On the Origins of mobile Antibiotic Resistance Genes

A comparative genomics approach

Stefan Ebmeyer

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

UNIVERSITY OF GOTHENBURG

Gothenburg 2021

Cover illustration: Stefan Ebmeyer

On the Origins of mobile Antibiotic Resistance Genes
A comparative genomics approach

stefan.ebmeyer@gu.se

ISBN 978-91-8009-304-0 (PRINT)
ISBN 978-91-8009-305-7 (PDF)

Printed in Borås, Sweden 2021
Printed by Stema Specialtryck AB, Borås

It is not our part to master all the tides of the world, but to do what is in us for the succor of those years wherein we are set, uprooting the evil in the fields that we know, so that those who live after may have clean earth to till.

- *J. R. R. Tolkien, The Return of the King*

ABSTRACT

Mobile antibiotic resistance genes (ARGs), transferable between bacterial cells, are major contributors to the antibiotic resistance crisis we are facing today. From which organisms pathogens acquired these genes is mostly unknown, yet knowledge about their origin is needed in order to limit the emergence and spread of novel ARGs in the future. Increasing the number of known origins of mobile resistance genes would allow us to investigate potential patterns that may hint towards the conditions that potentially promote the emergence of mobile ARGs. This thesis aims to identify from which taxa ARGs have been mobilized into pathogens, so that this knowledge may aid mitigations to limit the emergence of novel ARGs in the future.

We used comparative genomic methods on the large amount of publicly available sequenced bacterial genomes in order to identify bacterial taxa from which certain ARGs have been mobilized (paper I-IV). A literature review and the development of a computational pipeline (paper VI) to compare hundreds of genomic loci allowed us to scrutinize previously reported origins and analyze patterns among to-date identified ARG origins (paper V).

In this thesis, we have identified the recent origins of PER-type class A beta-lactamases as *Pararheinheimera* spp. (Paper I), the recent origins of CMY-1/MOX-1, MOX-2 and MOX-9 class C beta-lactamases as *Aeromonas sanarellii*, *Aeromonas caviae* and *Aeromonas media* respectively (Paper II), the recent origin of FOX-type class C beta-lactamases as *Aeromonas allosaccharophila* (Paper III), and the recent origin of GPC-1/BKC-1 carbapenemases as *Shinella* spp (Paper IV). In paper V, based on the amended and curated data from the literature, five criteria allowing for the confident identification of recent origins of mobile ARGs were identified. Of all recent origins identified on species level, all were Proteobacteria, >90% were identified as potential pathogens of humans and/or domestic animals, none of them known antibiotic producers themselves. However, all curated recent origins account for only about 4% of known mobile ARGs, indicating that environmental bacteria may represent a significant source of resistance genes. Finally, Paper VI presents a bioinformatics pipeline, GEnView, for comparative genomic analysis of gene loci among hundreds of genomes, developed throughout this thesis.

This thesis further elucidates the recent origins of several mobile resistance genes, identifies previously unrecognized patterns about their emergence and provides other researchers with the tools to investigate the origins of other resistance genes. This knowledge may prove valuable to guide future efforts trying to mitigate the emergence of additional ARGs in the clinics.

SAMMANFATTNING PÅ SVENSKA

Antibiotika är helt nödvändiga för stora delar av vår moderna sjukvård. De används inte bara för behandling av infektionssjukdomar orsakad av bakterier men också som förebyggande behandling vid t ex operationer och olika tillstånd som sätter ned immunförsvaret. Sjukdomsframkallande bakterier som har utvecklat resistens mot antibiotika och därmed är svårbehandlade blir allt vanligare. En av grunderna till denna utveckling är att många bakterier kan ta emot DNA från andra bakterier, och på det sätt skaffa gener som kodar för antibiotikaresistens. Nya mobila resistensgener som kan hoppa mellan arter upptäcks regelbundet, men det är oklart varifrån de kommer och hur de hamnar i sjukdomsframkallande bakterier. För att försöka att minska mängden av nya resistensgener som kan hamna i farliga bakterier är det viktigt att förstå varifrån mobila resistensgener kommer från början, och under vilka omständigheter de mobiliseras från de bakterierna där de har sina ursprung.

I denna avhandling har vi använt oss av jämförande genomik, en metod där vi har jämfört resistensgensekvenser och deras genetiska omgivning i olika bakteriegenom, för att hitta ursprunget för flera antibiotikaresistensgener. Vi har även sammanfattat, kritiskt granskat och analyserat litteraturen för att upptäcka mönster bland kända ursprung av resistensgener.

I studierna I-IV upptäckte vi att resistensgenerna *bla*$_{\text{PER}}$, *bla*$_{\text{CMY-1/MOX}}$, *bla*$_{\text{FOX}}$ och *bla*$_{\text{GPC-1/BKC-1}}$ har sina ursprung i vanligen vattenlevande släkten/arter (*Pararheinheimera*, *Aeromonas sanarellii, Aeromonas caviae, Aeromonas media, Aeromonas allosaccharophila* och *Shinella).* I studie V genomsökte vi den vetenskapliga litteraturen för att identifiera de artiklar som hade identifierat ursprung av andra resistensgener. Vi granskade datan jämsides med tillgängliga genom från tusentals arter, och upptäckte att alla hittills upptäckta ursprungsbakterier tillhör gruppen Proteobakterier, och att nästan alla av dessa åtminstone ibland orsakar infektioner i människor eller domesticerade djur. Efter en granskning av litteraturen formulerade vi kriterier som underlättar identifiering av fler ursprungsbakterier i framtiden. I studie VI utvecklade vi en mjukvara som gör det möjligt att visualisera och jämföra hundratals resistensgener och deras omgivning från olika genom med varandra.

LIST OF PAPERS

I. Ebmeyer S, Kristiansson E & Larsson D. G. J. **PER extended-spectrum β-lactamases originate from Pararheinheimera spp.** *Int. J. Antimicrob. Agents* **53**, 158–164 (2019).

II. Ebmeyer S, Kristiansson E & Larsson D. G. J. **CMY-1/MOX-family AmpC β-lactamases MOX-1, MOX-2 and MOX-9 were mobilized independently from three *Aeromonas* species**. *J. Antimicrob. Chemother.* (2019) doi:10.1093/jac/dkz025.

III. Ebmeyer S, Kristiansson E. & Larsson D. G. J. **The mobile FOX AmpC beta-lactamases originated in Aeromonas allosaccharophila**. *Int. J. Antimicrob. Agents* **54**, 798–802 (2019).

IV. Kieffer N, Ebmeyer S & Larsson D. G. J. **The Class A Carbapenemases BKC-1 and GPC-1 Both Originate from the Bacterial Genus Shinella**. *Antimicrob. Agents Chemother.* **64**, (2020).

V. Ebmeyer S, Kristiansson E & Larsson D. G. J. **A framework for identifying the recent origins of mobile antibiotic resistance genes**. *Commun. Biol.* **4**, 1–10 (2021).

VI. Ebmeyer S, Kristiansson E, Larsson DGJ. **GEnView: A gene-centered, phylogeny-based comparative genomics pipeline for bacterial genomes and plasmids**. Manuscript

OTHER PUBLICATIONS NOT INCLUDED IN THIS THESIS

1. Kraupner N, Ebmeyer S, Bengtsson-Palme J, Fick J, Kristiansson E, Flach CF, Larsson DGJ. **Selective concentration for ciprofloxacin resistance in Escherichia coli grown in complex aquatic bacterial biofilms.** *Environ. Int.* **116**, 255–268 (2018).

2. Rutgersson C, Ebmeyer S, Lassen SB, Karkman A, Fick J, Kristiansson E, Brandt K, Flach CF, Larsson DGJ. **Long-term application of Swedish sewage sludge on farmland does not cause clear changes in the soil bacterial resistome.** *Environ. Int.* **137**, 105339 (2020).

3. Kraupner N, Ebmeyer S, Hutinel M, Fick J, Flach CF, Larsson DGJ. **Selective concentrations for trimethoprim resistance in aquatic environments.** *Environ. Int.* **144**, 106083 (2020).

4. Berglund F, Böhm ME, Martinsson A, Ebmeyer S, Österlund T, Johnning A, Larsson DGJ, Kristiansson E. **Comprehensive screening of genomic and metagenomic data reveals a large diversity of tetracycline resistance genes.** *Microb. Genomics* **6**, 1–14 (2020).

CONTENT

ABBREVIATIONS

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic acid
ARG	Antibiotic Resistance Gene
IS	Insertion Sequence
IS*CR*	Insertion Sequence Common Region
MGE	Mobile Genetic Element
HGT	Horizontal Gene Transfer
VRE	Vancomycin Resistant *Enterococci*
3'CS	3' Conserved Segment
PCR	Polymerase Chain Reaction
WGS	Whole genome sequencing
NGS	Next generation sequencing
SMRT	Single Molecule Real Time
HGAP	Hierarchical Genome-Assembly Process
ARO	Antibiotic Resistance Ontology
MSA	Multiple Sequence Alignment
ML	Maximum Likelihood

1. INTRODUCTION

1.1 Antibiotics and antibiotic resistance

The large-scale introduction of antibiotics to the market in the 1940s revolutionized human medicine. Bacterial infections, previously one of the main causes of human mortality, suddenly became easily treatable[1]. Since then, antibiotics have become the foundation of modern health care systems. Being used not only for the treatment of bacterial infections, but also to prevent infection after surgery or during cancer treatment, the use of antibiotics is estimated to have extended the average human life span by several years[2]. Due to their effectiveness, antibiotic drugs are also widely used in agriculture and veterinary medicine. Today, there are multiple classes of antibiotics with distinct mechanisms of action, some of them containing several subclasses. While many of these are derived from natural products that are produced by fungi or bacteria, mostly from the bacterial taxon *Actinobacteria*[3], others are based on synthetic substances not found in the natural world.

Antibiotics act either through directly killing bacteria, or inhibiting bacterial growth by targeting essential metabolic processes in the cell, such as DNA synthesis (inhibited by e.g. fluoroquinolones), RNA synthesis (inhibited by rifamycins), Cell wall synthesis/maintenance (inhibited by e.g. beta-lactams) and protein synthesis (inhibited by e.g aminoglycosides)[4].

