other groups of the resistome, they provide little or no resistance function (Morar & Wright, 2010). For instance, a group of genes encoding protein kinases has high enzymatic similarity with aminoglycoside modifying enzymes and may be considered proto-resistance genes (see Fig. 3 in Paper II). Such genes could potentially confer resistance via a series of mutations. Proto-, silent-, and intrinsic resistance genes are dependent on horizontal gene transfer mechanisms (HGT) to become acquired ARGs. With further dissemination among versatile bacterial species, they could accumulate mutations to improve their resistance function and reside in clinical pathogens.

1.5 HORIZONTAL GENE TRANSFER

Bacteria can exchange genetic material to acquire adaptive phenotypes through various mechanisms that comprise horizontal gene transfer. In addition to the transfer of genetic material to a recipient cell, they should be inherited by the recipient offspring (Gillings, 2016). The acquisition of new ARGs and associated phenotypes might not be limited to the ARGs themselves, but could be facilitated by the simultaneous transfer of virulence factors or other groups of genes that could enhance the fitness of the bacteria in a new environment (co-selection). The HGT mechanisms are divided into three broad categories: conjugation (i.e., the transfer of conjugative and mobilizable plasmids), transformation (i.e., the uptake of free DNA or secreted membrane vesicles), and transduction (e.g., the transfer of genes by bacteriophage and gene transfer agents (GTA)) (Gillings, 2016; von Wintersdorff et al., 2016).

Exposure of bacteria to antibiotics could lead to the emergence of novel ARGs and the activation of HGT mechanisms. The SOS system is an ancient trait in bacteria in response to DNA damages that, for instance, are caused by the bactericidal effect of antibiotics (Gillings & Stokes, 2012). It activates error-prone DNA polymerases, leading to higher mutation rates that might cause, for instance, the modification or overproduction of the antibiotics' targets. Moreover, the SOS response can increase the rate of HGT mechanisms, including conjugations (Jutkina et al., 2016) and transduction (Allen et al., 2011), which, in turn, could increase the spread of ARGs within the community. Several steps are required for an ARG that is present on a chromosome in a non-pathogenic species to emerge in pathogens. The process of moving resistance determinants between immobile (e.g., chromosome) and mobile (e.g., conjugative plasmids) contexts is crucial. Transposable elements (TEs) are responsible for these types of movements of DNA within a bacterial cell.

1.6 TRANSPOSABLE ELEMENTS

Transposable elements often contain genes called transposases that encode enzymes responsible for the intra-cellular movement of DNA in different locations on genomes. Transposases are the most abundant genes in nature (Aziz et al., 2010) and significantly contributed to genome evolution in all branches of life. Transposable elements are classified as class I and II based on their transposition mechanisms (Lerat, 2010). The former, also known as retrotransposons, are mediated by an intermediate RNA and use replicative mechanisms, whereas the latter, known as DNA transposons, are DNA-mediated and primarily used the non-replicative mechanism. Moreover, TEs are divided into autonomous and non-autonomous groups, which are defined based on the presence or absence of self-encoding transposase, respectively.

Autonomous DNA transposons in bacterial genomes could be further classified into other types, including unit transposons and insertion sequences (ISs) (Partridge et al., 2018). Unit transposons are long (typically over 5 kb) TEs that mobilize several accessory genes in one unit (Brenner & Miller, 2014). The unit is flanked by inverted repeats (IRs) and target site duplications. In contrast, insertion sequences are smaller TEs (typically less than 3 kb) containing mostly a transposase gene surrounded by inverted repeats and target site duplications (Vandecraen et al., 2017). Pairs of ISs could form composite transposons and move the DNA region in between them. The boundary between unit transposons and ISs can be diffuse and confusing, as there are also examples of single IS units carrying accessory genes. Nevertheless, both

require a DNA binding domain to detect the IRs and catalytic domains for the excision and integration of mobile DNA. The catalytic mechanisms of transposase, which involve various nuclease activities, divide them into different groups, including phosphoserine, phosphotyrosine, HUH, and DD(E/D) transposes (Hickman & Dyda, 2015). Serine transposases (e.g., the IS*607* family) have arginine at the active sites and serine as a nucleophile, which acts on double-stranded DNA. Meanwhile, phosphotyrosine transposases (e.g., Tn*916*) cleave DNA using a single tyrosine residue and form a phosphotyrosine bond with DNA. Transposases with HUH domains (e.g., IS*91* and IS*200* families) contain two histidine residues and one non-conserved hydrophobic residue at the active site and use two nucleophilic tyrosines to move single-stranded DNA. They use rolling-circle transposition and, due to a lack of site-specificity, could target different sites on genomes. Transposases with DD(E/D) domains have three acetic residues (Asp, Asp, Glu) at their active site that are responsible for coordinating two metal ions needed for DNA cleavage and joining.

Insertion sequences containing DDE domains are the most abundant TEs on bacterial genomes (Siguier et al., 2014). They contribute to genome plasticity and adaptability by providing necessary resistance, virulence, pathogenic, and catabolic phenotypes (Vandecraen et al., 2017). They could decontextualize genes, act as promoters for silent genes, or inactivate gene products by interrupting the open reading frame or the promoter (Poirel et al., 2017). To name a few examples, ISs provide the promotor for an otherwise-silent expanded-spectrum beta-lactamase encoding gene (*bla*CTX-M) (Lartigue et al., 2006), increase the pathogenicity of a methicillin-resistant *Staphylococcus aureus* strain by interrupting a toxin production repressor (*rot* gene) (Benson et al., 2014), and enhance the fitness of *E. coli* carrying an interrupted *rpoS* gene in glucose- and phosphate-limited chemostats (Gaffé et al., 2011) Moreover, ISs often mobilize gene acquisition systems called integrons that are responsible for capturing and expressing acquired genes with adaptive phenotypes.

1.7 INTEGRONS

Integrons are ancient structures that shaped the evolution of bacteria by capturing and expressing genes, in the form of gene cassettes, to rapidly adapt to a changing environment. They have three main features: an integron integrase gene (*intI*), a cassette promoter (Pc), and a recombination site (*attI*) (Abella et al., 2015). The integrase gene has a tyrosine recombinase domain performing the insertion and excitation of circular gene cassettes at the *attI* site,

mostly integrating them in a reverse direction to itself. The expression of gene cassettes is derived by the Pc located within *intI* or between *intI* and the *attI* site. The strength of expression is dependent on the type of promoter and the proximity of the gene cassette and Pc. The expression of *intI* gene could lead to reshuffling of gene cassettes and potentially integrating several instances of a gene to enhance the expression and, consequently, the phenotype. The SOS response could also increase the expressions of *intI* and the subsequent acquisition and reshuffling of gene cassettes until a proper adaptive response is found (Escudero et al., 2015).

Integrons are divided into chromosomal and mobile integrons. The former have appeared on chromosomes of hundreds of bacterial species and contain up to 200 gene cassettes that mostly have unknown functions. In contrast, mobile integrons are carried by TEs and contain fewer than 10 gene cassettes (Gillings, 2014; Stalder et al., 2012). They are classified based on the homology of the *intI* gene. Among them, the clinical class 1 integron has been abundant in pathogenic *Proteobacteria* and carries mostly ARGs. This type of integron seems to have been mobilized originally from the chromosome of a *beta-proteobacterium* in a biofilm or freshwater environment by a Tn*402* transposon; through a series of mobilizations by other TEs (i.e., Tn*21*) and co-selection with biocide and metal resistance genes (i.e., mercury resistance), it has reached a recognizable genetic context. Its 3'-conserved segments (3'-CS), or downstream of gene cassette array, contain truncated *qacE (qacEΔ)*, *sul1* and a gene encoding an unknown function (Gillings, 2014).

Antibiotic pressure plays an important role in the dissemination of mobile integrons among pathogenic and human commensal bacteria. It is a driving force for the accumulation of resistance gene cassettes on bacterial genomes. Almost 6% of sequenced bacterial genomes have integrons (Cury et al., 2016) as a platform for potentially recruiting gene cassettes, including more than 130 identified cassettes that encode antibiotic resistance (Partridge et al., 2009). Mobile integrons are highly abundant in environments with a history of human activities. As an extreme example, as much as 80% of bacteria thriving in antibiotic-contaminated wastewater from drug manufacturing harbored integrons (Marathe et al., 2013). The vast pool of gene cassettes with unknown function, the presence of integron in diverse bacterial species, and their association with TEs could create a path for ARGs to emerge in human and animal pathogens.

1.8 EMERGENCE AND DISSEMINATION OF ARGS

The intertwined problems of ensuring good health for humans and domestic and wild animals, as well as our environments, have encouraged a collaborative effort to address the antimicrobial problems within a one-health perspective (McEwen & Collignon, 2018). In this context, one-health refers to the movement of bacteria and their genes between the environment and microbiota of humans and animals (Larsson et al., 2018). The root of concern is the antimicrobial use and abuse that contribute to the dissemination of resistance determinants. Data from 71 countries shows that, in 2010, more than 70 billion standard units of antibiotics (e.g., sold pills, capsules, and ampoules) were used (Van Boeckel et al., 2014). In 2010, around 63,000 tons of antibiotics were used in animal-framing; this amount could reach 105,000 tons per year by 2030 (Van Boeckel et al., 2015).

Antibiotics used by humans and animals could be partially excreted by urine and feces into the environment. For instance, up to 65% of erythromycin, 72% of fosfomycin, and 35% of ciprofloxacin that are administered orally could end up in the environment (Amábile-Cuevas, 2015). It has been estimated that, in 2010, humans released 15,000 tons of antibiotics in sewage (Amábile-Cuevas, 2015; Van Boeckel et al., 2014). In animal farming, the concentration of antibiotics in liquid waste and manure could reach up to a few hundred in ng/L and μg/kg, respectively (Xie et al., 2018; X. Zhang et al., 2014). However, industrial wastewater could discharge a staggering amount of antibiotic by-products into the environment (Larsson, 2014). For instance, up to 31 mg/L of ciprofloxacin was detected in the effluent of drug manufacturing wastewater in India (Larsson et al., 2007). Antibiotics are diluted and degraded at different rates depending on the properties of their residues and various features of the environment (e.g., pH, temperature, etc.) (Kumar et al., 2019). However, the constant discharge of antibiotics in the environment ensures the presence of sub-lethal concentrations in some places, which could subsequently put selective pressures on bacterial communities (Gullberg et al., 2014; Lundström et al., 2016).

The environment plays a major role in the development and dissemination of ARGs (Bengtsson-Palme et al., 2018). It is considered to be a source of resistance determinants that might end up in human pathogens. Antibiotic exposure could transform proto-resistance genes, silent, or intrinsic (chromosomal) ARGs into acquired ARGs and disseminate them further via HGT mechanisms (Perry et al., 2014). The novel ARGs could emerge and spread in our body under the therapeutic selection pressures of antibiotics or in external environments. The ARGs from external environments must pass

several critical steps to reach human pathogens. The first step is the movement of resistance determinants within genomes and also between other bacterial strains and species using TEs and HGT mechanisms. The positive selection and maintenance leads to their further spread between several species until they reach human pathogens. This also highlights the role of the environment as a transmission platform in which bacteria exchange genetic material with each other and move between humans and animals. Resistant pathogens could be transmitted from one host to another via direct contact or contaminated food (Marathe et al., 2017; Solomon et al., 2002; H. Wang et al., 2014). They could enhance the antibiotic resistance arsenal of environmental bacteria or pick up novel ARGs from them. Many factors are involved in the successful transmission of resistant pathogens and the dissemination of ARGs, including the environmental transmission medium (e.g., bacteria in aerosols are more likely to die than those in water), the adaptability of bacteria that carry ARGs to be colonized in different conditions (e.g., enduring different physical conditions like pH and also being able to live in different hosts), the association of ARGs with mobile genetic elements (e.g., plasmids, Integrative and Conjugative Elements (ICEs), transposases, and integrons), and the cost of novel ARGs in the recipient hosts (Andersson et al., 2020; Bengtsson-Palme et al., 2018). However, the presence of antibiotic selective pressures is a driving force for generating and maintaining resistance determinates that might be recruited by pathogens in the right time and conditions.