However, bacteria that had become resistant to the toxic effects of commonly used antibiotics already appeared shortly after the introduction of these drugs as clinical agents. Subsequent research on resistant isolates showed that resistance is usually mediated by a number of molecular mechanisms, the most common being **antibiotic efflux** (e.g. through efflux pumps), **decreased antibiotic uptake** (e.g. reduced expression of porins), **target alteration** (e.g. mutation of the *gyrA* gene leading to fluoroquinolone resistance in *E. coli*), **enzymatic inactivation** (e.g. neutralization of beta-lactam antibiotics through beta-lactamases) and **acquisition of alternative enzymes** (e.g. acquisition of *sul* genes in *Enterobacteriaceae*).

While some bacteria are intrinsically resistant to the effects of certain antibiotics, meaning that their physiology, such as an impermeable cell wall, lack of the target site, presence of antibiotic modifying enzymes or a combination of several such factors, renders a specific antibiotic or antibiotic class ineffective (e.g gram-negatives being resistant to glycopeptide antibiotics or species of *Enterococci* being intrinsically resistant to a multitude of beta-lactam antibiotics[5]). Thanks to the multitude of available antibiotic classes, intrinsic resistance can be circumvented in many cases simply

through use of another antibiotic class. However, bacteria have another trick up their sleeve – they are able to acquire resistance to antibiotics from other bacteria.

1.2 Acquired resistance – Horizontal gene transfer, mutation and mobile resistance genes

Bacteria can develop resistance to antibiotics either by mutation of e.g. an antibiotic target (such as mutation of *gyrA*, encoding topoisomerase II, can lead to fluoroquinolone resistance in *E. coli*), by acquisition of mobile antibiotic resistance genes (ARGs) through horizontal gene transfer (HGT), or a combination of these processes. Horizontal transfer of genetic material between bacterial cells occurs by three main mechanisms: transformation (uptake of free DNA from the extracellular environment), transduction (introduction of foreign DNA via bacteriophages) and conjugation (exchange of genetic material through physical connection of two cells). As the majority of clinically relevant mobile ARGs is found on plasmids, conjugative transposons and other elements transferred by conjugation, this process is arguably the most important for their dissemination. Though HGT between taxa is common in almost every environment, it has been shown that anthropogenic impact, such as selection pressure by antibiotics, can promote the spread of ARGs[6].

Both mutation and acquisition of foreign genetic material are frequent causes of resistance in clinical isolates. Whereas the number of resistance mutations, and thus the number of antibiotics a bacterium can acquire resistance to through mutation is usually limited, the number of resistance genes that a bacterium can acquire through HGT is *in theory* unlimited. This in turn means a bacterium potentially has the ability to develop resistance to all available antibiotic drugs (pan-resistance) through acquisition of mobile genes alone (though a combination of mutations and acquired genes is probably more common). Though pan-resistance is not common to date, such cases have already been observed in the clinics, such as pan-resistant *Klebsiella pneumoniae* isolates from India or China[7,8]. Thus, mobile ARGs represent a considerable risk to the efficiency of antibiotic treatment.

1.3 Insertion sequences as mobilizing agents

The risk that mobile ARGs represent opens the question how they gained mobility in the first place. ARGs can be mobilized from their 'native' origin locus by transposable elements, which are able to move the ARG onto mobile genetic elements (MGEs), such as plasmids or large transposons. Small transposable elements like insertion sequences

(IS) and insertion sequence common region elements (IS*CR*)[9–11] have been shown to play a key role in the mobilization and distribution of ARGs to different replicons[12–14], therefore it is necessary to discuss them in more detail here. IS elements usually encode only one or two transposase genes involved in their own transposition, are flanked by inverted repeats (IR_L and IR_R) and can transpose between different loci without the need for large regions of homology. They are divided into different groups based on the active site motif of the transposase gene(s) and their mechanism of transposition[15]. Some IS have been shown to not only transpose themselves, but also adjacent pieces of DNA, either single-handedly or with help of another identical or closely related IS. In the latter case a so-called composite transposon is formed, in which passenger DNA is flanked by the respective ISs on either side[16]. It has been experimentally verified on several occasions that IS are able to mobilize ARGs (or their progenitors) from the chromosome of their origins, such as for *bla*$_{CMY-2}$, *bla*$_{CTX-M}$ or the fluoroquinolone resistance gene *qnrB*[9,12,13]. Some types of insertion sequences have repeatedly been shown to be efficient mobilization machineries, one of the most noteworthy being IS*Ecp1*, which is suspected to have mobilized several ARGs from their origin onto conjugative vectors.

IS*CR* elements are atypical insertion sequences that lack terminal inverted repeats. They most likely transpose via a mechanism called rolling circle transposition[10], which enables them to single-handedly mobilize adjacent DNA sequences. Though different IS*CR*s have been associated with a variety of ARGs, IS*CR*1 is of special interest, as it is thought to be involved in to mobilization of several resistance genes. It is most often found linked to the 3' conserved segment (3'CS) of class I integrons (genetic structures that are able to capture and express mobile genes, so called gene cassettes), and associated with different ARGs[17]. As IS*CR*1 appears to repeatedly insert at the 3'CS region of class I integrons, it may be of special importance for the recruitment of novel ARGs into clinically relevant genetic contexts.

In addition to providing mobility, IS and IS*CR* elements also may alter the expression of adjacent genes, either through disrupting native regulators and creating hybrid promoters, providing promoters within their own sequence (such as the IS*CR*1-borne P_{OUT} promoters, whichs' role in the expression adjacent ARGs has been shown[18]) or increasing gene dosage (referring to the number of gene copies in a cell), of e.g. genes that are transferred to multicopy plasmids[18,19]. These mechanisms allow genes which do not provide clinical levels of antibiotic resistance in their native chromosomal context to be 're-functioned' as antibiotic resistance genes[20,21].

1.4 Human impact on antibiotic resistance evolution

Various forms of evidence produced during the past decades, such as positive correlations between antibiotic usage and resistance[22], the absence of resistance genes from pathogen-borne plasmids from the pre-antibiotic era[23] and high abundances of resistance genes at sites with high antibiotic selection pressure[24], have made it clear that anthropogenic use of antibiotics is the driving force of the high abundance and spread of mobile ARGs in pathogens that we observe today. As antibiotic exposure has been shown to promote HGT and processes favoring the recombination of DNA (such as the bacterial SOS response or transposition activity of intracellular MGEs)[25], exposure to antibiotics likely plays a critical role in the emergence of novel ARGs from their origins as well[9] – but it is still unclear in which environments these events take place, how much selection pressure they require, how frequent they are and how these novel ARGs make it into human pathogens.

As mobile ARGs play a crucial role in the emergence of difficult, if not impossible-to-treat bacterial infections, an important aspect of managing the antibiotic resistance crisis we are facing today is to limit the emergence of novel mobile ARGs in the clinics. In order to do that, we need to know what conditions and environments contribute to the emergence of such novel ARGs. To be able to investigate these factors, in turn, we need to know where these mobile genes come from, from where they have been mobilized onto transferrable vectors. In other words, we need to find their origin.

1.5 Origins of acquired resistance genes

Antibiotic molecules have existed since long before being used to treat bacterial infections. The enormous variety of ARGs and ARG-like genes found in pristine environmental samples[11] or ancient DNA recovered from permafrost[26], shows that the same is the case for resistance genes. One hypothesis concerning their natural function is that they serve as self-protection in antibiotic producing organisms, such as *Streptomyces*[27] or *Amycolatopsis*[28] spp.. Structural similarities in these genes and the mobile ARGs of resistant pathogens suggests that some ARGs, such as the mobile *vanHAX* operon, may have originated in those producer organisms[28,29]. However, the degree of divergence in sequence identities between ARG-like genes in these organisms and the mobile ARGs in pathogens indicates in most cases that potential gene mobilizations from producer organisms were not evolutionary recent[27] and existing evidence suggests that antibiotic producers are not the primary source of clinically relevant ARGs[30]. The mobile $bla_{CTX\text{-}M}$ genes, on the other hand, were shown to have been mobilized from the *Kluyvera* species *K. ascorbata*[31] and *K. georgiana*[32]. Sequence identities were up to 99% between not only the chromosomal and mobile $bla_{CTX\text{-}M}$, but

also parts of their genetic environments which have been co-mobilized from the *Kluyvera* chromosome. Experimental data indicate that CTX-M enzymes are only weakly expressed in their native content and require increased expression (such as may be induced by IS or IS*CR* elements) to confer high-level beta-lactam resistance to their host[33]. The chromosomal origins of other mobile ARGs that have been identified, such as *Shewanella algae* for the fluoroquinolone resistance gene *qnrA*[21] or the *Citrobacter freundii* group for *bla*$_{\text{CMY-2}}$[34], also are nearly identical in nucleotide identity (≥97%) towards their mobile counterparts, suggesting that their mobilization from the origin chromosome also was a more recent event. But how does one identify the origins of an ARG in the first place?

1.6 Identifying the origins of mobile antibiotic resistance genes – previous research and consideration of methods

To identify the origins of an ARG, the possibility to compare resistance determinants is fundamental. Early studies reported high similarities in biochemical activity between aminoglycoside inactivating enzymes identified in antibiotic-producing *Streptomyces* spp., prompting a first hypothesis about the origins of mobile ARGs in antibiotic-producing organisms[35]. This hypothesis received support when the presence of a *vanHAX*-like operon (which in its mobile form confers vancomycin resistance in *Enterococci* (VRE)) was discovered in antibiotic producing species of *Streptomyces* and *Amycolatopsis*[28], with amino acid identities of up to 64%.

The advancement of DNA sequencing methods and the establishment of public sequence databases greatly facilitated this process. Among the first mobile ARGs to be investigated were the mobile AmpC beta-lactamases, which provide their host with resistance to a range of beta-lactam antibiotics such as cephalosporins[36]. Already in 1990, high sequence similarity of the novel, mobile *ampC* gene *bla*$_{\text{MIR-1}}$ to parts of the chromosomal *Enterobacter cloacae ampC* gene were noticed – the comparison of the two genes was made based on similar resistance profiles[37]. In 1998, another transferrable beta-lactam resistance determinant was detected in *Salmonella enterica*. Again, the resistance patterns suggested the involvement of an AmpC beta-lactamase. The resistance determining sequence was, after recombination and subsequent selection, sequenced using the Sanger technique. A sequence search against the available public databases revealed 98,7% nucleotide identity of the resistance determinant (named DHA-1) towards the chromosomal *Morganella morganii ampC* gene. Furthermore, it was shown that another gene, *ampR*, which is involved in *ampC* regulation, was also present on the plasmid, 97% similar to *ampR* on the *M. morganii* chromosome – the sequences in between the *ampC* and *ampR* genes were 98% similar

between the recombinant plasmid and chromosome[38]. Since it was known that *ampR* and *ampC* are chromosomal genes in *M. morganii*, the results strongly suggested that the *M. morganii* chromosome was the origin of DHA-1. A year later, the origin of the mobile AmpC beta-lactamase CMY-2 was identified as the chromosomal *Citrobacter freundii* AmpC, using similar methodology. In this study, focus was also placed on the regions flanking the mobile $bla_{\text{CMY-2}}$ gene, which provided important evidence that CMY-2 had indeed been mobilized from the *C. freundii* chromosome[34]. The same methodology revealed *Hafina alvei* as origin of the mobile AmpC enzyme ACC-1 one year later[39]. As mentioned previously, it was already known that the *ampC* gene is encoded on the chromosome in these enterobacterial species – they were already then relatively well studied and known as (though in some cases rare) human pathogens. Thus, there was no question regarding the mobility of the *ampC* gene in these contexts. In other cases, where the potential origin is less well studied, methodology to assess the mobility of the ARG-like gene in the origin species is required, as mobile elements such as composite transposons can also transpose a mobilized ARG to the chromosome[40]. Hence, without assessing a genes mobility, it could in such cases be mistaken for a 'native' chromosomal gene.

In 2005, Poirel et al. investigated the origin of the mobile fluoroquinolone resistance gene *qnrA*. After an initial polymerase chain reaction (PCR) screening of several enterobacterial species for *qnrA*, several isolates of the aquatic species *Shewanella algae* was found to harbor genes highly similar to *qnrA*. In order to assess the mobility of the gene in *S. algae*, the I-Ceu-I endonuclease technique was applied. I-Ceu-I cuts bacterial DNA at the *rrl* gene, coding for the 23S ribosomal rRNA, thus cutting the targeted genome in several pieces[41]. Separated by pulse field gel electrophoresis, all fragments obtained from the *Shewanella algae* genome hybridized with DNA probes targeting the 16S and 23S rRNA genes – indicating that all fragments were derived from the *S. algae* chromosome. A probe targeting the *qnrA* gene hybridized with only one of the fragments, and the *S. algae* QnrA sequences were highly similar (two to four amino acid substitutions) to mobile QnrA. As the *qnrA* gene was known to be associated with the putative mobile element *orf513* in its mobile context (today known as IS*CR*1), the presence of *orf513* in the *Shewanella* isolates was investigated by PCR, which yielded negative results. The authors further noted that the GC-content of the *S. algae qnrA* gene matched that of the *S. algae* genome (52%)[21]. Based on these results, the authors identified the origin of mobile *qnrA* as *S. algae*. In further studies, the origins of OXA-181 and OXA-23 were identified as *S. xiamenensis* and *Acinetobacter radioresistens* respectively, using similar approaches. In these cases however, the potential origin of the mobile gene was also searched for genes that were encoded close to the respective ARG in its mobile context[42,43]. The presence of these genes, and synteny (conserved order of genes) between the mobile and chromosomal ARG-locus provide important evidence for each origin hypothesis.

The advancements in genome sequencing and the resulting increase in the availability of bacterial genome sequences in the following years led to the incorporation of genomic data in studies searching for ARG origins. To cover the known taxonomic diversity of *Acinetobacter* spp., Yoon et al. screened 133 *Acinetobacter* genomes for the presence of the aminoglycoside resistance determinant Aph(3')-IV. Three Aph(3')-IV-positive *Acinetobacter* spp. were identified and the genetic environments of the ARGs were analyzed, revealing the ARG as mobile (due to adjacent IS) in *A. parvus* and *A. baumanii*. No signs of mobility were found in the genomes of two *A. guillouiae* isolates – the ARG was encoded on large contigs that also encoded genes for ribosomal proteins, suggesting that the contigs were derived from chromosomal sequences. Due the low number of available genomes for this species, cultured isolates of *A. guillouiae* were analyzed using PCR based methodology, subsequently identifying *A. guillouiae* as the origin of mobile Aph(3')-IV.

These previous works have shown that it is possible to identify the origins of a mobile ARG by comparing genomic sequences of different species, with respect to both their nucleotide or amino acid sequences and their synteny. While assessing these questions using molecular methodology as previously described, has been shown to be successful, they leave some room for error – location of an ARG-like gene on a chromosomal DNA fragment for example could also be due to IS-mediated insertion of the gene to the chromosome. PCR-based techniques may not be able to identify unknown IS, and thus lead towards false assessments of an ARGs state of mobility. While these shortcomings are redeemable by analyzing a sufficient number of unrelated isolates, this is not always possible, as isolates of some species (especially novel ones) may be rare and difficult to obtain.

The rapidly growing number of bacterial genomes available in public sequence repositories[44,45] and the development of a wide array of tools for tasks such as bacterial genome assembly, large scale sequence search[46,47], sequence annotation[48,49] and clustering[50–52] provide the possibility to compare of ARG-loci in thousands of genomes from thousands of bacterial species. Being able to predict open reading frames (ORFs) in an ARGs genetic environment and sequence annotation with comprehensive or specialized reference databases, e.g containing sequences of ARGs or IS[53–55], allow us to assess both synteny between different genomes and mobility of the ARG in each individual genomic environment from sequencing data alone, without the need for time- and resource-consuming experiments. It furthermore provides the possibility to analyze the genomes of many species that have never been isolated from clinical settings without the repeated need for difficult culturing procedures.

This discussion would be incomplete without the mention of metagenomics – the sequencing of DNA obtained from not a single cell, but a community. The great potential of metagenomics for researching the origins of mobile ARGs is that it

completely circumvents the need for culturing. In theory, we can assemble the genomes of rare or uncultivatable species from metagenomes and use them in our analyses. In practice, there are major restrictions imposed by a number of parameters. To identify rare bacterial species and be able to assemble them, high sequencing depth is required, which is, despite decreasing sequencing prices, still costly. Related to high sequencing depth are the great computational resources and amount of time required to completely assemble deeply sequenced metagenomes. The main difficulty however is related to the nature of mobile ARGs – they are highly conserved and often exist in multiple genomic contexts. Due to the short read lengths generated by common sequencing methods (e.g. Illumina), most assemblers use algorithms in which individual reads are assembled based on overlap with previous reads. When assembling a gene that is present in different contexts (which may not only be true for the ARG itself, but also for mobile elements flanking it) in the same sample, this leads to multiple options as to which read 'fits' the assembled sequence. Since it is not possible to deduce which potential fit corresponds to the reads true genetic context, the assembly stops at this point – leaving the ARG on a short continuous sequence (contig) that carries little to no information about its state of mobility or its genetic environment. Though such repetitive regions can accurately be resolved by third generation long-read sequencing approaches, the sequencing volume needed for attempting to cover metagenomic communities is costly to date.

As more and more genomes from different species will be added to public repositories in the future, the possibility to identify origins of mobile ARGs from these data will grow as well. Therefore, the development of computational methodologies and frameworks to analyze whole genome sequencing data with respect to the origin of mobile ARGs will greatly facilitate their identification, contributing knowledge that can be used in the mitigation of antibiotic resistance.

2. AIMS OF THE THESIS

The overall aim of this thesis is to generate knowledge about where mobile ARGs are mobilized from, how they make their way into clinical isolates and which environments may play a role in these processes. To achieve this, the following aims are addressed in this thesis:

- To develop methods and tools for reliably identifying the origin of a mobile resistance gene
  (Papers I-VI)

- To identify the origins of single resistance genes, using the above methods
  (Papers I-IV)

- Identify patterns (if there are any) pointing towards bacterial taxa or environments that may play a role in the emergence of mobile ARGs
  (Paper V)

Through contributing with such knowledge, this thesis aims to pave the way to gaining more understanding about how we ultimately could act to reduce risks for the emergence of novel resistance genes into the clinics.

3. METHODS

3.1 DNA Sequencing

3.1.1 Background on Whole Genome Sequencing

The field of whole genome sequencing (WGS) has evolved rapidly in the recent years, producing several high throughput techniques for the analysis of prokaryotic genome data. Though we have not produced sequencing data ourselves during this thesis, the reliance of the here presented results on publicly available genome data, predominantly produced using next generation sequencing (NGS) technologies, requires an overview over short and long read sequencing techniques that gave rise to these data. Though there are many more platforms and providers available than described below, the following paragraphs describe the most frequently used ones.

3.1.2 Next generation sequencing

The trademark of NGS is the ability to process millions of DNA fragments in parallel[56]. While several providers and techniques are available, the Illumina sequencing platforms, providing high throughput methods for generating a large number of sequenced DNA fragments (up to 6 Tb on NovaSeq 6000) in a relatively short amount of time, dominate the market.

Illumina HiSeq and MiSeq, among the most commonly utilized Illumina platforms, differ from each other in number of reads produced per time unit and read-length (with a maximum of 300bp on MiSeq). Reads are produced using a sequencing-by-synthesis approach: The input DNA is randomly fragmented into pieces of a certain length, which are then combined with adapter sequences. The resulting DNA fragments are then attached to the surface of a glass flowcell, where each fragment is replicated through bridge amplification, leading to the generation of clusters of identical fragments at the same location on the flowcell. Now, primers and deoxyribonucleotide triphosphates (dNTPs) are added and DNA polymerase begins the synthesis of the complementary strand. The dNTPs are labeled with a reversible fluorescent ‘blocker’, that only allows the incorporation of one dNTP into the complementary strand – once that dNTP is incorporated and all remaining dNTPs washed away, the fluorescence of the incorporated dNTP reveals which nucleobase was incorporated. The fluorescent blocker is then chemically removed and another round of synthesis commences, until the fragment is fully sequenced[57]. While the generated reads are extremely accurate (with an error rate <0.1%), quality has been shown to decline in GC-rich regions. Another

difficulty for downstream analysis is the assembly of repetitive regions due to the relatively short read length.