1.9 DISCOVERING NOVEL ARGS

ARGs are often discovered through the exploration of genetic material using one of two broad approaches: genomics or metagenomics. The former is a culture-based approach focusing on bacterial isolates, whereas the latter is a culture-independent approach that explores complex microbial communities (Hadjadj et al., 2019). Both approaches could take advantage of next-generation sequencing. ARGs can be identified either by homology-based methods or by functionally assessing the resistance phenotypes they provide. Sequence-alignment algorithms like BLAST or optimized hidden Markov models could identify homolog genes/proteins using a set of known ARGs (Berglund et al., 2019; Schmieder & Edwards, 2012). In the functional approach, the bacterial isolates or the surrogate hosts containing metagenomic DNA are phenotypically assessed using selective growth media (Chistoserdova, 2009; Mullany, 2014). This could also involve mutagenesis techniques through the knocking out of genes conferring resistance (Hadjadj et al., 2019).

Genomic methods can provide high resolution to the genetic contexts containing ARGs. Thus, genomic analyses could clarify important factors regarding, for example, the risk of emergence in pathogens, the level of expression in the host bacterium, the association with TEs or integrons, and the presence on HGTs elements such as conjugative plasmids. However, we should consider that it is currently not feasible to culture the majority of bacterial species and that, even when it is, it may be tedious, time-consuming, and expensive to isolate and characterize bacterial strains one by one. In contrast, metagenomics approaches allow for the studying of a microbial community in a more time- and cost-efficient way, though often at the expense of the loss of the resolution of genetic contexts around the ARGs. The mobility of ARGs and their bacterial hosts are generally more difficult to identify with sequenced-based or functional metagenomics. In this thesis (Paper I), we seek to improve the metagenomics approach through targeted amplicon sequencing of gene cassettes, which, in turn, ensures that the identified novel ARGs are mobile.

The homology-based approach is an efficient way of identifying homologs of known ARGs, but it would be difficult to discover a previously unknown resistance mechanism that is unrelated to known ones. In contrast, functional assays could reveal completely novel resistance mechanisms through the phenotypical assessment of the recovered genetic materials (see section 3.4). However, besides the inability to address mobility, the results of such assays could be overwhelmed by the recovering of previously abundant known ARGs. In this thesis (Paper II), we combined the targeted amplicon sequencing with the functional assay of metagenomics DNA and *in silico* filtering of candidate genes to a) recover novel mobilized ARGs and b) bypass the limitations of finding rare novel resistance genes among much more commonly-known ARGs.

1.10 THE VALUE OF IDENTIFYING NOVEL ARGS

The transfer of ARGs from external environments to human pathogens involves several steps (see section 1.9) and bottlenecks that restrict the number of emerging ARGs in pathogens (Martínez, 2012). The mobility of ARGs, their positive selections and maintenance in the new recipient cells, and the ecological connectivity (i.e., shared habitats of environmental bacteria and pathogens) are among the important bottlenecks that could be used to understand and manage the emergence of ARGs in pathogens. Resistance genes that could bypass each of these barriers impose greater risks for human health (Bengtsson-Palme & Larsson, 2015; Martínez et al., 2015). For instance,

proto-resistance genes (e.g., some protein kinases and acetyltransferases) impose only a minor risk to us, as they are not mobile and do not confer resistance to antibiotics in their current form (Perry et al., 2014). The *bla*LRA-12 gene recovered from remote Alaskan soil imposes a higher risk than proto-resistance genes, as it encodes an active carbapenemase enzyme (Allen et al., 2009; Rodríguez et al., 2017). However, due to the lack of ecological connectivity with human pathogens, it could currently be less likely to cause treatment failures in clinics, in comparison to other ARGs encoding metallo-beta-lactamases, such as *bla*VIM, *bla*IMP, and *bla*NDM. These genes are constantly circulating in human pathogens and are responsible for resistance against our last-resort antibiotics. The emergence of novel ARGs with unknown resistance mechanisms against last-resort antibiotics in pathogens could impose the highest risks on human health (Bengtsson-Palme & Larsson, 2015).

Knowledge about ARGs that have the potential to become clinically relevant or that have already been accumulated in pathogens is valuable. It could facilitate surveillance, thereby enabling better detection and confinement of resistance determinants. For instance, the discovery of the *mcr-1* gene helped create an understanding of colistin-resistant bacteria in several hospital settings around the world (Caselli et al., 2018; Macesic et al., 2019) and initiated the passing on of advice and the implementation of regulations regarding the restricted use of colistin in animal sectors (EMA, 2016; Walsh & Wu, 2016; WHO, 2017b). Moreover, this knowledge could be utilized in molecular diagnostics (Tsalik et al., 2018). It enables assigning isolates as being resistant, based on gene or protein data, without the need to do a phenotypic test, thereby informing antibiotic choice (Evans et al., 2016). Knowledge of ARGs could also be integrated with drug discovery efforts to draw general policies and guide modifications of existing antibiotics, thereby circumventing critical resistance mechanisms. In 2017, the WHO used such information to draw up a priority pathogen list and guide research on the development of new antibiotics against these pathogens (WHO, 2017a). Moreover, a comparison of different classes of beta-lactamases provided clues regarding the modification of side chains of carbapenem to withstand the hydrolyzing effects of known carbapenemase enzymes (Papp-Wallace et al., 2011).

2. AIM

2.1 Hypothesis

- Exploring the context of mobile genetic elements can facilitate the discovery of novel mobile ARGs that are already present or have the potential to emerge in pathogens.

2.2 OVERALL AIM

- To discover novel mobile ARGs that are carried by integrons or transposable elements, both via *in silico* analyses of bacterial genomes and experimentally by recovering mobilized DNA sequences from environmental samples.

2.3 SPECIFIC AIMS

- Identifying novel resistance gene cassettes from polluted river sediments using the high-throughput metagenomic sequencing of amplified gene cassettes combined with a homology-based detection method (Paper I).
- Identifying novel mobile resistance genes with less similarity to known ARGs by functional metagenomics of amplified gene cassettes (Papers II and III).
- Identifying the spread of newly discovered ARGs in publicly available metagenomes and genomes (Papers I-III).
- Analyzing the genetic context around known ARGs and insertion sequences containing DDE domains in all publicly available sequenced bacterial genomes and the association between ARGs and ISs to facilitate the future discovery of novel mobile ARGs (Paper IV).

3. METHODS

3.1 DESCRIPTION OF SAMPLING SITES

In this thesis, the samples were collected from two polluted sites in India. A set of sediment samples was collected from the Mutha River flowing through Pune city in India (see Papers I and II). The river is highly polluted with mostly untreated urban waste and contains a large variety of resistant fecal bacteria. The relative abundance of ARGs in downstream sediments was 30-fold higher than that found upstream (Marathe et al., 2017). Humans and animals have direct contact with the river (i.e., via bathing and seasonal floods), which provides a shared habitat between pathogens and environmental bacteria to interact and possibly exchange ARGs. Moreover, such an environment could facilitate the transmission of resistant bacteria among different human and animal hosts. Identifying novel ARGs is more valuable in this kind of environment, which has lowered barriers for the introduction of resistance determinants into human pathogens.

Moreover, another set of sediment samples was collected from the Isakavagu/Nakkavagu River, which flows past an industrial wastewater treatment plant (Patancheru Enviro Tech Ltd.; PETL) near Hyderabad, India (Kristiansson et al., 2011; Larsson et al., 2007). The selective pressure from antibiotic by-products has led to the enrichment of ARGs. A nearby lake similarly affected by industrial wastewater contained a diverse range of ARGs that were around 7000 times more abundant than those in a Swedish lake (Bengtsson-Palme et al., 2014). Considering the high abundance of known ARGs, it is plausible to identify mobile unknown ARGs in such samples.

3.2 DNA EXTRACTION

The first step in studying a complex bacterial community is DNA extraction. It should be a sensitive and reproducible method that provides sufficient and high-quality DNA while preserving the heterogeneity of the community (Bag et al., 2016). Generally, the process starts with the lysis of the bacterial cell (i.e., chemical and/or physical lysis) and exposure of the double-stranded DNA. At the same time, to reduce DNA damage, the nucleases enzymes should be inactivated through the use of chemical agents and by increasing the pH and salt concentrations. Then, the products are subjected to DNA quantification assay before the downstream experiments or analyses. To determine the concentration of DNA, two approaches are used: one employing photometric measures and one employing fluorometric measures. The former is based on the amount of light (i.e., 260 nm wavelength) absorbed by DNA, and the latter is based on fluorescence signals provided by fluorogenic dyes that could bind to DNA in the sample. Some various commercial kits and protocols have been optimized for the recovery and quantification of DNA from different environments.

In this thesis, we have used the PowerSoil® DNA isolation kit, which is intended for use with environmental samples including sediment. It employs mechanical and chemical cell lysis and uses a silica membrane in a spinning column format. A fluorometric measure was used to calculate the concentration of the extracted DNA using a dsDNA High Sensitivity (HS) Assay kit on the Qubit® Fluorometer. This method has higher sensitivity and could selectively measure DNA in the presence of contamination. Then, the extracted DNA was amplified using the polymerase chain reaction (PCR) method with specific primer pairs.

3.3 POLYMERASE CHAIN REACTION

The polymerase chain reaction is a revolutionary method developed to make many copies of a specific DNA region *in vitro,* for various experiments that require a large quantity of DNA (Bartlett & Stirling, 2003). It is a three-step thermal cycle that utilizes the following key features: two short DNA templates or primer pairs, free nucleotides, a DNA polymerase enzyme, and a PCR buffer that encompasses all of them. The cycle begins with the separation of double-stranded DNA by raising the temperature to about 95°C. Then, in the second step, the temperature is reduced to about 55°C, which allows for the binding of primer pairs to the boundary regions. In the third steps, the temperature is raised to about 70°C and the polymerase enzyme starts sequentially adding the free nucleotides from the 3'-OH group to the other primer, creating a double-stranded DNA. The temperatures should be tuned according to primer pairs and PCR protocols. The required quantity of DNA is produced at an exponential rate by repeating the cycle.

In this thesis, we have used primer pairs that specifically target the content of integrons. The primer pair HS458-HS459 amplifies the region between the integron-integrase gene and the 3'-CS (i.e., *qacE*Δ-*sul1*) of clinical class 1 integrons (Holmes et al., 2003). It could recover all the gene cassettes that accumulated in the integron. The primer pair HS464-GCP2 recovers DNA regions between the *intI* gene and the conserved attachment sites (*attC*). It could amplify DNA with variable length due to the possible binding of GCP2 to any gene cassettes within the array, though shorter amplicons are often

amplified more efficiently (Elsaied et al., 2011). The primer pair MRG284-MRG285 was designed based on the pre-clinical integrons, and targets the regions between *attI* and the chromosomal site that was preserved after mobilization by Tn*402* (Gillings et al., 2009). By using different primer pairs, we could recover the genetic contexts of different integrons (e.g., environmental and clinical) from our metagenomic samples. After the amplification of gene cassettes, the products were sent for sequencing (Paper I) or were used for library preparation and functional metagenomic screening (Paper II and III).

3.4 FUNCTIONAL METAGENOMICS

Functional metagenomics is a culture-independent approach to investigating the functions of genes in microbial communities by cloning and expressing DNA fragments in surrogate hosts and, finally, screening for an acquired function of interest (Mullany, 2014). The first step is to recover genetic materials from environmental samples. The extracted DNA is selected by size and purified before removal of the overhanging bases or synthesizing their complementary strand (DNA blunting). The blunted, size-selected DNA is ligated to a linearized dephosphorylated (i.e., to avoid self-ligation) vector that harbors a constitutively active promoter. Then, the vector is transferred to the cloning host (e.g., *E. coli*) through transduction or transformation by a bacterial phage or electroporation, respectively. Finally, the transformants are screened for the function of interest, such as resistance phenotype.

Functional metagenomics have the ability to reveal completely novel resistance mechanisms. However, they have some limitations as well. Low-abundant genes in the metagenomic samples are often not captured during the library preparation and cloning steps. Lack of expression or lack of functionality of the resistance genes in *E. coli* could produce false negatives. Also, multi-gene resistance mechanisms are difficult to identify. Moreover, a mutated host with acquired resistance phenotypes could produce false positives.

In Paper II, gene cassettes recovered from Indian samples were first amplified and then subjected to a functional metagenomics approach. The recovered gene cassettes were ligated with vector pZE21-MCS1 containing prompter P_{bla}. Then, the library was electroporated to *E. coli* DH10β. The functional screening was performed by the culturing of surrogate hosts on various agar plates containing 13 antibiotics at three different concentrations. Next, all colonies were scraped off each plate, barcoded, and amplified before being sent for DNA sequencing.

3.5 DNA SEQUENCING

The development of different DNA sequencing platforms has been critical to the rapid development of the fields of microbiology and molecular biology in the last decades. Different next-generation sequencing (NGS) methods have provided ample opportunity to retrieve information in a massive and parallel way. In this thesis, we used NGS to study the genomes of individual bacterial isolates and complex microbial communities through the Illumina (Bentley, 2006) and single-molecule real-time (SMRT) (Levene et al., 2003) sequencing technologies. We have also utilized the conventional Sanger sequencing technique (Sanger et al., 1977).

3.5.1 SANGER SEQUENCING

The Sanger sequencing or chain-termination DNA sequencing method was introduced by Frederick Sanger and is based on the synthesis of a complementary DNA strand (Sanger et al., 1977). It incorporates additional features into PCR, called dideoxynucleotide triphosphates (ddNTPs), which stop the elongation of DNA. Original Sanger sequencing starts with the separate running of DNA synthesis with four dideoxynucleotides in parallel, which produces DNA fragments of varying lengths. They are separated and sorted by size using a gel. Then, the sequence of dideoxynucleotides is read sequentially from shorter to longer fragments in the gel to identify the sequence of the input DNA. However, in modern Sanger sequencing, fluorescent markers that emit lights at different wavelengths are attached to each of the four ddNTPs. This enables us to run all reactions in one tube, and then sort the products in one well using capillary electrophoresis that has an accuracy of one nucleotide. Then, the intensity of fluorescents is measured at each position using a laser beam and a charge-coupled device sensor (CCD) (Heather & Chain, 2016). Sanger sequencing is one of the most accurate sequencing technique but is also a tedious approach. Hence, it is not suitable for studying a whole bacterial genome or a complex microbial community. In this thesis, the Sanger sequencing technique was used only to confirm sequences of specific PCR products in Papers I and II.

3.5.2 ILLUMINA SEQUENCING

Illumina sequencing is a sequencing-by-synthesis technique that uses four fluorescently-labeled deoxyribonucleotide triphosphates (dNTPs) (Wanger et al., 2017). These altered nucleotides could release marker fluorophores representing four nucleotides during the cycle of DNA synthesis. Initially, the input DNA are randomly cut into fragments of around several hundred base pairs (bp), depending on the technology (e.g., miSeq® or hiSeq®). Then, short

DNA sequences (i.e., adaptors) are attached to each fragment. Next, they are immobilized on different regions of a surface called flow cell. The attached fragments are rapidly replicated through bridge amplification, which creates clusters of many identical single-stranded DNA templates. Then, the base calling and DNA synthesizing begin by releasing a single dNTP that synthesizes the alternative strand and releases a corresponding fluorophore, saving the image of the flow cell and, finally, enzymatically cleaving the fluorescent dye. This allows for the elongation of the next nucleotide. Identifying the sequences of the short fragments (i.e., reads) is possible by measuring the intensity of the signals in the image of each cycle. Unlike with Sanger sequencing, the use of Illumina sequencing allows for a whole bacterial genome and a complex bacterial community to be sequenced in a reasonable amount of time, and at a lower cost per gigabase (Gb). However, the fragmentation of DNA provides a shattered image of the input DNA. This requires computational analyses to evaluate the reads, which might be troublesome, particularly when one is dealing with repetitive DNA regions (e.g., integrase genes and gene cassettes). Assembling such complex regions of genomes without a reference sequence could result in chimeric contigs. In this thesis, we used miSeq® Illumina sequencing technology (producing 2×350 bp reads) for the whole-genome sequencing of a *Pseudomonas aeruginosa* isolate and also to correct long reads generated by SMRT sequencing technology.

3.5.3 SMRT SEQUENCING

Single-molecule real-time technology is a sequencing-by-synthesis technique offered by Pacific Biosciences (PacBio) (Eid et al., 2009). It uses four dNTPs and a nanophotonic structure called the zero-mode-waveguide (ZMW), which can detect a single fluorophore in real-time during DNA synthesis. The sequencing starts with library construction that creates closed circular DNA (SMRTbell) by ligating hairpin adaptors at the ends of the input DNA. Then, the SMRTbell and the adaptor bind to a fixed DNA polymerase at the bottom of ZMW, which is targeted with a laser beam from below. As the dNTPs binds to the DNA template, it diffuses the recognizable fluorophore, which is interpreted as the corresponding nucleotide. The PacBio sequencing offers different platforms, including RS II and Sequel. The latter contains 150,000 ZMWs, capable of producing up to 1 Gb of data per SMRT cell in less than six hours. Meanwhile, the former uses one million ZMWs, capable of producing up to 10 Gb per SMRT cell in the same amount of time. Both platforms provide long reads with an average length of over 10 kb. In this thesis, we have used both RS II sequencing (Paper I) and Sequel (Papers II and III).

The PacBio sequencing is not biased by the GC content and high-repeat regions, as are some other sequencing technologies. The resulting long reads enable us to study longer genetic contexts, such as integrons that are difficult to assemble accurately using only short Illumina reads (Roberts et al., 2013). However, the high error rate is the main drawback of PacBio sequencing. A single long read can reach up to 15% random errors. To overcome the problem, self- and hybrid-correction of PacBio reads are advisable.

3.6 ANALYZING SEQUENCING DATA

3.6.1 CORRECTING LONG READS

The self-correction of PacBio reads relies on the redundancy of the sequenced DNA templates in the final outputs. In PacBio, a DNA template is sequenced several times and the consensus is used to detect random insertions and deletions (indels) of nucleotides. This strategy could correct over 99% of indels of a single DNA template (i.e., not the entire dataset), but it might not be applicable if the coverage of the long reads is low (Eid et al., 2009).

In Papers II and III, we used a self-correction approach to identify novel ARGs. The output of PacBio Sequel sequencing provided us with the consensus of the sequenced integrons. Moreover, we utilized the redundancy of gene cassettes that resulted from the functional screening of the transformants. Resistant clones containing proper ARGs were selected based on their corresponding antibiotic selective plates, leading to an increased abundance of reads and, in turn, gene cassettes. This, therefore, helped us not only detect novel ARGs but also create consensus sequences of gene cassettes (not the entire read) and detect possible skipped indels.

The hybrid-correction utilizes the accurate short reads produced by Illumina sequencing technology to detect indels in long reads of PacBio (Fu et al., 2019; H. Zhang et al., 2019). It is divided into two broad methods: graph-based and alignment-based. The former constructs a de Bruijn graph from a set of *k*-mers. Then, it tries to find the best Eulerian path (i.e., a path visiting each edge exactly once) that matches the long read. However, the latter maps the short reads to the long reads and computes the consensus. Recently, new algorithms that use a combination of these two approaches have been proposed (Bao & Lan, 2017; Haghshenas et al., 2016).

In Paper I, we evaluated the output of three hybrid-correction methods: LoRDEC (Salmela & Rivals, 2014), LSC (Au et al., 2012), and Proovread (Hackl et al., 2014). LoRDEC identifies solid regions on long reads that match the frequent *k*-mers. Then, by traversing the De Bruijn graph calculated by

short reads, it finds the bridge paths that connect solid regions. It has a reasonable running time, but it trims the reads. We could also find indels especially at homopolymer regions. LSC is an alignment-based method that initially incorporates homopolymer compression on both short and long reads. Then, it concatenates the long reads to get a chromosome-size sequence, and, by mapping short reads to long reads, it corrects the errors. Finally, LSC decompresses homopolymer regions on the long reads. LSC failed to generate correct reads due to the long running time, which has also been recently reported from a benchmark experiment (H. Zhang et al., 2019). Proovread is also an alignment-based method that, through an iterative correction strategy, finds the consensus of mapped short reads and corrects the long reads. It uses an alignment scoring scheme customized for the PacBio error rates, in which substitution, deletion, and insertion have 1%, 5%, and 10% error rates, respectively. It can find the chimeric breakpoints on fused reads and also reports the Phred quality score for each nucleotide. Proovread utilizes high-performance alignment tools such as bowtie2 and SHRiMP2, which, in turn, provide a reasonable running time. In Paper I, we therefore chose to use Proovread to correct long reads and utilized the corresponding quality scores in downstream analyses.

3.6.2 ANNOTATING DNA SEQUENCES

Prodigal – predicting open reading frames (ORFs)

Identifying ORFs on the input DNA sequences is among the first steps in our annotation pipeline. In this thesis, we used Prodigal to predict ORFs on the studied DNA sequences (Hyatt et al., 2010). It is based on general rules that are identified by the study of almost 100 bacterial genomes in detail. The gene size, GC frame bias model, hexamer coding statistics, maximum overlap between two genes, and motifs of ribosomal binding sites (RBS) are among those rules that could be tuned by initial analyses of the input sequences. Prodigal uses dynamic programming for the training and gene calling phases. Dynamic programming is a class of algorithms that results in the best solution by transforming a complex problem into overlapping sub-problems and then optimally solving them. Prodigal considers valid starts and stop codons in each frame as building blocks (i.e., sub-problems) for finding ORFs. It uses different scoring schemes based on log-likelihood functions, bonuses, and penalty scores to assess different rules over intermediate ORFs and, finally, finds genes, intergenic space, and overlapping genes on the input sequences.

BLAST – basic local alignment search tools

To identify the functionality, predicted ORFs are searched against known protein/DNA databases using the BLAST algorithm (Altschul et al., 1990). This is a heuristic method for local alignment, with a seed-and-extend approach. Initially, the reference database is broken down into shorter sequences (i.e., words or seeds) stored in a lookup table. Next, the BLAST algorithm tries to find the seeds on the query sequences (seed finding) and connects them (extend) by using the reference sequences. Through the connecting of seeds, an alignment containing matches, mismatches, or gaps is produced between the two sequences. Also, through the use of a defined scoring matrix (e.g., BLOSUM or PAM), the similarity score of the sequences is calculated. The BLAST+ package is a well-known tool implemented BLAST algorithm (Camacho et al., 2009). In this thesis, we mostly used Diamond, a BLAST-like algorithm, which has significantly improved the speed through the better use of memory hierarchy (i.e., Disk, RAM, cache in CPU) and reduced alphabets, as well as through the use of longer seeds (Buchfink et al., 2015).

Probabilistic sequence alignments

Hidden Markov models (HMMs) have been used extensively for sequence alignment and for finding motifs on genomes. The basis for HMM is the Markov chain that strongly assumes that the prediction of the future in a sequence of events depends only on the present, and not the past. HMMs are probabilistic automatons consisted of three key features: a set of states, input alphabets, and a transition function (Rabin, 1963). The nucleotides/amino acids (i.e., alphabets) in multiple sequence alignments comprise *observed* states that are connected sequentially and that also have connections to *hidden* states representing insertions and deletions. The probabilities of transitions between states are calculated from a training set using the Baum-Welch algorithm. The likelihood of a particular sequence matching the profile of alignment is reported by traversing the HMM model and computing the joint probability using the forward algorithm (Durbin et al., 1998).

In this thesis, we used the HMMER package to create models of protein domains and to search them against input sequences to find distantly related homologues (Mistry et al., 2013). The *hmmbuild* program obtains a multiple sequence alignment as a training set and creates the HMM profile. The *hmmsearch* program accepts input HMM profiles and searches them against the query protein sequences. Moreover in Paper I, we used HattCI to detect the *attC* sites of integrons (Pereira et al., 2016). HattCI is an HMM design based on probabilistic context-free grammars. The grammars define a motif

representing a secondary structure of gene cassettes at the recombination site. The HMM grammars and training are based on a set of manually curated known *attC*s. In Papers II and III, we used IntegronFinder, which uses HMMs to detect the integrase genes and utilizes covariance models to identify *attC*s (Cury et al., 2016). In the latter, IntegronFinder adopted the RNA folding prediction (Eddy & Durbin, 1994). It uses a tree-based HMM that incorporates the following states: matching (i.e., pair of nucleotides or left/right bulge loops), deletion, and insertion.