3.1.3 Third generation sequencing

Third generation sequencing (or long-read sequencing) removes the need for DNA amplification, and produces reads that are multitudes longer than those of NGS platforms, though they currently have a lower throughput and a higher error rate. Different methodologies for the generation of reads have emerged, the most established and noteworthy at the time of writing being SMRT (single molecule real time) sequencing and Nanopore sequencing.

Applying, similar to Illumina, a sequencing-by-synthesis approach, Pacific Biosciences' (PacBio) SMRT sequencing technique generates reads with over 60kbp in length. To start the process, double stranded DNA is circulated using hairpin adapters. The construct is immobilized on the SMRT Cell, which contains a number of small wells called zero-mode waveguides (ZMW). The DNA construct is fixated in the ZMW, via a single polymerase bound to the bottom of the ZMW, which binds to the hairpin adapters. Distinctly fluorescently labeled dNTPs are the added to the SMRT Cell, and the polymerase starts the replication process. As dNTPs are incorporated, each emits a pulse of fluorescence in real time, indicating which dNTP was incorporated[58]. The number of generated reads depends on the used system, with the PacBio RSII producing about 55000 reads per SMRT cell, and the PacBio Sequel producing about 365000 reads per cell[59]. The throughput of PacBio systems is thus much lower than that of Illumina systems. Another drawback of the traditional PacBio systems is the error rate of the generated long reads, which can be up to 15%. These errors are however randomly distributed and can be corrected to <1% if the coverage is high (e.g. HiFi reads), but high coverage comes at cost of read length, as the lifetime of the polymerase is finite.

Nanopore sequencing, such as applied by Oxford Nanopore, sequences DNA by measuring changes in electric current as a DNA molecule is threaded through a nanopore, where each nucleobase causes a specific disruption in the current[60]. While SMRT sequencing produces reads >60kbp, the longest reported reads for nanopore sequencing exceed 2Mbp. In 2015, Oxford Nanopore launched a commercially available USB-sized, portable sequencer, the MinION, making long-read sequencing affordable for even small laboratories. A MinION flow cell currently contains 512 channels, meaning that 512 DNA molecules can be sequenced at the same time. Sequencing results can be obtained in real time, making the device interesting for use during epidemics or clinical diagnostics.

Irrespective of approach, a significant advantage of long reads is the ability to sequence long repetitive regions without the need for assembly later on, which greatly facilitates the study of e.g. mobile antibiotic resistance genes in their at times highly mosaic contexts. Improvements in error rates of long read sequencing methods have been significant since their early days, and are expected to continue in the future, making those techniques highly relevant for bacterial genomics.

3.2 Assembling bacterial genomes

Genome assembly is the process of joining the reads obtained from sequencing into longer, contiguous sequences (contigs), with the goal of reconstructing the genome of the sequenced organism. Assembly can be attempted *de novo*, meaning from scratch with only the reads to work with, or using a reference genome. Before the advent of long read sequencing techniques, bacterial genomes were assembled purely from short reads. Under the assumption that highly similar reads originate from the same genomic locus, several approaches, generally relying on overlap between single reads for genome assembly were developed, such as Greedy algorithms taking into account only the locally best matches, overlap-layout-consensus (OLC) algorithms sorting overlapping reads into matching pairs that subsequently are organized into graphs, or De Brujin graph-based algorithms utilizing substrings of reads to build a graph that is resolved with help of whole-length reads[62]. Due to the short length of short reads however, it is difficult to completely assemble bacterial genomes without gaps. Repetitive regions longer than the read length (such as rRNA operons or e.g IS present in several genomic locations) cannot effectively be resolved, such that the final assembly will be fragmented, consisting of several large contigs. Long reads generated by single molecule sequencing techniques effectively can solve this problem, as they can cover those regions completely, but they have a high error rate that is difficult for assemblers to handle. One solution to these problems are hybrid methods, to use for example highly accurate short reads for correction of long reads, or use both types of reads in hybrid assemblies. This is however costly and time consuming, as two DNA libraries have to be sequenced instead of one. To circumvent such problems, protocols such as the hierarchical genome-assembly process (HGAP) have been developed. This self-correcting process for long reads uses the longest reads as seed sequences, that are used to correct and preassemble all reads into highly accurate long reads that can then be assembled by suitable assemblers (such as e.g. Celera). This method has been shown to be suitable for the accurate *de novo* assembly of genomes using purely long reads[63].

3.3 Reference Databases

3.3.1 NCBI Assembly and RefSeq databases

The National Center for Biotechnology Informations' (NCBI) Assembly database contains unique identifiers and metadata for a set of assembled sequences that compromise a genome. Assemblies are classified into four subgroups, based on the degree of assembly: contig level assemblies, scaffold level assemblies, chromosome level assemblies and complete genome assemblies. Stored metadata for each assembly contain statistics such as sequence length, number of contigs, who submitted the genome, organism specific information and more. It contains assemblies from the International Nucleotide Sequence Database Collaboration (INSDC), and the NCBI RefSeq database, which is a curated, non-redundant set of protein, DNA and RNA sequences. The data are regularly updated and downloadable from the NCBI FTP site. At the time of writing, the assembly database contains 869264 bacterial genome assemblies. Plasmids are only included in the Assembly database if they are associated with a chromosome record. In this thesis, we used the NCBI Assembly database to obtain genome assemblies, and the curated plasmid sequences available from RefSeq at ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plasmid/ (all papers).

3.3.2 Antibiotic Resistance Gene Databases

Many ARG databases have arisen throughout the years, such as ARDB[64], CARD[54], ResFinder[65], ARG-annot[66] or MEGARes[67]. Some contain specific subsets of ARGs, such as ResFinder which contains only acquired ARGs, others are made for specific types of data, such as MEGARes, which specializes in metagenomics data. Apart from being classic databases containing sequence information, many also provide tools to identify ARGs in genomic data. In this thesis, we used CARD and Resfinder for large-scale identification of mobile ARGs from sequences, because of their comprehensiveness and structured annotation. CARD is a comprehensive, actively curated and highly structured database containing sequence information and metadata, supported through a structured Antibiotic Resistance Ontology (ARO), on intrinsic ARGs and dedicated ARGs as well as mutations. It also provides identification of ARGs in genomic data via its resistance gene identifier (RGI). In this thesis, we mostly used the sequences provided by CARDs protein homology model, in which most mobile ARGs are included (paper I, II, III and V). In paper VI, we searched the ResFinder database against CARD, to create a CARD subset containing only mobile ARGs, in order to make use of CARDs structured sequence annotations.

3.3.3 Genomic environment annotation – UniProtKB and ISFinder

Annotating an identified ARGs genetic environment, meaning the sequences surrounding it on a given locus, is essential in this thesis – The possibility of identifying a mobile ARGs origin is dependent on being able to differentiate between ARGs associated with mobile genetic elements and ARGs that are not associated with such elements. In order to do so, we make use of mainly two databases. The UniProt knowledgebase is a resource containing over 60 million protein sequences, which are mostly derived from the translation of nucleotide sequences submitted to the INSDC databases. UniProtKB consists of two sections: Sequences contained in the first section 'UniProtKB/Swiss-Prot' are manually curated and reviewed, whereas sequences contained in 'UniProtKB/TrEMBL' are annotated by an automatic pipeline[68]. This large repository of protein sequences is highly suitable to annotate genetic environments obtained from a multitude of different genomes.

ISFinder is a specialized database, focusing on bacterial IS. It is the most comprehensive sequence repository for ISs to date, the curators also assign name to novel ISs using a coherent naming system and provide background information on ISs and transposable elements. Sequences of novel ISs are submitted by the scientific community and curated by the authors[55]. The database is not available for large scale analysis and can only be accessed via the ISFinder website (www-is.biotoul.fr), which provides tools to search for IS in provided sequences. In this thesis, ISFinder was used to manually investigate the presence of IS in the vicinity of ARGs, or to investigate the identity of single transposases/IS-like genes (papers I-V).

3.4 Sequence Annotation

3.4.1 Sequence comparison with BLAST and DIAMOND

In order to identify ARGs in genomic data and annotate predicted ORFs, a method to evaluate similarities between two sequences is required. The Basic Local Alignment Search Tool (BLAST), is commonly used for such tasks. BLAST produces local alignments between two sequences using a seed-and-extend algorithm – The sequences in the reference database are split into smaller substrings (sometimes called k-mers), which are then searched against a query sequence. The algorithm tries to extend the matches of substrings in the query sequence (called seeds) based on the reference sequence, using a substitution matrix to assess the quality of the alignment, taking matches, mismatches and gaps between query and reference sequence into account[46] in

order to produce a local alignment. The NCBI provides an online platform hosting several BLAST algorithms for the comparison of different subject and query sequence types (e.g. BLASTN for nucleotide-nucleotide comparisons, BLASTX for protein-protein comparisons). However, BLAST is slow when comparing large numbers of sequence pairs, making it unfeasible for the large scale analyses conducted in this thesis, such as trying to identify ARGs in several hundred thousands of genomes. Therefore we used DIAMOND, an algorithm using a BLAST-like approach, but with significant speed improvement over BLAST. This improvement is achieved through use of a double-indexing algorithm that locates the seed sequence and their positions in both subject and query, whereas the traditional BLAST algorithm scans queries linearly with indexed subject seeds. Instead of 'traditional' seeds, DIAMOND uses 'spaced seeds', which are longer but do not use all positions in the seed sequence, in order to increase speed but maintain sensitivity. DIAMOND was, because of its speed and sensitivity, the most suitable sequence search algorithm for this thesis and was used in all included articles and manuscripts. Which identity threshold to use with such search algorithms highly depends on the research question. In order to identify ARGs and ARG-like genes more closely related to those observed in the clinics, we used identity thresholds >70% towards the reference sequence, as ARG-like genes above this threshold might contain clues about not only the taxonomic distribution of those genes, but also their more recent evolutionary history. When attempting to annotate genes in and ARGs/ARG-like genes environment, we used cutoffs between 40 and 60%. The rationale behind these relatively low cutoffs was the goal of reducing the number of hypothetical proteins and at the same time estimating the surrounding genes function, to see if they may be involved in mobility of the locus in some way.