Reference databases

Biological databases are collections of structured, indexed biological data that can be easily accessed and updated. They can be stored in different formats, from flat files (e.g., fasta, fastq, JSON, etc.) to relational databases managed by different softwares (e.g., SQLite, MySql, etc.). The National Center for Biotechnology Information (NCBI) and the European Molecular Biology Laboratory (EMBL) are two organizations that host biological databases and provide a range of tools (e.g., BLAST) for data retrieval, visualization, and entry. In this thesis, we used the following databases: NCBI assembly, GenBank, non-redundant protein and nucleotide collections, conserved domain database (CDD), PFAM, taxonomy, sequence read archive (SRA), EBI MGnify database, biocide and metal resistance database (Bacmet) (Pal et al., 2013), and virulence factor database (VFDB) (Chen et al., 2005).

Collections of ARGs served as important biological resources in the current thesis. We used two important ARG databases: the comprehensive antibiotic resistance database (CARD) (Alcock et al., 2020) and ResFinder (Zankari et al., 2012). The former is a set of DNA and protein sequences as well as tools such as antibiotic resistance gene ontology (ARO) and resistance gene identifiers (RGI) that enable better detection and analyses of resistome. The latter contains mainly DNA sequences of acquired and chromosomally mutated ARGs in separate flat files. Moreover, there are two Python wrapper functions that find these groups of ARGs in input sequences. Because CARD contains intrinsic and chromosomal ARGs (e.g., universal efflux pump *mexAB-oprM* or chromosomally mutated genes in PhoPQ system), we could better characterize clinically resistant strains that have, for example, chromosomally mutated genes by analyzing the entire genomes or specific parts of them. However, this mixed collection could produce overestimated abundances of ARGs in metagenomic analyses due to the false assignment of short reads that belong to genes with no resistance functions, to mutated resistance genes.

Sequence clustering

Clustering is an unsupervised learning method that groups data items. The groups (i.e., clusters) have higher intra-similarity than inter-similarity (Zadegan et al., 2013). Clustering DNA or protein sequences is an important step that helps identifying groups of homologue genes/proteins and also removes redundancy in the input datasets to reduce the computational burden of downstream analyses. The complexity of sequence clustering does not enable the finding of optimal clustering, e.g., through dynamic algorithms; instead, greedy algorithms are used to find the sub-optimal solutions. In this thesis, we used CD-HIT to cluster DNA and protein sequences (W. Li & Godzik, 2006). It uses a short word filtering method that estimates the overall identity by comparing short substrings in a pair of sequences. The clustering algorithm involves sorting the input sequences by length and iteratively comparing them to the representative (i.e., seed) of clusters. CD-HIT provides parameters to control the alignments (e.g., alignment coverage, local/global sequence alignments, etc.), DNA strands, memory usage, multi-processing, and output formats.

Phylogenetic tree

Phylogenetic trees show evolutionary relations of organisms, genes, or proteins (Gabaldón, 2005). They are a form of hierarchical clustering in which the similarity between data items is calculated based on sequence alignments. The tree could be constructed by a range of algorithms, from a simple bottom-up agglomerative clustering (e.g., UPMGA algorithm) to a maximum likelihood (ML) approach (Strimmer, 1997). In the latter, the probabilistic model describes how an ancestral sequence has evolved into other sequences. It incorporates parameters like tree topology, branch length, nucleotide/amino acid frequencies, and mutation rates. The probability of a given set of parameters and an input multiple sequence alignments determine the phylogenetic tree with the highest likelihood. For purposes of assessing the accuracy of the tree, it could be coupled with bootstrapping methods in which many trees are constructed from permutated input sequences (Efron et al., 1996). It produces a collection of tree topologies based on random sequences, which are compared with the original tree and which statistically calculate the confidence values for different clades.

In this thesis, we used MAFFT (Katoh et al., 2002) to create multiple sequence alignments and then FastTree to build the phylogenetic tree (Price et al., 2009). FastTree uses an ML approach with heuristics to simplify the problem and solve it in a reasonable time. The topology of the tree is initially created and improved by simpler algorithms like neighbor-joining and nearest-neighbor

interchanges. FastTree also uses substitution models like Jukes-Cantor and Jones-Taylor-Thornton for nucleotide and amino acid mutation rates, respectively. In this thesis, the visualization of trees is performed using Python packages like the ETE toolkit (in Paper I) (Huerta-Cepas et al., 2016) or webservers like iTOL (in Papers II and III) (Letunic & Bork, 2016).

3.6.3 ANALYZING METAGENOMES

Quality control

Checking the quality of short reads is the first step in analyzing metagenomic datasets. Various incidents, such as a disturbance in fluorescent signals, air bubbles on the chip surface, or a deficiency on sequencing chips, could create low-quality reads. Different software packages generate quality control reports and filter and trim single/paired-end reads to produce high-quality datasets. We used FastQC to assess the quality of the reads (Andrews, 2010). HTQC software (Paper I) (Yang et al., 2013) and Trim Galor software (Krueger, 2015) (Paper II) were used for filtering and trimming short paired-end reads.

Mapping short reads

To quantify the abundance of the reference genes/proteins (e.g., ARGs or ISs), we mapped the short reads to them using USEARCH (Paper I) (Edgar, 2010) and Diamond (Papers II-IV). Because the reads could potentially map to several highly similar reference sequences (e.g., *bla*TEM family or *bla*OXA family), it was important to adopt a policy to avoid over-estimations of their abundances. In Papers I-III, we searched the identified novel ARGs in different metagenomic datasets. By using high amino acid identity thresholds (100% for *sul4* and *gar*, and 95% for the *bla*IDC family), we tried to reduce incidences of the false assigning of short reads. In Paper IV, we clustered the reference proteins (i.e., ARGs and ISs) with a 90% identity threshold. Then, the representative of the clusters was searched against metagenomic datasets. In this way, the redundancy in the reference datasets was reduced and reads would be less likely to match several proteins.

Normalization

Metagenomic datasets contain systematic variability originating from the type of samples, DNA extraction techniques, sequencing platforms, and their resulting depth of sequencing (Jonsson et al., 2017; Pereira et al., 2018). Through normalization of the gene/protein count values, the effect of between-sample variability is diminished. Normalization by the total number of reads is the easiest way to reduce the biases. To make normalized values more readable,

they are multiplied by one million and reported as count values per million reads. However, the metagenomic sample could contain DNA from other branches of life, such as viruses, fungi, or eukaryotes, which could be worth considering when one is interpreting the abundances of genes. Hence, the abundance of ARGs in metagenomes is often normalized by the abundance of 16S rRNA genes that represent only bacterial genomes. In Papers I-III, we were interested in the presence of the novel ARGs in the metagenomes, which does not require normalization of the raw count values. In Paper IV, we normalized the ARGs and ISs by the total reads in each metagenomic dataset.

3.6.4 ASSEMBLY OF SHORT READS

The assembling of short reads produces larger continuous sequences called contigs, which could fully or partially reveal the genetic contexts of the recovered DNA (Khan et al., 2018). Reference-based assembly is guided by reference genomes, while de novo assembly produces contigs without any genomic reference. The assembler could use greedy, overlap-layout-consensus (OLC) and the de Bruijn graph approach (Lin et al., 2016). The greedy algorithm joins short reads with the best overlaps to create a longer sequence. The OLC approach finds the overlapped reads, creates a directed graph called layout, aligns all the relevant reads in the layout, and, finally, reports the consensus. The de Bruijn graph is based on the *k*-mers approach, in which the short reads are split into sequences of length k and the de Bruijn graph is created based on the head/tail nucleotide overlaps of length k-1. Then, Eulerian paths, which visit every edge exactly once, create longer contigs. In this thesis, we used two de novo assemblers—called SPAdes (Bankevich et al., 2012) and MEGAHIT (D. Li et al., 2015)—that employ the de Bruijn graph approach to assemble whole genome sequencing and metagenomic datasets, respectively.

3.7 GENE SYNTHESIS AND RECOMBINANT EXPRESSION

Gene synthesis is a chemical construction and assembling of a nucleotide sequence outside of a living cell. It starts with the elongating of modified nucleotides (i.e., nucleoside phosphoramidites) to form a short single-stranded DNA fragment called an oligonucleotide. Each modified nucleotide is incrementally added to a nano-well in which the growing chain of the oligonucleotide is fixed. In the next step, the overlapped oligonucleotides are connected. Then, through use of polymerase reactions, a double-stranded DNA of the input gene is produced. In this thesis, the candidate novel ARGs were

synthesized at Thermo Fisher Scientific, Germany, using its GeneArt Gene Synthesis service.

To assess the functionality of a gene, it is ligated into an expression vector. These vectors have several key features: the replication origin that allows the vector to multiply, selectable markers like ARGs that ensure maintenance of the vector by the bacterial host in the selective medium, cloning sites in which the gene of interest could be inserted, and, more importantly, the necessary transcriptional elements, such as promoter and, in some vectors, inducible promoter. In this thesis, the pZE21-MCS1 vector was used to express the candidate genes. It has a promoter that can be induced up to 5000-fold in the presence of anhydrotetracycline. The selective marker in this vector is kanamycin resistance genes, which might cause a problem in terms of testing candidate aminoglycoside resistance genes. Therefore, the PUC19 vector containing a beta-lactamase gene was also used. To test the antibiotic susceptibility, the vectors were transformed into different *E. coli* strains, like C600Z1 and BL21(DE3), by electroporation.

The resistance spectrum of bacteria is identified by agar plates or broth dilution, in which bacteria are cultured and exposed to gradients or serial dilutions of antibiotics. In the former, solid agar media (e.g., Mueller-Hinton) are streaked with bacteria containing the expression vectors. Then, the disk of antibiotics with different concentrations is placed to later measure the radius around them where bacteria could not grow. Alternatively, plastic strips are used on the agar, with each strip containing a predefined gradient of antibiotic (E-Test®) (Balouiri et al., 2016). In broth dilution, liquid media with different antibiotic concentrations are prepared and bacteria are cultured within them. The radius of the clean area around the disks or the concentration of antibiotics (i.e., in E-Test® or broth) determines the minimum inhibitory concentration (MIC). Based on the MIC, bacteria can be classified into the following groups: susceptible (i.e., inhibition of bacteria in vitro with a high certainty of therapeutic success), intermediate (i.e., an uncertain therapeutic effect despite the inhibition of bacteria in vitro), and resistant (i.e., a high likelihood of therapeutic failure). The MIC thresholds for each group related to different clinical bacterial species and antibiotics have been collected and defined by e.g., EUCAST (EUCAST, 2013).

4. RESULTS AND DISCCUSSION

In this thesis, we employed different strategies to recover novel mobilized ARGs. Previous high-throughput strategies such as shotgun or functional metagenomics could rarely address mobility due to challenges in providing the genetic context of identified ARGs. In some cases, such methods can recover fragments of DNA with indications of mobile genetic context such as the recognizable regions of a plasmid; however, revealing the mobility was not the purpose of their design. Our main contribution was to utilize the benefits of previous methods and, at the same time, address the mobility as an important risk factor for identifying novel ARGs that might accumulate in pathogens.

4.1 EXPLORING INTEGRONS WITH TARGTED AMPLICON SEQUENCING

In Paper I, we used a targeted amplicon sequencing method to identify novel ARGs in the form of gene cassettes. The high-throughput, cost-efficient metagenomic approach was focused on the content of integrons. In this way, the depth of analyses was increased as we discarded other genetic materials that were recovered from our environmental samples. This was the first high-throughput analysis of integrons, which could recover 19,723 unique gene cassettes (see table 2 in Paper I). We were able to not only discover a novel sulfonamide resistance genes, *sul4,* but also to prove its mobile contexts. Additionally, we identified novel putative ARG variants of aminoglycosides, beta-lactams, trimethoprim, rifampicin, and chloramphenicol. However, this approach limited us to detecting only homologues of known ARGs. Considering the number of ARGs in these polluted environments (Figure 1, Paper I) and the role of integrons in accumulating gene cassettes with a significant adaptive response, we hypothesized that some of the unknown gene cassettes could exhibit novel resistance mechanisms. In Paper II, we therefore screened gene cassettes recovered from metagenomic samples with a phenotypic assay against 13 different antibiotics at three different concentrations. Through this approach, we discovered a completely novel aminoglycoside resistance gene named *gar*. The protein encoded by *gar* contains a domain (i.e., P-loop NTPase) that was not recognized in ARG databases, including CARD and ResFinder. Therefore, despite its presence in the previous collection of gene cassettes (in Paper I) and even in the GenBank database, its function as a resistance gene had remained unknown. This highlights the benefits of functional metagenomics in discovering novel resistance functions. However, in comparison to the metagenomic approach, it requires more experimental efforts for library preparation and cloning the gene

cassettes. The success rate of discovering novel ARGs in both approaches is also dependent on the sample and the targeted amplified regions (e.g., integrons or TEs).