3.4.2 ORF identification using Prodigal

In this thesis, we used the Prokaryotic Dynamic Programming Gene-finding Algorithm (Prodigal) to predict ORFs in sequences flanking mobile antibiotic resistance genes. The rationale behind predicting and collecting short ORFs and comparing these to reference databases is a decrease in computation time compared to searching protein databases against whole genomes, as it allows for example for deduplication of sequences. Prodigal identifies prokaryotic genes based on a general set of rules created through examination of over hundred bacterial genomes from GenBank. These rules include start codon usage, ribosomal binding site (RBS) usage, maximum gene overlap and more. Start and stop codons are identified in the input sequence to be used as start and end of possible ORFs, and a frame bias model is built based on G/C positions in the different codons. During both a training and a final gene calling phase, Prodigal uses a

dynamic programming approach to evaluate different parameters and decide which ORFs most likely correspond to true genes.

3.5 Sequence alignments using MAFFT and MUSCLE

Sequence alignment is a method for identifying similar regions in DNA, RNA or protein sequences between two or more sequences. Used on its own or as the basis for phylogenetic analysis, sequence alignment may contain information about the evolutionary history of the respective sequences. In this thesis, we mainly used MAFFT to produce multiple sequence alignments prior to phylogenetic analysis (all papers). MAFFT is a commonly used tool for producing multiple sequence alignments, and is continually updated as new features are implemented. Alignment of even large sets of long sequences using MAFFT is relatively fast, due to the implementation of fast Fourier transform (FFT), which transforms sequences of amino acids into sequences of the volume and polarity of each amino acid residue in order to identify regions of similarity. Furthermore, a simplified scoring system reduces computing time and increases the accuracy for global alignments, even if the sequences differ in length [69]. Since 2018, MAFFT has parallelized some calculations, further increasing the speed for calculating large sequence alignments[70]. In paper I, multiple sequence alignments (MSA) were created using MUSCLE. Similar to MAFFT, MUSCLE uses a progressive alignment strategy in which input sequences are placed within a tree, created from a distance matrix that is computed based on similarity between pairs of sequences. The similarities are computed either by k-mer counting or a global alignment of the sequence pair[71].

3.6 Sequence clustering using CD-HIT and USEARCH/UCLUST

Sequence clustering is the process of sorting sequences into groups (clusters), based on sequence similarity. Clustering can be used to both investigate the degree of similarity between members of a protein family (e.g paper III), or to decrease the number of sequences to investigate in order to minimize computational time needed for further analyses (all papers). In this thesis, we used two different clustering approaches for different purposes: CD-HIT was used to cluster large protein databases like the above described UniProtKB, whereas USEARCH/UCLUST was used to cluster longer DNA sequences. CD-HIT is a fast, greedy clustering algorithm that uses short-word filtering as a means of identifying similar sequences. Short word (or k-mers, describing words of length k) filtering assumes that the number of k-mers common to two sequences is a function of their similarity – thus, the similarity is estimated based on the number of shared k-mers[50]. Though CD-Hit does not always find the most accurate clusters, it

makes up for its potential lack of accuracy in speed. While USEARCH also calculates the number of common k-mers between two sequences, it does not estimate identity between the two sequences based on the number of k-mers, but uses this number to prioritize which database sequences are compared to the query sequence[52]. The algorithms also use different scoring systems, where e.g mismatches and gaps are penalized differently. This leads to CD-HIT producing alignments with higher identities, potentially grouping sequences into one cluster that are slightly below the identity threshold. As in our approach the clustering of DNA sequences is meant to reduce redundancy in a set of sequences, USEARCH is better suited for that purpose, as CD-HIT may (though not especially likely) remove non-redundant sequences. In both algorithms, sequences have to be sorted by length and are then processed by decreasing length – each sequence is compared to the first one (called centroid) and becomes part of the cluster if its identity is above the specified threshold. Otherwise it becomes a centroid to which the remaining sequences are then compared.

3.7 Phylogenetic analysis

Phylogenetic analysis, often conducted through the construction of phylogenetic trees, is used to investigate evolutionary relationships of different subjects, which can be whole organisms, or merely DNA/protein sequences. In this thesis, we used phylogenetic analysis as a complementary measure to investigate the evolutionary relations between mobile ARGs in different organisms (papers I-III) and as a form of anchoring similar sequences together in the sequence visualizations (papers V and VI). To create precise phylogenetic trees from shorter sequences, such as single genes or proteins, we used the RAxML (Randomized Axelerated Maximum Likelihood) tool. RAxML uses a maximum likelihood (ML) approach, in which probability distributions are inferred on a range of possible phylogenetic trees in order to obtain the one that is most likely to represent the true evolutionary relationships of the sequences[72]. The method requires a substitution model, specifying the mutation rates of the input sequences, in order to infer probability to different trees. The general time reversible model (GTR) used in the phylogenetic analyses in this thesis, is a commonly used model and assumes different substitution rates and frequencies for each nucleobase. Another advantage of ML is that it allows for varying mutation rates across sequences, which is especially relevant for horizontally transferred ARGs. In order to obtain support values for specific branches of a tree, a process called bootstrapping can be used. During bootstrapping, confidence values for different clades are calculated from trees that are created from random subsamples of the input sequences.

For the calculation of phylogenetic trees from large numbers of sequences, we used FastTree, for its advantage in speed over RAxML. FastTree uses an ML approximation

approach and achieves a decrease in computation time by implementing heuristics coupled with neighbor-joining, nearest neighbor interchanges and bootstrapping.

3.8 Taxonomic classification of genomes

Misclassification of bacterial genomes is not uncommon in public sequence repositories, and often there is limited information on how specific genomes were classified. In this thesis, confirmation of the origin of a mobile ARG requires comparison of the mobile ARG-locus with the ARG locus of several members of the suspected origin species (if available). Irregularities in the results of such comparisons required reclassifications on several occasions (paper I-III). Potential classification methods, such as comparison of the universally conserved 16S rRNA genes or a combination of marker genes, may lack sensitivity for classification at species level, are susceptible to sequencing errors, errors in the reference databases[73] or incompleteness of genome assemblies. Therefore, we used ANIcalculator (paper II and III), a tool implementing a classification approach utilizing the combination of genome-wide average nucleotide identity (gANI) and alignment fraction (AF, describing the fraction of orthologous genes between two genomes) as a measure of relatedness between two genomes. For the calculation of the gANI, the sum of the nucleotide identities of shared genes is multiplied by the alignment length of the shared genes, divided by the cumulative length of all shared genes. The AF is calculated by dividing the sum of the length of all shared genes through the sum of the length of all genes. Using over 1 million genome pairs, the authors determined thresholds for assigning genomes to the same species, an AF >0.6 and gANI >96.5. These correlated well with traditional classification methods, such as 16S distance[74]. In paper I, we used dRep, which utilizes gANI for accurate genome comparison[75], and comparison of 16S signature nucleotides for genus assignment of *Rheinheimera* and *Pararheinheimera* genomes – due to lack of genomes for comparison.

4. RESULTS AND DISCUSSION

In this thesis, we identified the origins of several mobile antibiotic resistance genes exclusively from WGS data available from public sequencing repositories, using *in silico* comparative genomic methods, such as the large scale analysis and comparison of the flanking regions from ARG/ARG-like loci. Based on these findings and the summarized literature on the origins of mobile ARGs, we were able to formulate a framework containing criteria for the identification of the evolutionary recent origins of mobile ARGs, which can be used with both *in silico* and traditional molecular methods. We were to the best of our knowledge the first to analyze patterns in the to-date identified recent origin species, and finally provide a software that enables visual comparison of hundreds of gene loci at the same time. Thus, this thesis contributes to understanding from where and potentially under what conditions ARGs are mobilized from their origins' chromosome to mobile vectors.

4.1 Using WGS data to identify the origins of mobile ARGs

In paper I, we identified the origin of the *bla*$_{\text{PER}}$-type genes, a class A beta-lactamase causing resistance to certain groups of beta-lactam antibiotics. Under the working hypothesis that the regions flanking the mobile ARG would also be found in the ARGs original location, as shown before (e.g. Jacoby, Griffin, and Hooper 2011), we searched all genomes from The GenBank Assembly database for *bla*$_{\text{PER}}$-like genes and annotated and compared their genetic environment. This led to the identification of the genus *Pararheinheimera* as the origin of *bla*$_{\text{PER}}$-like genes, despite the availability of only three genomes at the time of writing. Furthermore, the genus *Pararheinheimera* had been recently split from the genus *Rheinheimera*, which required us to try to reclassify the three *bla*$_{\text{PER}}$-positive genomes based on the availability of 16S rRNA data. As we identified *Pararheinheimera* genomes that did not carry *bla*$_{\text{PER}}$-like genes, assessing the mobility of *bla*$_{\text{PER}}$ genes in the *Pararheinheimera* genomes was not only based on annotation of the genes genetic environment, but also on the phylogenies of several chromosomal genes to exclude any recent HGT of the *bla*$_{\text{PER}}$-like genes into *Pararheinheimera*. Despite the high nucleotide identity of 96% of the *Pararheinheimera* sp. KL1 *bla*$_{\text{PER}}$-like gene and its immediate genetic environment towards the clinical *bla*$_{\text{PER-1}}$ locus, we could not assign a species origin at the time of writing, simply because only one *bla*$_{\text{PER}}$-positive *Pararheinheimera* genome was assigned to a species – but ANI analysis did not suggest that *Pararheinheimera* sp. KL1 belonged the same species. Since then, more *Rheinheimera* and *Pararheinheimera* genomes have become available, and gANI and AF analysis shows that *Pararheinheimera* sp. KL1 shares 97.2% gANI and 0.85 AF with the genome of

Rheinheimera tangshanensis (GCA_008017875.1), which has recently been reclassified as *Pararheinheimera tangshanensis* [77]. Thus, based on established gANI and AF cutoffs[74], *Pararheinheimera* sp. KL1 appears to be *P. tangshanensis,* which most likely is the origin of mobile *bla*$_{PER}$-genes. However, more *P. tangshanensis* genomes are needed in order to further verify the classification of *P.* sp. KL1 as *P. tangshanensis*. To our knowledge, this was the first article utilizing purely bioinformatics analyses to identify the origin of a mobile resistance gene.

As our approach proved feasible, we next investigated the origin of the mobile AmpC beta-lactamases of the CMY-1/MOX family (paper II). Though some evidence was pointing towards the genus *Aeromonas,* the exact origin of these genes had not been resolved yet. Several variants that had been reported displayed relatively large sequence divergence to one another, indicating that they did not originate from the same species. Based on synteny, nucleotide identity comparison and phylogenetic analysis, we identified three distinct species of *Aeromonas* as the origins of three distinct CMY-1/MOX variants. As had been the case for previously determined origins and their mobilized genes (e.g paper I), the nucleotide identities of the mobile ARG locus and the ARG-like locus on the origin chromosome were nearly identical, and some mobile ARGs were associated with truncated genes co-mobilized from their original locus. This evidence of repeated mobilization of the chromosomal *Aeromonas* AmpC led us to hypothesize about the conditions that may favor such mobilizations – the scenario involving the least intermediate steps from origin to human commensals/pathogens being *Aeromonas* infection in humans/domestic animals treated antibiotics. This scenario involves both the selection pressure needed for IS/IS*CR* mediated mobilization and direct transfer possibility to human associated bacteria.

Having shown the most likely origin of the mobile MOX-2 AmpC in *A. caviae*, we investigated the origins of the mobile FOX-type AmpC in paper III, as these were also reported to have originated in *A. caviae*. The large degree of nucleotide divergence of *bla*$_{MOX-2}$ and *bla*$_{FOX}$-type genes however made this hypothesis unlikely, based on our observations from the literature and our previous studies, in which the origin loci of mobile ARGs were highly similar to their mobile counterparts. Comparing >230 *Aeromonas* AmpC-loci, we showed that the mobile *bla*$_{FOX}$ genes originate from *A. allosaccharophila*, and not from *A. caviae*. If we are to use the knowledge about the origin of today's mobile ARGs for mitigation purposes in the future, we have to understand under which conditions they are mobilized. To identify such conditions we also need to know in which habitats their origins thrive. For that, it is essential to know exactly from which species which ARGs have emerged. In order to see whether there are common patterns regarding e.g. the environments ARG origins are found in, we need to find as many origin species as possible.

While our comparative genomic approach enables the study of thousands of genomes without the need for further culturing, and the amount of different genomes in public databases is growing rapidly, it is obviously limited to what genomes are contained in the database at the timepoint a specific study is conducted. This may result in a lower taxonomic resolution than species level, such as in paper I at the time of writing. In paper IV we identified the genus *Shinella* as the origin of two recently described mobile class A carbapenemases, BKC-1 and GPC-1. Though the genetic environment of the BKC-1/GPC-1-like genes in different *Shinella* sp. genomes was quite conserved and displayed no signs of mobility, the nucleotide identities between the mobile genes and the chromosomal counterparts were about 90% at most for GPC-1 and 87% for BKC-1. A duplication within the BKC-1 sequence, that was not present in the BKC-1-like *Shinella* enzymes, was shown to lead to greater activity of the enzyme against beta-lactams, but it is unclear whether the duplication is present in the origin genome as well or whether it is a result of beta-lactam selection pressure on a already mobilized gene. The sequence similarity is much lower than what was described for other chromosomal/mobile gene pairs, suggesting that the genes either have not been mobilized recently, or that that (more likely) the genome of the species origin of these genes is simply not present in the NCBI Assembly Database yet. Still, identified origins on genus level provide important clues as to where the species origin of a certain gene will be found most likely. The results from papers I-IV show that our comparative genomics approach is suitable for the identification of the origins of at least some mobile ARGs.

4.2 Finding patterns in the origins of mobile ARGs

The ultimate aim of this thesis is to contribute towards providing insights that ultimately could be used to limit the emergence of novel forms of antibiotic resistance in pathogens. While it is important to identify single origins of mobile ARGs to build a knowledge base, for directing actions to reduce risks for future mobilization events, it is most important to see whether there are characteristics that the origin species of mobile ARGs have in common. In paper V of this thesis, we conducted a thorough literature research to identify articles proposing the origins of ARGs. Using the experience gained from the previously described work and insights from the literature, we formulated criteria which can be used to assess whether a bacterial species may be the origin of a certain mobile ARG. After scrutinizing each proposed origin with help of these criteria (amending data where needed), we then analyzed similarities among the curated, confirmed origins (on species level). As the term origin is used quite ambiguous, we found it necessary to define ‘recent origin’ in the context of this article as the *“bacterial taxon in which the ARG is widespread but commonly not associated*

with any of the MGEs that likely played a role in the ARGs transition into its clinically relevant context(s)". The following characteristics were common among many recent origin species:

- The great majority of mobilized ARGs had been mobilized by IS/IS*CR* elements, many retaining non-ARG sequences from their 'native' environment

- Nucleotide identities between the non-mobile ARG progenitor locus and the mobile ARG-locus were usually ≥96%

- Variants from many ARG-families appear to have been mobilized repeatedly and independently from one another (e.g. CTX-M, SHV, CMY-1/MOX, QnrB)

- All of the identified origins are Proteobacteria

- The great majority of origins has been associated with infection in humans or domesticated animals
- None of the origin species is known to be an antibiotic producer

Though the mobile ARGs which have an origin assigned based on this study account for only roughly 4% of all known mobile ARGs, these similarities may already hint towards what conditions and environments could contribute to the mobilization of at least a subgroup of mobile ARGs. The association with IS/IS*CR* elements provides high mobility, as mobilized ARGs can move to different chromosomal locations or plasmids. Furthermore, IS/IS*CR* elements have been shown to increase the expression of adjacent genes, which may transform the role of a non-(clinical) resistance providing enzyme into that of an ARG, given the right enzymatic properties. Based on that, antibiotic selection pressure in the origin species' direct environment is likely crucial for the mobilization and subsequent selection of an ARG. This is supported by the observation that the great majority of known recent origin species are at least sometimes associated with disease in humans/domestic animals – which may trigger antibiotic use, thus selecting for the IS/IS*CR*/ARG combination.

Though bacteria in certain ecological niches (e.g. soil) have been exposed to antibiotic selection pressure by antibiotics since ancient times[78], the selection pressure exerted by clinical antibiotic concentrations or antibiotic pollution at manufacturing sites most likely greatly exceeds what is present in natural environments. ARGs have likely been transferred horizontally even before the antibiotic era [79], but there is evidence that ARGs on plasmids were uncommon in human pathogens at that time[23], whereas they can be found in almost all kinds of environments today[24]. Is their presence now the result of selection of ancient genes that have been mobilized to MGEs long before the antibiotic

era, and are selected for only now, or are novel ARGs constantly mobilized from the chromosomes of modern bacterial species? And if both scenarios occur, is one more common than the other, and thus more relevant for mitigation? The high nucleotide identities of both mobile ARGs and co-mobilized genes in their immediate genetic environment speak for the latter – the co-mobilized genes are in many cases truncated, and do thus likely not contribute to the resistance function of the mobilized DNA unit. These sequences would not be present on MGEs (at least not with such a high degree of identity) if they had been mobilized there in ancient times, as there is no obvious selection pressure acting on them. Thus, the association and subsequent mobilization of these ARG progenitors to MGEs where they act as resistance determinants, is likely as recent as during the antibiotic era. Recent mobilization of ARGs to MGEs has already been suggested by other authors[78,79], but verifying an exact time point is difficult. Phylogenetic methods for estimating time spans across phylogenetic events exist and have been applied to estimate the emergence time point of mobile ARGs such as MCR-1[80], based on sequence alignments of a great number of mobile MCR-1 loci and assumptions of different molecular clocks. The resulting estimate approximately fits with other data, such as when certain mobile variants were detected, but the approach assumes a single emergence event of mobile MCR-1. Though this may be accurate in the case of MCR-1, the data analyzed in this thesis suggest that many ARGs have been mobilized more than once, from either one or several closely related species. Thus, estimating emergence time based on comparison of mobile loci to the origin locus appears less feasible to give a reliable estimate, as multiple mobilization events and absence of the exact origin strain may confound the analysis.

For some ARGs, there is evidence that resistance genes that are now fixed on the chromosome of certain taxa have been acquired by those taxa before the genes mobilization into clinically relevant contexts (e.g. *bla*$_{\text{PER}}$-type genes, paper I), others have been suggested to have been plasmid-borne at several time points in their ancient evolution[79]. To distinguish these ancient origins, referring to where an ARG progenitor evolved, from the taxon from where a resistance gene was mobilized into clinically relevant contexts, we introduced the term ‘recent origin’ for the latter case.

Surprisingly, all identified recent origins were Proteobacteria, none of them a known producer of antibiotics. The respective mobilized ARGs from these recent origins are mostly found in proteobacterial pathogens as well, so this finding, which is at least for this set of mobile ARGs in contrast to the producer hypothesis, may highlight transfer barriers between more distantly related taxa. This could suggest that mobile ARGs are mostly transferred between taxa that are more closely related, as e.g. genes that have evolved in one taxon may not function efficiently in a too genetically distant receiver[81]. However, most of these recent origin species might also have been discovered precisely because they are Proteobacteria and research has been biased towards this taxon as they

are frequent human pathogens – they have been studied and sequenced extensively, whereas other organisms, which are not of (as much) clinical relevance, are less investigated. The overrepresentation analysis we conducted showed however that the majority of origin species being associated with infection in humans or domestic animals was not a product of database bias – but that a proteobacterial species that is infectious had higher odds of being an origin than non-infectious proteobacterial species. This suggests the human/animal body undergoing antibiotic treatment as a potential hotspot for the mobilization of novel resistance genes. Factors that likely are important for the emergence of a novel mobile ARG are present in the body – Antibiotic selection pressure, potentially pathogenic/commensal recipients, and mobile genetic elements that can serve as vectors for the novel ARG.