The selective pressures in the environments dictate the significance of gene cassettes and subsequently contribute to their emergence and abundance. For instance, the profile of gene cassette functions recovered from the Sydney Tar Ponds is quite different from that of our Indian samples. These sites in Sydney are contaminated by petroleum toxic agents and heavy metals from steel production processes (Koenig et al., 2009). Around 22% of the recovered gene cassettes encoded proteins with identifiable functions, such as *lysR* (i.e., which regulates the catabolism of pollutants like aromatic compounds), heavy metal resistances (e.g., mercury resistance genes), and proteins involved in transportations (e.g., ModA, benzoate transport proteins, etc.). In contrast, almost 89% of gene cassettes recovered in Paper I encode proteins with resistance functions. This is most likely due to the antibiotic selective pressures in the environment (PETL) or in the human population that contributed bacterial DNA to the sediment samples (Pune). Hence, applying these methods to complex bacterial communities with a history of antibiotic exposure could hugely increase the chance of discovering novel ARGs. Moreover, integrons are well-studied genetic systems that have a strong association with ARGs in such environments. However, the relation of other potential targets, including specific TEs with ARGs, has not been thoroughly investigated. In Paper IV, we sought to expand our knowledge about abundant TEs in bacterial genomes and propose novel candidate targets beyond integrons for use in the search for novel, mobile ARGs.

4.2 CHARACTERISTICS OF NOVEL ARGS

The discovered *sul4* is a relative of the *folp* gene in the folate biosynthesis pathway but, like other sulfonamide resistance proteins, it has a lower affinity to sulfonamides molecules than the native protein. By expressing *sul4*, the production of dihydropteroate is not interrupted by the antibiotic, which leads to the resistance phenotype (Sköld, 2000). It has less than 33% identity with the previous mobile sulfonamide resistance genes *sul1*, *sul2*, and *sul3*. The phylogenetic tree of *sul4* and other chromosomal *folp* genes suggested that it was mobilized from a bacterium in the Phylum *Chloroflexi*. Sulfonamides form one of the antibiotic families that, nowadays, is used in animal farming (Agency & Consumption, 2017; Suzuki & Hoa, 2012) and human medicine in combination with trimethoprim (Church et al., 2015). We hypothesize that the emergence of *sul4* could further diminish the effect of sulfonamides and,

thereby, contribute to the shifting of the use of sulfonamides to other antibiotic classes, such as tetracyclines and macrolides. This could lead to developments of resistance determinants against such antibiotics which are more important in human medicine. Moreover, a stronger dose of sulfonamide might be used to overcome the extended resistance, which, in turn, would increase the excretion of sulfonamide and its toxic metabolites into the environment (García-Galán et al., 2012; Ou et al., 2015).

In Paper II, we discovered *gar*, which conferred resistance to several aminoglycoside antibiotics, including gentamicin. Its resistance profile against different sub-families of aminoglycosides (Figure 1 in Paper II) and the presence of specific motifs (Walker A NTP-binding and two DxD motifs, see Figure 2 in Paper II) suggest that GAR modifies garosamine-specific aminoglycosides. By comparing the structures of the antibiotics (Table S2 in Paper II), we hypothesized that GAR adds a phosphoryl group to the 4" carbon atom, which has not previously been reported as a target. The phylogenetic tree of GAR, its close homologs, and known aminoglycoside resistance enzymes showed that GAR represents a novel sub-family within aminoglycoside resistance proteins. The emergence of *gar* could expand the resistant spectrum of bacteria against aminoglycosides, particularly diminishing the effect of garosamine-containing antibiotics such as plazomicin, which has recently been developed.

In Paper III, two novel class C beta-lactamases, named *bla*IDC-1 and *bla*IDC-2, were discovered. Both genes provide resistance against penicillins, clavulanic acid, second- and third-generation cephalosporins, and monobactams (Tables 1 and 2 in Paper III). The closest homologs from the CARD database are *bla*LRA-18 and *bla*LRA-10, which share around 55% of the amino acid identity. The phylogenetic trees of these genes, together with all the known beta-lactamases, showed that *bla*IDC represents a sub-group within class C beta-lactamases.

4.3 MOBILIZATION OF NOVEL ARGS

Integrons are present in many bacterial species and have associations with various TEs on horizontally transferable elements, such as conjugative plasmids and ICEs. The identified novel ARGs (i.e., *sul4*, *gar*, *bla*IDC) are among the first of their kind to be discovered as gene cassettes. They are an accessible form of ARGs that could be integrated into virtually all integrons that are widespread on bacterial genomes, in response to a changing environment. Reshuffling gene cassettes could create an optimal and stable

array of different adaptive functions, with the potential to be mobilized and co-selected under various antibiotic selective forces.

In Paper III, we discussed the presence of the *bla*IDC group in the form of gene cassettes in comparison to chromosomal and IS-mediated *ampC*s, including how it leads to an adaptive, cost-efficient expression, which, in turn, imposes a higher risk of transfer to pathogens. Similarly, integrons provided *gar* with the same mode of genetic mobility, co-selection, and expression opportunities. The *gar* gene has been found in two different pathogenic bacteria, and in arrays of gene cassettes in conjunction with ARGs providing resistance against beta-lactam (e.g., *bla*Vim-1, *bla*OXA-2), chloramphenicol (*catB3*), and aminoglycoside (*ant*(3")-la, *aac*(6')-lb) antibiotics.

The previously identified mobile sulfonamide resistance genes (i.e., *sul1*, *sul2*, and *sul3*) have not been found in the form of gene cassettes. The *sul1* gene is, in most cases, located downstream of *qacE*Δ at the conserved segment (3'-CS) of class 1 integron. Hence, unlike gene cassettes, it cannot be excised, integrated, or reshuffled in response to changing environments. The *sul2* gene is usually located on small non-conjugative plasmids and has associations with insertion sequences, including IS*6* and Tn*3* (Byrne-Bailey et al., 2009). Finally, *sul3* is identified on conjugative plasmids but, like *sul1*, it is located downstream of integrons in associations with *qacH* and insertion sequences such as IS*26* and IS*440* (Antunes et al., 2007). In contrast, we have identified *sul4* as part of a larger unit together with a partial *folk* gene, both as a gene cassette and in association with IS*CR20* outside of integrons (Figure 2 (c) and (f) in Paper I). Recent data have supported the hypothesis that *sul4* is mobilized both as an integron gene cassette and by insertion sequences. After the publication of Paper I, we have found *sul4* within an integron located on a plasmid of a *Sphingobium yanoikuyae* strain (GenBank: CP033227.1) and also on a chromosome of a *Moraxella osloensis* strain (GenBank: CP040257.1). The latter has no integrase gene and *attC* sites around the immediate genetic contexts of *sul4*. The precedence of these mobilization events is not clear, nor is it clear whether the ISs moved *sul4* into or from integrons. However, its presence in taxonomically distant bacteria (i.e., *Chloroflexi* Phylum, *Gamma-* and *Alpha-proteobacteria*) reveals the ample opportunities of spreading via integrons and ISs.

4.4 ABUNDANCE IN METAGENOMIC DATASETS

Searching metagenomes helped us estimate the spread of the novel ARGs in different environments and geographic locations. The most and least abundant genes in the metagenomic datasets were *sul4* and *gar*, respectively. The *sul4*

gene was detected in diverse environments including wastewater/sludge, river sediments, air, and freshwater in seven different countries. The metagenomic samples from algal blooms in Kolkata, India had a unique profile of *sul* genes, in which only *sul4* was present. Given that *sul1-3* genes are indicators of fecal contamination, and given the dominance of phylum *Chloroflexi* in alga bloom, we hypothesized that *sul4* was decontextualized from its original host belonging to phylum *Chloroflexi* in such an environment. The *gar* and *bla*IDC genes were abundant mostly in wastewater/sludge environments, and not in metagenomes from human individuals. This could be explained by sewage metagenomes that represent a collection of many thousands of individual fecal metagenomes. Genes that are completely absent in the vast majority of individuals would, hence, be much more likely to be detected in sewage metagenomes and not in samples from individuals. Moreover, while *gar* was detected in geographically distant locations, including Asia, Europe, Australia, and Africa, the *bla*IDC family was found primarily in Asia.

4.5 RISKS ASSOCIATED WITH THE NOVEL ARGS

The three identified novel ARGs are mobilized in the form of gene cassettes with the potential to be widely spread among diverse bacterial species. Integrons enable an adaptive expression mechanism by reshuffling and by incorporating promoters with different levels of strength. These mechanisms could also enable an accessible integration of resistance gene cassettes or their maintenance on bacterial genomes through co-selection. This could be interpreted as a mechanism for tuning the expression of the novel ARGs in different bacterial hosts and bypassing one of the bottlenecks of reaching human pathogens. Moreover, the abundance of the identified ARGs in metagenomes from wastewater/sludge environments likely reflects their hosts' opportunities to interact with human-related bacteria and possibly share genetic material through HGT mechanisms (i.e., ecological connectivity).

Among the identified novel ARGs, *gar* has been found in several bacterial hosts, including *P. aeruginosa* extracted from a patient in a rehabilitation facility in northern Italy, *Lysobacter oculi sp*. isolates recovered from human Meibomian gland secretions in China, and *Salmonella enterica* from poultry products in the USA. The *P. aeruginosa* isolates (sequence types ST235 and ST111) are multi-resistant bacteria that are distributed worldwide and that cause nosocomial infections with high mortality rates. The *Lysobacter* isolate (i.e., it was initially classified as *Luteimonas* because they share a high taxonomical similarity (Bai et al., 2020)) thrives primarily in soil and plants, but also sludge reactors. Members of *Lysobacter* could inhibit the growth of

plant pathogens and have the potential for use as biological control agents (Folman et al., 2003). *Salmonella enterica* is an important human pathogen that could colonize and infect the intestinal tract of humans and domestic animals. The presence and dissemination of *gar* among human commensals and pathogens have imposed an apparent risk on our health. Though *gar* is not active against all aminoglycosides (including amikacin), it could limit our treatment options by expanding the resistance spectrum and forcing us to use more antibiotics of last resort against multi-drug-resistant bacteria like *P. aeruginosa*.

To date, *sul4* has been found on the chromosome of *Moraxella osloensis* and a plasmid recovered from *Sphingobium yanoikuyae* isolate. The former is an opportunistic human pathogen that could be recovered from our natural flora in the skin, mucus membranes, and respiratory tract (Shah et al., 2000). A few case reports suggest that *Moraxella osloensis* could cause infection in humans, such as endocarditis, osteomyelitis, and meningitis (Hadano et al., 2012). It could live in hospital environments, including anesthetic agents and sink traps, which increases its opportunity to interact with other human pathogens. Moreover, *Moraxella osloensis* is found in symbiosis with a nematode parasite of slugs (Tan & Grewal, 2001). Both are used as biological controls for the grey garden slug, which is a significant agricultural pest.

Members of the genus *Sphingobium* are known to degrade and utilize various pollutants (e.g., biphenyls, chloroxylenol, naphthalene, etc.) as a source of carbon (Choi & Oh, 2019; Y. Wang et al., 2018). The abundance of such pollutants in wastewater/sludge environments provides an increased chance for these bacteria to thrive and may have acquired *sul4* by interacting with other bacteria in such an environment.

Moreover, a PCR-based approach has recently been used to detect *sul4* in clinical isolates (Xu et al., 2020). Two isolates—a *Salmonella spp.* and an *Escherichia coli*—were reported to harbor *sul4*. However, with the amplification of only 25% of *sul4,* the genetic contexts are still unknown. More analyses, including sequencing of the whole isolates or part of the isolates, are required to first confirm the presence of *sul4* and then characterize its genetic contexts. Nevertheless, the direct human health risk of an acquired *sul4* gene might not be as high as that with *gar*, as sulfonamides are rarely or never the only treatment option. Still, *sul4* could redirect the use to other classes of antibiotics, particularly in animal sectors, and, thereby, indirectly poses a risk to human health.

So far, we have not found a bacterial host for the identified *bla*IDC genes from the recovered genetic materials and the repositories of sequenced bacterial genomes. Subsequently, it is difficult to estimate the associated risk to our

health. However, we know that if these genes are acquired by human pathogens, it could clearly limit our treatment options in many cases. Furthermore, it is not unrealistic that it could develop mutations to inactivate carbapenems under the clinical selection pressures of antibiotics, as has been the case with several other cephalosporinases.