It remains to say that, to the best of our knowledge, no origin is known for genes conferring resistance to certain classes of antibiotics, such as tetracyclines or macrolides. The most likely explanation for this is in our opinion that their origins have not been sequenced yet, though it is theoretically possible that their mobilization happened a long time ago and the mobile variants and their original locus have diverged to such an extent that it is no longer possible to identify a recent origin.

4.3 GEnView – comparing the genetic environment of target genes from hundreds of genomes

In the previously discussed articles, we have shown that comparing the immediate genetic environments of mobile ARGs is a powerful way to identify their origins. However, the concept is not only applicable to ARGs, but can be applied to any kind of sequence in order to investigate a multitude of research questions, from the conservation of loci among different taxa towards the contexts of any gene across a range of taxa. In paper VI of this thesis, we describe GEnView, a software that enables the visualization of the contexts of user-specified genes in hundreds of bacterial genomes in a single figure. Through accessing the NCBI Assembly and RefSeq databases, GEnView can search for genes in to date >800.000 genomes and extract the immediate genetic environments of the target gene. ORFs are predicted and annotated using large reference databases such as UniProtKB and visualized using phylogenetic trees to cluster similar sequences together. Besides the visualization, GEnView provides additional files, containing e.g. the extracted sequences in FASTA format or annotation metadata such as position and sequence of each annotated gene, enabling more detailed manual sequence comparison. In order to reduce storage requirements and computational load, users are able to subset the data through specification of desired taxa or providing a custom list of assembly accessions. GEnView will greatly facilitate the comparison of

specific genomic loci in a large number of genomes, enabling researchers to utilize the rapidly growing number of bacterial genomes in public databases.

5. CONCLUSION

In papers I-IV, we developed and used a comparative genomics approach to identify the recent origins of mobile antibiotic resistance genes. We identified *Pararheinheimera* sp. as the origin of PER-family beta-lactamases, the *Aeromonas* species *A. sanarellii, A. caviae* and *A. media* as the origins of CMY-1/MOX-1, MOX-2 and MOX-9 AmpC beta-lactamases, *A. allosaccharophila* as the origin of FOX-family beta-lactamases and the genus *Shinella* as the likely origin of GPC-1 and BKC-1-family carbapenemases. We thus managed to contribute knowledge on from which bacteria mobile ARGs have been mobilized.

In paper V, we summarized, scrutinized and set the current literature on the origins of mobile ARGs into context, and identified several patterns that most origin species had in common. These led us to hypothesize that the analyzed ARGs were mobilized during the antibiotic era, and indicate that the human/animal body during antibiotic treatment may be a hotspot for the emergence of novel ARGs. The fact that we do not know >90% of the recent origins of known mobile ARGs points to unsequenced environmental bacteria as the origins of clinically relevant ARGs. We also formulated a framework that will aid the confident identification of the origins of mobile ARGs in the future. To correctly identify the origins of mobile ARGs is crucial if this knowledge is to be used in the mitigation of future ARGs.

In paper VI, we developed GEnView, a comparative genomics software that enables researchers to visually compare of hundreds of gene loci simultaneously, and provides data for further in depth analysis of the respective loci.

6. FUTURE PERSPECTIVES

The number of bacterial genomes and plasmids in public sequence databases is rapidly increasing, which in turn will increase the possibility to identify the origins of mobile ARGs using the framework and methods provided in this thesis, as the lack of origin genomes is the main limitation to date. The resulting increase in the number of identified origins will provide further insights into the patterns that we began to get a glimpse of in this thesis – Are the majority of recent ARG origins indeed Proteobacteria, or will we identify antibiotic producing species as recent origins of mobile ARGs as well? Does infection or connection with the human/animal body indeed increase the chance of a taxon to become the recent origin of a mobile ARG? What natural barriers limit the spread of newly mobilized ARGs? Decreasing costs of long-read sequencing techniques in the future may make the use of such methods in metagenomics studies more feasible, and enable the assembly of an ARGs native contexts in metagenomics samples, even in the presence of mobile variants – Though it may not be possible to assemble the whole genome of the origin taxon from such samples, this would enable the scientific community to answer another question: In which environments do the origins of mobile ARGs thrive? Alternatively, screenings of large collections of bacterial isolates could be used to identity further origins, an approach that, though more time intensive, has been shown to be successful in the past [82].

In order to use our knowledge about the recent origins of mobile ARGs for mitigation purposes and asses risks associated with different environments, it is crucial to know which environments these species can thrive in and in which abundances they are found in different environments. To date, it is unknown to what extent species that are known to inhabit e.g. the human gut, such as *Klebsiella pneumonia* or *Citrobacter freundii*, thrive in other environments. Based on the abundance of origin species and other factors, such as e.g. antibiotic selection pressure or ecological connectivity, certain environments are likely to present a higher risk for the emergence of novel mobile ARGs in the future than others. In order to focus mitigation efforts, we need to know which environments and conditions present the highest risk. Usage of deeply sequenced metagenomes and taxonomic classification tools available today might be able to provide insights into these follow-up questions, and enable us to try and slow down the emergence of novel mobile ARGs in clinical settings.

ACKNOWLEDGEMENTS

I am truly grateful to my main supervisor, **Joakim Larsson**, for giving me the chance of performing my PhD in his lab. Thoughtful and lively discussions on both research and life (which very much includes fishing), the possibility to bounce ideas at any time, constructive critique and your thoroughness in teaching me both science and writing provided me with the best imaginable environment to conduct my research in.

I am also indebted to my co-supervisor **Erik Kristiansson**. You saw the best in all datasets no matter how crappy they appeared to me, helped me tremendously to wade the deep and foggy marshes of advanced statistics more than once and managed to motivate me time and time again.

Johan Bengtsson-Palme, you were a constant mentor to me especially during my first two years, and I truly appreciate all support, patience, advice and thought material you have given me over the years.

There are many others I have met over those last years and that have influenced me in one way or another – my fellow (ex-) PhD students **Marion Hutinel** and **Mohammad Razavi**, my constant, patient co-author **Nadine Kraupner**, and all other present and past members of Joakims' and Eriks groups: **Carl Fredrik Flach, Marie Elisabeth Böhm, Antti Karkman, Carolin Ruttgersson, Maja Genheden, Patricia Huijbers, Nachiket Marathe, Kaisa Thorell, Declan Grey, Jenni Sjöhamn, Fanny Berglund, Nicolas Kieffer, Tobias Österlund, Carl-Johan Svensson, Jekatarina Jutkina, Sazzad Karim, Rosmarie Frieman, Anna Johnning and Fredrik Bolund.**

I also want to give a heartfelt thank you to all **my friends** which are not included in the above list, be they in Gothenburg or anywhere else in the world right now – You all had a great part in enabling me to accomplish this, keeping me sane during the difficult times and laughing (and doing whatever else we did) with me during the happy times.

Of course, I want to thank my **mother Angelika,** my **father Volker and** my **sister Susanne** for supporting me during not only my PhD, but my whole life. Whatever I do, I know that I can always rely on you.

Each of you have taught me, made me reflect and enabled me to grow during those years. My life is richer for knowing each one of you, and I truly appreciate that. Thank you!

REFERENCES

1. Davies, J. & Davies, D. Origins and Evolution of Antibiotic Resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).

2. Hutchings, M., Truman, A. & Wilkinson, B. Antibiotics: past, present and future. *Current Opinion in Microbiology* vol. 51 72–80 (2019).

3. Baltz, R. H. Renaissance in antibacterial discovery from actinomycetes. *Current Opinion in Pharmacology* vol. 8 557–563 (2008).

4. Kohanski, M. A., Dwyer, D. J. & Collins, J. J. How antibiotics kill bacteria: from targets to networks. *Nat. Rev. Microbiol.* **8**, 423–35 (2010).

5. Ahmed, M. O. & Baptiste, K. E. Vancomycin-Resistant Enterococci: A Review of Antimicrobial Resistance Mechanisms and Perspectives of Human and Animal Health. *Microbial Drug Resistance* vol. 24 590–606 (2018).

6. Gillings, M. R. & Stokes, H. W. Are humans increasing bacterial evolvability? *Trends Ecol. Evol.* **27**, 346–352 (2012).

7. Chen, L., Todd, R., Kiehlbauch, J., Walters, M. & Kallen, A. *Notes from the Field* : Pan-Resistant New Delhi Metallo-Beta-Lactamase-Producing *Klebsiella pneumoniae* — Washoe County, Nevada, 2016. *MMWR. Morb. Mortal. Wkly. Rep.* **66**, 33 (2017).

8. Xu, J., Zhao, Z., Ge, Y. & He, F. Rapid emergence of a pandrug-resistant Klebsiella pneumoniae ST11 isolate in an inpatient in a teaching hospital in China after treatment with multiple broad-spectrum antibiotics. *Infect. Drug Resist.* **13**, 799–804 (2020).

9. Lartigue, M.-F., Poirel, L., Aubert, D. & Nordmann, P. In vitro analysis of ISEcp1B-mediated mobilization of naturally occurring beta-lactamase gene blaCTX-M of Kluyvera ascorbata. *Antimicrob. Agents Chemother.* **50**, 1282–6 (2006).

10. Toleman, M. A., Bennett, P. M. & Walsh, T. R. ISCR elements: novel gene-capturing systems of the 21st century? *Microbiol. Mol. Biol. Rev.* **70**, 296–316 (2006).

11. Perry, J. A. & Wright, G. D. The antibiotic resistance "mobilome": searching for the link between environment and clinic. *Front. Microbiol.* **4**, 138 (2013).

12. Fang, L.-X. *et al.* ISEcp1-mediated transposition of chromosome-borne blaCMY-2 into an endogenous ColE1-like plasmid in Escherichia coli. *Infect. Drug Resist.* **11**, 995–1005 (2018).