4.6 INSERTION SEQUENCES – NOVEL TARGETS

In Paper IV, we studied the mobile genetic contexts beyond integrons with the long-term intent of utilizing the knowledge to discover novel mobile ARGs. Insertion sequences with DDE domains are abundant transposases that have substantial variability in the overall protein sequences on bacterial genomes. They could mobilize individual ARGs and even entire integrons. There is a knowledge gap in their association with ARGs and whether they can recover novel ARGs in the form of composite transposons.

The genetic context around ARGs and ISs on sequenced bacterial genomes showed that ARGs have formed clusters on bacterial genomes and have strong associations with ISs. On the other hand, ISs have weaker associations with ARGs as ISs having other roles in bacterial genomes than mobilizing ARGs. This prompted us to analyze genomic and metagenomic data to measure the association between individual ISs and ARGs and to rank their importance in terms of mobility and abundance in relevant environments.

The association of ISs with ARGs could be falsely identified as strong because of their presence in short bacterial contigs that have been deposited in the GenBank database or a high copy number of an IS on a particular genome. To remove such anomalies, we have statistically assessed these associations on all the genomes using a permutation test and then calculated the significantly associated ARGs for each IS. Moreover, tentative composite transposons were defined and the variability of their genetic contexts was measured. The abundance of ISs and ARGs was calculated in 1891 metagenomes representing 12 different environments. We showed that the strong associations between ISs and ARGs that were identified from genome data are maintained in complex bacterial communities.

Insertion sequences that have strong associations with ARGs, with varying composite transposons, are among the best candidates. Because ISs mobilize many other genes, it is important to consider proper samples for an approach involving amplicon sequencing. Moreover, PCR-based approaches are often biased toward amplifying shorter products. This might create difficulty in recovering longer composite transposons potentially containing novel ARGs.

4.7 ASSOCIATIONS REPOSITORY OF ISS AND ARGS (ARIA)

The approaches developed through the analyses in Paper IV provided opportunities to study the genetic contexts and mobilization mechanisms of ISs and ARGs. Because the study of such associations can be valuable for many different purposes, we have an ambition to make the results of analyses accessible to everyone in a user-friendly application. Hence, we are currently working on an open-access web-resource, tentatively called ARIA, to facilitate the search for, and retrieval of, information (Figure 1). Users will be able to search the occurrence and contexts of ISs and ARGs in bacterial genomes and filter their results based on different criteria, such as the sequence identity and taxonomical rank of the bacterial hosts. They will also be able to retrieve the frequency of the putative functions of genes (i.e., including biocide and metal resistance genes, virulence factors, integrons, and other transposons) around ARGs and ISs. In addition, it will be possible to retrieve information from the ISfinder, CARD, PFAM, and NCBI GenBank databases. With a novel sequence viewer, users can browse the genetic contexts of ISs and ARGs on different bacterial hosts, search tentative composite transposons, and use the extensive filtering options to see and download the genetic contexts of interest. ARIA processes the raw sequencing data from the NCBI GenBank database and integrates it with the information in reference databases to create knowledge useful for studying ARGs and ISs.

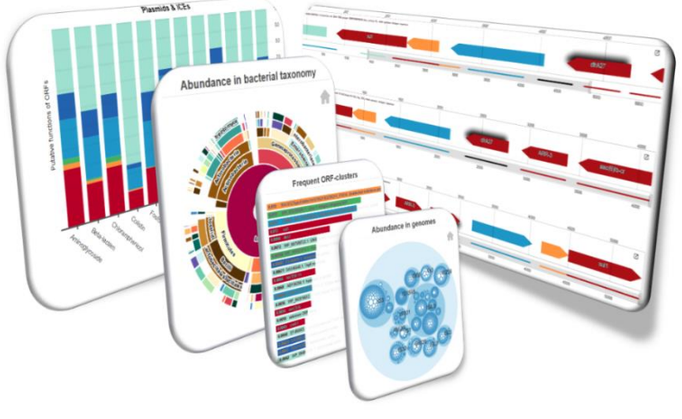


Figure 1. Snapshot of different web pages of ARIA (pilot version). The combination of interactive plots and extensive filtering options enables efficient retrieval of information about ISs and ARGs.

5. CONCLUSIONS

In Paper I, we explored the content of integrons in two polluted environments using a targeted amplicon sequencing approach combined with a homology-based search. We found the first novel mobile sulfonamide resistance gene discovered since 2003, providing a high level of resistance when expressed in *E. coli*. We also identified novel putative ARGs to aminoglycosides, beta-lactams, trimethoprim, rifampicin, and chloramphenicol.

In Paper II, we improved our screening technique by combining integron amplification with functional metagenomics to identify genes without any homology to known ARGs. Through this approach, we identified and characterized a completely new integron-borne aminoglycoside resistance gene. By searching public databases, we found that the gene was already present in multi-resistant clinical isolates collected from patients in Italy, the human microbiome in China, and two food-borne *Salmonella enterica* isolates from the USA. In all cases, it escaped discovery as a resistance gene until now.

In Paper III, we described and characterized a new class C β-lactamase (*ampC*) that is expressed as a gene cassette in an integron. This context has not been described earlier for any *ampC* gene. It is an important finding, as it provides an additional mode of regulation with increased transmission opportunities for this clinically important group of β-lactamases.

In total, several thousand metagenomic runs were analyzed to find the abundance of the identified novel ARGs. These metagenomes were selected because they represent different environments and geographical locations, to estimate the spread of those ARGs. To retrieve various genetic contexts containing novel ARGs, some of the metagenomic datasets were assembled and the resulted contigs were analyzed.

In Paper IV, we studied the genetic contexts around known ARGs, and ISs that are often responsible for the mobilization of ARGs. The significant associations of ISs with ARGs were statistically identified. The composite transposons with varying genetic contexts were identified and the information necessary for assessing their importance in mobilizing novel ARGs was presented. A search of both groups in metagenomic datasets from different environments showed, firstly, the identified association of ISs with ARGs in bacterial communities and, secondly, the abundance of ISs in different environments.

Ongoing work includes web-application software based on data from Paper IV. When implemented, it could empower other researchers to search for ISs,

ARGs, and their genetic contexts in all the sequenced bacterial genomes through user-friendly interfaces.

6. FUTURE PERSPECTIVES

In the future, amplicon sequencing of integrons combined with long-read sequencing technologies could be widely employed to identify mobile genes, including ARGs in the form of gene cassette. The process of cassette formation is an important yet still unanswered question. The more newly formed gene cassettes we identify, the more comprehensive datasets are provided for studying cassette formation by using, e.g., synteny analyses on sequenced bacterial genomes.

Functional metagenomics of amplified gene cassettes have proved to be a successful approach to identifying novel ARGs. This could address both the novelty and mobility of the recovered resistance gene cassettes, without being dependent on similarities to known ARGs. The analysis of metagenomic samples from other environments will provide a clearer picture of the antibiotic resistome and empower us to prepare early confinement actions on novel resistance determinants. However, optimization of the protocols could improve the results. For instance, the use of a cloning vector carrying chloramphenicol resistance genes, and not beta-lactamases, reduces the satellite colonies on the selective medium. In addition, other cloning hosts could be used to bypass the incompatibility of the novel ARGs and *E. coli* cell (Chistoserdova, 2009).

Identification of bacterial hosts carrying novel ARGs could provide us with valuable information for assessing their corresponding risk to human health. So far, both methods mentioned in this thesis have not been able to provide such information. Epic-PCR could recover a fused PCR product of 16S rRNA of the bacterial host and the target gene (Spencer et al., 2016). However, it requires intact bacterial cells and that the target gene is reasonably common. More improvements are needed to efficiently address the bacterial hosts of novel ARGs in the metagenomic samples, or at least to distinguish the chromosomal or mobile genetic contexts (e.g., conjugative plasmids) using markers other than 16S rRNA, like plasmid mobility systems.

In the future, it would be valuable to study the genetic contexts of composite transposons that have strong associations with ARGs, using a well-designed approach. Amplicon sequencing is an option; however, empty composite transposons might be dominant in the final PCR products due to the presence of several copies of ISs in each other's vicinity. This might require the sorting of intermediate PCR products by gel and the extracting of an initial batch of longer products. To acquire a sufficient quantity of the products, it is possible to tag the ends and design new primers that exclusively amplify the newly tagged DNA region. Moreover, Oxford Nanopore Technology (ONT) could be

used to sequence a long stretch of DNA without the trouble of assembling short reads. The accuracy of the reads might be an issue. However, with high coverage of reads or the ORFs (i.e., that appear in different reads), a self-correction method could be employed.

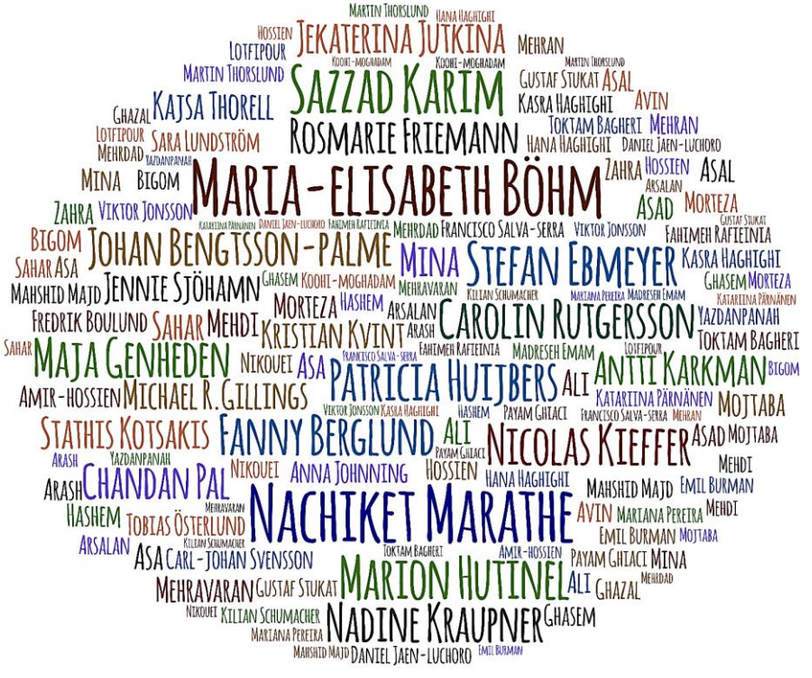
The web application under development will be expanded through improvements to the back-end capacity to handle many requests and through the addition of functionalities to the user interface for more powerful searches. We would like to include other mobile genetic elements, such as ISs containing other transposition domains (e.g., HUH), ISCRs, and metal and biocide resistance genes. Moreover, a pipeline should be implemented to regularly fetch sequenced genomes from NCBI GenBank, annotate them, and update the database behind ARIA. The abundance of ISs and ARGs in metagenomic datasets will be added to provide a full-stack search from bacterial genomes to complex communities.

ACKNOWLEDGEMENTS

Very special thanks to the University of Gothenburg and funding agencies for providing the necessary platform for performing this research.

I wish to express my deepest gratitude to my main supervisor, **Joakim Larsson,** a patient teacher, diligent colleague and generous friend who supported and guided me in accomplishing this work. I am extremely grateful to my co-supervisors: **Erik Kristiansson** and **Carl-Fredrik Flach** for their tremendous support, positive encouragement and helpful critiques.

I would like to thank the co-authors of my manuscripts, my colleagues and friends for their invaluable help, precious advice and heartfelt friendship.



I would especially like to thank my wonderful wife, **Mina**. Her endless support and encouragement helped me to arrive at this point. My parents and brothers deserve my deepest gratitude for their unconditional love, care, and support throughout my life.

REFERENCES

Abella, J., Bielen, A., Huang, L., Delmont, T. O., Vujaklija, D., Duran, R., & Cagnon, C. (2015). Integron diversity in marine environments. Environmental Science and Pollution Research, 22(20), 15360-15369.

Agency, E. M., & Consumption, E. S. o. V. A. (2017). Sales of veterinary antimicrobial agents in 30 European countries in 2015. 178p.

Alcock, B. P., Raphenya, A. R., Lau, T. T., Tsang, K. K., Bouchard, M., Edalatmand, A., . . . Liu, S. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Research, 48(D1), D517-D525.