13. Cattoir, V., Nordmann, P., Silva-Sanchez, J., Espinal, P. & Poirel, L. ISEcp1-mediated transposition of qnrB-like gene in Escherichia coli. *Antimicrob. Agents Chemother.* **52**, 2929–32 (2008).

14. Zong, Z. The Complex Genetic Context of blaPER-1 Flanked by Miniature Inverted-Repeat Transposable Elements in Acinetobacter johnsonii. *PLoS One* **9**, e90046 (2014).

15. P, S., E, G., A, V., B, T.-H. & M, C. Everyman's Guide to Bacterial Insertion Sequences. in *Mobile DNA III* vol. 3 555–590 (American Society of Microbiology, 2015).

16. Razavi, M. Identification of novel antibiotic resistance genes through the exploration of mobile genetic elements. (University of Gothenburg, 2020).

17. Rui, Y. *et al.* Integrons and insertion sequence common region 1 (ISCR1) of carbapenem-non-susceptible Gram-negative bacilli in fecal specimens from 5000 patients in southern China. *Int. J. Antimicrob. Agents* **52**, 571–576 (2018).

18. Lallement, C., Pasternak, C., Ploy, M.-C. & Jové, T. The Role of ISCR1-Borne POUT Promoters in the Expression of Antibiotic Resistance Genes. *Front. Microbiol.* **9**, 2579 (2018).

19. Poirel, L., Cabanne, L., Vahaboglu, H. & Nordmann, P. Genetic environment and expression of the extended-spectrum beta-lactamase blaPER-1 gene in gram-negative bacteria. *Antimicrob. Agents Chemother.* **49**, 1708–13 (2005).

20. Cantón, R., González-Alba, J. M. & Galán, J. C. CTX-M Enzymes: Origin and Diffusion. *Front. Microbiol.* **3**, 110 (2012).

21. Poirel, L., Rodriguez-Martinez, J.-M., Mammeri, H., Liard, A. & Nordmann, P. Origin of plasmid-mediated quinolone resistance determinant QnrA. *Antimicrob. Agents Chemother.* **49**, 3523–5 (2005).

22. Sun, L., Klein, E. Y. & Laxminarayan, R. Seasonality and Temporal Correlation between Community Antibiotic Use and Resistance in the United States. *Clin. Infect. Dis.* **55**, 687–694 (2012).

23. Hughes, V. M. & Datta, N. Conjugative plasmids in bacteria of the 'pre-antibiotic' era [24]. *Nature* vol. 302 725–726 (1983).

24. Pal, C., Bengtsson-Palme, J., Kristiansson, E. & Larsson, D. G. J. The structure and diversity of human, animal and environmental resistomes. *Microbiome* **4**, 54 (2016).

25. Miller, C. *et al.* SOS response induction by β-lactams and bacterial defense against antibiotic lethality. *Science (80-.).* **305**, 1629–1631 (2004).

26. D'Costa, V. M. *et al.* Antibiotic resistance is ancient. *Nature* **477**, 457–461 (2011).

27. Wright, G. D. The origins of antibiotic resistance. *Handb. Exp. Pharmacol.* **211**, 13–30 (2012).

28. Marshall, C. G., Lessard, I. A., Park, I. & Wright, G. D. Glycopeptide antibiotic resistance genes in glycopeptide-producing organisms. *Antimicrob. Agents Chemother.* **42**, 2215–20 (1998).

29. Wright, G. D. The antibiotic resistome: The nexus of chemical and genetic diversity. *Nature Reviews Microbiology* vol. 5 175–186 (2007).

30. Peterson, E. & Kaur, P. Antibiotic Resistance Mechanisms in Bacteria: Relationships Between Resistance Determinants of Antibiotic Producers, Environmental Bacteria, and Clinical Pathogens. *Front. Microbiol.* **9**, 2928 (2018).

31. Humeniuk, C. *et al.* Beta-lactamases of Kluyvera ascorbata, probable progenitors of some plasmid-encoded CTX-M types. *Antimicrob. Agents Chemother.* **46**, 3045–9 (2002).

32. Rodriguez, M. M., Ghiglione, B., Power, P., Naas, T. & Gutkind, G. Proposing Kluyvera georgiana as the Origin of the Plasmid-Mediated Resistance Gene fosA4. *Antimicrob. Agents Chemother.* **62**, e00710-18 (2018).

33. Cantón, R., González-Alba, J. M. & Galán, J. C. CTX-M Enzymes: Origin and Diffusion. *Front. Microbiol.* **3**, 110 (2012).

34. Wu, S. W., Dornbusch, K., Kronvall, G. & Norgren, M. Characterization and nucleotide sequence of a Klebsiella oxytoca cryptic plasmid encoding a CMY-type beta-lactamase: confirmation that the plasmid-mediated cephamycinase originated from the Citrobacter freundii AmpC beta-lactamase. *Antimicrob. Agents Chemother.* **43**, 1350–7 (1999).

35. Benveniste, R. & Davies, J. Aminoglycoside antibiotic-inactivating enzymes in actinomycetes similar to those present in clinical isolates of antibiotic-resistant bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 2276–80 (1973).

36. Philippon, A., Arlet, G. & Jacoby, G. A. Plasmid-determined AmpC-type beta-lactamases. *Antimicrob. Agents Chemother.* **46**, 1–11 (2002).

37. Papanicolaou, G. A., Medeiros, A. A. & Jacoby, G. A. Novel plasmid-mediated beta-lactamase (MIR-1) conferring resistance to oxyimino- and alpha-methoxy beta-lactams in clinical isolates of Klebsiella pneumoniae. *Antimicrob. Agents Chemother.* **34**, 2200–9 (1990).

38. Barnaud, G. *et al.* Salmonella enteritidis: AmpC plasmid-mediated inducible beta-lactamase (DHA-1) with an ampR gene from Morganella morganii. *Antimicrob. Agents Chemother.* **42**, 2352–8 (1998).

39. Nadjar, D. *et al.* Outbreak of Klebsiella pneumoniae producing transferable AmpC-type beta-lactamase (ACC-1) originating from Hafnia alvei. *FEMS Microbiol. Lett.* **187**, 35–40 (2000).

40. Mancini, S., Poirel, L., Kieffer, N. & Nordmann, P. Transposition of Tn1213 Encoding the PER-1 Extended-Spectrum β-Lactamase. *Antimicrob. Agents Chemother.* **62**, e02453-17 (2018).

41. Liu, S. L., Hessel, A. & Sanderson, K. E. Genomic mapping with I-Ceu I, an intron-encoded endonuclease specific for genes for ribosomal RNA, in Salmonella spp., Escherichia coli, and other bacteria. *Proc. Natl. Acad. Sci. U.*

S. A. **90**, 6874–6878 (1993).

42. Poirel, L., Figueiredo, S., Cattoir, V., Carattoli, A. & Nordmann, P. Acinetobacter radioresistens as a silent source of carbapenem resistance for Acinetobacter spp. *Antimicrob. Agents Chemother.* **52**, 1252–6 (2008).

43. Potron, A., Poirel, L. & Nordmann, P. Origin of OXA-181, an emerging carbapenem-hydrolyzing oxacillinase, as a chromosomal gene in Shewanella xiamenensis. *Antimicrob. Agents Chemother.* **55**, 4405–7 (2011).

44. Kitts, P. A. *et al.* Assembly: A resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44**, D73–D80 (2016).

45. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **41**, D8–D20 (2013).

46. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10 (1990).

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

48. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

49. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

50. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–9 (2006).

51. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

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

53. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699–2699 (2018).

54. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).

55. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32-6 (2006).

56. Behjati, S. & Tarpey, P. S. What is next generation sequencing? *Arch. Dis. Child. Educ. Pract. Ed.* **98**, 236–238 (2013).

57. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).

58. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics* vol. 13 278–289 (2015).

59. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics. *Nucleic Acids Research* vol. 46 2159–2168 (2018).

60. Lu, H., Giordano, F. & Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics and Bioinformatics* vol. 14 265–279 (2016).

61. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* vol. 21 1–16 (2020).

62. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nature Reviews Genetics* vol. 14 157–167 (2013).

63. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).

64. Liu, B. & Pop, M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **37**, D443-7 (2009).

65. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640 (2012).

66. Gupta, S. K. *et al.* ARG-annot, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.* **58**, 212–220 (2014).

67. Doster, E. *et al.* MEGARes 2.0: A database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Res.* **48**, D561–D569 (2020).

68. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

69. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–66 (2002).

70. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).

71. Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

72. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

73. Hugenholtz, P., Skarshewski, A. & Parks, D. H. Genome-based microbial taxonomy coming of age. *Cold Spring Harb. Perspect. Biol.* **8**, (2016).

74. Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* **43**, 6761–6771 (2015).

75. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

76. Jacoby, G. A., Griffin, C. M. & Hooper, D. C. Citrobacter spp. as a source of qnrB Alleles. *Antimicrob. Agents Chemother.* **55**, 4979–84 (2011).

77. Sisinthy, S., Chakraborty, D., Adicherla, H. & Gundlapally, S. R. Emended description of the family Chromatiaceae, phylogenetic analyses of the genera Alishewanella, Rheinheimera and Arsukibacterium, transfer of Rheinheimera longhuensis LH2-2T to the genus Alishewanella and description of Alishewanella alkalitolerans sp. *Antonie Van Leeuwenhoek* **110**, 1227–1241 (2017).

78. Aminov, R. I. & Mackie, R. I. Evolution and ecology of antibiotic resistance genes. *FEMS Microbiol. Lett.* **271**, 147–161 (2007).

79. Barlow, M. & Hall, B. G. Phylogenetic Analysis Shows That the OXA b-Lactamase Genes Have Been on Plasmids for Millions of Years. *J. Mol. Evol.* **55**, 314–321 (2002).

80. Wang, R. *et al.* The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nat. Commun.* **9**, 1179 (2018).

81. Waglechner, N. & Wright, G. D. Antibiotic resistance: it's bad, but why isn't it worse? *BMC Biol.* **15**, 84 (2017).

82. Yoon, E.-J. *et al.* Origin in Acinetobacter guillouiae and dissemination of the aminoglycoside-modifying enzyme Aph(3')-VI. *MBio* **5**, e01972-14 (2014).