Allen, H. K., Looft, T., Bayles, D. O., Humphrey, S., Levine, U. Y., Alt, D., & Stanton, T. B. (2011). Antibiotics in feed induce prophages in swine fecal microbiomes. MBio, 2(6), e00260-00211.

Allen, H. K., Moe, L. A., Rodbumrer, J., Gaarder, A., & Handelsman, J. (2009). Functional metagenomics reveals diverse β-lactamases in a remote Alaskan soil. The ISME journal, 3(2), 243-251.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403-410.

Amábile-Cuevas, C. F. (2015). Antibiotics and antibiotic resistance in the environment: CRC Press.

Andersson, D. I., Balaban, N. Q., Baquero, F., Courvalin, P., Glaser, P., Gophna, U., . . . Tønjum, T. (2020). Antibiotic resistance: turning evolutionary principles into clinical reality. FEMS Microbiology Reviews.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. In: Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Antunes, P., Machado, J., & Peixe, L. (2007). Dissemination of sul3-containing elements linked to class 1 integrons with an unusual 3′ conserved sequence region among Salmonella isolates. Antimicrobial Agents and Chemotherapy, 51(4), 1545-1548.

Au, K. F., Underwood, J. G., Lee, L., & Wong, W. H. (2012). Improving PacBio long read accuracy by short read alignment. PloS One, 7(10).

Aziz, R. K., Breitbart, M., & Edwards, R. A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Research, 38(13), 4207-4217.

Bag, S., Saha, B., Mehta, O., Anbumani, D., Kumar, N., Dayal, M., . . . Allin, K. H. (2016). An improved method for high quality metagenomics DNA extraction from human and environmental samples. Scientific Reports, 6, 26775.

Bai, H., Lv, H., Deng, A., Jiang, X., Li, X., & Wen, T. (2020). Lysobacter oculi sp. nov., isolated from human Meibomian gland secretions. Antonie van Leeuwenhoek, 113(1), 13-20.

Balouiri, M., Sadiki, M., & Ibnsouda, S. K. (2016). Methods for in vitro evaluating antimicrobial activity: A review. Journal of Pharmaceutical Analysis, 6(2), 71-79.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., . . . Prjibelski, A. D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19(5), 455-477.

Bao, E., & Lan, L. (2017). HALC: High throughput algorithm for long read error correction. BMC Bioinformatics, 18(1), 204.

Bartlett, J. M., & Stirling, D. (2003). A short history of the polymerase chain reaction. In PCR protocols (pp. 3-6): Springer.

Bengtsson-Palme, J., Boulund, F., Fick, J., Kristiansson, E., & Larsson, D. G. J. (2014). Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. Frontiers in Microbiology, 5, 648.

Bengtsson-Palme, J., Kristiansson, E., & Larsson, D. J. (2018). Environmental factors influencing the development and spread of antibiotic resistance. FEMS Microbiology Reviews, 42(1), fux053.

Bengtsson-Palme, J., & Larsson, D. G. J. (2015). Antibiotic resistance genes in the environment: prioritizing risks. Nature Reviews. Microbiology, 13(6), 396.

Benson, M. A., Ohneck, E. A., Ryan, C., Alonzo III, F., Smith, H., Narechania, A., . . . Sebra, R. (2014). Evolution of hypervirulence by a MRSA clone through acquisition of a transposable element. Molecular Microbiology, 93(4), 664-681.

Bentley, D. R. (2006). Whole-genome re-sequencing. Current Opinion in Genetics & Development, 16(6), 545-552.

Berglund, F., Österlund, T., Boulund, F., Marathe, N. P., Larsson, D. J., & Kristiansson, E. (2019). Identification and reconstruction of novel antibiotic resistance genes from metagenomes. Microbiome, 7(1), 52.

Brenner, S., & Miller, J. H. (2014). Bacterial genetics. In Brenner's encyclopedia of genetics (pp. 162): Elsevier Science.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1), 59.

Byrne-Bailey, K., Gaze, W., Kay, P., Boxall, A., Hawkey, P., & Wellington, E. (2009). Prevalence of sulfonamide resistance genes in bacterial isolates from manured agricultural soils and pig slurry in the United Kingdom. Antimicrobial Agents and Chemotherapy, 53(2), 696-702.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. BMC Bioinformatics, 10(1), 1.

Caselli, E., D'Accolti, M., Soffritti, I., Piffanelli, M., & Mazzacane, S. (2018). Spread of mcr-1–Driven Colistin Resistance on Hospital Surfaces, Italy. Emerging infectious diseases, 24(9), 1752.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. Nucleic Acids Research, 33(suppl_1), D325-D328.

Chistoserdova, L. (2009). Functional metagenomics: recent advances and future challenges. Biotechnology and Genetic Engineering Reviews, 26(1), 335-352.

Choi, D., & Oh, S. (2019). Removal of Chloroxylenol Disinfectant by an Activated Sludge Microbial Community. Microbes and Environments, 34(2), 129-135.

Church, J. A., Fitzgerald, F., Walker, A. S., Gibb, D. M., & Prendergast, A. J. (2015). The expanding role of co-trimoxazole in developing countries. The Lancet infectious diseases, 15(3), 327-339.

Cox, G., & Wright, G. D. (2013). Intrinsic antibiotic resistance: mechanisms, origins, challenges and solutions. International journal of medical microbiology, 303(6-7), 287-292.

Cury, J., Jové, T., Touchon, M., Néron, B., & Rocha, E. P. (2016). Identification and analysis of integrons and cassette arrays in bacterial genomes. Nucleic Acids Research, 44(10), 4539-4550.

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids: Cambridge university press.

Eddy, S. R., & Durbin, R. (1994). RNA sequence analysis using covariance models. Nucleic Acids Research, 22(11), 2079-2088.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460-2461.

Efron, B., Halloran, E., & Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. Proceedings of the National Academy of Sciences, 93(23), 13429-13429.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., . . . Bettman, B. (2009). Real-time DNA sequencing from single polymerase molecules. Science, 323(5910), 133-138.

Elsaied, H., Stokes, H. W., Kitamura, K., Kurusu, Y., Kamagata, Y., & Maruyama, A. (2011). Marine integrons containing novel integrase genes, attachment sites, attI, and associated gene cassettes in polluted sediments from Suez and Tokyo Bays. The ISME journal, 5(7), 1162-1177.

EMA. (2016). Updated advice on the use of colistin products in animals within the European Union: development of resistance and possible impact on human and animal health. In: EMA London, United Kingdom.

Escudero, J. A., Loot, C., Nivina, A., & Mazel, D. (2015). The integron: adaptation on demand. In Mobile DNA III (pp. 139-161): American Society of Microbiology.

EUCAST. (2013). EUCAST guidelines for detection of resistance mechanisms and specific resistances of clinical and/or epidemiological importance. EUCAST, Basel, Switzerland: http://www. eucast. org/clinical_breakpoints.

Evans, S. R., Hujer, A. M., Jiang, H., Hujer, K. M., Hall, T., Marzan, C., . . . Manca, C. (2016). Rapid molecular diagnostics, antibiotic treatment decisions, and developing approaches to inform empiric therapy: PRIMERS I and II. Clinical Infectious Diseases, 62(2), 181-189.

Folman, L. B., Postma, J., & van Veen, J. A. (2003). Characterisation of Lysobacter enzymogenes (Christensen and Cook 1978) strain 3.1 T8, a powerful antagonist of fungal diseases of cucumber. Microbiological Research, 158(2), 107-115.

Fu, S., Wang, A., & Au, K. F. (2019). A comparative evaluation of hybrid error correction methods for error-prone long reads. Genome Biology, 20(1), 26.

Gabaldón, T. (2005). Evolution of proteins and proteomes: a phylogenetics approach. Evolutionary Bioinformatics, 1, 117693430500100004.

Gaffé, J., McKenzie, C., Maharjan, R. P., Coursange, E., Ferenci, T., & Schneider, D. (2011). Insertion sequence-driven evolution of Escherichia coli in chemostats. Journal of Molecular Evolution, 72(4), 398-412.

García-Galán, M. J., Blanco, S. G., Roldán, R. L., Díaz-Cruz, S., & Barceló, D. (2012). Ecotoxicity evaluation and removal of sulfonamides and their acetylated metabolites during conventional wastewater treatment. Science of the Total Environment, 437, 403-412.

Gillings, M. R. (2014). Integrons: past, present, and future. Microbiology and Molecular Biology Reviews, 78(2), 257-277.

Gillings, M. R. (2016). Lateral gene transfer, bacterial genome evolution, and the Anthropocene. Annals of the New York Academy of Sciences.

Gillings, M. R., & Stokes, H. (2012). Are humans increasing bacterial evolvability? Trends in ecology & evolution, 27(6), 346-352.

Gillings, M. R., Xuejun, D., Hardwick, S. A., Holley, M. P., & Stokes, H. (2009). Gene cassettes encoding resistance to quaternary ammonium compounds: a role in the origin of clinical class 1 integrons? The ISME journal, 3(2), 209-215.

Gullberg, E., Albrecht, L. M., Karlsson, C., Sandegren, L., & Andersson, D. I. (2014). Selection of a multidrug resistance plasmid by sublethal levels of antibiotics and heavy metals. MBio, 5(5), e01918-01914.

Hackl, T., Hedrich, R., Schultz, J., & Förster, F. (2014). proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics, 30(21), 3004-3011.

Hadano, Y., Ito, K., Suzuki, J., Kawamura, I., Kurai, H., & Ohkusu, K. (2012). Moraxella osloensis: an unusual cause of central venous catheter infection in a cancer patient. International Journal of General Medicine, 5, 875.

Hadjadj, L., Baron, S. A., Diene, S. M., & Rolain, J.-M. (2019). How to discover new antibiotic resistance genes? Expert Review of Molecular Diagnostics, 19(4), 349-362.

Haghshenas, E., Hach, F., Sahinalp, S. C., & Chauve, C. (2016). Colormap: Correcting long reads by mapping short reads. Bioinformatics, 32(17), i545-i551.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1-8.

Hibbing, M. E., Fuqua, C., Parsek, M. R., & Peterson, S. B. (2010). Bacterial competition: surviving and thriving in the microbial jungle. Nature Reviews Microbiology, 8(1), 15-25.

Hickman, A. B., & Dyda, F. (2015). Mechanisms of DNA transposition. Microbiology spectrum, 3(2).

Holmes, A. J., Gillings, M. R., Nield, B. S., Mabbutt, B. C., Nevalainen, K., & Stokes, H. (2003). The gene cassette metagenome is a basic resource for bacterial genome evolution. Environmental Microbiology, 5(5), 383-394.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Molecular Biology and Evolution, 33(6), 1635-1638.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11(1), 1.

Jonsson, V., Österlund, T., Nerman, O., & Kristiansson, E. (2017). Variability in metagenomic count data and its influence on the identification of differentially abundant genes. Journal of Computational Biology, 24(4), 311-326.

Jutkina, J., Rutgersson, C., Flach, C.-F., & Larsson, D. G. J. (2016). An assay for determining minimal concentrations of antibiotics that drive horizontal transfer of resistance. Science of the Total Environment, 548, 131-138.

Katoh, K., Misawa, K., Kuma, K. i., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14), 3059-3066.

Khan, A. R., Pervez, M. T., Babar, M. E., Naveed, N., & Shoaib, M. (2018). A comprehensive study of de novo genome assemblers: current challenges and future prospective. Evolutionary Bioinformatics, 14, 1176934318758650.

Koenig, J. E., Sharp, C., Dlutek, M., Curtis, B., Joss, M., Boucher, Y., & Doolittle, W. F. (2009). Integron gene cassettes and degradation of compounds associated with industrial waste: the case of the Sydney tar ponds. PloS One, 4(4), e5276.

Kristiansson, E., Fick, J., Janzon, A., Grabic, R., Rutgersson, C., Weijdegård, B., . . . Larsson, D. G. J. (2011). Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. PloS One, 6(2), e17038.

Krueger, F. (2015). Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

Kumar, M., Jaiswal, S., Sodhi, K. K., Shree, P., Singh, D. K., Agrawal, P. K., & Shukla, P. (2019). Antibiotics bioremediation: Perspectives on its ecotoxicity and resistance. Environment International, 124, 448-461.

Larsson, D. G. J. (2014). Pollution from drug manufacturing: review and perspectives. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1656), 20130571.

Larsson, D. G. J., Andremont, A., Bengtsson-Palme, J., Brandt, K. K., de Roda Husman, A. M., Fagerstedt, P., . . . Kuroda, M. (2018). Critical knowledge gaps and research needs related to the environmental dimensions of antibiotic resistance. Environment International, 117, 132-138.

Larsson, D. G. J., de Pedro, C., & Paxeus, N. (2007). Effluent from drug manufactures contains extremely high levels of pharmaceuticals. Journal of Hazardous Materials, 148(3), 751-755.

Lartigue, M.-F., Poirel, L., & Nordmann, P. (2006). Diversity of genetic environment of bla CTX-M genes. FEMS Microbiology Letters, 234(2), 201-207.

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity, 104(6), 520-533.

Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Research, 44(W1), W242-W245.

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., & Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. Science, 299(5607), 682-686.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 31(10), 1674-1676.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658-1659.

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M. W., Chaisson, M., & Pevzner, P. A. (2016). Assembly of long error-prone reads using de Bruijn graphs. Proceedings of the National Academy of Sciences, 113(52), E8396-E8405.

Lundström, S. V., Östman, M., Bengtsson-Palme, J., Rutgersson, C., Thoudal, M., Sircar, T., . . . Flach, C.-F. (2016). Minimal selective concentrations of tetracycline in complex aquatic bacterial biofilms. Science of the Total Environment, 553, 587-595.

Macesic, N., Khan, S., Giddins, M. J., Freedberg, D. E., Whittier, S., Green, D. A., . . . Gomez-Simmonds, A. (2019). Escherichia coli harboring mcr-1 in a cluster of liver transplant recipients: detection through active surveillance and

whole-genome sequencing. Antimicrobial Agents and Chemotherapy, 63(6), e02680-02618.

Maier, R. M., Pepper, I. L., & Gerba, C. P. (2009). Environmental microbiology (Vol. 397): Academic press.

Marathe, N. P., Chandan, P., Gaikwad, S. S., Jonsson, V., Kristiansson, E., & Larsson, D. G. J. (2017). Untreated urban waste contaminates Indian river sediments with resistance genes to last resort antibiotics. Water Research, 124, 388-397. doi:http://dx.doi.org/10.1016/j.watres.2017.07.060

Marathe, N. P., Regina, V. R., Walujkar, S. A., Charan, S. S., Moore, E. R., Larsson, D. G. J., & Shouche, Y. S. (2013). A treatment plant receiving waste water from multiple bulk drug manufacturers is a reservoir for highly multi-drug resistant integron-bearing bacteria. PloS One, 8(10), e77310.

Martínez, J. L. (2008). Antibiotics and antibiotic resistance genes in natural environments. Science, 321(5887), 365-367.

Martínez, J. L. (2012). Bottlenecks in the transferability of antibiotic resistance from natural ecosystems to human bacterial pathogens. Frontiers in Microbiology, 2, 265.

Martínez, J. L., Coque, T. M., & Baquero, F. (2015). What is a resistance gene? Ranking risk in resistomes. Nature Reviews. Microbiology, 13(2), 116-123.

McEwen, S. A., & Collignon, P. J. (2018). Antimicrobial resistance: a One Health perspective. Antimicrobial Resistance in Bacteria from Livestock and Companion Animals, 521-547.

Mingeot-Leclercq, M.-P., Glupczynski, Y., & Tulkens, P. M. (1999). Aminoglycosides: activity and resistance. Antimicrobial Agents and Chemotherapy, 43(4), 727-737.

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Research, 41(12), e121-e121.

Morar, M., & Wright, G. D. (2010). The genomic enzymology of antibiotic resistance. Annual Review of Genetics, 44, 25-51.

Mullany, P. (2014). Functional metagenomics for the investigation of antibiotic resistance. Virulence, 5(3), 443-447.

Ou, D., Chen, B., Bai, R., Song, P., & Lin, H. (2015). Contamination of sulfonamide antibiotics and sulfamethazine-resistant bacteria in the downstream and estuarine areas of Jiulong River in Southeast China. Environmental Science and Pollution Research, 22(16), 12104-12113.

Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., & Larsson, D. J. (2013). BacMet: antibacterial biocide and metal resistance genes database. Nucleic Acids Research, 42(D1), D737-D743.

Papp-Wallace, K. M., Endimiani, A., Taracila, M. A., & Bonomo, R. A. (2011). Carbapenems: past, present, and future. Antimicrobial Agents and Chemotherapy, 55(11), 4943-4960.

Partridge, S. R., Kwong, S. M., Firth, N., & Jensen, S. O. (2018). Mobile genetic elements associated with antimicrobial resistance. Clinical Microbiology Reviews, 31(4), e00088-00017.

Partridge, S. R., Tsafnat, G., Coiera, E., & Iredell, J. R. (2009). Gene cassettes and cassette arrays in mobile resistance integrons. FEMS Microbiology Reviews, 33(4), 757-784.

Pereira, M. B., Wallroth, M., Jonsson, V., & Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. BMC genomics, 19(1), 274.

Pereira, M. B., Wallroth, M., Kristiansson, E., & Axelson-Fisk, M. (2016). HattCI: Fast and Accurate attC site Identification Using Hidden Markov Models. Journal of Computational Biology, 23(11), 891-902.

Perry, J. A., Westman, E. L., & Wright, G. D. (2014). The antibiotic resistome: what's new? Current Opinion in Microbiology, 21, 45-50.

Poirel, L., Kieffer, N., Fernandez-Garayzabal, J. F., Vela, A. I., Larpin, Y., & Nordmann, P. (2017). MCR-2-mediated plasmid-borne polymyxin resistance most likely originates from Moraxella pluranimalium. Journal of Antimicrobial Chemotherapy, 72(10), 2947-2949.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution, 26(7), 1641-1650.

Rabin, M. O. (1963). Probabilistic automata. Information and control, 6(3), 230-245.

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. Genome Biology, 14(7), 405.

Rodríguez, M. M., Herman, R., Ghiglione, B., Kerff, F., González, G. D. A., Bouillenne, F., . . . Gutkind, G. (2017). Crystal structure and kinetic analysis of the class B3 di-zinc metallo-β-lactamase LRA-12 from an Alaskan soil metagenome. PloS One, 12(7).

Salmela, L., & Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. Bioinformatics, 30(24), 3506-3514.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, 74(12), 5463-5467.

Schmieder, R., & Edwards, R. (2012). Insights into antibiotic resistance through metagenomic approaches. Future Microbiology, 7(1), 73-89.

Sengupta, S., Chattopadhyay, M. K., & Grossart, H.-P. (2013). The multifaceted roles of antibiotics and antibiotic resistance in nature. Frontiers in Microbiology, 4, 47.

Shah, S. S., Ruth, A., & Coffin, S. E. (2000). Infection due to Moraxella osloensis: case report and review of the literature. Clinical Infectious Diseases, 30(1), 179-181.

Siguier, P., Gourbeyre, E., & Chandler, M. (2014). Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiology Reviews, 38(5), 865-891.

Sköld, O. (2000). Sulfonamide resistance: mechanisms and trends. Drug Resistance Updates, 3(3), 155-160.

Solomon, E. B., Yaron, S., & Matthews, K. R. (2002). Transmission of Escherichia coli O157: H7 from contaminated manure and irrigation water to lettuce plant tissue and its subsequent internalization. Applied and Environmental Microbiology, 68(1), 397-400.

Spencer, S. J., Tamminen, M. V., Preheim, S. P., Guo, M. T., Briggs, A. W., Brito, I. L., . . . Virta, M. P. (2016). Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. The ISME journal, 10(2), 427-436.

Stalder, T., Barraud, O., Casellas, M., Dagot, C., & Ploy, M.-C. (2012). Integron involvement in environmental spread of antibiotic resistance. Frontiers in Microbiology, 3, 119.

Strimmer, K. S. (1997). Maximum likelihood methods in molecular phylogenetics: Herbert Utz Verlag.

Suzuki, S., & Hoa, P. T. P. (2012). Distribution of quinolones, sulfonamides, tetracyclines in aquatic environment and antibiotic resistance in Indochina. Frontiers in Microbiology, 3, 67.

Tan, L., & Grewal, P. S. (2001). Pathogenicity of Moraxella osloensis, a bacterium associated with the nematode Phasmarhabditis hermaphrodita, to the slug Deroceras reticulatum. Applied and Environmental Microbiology, 67(11), 5010-5016.

Troeger, C., Blacker, B., Khalil, E., & Collaborators, G. L. R. I. (2018). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet. Infectious diseases, 18(11), 1191.

Troeger, C., Forouzanfar, M., Rao, P. C., Khalil, I., Brown, A., Reiner Jr, R. C., . . . Ahmed, M. (2017). Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet infectious diseases, 17(9), 909-948.

Tsalik, E. L., Bonomo, R. A., & Fowler Jr, V. G. (2018). New molecular diagnostic approaches to bacterial infections and antibacterial resistance. Annual Review of Medicine, 69, 379-394.

Walsh, T. R., & Wu, Y. (2016). China bans colistin as a feed additive for animals. The Lancet infectious diseases, 16(10), 1102.

Van Boeckel, T. P., Brower, C., Gilbert, M., Grenfell, B. T., Levin, S. A., Robinson, T. P., . . . Laxminarayan, R. (2015). Global trends in antimicrobial use in food animals. Proceedings of the National Academy of Sciences, 112(18), 5649-5654.

Van Boeckel, T. P., Gandra, S., Ashok, A., Caudron, Q., Grenfell, B. T., Levin, S. A., & Laxminarayan, R. (2014). Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. The Lancet infectious diseases, 14(8), 742-750.

Vandecraen, J., Chandler, M., Aertsen, A., & Van Houdt, R. (2017). The impact of insertion sequences on bacterial genome plasticity and adaptability. Critical Reviews in Microbiology, 43(6), 709-730.

Wang, H., Wang, T., Zhang, B., Li, F., Toure, B., Omosa, I. B., . . . Pradhan, M. (2014). Water and wastewater treatment in Africa–current practices and challenges. CLEAN–Soil, Air, Water, 42(8), 1029-1035.

Wang, Y., Liu, H., Tong, L., Feng, L., & Ma, K. (2018). Characterization of the diethyl phthalate-degrading bacterium Sphingobium yanoikuyae SHJ. Data in brief, 20, 1758-1763.

Wanger, A., Chavez, V., Huang, R., Wahed, A., Actor, J., & Dasgupta, A. (2017). Chapter 12—Overview of Molecular Diagnostics Principles. Microbiology and Molecular Diagnosis in Pathology; Wanger, A., Chavez, V., Huang, R., Wahed, A., Dasgupta, A., Actor, J., Eds, 233-257.

WHO. (2017a). Antibacterial agents in clinical development: an analysis of the antibacterial clinical development pipeline, including tuberculosis. Retrieved from

WHO. (2017b). Critically important antimicrobials for human medicine: ranking of antimicrobial agents for risk management of antimicrobial resistance due to non-human use.

WHO. (2019). Global tuberculosis report 2019. Geneva, Switzerland: World Health Organization; 2019. In.

von Wintersdorff, C. J., Penders, J., van Niekerk, J. M., Mills, N. D., Majumder, S., van Alphen, L. B., . . . Wolffs, P. F. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. Frontiers in Microbiology, 7, 173.

Xie, W. Y., Shen, Q., & Zhao, F. (2018). Antibiotics and antibiotic resistance from animal manures to soil: a review. European Journal of Soil Science, 69(1), 181-195.

Xu, F., Min, F., Wang, J., Luo, Y., Huang, S., Chen, M., . . . Zhang, Y. (2020). Development and evaluation of a Luminex xTAG assay for sulfonamide resistance genes in Escherichia coli and Salmonella isolates. Molecular and Cellular Probes, 49, 101476.

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., . . . Zhu, B. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. BMC Bioinformatics, 14(1), 1.

Zadegan, S. M. R., Mirzaie, M., & Sadoughi, F. (2013). Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. Knowledge-Based Systems, 39, 133-143.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., . . . Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy, 67(11), 2640-2644.

Zhang, H., Jain, C., & Aluru, S. (2019). A comprehensive evaluation of long read error correction methods. BioRxiv, 519330.

Zhang, X., Li, Y., Liu, B., Wang, J., Feng, C., Gao, M., & Wang, L. (2014). Prevalence of veterinary antibiotics and antibiotic-resistant Escherichia coli in the surface water of a livestock production region in northern China. PloS One, 9(11).

I